Ann Inst Stat Math (2010) 62:209–234
DOI 10.1007/s10463-009-0237-1

Bias and variance reduction techniques for bootstrap information criteria

Genshiro Kitagawa · Sadanori Konishi

Received: 17 November 2008 / Revised: 23 January 2009 / Published online: 29 July 2009

Abstract We discuss the problem of constructing information criteria by applying the bootstrap methods. Various bias and variance reduction methods are presented for improving the bootstrap bias correction term in computing the bootstrap information criterion. The properties of these methods are investigated both in theoretical and numerical aspects, for which we use a statistical functional approach. It is shown that the bootstrap method automatically achieves the second-order bias correction if the bias of the first-order bias correction term is properly removed. We also show that the variance associated with bootstrapping can be considerably reduced for various model estimation procedures without any analytical argument. Monte Carlo experiments are conducted to investigate the performance of the bootstrap bias and variance reduction techniques.

Keywords Kullback–Leibler information · AIC · Information criteria · Bootstrapping · Statistical functional · Variance reduction · Higher-order bias correction

1 Introduction

Akaike information criterion (AIC), was introduced for the evaluation of various types of statistical models (Akaike 1973, 1974). It facilitated to evaluate and compare various

G. Kitagawa (✉)
The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan
e-mail: kitagawa@ism.ac.jp

S. Konishi
Faculty of Mathematics, Kyushu University, 6-10-1 Hakozaki, Fukuoka 812-8581, Japan
e-mail: konishi@math.kyushu-u.ac.jp

statistical models quite freely, and promoted the development of statistical models in various fields of statistics (e.g., Bozdogan 1994).

Although direct application of AIC is limited to the models with parameters estimated by the maximum likelihood methods, Akaike's basic idea of bias correction for the log-likelihood can be applied to a wider class of models defined by statistical functionals (Konishi and Kitagawa 1996, 2003, 2008). Another direction of the extension is based on the bootstrapping. In Ishiguro et al. (1997), the bias of the log-likelihood was estimated by using the bootstrap method (Efron 1979). By this method, we can apply the information criterion to various types of statistical models. This bootstrap approach can be trace back to Wong (1983) and Efron (1983) where they use the bootstrap bias correction for the log-likelihood in determining the kernel width. Bootstrapping log-likelihood is also considered in Cavanaugh and Shumway (1997), Davison and Hinkley (1992), and Shibata (1997).

The bootstrap information criteria have a significant merit that it can be applied to almost any type of models and estimation procedures under very weak assumption. Beside this merit, unlike other analytic information criteria such as TIC and GIC, it is not necessary to derive the bias correction term analytically for each specific model and estimation procedure. It should be emphasized here that the AIC is free from troublesome analytic derivation of the bias correction term in actual modeling and, in some sense, the EIC inherits this nice property.

On the other hand, the bias correction term of EIC inevitably suffers from the bootstrap sampling fluctuation. In its simplest definition, it can be shown that the variance of this fluctuation is proportional to the sample size. Therefore, unlike the usual estimation problems, the accuracy of the bias correction term gets worse as the sample size increases. Therefore, to make the bootstrap information criterion practical, it is indispensable to develop a computationally efficient methods for estimating the bootstrap bias correction term.

In this paper we first consider methods of increasing the accuracy of the bootstrap bias correction term from two different approaches. The first is concerned with the reduction of the bias of the correction term in estimating the expected log-likelihood. It will be shown that the bootstrap method automatically performs higher-order bias correction but the bias of the first-order correction term remains. It means that by removing the bias of the first-order bias correction term, bootstrap method automatically achieve second-order bias correction. This can be realized by bootstrapping first-order bias corrected log-likelihood or double bootstrapping. By numerical examples, we show that bootstrap method has less bias than the analytic method, although it suffers from the increase of variance due to bootstrapping.

The second approach is the reduction of the variance of the bootstrap estimate of the bias term. This is achieved by the decomposition of the bias correction term and by omitting a term which has zero mean and the largest variance. The decomposition method yields a significant variance reduction in bootstrap bias estimation. This clearly shows the advantage of the use of the variance reduction method in bootstrapping.

This paper is organized as follows. In Sect. 2 we give a brief review of constructing information criteria. Section 3 discusses the theoretical evaluation of the asymptotic accuracy of various information criteria as an estimator of the expected log-likelihood, using a statistical functional approach. Section 4 introduces the bootstrap bias corrected

version of a log-likelihood with theoretical justification. In Sect. 5 we present the variance reduction method for bootstrap simulation with theoretical improvement. In Sect. 6 we conduct Monte Carlo simulations to illustrate the efficiency of our variance reduction technique. Some concluding remarks are given in Sect. 7.

2 Preliminaries: A brief review of information criteria

Suppose that the data $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ are generated from an unknown true distribution function $G(x)$ with probability density $g(x)$. We regard $g(x)$ as a true probability mechanism generating the data. The objective of statistical modeling is to build a reasonable model based on the observed data. In practical situations it is difficult to estimate the true density $g(x)$ precisely from a finite number of observations. Hence we usually make use of an approximating model which consists of a family of probability densities $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ is the p-dimensional vector of unknown parameters and Θ is an open subset of R^p. The parametric model is estimated by finding suitable estimate, $\hat{\boldsymbol{\theta}}$, of unknown parameter vector $\boldsymbol{\theta}$ and replacing $\boldsymbol{\theta}$ in $f(x|\boldsymbol{\theta})$ by the estimate $\hat{\boldsymbol{\theta}}$. We call this probability density $f(x|\hat{\boldsymbol{\theta}})$ a statistical model.

The problem here is how to evaluate the goodness of the statistical model when it is used to predict a future observation z from the true density $g(z)$.

In order to assess the closeness of $f(z|\hat{\boldsymbol{\theta}})$ to the true density $g(z)$, we use the Kullback–Leibler information (Kullback and Leibler 1951)

$$\begin{aligned} I\{g(z); f(z|\hat{\boldsymbol{\theta}})\} &= E_{G(z)}\left[\log \frac{g(Z)}{f(Z|\hat{\boldsymbol{\theta}})}\right] \\ &= E_{G(z)}[\log g(Z)] - E_{G(z)}[\log f(Z|\hat{\boldsymbol{\theta}})], \end{aligned} \tag{1}$$

where expectation is taken over the distribution of z, conditional on the observed data $\boldsymbol{x}_n$. The KL-information $I\{g(z); f(z|\hat{\boldsymbol{\theta}})\}$ always takes a positive value, unless $f(z|\hat{\boldsymbol{\theta}}) = g(z)$ holds almost everywhere. We choose the model that minimizes $I\{g(z); f(z|\hat{\boldsymbol{\theta}})\}$ among candidate statistical models. We note that it is sufficient to consider only the second term on the right-hand side in Eq. (1), since the first term, $E_{G(z)}[\log g(Z)]$, is a constant that depends solely on the true distribution $g(z)$. The term

$$\eta(G; \hat{\boldsymbol{\theta}}) \equiv E_{G(z)}[\log f(Z|\hat{\boldsymbol{\theta}})] = \int \log f(z|\hat{\boldsymbol{\theta}}) dG(z) \tag{2}$$

is conditional on the observed data $\boldsymbol{x}_n$ through the estimator, $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)$, and depends on the unknown true distribution G. It is called the expected log-likelihood. The larger this value is for a statistical model, the smaller its Kullback–Leibler information is and hence the better the model is.

The expected log-likelihood still contains the unknown true distribution $g(z)$. Therefore, we cannot directly evaluate the expected log-likelihood but it can be estimated from the observations. It is known that an obvious estimator of $\eta(G; \hat{\boldsymbol{\theta}})$ is $(1/n$ of) the log-likelihood

$$\eta(\hat{G}; \hat{\boldsymbol{\theta}}) = \int \log f(z|\hat{\boldsymbol{\theta}}) d\hat{G}(z) = \frac{1}{n} \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}) \equiv \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}), \tag{3}$$

obtained by replacing the unknown distribution G in $\eta(G; \hat{\boldsymbol{\theta}})$ by the empirical distribution, $\hat{G}$, defined by putting mass $1/n$ on each observation x_α. Throughout this paper, because of the order of the expansion formula, we refer to the above equation divided by n as the log-likelihood.

In general, the log-likelihood provides an optimistic assessment (over-estimation) of the expected log-likelihood $\eta(G; \hat{\boldsymbol{\theta}})$, that is, $\eta(G; \hat{\boldsymbol{\theta}}) < \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}})$ always holds, because the same data are used both to estimate the parameters of the model and to evaluate $\eta(G; \hat{\boldsymbol{\theta}})$. We therefore consider the bias correction of the log-likelihood $\ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}})$.

The bias of the log-likelihood in estimating the expected log-likelihood is defined by

$$b(G) = E_{G(\boldsymbol{x})}\left[\ell_n(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}) - \int \log f(z|\hat{\boldsymbol{\theta}}) dG(z)\right], \tag{4}$$

where expectation is taken over the joint distribution of $\boldsymbol{x}_n$, that is, $\prod_{\alpha=1}^{n} dG(x_\alpha)$. If the bias can be estimated by an appropriate procedure, then an information criterion is defined as a bias corrected log-likelihood

$$\begin{aligned} \mathrm{IC}(\boldsymbol{x}_n; \hat{\theta}) &= -2n\left\{\ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) - \hat{b}(G)\right\} \\ &= -2\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}) + 2n\hat{b}(G), \end{aligned} \tag{5}$$

where $\hat{b}(G)$ is an estimator of the bias $b(G)$ in Eq. (4).

The bias correction term $\hat{b}(G)$ is usually given as an asymptotic bias in (4). According to the assumptions made on model estimation and the relationship between the specified model and the true density, the asymptotic bias takes a different form, and consequently we obtain various information criteria proposed previously including AIC.

