Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 97 (2006) 1965 – 1975

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Bias correction of cross-validation criterion based on Kullback–Leibler information under a general condition

Hirokazu Yanagihara[a,*], Tetsuji Tonda[b], Chieko Matsumoto[c]

[a] *Department of Social Systems and Management, Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan*

[b] *Department of Environmetrics and Biometrics, Research Institute for Radiation Biology and Medicine, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima, Hiroshima 734-8553, Japan*

[c] *Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan*

Received 11 April 2005
Available online 7 July 2006

Abstract

This paper deals with the bias correction of the cross-validation (CV) criterion to estimate the predictive Kullback–Leibler information. A bias-corrected CV criterion is proposed by replacing the ordinary maximum likelihood estimator with the maximizer of the adjusted log-likelihood function. The adjustment is just slight and simple, but the improvement of the bias is remarkable. The bias of the ordinary CV criterion is $O(n^{-1})$, but that of the bias-corrected CV criterion is $O(n^{-2})$. We verify that our criterion has smaller bias than the AIC, TIC, EIC and the ordinary CV criterion by numerical experiments.

AMS 2000 subject classification: primary 62H12; secondary 62F07

Keywords: Bias correction; Cross-validation; Predictive Kullback–Leibler information; Model misspecification; Model selection; Robustness; Weighted log-likelihood function

* Corresponding author. Fax: +81 29 853 6451.
E-mail address: yanagi@sk.tsukuba.ac.jp (H. Yanagihara).

0047-259X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2005.10.009

1. Introduction

Let $\varphi(\boldsymbol{y})$ be the probability density function and $\boldsymbol{y}_i$ $(i = 1, \ldots, n)$ be a $p \times 1$ observation vector drawn from φ, where n is the sample size. The true model is expressed as

$$M^* : \boldsymbol{y}_i \sim i.i.d.\ \varphi(\boldsymbol{y}_i), \quad (i = 1, \ldots, n). \tag{1}$$

We consider a family of parametric models $\mathcal{F} = \{f(\boldsymbol{y}|\boldsymbol{\theta});\ \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^q\}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ is the q-dimensional vector of unknown parameters. A candidate model is expressed as

$$M : \boldsymbol{y}_i \sim i.i.d.\ f(\boldsymbol{y}_i|\boldsymbol{\theta}), \quad (i = 1, \ldots, n). \tag{2}$$

Akaike's information criterion (AIC) proposed by Akaike [1,2] has been used universally for choosing the best model from the candidate models. It is well known that the AIC is an estimator of the risk based on the predictive Kullback–Leibler (K–L) information (see [8]), which measures the discrepancy between the true model M^* and the candidate model M. However, AIC has a constant bias for the risk when the candidate model is misspecified (see e.g., [11,12]). Takeuchi [12] revaluated the bias correction term of AIC under the situation that $\mathcal{F}$ does not contain $\varphi(\boldsymbol{y})$, and proposed the Takeuchi's information criterion (TIC) by replacing the AIC's bias correction term with the revaluated term. The TIC is an asymptotically unbiased estimator for the risk if $\boldsymbol{y}_i$'s are *i.i.d.* However, Fujikoshi et al. [5] pointed out that TIC in normal regression models hardly corrects the bias in actual use, because its bias correction term mainly consists of an estimator of the fourth cumulant of the true distribution. Such an estimator tends to underestimate too much, even if the sample size n is moderate (see [15]).

Like TIC, the cross-validation (CV) criterion proposed by Stone [10] is known as an asymptotically unbiased estimator for the risk (see [11]), although there are no estimators of higher-order cumulants in the CV criterion. Therefore, unlike TIC, the CV criterion can correct the bias efficiently. Using the better property of the CV criterion, Yanagihara [14,15] proposed new criteria which are partially constructed by the cross-validation method, and which are slightly influenced by the difference between $\varphi(\boldsymbol{y})$ and $f(\boldsymbol{y}|\boldsymbol{\theta})$. However, a bias for the risk exists also in the CV criterion. Fujikoshi et al. [4] corrected the biases of the CV criteria in normal multivariate regression and GMANOVA models. The purpose of our paper is to reduce the bias in the CV criterion under a general condition without adding any correction terms. We replaced the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ with the maximizer of the adjusted log-likelihood function, and thus propose a bias-corrected CV (corrected CV, CCV) criterion. The adjustment is merely to change a weight of the weighted log-likelihood function, but the improvement of the bias is remarkable. The bias of the ordinary CV criterion is $O(n^{-1})$, but that of the CCV criterion is $O(n^{-2})$.

