Iterative Bias Correction of the Cross-Validation Criterion

(Last Modified: February 29 2008)

Hirokazu YANAGIHARA[1] AND Hironori FUJISAWA[2]

[1] *Department of Mathematics, Graduate School of Science, Hiroshima University*
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

[2] *Department of Mathematical Analysis and Statistical Inference*
The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan

Abstract

The cross-validation (CV) criterion is known to be a second-order unbiased estimator of the risk function measuring the discrepancy between the candidate model and the true model, as well as the generalised information criterion (GIC) and the extended information criterion (EIC). In the present paper, we show that the $2k$th-order unbiased estimator can be obtained using a linear combination from the leave-one-out CV criterion to the leave-k-out CV criterion. The proposed scheme is unique in that a bias smaller than that of a jackknife method can be obtained without any analytic calculation, i.e., it is not necessary to obtain the explicit form of several terms in an asymptotic expansion of the bias. Furthermore, the proposed criterion can be regarded as a finite correction of the bootstrap iterative type of CV criterion.

AMS 2000 *subject classifications*: Primary 62H12; Secondary 62F07.
Key words: Asymptotic expansion, Bias correction, Bootstrap iteration, Cross-validation criterion, EIC, GIC, Leave-k-out cross-validation, Model selection.

1. Introduction

Let $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ be the p-dimensional independent random vectors from $\boldsymbol{y}$, where n is the sample size. Assume that $\boldsymbol{y}$ is a random vector having an unknown probability density function $\varphi(\boldsymbol{y})$. Hence, the true model is expressed as

$$M_\varphi : \ \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \sim i.i.d. \ \varphi(\boldsymbol{y}). \tag{1.1}$$

[1] Corresponding author, E-mail: *yanagi@math.sci.hiroshima-u.ac.jp*

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

$$M:\ \boldsymbol{y}_1, \dots, \boldsymbol{y}_n \sim i.i.d.\ f(\boldsymbol{y}|\boldsymbol{\theta}). \tag{1.2}$$

The best model is chosen by minimizing the risk function of the model. The risk function is based on the predictive discrepancy between $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ and $\varphi(\boldsymbol{y})$. Here, $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$, which is obtained by minimizing the discrepancy between M and M_φ. These two discrepancies are sometimes different, for example, $\boldsymbol{\theta}$ is estimated by a penalized log-likelihood function, and the risk function is generated by the Kullback-Leibler (K-L) discrepancy (Kullback & Leibler, 1951). Therefore, we assume that the two discrepancies are not always the same. Several estimates of the risk function have been proposed, e.g., the cross-validation (CV) criterion (Stone, 1974, 1977), the generalised information criterion (GIC; Konishi & Kitagawa, 1996) and the extended information criterion (EIC; Ishiguro, Sakamoto & Kitagawa, 1997). The three criteria are the second-order unbiased estimators of the risk function, i.e., the biases of criteria are $O(n^{-2})$. The purpose of the present paper is to propose a higher order unbiased estimator of the risk function without complicated analytic calculations.

A bias correction is generally achieved by subtracting an estimated bias from the target estimator. The estimated bias usually depends on estimators of the higher-order cumulants of φ. Note that it is difficult to obtain good estimates of the higher-order cumulants even when n is relatively large (see Yanagihara, 2007, for the case of kurtosis). Therefore, in the present paper, we attempt to reduce the bias of the CV criterion without estimating higher-order cumulants.

The CV criterion is based on the leave-one-out concept. A leave-k-out CV criterion (see e.g., Shao, 1993) can also be proposed by extending this concept. Using a linear combination from the leave-one-out CV criterion to the leave-k-out CV criterion, we propose a bias-corrected CV criterion, which becomes the $2k$th-order unbiased estimator of the risk function, i.e., the bias of the proposed criterion is $O(n^{-2k})$. The proposed scheme is unique in that a bias that is smaller than that of a jackknife method can be obtained without any analytic calculation, i.e., it is not necessary to obtain the explicit form of several terms in an asymptotic expansion of the bias. We can sometimes reduce the bias by the bootstrap iteration (see e.g., Efron, 1983; Hall & Martin, 1988; Hall, 1992, p.

28). The bias-corrected criterion obtained by the kth bootstrap iteration is the $(k+1)$th-order unbiased estimator for the risk function. However, the proposed criterion is the $2k$th-order unbiased estimator when we use the leave-j-out CV criterion for $j = 1, \ldots, k$. We will see that coefficients of the kth bootstrap iteration are asymptotically the same as those of the kth bias-corrected CV criterion. In other words, the coefficients of the kth bias-corrected CV criterion are finitely adjusted versions of the coefficients of the kth bootstrap iteration.

The remainder of the present paper is organized as follows. In Section 2, we propose the kth bias-corrected CV criterion by using a linear combination from the leave-one-out CV criterion to the leave-k-out CV criterion. Then, we describe the relation between the proposed criterion and the bias-corrected criterion obtained by bootstrap iteration. In Section 3, we investigate the performances of the proposed criteria by conducting numerical simulations. In Section 4, we present a discussion and our conclusions. Technical details are provided in the Appendix.

2. Higher-Order Bias-Corrected CV

2.1. Preliminary

Assume that $\psi(\boldsymbol{y}|\boldsymbol{\theta})$ is a discrepancy function for the candidate model M in (1.2). A typical example is the K-L type, i.e., $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta})$. Let $\mathcal{I}_k$ be a set of indices on observation vectors, which is given by

$$\mathcal{I}_k = \{i_1, \ldots, i_k \in \mathbb{N} : 1 \le i_1 < \cdots < i_k \le n\}.$$

The leave-k-out estimator of $\boldsymbol{\theta}$ is defined by

$$\hat{\boldsymbol{\theta}}_{[-\mathcal{I}_k]} = \arg\min_{\boldsymbol{\theta}} \sum_{i \notin \mathcal{I}_k} \psi(\boldsymbol{y}_i|\boldsymbol{\theta}). \tag{2.1}$$

Note that $\hat{\boldsymbol{\theta}}_{[-\mathcal{I}_0]}$ and $\hat{\boldsymbol{\theta}}_{[-\mathcal{I}_1]}$ denote the ordinary estimator $\hat{\boldsymbol{\theta}}$ obtained by the full sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ and the ith jackknife estimator $\hat{\boldsymbol{\theta}}_{[-i]}$ obtained by ith jackknife sample, which is derived from $\boldsymbol{Y}$ by deleting $\boldsymbol{y}_i$. It is easy to see that $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ when $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta})$ and that $\hat{\boldsymbol{\theta}}$ is the penalized MLE, if $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta}) + \lambda\pi(\boldsymbol{\theta})/2$, where $\pi(\boldsymbol{\theta})$ is some function of $\boldsymbol{\theta}$ and λ is the hyper-parameter. Basu *et al.* (1996) proposed the density power divergence with the tuning parameter α, which includes the K-L divergence and the L_2 distance as special cases, i.e., $\alpha = 0$ and 1, respectively. The corresponding discrepancy is given by

$$\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\frac{1}{\alpha} f(\boldsymbol{y}|\boldsymbol{\theta})^{\alpha} + \frac{1}{1+\alpha} \int \{f(\boldsymbol{x}|\boldsymbol{\theta})\}^{1+\alpha} d\boldsymbol{x}. \tag{2.2}$$

As other discrepancies, we can also use, for example, the residuals of sum of square (RSS), the quasi-likelihood (Wedderburn, 1974), the extended quasi-likelihood (Nelder & Pregibon, 1987), and the γ-divergence (Fujisawa & Eguchi, 2008).

We evaluate the fit of the model by another discrepancy $\gamma(\boldsymbol{y}|\boldsymbol{\theta})$. Note that $\psi(\boldsymbol{y}|\boldsymbol{\theta})$ and $\gamma(\boldsymbol{y}|\boldsymbol{\theta})$ are not always the same. For example, we estimate $\boldsymbol{\theta}$ by the penalized log-likelihood $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta}) + \lambda\pi(\boldsymbol{\theta})$ and evaluate the model by the K-L discrepancy $\gamma(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta})$ (see e.g., Konishi & Kitagawa, 1996, 2003, 2008; Imoto & Konishi, 2003). In addition, we estimate $\boldsymbol{\theta}$ by (2.2) and evaluate the model by another discrepancy (Fujisawa & Eguchi, 2006). Other examples are also shown in Ray and Lindsay (2008) and Lindsay *et al.* (2008). The $\gamma(\boldsymbol{y}|\boldsymbol{\theta})$ yields the target risk function as follows. Let $\rho(\boldsymbol{\theta})$ denote the expectation of the discrepancy $\gamma(\boldsymbol{y}|\boldsymbol{\theta})$, i.e.,

$$\rho(\boldsymbol{\theta}) = E_{\varphi}[\gamma(\boldsymbol{y}|\boldsymbol{\theta})],$$

where E_{φ} is the expectation under the true model M_{φ} in (1.1). Then, the risk function of the model $f(\boldsymbol{y}|\hat{\boldsymbol{\theta}})$ is defined by

$$R_{\gamma} = E_{\varphi}[\rho(\hat{\boldsymbol{\theta}})]. \tag{2.3}$$

In the model selection based on $\gamma(\boldsymbol{y}|\boldsymbol{\theta})$, we regard the model having the smallest R_{γ} as the best model, which is typically different from the true model. In many contexts of statistical modeling, the aim is to determine the best model. Obtaining an unbiased estimator of R_{γ} will allow us to correctly evaluate the discrepancy between the data and the model, which will further facilitate the selection of the best model.

The ordinary CV criterion proposed by Stone (1974, 1977) is an asymptotic unbiased estimator of R_{γ}. The CV criterion is defined by

$$\mathrm{CV} = \frac{1}{n} \sum_{i=1}^{n} \gamma(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]}). \tag{2.4}$$

From Theorem A.1 in Appendix A.1, we can see that CV is the second-order unbiased estimator of R_{γ}, i.e., $E_{\varphi}[\mathrm{CV}] = R_{\gamma} + O(n^{-2})$. Thus, we assume that the expectation of CV can be expanded as follows:

Assumption: $E_\varphi[\mathrm{CV}]$ can be expanded until the n^{-L} term as

$$\begin{aligned} E_\varphi[\mathrm{CV}] &= R_\gamma + \frac{1}{n^2}\beta_2 + \frac{1}{n^3}\beta_3 + \frac{1}{n^4}\beta_4 + \frac{1}{n^5}\beta_5 + \cdots \\ &= R_\gamma + \sum_{l=2}^{L} \frac{\beta_l}{n^l} + O(n^{-L-1}). \end{aligned} \tag{2.5}$$

An explicit form of β_2 is given by (A.6) in Appendix A.1. It is possible to obtain the coefficients β_l $(l \geq 3)$ by repeating tedious calculations, as described in Appendix A.1. However, explicit forms of β_l are not needed in the subsequent discussion.