The objective of constructing information criteria is to estimate the quantity (2) from observed data as accurately as possible. If there exists an estimator $\bar{\eta}(\hat{G}; \hat{\boldsymbol{\theta}})$ such that

$$E_{G(\boldsymbol{x})}\left[\bar{\eta}(\hat{G}; \hat{\boldsymbol{\theta}}) - \eta(G; \hat{\boldsymbol{\theta}})\right] = 0, \tag{6}$$

then $\bar{\eta}(\hat{G};\hat{\boldsymbol{\theta}})$ is an unbiased estimator of the expected log-likelihood $\eta(G;\hat{\boldsymbol{\theta}})$. In fact, Sugiura (1978) and Hurvich and Tsai (1989) gave such an unbiased estimator for a Gaussian regression model under the assumptions that (i) the parametric model is estimated by the method of maximum likelihood and (ii) the specified parametric family of probability distributions contains the true density, that is, $g(z) = f(z|\boldsymbol{\theta}_0)$; (i.e., $G = F_{\boldsymbol{\theta}_0}$) for some $\boldsymbol{\theta}_0$ in Θ.

In a general framework, it is however difficult to obtain an information criterion as an unbiased estimator of the expected log-likelihood. Hence, it is desirable to obtain an estimator $\eta_{\mathrm{IC}}(\hat{G};\hat{\boldsymbol{\theta}})$ of $\eta(G;\hat{\boldsymbol{\theta}})$ that satisfies the condition

$$E_{G(\boldsymbol{x})}\big[\eta_{\mathrm{IC}}(\hat{G};\hat{\boldsymbol{\theta}}) - \eta(G;\hat{\boldsymbol{\theta}})\big] = O(n^{-j}) \tag{7}$$

for j as large as possible. For example, if $j = 2$, (7) indicates that the estimator agrees up to a term of order $1/n$ in the target quantity $E_{G(\boldsymbol{x})}[\eta(G;\hat{\boldsymbol{\theta}})]$.

We recall that a random sample $\boldsymbol{x}_n$ is drawn from an unknown true distribution $G(x)$ with density $g(x)$ and that the true density $g(x)$ is in the neighborhood of the specified parametric family of probability densities $\{f(x|\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta \subset R^p\}$. Then the p-dimensional parameter vector $\boldsymbol{\theta}$ is estimated based on data from $G(x)$, not $f(x|\boldsymbol{\theta})$. Hence we employ a p-dimensional functional estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$, where $\boldsymbol{T}(\cdot) = (T_1(\cdot),\ldots,T_p(\cdot))^T \in R^p$ is a functional vector on the space of distribution functions and $\hat{G}$ is the empirical distribution function. For example, the sample mean $\bar{x}_n = \sum_{\alpha=1}^n x_\alpha/n$ can be written as $\bar{x}_n = T(\hat{G})$ for the functional given by $T(G) = \int x dG(x)$.

Within the framework of regular functional, Konishi and Kitagawa (1996) showed that the asymptotic bias of the log-likelihood in the estimation of the expected log-likelihood is given by

$$\begin{aligned} b(G) &= E_{G(\boldsymbol{x})}\left[\ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) - \eta(G;\hat{\boldsymbol{\theta}})\right] \\ &= \frac{1}{n}\mathrm{tr}\left\{\int \boldsymbol{T}^{(1)}(z;G)\left.\frac{\partial \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{T}(G)} dG(z)\right\} + O(n^{-2}), \end{aligned} \tag{8}$$

where $\boldsymbol{T}^{(1)}(z;G) = (T_1^{(1)}(z;G),\ldots,T_p^{(1)}(z;G))^T$ and $T_i^{(1)}(z;G)$ is the influence function defined by

$$T_i^{(1)}(z;G) = \lim_{\varepsilon\to 0}\frac{T_i((1-\varepsilon)G + \varepsilon\delta_z) - T_i(G)}{\varepsilon}, \tag{9}$$

with δ_z being a point mass at z. By replacing the unknown true probability distribution G with the empirical distribution function $\hat{G}$, we estimate the asymptotic bias, and then an information criterion based on the bias-corrected version of the log-likelihood is given by

$$\mathrm{GIC} = -2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) + \frac{2}{n}\sum_{\alpha=1}^n \mathrm{tr}\left\{\boldsymbol{T}^{(1)}(X_\alpha;\hat{G})\left.\frac{\partial \log f(x_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right|_{\hat{\boldsymbol{\theta}}}\right\}. \tag{10}$$

Now we consider a statistical model $f(x|\hat{\boldsymbol{\theta}}_{ML})$ estimated by the maximum likelihood methods, where $\hat{\boldsymbol{\theta}}_{ML}$ is a p-dimensional maximum likelihood estimator based on data generated from the true distribution $G(x)$. The maximum likelihood estimator, $\hat{\boldsymbol{\theta}}_{ML}$, given as the solution of the likelihood equation can be written as $\hat{\boldsymbol{\theta}}_{ML} = \boldsymbol{T}_{ML}(\hat{G})$, where $\boldsymbol{T}_{ML}$ is the p-dimensional functional implicitly defined by

$$\int \frac{\partial \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{T}_{ML}(G)} dG(z) = \mathbf{0}. \tag{11}$$

Replacing G in (11) by $G_\varepsilon = (1-\varepsilon)G + \varepsilon\, \delta_z$ and differentiating with respect to ε yield the p-dimensional influence function of the maximum likelihood estimator in the form

$$\boldsymbol{T}^{(1)}_{ML}(z; G) = J(G)^{-1} \frac{\partial \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{T}_{ML}(G)}, \tag{12}$$

where $J(G)$ is given by

$$J(G) = -\int \frac{\partial^2 \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{T}(G)} dG(z). \tag{13}$$

Substituting the influence function (12) into the Eq. (8), we have the asymptotic bias of the log-likelihood $\ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_{ML})$ in the form

$$b(G) = \frac{1}{n}\mathrm{tr}\{J(G)^{-1} I(G)\} + O(n^{-2}), \tag{14}$$

where

$$I(G) = \int \frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{T}_{ML}(G)} dG(z). \tag{15}$$

By correcting the asymptotic bias estimate, an information criterion for evaluating a statistical model estimated by the maximum likelihood method is given by

$$\mathrm{TIC} = -2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_{ML}) + 2\mathrm{tr}\{J(\hat{G})^{-1} I(\hat{G})\}. \tag{16}$$

This criterion was originally introduced by Takeuchi (1976) (see also Stone 1977).

If the parametric model is estimated by the maximum likelihood methods and the true distribution belongs to the parametric family of densities $\{f(x|\boldsymbol{\theta});\ \boldsymbol{\theta} \in \boldsymbol{\theta}\}$, we have the well-known result $I(F_{\boldsymbol{\theta}}) = J(F_{\boldsymbol{\theta}})$, where $F_{\boldsymbol{\theta}}$ is the distribution function of $f(x|\boldsymbol{\theta})$. Then the asymptotic bias is given by $\mathrm{tr}\left\{J(F_{\boldsymbol{\theta}})^{-1} I(F_{\boldsymbol{\theta}})\right\} = p$, and consequently we have Akaike (1973, 1974) information criterion, known as AIC, in the form:

$$\mathrm{AIC} = -2\log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}_{ML}) + 2p, \tag{17}$$

where p is the number of estimated parameters within the model.

It is worth pointing out that the asymptotic bias correction term in AIC does not depend on any unknown parameters and has no variability. Also AIC may be applied in an automatic way in various situations for the evaluation of statistical models estimated by the maximum likelihood methods.

3 Asymptotic accuracy of information criteria

Information criteria were constructed as approximately unbiased estimators of the expected log-likelihood $\eta(G;\hat{\boldsymbol{\theta}}) = E_{G(z)}[\log f(Z|\hat{\boldsymbol{\theta}})]$ or, equivalently, the Kullback–Leibler information discrepancy between the true distribution $g(z)$ and a statistical model $f(z|\hat{\boldsymbol{\theta}})$ from a predictive point of view. In this section we discuss, in a general framework, the theoretical evaluation of the asymptotic accuracy of various types of information criteria as an estimator of the expected log-likelihood, based on the functional approach developed by Konishi and Kitagawa (1996), Konishi and Kitagawa (2003), and Konishi and Kitagawa (2008).

The expected log-likelihood is conditional on the observed data $\boldsymbol{x}_n$ through $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)$ and also depends on the unknown true distribution $G(z)$ generating the data. It can be shown (see Konishi and Kitagawa 2008, Chapter 7) that under certain regularity conditions, the expectation of $\eta(G;\hat{\boldsymbol{\theta}})$ over the sampling distribution G of $\boldsymbol{X}_n$ is expanded in the form

$$\begin{aligned} E_{G(\boldsymbol{x})}\left[\eta(G;\hat{\boldsymbol{\theta}})\right] &= E_{G(\boldsymbol{x})}\left[E_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}})\right]\right] \\ &= \int \log f(z|\boldsymbol{T}(G))dG(z) + \frac{1}{n}\eta_1(G) + \frac{1}{n^2}\eta_2(G) + O(n^{-3}), \end{aligned} \tag{18}$$

where

$$\eta_1(G) = \boldsymbol{s}(G)^T \int \left.\frac{\partial \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{T}(G)} dG(z) - \frac{1}{2}\mathrm{tr}\{J(G)\Sigma(G)\}. \tag{19}$$

Here, $\boldsymbol{s}(G)$ and $\Sigma(G)$ are, respectively, the asymptotic bias and variance-covariance matrix of the p dimensional estimator $\hat{\boldsymbol{\theta}}$ given by

$$\begin{aligned} E_{G(\boldsymbol{x})}\left[\hat{\boldsymbol{\theta}} - \boldsymbol{T}(G)\right] &= \frac{1}{n}\boldsymbol{s}(G) + O(n^{-2}), \\ E_{G(\boldsymbol{x})}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{T}(G))(\hat{\boldsymbol{\theta}} - \boldsymbol{T}(G))^T\right] &= \frac{1}{n}\Sigma(G) + O(n^{-2}) \end{aligned} \tag{20}$$

and $J(G)$ is defined by Eq. (13).

