Bias Correction of Cross-Validation Criterion Based on Kullback-Leibler Information under a General Condition

Hirokazu YANAGIHARA[1], Tetsuji TONDA[2] AND Chieko MATSUMOTO[3]

[1]*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

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

[3]*Institute of Economic Research*
Hitotsubashi University
2-1 Naka, Kunitachi, Tokyo 186-8603, Japan

Abstract

This paper deals with the bias correction of the cross-validation (CV) criterion for a choice of models. The bias correction is based on the predictive Kullback-Leibler information, which measures the discrepancy between the distributions of an observation for a candidate model and the true model. By replacing an ordinary maximum likelihood estimator with an estimator obtained by maximizing a weighted log-likelihood function, a bias-corrected CV criterion is proposed. This criterion always corrects the bias to $\mathrm{O}(n^{-2})$ under a general condition. We verify that our criterion has smaller bias than the AIC, TIC, EIC and CV criteria by conducting numerical experiments.

[1]Corresponding author, *e-mail*: *yanagi@sk.tsukuba.ac.jp*, tel. & fax: +81-29-853-6451.
(Last Modified: March 30, 2005)

AMS 2000 *subject classifications.* Primary 62H12; Secondary 62F07.
Key words: Bias correction, Cross-validation, Predictive Kullback-Leibler information, Model misspecification, Model selection, Robustness, Weighted log-likelihood function.

1. Introduction

Let $\boldsymbol{y}_i$ $(i = 1, \ldots, n)$ be a $p \times 1$ observation vector, where n is the sample size. Suppose that each $\boldsymbol{y}$ is an independent random sample from an unknown distribution having a density function $\varphi(\boldsymbol{y}_i)$, that is, the true model is

$$M^* : \ \boldsymbol{y}_i \sim i.i.d.\ \varphi(\boldsymbol{y}_i), \ (i = 1, \ldots, n). \tag{1.1}$$

Then we consider a parametric model which consists of a family of probability distributions $\{f(\boldsymbol{y}|\boldsymbol{\theta}); \ \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ is the q-dimensional vector of unknown parameters, and Θ is an open subset of $\Re^q$. Therefore, a candidate model is

$$M : \ \boldsymbol{y}_i \sim i.i.d.\ f(\boldsymbol{y}_i|\boldsymbol{\theta}), \ (i = 1, \ldots, n). \tag{1.2}$$

The Akaike information criterion (AIC) proposed by Akaike (1973) is being used universally for choosing the best model in all candidate models (1.2). It is well known that AIC is an estimator of a risk based on the predictive Kullback-Leibler (K-L) information (Kullback & Leibler, 1951), which measures the discrepancy between $\varphi(\cdot)$ and $f(\cdot|\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$. However, AIC has a constant bias for the risk when $\varphi(\cdot)$ is not consistent with $f(\cdot|\boldsymbol{\theta})$, because Akaike derived AIC only under the assumption that $\varphi(\cdot)$ and $f(\cdot|\boldsymbol{\theta})$ are equal. Takeuchi (1976) revaluated the bias correction term of AIC under the assumption that $\varphi(\cdot)$ and $f(\cdot|\boldsymbol{\theta})$ are equal, and proposed the Takeuchi information criterion (TIC) by replacing the bias correction term of AIC with the revaluated term. The TIC is an asymptotic unbiased estimator for the risk in any model if the true distribution of $\boldsymbol{y}_i$ is *i.i.d.* However, Fujikoshi, Yanagihara and Wakaki

(2005) pointed out that TIC for selecting variables in normal regression models hardly corrects the bias in actual use, because its bias correction term is based on 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 (Yanagihara, 2004b). Like TIC, the cross-validation (CV) criterion proposed by Stone (1974) is known as an asymptotic unbiased estimator for the risk (Stone, 1977), though 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 instead of those of TIC, Yanagihara (2004a, 2005) proposed a new criterion which is partially constructed by the cross-validation method, and which is slightly influenced by the difference of the distributions. However, a bias for the risk exists also in the CV criterion. Fujikoshi et al. (2003) corrected the biases of the CV criteria for selecting 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 several correction terms. We replaced $\hat{\boldsymbol{\theta}}$ with an estimator obtained by maximizing a weighted log-likelihood function, and thus propose a bias-corrected CV criterion. The bias of our criterion is always corrected to $\mathrm{O}(n^{-2})$.

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 our proposed criterion and its asymptotic property. In Section 4, by conducting numerical experiments, we verify that our criterion has smaller bias than other criteria, namely, the AIC, TIC, EIC and CV criterion.