The leave-k-out CV criterion (see e.g., Shao, 1993) is defined by

$$\mathrm{CV}_k = \frac{1}{k_n C_k} \sum_{\mathcal{I}_k \subset \mathcal{I}_n} \sum_{i \in \mathcal{I}_k} \gamma(\boldsymbol{y}_i | \hat{\boldsymbol{\theta}}_{[-\mathcal{I}_k]}), \tag{2.6}$$

where ${}_nC_k$ is the binominal coefficient given by ${}_nC_k = n!/\{(n-k)!k!\}$. The leave-one-out CV criterion is the ordinary CV criterion, i.e., $\mathrm{CV}_1 = \mathrm{CV}$. Let $\hat{\boldsymbol{\theta}}_n$ denote an estimator of $\boldsymbol{\theta}$ evaluated from n observations. Note that $\boldsymbol{y}_i$ and $\hat{\boldsymbol{\theta}}_{[-\mathcal{I}_k]}$ are mutually independent when $i \in \mathcal{I}_k$. It follows that $E_\varphi[\gamma(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]})] = \rho(\hat{\boldsymbol{\theta}}_{[-i]})$, and then $E_\varphi[\mathrm{CV}_k] = \sum_{\mathcal{I}_k \subset \mathcal{I}_n} E_\varphi[\rho(\hat{\boldsymbol{\theta}}_{[-\mathcal{I}_k]})]/{}_nC_k = E_\varphi[\rho(\hat{\boldsymbol{\theta}}_{n-k})]$. Recall that $R_\gamma = E_\varphi[\rho(\hat{\boldsymbol{\theta}}_n)]$. Therefore, we obtain the following expansion from the assumption (2.5):

$$E_\varphi[\rho(\hat{\boldsymbol{\theta}}_{n-1})] = E_\varphi[\rho(\hat{\boldsymbol{\theta}}_n)] + \sum_{l=2}^{L} \frac{\beta_l}{n^l} + O(n^{-(L+1)}).$$

It then holds that

$$\begin{aligned} E_\varphi[\mathrm{CV}_{j+1}] &= \frac{1}{(j+1)_n C_k} \sum_{\mathcal{I}_{j+1} \subset \mathcal{I}_n} \sum_{i \in \mathcal{I}_{j+1}} E_\varphi\left[\gamma(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-\mathcal{I}_{j+1}]})\right] \\ &= E_\varphi[\rho(\hat{\boldsymbol{\theta}}_{n-j-1})] \\ &= E_\varphi[\rho(\hat{\boldsymbol{\theta}}_{n-j})] + \sum_{l=2}^{L} \frac{\beta_l}{(n-j)^l} + O(n^{-L-1}) \\ &= E_\varphi[\mathrm{CV}_j] + \sum_{l=2}^{L} \frac{\beta_l}{n^l} a_j^l + O(n^{-L-1}), \end{aligned} \tag{2.7}$$

where the coefficient a_j is given by

$$a_j = \frac{n}{n-j}. \tag{2.8}$$

2.2. Basic Concept of Bias Correction

If we use an ordinary procedure for the bias correction, we propose a bias-corrected CV criterion by subtracting $\hat{\beta}_2/n^2$ from CV, where $\hat{\beta}_2$ is a consistent estimator of β_2. However, this procedure does not work well, even when n is relatively large, because $\hat{\beta}_2$ usually depends on the higher-order cumulants of φ. Therefore, we attempt to correct the bias of the CV criterion without directly estimating β_l.

First, we show the bias correction method using CV_1 and CV_2. We see that the bias of the CV criterion may be improved from $O(n^{-2})$ to $O(n^{-3})$. Through a detailed verification, we furthermore observe that the bias of the CV criterion is automatically corrected to $O(n^{-4})$. Let $\Delta_{1,j}$ be the difference between CV_{j+1} and CV_j standardized by a_j^2, i.e.,

$$\Delta_{1,j} = \frac{1}{a_j^2}\left(\mathrm{CV}_{j+1} - \mathrm{CV}_j\right). \tag{2.9}$$

From (2.7), we can see that

$$\begin{aligned} E_\varphi\left[\Delta_{1,1}\right] &= \frac{1}{n^2}\beta_2 + \frac{1}{n^3}a_1\beta_3 + \frac{1}{n^4}a_1^2\beta_4 + \frac{1}{n^5}a_1^3\beta_5 + O(n^{-6}) \\ &= \frac{1}{n^2}\beta_2 + \frac{1}{n^3}a_1\beta_3 + O(n^{-4}). \end{aligned}$$

It follows that the first term in the above expansion is equal to the first term in the expansion of the bias of CV in (2.5). Therefore, when we define the bias-corrected CV criterion as $\mathrm{CV} - \Delta_{1,1}$, its expectation is expanded as

$$\begin{aligned} E_\varphi\left[\mathrm{CV} - \Delta_{1,1}\right] &= R_\gamma + \frac{1}{n^3}(1-a_1)\beta_3 + \frac{1}{n^4}(1-a_1^2)\beta_4 + \frac{1}{n^5}(1-a_1^3)\beta_5 + O(n^{-6}) \\ &= R_\gamma + \frac{1}{n^3}(1-a_1)\beta_3 + O(n^{-4}). \end{aligned} \tag{2.10}$$

Note that $1-a_1 = O(n^{-1})$. This implies that $E_\varphi[\mathrm{CV}-\Delta_{1,1}] = R_\gamma + O(n^{-4})$. Consequently, $\mathrm{CV} - \Delta_{1,1}$ becomes the fourth-order, rather than the third-order, unbiased estimator of R_γ.

Next, we show the bias correction method using CV_1, CV_2 and CV_3. The bias of CV criterion may be improved from $O(n^{-2})$ to $O(n^{-4})$. Through a detailed verification, we furthermore observe that the bias of the CV criterion is automatically corrected to $O(n^{-6})$. Let us define $\Delta_{k,j}$ as the following recursion formula:

$$\Delta_{k,j} = \frac{1}{a_j - a_{j-k+1}}\left(\Delta_{k-1,j} - \Delta_{k-1,j-1}\right), \quad \text{for } 2 \le k < j. \tag{2.11}$$

Using (2.11), we have

$$\Delta_{2,2} = \frac{1}{a_2 - a_1}\left\{\frac{1}{a_2^2}(\mathrm{CV}_3 - \mathrm{CV}_2) - \frac{1}{a_1^2}(\mathrm{CV}_2 - \mathrm{CV}_1)\right\}.$$

From equation (2.7), we can see that

$$\frac{1}{a_j^2} E_\varphi[\mathrm{CV}_{j+1} - \mathrm{CV}_j] = \frac{1}{n^2}\beta_2 + \frac{1}{n^3} a_j \beta_3 + \frac{1}{n^4} a_j^2 \beta_4 + \frac{1}{n^5} a_j^3 \beta_5 + O(n^{-6}).$$

Hence, the expectation of $\Delta_{2,2}$ is expanded as

$$\begin{aligned} E_\varphi\left[\Delta_{2,2}\right] &= \frac{1}{a_2 - a_1}\left\{\frac{1}{n^3}(a_2 - a_1)\beta_3 + \frac{1}{n^4}(a_2^2 - a_1^2)\beta_4 + \frac{1}{n^5}(a_2^3 - a_1^3)\beta_5\right\} + O(n^{-6}) \\ &= \frac{1}{n^3}\beta_3 + \frac{1}{n^4}(a_1 + a_2)\beta_4 + \frac{1}{n^5}(a_1^2 + a_1 a_2 + a_2^2)\beta_5 + O(n^{-6}). \end{aligned}$$

It follows that the quantity $(1-a_1)$ times the first term in the above expansion is equivalent to the first term in the expansion of the bias of $\mathrm{CV} - \Delta_{1,1}$ in (2.10). Therefore, when we define the second bias-corrected CV criterion as $\mathrm{CV} - \Delta_{1,1} - (1-a_1)\Delta_{2,2}$, its expectation is expanded as

$$\begin{aligned} &E_\varphi\left[\mathrm{CV} - \Delta_{1,1} - (1-a_1)\Delta_{2,2}\right] \\ &\qquad = R_\gamma + \frac{1}{n^4}(1-a_1)(1-a_2)\beta_4 + \frac{1}{n^5}(1-a_1)(1-a_2)(1+a_1+a_2)\beta_5 + O(n^{-6}). \end{aligned}$$

Note that $(1-a_1)(1-a_2) = O(n^{-2})$. This implies that $E_\varphi[\mathrm{CV} - \Delta_{1,1} - (1-a_1)\Delta_{2,2}] = R_\gamma + O(n^{-6})$. Consequently, $\mathrm{CV} - \Delta_{1,1} - (1-a_1)\Delta_{2,2}$ becomes the sixth-order, rather than the fourth-order, unbiased estimator of R_γ.

2.3. General Formula

By repeating the technique described in Section 2.2, it is possible to make an even higher-order bias-corrected CV criterion. The general formula is defined as follows:

DEFINITION: The kth bias-corrected CV (Corrected CV; CCV) criterion is defined as

$$\mathrm{CCV}_k = \mathrm{CCV}_{k-1} - c_{k-2}\Delta_{k-1,k-1} = \mathrm{CV} + \sum_{j=1}^{k-1} c_{j-1}\Delta_{j,j}, \tag{2.12}$$

where the coefficient c_j is given by

$$c_j = \begin{cases} 1 & (j = 0) \\ \prod_{l=1}^{j}(1-a_l) & (j \geq 1) \end{cases},$$

and $\Delta_{k,j}$ is given by (2.9) and (2.11).

The order of the bias of CCV_k is shown in the following theorem. The proof is given in Appendix A.2.