On the other hand, since the log-likelihood $\ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}})$ in Eq. (3) being an estimator of the expected log-likelihood $\eta(G;\hat{\boldsymbol{\theta}})$, the expectation of the log-likelihood gives a valid expansion of the following form:

$$E_{G(\boldsymbol{x})}\left[\ell_n(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}})\right] = \int \log f(z|\boldsymbol{T}(G))dG(z) + \frac{1}{n}L_1(G) + \frac{1}{n^2}L_2(G) + O(n^{-3}), \tag{21}$$

where

$$L_1(G) = \mathrm{tr}\left\{\int \boldsymbol{T}^{(1)}(z;G)\left.\frac{\partial \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{T}(G)} dG(z)\right\} + \eta_1(G), \tag{22}$$

with $\eta_1(G)$ given by (19). Hence, comparing the Eq. (18) with Eq. (21), the log-likelihood as an estimator of the expected log-likelihood only agrees in the first term, and the term of order $1/n$ remains as a bias. This implies that the bias of the log-likelihood in the estimation of the expected log-likelihood can be expanded as

$$b(G) = E_{G(\boldsymbol{x})}\left[\ell_n(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}) - \eta(G;\hat{\boldsymbol{\theta}})\right] = \frac{1}{n}b_1(G) + \frac{1}{n^2}b_2(G) + O(n^{-3}), \tag{23}$$

where

$$\begin{aligned} b_1(G) &= L_1(G) - \eta_1(G) = \mathrm{tr}\left\{\int \boldsymbol{T}^{(1)}(z;G)\left.\frac{\partial \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right|_{\boldsymbol{T}(G)} dG(z)\right\}, \\ b_2(G) &= L_2(G) - \eta_2(G). \end{aligned} \tag{24}$$

By replacing the unknown true probability distribution G with the empirical distribution function $\hat{G}$, we estimate the asymptotic bias, and then an information criterion based on the asymptotic bias-corrected version of the log-likelihood is given by

$$\eta_{\mathrm{GIC}}(\hat{G};\hat{\boldsymbol{\theta}}) = \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) - \frac{1}{n}b_1(\hat{G}). \tag{25}$$

Noting that the difference between $E_G[b_1(\hat{G})]$ and $b_1(G)$ is usually of order n^{-1}, that is,

$$\begin{aligned} E_{G(\boldsymbol{x})}[b_1(\hat{G})] &= b_1(G) + \frac{1}{n}\Delta b_1(G) + O(n^{-2}) \\ &= L_1(G) - \eta_1(G) + \frac{1}{n}\Delta b_1(G) + O(n^{-2}), \end{aligned} \tag{26}$$

we have

$$E_{G(\boldsymbol{x})}\left[\eta_{\mathrm{GIC}}(\hat{G};\hat{\boldsymbol{\theta}})\right] = \int \log f(z|\boldsymbol{T}(G))dG(z) + \frac{1}{n}\eta_1(G) + \frac{1}{n^2}\{L_2(G) - \Delta b_1(G)\} + O(n^{-3}), \tag{27}$$

and

$$E_{G(\boldsymbol{x})}\left[\eta_{\mathrm{GIC}}(\hat{G};\hat{\boldsymbol{\theta}})-\eta(G;\hat{\boldsymbol{\theta}})\right]=\frac{1}{n^2}\{b_2(G)-\Delta b_1(G)\}+O(n^{-3}). \tag{28}$$

Hence the bias corrected log-likelihood $\eta_{\mathrm{GIC}}(\hat{G};\hat{\boldsymbol{\theta}})$ is second-order correct or accurate for the expected log-likelihood $\eta(G;\hat{\boldsymbol{\theta}})$ in the sense that the expectations of two quantities are in agreement up to and including the term of order n^{-1}, and that the order of the remainder is n^{-2}. It also implies that $\eta_{\mathrm{GIC}_2}(\hat{G};\hat{\boldsymbol{\theta}})$ defined by

$$\eta_{\mathrm{GIC}_2}(\hat{G};\hat{\boldsymbol{\theta}})=\ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}})-\frac{1}{n}b_1(\hat{G})-\frac{1}{n^2}\left\{b_2(\hat{G})-\Delta b_1(\hat{G})\right\} \tag{29}$$

is third-order correct.

If the specified parametric family of densities includes the true distribution and the maximum likelihood method is used to estimate the underlying density, then the asymptotic bias of the log-likelihood is the number of estimated parameters, giving $\mathrm{AIC}=-2n\{\eta_{\mathrm{AIC}}(\hat{F};\hat{\boldsymbol{\theta}}_{ML})-p/n\}$ with second-order accuracy for $\eta(F;\hat{\boldsymbol{\theta}}_{ML})$. Moreover the bias-corrected version of the log-likelihood defined by

$$\eta_{\mathrm{AIC}_2}(\hat{F};\hat{\boldsymbol{\theta}}_{ML})=\eta(\hat{F};\hat{\boldsymbol{\theta}}_{ML})-\frac{1}{n}p-\frac{1}{n^2}b_2(\hat{F}) \tag{30}$$

is third-order correct for $\eta(F;\hat{\boldsymbol{\theta}}_{ML})$, since it can be readily checked that

$$\mathrm{E}_{F(\boldsymbol{x})}\left[\eta_{\mathrm{AIC}_2}(\hat{F};\hat{\boldsymbol{\theta}}_{ML})-\eta(F;\hat{\boldsymbol{\theta}}_{ML})\right]=O(n^{-3}). \tag{31}$$

4 Bootstrap information criterion

In practice, we need to derive the bias correction terms analytically for each estimator. The bootstrap method offers an alternative approach to estimate the biases numerically.

In the bootstrap methods, the true distribution $G(x)$ is first estimated by an empirical distribution function $\hat{G}(x)$. A random sample from the empirical distribution function $\hat{G}(x)$ is referred to as a bootstrap sample and is denoted as $\boldsymbol{x}_n^*=\{x_1^*,x_2^*,\ldots,x_n^*\}$. Then a statistical model $f(x|\hat{\boldsymbol{\theta}}^*)$ is constructed based on the bootstrap sample $\boldsymbol{x}_n^*$, with $\hat{\boldsymbol{\theta}}^*=\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n^*)$.

The expected log-likelihood of the statistical model $f(x|\hat{\boldsymbol{\theta}}^*)$ when the empirical distribution function $\hat{G}(x)$ is considered as the true distribution is given by

$$\begin{aligned}E_{\hat{G}(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}^*)\right]&=\int \log f(z|\hat{\boldsymbol{\theta}}^*)d\hat{G}(z)\\&=\frac{1}{n}\sum_{\alpha=1}^{n}\log f(x_\alpha|\hat{\boldsymbol{\theta}}^*)\equiv\ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}^*).\end{aligned} \tag{32}$$

On the other hand, the log-likelihood, which is an estimator of the expected log-likelihood, is defined by re-using the bootstrap sample $\boldsymbol{x}_n^*$ as follows:

$$E_{\hat{G}^*(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}^*)\right] = \int \log f(z|\hat{\boldsymbol{\theta}}^*) d\hat{G}^*(z) = \frac{1}{n}\sum_{\alpha=1}^{n} \log f(x_\alpha^*|\hat{\boldsymbol{\theta}}^*) \equiv \ell_n(\boldsymbol{x}_n^*|\hat{\boldsymbol{\theta}}^*), \tag{33}$$

where $\hat{G}^*(z)$ is an empirical distribution function based on the bootstrap sample $\boldsymbol{x}_n^*$. Consequently, the bootstrap estimate of the bias of the log-likelihood in estimating the expected log-likelihood is

$$b^*(\hat{G}) = E_{\hat{G}(\boldsymbol{x}^*)}\left[\ell_n(\boldsymbol{x}_n^*|\hat{\boldsymbol{\theta}}^*) - \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}^*)\right] = \int\cdots\int \left\{\ell_n(\boldsymbol{x}_n^*|\hat{\boldsymbol{\theta}}^*) - \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}^*)\right\} \prod_{\alpha=1}^{n} d\hat{G}(x_\alpha^*). \tag{34}$$

Then the bootstrap bias corrected version of the log-likelihood is defined by

$$\eta_{\mathrm{EIC}}(\hat{G};\hat{\boldsymbol{\theta}}) = \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) - b^*(\hat{G}). \tag{35}$$

It follows from (23) that the bootstrap bias estimate of the log-likelihood in the estimation of the expected log-likelihood can be expanded as

$$b^*(\hat{G}) = E_{\hat{G}(\boldsymbol{x}^*)}\left[\ell_n(\boldsymbol{x}_n^*|\hat{\boldsymbol{\theta}}^*) - \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}^*)\right] = \frac{1}{n}b_1(\hat{G}) + \frac{1}{n^2}b_2(\hat{G}) + O(n^{-3}), \tag{36}$$

where

$$b_1(\hat{G}) = L_1(\hat{G}) - \eta_1(\hat{G}), \quad b_2(\hat{G}) = L_2(\hat{G}) - \eta_2(\hat{G}). \tag{37}$$

Taking expectation of $\eta_{\mathrm{EIC}}(\hat{G};\hat{\boldsymbol{\theta}})$ over the sampling distribution of $\boldsymbol{x}_n$, and using $E_{G(\boldsymbol{x})}[b^*(\hat{G})] = E_{G(\boldsymbol{x})}[b(\hat{G})]$, we have

$$E_{G(\boldsymbol{x})}\left[\eta_{\mathrm{EIC}}(\hat{G};\hat{\boldsymbol{\theta}})\right] = \int \log f(z|\boldsymbol{T}(G)) dG(z) + \frac{1}{n}\eta_1(G) + \frac{1}{n^2}\left\{\eta_2(G) - \Delta b_1(G)\right\} + O(n^{-3}), \tag{38}$$

and thus

$$E_{G(\boldsymbol{x})}\left[\eta_{\mathrm{EIC}}(\hat{G};\hat{\boldsymbol{\theta}}) - \eta(G;\hat{\boldsymbol{\theta}})\right] = -\frac{1}{n^2}\Delta b_1(G) + O(n^{-3}). \tag{39}$$

On the other hand, as shown in the previous section, the information criterion GIC, correcting the asymptotic bias $b_1(\hat{G})$, yields the corresponding result

$$E_{G(\boldsymbol{x})}\left[\eta_{\mathrm{GIC}}(\hat{G};\hat{\boldsymbol{\theta}})\right] = \int \log f(z|\boldsymbol{T}(G))dG(z) + \frac{1}{n}\eta_1(G) + \frac{1}{n^2}\{L_2(G) - \Delta b_1(G)\} + O(n^{-3}). \quad (40)$$

Therefore, we have

$$E_{G(\boldsymbol{x})}\left[\eta_{\mathrm{GIC}}(\hat{G};\hat{\boldsymbol{\theta}}) - \eta(G;\hat{\boldsymbol{\theta}})\right] = \frac{1}{n^2}\{b_2(G) - \Delta b_1(G)\} + O(n^{-3}). \quad (41)$$

By comparing the target quantity, that is, the expectation of the expected log-likelihood $\eta(G;\hat{\boldsymbol{\theta}})$ given by Eq. (18), we know that GIC and the bootstrap bias corrected log-likelihood are both second-order accurate for the expected log-likelihood. It is, however, noticed from (39) and (41) that the term of order n^{-2} in EIC seems to be smaller than that of GIC, if $|\Delta b_1(G)| < |b_2(G) - \Delta b_1(G)|$. We investigate this finding through Monte Carlo simulation study in Sect. 6.