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

$$L(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{d}) = \sum_{i=1}^{n} d_i \log f(\boldsymbol{y}_i|\boldsymbol{\theta}), \tag{2.1}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ and $\boldsymbol{d} = (d_1, \ldots, d_n)'$. For simplicity, if a weight vector $\boldsymbol{d}$ is $\mathbf{1}_n$, where $\mathbf{1}_n$ is an $n \times 1$ vector, all of whose elements are 1, we omit writing $\boldsymbol{d}$ into equation (2.1), i.e., $L(\boldsymbol{\theta}|\boldsymbol{Y}) = L(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n)$. When we assume that each $\boldsymbol{y}_i$ is identically and independently distributed according to a distribution having a density function $f(\boldsymbol{y}_i|\boldsymbol{\theta})$ in (1.2), then an MLE of $\boldsymbol{\theta}$ is obtained by maximizing an ordinary log-likelihood function $L(\boldsymbol{\theta}|\boldsymbol{Y})$, i.e.,

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{Y}). \tag{2.2}$$

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 each $\boldsymbol{u}_i$ is independently distributed according to the same distribution of $\boldsymbol{y}_i$. Then, a risk based on the predictive K-L information, which measures the discrepancy between the true model (1.1) and the candidate model, (1.2) is defined by

$$R_{KL} = \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{u}}\left[-2L(\hat{\boldsymbol{\theta}}|\boldsymbol{U})\right], \tag{2.3}$$

where E^* means an expectation under the true model M^* (1.1).

The AIC proposed by Akaike (1973) is a simple estimator of the risk R_{KL} (2.3). Under the candidate model M (1.2), AIC is defined by

$$AIC = -2L(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + 2q. \tag{2.4}$$

However, if $f(\cdot|\boldsymbol{\theta})$ is not equal to $\varphi(\cdot)$, AIC has a constant bias, i.e.,

$$B_{AIC} = R_{KL} - \mathrm{E}^*_{\boldsymbol{y}}[AIC] = \mathrm{O}(1). \tag{2.5}$$

This is mainly because Akaike derived AIC only under the assumption that $\varphi(\cdot)$ and $f(\cdot|\boldsymbol{\theta})$ are equal. Takeuchi (1976) revaluated the bias correction term of AIC, $2q$, under the inconsistency with $\varphi(\cdot)$ and $f(\cdot|\boldsymbol{\theta})$, and proposed TIC by his revaluation. Under the candidate model M (1.2), TIC is defined by

$$TIC = -2L(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + 2\mathrm{tr}(\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}})), \tag{2.6}$$

where

$$\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}), \quad \hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'. \tag{2.7}$$

Here, $\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})$ and $\boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})$ are a $q\times 1$ vector and a $q\times q$ matrix, respectively, which are based on the partial derivatives up to the second order, that is,

$$\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{y}_i|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \quad \boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(\boldsymbol{y}_i|\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{2.8}$$

Takeuchi (1976) showed that TIC is an asymptotic unbiased estimator for the risk (2.3) in any model if the distribution of $\boldsymbol{y}_i$ is *i.i.d.*, i.e.,

$$B_{TIC} = R_{KL} - \mathrm{E}^*_{\boldsymbol{y}}[TIC] = \mathrm{O}(n^{-1}). \tag{2.9}$$

On the other hand, Stone (1974) proposed the CV criterion for a choice of models (1.2) in the following way. Let $\hat{\boldsymbol{\theta}}_{[-i]}$ be an estimator of $\boldsymbol{\theta}$ obtained by maximizing $\sum_{j\neq i}^{n}\log f(\boldsymbol{y}_j|\boldsymbol{\theta})$, that is,

$$\hat{\boldsymbol{\theta}}_{[-i]} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{1}_n - \boldsymbol{e}_i), \tag{2.10}$$

where $\boldsymbol{e}_i$ is an $n\times 1$ vector whose i-th element is 1 and the other elements are 0. Using $\hat{\boldsymbol{\theta}}_{[-i]}$ (2.10), the CV criterion is given by

$$CV = -2\sum_{i=1}^{n}\log f(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]}). \tag{2.11}$$

Stone (1977) pointed out that the TIC (2.6) and CV criteria (2.11) are asymptotically equivalent, i.e., $CV = TIC + \mathrm{O}_p(n^{-1})$. Therefore, from (2.9), the bias of the CV criterion is given by

$$B_{CV} = R_{KL} - \mathrm{E}^*_{\boldsymbol{y}}[CV] = \mathrm{O}(n^{-1}). \tag{2.12}$$

By comparing (2.9) and (2.12), we can see that the orders of B_{CV} and B_{TIC} are the same. However, Yanagihara (2004a) showed that the CV criterion for selecting variables in normal regression models has smaller bias than TIC.