This paper is organized in the following way. In Section 2, we describe the risk based on the K–L information and usual information criteria. In Section 3, we state the derivation of CCV criterion and its asymptotic property. In Section 4, we verify that CCV criterion has smaller bias than other criteria, namely, the AIC, TIC, (the extended information criterion (EIC) [6]) and CV criterion by numerical experiments.

2. Risk and usual information criteria

Let $L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d})$ be a weighted log-likelihood function on $f(\boldsymbol{y}_i|\boldsymbol{\theta})$ given by

$$L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d}) = \sum_{i=1}^{n} d_i \log f(\boldsymbol{y}_i|\boldsymbol{\theta}), \tag{3}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ and $\boldsymbol{d} = (d_1, \ldots, d_n)'$. Let $\mathbf{1}_n$ be the $n \times 1$ vector, all of which elements are 1. For simplicity, let us express $L(\boldsymbol{\theta}|\boldsymbol{Y}) = L(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n)$. The MLE of $\boldsymbol{\theta}$ is obtained by maximizing the ordinary log-likelihood function $L(\boldsymbol{\theta}|\boldsymbol{Y})$, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{Y}). \tag{4}$$

Let $\boldsymbol{u}_i$ be a $p \times 1$ future observation vector and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)'$. We assume that $\boldsymbol{U}$ is independent of $\boldsymbol{Y}$ and $\boldsymbol{u}_i$'s are independently and identically distributed according to φ. Note that the distribution of $\boldsymbol{u}_i$ is the same as that of $\boldsymbol{y}_i$. Then, the risk based on the predictive K–L information, which measures the discrepancy between the true model M^* and the candidate model M is defined by

$$R_{\mathrm{KL}} = E^*_{\boldsymbol{y}} E^*_{\boldsymbol{u}} \left[-2L(\hat{\boldsymbol{\theta}}|\boldsymbol{U}) \right], \tag{5}$$

where E^* means an expectation under the true model M^*.

The AIC proposed by Akaike [1,2] is a simple estimator of the risk R_{KL}, and is given by

$$\mathrm{AIC} = -2L(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + 2q. \tag{6}$$

However, if the candidate model is misspecified, AIC has a constant bias, i.e.,

$$B_{\mathrm{AIC}} = R_{\mathrm{KL}} - E^*_{\boldsymbol{y}}[\mathrm{AIC}] = O(1) \tag{7}$$

(see e.g., [11,12]). This is mainly because Akaike [1,2] derived AIC under the condition that the candidate model is correctly specified. Takeuchi [12] revaluated the AIC's bias correction term, $2q$, under the situation that $\mathcal{F}$ does not contain $\varphi(\boldsymbol{y})$, and proposed TIC as follows: Let

$$\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i|\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad \boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}) = \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \log f(\boldsymbol{y}_i|\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{8}$$

Under the candidate model M, TIC is given by

$$\mathrm{TIC} = -2L(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + 2\,\mathrm{tr}(\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1} \hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})), \tag{9}$$

where

$$\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}), \quad \hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}) \boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'. \tag{10}$$

Takeuchi [12] showed that TIC is an asymptotically unbiased estimator for the risk R_{KL} if $\boldsymbol{y}_i$'s are *i.i.d.*, i.e.,

$$B_{\mathrm{TIC}} = R_{\mathrm{KL}} - E^*_{\boldsymbol{y}}[\mathrm{TIC}] = O(n^{-1}). \tag{11}$$

On the other hand, Stone [10] proposed the CV criterion in the following way. Let $\hat{\boldsymbol{\theta}}_{[-i]}$ be the MLE of $\boldsymbol{\theta}$ evaluated from *i*th jackknife sample. Note that $\hat{\boldsymbol{\theta}}_{[-i]}$ is the maximizer of $\sum_{j \neq i}^{n} \log f(\boldsymbol{y}_j|\boldsymbol{\theta})$, i.e.,

$$\hat{\boldsymbol{\theta}}_{[-i]} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n - \boldsymbol{e}_i), \tag{12}$$

where $\boldsymbol{e}_i$ is the $n \times 1$ vector whose ith element is 1 and the other elements are 0. The CV criterion is given by

$$\mathrm{CV} = -2 \sum_{i=1}^{n} \log f(\boldsymbol{y}_i | \hat{\boldsymbol{\theta}}_{[-i]}). \tag{13}$$

Stone [11] pointed out that the CV criterion is an asymptotically unbiased estimator for the risk. From the result in Stone [11], we can see that the TIC and CV criteria are asymptotically equivalent, i.e., CV = TIC + $O_p(n^{-1})$. Therefore, from Eq. (11), the bias of the CV criterion is given by

$$B_{\mathrm{CV}} = R_{\mathrm{KL}} - E_{\boldsymbol{y}}^{*}[\mathrm{CV}] = O(n^{-1}). \tag{14}$$

We can see that the order of bias of CV is the same as that of TIC. However, Yanagihara [15] showed that the CV criterion in normal regression models has smaller bias than TIC by investigating the asymptotic expansions of biases for the risk. In order to get TIC, we must estimate higher-order cumulants. However, the ordinary estimators of higher-order cumulants tend to underestimate too much, even if the sample size n is moderate. Consequently, TIC tends to have large bias. Needless to say, we can obtain the CV criterion without estimating higher-order cumulants.