The most significant feature of the use of the bootstrap method is that the integral in Eq. (34) can be approximated numerically by the Monte Carlo method. Let us extract B sets of bootstrap samples of size n and write the ith bootstrap sample as $\boldsymbol{x}_n^*(i) = \{x_1^*(i), x_2^*(i), \ldots, x_n^*(i)\}$. We denote the difference between (32) and (33) with respect to the sample $\boldsymbol{x}_n^*(i)$ as

$$D^*(i) = \ell_n(\boldsymbol{x}_n^*(i)|\hat{\boldsymbol{\theta}}^*(i)) - \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}^*(i)), \quad (42)$$

where $\hat{\boldsymbol{\theta}}^*(i)$ is an estimate of $\boldsymbol{\theta}$ obtained from the ith bootstrap sample. Then, the expectation in (34) based on B bootstrap samples can be numerically approximated as

$$b^*(\hat{G}) \approx \frac{1}{B}\sum_{i=1}^{B} D^*(i) \equiv b_B(\hat{G}). \quad (43)$$

The quantity $b_B(\hat{G})$ is the bootstrap estimate of the bias $b(G)$ of the log-likelihood. Consequently, the bootstrap methods yield an information criterion as follows:

$$\eta_{\mathrm{EIC}}(\hat{G};\hat{\boldsymbol{\theta}}) = \ell_n(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}) - b_B(\hat{G}). \quad (44)$$

The information criterion based on the bootstrap method was referred to as the extended information criterion (EIC) by Ishiguro et al. (1997).

It might be noticed that the EIC suffers from two types of variances. The first is the variance of the bootstrap estimate $b^*(\hat{G})$ due to the sample or the empirical

distribution function $\hat{G}$ and is common to all other information criteria unless it is independent from $\hat{G}$ like AIC. The second is the bootstrap variance, caused by the bootstrap simulation,

$$\mathrm{Var}\left\{b_B(\hat{G})\right\} = E_{\hat{G}(\boldsymbol{x}^*)}\left[\left\{(b_B(\hat{G}) - b^*(\hat{G})\right\}^2\right], \tag{45}$$

which is inverse proportional to the number of bootstrap resampling, B, i.e., $B^{-1}\mathrm{V}^*$ where

$$\mathrm{V}^* = \mathrm{Var}\left\{D^*(i)\right\} = \frac{1}{B}\sum_{i=1}^{B}\left\{D^*(i) - b_B(\hat{G})\right\}^2. \tag{46}$$

We shall discuss a method of reducing this bootstrap variance in the following sections.

5 Efficient bootstrap simulation

5.1 Variance reduction for bootstrap simulation

The bootstrap method can be applied without analytically cumbersome procedures under very weak assumptions, that is, the estimator is invariant with respect to the re-ordering of the sample. In applying the bootstrap method, however, care should be paid to the magnitude of the fluctuation due to bootstrap simulation and approximation error, in addition to the sample fluctuation of the bias estimate itself.

For a set of given observations, the bootstrap bias approximation $b_B(\hat{G})$ in (43) converges to the bootstrap bias estimate $b^*(\hat{G})$ in (34), with probability one, when the number of bootstrap resampling B goes to infinity. However, because simulation errors occur for finite B, a procedure must be devised to reduce the error. This can be considered reduction of simulation error for $b_B(\hat{G})$ for a given sample. The variance reduction method described in this section, called the efficient bootstrap simulation method or the efficient resampling method, provides an effective, yet extremely simple method of reducing any fluctuation in the bootstrap bias estimation of log-likelihood. The variance of the bootstrap estimate of the bias defined in (45) can be significantly reduced by the decomposition of the bias term (Konishi and Kitagawa 1996, 2008).

Let $D(\boldsymbol{X}_n; G)$ be the difference between the log-likelihood and the expected log-likelihood of a statistical model. Then $D(\boldsymbol{X}_n; G)$ can be decomposed into three terms as follows:

$$\begin{aligned} D(\boldsymbol{X}_n; G) &= \ell_n(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}) - \int \log f(z|\hat{\boldsymbol{\theta}})dG(z) \\ &= D_1(\boldsymbol{X}_n; G) + D_2(\boldsymbol{X}_n; G) + D_3(\boldsymbol{X}_n; G) \end{aligned} \tag{47}$$

where

$$
\begin{aligned}
D_1(\boldsymbol{X}_n; G) &= \ell_n(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}) - \ell_n(\boldsymbol{X}_n|\boldsymbol{T}(G)),\\
D_2(\boldsymbol{X}_n; G) &= \ell_n(\boldsymbol{X}_n|\boldsymbol{T}(G)) - \int \log f(z|\boldsymbol{T}(G))dG(z),\\
D_3(\boldsymbol{X}_n; G) &= \int \log f(z|\boldsymbol{T}(G))dG(z) - \int \log f(z|\hat{\boldsymbol{\theta}})dG(z)
\end{aligned}
\tag{48}
$$

with $\hat{\boldsymbol{\theta}}$ being the functional estimator such that $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$.

For a general estimator $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\hat{G})$ defined by a statistical functional, the means and the variances of these terms are given by

$$
\begin{aligned}
&E_{G(\boldsymbol{x})}\left[D_1(\boldsymbol{X}_n; G)\right] = \frac{1}{n}\left[c + \boldsymbol{s}^T\boldsymbol{v} - \frac{1}{2}\mathrm{tr}\{\Sigma(G)J(G)\}\right] + O(n^{-2})\\
&E_{G(\boldsymbol{x})}\left[D_3(\boldsymbol{X}_n; G)\right] = -\frac{1}{n}\left[\boldsymbol{s}^T\boldsymbol{v} - \frac{1}{2}\mathrm{tr}\{\Sigma(G)J(G)\}\right] + O(n^{-2})\\
&\mathrm{Var}\left\{D_1(\boldsymbol{X}_n; G)\right\} = \mathrm{Var}\left\{D_3(\boldsymbol{X}_n; G)\right\} = \frac{1}{n}\boldsymbol{v}^T\Sigma(G)\boldsymbol{v} + O(n^{-2})\\
&E_{G(\boldsymbol{x})}\left[D_2(\boldsymbol{X}_n; G)\right] = 0, \quad \mathrm{Var}\left\{D_2(\boldsymbol{X}_n; G)\right\} = \frac{a}{n} + O(n^{-2}),
\end{aligned}
\tag{49}
$$

where $\boldsymbol{s}$ and $\Sigma(G)$ are, respectively, the asymptotic bias and variance-covariance matrix of the estimator $\hat{\boldsymbol{\theta}}$ given by Eq. (20) and

$$
\begin{aligned}
&c = \mathrm{tr}\left\{\int \boldsymbol{T}^{(1)}(z; G)\frac{\partial \log f(z|\boldsymbol{T}(G))}{\partial \boldsymbol{\theta}^T}dG(z)\right\}, \quad \boldsymbol{v} = \int \frac{\partial \log f(z|\boldsymbol{T}(G))}{\partial \boldsymbol{\theta}}dG(z),\\
&a = \int \left\{\log f(z|\boldsymbol{T}(G))\right\}^2 dG(z) - \left\{\int \log f(z|\boldsymbol{T}(G))dG(z)\right\}^2,
\end{aligned}
$$

(see Appendix).

For the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML} = \boldsymbol{T}_{ML}(\hat{G})$, noting that $\boldsymbol{v} = 0$ and $c = \mathrm{tr}\{J(G)^{-1}I(G)\}$, the corresponding results are given by

$$
\begin{aligned}
&E_{G(\boldsymbol{x})}\left[D_1(\boldsymbol{X}_n; G)\right] = E_{G(\boldsymbol{x})}\left[D_3(\boldsymbol{X}_n; G)\right] = \frac{1}{2n}\mathrm{tr}\left\{\Sigma(G)J(G)\right\} + O(n^{-2}),\\
&\mathrm{Var}\left\{D_1(\boldsymbol{X}_n; G)\right\} = \mathrm{Var}\left\{D_3(\boldsymbol{X}_n; G)\right\} = O(n^{-2}),\\
&E_{G(\boldsymbol{x})}\left[D_2(\boldsymbol{X}_n; G)\right] = 0, \quad \mathrm{Var}\left\{D_2(\boldsymbol{X}_n; G)\right\} = \frac{a}{n} + O(n^{-2}),
\end{aligned}
\tag{50}
$$

where $\Sigma(G) = J(G)^{-1}I(G)J(G)^{-1}$.

We observe that in both cases, the expectation of $D_2(\boldsymbol{X}_n; G)$ is zero and that the bootstrap estimate $E_{\hat{G}(\boldsymbol{x}^*)}[D(\boldsymbol{X}_n^*; \hat{G})]$ is the same as

$$
E_{\hat{G}(\boldsymbol{x}^*)}\left[D_1(\boldsymbol{X}_n^*; \hat{G}) + D_3(\boldsymbol{X}_n^*; \hat{G})\right], \tag{51}
$$

where

$$D_1(\boldsymbol{X}_n^*; \hat{G}) = \ell_n(\boldsymbol{X}_n^*|\hat{\boldsymbol{\theta}}^*) - \ell_n(\boldsymbol{X}_n^*|\hat{\boldsymbol{\theta}}),$$
$$D_3(\boldsymbol{X}_n^*; \hat{G}) = \ell_n(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}) - \ell_n(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}^*).$$

It is also worth noting that $\mathrm{Var}\{D(\boldsymbol{X}_n; G)\} = O(n^{-1})$ and in contrast

$$\mathrm{Var}\{D_1(\boldsymbol{X}_n; G) + D_3(\boldsymbol{X}_n; G)\} = O(n^{-2}), \tag{52}$$

(see (81) and (78) in the Appendix). Therefore, the use of the decomposition in Eq. (51) yields a significant reduction of the variance not only for the maximum likelihood estimators but also for any estimators defined by statistical functional.