This is caused by the necessity to estimate higher-order cumulants, because an ordinary estimator of higher-order cumulants tends to underestimate too much, even if the sample size n is moderate. Needless to say, we can obtain the CV criterion without estimating higher-order cumulants. Therefore, even if a calculation is troublesome, we support using the CV criterion for model selection rather than TIC.

3. Bias Correction of the CV Criterion

3.1. Asymptotic Expansion of the Bias of the CV Criterion

In this section, we propose a bias-corrected CV (Corrected CV; CCV) criterion by replacing $\hat{\boldsymbol{\theta}}_{[-i]}$ (2.10) in the CV criterion (2.11) with an estimator which is obtained by maximizing 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 a $q \times 1$ vector which satisfies the following equation.

$$\mathrm{E}_{\boldsymbol{y}}^{*}\left[\boldsymbol{g}(\boldsymbol{y}|\boldsymbol{\theta}_0)\right] = \mathbf{0}. \tag{3.1}$$

Note that the MLE $\hat{\boldsymbol{\theta}}$ (2.2) converges to $\boldsymbol{\theta}_0$ when n is large, i.e., $\lim_{n\to\infty}\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. 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 a regularity condition, the bias of the CV criterion is expanded as*

$$B_{CV} = R_{KL} - \mathrm{E}_{\boldsymbol{y}}^{*}[CV] = -\frac{1}{n}\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) + \mathrm{O}(n^{-2}), \tag{3.2}$$

where

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

(PROOF) We define a $q \times q^2$ matrix based on the partial derivatives up to the third order as

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

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

$$\hat{\boldsymbol{\theta}}_{[-i]} = \hat{\boldsymbol{\theta}} - \frac{1}{n}\boldsymbol{z}_{1,i} - \frac{1}{n^2}\boldsymbol{z}_{2,i} + \mathrm{O}_p(n^{-3}), \tag{3.3}$$

where

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

Here, $\boldsymbol{g}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})$ and $\boldsymbol{H}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})$ are given by (2.8), and $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})$ is given by (2.7). Since the distributions of $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ are the same, the commutative equation $\mathrm{E}^*_{\boldsymbol{y}}[\log f(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]})] = \mathrm{E}^*_{\boldsymbol{y}}\mathrm{E}^*_{\boldsymbol{u}}[\log f(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}}_{[-i]})]$ is held. Therefore, using the Taylor expansion and the equation (3.3), the expectation of the CV criterion is expanded as

$$\mathrm{E}^*_{\boldsymbol{y}}[CV] = R_{KL} - R_1 - \frac{1}{n}R_2 + \mathrm{O}(n^{-2}), \tag{3.4}$$

where

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

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

$$R_1 = \frac{2}{n}\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\mathrm{E}^*_{\boldsymbol{u}}\left[\boldsymbol{g}(\boldsymbol{u}_i|\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{1,i}\Big|\,\boldsymbol{Y}\right]\right] = \frac{2}{n}\sum_{i=1}^{n}\mathrm{E}^*_{\boldsymbol{y}}\left[\boldsymbol{\eta}'\boldsymbol{z}_{1,i}\right] = 0. \tag{3.5}$$

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

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

From (3.1), the first term on the right side of the above equation disappears. Moreover, using equation $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) \to \boldsymbol{J}(\boldsymbol{\theta}_0)$ and $\hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}) \to \boldsymbol{I}(\boldsymbol{\theta}_0)$ $(n \to \infty)$, we derive the following equation.

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

Therefore, we can see that

$$R_2 = -\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) + \mathrm{O}(n^{-1}). \tag{3.6}$$

Substituting R_1 (3.5) and R_2 (3.6) into (3.4) yields

$$\mathrm{E}_{\boldsymbol{y}}^{*}[CV] = R_{KL} + \frac{1}{n}\mathrm{tr}(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}(\boldsymbol{\theta}_0)) + \mathrm{O}(n^{-2}). \tag{3.7}$$

Consequently, the result (3.2) in Theorem 1 is obtained.

3.2. Corrected CV Criterion

Next, we propose a new criterion, a corrected CV criterion, which always corrects the bias for the risk (2.3) to $\mathrm{O}(n^{-2})$. 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 (2.10), 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} (2.3) is not the predictive K-L information measuring the discrepancy between $\varphi(\cdot)$ and $f(\cdot|\hat{\boldsymbol{\theta}}_{[-i]})$, but, rather, $\varphi(\cdot)$ and $f(\cdot|\hat{\boldsymbol{\theta}})$. Thus, we use an 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 bias-corrected CV criterion.