Theorem 2.1: *Assume that Assumption* (2.5) *holds. Then, the* CCV_k *becomes the* $2k$*th-order unbiased estimator of* R_γ*, i.e.,* $E_\varphi[\mathrm{CCV}_k] = R_\gamma + O(n^{-2k})$.

When we apply the general formula (2.12) to the actual situation, it may be somewhat troublesome to calculate CCV_k because the definition (2.12) is based on a recursive formula. However, the formula can be rewritten as in the following theorem. The proof is given in Appendix A.3.

Theorem 2.2: CCV_k *is rewritten as the following linear combination from* CV_1 *to* CV_k*:*

$$\mathrm{CCV}_k = \sum_{j=1}^{k} m_{k,j} \mathrm{CV}_j, \tag{2.13}$$

where the coefficient $m_{k,j}$ *is given by*

$$m_{k,j} = \begin{cases} (-1)^{j+1} \left({}_{k-1}C_{j-1} a_{j-1}^{-k} + {}_{k-1}C_j a_j^{-k} \right) & (1 \le j \le k-1) \\ (-1)^{k+1} a_{k-1}^{-k} & (j = k) \end{cases} \tag{2.14}$$

Since the coefficient $m_{1,1}$ becomes 1 when $k = 1$, CCV_1 is equal to the ordinary CV in (2.4). Moreover, when $k = 2$ and $k = 3$, the coefficient $m_{k,j}$ becomes

$$m_{2,1} = \frac{n^2 + (n-1)^2}{n^2}, \quad m_{2,2} = -\left(\frac{n-1}{n} \right)^2,$$
$$m_{3,1} = \frac{n^3 + 2(n-1)^3}{n^3}, \quad m_{3,2} = -\frac{2(n-1)^3 + (n-2)^3}{n^3}, \quad m_{3,3} = \left(\frac{n-2}{n} \right)^3.$$

Note that the coefficient $m_{k,j}$ does not depend on unknown parameters, but depends only on the sample size. Therefore, we could correct the bias of the CV criterion without estimating higher-order cumulants of φ.

2.4. Relation between CCV_k and Bootstrap Iteration

Let $\hat{\boldsymbol{\theta}}(\boldsymbol{w})$ be an estimator of $\boldsymbol{\theta}$ obtained by minimizing a weighted discrepancy function, which is given by

$$\hat{\boldsymbol{\theta}}(\boldsymbol{w}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} w_i \psi(\boldsymbol{y}_i | \boldsymbol{\theta}),$$

where $\boldsymbol{w} = (w_1, \ldots, w_n)'$ and $\sum_{i=1}^{n} w_i = n$. Moreover, let $\Gamma(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{w})$ be another weighted discrepancy given by

$$\Gamma(\boldsymbol{\theta}|\boldsymbol{Y}; \boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} w_i \gamma(\boldsymbol{y}_i | \boldsymbol{\theta}). \tag{2.15}$$

For simplicity, we express $\Gamma(\boldsymbol{\theta}|\boldsymbol{Y}) = \Gamma(\boldsymbol{\theta}|\boldsymbol{Y}; \mathbf{1}_n)$, where $\mathbf{1}_n$ is the n-dimensional vector, the elements of which are 1.

Let $\mathrm{Multi}_n(m : \boldsymbol{p})$ denote the multinomial distribution, where the number of events is m and the cell probability vector is $\boldsymbol{p} = (p_1, \ldots, p_n)'$. Assume that $\boldsymbol{d}_b = (d_{b1}, \ldots, d_{bn})'$ $(b = 1, \ldots, B)$ are the n-dimensional independent random vectors from $\mathrm{Multi}_n(n : \mathbf{1}_n/n)$. From Yanagihara *et al.* (2008), the EIC-type criterion can be expressed as

$$\mathrm{EIC} = \Gamma(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + \frac{1}{B}\sum_{b=1}^{B} \Gamma(\hat{\boldsymbol{\theta}}(\boldsymbol{d}_b)|\boldsymbol{Y}; \mathbf{1}_n - \boldsymbol{d}_b). \tag{2.16}$$

When $B \to \infty$, the bias of EIC has the same order as that of CV, i.e., $E_\varphi[\mathrm{EIC}] = R_\gamma + O(n^{-2})$.

By using the bootstrap iteration, we can improve the bias of EIC. Let $\boldsymbol{d}_{(b_1,b_2)}$ $(b_2 = 1, \ldots, B)$ be the n-dimensional independent random vectors from $\mathrm{Multi}_n(n : \boldsymbol{d}_{(b_1)}/n)$, where $\boldsymbol{d}_{(b_1)}$ $(b_1 = 1, \ldots, B)$ are also the n-dimensional independent random vectors from $\mathrm{Multi}_n(n : \mathbf{1}_n/n)$. Then, we define the second bootstrap bias correction term as

$$D_2 = \frac{1}{B^2}\sum_{b_1,b_2}^{B} \Gamma(\hat{\boldsymbol{\theta}}(\boldsymbol{d}_{(b_1,b_2)})|\boldsymbol{Y}; \boldsymbol{d}_{(b_1)} - \boldsymbol{d}_{(b_1,b_2)}),$$

where the notation $\sum_{b_1,\ldots,b_k}^{B}$ means $\sum_{b_1=1}^{B} \cdots \sum_{b_k=1}^{B}$. Let $\boldsymbol{b}_k = (b_1, \ldots, b_k)$. In the same manner as in definition of D_2, we define the kth bootstrap bias correction term as

$$D_k = \frac{1}{B^k}\sum_{b_1,\ldots,b_k}^{B} \Gamma(\hat{\boldsymbol{\theta}}(\boldsymbol{d}_{\boldsymbol{b}_k})|\boldsymbol{Y}; \boldsymbol{d}_{\boldsymbol{b}_{k-1}} - \boldsymbol{d}_{\boldsymbol{b}_k}), \tag{2.17}$$

where $\boldsymbol{d}_{\boldsymbol{b}_0} = \mathbf{1}_n$. By applying the general formula of the bias correction obtained by kth bootstrap iteration (see e.g., Efron, 1983; Hall & Martin, 1988; Hall, 1992, p. 28) to EIC, the kth bias-corrected EIC (Corrected EIC; CEIC) can be given by

$$\mathrm{CEIC}_k = \Gamma(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + \sum_{j=1}^{k} m_{k,j}^{(0)} D_k, \tag{2.18}$$

where $m_{k,j}^{(0)} = (-1)^{j+1}{}_kC_j$. When $B \to \infty$, the bias of CEIC_k is $O(n^{-k-1})$.

Note that $\lim_{n\to\infty} a_j = 1$. Therefore, the coefficient $m_{k,j}$ in CCV_k converges to $(-1)^{j+1}({}_{k-1}C_{j-1} + {}_{k-1}C_j) = (-1)^{j+1}{}_kC_j$. This implies that $\lim_{n\to\infty} m_{k,j} = m_{k,j}^{(0)}$. Consequently, we can regard $m_{k,j}$ as the finite correction of $m_{k,j}^{(0)}$. In fact, by replacing $m_{j,k}$ in (2.13) with $m_{k,j}^{(0)}$, we can also define another bias-corrected CV criterion, i.e.,

$\mathrm{CCV}_k^{(0)} = \sum_{j=1}^{k} m_{k,j}^{(0)}\mathrm{CV}_j$. This criterion corrects the bias of CV to $O(n^{-k-1})$, and its order is the same as in CEIC_k. The proof is omitted because it can be obtained by a slight modification of the proofs in Appendix A.2. However, we have already shown that the bias of CCV_k is $O(n^{-2k})$. The order of the bias of $\mathrm{CCV}_k^{(0)}$ is lower than that of CCV_k. In other words, we can regard CCV_k as the finite correction of $\mathrm{CCV}_k^{(0)}$ by using only the sample size n.

3. Numerical Studies

In this section, the performance of CCV_k is investigated by simulations. We examined the bias and the square root mean square error (RMSE) of the information criterion as well as the frequency of selecting the best model having the smallest R_γ. The bias and RMSE were evaluated with $N = 20,000$ repetitions.

The true model was generated by the following four distributions (κ_3 and κ_4 denote the skewness and kurtosis of the distribution, respectively):

Distribution 1: Standard normal distribution ($\kappa_3 = 0$ and $\kappa_4 = 0$).

Distribution 2: Laplace distribution with mean 0 and standard deviation 1 ($\kappa_3 = 0$ and $\kappa_4 = 3$).

Distribution 3: Uniform distribution on $(-1, 1)$ divided by the standard deviation $1/\sqrt{3}$ ($\kappa_3 = 0$ and $\kappa_4 = -1.2$).

Distribution 4: Skew-Laplace distribution with location parameter 0, dispersion parameter 1, and skew parameter 1, standardized by mean 3/4 and standard deviation $\sqrt{23}/4$ ($\kappa_3 \approx 1.06$ and $\kappa_4 \approx 3.26$).

The skew-Laplace distribution was proposed by Balakrishnan and Ambagaspitiya (1994). For the probability density function, see, e.g., Yanagihara and Yuan (2005).