Therefore, instead of $\frac{1}{B}\sum_{i=1}^{B} D(\boldsymbol{X}_n^*(i); \hat{G})$, we may use

$$b_B(\hat{G}) = \frac{1}{B}\sum_{i=1}^{B}\left\{D_1(\boldsymbol{X}_n^*(i); \hat{G}) + D_3(\boldsymbol{X}_n^*(i); \hat{G})\right\}$$

as a bootstrap bias estimate.

5.2 Higher-order bootstrap bias correction

An information criterion that yields more refined results may be obtained by deriving a second-order bias correction term in Eq. (28). In practical situations the bootstrap method offers an approach to estimate it numerically. If the asymptotic bias $b_1(G)$ is evaluated analytically, then the bootstrap estimate of the second-order bias correction term can be obtained by using

$$\frac{1}{n}b_2^*(\hat{G}) = E_{\hat{G}(\boldsymbol{x}^*)}\left[\log f(\boldsymbol{X}_n^*|\hat{\boldsymbol{\theta}}^*) - b_1(\hat{G}) - nE_{\hat{G}(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}^*)\right]\right]. \tag{53}$$

On the other hand, in situations where it is difficult to analytically determine the first-order correction term $b_1(G)$, an estimate of the second-order correction term can be obtained by employing the following two-step bootstrap methods:

$$\frac{1}{n}b_2^{**}(\hat{G}) = E_{\hat{G}(\boldsymbol{x}^*)}\left[f(\boldsymbol{X}_n^*|\hat{\boldsymbol{\theta}}^*) - b_B^*(\hat{G}) - nE_{\hat{G}(z^*)}\log f(Z^*|\hat{\boldsymbol{\theta}}^*)\right], \tag{54}$$

where $b_B^*(\hat{G})$ is the bootstrap estimate of the first-order correction term obtained by (43). This bias estimate gives the third-order bias corrected information criterion.

Theoretically, we may obtain the higher-order bias corrected information criterion by using the bootstrap repeatedly. However, care should be paid to the large variability due to the bootstrap simulations and also the approximation errors.

It is worth noting that if the AIC gives exact asymptotic bias correction term, we can automatically attain the second order bias correction by bootstrapping $D_1(X_n^*; \hat{G}) + D_3(X_n^*; \hat{G})$.

6 Numerical examples

We conducted Monte Carlo simulations to illustrate the effect of the variance reduction method and the higher order bias correction for the log-likelihood of a statistical model. As a model we considered a parametric family of normal distributions $f(x|\mu, \sigma^2)$ with mean μ and variance σ^2. Sample data were generated from the standard normal or Laplace distributions.

6.1 Variance reduction in bootstrapping

In our Monte Carlo experiments, we consider three situations: (i) the specified model contains the true distribution, (ii) the specified model does not contain the true distribution, and (iii) the model parameters are estimated by procedures other than the maximum likelihood method. In each table, the notations stand for the following:

$$C_1^* = nE_{\hat{G}}[D_1(\boldsymbol{X}_n^*; \hat{G})], \quad C_2^* = nE_{\hat{G}}[D_2(\boldsymbol{X}_n^*; \hat{G})], \quad C_3^* = nE_{\hat{G}}[D_3(\boldsymbol{X}_n^*; \hat{G})],$$
$$C_{123}^* = C_1^* + C_2^* + C_3^*, \quad C_{13}^* = C_1^* + C_3^*.$$

In Tables 1–3, the values in the first row for each sample size n show the averages of the bootstrap estimates over S samples for $C_{123}^*, C_{13}^*, C_1^*, C_2^*, C_3^*$. The second row with parenthesis shows the average of the bootstrap variances with B resamplings, i.e., $B^{-1}\sum_{i=1}^{B}(D^*(i) - b_B(\hat{G}))^2$ over S repetitions, where $D^*(i)$ and $b_B(\hat{G})$ are defined in (42) and (43), respectively. The variance of $b_B(\hat{G})$ as an estimate of $b^*(\hat{G})$ is inverse proportional to the number of bootstrap resampling, B. Namely, in actual computation, the bootstrap variance from B resamplings is obtained by dividing these values by B. On the other hand, the third row with parenthesis shows the variance of the bootstrap estimate $b^*(\hat{G})$ in S repetitions when B is set to 1,000. Note that these variances include both the fluctuation caused by bootstrap approximation and the fluctuation caused by the observed data, namely $\hat{G}(x)$.

To reduce the effect of the sample data, random samples $\boldsymbol{X}_n$ were generated for $S = 1{,}000{,}000$ times and the average of these estimates were computed, for sample size $n = 25$, 100, 400 and 1,600.

Example 1 (True distribution: Normal) Random samples were generated from the standard normal distribution, $N(0, 1)$. This is a typical situation where the specified model contains the true distribution. Note that in this case, the exact biases can be easily evaluated analytically and are given by $nb(G) = 2n/(n-3)$, i.e., 2.273, 2.063, 2.014 and 2.003 for $n = 25$, 100, 400 and 1,600, respectively.

It can be seen from Table 1 that the means of C_{123}^* and C_{13}^* accord each other. However the variances of C_{123}^* are significantly larger than those of C_{13}^*, in particular

Table 1 Bootstrap estimates of biases for sample sizes $n = 25$, 100, 400 and 1,600 when the normal distribution model is fitted to data generated by a standard normal distribution

n	C^*_{123}	C^*_{13}	C^*_1	C^*_2	C^*_3	$\mathrm{tr}(\hat{I}\hat{J}^{-1})$
25	2.227	2.227	0.999	0.000	1.228	1.885
	(24.5021)	(8.6477)	(1.1913)	(11.0499)	(3.6890)	–
	(0.2891)	(0.2573)	(0.0405)	(0.0111)	(0.1044)	(0.1337)
100	2.038	2.037	0.996	0.000	1.041	1.970
	(56.63)	(4.6681)	(1.0425)	(48.4716)	(1.3379)	–
	(0.1128)	(0.0605)	(0.0138)	(0.0485)	(0.0165)	(0.0517)
400	2.008	2.008	0.999	0.000	1.009	1.993
	(205.63)	(4.1491)	(1.0108)	(198.34)	(1.0730)	–
	(0.2202)	(0.0188)	(0.0046)	(0.1982)	(0.0048)	(0.0145)
1600	2.002	2.002	1.000	0.000	1.002	1.998
	(804.76)	(4.0330)	(1.0019)	(797.69)	(1.0168)	–
	(0.8099)	(0.0077)	(0.0019)	(0.7989)	(0.0019)	(0.004)

Table 2 Bootstrap estimates of biases for sample sizes $n = 25$, 100, 400 and 1,600 when the normal distribution model is fitted to data generated by Laplace distribution

n	C^*_{123}	C^*_{13}	C^*_1	C^*_2	C^*_3	$\mathrm{tr}(\hat{I}\hat{J}^{-1})$
25	3.433	3.433	1.413	0.000	2.020	2.594
	(70.1861)	(33.6816)	(2.7615)	(19.9007)	(18.8737)	–
	(3.4377)	(3.4018)	(0.2929)	(0.0199)	(1.7401)	(0.7692)
100	3.330	3.330	1.597	0.000	1.733	3.165
	(137.69)	(16.4109)	(3.2991)	(108.07)	(5.3728)	–
	(1.3325)	(1.2095)	(0.2453)	(0.1082)	(0.3668)	(0.9892)
400	3.430	3.430	1.700	0.000	1.730	3.402
	(502.92)	(14.9768)	(3.6197)	(479.90)	(3.9659)	–
	(1.0015)	(0.5133)	(0.1252)	(0.4806)	(0.1316)	(0.5150)
1600	3.479	3.479	1.736	0.000	1.743	3.474
	(1996.68)	(14.6785)	(3.6472)	(1977.28)	(3.7145)	–
	(2.1575)	(0.1760)	(0.0440)	(1.9757)	(0.0440)	(0.1650)

for large sample size n. This is due to the properties shown in (50) that the variance of C^*_2 is proportional to the sample size n, and in this case $nE_{G(\boldsymbol{x})}[\{\log f(\boldsymbol{X}_n|\boldsymbol{T}(G)) - E_{G(z)} \log f(Z|\boldsymbol{T}(G))\}^2] = n/2$, whereas the variance of C^*_{13} is of order $O(1)$ and converges to a constant as n gets large. It can be also seen that, as n becomes larger, the bootstrap variances of both C^*_1 and C^*_3 converge to one, the asymptotic variances of C_1 and C_3. The variance of C^*_2 is close to $n/2$ and clearly shows the efficiency of using C^*_{13} instead of C^*_{123} in the bootstrap bias estimate.

Table 3 Bootstrap estimates of biases for sample sizes $n = 25$, 100, 400 and 1,600 when the normal distribution model is fitted to data generated by a standard normal distribution and the model parameters are estimated by a robust procedure

n	C^*_{123}	C^*_{13}	C^*_1	C^*_2	C^*_3	$\mathrm{tr}(\hat{I}\hat{J}^{-1})$
25	2.275	2.581	−0.162	−0.306	2.743	1.884
	(40.1966)	(20.2358)	(25.4973)	(23.8070)	(37.1099)	–
	(0.8571)	(0.7607)	(8.3050)	(1.0766)	(6.5114)	(0.1334)
100	2.247	2.248	−0.365	−0.002	2.613	1.970
	(74.9426)	(11.5065)	(16.2370)	(57.0646)	(26.0014)	–
	(0.9291)	(0.2951)	(0.9035)	(0.5756)	(1.5505)	(0.0515)
400	2.008	2.008	−0.195	0.000	2.253	1.992
	(203.7201)	(4.1176)	(9.8470)	(204.7648)	(13.4178)	–
	(2.0650)	(0.0565)	(0.4017)	(2.0640)	(0.6322)	(0.0143)
1600	2.007	2.008	−0.156	0.000	2.163	2.00
	(812.2547)	(6.7168)	(8.6334)	(805.2575)	(11.0462)	–
	(0.8347)	(0.0343)	(0.1515)	(0.8059)	(0.2343)	(0.004)

Two variances in the table are suggestive about the selection of the number of bootstrap resamplings. For small sample size such as $n = 25$, the bootstrap variance $8.6477/100$ is significantly smaller than the variance of $C^*_{13} = 0.2573$. This means that the bootstrap resampling over 100 times is not so effective compared with its cost for small n. On the other hand, for $n = 1{,}600$, the bootstrap variances with $B = 1{,}000$ still contribute a main part of the variances shown in the third row.