DEFINITION. *Let $\tilde{\boldsymbol{\theta}}_i$ be an estimator of $\boldsymbol{\theta}$ by maximizing a weighted log-likelihood function as*

$$\tilde{\boldsymbol{\theta}}_i = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{Y}, \mathbf{1}_n - c_n \boldsymbol{e}_i), \tag{3.8}$$

where $c_n = \sqrt{n/(n+1)}$. Then, we propose the bias-corrected CV (CCV) criterion as

$$CCV = -2\sum_{i=1}^{n} \log f(\boldsymbol{y}_i|\tilde{\boldsymbol{\theta}}_i). \tag{3.9}$$

From the definition of CCV, we can see that any estimators of higher-order cumulants are not necessary for obtaining CCV. However, CCV always corrects the bias to $\mathrm{O}(n^{-2})$, 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 (3.9). The order of bias of the CCV criterion is obtained in the following theorem.

THEOREM 2. *Under a regularity condition, the order of a bias of the CCV criterion is given by*

$$B_{CCV} = R_{KL} - \mathrm{E}^*_{\boldsymbol{y}}[CCV] = \mathrm{O}(n^{-2}). \tag{3.10}$$

(PROOF) From the definition of $\tilde{\boldsymbol{\theta}}_i$ (3.8) and the Taylor expansion, we expand $\tilde{\boldsymbol{\theta}}_i$ as

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

Thus, the perturbation expansion of CCV is given by

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

Therefore, we calculate the expectation of CCV as

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

Substituting equation (3.7) into the above equation yields the equation (3.10) in Theorem 2.

4. Numerical Examination

In this section, we examine the numerical studies for average biases and frequencies of the selected model according to the criteria. First, we prepare another bias-corrected criterion constructed by the bootstrap method, which was named the empirical information criterion (EIC) by Ishiguro, Sakamoto and Kitagawa (1997). Let $\boldsymbol{Y}_b^*$ $(b = 1, \ldots, B)$ be the b-th bootstrap data matrix by resampling, and $\hat{\boldsymbol{\theta}}_b^*$ be the MLE of $\boldsymbol{\theta}$ based on $\boldsymbol{Y}_b^*$, i.e.,

$$\hat{\boldsymbol{\theta}}_b^* = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\boldsymbol{Y}_b^*). \tag{4.1}$$

Then, the EIC for the selected model (1.2) is given by

$$EIC = -2L(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) - \frac{2}{B}\sum_{b=1}^{B}\left\{L(\hat{\boldsymbol{\theta}}_b^*|\boldsymbol{Y}) - L(\hat{\boldsymbol{\theta}}_b^*|\boldsymbol{Y}_b^*)\right\}, \tag{4.2}$$

(see e.g., Konishi, 1999). Through the simulation, we compare the biases and frequencies of the selected model in our proposed CCV criterion (3.9), and also the AIC (2.4), TIC (2.6), EIC (4.2) and CV criteria (2.11).

In this paper, we deal with the selection of the best model of the candidate models (1.2) having an elliptical distribution, i.e.,

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}) = c_p|\boldsymbol{\Lambda}|^{-1/2}g((\boldsymbol{y}_i - \boldsymbol{\mu})'\boldsymbol{\Lambda}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})), \quad (i = 1, \ldots, 20), \tag{4.3}$$

where $g(r)$ is a known non-negative function and c_p is a positive constant depending on the dimension p (see e.g., Fang, Kotz & Ng, 1990). We choose the best $g(r)$ and c_p in the candidate models by minimizing the information criteria. The candidate models considered are as follows.

Model 1: *Multivariate Normal Distribution*,

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

Model 2: *Multivariate Logistic Distribution*,

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

Model 3: Multivariate Cauchy Distribution,

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

where $\Gamma(\cdot)$ 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 an information criterion, since the normal distribution is in the candidate 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 $L(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{d})$ (2.1) under the candidate model (4.3), i.e.,

$$\boldsymbol{m} = \frac{1}{\mathrm{tr}(\boldsymbol{WD})}\boldsymbol{Y}'\boldsymbol{DW}\boldsymbol{1}_n, \ \boldsymbol{S} = -\frac{2}{\mathrm{tr}(\boldsymbol{D})}(\boldsymbol{Y}-\boldsymbol{1}_n\boldsymbol{m}')'\boldsymbol{DW}(\boldsymbol{Y}-\boldsymbol{1}_n\boldsymbol{m}'), \quad (4.4)$$

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 $\hat{\boldsymbol{\theta}}$ (2.2), $\hat{\boldsymbol{\theta}}_{[-i]}$ (2.10), $\tilde{\boldsymbol{\theta}}_i$ (3.8) and $\hat{\boldsymbol{\theta}}_b^*$ (4.1) from the formula (4.4). On the other hand, we prepare the following four distributions for the true model (1.1).