First, in order to study the behavior of CCV_k for $k = 1, \ldots, 5$ as the estimator of R_γ, the biases and RMSE of CCV_k were examined. We simulated N data sets consisting of 20 ($= n$) observations from $\boldsymbol{y}$ with the dimension $p = 5$. Each element of $\boldsymbol{y}$ was generated independently from distributions 1 to 4. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the probability density function of the p-dimensional normal distribution as

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \left(\frac{1}{2\pi}\right)^{p/2} |\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\},$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathrm{vech}(\boldsymbol{\Sigma}))'$. To estimate the parameter $\boldsymbol{\theta}$, we used the following log-likelihood function:

$$\psi(\boldsymbol{y}|\boldsymbol{\theta}) = -\log f(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{1}{2}\left\{p\log 2\pi + \log|\boldsymbol{\Sigma}| + (\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}. \tag{3.1}$$

Then, the estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ become MLEs under the normal assumption. On the other hand, to measure the goodness of fit of the model, we prepared the following two discrepancies:

$$\begin{aligned}
&\text{Case of K-L} : \gamma(\boldsymbol{y}|\boldsymbol{\theta}) = \psi(\boldsymbol{y}|\boldsymbol{\theta}) \text{ in (3.1)},\\
&\text{Case of } L_2 \quad : \gamma(\boldsymbol{y}|\boldsymbol{\theta}) = -f(\boldsymbol{y}|\boldsymbol{\theta}) + \frac{1}{2}\int\{f(\boldsymbol{z}|\boldsymbol{\theta})\}^2 d\boldsymbol{z}\\
&\qquad\qquad\qquad = \left(\frac{1}{2\pi}\right)^{p/2}|\boldsymbol{\Sigma}|^{-1/2}\left[\frac{1}{2^{(p+2)/2}} - \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}\right].
\end{aligned}$$

Figure 1 shows the plots of relative biases ($= 100\times \mathrm{Bias}/|R_\gamma|$ (%)) and relative RMSEs ($= 100\times \mathrm{RMSE}/|R_\gamma|$ (%)) for the K-L and L_2 discrepancies, respectively. The plots of the relative biases are on the left-hand side of the figure, and the plots of the relative RMSEs are on the right-hand side of the figure. The bias approached 0 as k moved towards 5. In particular, the bias of CCV$_k$ was dramatically improved when k changed from 1 to 2. Moreover, in the case of K-L, the CCV$_k$ for $k = 2, \ldots, 5$ improved not only the bias of CV but also the RMSE. In the case of L_2, the RMSE of CCV$_k$ became slightly larger than that of CV, except for the case of a uniform distribution. In the case of K-L, the biases and RMSEs appeared to increase when the kurtosis (κ_4) increased. On the other hand, there was no apparent trend with respect to the kurtosis in the case of L_2. Comparing the plots for the Laplace and skew-Laplace distributions, the size of skewness (κ_3) had less of an effect on the bias and RMSE than the kurtosis. The investigation of several other models yielded similar results (not shown).

Insert Figure 1 around here

Next, a selection of the ridge parameter λ in the ridge regression model was investigated. We simulated N data sets consisting of 20 ($= n$) observations from the model

$$z = \boldsymbol{x}'\boldsymbol{\theta}_* + 5\varepsilon,$$

where $\boldsymbol{\theta}_* = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and ε is generated from the same four distributions as in Example 1. The covariate variable $\boldsymbol{x} = (x_1, \ldots, x_q)'$ is independent of ε, and each element of $\boldsymbol{x}$ is normal with mean 0 and standard deviation 3. The correlation between x_a and x_b was $(0.5)^{|a-b|}$. The setting of $\boldsymbol{\theta}_*$ and the correlation of $\boldsymbol{x}$ were the same as in Tibshirani (1996). Let $\boldsymbol{y} = (z, \boldsymbol{x}')'$. Our result can easily be applied to the ridge regression model, i.e., the following two discrepancies can be used:

$$\psi(\boldsymbol{y}|\boldsymbol{\theta}) = (z - \boldsymbol{x}'\boldsymbol{\theta})^2 + \lambda\|\boldsymbol{\theta}\|^2, \quad \gamma(\boldsymbol{y}|\boldsymbol{\theta}) = (z - \boldsymbol{x}'\boldsymbol{\theta})^2. \tag{3.2}$$

Then, $\hat{\boldsymbol{\theta}}$ becomes the penalized least square estimator, i.e., $\hat{\boldsymbol{\theta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_q)^{-1}\boldsymbol{X}'\boldsymbol{z}$, where $\boldsymbol{z} = (z_1, \ldots, z_n)'$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$. In order to compare the proposed criterion with other criteria, we prepared the following two criteria: the GCV proposed by Craven and Wahba (1979), which is commonly used to select λ in ridge regression, and the GIC, which is obtained by applying equation (A.7) to equation (3.2). Although the EIC is also a well known criterion, we did not use the EIC herein because several calculations are required in order to obtain the EIC.

Let $\hat{\sigma}^2 = (\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\theta}})'(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\theta}})/n$ and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_q)^{-1}\boldsymbol{X}'$. Then, the GCV and GIC are given by

$$\mathrm{GCV} = \frac{\hat{\sigma}^2}{1 - \mathrm{tr}(\boldsymbol{H})/n}, \quad \mathrm{GIC} = \hat{\sigma}^2 + \frac{2}{n}\left\{\mathrm{tr}(\boldsymbol{D}\boldsymbol{H}) - \lambda\hat{\boldsymbol{e}}'\boldsymbol{H}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{\theta}}\right\},$$

where $\boldsymbol{D} = \mathrm{diag}(\hat{e}_1, \ldots, \hat{e}_n)$, $\hat{\boldsymbol{e}} = (\hat{e}_1, \ldots, \hat{e}_n)'$, and $\hat{e}_i = z_i - \boldsymbol{x}_i'\hat{\boldsymbol{\theta}}$ $(i = 1, \ldots, n)$.

The averages of GCV, GIC, CV, and CCV$_2$ with $\lambda_j = 3(j-1)/99$ $(j = 1, \ldots, 100)$ are shown in Figure 2. The biases of GCV and GIC were very larger than the bias of CV. However, there was a clear bias in CV. The CCV$_2$ reduced this bias to nearly 0. The RMSEs of GCV, GIC, CV, and CCV$_2$ are shown in Figure 3. The CCV$_2$ corrected not only the bias of CV but also the RMSE of CV. The frequencies of λ selected by the information criteria are shown in Figure 4. The best ridge parameter λ_0 minimizing R_γ was 0.78 for all distributions, because the risk function depended only on the first- and second-order moments of error distributions (and they were the same for all of the error distributions). The modes of frequencies of λ selected by GCV and GIC were not near λ_0, but the modes of λ selected by CV and CCV$_2$ were near λ_0. Table 1 shows the means of the selected λ. The mean of λ selected by CCV$_2$ was closer to λ_0 than that selected by CV. The CV tends to choose a more inflexible model as the best model. In the ridge regression model, an inflexible model denotes the model having a large ridge parameter. The CCV$_2$ also improved the disadvantage of the CV.

Insert Figures 2, 3, 4 and Table 1 around here

4. Conclusion and Discussion

In the present paper, we proposed the kth bias-corrected CV criterion (CCV_k), which is defined by the linear combination from CV_1 to CV_k. The CCV_k reduces the bias of the CV criterion to $O(n^{-2k})$ without estimating any higher-order cumulants, without obtaining any explicit form of the asymptotic expansion for the bias of CV criterion, and without calculating any partial derivatives of $\psi(\boldsymbol{y}|\boldsymbol{\theta})$ or $\gamma(\boldsymbol{y}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The Monte Carlo results presented in the previous section verified the advantage of CCV_k, and in particular that of CCV_2. In many cases, the CCV_k improves not only the bias of CV but also the RMSE and the frequency, where the model with the smallest R_γ is selected as the best model. Several second-order bias-corrected information criteria have been proposed by, e.g., Hurvich, Simonoff and Tsai (1998), Hurvich and Tsai (1998), Simonoff and Tsai (1999), Naik and Tsai (2001), Yanagihara, Sekiguchi and Fujikoshi (2003) and Chiou and Tsai (2006). However, these criteria were obtained under specified models and distributions. On the other hand, Yanagihara, Tonda and Matsumoto (2006) and Yanagihara *et al.* (2008) proposed second-order bias-corrected CV criteria without specifying models and distributions. However, their results can be applied only when $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = \gamma(\boldsymbol{y}|\boldsymbol{\theta})$. The CCV_k criterion presented herein is obtained under a more general assumption, and can be applied broadly.

Using the bootstrap iteration, we can improve the order of the bias of CV up to a higher order. However, the order of the bias of CEIC_k is lower than that of CCV_k, and the computational task of CEIC_k is larger than that of CCV_k in most cases. We can reduce the computational task of CV_k more effectively by using a Monte Carlo CV_k (see e.g., Picard & Cook, 1984; Shao, 1993). However, we should note that the MSE of a bias-corrected CV criterion based on Monte Carlo CV_k might become larger than that of CCV_k.

Furthermore, a serious issue remains regarding the selection of k. A theoretical solution to this issue will be very difficult to obtain. From the simulation studies and the viewpoint of the computational task, we recommend the use of $k = 2$.

Appendix

A.1. Asymptotic Expansions of $E_\varphi[\mathrm{CV}]$

First, we define the following vectors and matrices, which are based on the partial derivatives up to the third order:

$$\boldsymbol{g}_\psi(\boldsymbol{y}|\boldsymbol{\vartheta}) = \left.\frac{\partial}{\partial \boldsymbol{\theta}}\psi(\boldsymbol{y}|\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}}, \qquad \boldsymbol{g}_\gamma(\boldsymbol{y}|\boldsymbol{\vartheta}) = \left.\frac{\partial}{\partial \boldsymbol{\theta}}\gamma(\boldsymbol{y}|\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}},$$
$$\boldsymbol{H}_\psi(\boldsymbol{y}|\boldsymbol{\vartheta}) = \left.\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\psi(\boldsymbol{y}|\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}}, \qquad \boldsymbol{H}_\gamma(\boldsymbol{y}|\boldsymbol{\vartheta}) = \left.\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\gamma(\boldsymbol{y}|\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}},$$
$$\boldsymbol{L}_\psi(\boldsymbol{y}|\boldsymbol{\vartheta}) = \left.\left(\frac{\partial}{\partial \boldsymbol{\theta}'} \otimes \frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)\psi(\boldsymbol{y}|\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\vartheta}},$$

and

$$\boldsymbol{r}_\psi(\boldsymbol{\theta}) = E_\varphi[\boldsymbol{g}_\psi(\boldsymbol{y}|\boldsymbol{\theta})], \quad \boldsymbol{I}_\psi(\boldsymbol{\theta}) = E_\varphi[\boldsymbol{g}_\psi(\boldsymbol{y}|\boldsymbol{\theta})\boldsymbol{g}_\psi(\boldsymbol{y}|\boldsymbol{\theta})'], \quad \boldsymbol{J}_\psi(\boldsymbol{\theta}) = E_\varphi[\boldsymbol{H}_\psi(\boldsymbol{y}|\boldsymbol{\theta})],$$
$$\boldsymbol{J}_\gamma(\boldsymbol{\theta}) = E_\varphi[\boldsymbol{H}_\gamma(\boldsymbol{y}|\boldsymbol{\theta})], \quad \boldsymbol{K}_\psi(\boldsymbol{\theta}) = E_\varphi[\boldsymbol{L}_\psi(\boldsymbol{y}|\boldsymbol{\theta})].$$

Assume that $\boldsymbol{\theta}_0$ is a $q \times 1$ vector such that $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $n \to \infty$. Under the proper regularity conditions, as specified in White (1982), $\boldsymbol{\theta}_0$ satisfies $\boldsymbol{r}_\psi(\boldsymbol{\theta}_0) = \boldsymbol{0}_q$, where $\boldsymbol{0}_q$ is a vector of q zeros.