Example 2 (True distribution: Laplace) Table 2 shows the case when the same normal distribution model is fitted to the data generated from the Laplace distribution

$$g(x) = \frac{1}{\sqrt{2}} \exp\left\{-\sqrt{2}|x|\right\}, \tag{55}$$

for which $\sigma^2 = 2$, $\mu_4 = 6$.

The bootstrap estimates of the bias, C^*_{123} and C^*_{13} are larger than those of Table 1. However, this simulation result also shows that the means of C^*_2 are zero and the variances are proportional to the sample size n. Actually, in this case, the variance of C^*_2 is evaluated as $(\mu_4/\sigma^4 - 1)n/4 = 5n/4$. It can be seen that for small sample such as $n = 25$, the variance of C^*_3 is significantly larger than that of C^*_1. Also comparing two variances of C^*_{13}, we understand that at least for $n = 25$, the number of bootstrap resampling $B = 100$ is already sufficiently large. For large sample sizes such as $n = 400$ and 1,600, C^*_1 and C^*_3 are very close each other and C^*_{13} yields similar values as $\mathrm{tr}(\hat{I}\hat{J}^{-1})$.

In both Tables 1 and 2, the maximum likelihood estimators are used for the estimation of the parameters of the model. As shown in the previous section, for this

situation, the variances of C^*_{123} and C^*_2 are of order $O(n)$, whereas those of C^*_1, C^*_2 and C^*_{13} are of order $O(1)$.

Example 3 (Robust estimation) Table 3 shows the case when the parameters of the normal distribution model, μ and σ, are estimated by using a robust procedure. In this example, the median $\hat{\mu}_m = \text{med}_i\{X_i\}$ and the median absolute deviation $\hat{\sigma}_m = c^{-1}\text{med}_i\{|X_i - \text{med}_j\{X_j\}|\}$, where $c = \Phi^{-1}(0.75)$, are used. To reduce computing time, $S = 100{,}000$ was used for the sample size $n = 1{,}600$.

The bootstrap method can be applied to even these types of estimates as well. In this case, Table 3 shows that C^*_1 and C^*_3 take entirely different values. It is noteworthy, however, that for large sample cases such as $n = 400$ and larger, the bias correction term by AIC (=2) provides an appropriate approximation to the bootstrap bias estimates. This is due to the fact that the asymptotic bias may coincide with the number of parameters (Konishi and Kitagawa 1996, and page 132 of (2008)). As shown in (52), although the variances of C^*_1 and C^*_3 are large, variance of C^*_{13} is of order $O(1)$ and thus the variance reduction method is effective even for this situation.

6.2 Higher order bias corrections

In this subsection, we consider the effect of higher-order bias correction for bootstrap information criteria. To fully understand the results of Monte Carlo study, we shall use the parameters that can be expressed by statistical functionals and utilize explicit representations of various bias correction terms.

For a normal distribution model with unknown mean μ and the variance σ^2, the maximum likelihood estimators are explicitly given by statistical functionals (Konishi and Kitagawa 2008),

$$T_\mu(G) = \int x dG(x), \quad T_{\sigma^2}(G) = \int (x - T_\mu(G))^2 dG(x). \tag{56}$$

For these estimators, the functional derivatives are given by

$$\begin{aligned} T^{(1)}_\mu(x : G) &= x - \mu, \quad T^{(j)}_\mu(x_1, \ldots, x_j : G) = 0 \ (j \geq 2), \\ T^{(1)}_{\sigma^2}(x : G) &= (x - \mu)^2 - \sigma^2, \\ T^{(2)}_{\sigma^2}(x, y : G) &= -2(x - \mu)(y - \mu), \quad T^{(j)}_{\sigma^2}(x_1, \ldots, x_j : G) = 0 \ (j \geq 3). \end{aligned} \tag{57}$$

Using these results, the second-order bias correction term and the bias of the first order bias correction term are explicitly given by

$$b_1(G) = \frac{1}{2}\left(1 + \frac{\mu_4}{\sigma^4}\right), \tag{58}$$

$$b_2(G) = 3 - \frac{\mu_4}{\sigma^4} - \frac{1}{2}\frac{\mu_6}{\sigma^6} + 4\frac{\mu_3^2}{\sigma^6} + \frac{3}{2}\frac{\mu_4^2}{\sigma^8}, \tag{59}$$

Table 4 Bias correction terms and their estimates for normal distribution models

Sample size n	25	50	100	200	400	800
True bias $b(G)$	2.273	2.128	2.061	2.030	2.015	2.008
$b_1(G)$	2.000	2.000	2.000	2.000	2.000	2.000
$b_1(G)+n^{-1}b_2(G)$	2.240	2.120	2.060	2.030	2.015	2.008
$b_1(\hat{G})$	1.884	1.941	1.970	1.985	1.992	1.996
$b_1(\hat{G})+n^{-1}b_2(\hat{G})$	2.179	2.079	2.035	2.017	2.008	2.004
$b_1(\hat{G})+n^{-1}(b_2(\hat{G})-\Delta b_1(\hat{G}))$	2.177	2.103	2.056	2.029	2.015	2.008
$2+n^{-1}b_2(\hat{G})$	2.294	2.138	2.065	2.031	2.015	2.008
B_1^*	2.217	2.087	2.037	2.017	2.008	2.004
B_2^*	2.255	2.119	2.059	2.030	2.015	2.008
B_2^{**}	2.258	2.110	2.055	2.028	2.014	2.008
B_3^*	2.264	2.126	2.061	2.031	2.015	2.008

$$\Delta b_1(G) = 3 - \frac{3}{2}\frac{\mu_4}{\sigma^4} - \frac{\mu_6}{\sigma^6} + 4\frac{\mu_3^2}{\sigma^6} + \frac{3}{2}\frac{\mu_4^2}{\sigma^8}, \tag{60}$$

$$b_2(G) - \Delta b_1(G) = \frac{\mu_4}{\sigma^4} + \frac{\mu_6}{\sigma^6}, \tag{61}$$

where μ_j is the jth central moment of the distribution generating the data.

In the following, we consider two cases where the true distribution generating data is the standard normal distribution and the Laplace distribution.

Example 4 (True distribution: Normal) If the true distribution is the normal $N(0, \sigma^2)$, we have that $\mu_3 = 0$, $\mu_4 = 3\sigma^4$ and $\mu_6 = 15\sigma^6$. Therefore in this case, we have

$$\begin{aligned} b_1(G) &= 2, \quad \frac{1}{n}b_2(G) = \frac{6}{n}, \\ \Delta b_1(G) &= -\frac{3}{n}, \quad \frac{1}{n}(b_2(G) - \Delta b_1(G)) = \frac{9}{n}. \end{aligned} \tag{62}$$

Table 4 shows the bias correction term obtained by running Monte Carlo trials $S = 1{,}000{,}000$ times for six sample sizes, $n = 25$, 50, 100, 200, 400 and 800, when the true distribution generating the data is the standard normal distribution $N(0, 1)$. The $b(G)$ represents the exact bias that can be evaluated analytically as $2n/(n-3)$.

In this case, since the model contains the true distribution, the asymptotic bias correction term $b_1(G)$ is identical to the bias correction term of AIC, i.e., the number of parameters (=2). The second order bias correction term $b_1(G)+n^{-1}b_2(G) = 2+6n^{-1}$ gives an extremely accurate approximation to the true bias $b(G)$ except for very small sample size such as $n = 25$.

However, it should be noted that, in actual use, the true distribution G or the true moments μ_j are unknown, and we have to estimate them from data. The row indicated by $b_1(\hat{G})$ in Table 4 shows the values of the first order (asymptotic) bias correction

terms obtained by substituting the sample central moments to Eq. (58). It can be seen that they underestimate the $b_1(G)$. This bias of $b_1(\hat{G})$ is clearly explained by the bias of the first-order bias correction term $\Delta b_1(G)$. Note that in this case, it is given by $\Delta b_1(G) = -3/n$ and are -0.120, -0.030 and -0.008 for $n = 25$, 100 and 400, respectively. It can be seen from the table, that the corresponding bias $b_1(\hat{G}) + n^{-1}b_2(\hat{G})$ has a negative bias and $b_1(\hat{G}) + n^{-1}(b_2(\hat{G}) + \Delta b_1(\hat{G}))$ yields significantly better estimates of $b_1(G) + n^{-1}b_2(G)$ at least for n larger than or equals to 100. Further, as mentioned in the previous section, in this case where the model contains the true distribution, the bias correction of AIC attains the exact asymptotic bias correction term and "the number of parameters" $+ n^{-1}b_2(\hat{G})$ yields a better estimate of $b_1(G) + n^{-1}b_2(G)$ than $b_1(\hat{G}) + n^{-1}b_2(\hat{G})$. This can be checked from the row indicated by $2 + n^{-1}b_2(\hat{G})$ in Table 4.

The last four rows in Table 4 show the results of bootstrap bias correction. The notations B_1^*, B_2^* and B_2^{**} represent the bootstrap bias corrections terms obtained by Eqs. (51), (53) and (54), respectively. As pointed out previously, the bootstrap bias correction automatically achieves higher-order bias correction and actually it yields close approximations to the (sample) second order correction term $b_1(\hat{G}) + n^{-1}b_2(\hat{G})$. From the table, it can be seen that B_1^* is less biased than $b_1(\hat{G})$.

The second-order bootstrap bias correction terms, B_2^* and B_2^{**}, obtained by Eqs. (53) and (54) give close values of $b_1(\hat{G}) + n^{-1}(b_2(\hat{G}) - \Delta b_1(\hat{G}))$ and the true bias $b(G)$. It is worth noting that for small sample sizes such as $n = 25$, 50 and 100, they yield closer values of the true bias $b(G)$ than $b_1(\hat{G}) + n^{-1}(b_2(\hat{G}) - \Delta b_1(\hat{G}))$ and even the $b_1(G) + n^{-1}b_2(G)$ for some case. It is probably because the third order term $n^{-2}b_3(G)$ is automatically estimated by bootstrapping. Finally, B_3^* shows the estimates obtained by bootstrapping the second-order bias corrected log-likelihood using $b_1(\hat{G}) + n^{-1}(b_2(\hat{G}) - \Delta b_1(\hat{G}))$. We anticipated that it yields an estimate of the third-order bias correction term. Actually, very good estimates of the true bias are obtained even for smaller sample sizes.