3. Bias correction of the CV criterion

3.1. Asymptotic expansion of the bias of the CV criterion

In this section, we propose the bias-corrected CV (CCV) criterion by replacing $\hat{\boldsymbol{\theta}}_{[-i]}$ in the CV criterion with the maximizer of another weighted log-likelihood function. First, in order to correct the bias of the CV criterion, we derive an asymptotic expansion of its bias up to the order n^{-1}. Let $\boldsymbol{\theta}_0$ be the $q \times 1$ vector such that

$$E_{\boldsymbol{y}}^{*}\left[\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)\right] = \boldsymbol{0}_q, \tag{15}$$

where $\boldsymbol{0}_q$ is the $q \times 1$ vector, all of which elements are 0. Note that $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ almost surely as n goes to infinity (see [13]). Then, we obtain an asymptotic expansion of the bias of the CV criterion up to the order n^{-1} in the following theorem.

Theorem 1. *Under the regularity conditions, the bias of the CV criterion is expanded as*

$$B_{\mathrm{CV}} = R_{\mathrm{KL}} - E_{\boldsymbol{y}}^{*}[\mathrm{CV}] = -\frac{1}{n}\operatorname{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) + O(n^{-2}), \tag{16}$$

where

$$\boldsymbol{J}(\boldsymbol{\theta}_0) = -E_{\boldsymbol{y}}^{*}\left[\boldsymbol{H}(\boldsymbol{y}|\boldsymbol{\theta}_0)\right], \quad \boldsymbol{I}(\boldsymbol{\theta}_0) = E_{\boldsymbol{y}}^{*}\left[\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)'\right].$$

Proof. Let

$$\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial}{\partial \boldsymbol{\theta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right)\log f(\boldsymbol{y}_i|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Using the Taylor expansion, we obtain the perturbation expansion of $\hat{\boldsymbol{\theta}}_{[-i]}$ as

$$\hat{\boldsymbol{\theta}}_{[-i]} = \hat{\boldsymbol{\theta}} - \frac{1}{n}\boldsymbol{z}_{1,i} - \frac{1}{n^2}\boldsymbol{z}_{2,i} + O_p(n^{-3}), \tag{17}$$

where

$$z_{1,i} = \hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}), \quad z_{2,i} = \hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\left\{\frac{1}{2}\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})\mathrm{vec}(z_{1,i}z_{1,i}') - \boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})z_{1,i}\right\}.$$

Here, $\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})$ and $\boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})$ are given by (8), and $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})$ is given by (10). Since the distributions of $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ are the same, the commutative equation $E_{\boldsymbol{y}}^*[\log f(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]})] = E_{\boldsymbol{y}}^*E_{\boldsymbol{u}}^*[\log f(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}}_{[-i]})]$ holds. Therefore, using the Taylor expansion and Eq. (17), $E_{\boldsymbol{y}}^*[\mathrm{CV}]$ is expanded as

$$E_{\boldsymbol{y}}^*[\mathrm{CV}] = R_{\mathrm{KL}} + R_1 + \frac{1}{n}R_2 + O(n^{-2}), \tag{18}$$

where

$$R_1 = \frac{2}{n}\sum_{i=1}^{n} E_{\boldsymbol{y}}^*E_{\boldsymbol{u}}^*\left[\boldsymbol{g}(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})'z_{1,i}\right],$$
$$R_2 = \frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{y}}^*E_{\boldsymbol{u}}^*\left[2\boldsymbol{g}(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})'z_{2,i} - z_{1,i}'\boldsymbol{H}(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})z_{1,i}\right].$$