Normal Distribution: Each of the p variables is generated independently from N(0,1) ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 0$ and $\kappa_4^{(1)} = 0$),

Laplace Distribution: Each of the p variables is generated independently from the Laplace distribution $\mathrm{L}(0,1)$ divided by the standard deviation $\sqrt{2}$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} = 0$ and $\kappa_4^{(1)} = 2p$),

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 $\sqrt{6}$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} \approx 1.63 \times p$ and $\kappa_4^{(1)} = 4p$),

Log-Normal Distribution: Each of the p variables is generated independently from a log-normal distribution $\mathrm{LN}(0, 1/4)$ standardized by the mean $e^{1/8}$ and standard deviation $e^{1/2}\sqrt{e^{1/4}-1}$ ($\kappa_{3,3}^{(1)} = \kappa_{3,3}^{(2)} \approx 1.71 \times p$ and $\kappa_4^{(1)} \approx 8.90 \times p$).

Table 1 lists the average risk, the biases of the CCV criterion along with the AIC, TIC EIC and CV criteria, and the frequencies of the model selected by the criteria in the cases of $p = 2$ and $p = 6$. These average values were obtained after 10,000 iterations, and the EIC was obtained by resampling 100 times. 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.

Insert Table 1 around here

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 (3.9), which were obtained by adding some bias correction terms.

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

Note that the CCV' and CCV'' criteria correct the biases to $\mathrm{O}(n^{-2})$ 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.

Insert Table 2 around here

Acknowledgment

We wish to express our deepest gratitude to Professor Yasunori Fujikoshi of Hiroshima University for his valuable advice and encouragement.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csáki, pp. 267–281. Akadémiai Kiadó, Budapest.

[2] Fang, K. T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions*. Chapman & Hall/CRC, London.

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

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

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

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

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

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

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

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

[11] Yanagihara, H. (2004a). Corrected version of *AIC* for selecting multivariate normal linear regression models in a general nonnormal case. Technical Report, No. 04-04, *Statistical Research Group*, Hiroshima University, Hiroshima.

[12] Yanagihara, H. (2004b). A family of estimators for multivariate kurtosis in a nonnormal linear regression model. Technical Report, No. 04-08, *Statistical Research Group*, Hiroshima University, Hiroshima.

[13] Yanagihara, H. (2005). 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** (in press, in Japanese).

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

Distribution	Criterion		$p=2$ Model 1	$p=2$ Model 2	$p=2$ Model 3	$p=6$ Model 1	$p=6$ Model 2	$p=6$ Model 3
Normal		Risk	120.50*	123.45	132.31	399.91*	405.21	404.71
	AIC	Bias	2.30	5.54	1.77	37.43	43.81	15.90
		Freq.	(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
		Freq.	(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**
		Freq.	(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
		Freq.	(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
		Freq.	(71.0)	(22.5)	(6.5)	(62.8)	(5.5)	(31.7)
Laplace		Risk	125.27	135.24	121.32*	422.27	434.96	387.84*
	AIC	Bias	9.93	17.42	1.47	67.59	79.65	17.69
		Freq.	(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
		Freq.	(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**
		Freq.	(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
		Freq.	(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
		Freq.	(47.9)	(3.1)	(49.0)	(19.9)	(0.3)	(79.8)
Chi-Square		Risk	126.92	135.41	122.52*	426.64	437.16	387.63*
	AIC	Bias	12.06	18.76	2.82	71.44	82.00	17.15
		Freq.	(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
		Freq.	(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
		Freq.	(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
		Freq.	(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**
		Freq.	(38.8)	(14.4)	(46.8)	(20.1)	(0.9)	(79.0)
Log-Normal		Risk	128.07	137.34	121.90*	427.94	439.32	384.36*
	AIC	Bias	13.85	21.38	2.71	75.71	87.04	17.12
		Freq.	(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
		Freq.	(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**
		Freq.	(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
		Freq.	(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
		Freq.	(41.6)	(11.9)	(46.5)	(19.9)	(1.2)	(78.9)

* denotes the smallset 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.