Example 5 (True distribution: Laplace) We consider the case when the true distribution is the Laplace distribution (two-sided exponential distribution),

$$g(x) = \frac{1}{2}\exp(-|x|). \tag{63}$$

In this case, the central moments are $\mu_3 = 0, \mu_4 = 6, \mu_6 = 90$, and the bias correction terms are given by

$$\begin{aligned} &b_1(G) = \frac{7}{2}, \quad b_1(G) + \frac{1}{n}b_2(G) = \frac{7}{2} + \frac{6}{n}, \\ &\Delta b_1(G) = -\frac{42}{n}, \quad b_1(G) + \frac{1}{n}(b_2(G) - \Delta b_1(G)) = \frac{7}{2n} + \frac{48}{n^2}. \end{aligned} \tag{64}$$

Different from the case of normal distribution, the bias of $b_1(G)$, $\Delta b_1(G)$, is very large and is seven times of the second-order bias correction term $b_2(G)$. Therefore, in general, it would be meaningless to correct for only $b_2(G)$.

Table 5 Bias correction terms and their estimates for Laplace distribution models

Sample size n	25	50	100	200	400	800
True bias $b(G)$	3.875	3.661	3.572	3.533	3.515	3.508
$b_1(G)$	3.500	3.500	3.500	3.500	3.500	3.500
$b_1(G) + n^{-1}b_2(G)$	3.740	3.620	3.560	3.530	3.515	3.508
$b_1(\hat{G})$	2.594	2.929	3.166	3.313	3.402	3.449
$b_1(\hat{G}) + n^{-1}b_2(\hat{G})$	3.296	3.310	3.343	3.386	3.431	3.460
$b_1(\hat{G}) + n^{-1}(b_2(\hat{G}) - \Delta b_1(\hat{G}))$	3.283	3.434	3.494	3.507	3.509	3.505
B_1^*	3.433	3.301	3.332	3.383	3.430	3.460
B_2^*	3.650	3.511	3.505	3.505	3.507	3.504
B_2^{**}	3.608	3.513	3.490	3.494	3.501	3.501
B_3^*	3.767	3.596	3.553	3.526	3.515	3.507

Table 5 shows the true bias and various bias correction terms for six sample sizes, $n = 25, 50, 100, 200, 400$ and 800. The true biases $b(G)$ are estimated by conducting $S = 100{,}000{,}000$ Monte Carlo simulations.

The asymptotic correction term is $b_1(G) = 3.5$ and the second-order correction term $b_1(G)+n^{-1}b_2(G)$ yields relatively good approximations to $b(G)$. However, their estimates $b_1(\hat{G})$ and $b_1(\hat{G}) + n^{-1}b_2(\hat{G})$ have significant biases because of the large value of the bias of the asymptotic bias estimate $b_1(\hat{G})$, $\Delta b_1(G) = -42/n$. In fact, the bias-corrected second-order bias correction term $b_1(\hat{G}) + n^{-1}(b_2(\hat{G}) - \Delta b_1(\hat{G}))$ yields fairly accurate approximations to the true bias. From the table, we also notice that while the correction with $\Delta b_1(G)$ significantly improves the estimates for $n = 50$ or larger, it is virtually useless when $n = 25$. This is probably due to a poor estimation accuracy on the first-order corrected bias and seems to indicate a limitation of the analytic second-order bias correction.

In this case the bootstrap estimate B_1^* also gives a better approximation to the bias $b(G)$ than does $b(\hat{G})$ and is actually close to the bias-unadjusted second-order bias correction term $b_1(\hat{G}) + n^{-1}b_2(\hat{G})$ except for $n = 25$. B_2^* and B_2^{**} are second-order bootstrap bias correction terms by (53) and (54). They yield fairly good approximation to the second-order bias $b_1(\hat{G}) + n^{-1}(b_2(\hat{G}) - \Delta b_1(\hat{G}))$ for sample sizes larger than or equal to 100. However, for $n = 25$ and 50, B_2^* and B_2^{**} yield rather better approximations of the true bias $b(G)$ than $b_1(\hat{G}) + n^{-1}(b_2(\hat{G}) - \Delta b_1(\hat{G}))$. Finally, B_3^*, obtained by bootstrapping the second-order bias corrected log-likelihood, yields closer estimates of the true bias $b(G)$ than the second order corrected terms B_2^* and B_2^{**}.

A comment on the bootstrapping with small sample sizes, such as $n = 25$ or smaller, will be in order here. In such cases, there is a positive probability of obtaining bootstrap samples with variance 0 or very small compared with the original sample that may cause unexpectedly large bias estimate. Therefore, in obtaining the bootstrap bias correction term for small sample sizes, we have to set a threshold to exclude such pathological cases.

7 Summary and conclusions

Statistical properties of the bootstrap bias correction terms used for defining the bootstrap information criteria are analyzed based on the statistical functional approach. Various bias and variance reduction methods for the bias correction terms are considered in details both theoretically and by simulation study.

It was shown that

(1) The variance reduction method can remove the dominant $O(n^{-1})$ bootstrap sampling error. This method can be applied to any types of models and estimation procedures and increases the accuracy of the bootstrap bias correction term, in particular for large sample size.
(2) In actual estimation, the simple second-order bias correction cannot eliminate the second-order bias entirely and we need a correction for the bias of the first-order bias correction term.
(3) Bootstrap bias correction term automatically performs the higher-order bias correction. However, it also suffers from the bias of the first-order bias correction term. This bias also can be removed by bootstrapping first-order corrected log-likelihood or by double bootstrapping. By the Monte Carlo study, it was shown that the first-order and second-order bootstrap bias corrections outperform the first and second order analytic bias corrections, respectively.

From these analysis, it can be concluded that the proposed bias and variance reduction methods are quite effective in computing the bootstrap information criterion.

Acknowledgments We would like to thank the reviewer for careful reading and helpful comments that improved the quality of this paper considerably.

Appendix: Derivation of the variances of the bias correction terms

In this subsection we briefly review the evaluation of the first order bias term used for the derivation of GIC. Assume that the estimator $\hat{\theta}_p$ of $\theta = (\theta_1, \ldots, \theta_m)$ is defined by $\hat{\theta}_p = T_p(\hat{G})$. Then the Taylor series expansion for $\hat{\theta}_p = T_p(\hat{G})$ up to order n^{-1} is,

$$\hat{\theta}_p = T_p(G) + \frac{1}{n}\sum_{\alpha=1}^{n} T_p^{(1)}(X_\alpha; G) + \frac{1}{2n^2}\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} T_p^{(2)}(X_\alpha, X_\beta; G) + o(n^{-1}), \quad (65)$$

where $T_p^{(j)}$ is defined as symmetric functions such that, for any arbitrary distribution H,

$$\frac{d^j}{d\varepsilon^j} T_p\{(1-\varepsilon)G + \varepsilon H\} = \int \cdots \int T_p^{(j)}(x_1, \ldots, x_j; G) \times \prod_{i=1}^{j} d\{H(x_i) - G(x_i)\}, \quad (66)$$

at $\varepsilon = 0$ and for any i such that $1 \le i \le j$,

$$\int T_p^{(j)}(x_1, \ldots, x_j; G) dG(x_i) = 0. \tag{67}$$

Expanding $\int \log f(y|\hat{\theta}) dG(y)$ in a Taylor series around $\theta = T(G) = (T_1(G), \ldots, T_m(G))$,

$$\begin{aligned}\int \log f(y|\hat{\theta}) dG(y) = &\int \log f(y|T(G)) dG(y) \\ &+ \sum_{p=1}^{m} (\hat{\theta}_p - T_p(G)) \int \frac{\partial}{\partial \theta_p} \log f(y|T(G)) dG(y) \\ &+ \frac{1}{2} \sum_{p=1}^{m} \sum_{q=1}^{m} (\hat{\theta}_p - T_p(G))(\hat{\theta}_q - T_q(G)) \\ &\times \int \frac{\partial^2}{\partial \theta_p \partial \theta_q} \log f(y|T(G)) dG(y) + o(n^{-1}). \end{aligned} \tag{68}$$

For simplicity, hereafter we shall use the following notations:

$$\begin{aligned} &T_{p\alpha}^{(1)} = T_p^{(1)}(X_\alpha; G), \quad T_{p\alpha\beta}^{(2)} = T_p^{(2)}(X_\alpha, X_\beta; G) \\ &S_p^{(2)} = \int T_p^{(2)}(x, x; G) dG(x), \quad S_{pq}^{(11)} = \int T_p^{(1)}(x; G) T_q^{(1)}(x; G) dG(x) \\ &f_{p\alpha}^{(1)} = \frac{\partial \log f(X_\alpha; G)}{\partial \theta_p}, \quad f_{pq\alpha}^{(2)} = \frac{\partial^2 \log f(X_\alpha; G)}{\partial \theta_p \partial \theta_q}, \\ &F_p^{(1)} = \int \frac{\partial \log f(y; G)}{\partial \theta_p} dG(y), \quad F_{pq}^{(2)} = \int \frac{\partial^2 \log f(y; G)}{\partial \theta_p \partial \theta_q} dG(y). \end{aligned} \tag{69}$$

Note that $F_{pq}^{(2)} = -J(G)$. Further, in the following expressions, we omit the symbol of summation and assume that we will always take summation over all subscripts; from 1 to n for Greek subscripts and from 1 to m for Roman subscripts. For example, $T_{p\alpha}^{(1)} T_{q\beta}^{(1)} f_{pq\gamma}^{(2)}$ means

$$\sum_{p=1}^{m} \sum_{q=1}^{m} \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \sum_{\gamma=1}^{n} T_{p\alpha}^{(1)} T_{q\beta}^{(1)} f_{pq\gamma}^{(2)}.$$