Note that $\sum_{i=1}^{n} z_{1,i} = \boldsymbol{0}_q$ because $\hat{\boldsymbol{\theta}}$ is the MLE, i.e., $\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}) = \boldsymbol{0}_q$. Therefore, taking a conditional expectation of $\boldsymbol{g}(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})$ for $\boldsymbol{Y}$ as $\boldsymbol{\eta} = E_{\boldsymbol{u}}^*[\boldsymbol{g}(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})|\boldsymbol{Y}]$, we obtain

$$R_1 = \frac{2}{n}\sum_{i=1}^{n} E_{\boldsymbol{y}}^*\left[E_{\boldsymbol{u}}^*\left[\boldsymbol{g}(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})'z_{1,i}\middle|\boldsymbol{Y}\right]\right] = \frac{2}{n}\sum_{i=1}^{n} E_{\boldsymbol{y}}^*\left[\boldsymbol{\eta}'z_{1,i}\right] = 0. \tag{19}$$

On the other hand, by using the equation $\hat{\boldsymbol{\theta}} \overset{a.s.}{\to} \boldsymbol{\theta}_0$ $(n \to \infty)$, R_2 is expanded as

$$R_2 = \frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{y}}^*E_{\boldsymbol{u}}^*\left[2\boldsymbol{g}(\boldsymbol{u}_i|\boldsymbol{\theta}_0)'z_{2,i} - z_{1,i}'\boldsymbol{H}(\boldsymbol{u}_i|\boldsymbol{\theta}_0)z_{1,i}\right] + O(n^{-1}).$$

From Eq. (15), the first term on the right side of the above equation disappears. Moreover, using equation $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) \overset{a.s.}{\to} \boldsymbol{J}(\boldsymbol{\theta}_0)$ and $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) \overset{a.s.}{\to} \boldsymbol{I}(\boldsymbol{\theta}_0)$ $(n \to \infty)$, we derive the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} E_{\boldsymbol{y}}^*E_{\boldsymbol{u}}^*\left[z_{1,i}'\boldsymbol{H}(\boldsymbol{u}_i|\boldsymbol{\theta}_0)z_{1,i}\right] = -\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) + O(n^{-1}).$$

Therefore, R_2 is expanded as

$$R_2 = \mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) + O(n^{-1}). \tag{20}$$

Substituting Eqs. (19) and (20) into (18) yields

$$E_{\boldsymbol{y}}^*[\mathrm{CV}] = R_{\mathrm{KL}} + \frac{1}{n}\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) + O(n^{-2}). \tag{21}$$

Consequently, the result (16) in Theorem 1 is obtained. □

Because $-\log f(\boldsymbol{y}|\boldsymbol{\theta})$ is strictly convex with respect to $\boldsymbol{\theta}$, the matrix $\boldsymbol{H}(\boldsymbol{y}|\boldsymbol{\theta})$ is positive definite (see e.g., [9, p. 49]). It makes the inequality $\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) > 0$. Consequently, the CV criterion tends to overestimate for the risk R_{KL}.

3.2. Corrected CV criterion

Next, we propose a new criterion, CCV criterion, which always corrects the bias for the risk R_{KL} to $O(n^{-2})$. Theoretically, we can correct the bias in the CV criterion by subtracting the term $n^{-1}\,\mathrm{tr}(\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}))$ from the CV criterion. However, we can easily forecast that the bias is not fully corrected by such a plug-in estimator because $\mathrm{tr}(\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}))$ must have a large bias for $\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0))$, even if the sample size n is moderate. The reason for this is the same as the reason that TIC does not reduce the bias enough in actual use. Therefore, we need to prepare other methods to correct the bias without estimating $\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0))$. From Eq. (12), we notice that $\hat{\boldsymbol{\theta}}_{[-i]}$ removes the influence of $\boldsymbol{y}_i$ perfectly. However, we consider that the effect of $\boldsymbol{y}_i$ should not be removed completely because R_{KL} in (5) is not the predictive K–L information measuring the discrepancy between $\varphi(\boldsymbol{u})$ and $f(\boldsymbol{u}|\hat{\boldsymbol{\theta}}_{[-i]})$, but, rather, $\varphi(\boldsymbol{u})$ and $f(\boldsymbol{u}|\hat{\boldsymbol{\theta}})$. Thus, we use the estimator obtained by maximizing another weighted log-likelihood function, in which the influence of $\boldsymbol{y}_i$ remains for a while. Consequently, we propose the following CCV criterion.

Definition. Let $\tilde{\boldsymbol{\theta}}_i$ be the estimator of $\boldsymbol{\theta}$ by maximizing the weighted log-likelihood function as

$$\tilde{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{1}_n - c_n\boldsymbol{e}_i), \tag{22}$$

where c_n is any constant which can be expanded as $c_n = 1 - 1/(2n) + O(n^{-2})$. Then, we propose the bias-corrected CV (CCV) criterion as

$$\mathrm{CCV} = -2\sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\tilde{\boldsymbol{\theta}}_i). \tag{23}$$

We can see that any estimators of higher-order cumulants are not necessary for obtaining CCV criterion. However, CCV criterion always corrects the bias to $O(n^{-2})$, even though there is no term based on $\mathrm{tr}(\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}))$ in the formula (23). The order of bias of the CCV criterion is obtained in the following theorem.