Let

$$\hat{\boldsymbol{J}}_\psi(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{H}_\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}), \qquad \hat{\boldsymbol{K}}_\psi(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{L}_\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}). \tag{A.1}$$

From Yanagihara *et al.* (2008), we have the stochastic expansion of $\hat{\boldsymbol{\theta}}_{[-i]}$ as

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

where

$$\boldsymbol{z}_{1,i} = \hat{\boldsymbol{J}}_\psi(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}_\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}), \quad \boldsymbol{z}_{2,i} = \hat{\boldsymbol{J}}_\psi(\hat{\boldsymbol{\theta}})^{-1}\left\{\boldsymbol{H}_\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\boldsymbol{z}_{1,i} - \frac{1}{2}\hat{\boldsymbol{K}}_\psi(\hat{\boldsymbol{\theta}})\mathrm{vec}(\boldsymbol{z}_{1,i}\boldsymbol{z}_{1,i}')\right\}. \tag{A.3}$$

Since $\boldsymbol{y}_i$ and $\hat{\boldsymbol{\theta}}_{[-i]}$ are mutually independent, the equation $E_\varphi[\gamma(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}_{[-i]})] = E_\varphi[\rho(\hat{\boldsymbol{\theta}}_{[-i]})]$ holds. Note that $E_\varphi[\mathrm{CV}] = n^{-1}\sum_{i=1}^n E_\varphi[\rho(\hat{\boldsymbol{\theta}}_{[-i]})]$. Therefore, applying the Taylor expansion of $E_\varphi[\rho(\hat{\boldsymbol{\theta}}_{[-i]})]$ at $\hat{\boldsymbol{\theta}}$, we obtain

$$E_\varphi[\mathrm{CV}] = R_\gamma + \frac{1}{n}R_1 + \frac{1}{n^2}R_2 + O(n^{-3}), \tag{A.4}$$

where

$$R_1 = \frac{1}{n}\sum_{i=1}^n E_\varphi\left[\boldsymbol{r}_\gamma'(\hat{\boldsymbol{\theta}})\boldsymbol{z}_{1,i}\right], \quad R_2 = \frac{1}{n}\sum_{i=1}^n E_\varphi\left[\boldsymbol{r}_\gamma(\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{2,i} + \frac{1}{2}\boldsymbol{z}_{1,i}'\boldsymbol{J}_\gamma(\hat{\boldsymbol{\theta}})\boldsymbol{z}_{1,i}\right].$$

Note that since $\hat{\boldsymbol{\theta}}$ is the minimum of $\sum_{i=1}^{n} \psi(\boldsymbol{y}_i|\boldsymbol{\theta})$, the equation $\sum_{i=1}^{n} \boldsymbol{g}_\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}}) = \boldsymbol{0}_q$ holds. Thus, we have $\sum_{i=1}^{n} \boldsymbol{z}_{1,i} = \boldsymbol{0}_q$. Consequently, R_1 becomes 0. By using relations $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$, $\hat{\boldsymbol{J}}_\psi(\hat{\boldsymbol{\theta}}) \xrightarrow{a.s.} \boldsymbol{J}_\psi(\boldsymbol{\theta}_0)$, and $\hat{\boldsymbol{K}}_\psi(\hat{\boldsymbol{\theta}}) \xrightarrow{a.s.} \boldsymbol{K}_\psi(\boldsymbol{\theta}_0)$, R_2 is expanded as

$$R_2 = \eta_1 - \frac{1}{2}\eta_2 + \frac{1}{2}\eta_3 + O(n^{-1}),$$

where

$$\begin{aligned}
\eta_1 &= \boldsymbol{r}_\gamma(\boldsymbol{\theta}_0)'\boldsymbol{J}_\psi(\boldsymbol{\theta}_0)^{-1}E_\varphi[\boldsymbol{H}_\psi(\boldsymbol{y}|\boldsymbol{\theta}_0)\boldsymbol{J}_\psi(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}_\psi(\boldsymbol{y}|\boldsymbol{\theta}_0)],\\
\eta_2 &= \boldsymbol{r}_\gamma(\boldsymbol{\theta}_0)'\boldsymbol{J}_\psi(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}_\psi(\boldsymbol{\theta}_0)\{\boldsymbol{J}_\psi(\boldsymbol{\theta}_0)^{-1} \otimes \boldsymbol{J}_\psi(\boldsymbol{\theta}_0)^{-1}\}\mathrm{vec}(\boldsymbol{I}_\psi(\boldsymbol{\theta}_0)),\\
\eta_3 &= \mathrm{tr}\left\{\boldsymbol{I}_\psi(\boldsymbol{\theta}_0)\boldsymbol{J}_\psi(\boldsymbol{\theta}_0)^{-1}\boldsymbol{J}_\gamma(\boldsymbol{\theta}_0)\boldsymbol{J}_\psi(\boldsymbol{\theta}_0)^{-1}\right\}.
\end{aligned} \tag{A.5}$$

Substituting above the expansion of R_2 and $R_1 = 0$ into (A.4) yields the following theorem.

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

$$R_\gamma - E_\varphi[\mathrm{CV}] = -\frac{1}{2n^2}(2\eta_1 - \eta_2 + \eta_3) + O(n^{-3}), \tag{A.6}$$

where η_1, η_2, and η_3 are given by (A.5).

From Theorem A.1, it is easy to see that β_2 in (2.5) is $-\eta_2 + \eta_2/2 - \eta_3/2$.

By using Theorem A.1, we can easily obtain an expansion of the bias of GIC. Under our setting of the model selection, GIC is written as

$$\mathrm{GIC} = \Gamma(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + \frac{1}{n}\mathrm{tr}\left\{\hat{\boldsymbol{I}}_{\psi\gamma}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}_\psi(\hat{\boldsymbol{\theta}})^{-1}\right\}, \tag{A.7}$$

where $\Gamma(\cdot|\boldsymbol{Y})$ and $\hat{\boldsymbol{J}}_\psi(\hat{\boldsymbol{\theta}})$ are given by (2.15) and (A.1), respectively, and $\hat{\boldsymbol{I}}_{\psi\gamma}(\hat{\boldsymbol{\theta}})$ is given by

$$\hat{\boldsymbol{I}}_{\psi\gamma}(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_\psi(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\boldsymbol{g}_\gamma(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'.$$

By applying the Taylor expansion of CV at $\hat{\boldsymbol{\theta}}$, we have

$$\mathrm{CV} = \Gamma(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}) + \frac{1}{n}G_1 + \frac{1}{n^2}\left(G_2 + \frac{1}{2}G_3\right) + O_p(n^{-3}),$$

where G_1, G_2, and G_3 are given by

$$G_1 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_\gamma(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{1,i},\ G_2 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_\gamma(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{2,i},\ G_3 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_{1,i}'\boldsymbol{H}_\gamma(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\boldsymbol{z}_{1,i}. \tag{A.8}$$

Note that $G_1 = \mathrm{tr}\{\hat{\boldsymbol{I}}_{\psi\gamma}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}})^{-1}\}$. Thus,

$$E_{\varphi}[\mathrm{GIC}] = E_{\varphi}[\mathrm{CV}] - \frac{1}{n^2}\left(E_{\varphi}[G_2] + \frac{1}{2}E_{\varphi}[G_3]\right) + O(n^{-3}).$$

By using the relations $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$, $\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}}) \xrightarrow{a.s.} \boldsymbol{J}_{\psi}(\boldsymbol{\theta}_0)$, and $\hat{\boldsymbol{K}}_{\psi}(\hat{\boldsymbol{\theta}}) \xrightarrow{a.s.} \boldsymbol{K}_{\psi}(\boldsymbol{\theta}_0)$, we obtain

$$\begin{aligned}
E_{\varphi}[G_2] &= \frac{1}{n}\sum_{i=1}^{n}\Big\{E_{\varphi}\left[\boldsymbol{g}_{\gamma}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\psi}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}_{\psi}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\right] \\
&\quad -\frac{1}{2}\,E_{\varphi}\left[\boldsymbol{g}_{\gamma}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{K}}_{\psi}(\hat{\boldsymbol{\theta}})\left\{\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}})^{-1}\otimes\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}})^{-1}\right\}\mathrm{vec}(\boldsymbol{g}_{\psi}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\boldsymbol{g}_{\psi}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})')\right]\Big\} \\
&= \eta_4 - \frac{1}{2}\eta_5 + O(n^{-1}), \\
E_{\varphi}[G_3] &= \frac{1}{n}\sum_{i=1}^{n}E_{\varphi}\left[\boldsymbol{g}_{\psi}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{H}_{\gamma}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}_{\psi}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{g}_{\psi}(\boldsymbol{y}_i|\hat{\boldsymbol{\theta}})\right] \\
&= \eta_6 + O(n^{-1}),
\end{aligned}$$

where

$$\begin{aligned}
\eta_4 &= E_{\varphi}\left[\boldsymbol{g}_{\gamma}(\boldsymbol{y}|\boldsymbol{\theta}_0)'\boldsymbol{J}_{\psi}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}_{\psi}(\boldsymbol{y}|\boldsymbol{\theta}_0)\boldsymbol{J}_{\psi}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}_{\psi}(\boldsymbol{y}|\boldsymbol{\theta}_0)\right], \\
\eta_5 &= E_{\varphi}\left[\boldsymbol{g}_{\gamma}(y|\boldsymbol{\theta}_0)'\boldsymbol{J}_{\psi}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{K}_{\psi}(\boldsymbol{\theta}_0)\left\{\boldsymbol{J}_{\psi}(\boldsymbol{\theta}_0)^{-1}\otimes\boldsymbol{J}_{\psi}(\boldsymbol{\theta}_0)^{-1}\right\}\mathrm{vec}(\boldsymbol{g}_{\psi}(y|\boldsymbol{\theta}_0)\boldsymbol{g}_{\psi}(y|\boldsymbol{\theta}_0)')\right], \quad \text{(A.9)} \\
\eta_6 &= E_{\varphi}\left[\boldsymbol{g}_{\psi}(\boldsymbol{y}|\boldsymbol{\theta}_0)'\boldsymbol{J}_{\psi}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{H}_{\gamma}(\boldsymbol{y}|\boldsymbol{\theta}_0)\boldsymbol{J}_{\psi}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{g}_{\psi}(\boldsymbol{y}|\boldsymbol{\theta}_0)\right].
\end{aligned}$$

Thus,

$$E_{\varphi}[\mathrm{GIC}] = E_{\varphi}[\mathrm{CV}] - \frac{1}{2n^2}(2\eta_4 - \eta_5 + \eta_6) + O(n^{-3}). \tag{A.10}$$

Substituting (A.6) into (A.10) yields the following theorem.