Then substituting (68) into (48), the difference between the expected log-likelihoods, $D_3(G)$ is expressed as follows:

$$\begin{aligned} D_3(G) &= \int \log f(z|T(G)) dG(z) - \int \log f(z|\hat{\theta}) dG(z) \\ &= -\frac{1}{n} T_{p\alpha}^{(1)} F_p^{(1)} - \frac{1}{2n^2} \left(T_{p\alpha\beta}^{(2)} F_p^{(1)} + T_{p\alpha}^{(1)} T_{q\beta}^{(1)} F_{pq}^{(2)} \right) + o\left(n^{-1}\right), \end{aligned} \tag{70}$$

Similarly the expansion for the difference between the log-likelihoods is given by

$$D_1(G) = \frac{1}{n}\sum_{\alpha=1}^{n}\log f(X_\alpha|\hat{\theta}) - \frac{1}{n}\sum_{\alpha=1}^{n}\log f(X_\alpha|T(G))$$
$$= \frac{1}{n^2}T^{(1)}_{p\alpha}f^{(1)}_{p\beta} + \frac{1}{2n^3}\left(T^{(2)}_{p\alpha\beta}f^{(1)}_{p\gamma} + T^{(1)}_{p\alpha}T^{(1)}_{q\beta}f^{(2)}_{pq\gamma}\right) + o\left(n^{-1}\right). \quad (71)$$

Taking expectation term by term yields

$$E_G[D_1(G)] = \frac{1}{n}\left(\frac{1}{n}\int T^{(1)}_{p\alpha}f^{(1)}_{p\alpha}dG(z) + \frac{1}{2}S^{(2)}_{p}F^{(1)}_{p} + \frac{1}{2}\mathrm{tr}\{S^{(11)}_{pq}F^{(2)}_{pq}\}\right) + o(n^{-1}), \quad (72)$$

$$E_G[D_3(G)] = -\frac{1}{2n}\left(S^{(2)}_{p}F^{(1)}_{p} + \mathrm{tr}\{S^{(11)}_{pq}F^{(2)}_{pq}\}\right) + o(n^{-1}), \quad (73)$$

and we obtain

$$b_1(G) = E[D(G)] = E[D_1(G)] + E[D_3(G)]$$
$$= \frac{1}{n^2}\int T^{(1)}_{p\alpha}f^{(1)}_{p\alpha}dG(z) + o(n^{-1}). \quad (74)$$

Since $\{E_G[D_1(G)]\}^2 = O(n^{-2})$ and $\{E_G[D_3(G)]\}^2 = O(n^{-2})$,

$$\mathrm{Var}(D_3(G)) = E_G[\{D_3(G)\}^2] + O(n^{-2})$$
$$= E_G\left[\left\{\frac{1}{n}T^{(1)}_{p\alpha}F^{(1)}_{p} + \frac{1}{2n^2}T^{(2)}_{p\alpha\beta}F^{(1)}_{p} + \frac{1}{2n^2}T^{(1)}_{p\alpha}T^{(1)}_{p\beta}F^{(2)}_{pq}\right\}^2\right] + O(n^{-2})$$
$$= E_G\left[\frac{1}{n^2}T^{(1)}_{p\alpha}T^{(1)}_{q\beta}F^{(1)}_{p}F^{(1)}_{q} + \frac{1}{4n^4}T^{(2)}_{p\alpha\beta}T^{(2)}_{q\varepsilon\delta}F^{(1)}_{p}F^{(1)}_{q}\right.$$
$$+\frac{1}{4n^4}T^{(1)}_{p\alpha}T^{(1)}_{p\beta}T^{(1)}_{r\varepsilon}T^{(1)}_{r\delta}F^{(2)}_{pq}F^{(2)}_{rs} + \frac{1}{n^3}T^{(1)}_{p\alpha}T^{(2)}_{q\varepsilon\delta}F^{(1)}_{p}F^{(1)}_{q}$$
$$\left.+\frac{1}{n^3}T^{(1)}_{p\alpha}T^{(1)}_{q\beta}T^{(1)}_{\gamma\varepsilon}F^{(1)}_{p}F^{(2)}_{qr} + \frac{1}{2n^4}T^{(2)}_{p\alpha\beta}T^{(1)}_{q\varepsilon}T^{(1)}_{q\delta}F^{(1)}_{p}F^{(2)}_{qr}\right] + O(n^{-2})$$
$$= \frac{1}{n}\left\{F^{(1)}_{p}S^{(11)}_{pq}F^{(1)}_{q}\right\} + O(n^{-2}). \quad (75)$$

Hereafter, we shall consider the main terms of the $\mathrm{Var}(D_3(G))$, $\mathrm{Var}(D_1(G))$ and $\mathrm{Var}(D(G))$.

$$\begin{aligned}
\mathrm{Var}(D_1(G)) &= E_G[\{D_1(G)\}^2] + O(n^{-2}) \\
&= E_G\left[\left\{\frac{1}{n^2}T^{(1)}_{p\alpha} f^{(1)}_{p\beta} + \frac{1}{2n^3}\left(T^{(2)}_{p\alpha\beta} f^{(1)}_{p\beta} + T^{(1)}_{p\alpha} T^{(1)}_{q\beta} f^{(2)}_{pq\gamma}\right)\right\}^2\right] + O(n^{-2}) \\
&= E_G\left[\frac{1}{n^4}T^{(1)}_{p\alpha} T^{(1)}_{q\beta} f^{(1)}_{p\gamma} f^{(1)}_{p\delta} + \frac{1}{4n^6}T^{(2)}_{p\alpha\beta} T^{(2)}_{q\gamma\delta} f^{(1)}_{p\beta} f^{(1)}_{q\delta}\right. \\
&\quad + \frac{1}{4n^6}T^{(1)}_{p\alpha} T^{(1)}_{q\beta} T^{(1)}_{r\gamma} T^{(1)}_{s\delta} f^{(2)}_{pq\varepsilon} f^{(2)}_{rs\eta} + \frac{1}{n^5}T^{(1)}_{p\alpha} T^{(2)}_{q\alpha\gamma} f^{(1)}_{p\beta} f^{(1)}_{p\delta} \\
&\quad \left. + \frac{1}{n^5}T^{(1)}_{p\alpha} T^{(1)}_{q\gamma} T^{(1)}_{r\delta} f^{(1)}_{p\beta} f^{(2)}_{qr\varepsilon} + \frac{1}{2n^6}T^{(1)}_{p\alpha} T^{(1)}_{q\beta} T^{(2)}_{r\delta\varepsilon} f^{(1)}_{r\varepsilon} f^{(2)}_{pq\gamma}\right] + O(n^{-2}) \\
&= \frac{1}{n}\left\{F^{(1)}_p S^{(11)}_{pq} F^{(1)}_q\right\} + O(n^{-2}). \qquad (76)
\end{aligned}$$

Therefore, the mean and the variance of $D_1(G) + D_3(G)$ are given by

$$\begin{aligned}
E[D_1(G) + D_3(G)] &= E[D_1(G)] + E[D_3(G)] \\
&= \frac{1}{n}\int T^{(1)}_{p\alpha} f^{(1)}_{p\alpha} dG(x) + o(n^{-1}) \qquad (77)
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\{D_1(G) + D_3(G)\} &= E\left\{\left(\frac{1}{n^2}T^{(1)}_{p\alpha} f^{(1)}_{p\beta} - \frac{1}{n}T^{(1)}_{p\alpha} F^{(1)}_p\right)\right. \\
&\quad + \left(\frac{1}{2n^3}T^{(2)}_{p\alpha\beta} f^{(1)}_{p\gamma} - \frac{1}{2n^2}T^{(2)}_{p\alpha\beta} F^{(1)}_p\right) \\
&\quad \left. + \left(\frac{1}{2n^3}T^{(1)}_{p\alpha} T^{(1)}_{q\beta} f^{(2)}_{pq\gamma} - \frac{1}{2n^2}T^{(1)}_{p\alpha} T^{(1)}_{q\beta} F^{(2)}_{pq}\right)\right\}^2 + O(n^{-2}) \\
&= O(n^{-2}). \qquad (78)
\end{aligned}$$

On the other hand, from

$$D_2(G) = \frac{1}{n}\sum_{\alpha} \log f(X_\alpha | T(G)) - \int \log f(x|T(G)) dG(x), \qquad (79)$$

it is obvious that the mean and variance of $D_2(G)$ are given by

$$\begin{aligned}
E[D_2] &= 0 \qquad (80) \\
\mathrm{Var}\{D_2\} &= \frac{1}{n}\mathrm{Var}(\log f(X|T(G))) \\
&= \frac{1}{n}E_{G(x)}\left[\left\{\log f(X|T(G)) - E_{G(y)}[\log f(Y|T(G))]\right\}^2\right]. \qquad (81)
\end{aligned}$$

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd international symposium in information theory* (pp. 267–281). Budapest: Akademiai Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723.

Bozdogan, H. (1994). *Proceeding of the first US/Japan conference on the frontiers of statistical modeling: An informational approach*. The Netherlands: Kluwer.

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

Davison, A. C. (1986). Approximate predictive likelihood. *Biometrika, 73*, 323–332.

Davison, A. C., Hinkley, D. V. (1992). Bootstrap likelihoods. *Biometrika, 79*(1), 113–130.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *Annals of Statistics, 7*, 1–26.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association, 78*(382), 316–331.

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

Ishiguro, M., Sakamoto, Y., Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics, 49*(3), 411–434.

Konishi, S., Kitagawa, G. (1996). Generalized information criteria in model selection. *Biometrika, 83*(4), 875–890.

Konishi, S., Kitagawa, G. (1998). Second order bias correction for generalized information criterion. *Research Memorandom*, No.661. The Institute of Statistical Mathematics.

Konishi, S., Kitagawa, G. (2003). Asymptotic theory for information criteria in model selection—functional approach. *Journal of Statistical Planning and Inference, 114*, 45–61.

Konishi, S., Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York: Springer.

Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

Sakamoto, Y., Ishiguro, M., Kitagawa, G. (1986). *Akaike information criterion statistics*. Dordrecht: D. Reidel Publishing Company.

Shibata, R. (1997). Bootstrap estimate of Kullback–Leibler information for model selection. *Statistica Sinica, 7*, 375–394.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, B39*, 44–47.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics Series A, 7(1)*, 13–26.

Takeuchi, K. (1976). Distributions of information statistics and criteria for adequacy of models. *Mathematical Science, 153*, 12–18 (in Japanese).

Wong, W. (1983). A note on the modified likelihood for density estimation. *Journal of the American Statistical Association, 78*(382), 461–463.