Theorem 2. *Under the regularity conditions, the order of the bias of the CCV criterion is given by*

$$B_{\mathrm{CCV}} = R_{\mathrm{KL}} - E_{\boldsymbol{y}}^{*}[\mathrm{CCV}] = O(n^{-2}). \tag{24}$$

Proof. From the definition of $\tilde{\boldsymbol{\theta}}_i$ and the Taylor expansion, we can expand $\tilde{\boldsymbol{\theta}}_i$ as

$$\tilde{\boldsymbol{\theta}}_i = \hat{\boldsymbol{\theta}}_{[-i]} - \frac{1}{2n^2}\left\{\frac{1}{n-1}\sum_{j\neq i}^{n} \boldsymbol{H}(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}}_{[-i]})\right\}^{-1} \boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]}) + O_p(n^{-3}).$$

Thus, the perturbation expansion of CCV is given by

$$\mathrm{CCV} = \mathrm{CV} + \frac{1}{n^2}\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]})'\left\{\frac{1}{n-1}\sum_{j\neq i}^{n} \boldsymbol{H}(\boldsymbol{y}_j|\hat{\boldsymbol{\theta}}_{[-i]})\right\}^{-1} \boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]}) + O_p(n^{-2}).$$

Therefore, we calculate $E^*_{\boldsymbol{y}}[\mathrm{CCV}]$ as

$$E^*_{\boldsymbol{y}}[\mathrm{CCV}] = E^*_{\boldsymbol{y}}[\mathrm{CV}] - \frac{1}{n}\,\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) + O(n^{-2}).$$

Substituting Eq. (21) into the above equation yields Eq. (24) in Theorem 2. □

4. Numerical examination

First, we prepare another bias-corrected criterion constructed by the bootstrap method, which was named EIC by Ishiguro et al. [6]. Let $\boldsymbol{Y}^\star_b$ $(b = 1, \ldots, B)$ be the bth bootstrap sample by resampling, and $\hat{\boldsymbol{\theta}}^\star_b$ be the MLE of $\boldsymbol{\theta}$ evaluated from $\boldsymbol{Y}^\star_b$, i.e.,

$$\hat{\boldsymbol{\theta}}^\star_b = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{Y}^\star_b). \tag{25}$$

Then, the EIC is given by

$$\mathrm{EIC} = -2L(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) - \frac{2}{B}\sum_{b=1}^{B}\left\{L(\hat{\boldsymbol{\theta}}^\star_b|\boldsymbol{Y}) - L(\hat{\boldsymbol{\theta}}^\star_b|\boldsymbol{Y}^\star_b)\right\}, \tag{26}$$

(see e.g., [7]). Through the simulation, we compare the biases and frequencies of the selected model in our proposed CCV criterion, and also the AIC, TIC, EIC and CV criterion. In this paper, we deal with the selection of the best model from the candidate models having the elliptical distribution, i.e.,

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}) = a_p|\boldsymbol{\Lambda}|^{-1/2}g((\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})), \quad (i = 1, \ldots, 20), \tag{27}$$

where $g(r)$ is a known non-negative function and a_p is a positive constant depending on the dimension p (see e.g., [3]). We choose the best $g(r)$ and a_p from the candidate models by minimizing the information criteria. The candidate models considered are as follows:

Model 1: *Multivariate normal distribution*,

$$a_p = (2\pi)^{-p/2}, \quad g(r) = e^{-r/2}.$$

Model 2: *Multivariate logistic distribution*,

$$a_p = (2\pi)^{-p/2}\left\{\sum_{j=1}^{\infty}(-1)^{j-1}j^{1-p/2}\right\}^{-1}, \quad g(r) = \frac{e^{-r/2}}{\{1+e^{-r/2}\}^2}.$$

Model 3: *Multivariate Cauchy distribution*,

$$a_p = \frac{\Gamma((p+1)/2)}{\pi^{(p+1)/2}}, \quad g(r) = (1+r)^{-(p+1)/2},$$

where $\Gamma(x)$ is the gamma function.