Theorem A.2. *Under proper regularity conditions, the bias of GIC is expanded as*

$$R_{\gamma} - E_{\varphi}[\mathrm{GIC}] = -\frac{1}{2n^2}\left\{2(\eta_1 - \eta_4) - (\eta_2 - \eta_5) + (\eta_3 - \eta_6)\right\} + O(n^{-3}), \tag{A.11}$$

where η_1, η_2, and η_3 are given by (A.5), *and η_4, η_5, and η_6 are given by* (A.9).

Konishi and Kitagawa (2003) obtained the asymptotic expansion of the bias of GIC up to the same order of as that of our result in Theorem A.2. However, our result is simpler than that of Konishi and Kitagawa, although they consider a more general situation. Theorems A.1 and A.2 in Appendix A.2 show that the n^{-2} term in the bias of GIC contains more higher-order moments than that of the CV criterion. On the other hand, if the equation $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = \gamma(\boldsymbol{y}|\boldsymbol{\theta})$ is satisfied, the GIC is equivalent to Takeuchi's information criterion (TIC; Takeuchi, 1976). Then, $\boldsymbol{r}_\psi(\hat{\boldsymbol{\theta}})$ becomes $\mathbf{0}_q$. This yields the relation $\eta_1 = \eta_2 = 0$. Moreover, when the equation $\psi(\boldsymbol{y}|\boldsymbol{\theta}) = \gamma(\boldsymbol{y}|\boldsymbol{\theta})$ is satisfied, η_4 is equal to η_6 because $\boldsymbol{g}_\psi(\boldsymbol{y}|\boldsymbol{\theta}_0) = \boldsymbol{g}_\gamma(\boldsymbol{y}|\boldsymbol{\theta}_0)$ and $\boldsymbol{H}_\psi(\boldsymbol{y}|\boldsymbol{\theta}_0) = \boldsymbol{H}_\gamma(\boldsymbol{y}|\boldsymbol{\theta}_0)$ hold. Consequently, the expansions in Theorems A.1 and A.2 when $\psi(\boldsymbol{y}|\boldsymbol{\theta})$ is equal to $\gamma(\boldsymbol{y}|\boldsymbol{\theta})$ coincide with the results in Yanagihara *et al.* (2008).

A.2. Proof of Theorem 2.1

In this sub-section, we present the proof of Theorem 2.1. First, we present propositions and lemmas to be used in the proof of Theorem 2.1.

PROPOSITION A.1. *Let* $b_{j,l}^{(1)} = a_j^{l-2}$ $(l \geq 2)$ *and*

$$b_{k,j}^{(l)} = \frac{b_{k-1,j}^{(l)} - b_{k-1,j-1}^{(l)}}{a_j - a_{j-k+1}}, \qquad (k \geq 2,\ l \geq k+1), \tag{A.12}$$

where a_j *is given by* (2.8). *It then holds that*

$$b_{k,j}^{(l)} = \sum_{\alpha_{j-k+1}+\cdots+\alpha_j = l-k-1} a_{j-k+1}^{\alpha_{j-k+1}} \cdots a_j^{\alpha_j}, \tag{A.13}$$

where α_* *is a nonnegative integer. In particular,* $b_{k,j,k}^{(k+1)} = 1$.

PROOF. The case in which $k = 1$ on equation (A.13) holds because

$$b_{1,j}^{(l)} = \sum_{\alpha_j = l-2} a_j^{\alpha_j} = a_j^{l-2}.$$

The case in which $k \geq 2$ can be proved by mathematical induction. Assume that equation (A.13) holds until k, and then consider the case in which $k+1$. We can see that

$$\begin{aligned}
b_{k+1,j}^{(l)} &= \frac{b_{k,j}^{(l)} - b_{k,j-1}^{(l)}}{a_j - a_{j-k}} \\
&= \frac{1}{a_j - a_{j-k}} \left(\sum_{\alpha_{j-k+1}+\cdots+\alpha_j = l-k-1} a_{j-k+1}^{\alpha_{j-k+1}} \cdots a_j^{\alpha_j} - \sum_{\alpha_{j-k}+\cdots+\alpha_{j-1} = l-k-1} a_{j-k}^{\alpha_{j-k}} \cdots a_{j-1}^{\alpha_{j-1}} \right) \\
&= \frac{1}{a_j - a_{j-k}} \sum_{u=0}^{l-k-1} \sum_{\alpha_{j-k+1}+\cdots+\alpha_{j-1} = u} a_{j-k+1}^{\alpha_{j-k+1}} \cdots a_{j-1}^{\alpha_{j-1}} \left(a_j^{l-k-1-u} - a_{j-k}^{l-k-1-u} \right) \\
&= \sum_{u=0}^{l-k-1} \sum_{\alpha_{j-k+1}+\cdots+\alpha_{j-1} = u} a_{j-k+1}^{\alpha_{j-k+1}} \cdots a_{j-1}^{\alpha_{j-1}} \frac{a_j^{l-k-1-u} - a_{j-k}^{l-k-1-u}}{a_j - a_{j-k}}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{u=0}^{l-k-2} \sum_{\alpha_{j-k+1}+\cdots+\alpha_{j-1}=u} a_{j-k+1}^{\alpha_{j-k+1}} \cdots a_{j-1}^{\alpha_{j-1}} \sum_{\alpha_j+\alpha_{j-k}=l-k-2-u} a_j^{\alpha_j} a_{j-k+1}^{\alpha_{j-k}} \\
&= \sum_{\alpha_{j-k}+\cdots+\alpha_j=l-k-2} a_{j-k}^{\alpha_{j-k}} \cdots a_j^{\alpha_j}.
\end{aligned}
$$

The proof of equation (A.13) is complete. Furthermore, it follows from (A.13) that

$$
b_{k,j}^{(k+1)} = \sum_{\alpha_{j-k+1}+\cdots+\alpha_j=0} a_{j-k+1}^{\alpha_{j-k+1}} \cdots a_j^{\alpha_j} = 1. \qquad \square
$$

Lemma A.1. *It holds that*

$$
E_\varphi\left[\Delta_{k,j}\right] = \sum_{l=k+1}^{L} \frac{\beta_l}{n^l} b_{k,j}^{(l)} + O(n^{-L-1}), \tag{A.14}
$$

where $\Delta_{k,j}$ is given by (2.9) *and* (2.11), *and $b_{k,j}^{(l)}$ is defined in Proposition A.1.*

Proof. The case in which $k = 1$ follows from equation (2.7). The case in which $k \geq 2$ can be proved by mathematical induction. Assume that equation (A.14) holds until k, and then consider the case in which $k+1$. We can see that

$$
\begin{aligned}
E_\varphi\left[\Delta_{k+1,j}\right] &= \frac{1}{a_j - a_{j-k}} E_\varphi\left[\Delta_{j,k} - \Delta_{k,j-1}\right] \\
&= \sum_{l=k+1}^{L} \frac{\beta_l}{n^l} \left(\frac{b_{k,j}^{(l)} - b_{k,j-1}^{(l)}}{a_j - a_{j-k}} \right) + O(n^{-L-1}) \\
&= \sum_{l=k+2}^{L} \frac{\beta_l}{n^l} b_{k+1,j}^{(l)} + O(n^{-L-1}).
\end{aligned}
$$

The case in which $l = k+1$ vanishes because $b_{k,j}^{(l)} - b_{k,j-1}^{(l)} = 0$ from $b_{k,j'}^{(k+1)} = 1$ for any j' by Proposition A.1. The proof is complete. $\square$

Proposition A.2. *Let*

$$
Q_k^{(l)} = \sum_{j=1}^{k} c_{j-1} b_{j,j}^{(l)}. \tag{A.15}
$$

It then holds that

$$
Q_k^{(l)} = 1 - c_k \sum_{u=0}^{l-k-2} \sum_{\alpha_1+\cdots+\alpha_k=u} a_1^{\alpha_1} \cdots a_k^{\alpha_k}, \qquad (l \geq k+1), \tag{A.16}
$$

where α_* is a nonnegative integer. In particular, $Q_{l-1}^{(l)} = \sum_{j=1}^{l-1} c_{j-1} b_{j,j}^{(l)} = 1$. Furthermore, $1 - Q_k^{(l)} = O(n^{-k})$.