Choosing the best model is equivalent to determining the best weight function in the *M*-estimation. Therefore, we will judge whether or not the robust estimation should be performed through minimizing the information criterion, since the normal distribution is included in the candidate

Table 1
Biases and frequencies of the selected model according to the criteria

Distribution	Criterion		$p=2$			$p=6$		
			Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
Normal		Risk	120.50[a]	123.45	132.31	399.91[a]	405.21	404.71
	AIC	Bias	2.30	5.54	1.77	37.43	43.81	15.90
		Frequency	(40.7)	(58.5)	(0.8)	(14.4)	(85.5)	(0.1)
	TIC	Bias	3.07	5.34	−0.60	42.04	49.58	−11.84
		Frequency	(55.8)	(43.2)	(1.0)	(13.3)	(86.7)	(0.0)
	EIC	Bias	**−0.06**	0.43	−0.74	1.92	2.43	**−0.24**
		Frequency	(65.7)	(28.7)	(5.6)	(55.0)	(21.0)	(24.0)
	CV	Bias	−0.73	−0.93	−0.40	−4.35	−4.71	−2.37
		Frequency	(71.4)	(21.6)	(7.1)	(57.4)	(4.0)	(38.6)
	CCV	Bias	−0.27	**−0.30**	**−0.11**	**0.49**	**0.66**	−0.79
		Frequency	(71.0)	(22.5)	(6.5)	(62.8)	(5.5)	(31.7)
Laplace		Risk	125.27	135.24	121.32[a]	422.27	434.96	387.84[a]
	AIC	Bias	9.93	17.42	1.47	67.59	79.65	17.69
		Frequency	(62.8)	(14.0)	(23.3)	(57.2)	(36.4)	(6.4)
	TIC	Bias	8.52	13.66	−0.59	68.07	79.79	−10.49
		Frequency	(65.4)	(9.0)	(25.6)	(60.7)	(38.9)	(0.4)
	EIC	Bias	2.49	4.35	−0.80	11.68	14.22	**0.21**
		Frequency	(48.1)	(7.0)	(44.9)	(14.1)	(3.6)	(82.3)
	CV	Bias	**−0.31**	**−0.57**	−0.39	−6.27	−7.12	−1.01
		Frequency	(46.9)	(2.9)	(50.2)	(17.0)	(0.2)	(82.8)
	CCV	Bias	0.71	0.92	**−0.11**	**2.74**	**3.09**	0.58
		Frequency	(47.9)	(3.1)	(49.0)	(19.9)	(0.3)	(79.8)
Chi-square		Risk	126.92	135.41	122.52[a]	426.64	437.16	387.63[a]
	AIC	Bias	12.06	18.76	2.82	71.44	82.00	17.15
		Frequency	(39.3)	(36.5)	(24.2)	(35.2)	(56.5)	(8.3)
	TIC	Bias	10.61	15.74	-0.31	71.96	82.90	−13.39
		Frequency	(47.7)	(27.8)	(24.5)	(38.7)	(61.1)	(0.2)
	EIC	Bias	2.94	4.76	−0.51	9.74	12.46	−2.21
		Frequency	(39.5)	(18.0)	(42.5)	(12.7)	(6.3)	(81.0)
	CV	Bias	−0.76	−1.36	**−0.11**	−11.56	−13.06	−3.72
		Frequency	(38.2)	(13.8)	(48.0)	(16.3)	(0.8)	(82.9)
	CCV	Bias	**0.65**	**0.67**	0.20	**−0.21**	**−0.34**	**−2.09**
		Frequency	(38.8)	(14.4)	(46.8)	(20.1)	(0.9)	(79.0)
Log-normal		Risk	128.07	137.34	121.90[a]	427.94	439.32	384.36[a]
	AIC	Bias	13.85	21.38	2.71	75.71	87.04	17.12
		Frequency	(42.9)	(33.3)	(23.8)	(39.4)	(52.5)	(8.1)
	TIC	Bias	12.31	18.17	−0.03	76.01	87.57	−12.89
		Frequency	(49.7)	(24.9)	(25.4)	(43.7)	(55.4)	(0.9)
	EIC	Bias	4.68	7.00	−0.19	14.07	16.86	**−1.42**
		Frequency	(41.5)	(16.2)	(42.3)	(13.5)	(4.6)	(81.9)
	CV	Bias	**−0.24**	**−0.83**	**0.17**	−12.92	−14.57	−3.28
		Frequency	(41.1)	(11.3)	(47.6)	(15.7)	(0.7)	(83.6)
	CCV	Bias	1.51	1.70	0.47	**−0.31**	**−0.41**	−1.65
		Frequency	(41.6)	(11.9)	(46.5)	(19.9)	(1.2)	(78.9)

[a] Denotes the smallest risk in all the candidate models, and the smallest bias in all the criteria is in bold.

Table 2
Biases of CV, CCV, CCV′ and CCV″ criteria

Distribution	Criterion	$p=2$			$p=6$			Average
		Model 1	Model 2	Model 3	Model 1	Model 2	Model 3	
Normal	CV	−0.54	−0.73	−0.48	−6.39	−7.01	−2.71	−2.98
	CCV	**−0.08**	**−0.09**	−0.19	**−1.49**	**−1.57**	−1.14	**−0.76**
	CCV′	−0.31	−0.47	**−0.17**	−5.16	−5.81	**−0.67**	−2.10
	CCV″	−0.21	−0.31	−0.18	−3.99	−4.43	−0.90	−1.67
Laplace	CV	−1.20	−1.79	−0.31	−8.61	−9.62	−2.75	−4.04
	CCV	**−0.17**	**−0.29**	−0.02	**0.16**	**0.31**	−1.16	**−0.20**
	CCV′	−0.91	−1.45	**−0.01**	−7.27	−8.27	**−0.69**	−3.10
	CCV″	−0.68	−1.09	**−0.01**	−5.43	−6.13	−0.93	−2.38
Chi-square	CV	−0.57	−1.13	**−0.03**	−10.78	−12.07	−2.86	−4.57
	CCV	0.79	0.81	0.28	**0.36**	**0.41**	−1.23	**0.24**
	CCV′	−0.29	−0.80	0.30	−9.44	−10.74	**−0.74**	−3.62
	CCV″	**−0.07**	**−0.38**	0.30	−7.35	−8.34	−0.99	−2.80
Log-normal	CV	−1.14	−2.01	−0.35	−12.91	−14.78	−2.95	−5.69
	CCV	0.65	**0.58**	−0.05	**−0.35**	**−0.67**	−1.33	**−0.20**
	CCV′	−0.85	−1.68	**−0.03**	−11.57	−13.45	**−0.86**	−4.74
	CCV″	**−0.53**	−1.20	−0.04	−9.36	−10.90	−1.10	−3.85

The smallest bias in all the criteria is in bold.

models. Let $\boldsymbol{m}$ and $\boldsymbol{S}$ be the $p \times 1$ vector and $p \times p$ matrix obtained by maximizing the weighted log-likelihood function $L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d})$ in (3), i.e.,

$$\boldsymbol{m} = \frac{1}{\mathrm{tr}(\boldsymbol{W}\boldsymbol{D})}\boldsymbol{Y}'\boldsymbol{D}\boldsymbol{W}\boldsymbol{1}_n, \quad \boldsymbol{S} = -\frac{2}{\mathrm{tr}(\boldsymbol{D})}(\boldsymbol{Y} - \boldsymbol{1}_n\boldsymbol{m}')'\boldsymbol{D}\boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{1}_n\boldsymbol{m}'), \tag{28}$$

where $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_n)$ and $\boldsymbol{W} = \mathrm{diag}(w(r_1), \ldots, w(r_n))$. Here, $w(r) = \{dg(r)/dr\}/g(r)$ and $r_i = (\boldsymbol{y}_i - \boldsymbol{m})'\boldsymbol{S}^{-1}(\boldsymbol{y}_i - \boldsymbol{m})$. We can obtain $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}_{[-i]}$, $\tilde{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\theta}}^{\star}_b$ from formula (28). On the other hand, we prepare the following four distributions for the true model.

- *Normal distribution*: Each of the p variables is generated independently from $N(0, 1)$ ($\kappa^{(1)}_{3,3} = \kappa^{(2)}_{3,3} = 0$ and $\kappa^{(1)}_4 = 0$),
- *Laplace distribution*: Each of the p variables is generated independently from the Laplace distribution $L(0, 1)$ divided by the standard deviation $\sqrt{2}$ ($\kappa^{(1)}_{3,3} = \kappa^{(2)}_{3,3} = 0$ and $\kappa^{(1)}_4 = 2p$),
- *Chi-square distribution*: Each of the p variables is generated independently from the χ^2 distribution with 3 degrees of freedom standardized by the mean 3 and standard deviation $\sqrt{6}$ ($\kappa^{(1)}_{3,3} = \kappa^{(2)}_{3,3} \approx 1.63 \times p$ and $\kappa^{(1)}_4 = 4p$),
- *Log-normal distribution*: Each of the p variables is generated independently from a log-normal distribution $LN(0, 1/4)$ standardized by the mean $e^{1/8}$ and standard deviation $e^{1/8}\sqrt{e^{1/4} - 1}$ ($\kappa^{(1)}_{3,3} = \kappa^{(2)}_{3,3} \approx 1.71 \times p$ and $\kappa^{(1)}_4 \approx 8.90 \times p$).