Proof. The case in which $k = 1$ holds because expressions (A.15) and (A.16) are given by

$$\begin{aligned}
Q_1^{(l)} &= c_0 b_{1,1}^{(l)} = \sum_{\alpha_1 = l-2} a_1^{\alpha_1} = a_1^{l-2} \\
Q_1^{(l)} &= 1 - c_1 \sum_{u=0}^{l-3} \sum_{\alpha_1 = u} a_1^{\alpha_1} = 1 - (1 - a_1) \sum_{u=0}^{l-3} a_1^u \\
&= \ 1 - (1 - a_1) \frac{1 - a_1^{l-2}}{1 - a_1} = a_1^{l-2}.
\end{aligned}$$

The case in which $k \geq 2$ can be proved by mathematical induction. Assume that equation (A.14) holds until k, and then consider the case in which $k + 1$. We can see that

$$\begin{aligned}
\frac{Q_{k+1}^{(l)} - 1}{c_k} &= \frac{Q_k^{(l)} - 1 + c_k b_{k+1,k+1}^{(l)}}{c_k} = \frac{Q_k^{(l)} - 1}{c_k} + b_{k+1,k+1}^{(l)} \\
&= - \sum_{u=0}^{l-k-2} \sum_{\alpha_1 + \cdots + \alpha_k = u} a_1^{\alpha_1} \cdots a_k^{\alpha_k} + \sum_{\alpha_1 + \cdots + \alpha_{k+1} = l-k-2} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha_{k+1}}
\end{aligned}$$

and then

$$\begin{aligned}
&\frac{Q_{k+1}^{(l)} - 1}{c_k} - \frac{c_{k+1}}{c_k} \sum_{u=0}^{l-k-3} \sum_{\alpha_1 + \cdots + \alpha_{k+1} = u} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha_k + 1} \\
&= - \sum_{u=0}^{l-k-2} \sum_{\alpha_1 + \cdots + \alpha_k = u} a_1^{\alpha_1} \cdots a_k^{\alpha_k} + \sum_{\alpha_1 + \cdots + \alpha_{k+1} = l-k-2} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha_{k+1}} \\
&\quad + (1 - a_{k+1}) \sum_{u=0}^{l-k-3} \sum_{\alpha_1 + \cdots + \alpha_{k+1} = u} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha_{k+1}} \\
&= - \sum_{u=0}^{l-k-2} \sum_{\alpha_1 + \cdots + \alpha_k = u} a_1^{\alpha_1} \cdots a_k^{\alpha_k} + \sum_{u=0}^{l-k-2} \sum_{\alpha_1 + \cdots + \alpha_{k+1} = u} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha_{k+1}} \\
&\quad - a_{k+1} \sum_{u=0}^{l-k-3} \sum_{\alpha_1 + \cdots + \alpha_{k+1} = u} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha_k + 1} \\
&= - \sum_{u=0}^{l-k-2} \sum_{\alpha_1 + \cdots + \alpha_k = u} a_1^{\alpha_1} \cdots a_k^{\alpha_k} + \sum_{u=0}^{l-k-2} \sum_{\alpha_1 + \cdots + \alpha_k = u} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha_{k+1}} \\
&\quad + \sum_{u'=0}^{l-k-3} \sum_{\alpha_1 + \cdots + \alpha'_{k+1} = u'} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha'_{k+1} + 1} - \sum_{u=0}^{l-k-3} \sum_{\alpha_1 + \cdots + \alpha_{k+1} = u} a_1^{\alpha_1} \cdots a_{k+1}^{\alpha_{k+1} + 1} \\
&= 0,
\end{aligned}$$

where on the third equality, the second term is divided into two cases, where $\alpha_{k+1} = 0$ and $\alpha_{k+1} \geq 1$, and then we use the transformation $\alpha_{k+1} - 1 = \alpha'_{k+1}$ and $u' = u - 1$. It is clear that $Q_{l-1}^{(l)} = 1$ by equation (A.16) and that $1 - Q_k^{(l)} = O(n^{-k})$ because $c_k = O(n^{-k})$. The proof is complete. □

Proposition A.3. *Let*

$$A_k^{(l)} = 1 - \sum_{j=1}^{\min(l-1,k)} c_{j-1} b_{j,j}^{(l)}.$$

It then holds that

$$A_k^{(l)} = \begin{cases} 0 & (\textit{if } l \leq k+1) \\ O(n^{-k}) & (\textit{if } l \geq k+2) \end{cases}$$

Proof. First consider the case in which $l \leq k + 1$. It follows from Proposition A.2 that

$$A_k^{(l)} = 1 - \sum_{j=1}^{l-1} c_{j-1} b_{j,j}^{(l)} = 1 - Q_{l-1}^{(l)} = 1 - 1 = 0.$$

Next consider the case in which $l \geq k + 2$. It follows from Proposition A.2 that

$$A_k^{(l)} = 1 - \sum_{j=1}^{k} c_{j-1} b_{j,j}^{(l)} = 1 - D_k^{(l)} = O(n^{-k}). \qquad \square$$

By using Lemma A.1 and Proposition A.2, we give the proof of Theorem 2.1 as follows:

Proof of Theorem 2.1. It follows from Lemma A.1 and Proposition A.2 that

$$\begin{aligned} E_{\varphi}[\mathrm{CCV}_k] &= E_{\varphi}[\mathrm{CV}] - \sum_{j=1}^{k-1} c_{j-1} E_{\varphi}[\Delta_{j,j}] \\ &= R_{\gamma} + \sum_{l=2}^{L} \frac{\beta_l}{n^l} - \sum_{j=1}^{k-1} c_{j-1} \sum_{l=j+1}^{L} \frac{\beta_l}{n^l} b_{j,j}^{(l)} + O(n^{-L-1}) \\ &= R_{\gamma} + \sum_{l=2}^{L} \frac{\beta_l}{n^l} - \sum_{l=2}^{L} \sum_{j=1}^{\min(l-1,k-1)} \frac{\beta_l}{n^l} c_{j-1} b_{j,j}^{(l)} + O(n^{-L-1}) \\ &= R_{\gamma} + \sum_{l=2}^{L} \frac{\beta_l}{n^l} A_{k-1}^{(l)} + O(n^{-L-1}) \\ &= R_{\gamma} + \sum_{l=k+1}^{L} \frac{\beta_l}{n^l} A_{k-1}^{(l)} + O(n^{-L-1}) \\ &= R_{\gamma} + O(n^{-2k}). \qquad \square \end{aligned}$$

A.3. Proof of Theorem 2.2

In this sub-section, we present the proof of Theorem 2.2. First, we present propositions and lemmas that will be used in the proof of Theorem 2.2.

Lemma A.2. *It holds that*

$$\Delta_{k,j} = \sum_{l=0}^{k} \xi_{k,j}^{(l)} \mathrm{CV}_{j+1-l}, \tag{A.17}$$

where

$$\begin{aligned}
\xi_{k,j}^{(l)} &= \frac{(-1)^l n^{k-1}}{(k-1)! \prod_{\alpha=1}^{k} a_{j-\alpha+1}} \left(\frac{{}_{k-1}C_{l-1}}{a_{j-l+1}^k} + \frac{{}_{k-1}C_l}{a_{j-l}^k} \right), \ (l = 1, \ldots, k-1), \\
\xi_{k,j}^{(0)} &= \frac{n^{k-1}}{(k-1)! a_j^k \prod_{\alpha=1}^{k} a_{j-\alpha+1}}, \\
\xi_{k,j}^{(k)} &= \frac{(-1)^k n^{k-1}}{(k-1)! a_{j-k+1}^k \prod_{\alpha=1}^{k} a_{j-\alpha+1}},
\end{aligned}$$

and a_j is given by (2.8).

Proof. The case in which $k = 1$ in equation (A.17) holds because

$$\Delta_{1,j} = \sum_{l=0}^{1} \xi_{1,j}^{(l)} \mathrm{CV}_{j+1-l} = \frac{1}{a_j^2} \mathrm{CV}_{j+1} - \frac{1}{a_j^2} \mathrm{CV}_j,$$

which is the same as (2.9). The case in which $k \geq 2$ can be proven by mathematical induction. Assume that equation (A.17) holds until k, and then consider the case in which $k+1$. It follows from equation (2.11) that

$$\begin{aligned}
\Delta_{k+1,j} &= \frac{1}{a_j - a_{j-k}} \left(\Delta_{k,j} - \Delta_{k,j-1} \right) \\
&= \frac{n}{k a_j a_{j-k}} \left(\Delta_{k,j} - \Delta_{k,j-1} \right) \\
&= \frac{n}{k a_j a_{j-k}} \left\{ \sum_{l=0}^{k} \xi_{k,j}^{(l)} \mathrm{CV}_{j+1-l} - \sum_{l=0}^{k} \xi_{k,j-1}^{(l)} \mathrm{CV}_{j-l} \right\} \\
&= \frac{n}{k a_j a_{j-k}} \left\{ \mathrm{CV}_{j+1} \xi_{k,j}^{(0)} - \mathrm{CV}_{j-k} \xi_{k,j-1}^{(k)} + \sum_{l=0}^{k} \mathrm{CV}_{j+1-l} \left(\xi_{k,j}^{(l)} - \xi_{k,j-1}^{(l-1)} \right) \right\}.
\end{aligned}$$

The coefficients of CV_{j+1} and CV_{j-k} can be easily expressed as follows:

$$\frac{n}{k a_j a_{j-k}} \xi_{k,j}^{(0)} = \frac{n}{k a_j a_{j-k}} \frac{n^{k-1}}{(k-1)! a_j^k \prod_{\alpha=1}^{k} a_{j-\alpha+1}}$$

$$
\begin{aligned}
&= \frac{n^k}{k! a_j^{k+1} \prod_{\alpha=1}^{k+1} a_{j-\alpha+1}} \\
&= \xi^{(0)}_{k+1,j}, \\
\frac{n}{k a_j a_{j-k}} \xi^{(k)}_{k,j-1} &= -\frac{n}{k a_j a_{j-k}} \frac{(-1)^k n^{k-1}}{(k-1)! a^k_{k-k+1} \prod_{\alpha=1}^{k} a_{j-\alpha+1}} \\
&= \frac{(-1)^{k+1} n^k}{k! a^{k+1}_{j-k+1} \prod_{\alpha=1}^{k+1} a_{j-\alpha+1}} \\
&= \xi^{(k+1)}_{k+1,j}.
\end{aligned}
$$

The coefficient of CV_{j+1-l} can be rewritten as follows:

$$
\begin{aligned}
&\frac{n}{k a_j a_{j-k}} \left\{ \xi^{(l)}_{k,j} - \xi^{(l-1)}_{k,j-1} \right\} \\
&= \frac{n}{k a_j a_{j-k}} \left\{ \frac{(-1)^l n^{k-1}}{(k-1)! \prod_{\alpha=1}^{k} a_{j-\alpha+1}} \left(\frac{{}_{k-1}C_{l-1}}{a^k_{j-l+1}} + \frac{{}_{k-1}C_{l}}{a^k_{j-l}} \right) \right. \\
&\qquad \left. - \frac{(-1)^{l-1} n^{k-1}}{(k-1)! \prod_{\alpha=1}^{k} a_{j-\alpha}} \left(\frac{{}_{k-1}C_{l-2}}{a^k_{j-l+1}} + \frac{{}_{k-1}C_{l-1}}{a^k_{j-l}} \right) \right\} \\
&= \frac{(-1)^l n^k}{k! \prod_{\alpha=1}^{k+1} a_{j-\alpha}} \left\{ \frac{1}{a^k_{j-l+1}} \left(\frac{{}_{k-1}C_{l-1}}{a_j} + \frac{{}_{k-1}C_{l-2}}{a_{j-k}} \right) + \frac{1}{a^k_{j-l}} \left(\frac{{}_{k-1}C_{l}}{a_j} + \frac{{}_{k-1}C_{l-1}}{a_{j-k}} \right) \right\}.
\end{aligned}
$$