Table 1 lists the average risk, the biases of the CCV criterion along with the AIC, TIC, EIC and CV criterion, and the frequencies of the model selected by the criteria in the cases of $p = 2$ and 6. These average values were obtained after 10,000 iterations, and the EIC was obtained by resampling 100 times. Moreover, we use $c_n = \sqrt{n/(n+1)}$ as the constant for adjusting $\tilde{\boldsymbol{\theta}}_i$. From the table, we can see that the biases of AIC were large in all the cases. TIC hardly corrected the bias in Models 1 and 2. On the other hand, the biases of the CV criterion were smaller than the biases of AIC and TIC. Especially, the biases of the CV criterion were smaller than the biases of EIC in most cases. Moreover, when we use the CV criterion for model selection, the frequencies of the model with the smallest risk selected was the highest in all the criteria. However, the bias of the CV criterion became large when the dimension p increased. We can see that CCV corrected the bias efficiently.

Next, we compared several methods for correcting the bias in the CV criterion. We prepared the following two different bias-corrected CV criteria from the CCV criterion, which were obtained by adding some bias correction terms

$$\mathrm{CCV}' = \mathrm{CV} - \frac{1}{n}\,\mathrm{tr}(\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})), \quad \mathrm{CCV}'' = \left(1 - \frac{1}{2n}\right)\mathrm{CV} - \frac{1}{n}\,L(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}).$$

Note that the CCV$'$ and CCV$''$ criteria correct the biases to $O(n^{-2})$ as well as the CCV criterion. Table 2 shows the biases of the CV, CCV, CCV$'$ and CCV$''$ criteria. From the table, we can see that CCV$'$ and CCV$''$ did not reduce the bias fully when the bias is large. Therefore, the methods for reducing the bias by adding correction terms should not be used for bias correction. We have studied several other models and have obtained similar results.

Acknowledgments

We wish to express our deepest gratitude to Professor Yasunori Fujikoshi of Chuo University for his valuable advices and encouragements. Also we would like to thank Dr. H. Fujisawa of Institute of Statistical Mathematics and the two referees for several useful comments and suggestions. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), #17700274, 2005–2007.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, In: B.N. Petrov, F. Csáki (Eds.), Second International Symposium on Information Theory, Akadémiai Kiadó, Budapest, 1973, pp. 267–281.

[2] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Control AC-19 (1974) 716–723.

[3] K.T. Fang, S. Kotz, K.W. Ng, Symmetric Multivariate and Related Distributions, Chapman & Hall, CRC, London, 1990.

[4] Y. Fujikoshi, T. Noguchi, M. Ohtaki, H. Yanagihara, Corrected versions of cross-validation criteria for selecting multivariate regression and growth curve models, Ann. Inst. Statist. Math. 55 (2003) 537–553.

[5] Y. Fujikoshi, H. Yanagihara, H. Wakaki, Bias corrections of some criteria for selection multivariate linear regression models in a general case, Amer. J. Math. Management Sci. 25 (2005) 221–258.

[6] M. Ishiguro, Y. Sakamoto, G. Kitagawa, Bootstrapping log likelihood and EIC, an extension of AIC, Ann. Inst. Statist. Math. 49 (1997) 411–434.

[7] S. Konishi, Statistical model evaluation and information criteria, in: S. Ghosh (Ed.), Multivariate Analysis, Design of Experiments, and Survey Sampling, Marcel Dekker, New York, 1999, pp. 369–399.

[8] S. Kullback, R.A. Leibler, On information and sufficiency, Ann. Math. Statist. 22 (1951) 79–86.

[9] E.L. Lehmann, G. Casella, Theory of Point Estimation, second ed., Springer, New York, 1998.

[10] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. Roy. Statist. Soc. Ser. B 36 (1974) 111–147.

[11] M. Stone, An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion, J. Roy. Statist. Soc. Ser. B 39 (1977) 44–47.

[12] K. Takeuchi, Distribution of information statistics and criteria for adequacy of models, Math. Sci. 153 (1976) 12–18 (in Japanese).

[13] H. White, Maximum likelihood estimation of misspecified models, Econometrica 50 (1982) 1–25.

[14] H. Yanagihara, Selection of covariance structure models in nonnormal data by using information criterion: an application to data from the survey of the Japanese national character, Proc. Inst. Statist. Math. 53 (2005) 133–157 (in Japanese).

[15] H. Yanagihara, Corrected version of *AIC* for selecting multivariate normal linear regression models in a general nonnormal case, J. Multivariate Anal. 97 (2006) 1070–1089.

[16] H. Yanagihara, A family of estimators for multivariate kurtosis in a nonnormal linear regression model, J. Multivariate Anal., in press.