We can see that

$$
\begin{aligned}
&\frac{{}_{k-1}C_{l-1}}{a_j} + \frac{{}_{k-1}C_{l-2}}{a_{j-k}} \\
&= \frac{(k-1)!}{(k-l)!(l-1)!} \frac{n-j}{n} + \frac{(k-1)!}{(k-l+1)!(l-2)!} \frac{n-j+k}{n} \\
&= \frac{(k-1)!}{(k-l+1)!(l-1)!} \frac{(k-l+1)(n-j) + (l-1)(n-j+k)}{n} \\
&= \frac{(k-1)!}{(k-l+1)!(l-1)!} \frac{k(n-j+l-1)}{n} \\
&= \frac{{}_{k}C_{l-1}}{a_{j-l+1}}.
\end{aligned}
$$

Hence, the coefficient of CV_{j+1-l} can be given by

$$
\frac{(-1)^l n^k}{k! \prod_{\alpha=1}^{k+1} a_{j+\alpha-1}} \left(\frac{1}{a^k_{j-l+1}} \frac{{}_{k}C_{l-1}}{a_{j-l+1}} + \frac{1}{a^k_{j-l}} \frac{{}_{k}C_{l}}{a_{j-l}} \right) = \xi^{(l)}_{k+1,j}.
$$

Therefore, the proof is complete. □

By using Lemma A.2, we present the proof of Theorem 2.1 as follows:

Proof of Theorem 2.2. The case in which $k = 1$ in equation (2.10) is clear because $m_{1,1} = 1$. The case in which $k \geq 2$ can be proven by mathematical induction. Assume that equation (A.17) holds until k, and then consider the case in which $k+1$. We can see that

$$c_j = \prod_{l=1}^{j}(1-a_l) = \prod_{l=1}^{j}\left(1-\frac{n}{n-l}\right) = \prod_{l=1}^{j}\frac{-l}{n-l} = \frac{(-1)^j j!}{n^j}\prod_{l=1}^{j} a_l.$$

It then follows from equation (2.12) that

$$\begin{aligned}\mathrm{CCV}_{k+1} &= \mathrm{CCV}_k - c_{k-1}\Delta_{k,k}\\ &= \sum_{j=1}^{k} m_{k,j}\mathrm{CV}_j - c_{k-1}\sum_{l=0}^{k}\xi_{k,k}^{(l)}\mathrm{CV}_{k+1-l}\\ &= \sum_{j=1}^{k}\left(m_{k,j} - c_{k-1}\xi_{k,k}^{(k+1-j)}\right)\mathrm{CV}_j - c_{k-1}\xi_{k,k}^{(0)}\mathrm{CV}_{k+1}.\end{aligned}$$

The coefficient of CV_{k+1} and CV_1 can be expressed as follows:

$$\begin{aligned}-c_{k-1}\xi_{k,k}^{(0)} &= -\frac{(-1)^{k-1}(k-1)!}{n^{k-1}}\left(\prod_{l=1}^{k-1}a_l\right)\frac{n^{k-1}}{(k-1)!a_k^k\prod_{\alpha=1}^{k}a_\alpha}\\ &= \frac{(-1)^{k+2}}{a_k^{k+1}} = m_{k+1,k+1}.\\ m_{k,1} - c_{k-1}\xi_{k,k}^{(k)} &= \left(1+\frac{k-1}{a_1^k}\right) - \frac{(-1)^{k-1}(k-1)!}{n^{k-1}}\left(\prod_{l=1}^{k-1}a_l\right)\frac{(-1)^k n^{k-1}}{(k-1)!a_1^k\prod_{\alpha=1}^{k}a_\alpha}\\ &= \left(1+\frac{k-1}{a_1^k}\right) + \frac{1}{a_k a_1^k} = 1 + \frac{k}{a_1^{k+1}} = m_{k+1,1}.\end{aligned}$$

The coefficient of CV_j for $j \geq 2$ can be rewritten as follows:

$$\begin{aligned}&m_{k,j} - c_{k-1}\xi_{k,k}^{(k+1-j)}\\ &\quad= (-1)^j\left(\frac{{}_{k-1}C_{j-1}}{a_{j-1}^k} + \frac{{}_{k-1}C_j}{a_j^k}\right)\\ &\qquad -\frac{(-1)^{k-1}(k-1)!}{n^{k-1}}\left(\prod_{l=1}^{k-1}a_l\right)\frac{(-1)^{k+1-j}n^{k-1}}{(k-1)!\prod_{\alpha=1}^{k}a_\alpha}\left(\frac{{}_{k-1}C_{j-1}}{a_j^k} + \frac{{}_{k-1}C_j}{a_{j-1}^k}\right)\\ &\quad= (-1)^{j+1}\left\{\frac{1}{a_j^k}\left({}_{k-1}C_j + \frac{{}_{k-1}C_{j-1}}{a_k}\right) + \frac{1}{a_{j-1}^k}\left({}_{k-1}C_{j-1} + \frac{{}_{k-1}C_{j-2}}{a_k}\right)\right\}.\end{aligned}$$

We can see that

$${}_{k-1}C_j + \frac{{}_{k-1}C_{j-1}}{a_k} = \frac{(k-1)!}{(k-1-j)!j!} + \frac{(k-1)!}{(k-j)!(j-1)!}\frac{n-k}{n}$$

$$
\begin{aligned}
&= \frac{(k-1)!}{(k-j)!j!}\left\{(k-j) + j\frac{n-k}{n}\right\} \\
&= \frac{(k-1)!}{(k-j)!j!}\frac{k(n-j)}{n} \\
&= \frac{{}_kC_j}{a_j}.
\end{aligned}
$$

Hence, the coefficient of CV_j can be given by

$$(-1)^{j+1}\left(\frac{1}{a_j^k}\frac{{}_kC_j}{a_j} + \frac{1}{a_{j-1}^k}\frac{{}_kC_{j-1}}{a_{j-1}}\right) = m_{k+1,j}.$$

Therefore, the proof is complete. □

Acknowledgments

Hirokazu Yanagihara's research was supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Young Scientists (B), #19700265, 2007–2010. Hironori Fujisawa's research was supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Young Scientists (B), #19700270, 2007–2010.

References

[1] Balakrishnan, N., & Ambagaspitiya, R. S. (1994). On skew Laplace distribution. *Technical Report, Department of Mathematics & Statistics, McMaster University*, Hamilton, Ontario, Canada.

[2] Basu, A., Harris, I. R., Hjort, N. L. & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika*, **85**, 549–559.

[3] Chiou, J.-M. & Tsai, C.-L. (2006). Smoothing parameter selection in quasi-likelihood models. *J. Nonparametr. Stat.*, **18**, 307–314.

[4] Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

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

[6] Fujisawa, H. & Eguchi, S. (2006). Robust estimation in the normal mixture model. *J. Statist. Plann. Inference*, **136**, 3989–4011.

[7] Fujisawa, H. & Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination *J. Multivariate Anal.* (in press).

[8] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

[9] Hall, P. & Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika*, **75**, 661–671.

[10] Hurvich, C. M. & Tsai, C.-L. (1998). A crossvalidatory AIC for hard wavelet thresholding in spatially adaptive function estimation. *Biometrika*, **85**, 701–710.

[11] Hurvich, C. M., Simonoff, J. S. & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. Ser.* **B**, **60**, 271–293.

[12] Imoto, S. & Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Ann. Inst. Statist. Math.*, **55**, 671–687.

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

[14] Konishi, S. & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

[15] Konishi, S. & Kitagawa, G. (2003). Asymptotic theory for information criteria in model selection – functional approach. *J. Statist. Plann. Inference*, **114** (2003), 45–61.

[16] Konishi, S. & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* Springer Science+Business Media, LLC, New York.

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

[18] Lindsay, B. G., Markatou, M., Ray, S., Yang, K. & Chen, S.-C. (2008). Quadratic distances on probabilities: a unified foundation. *Ann. Statist.* (in press).

[19] Naik, P. A. & Tsai, C.-L. (2001). Single-index model selections. *Biometrika*, **88**, 821–832.

[20] Nelder, J. A. & Pregibon, D. (1987). An extended quasilikelihood function. *Biometrika*, **74**, 221–232.

[21] Picard, R. R. & Cook, R. D. (1984). Cross-validation of regression models. *J. Amer. Statist. Assoc.*, **79**, 575–583.

[22] Ray, S. & Lindsay, B. G. (2008). Model selection in high dimensions: a quadratic-risk-based approach. *J. Roy. Statist. Soc. Ser.* **B**, **70**, 95–118.

[23] Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, **88**, 486–494.

[24] Simonoff, J. S. & Tsai, C.-L. (1999). Semiparametric and additive model selection using an improved Akaike information criterion. *J. Comput. Graph. Statist.*, **8**, 22–40.

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

[26] 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.

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

[28] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser.* **B**, **58**, 267–288.

[29] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

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

[31] Yanagihara, H. (2007). A family of estimators for multivariate kurtosis in a nonnormal linear regression model. *J. Multivariate Anal.*, **98**, 1–29.

[32] Yanagihara, H. & Yuan, K.-H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis. *British J. Math. Statist. Psych.*, **58**, 209–237.

[33] Yanagihara, H., Sekiguchi, R. & Fujikoshi, Y. (2003). Bias correction of AIC in logistic regression models. *J. Statist. Plann. Inference*, **115**, 349–360.

[34] Yanagihara, H., Tonda, T. & Matsumoto, C. (2006). Bias correction of cross-validation criterion based on Kullback-Leibler information under a general condition. *J. Multivariate Anal.*, **97**, 1965–1975.

[35] Yanagihara, H., Yuan, K.-H., Fujisawa, H. & Hayashi, K. (2008). A class of cross-validatory model selection criteria (submitted for publication).

Table 1. Mean of selected ridge parameter

		Criteria			
Distribution	λ_0	GCV	GIC	CV	CCV_2
Normal	0.7800	0.4129	0.2761	0.8989	0.8601
Laplace	0.7800	0.4090	0.2739	0.8865	0.8489
Uniform	0.7800	0.4170	0.2789	0.9039	0.8645
Skew-Laplace	0.7800	0.4060	0.2733	0.8816	0.8446

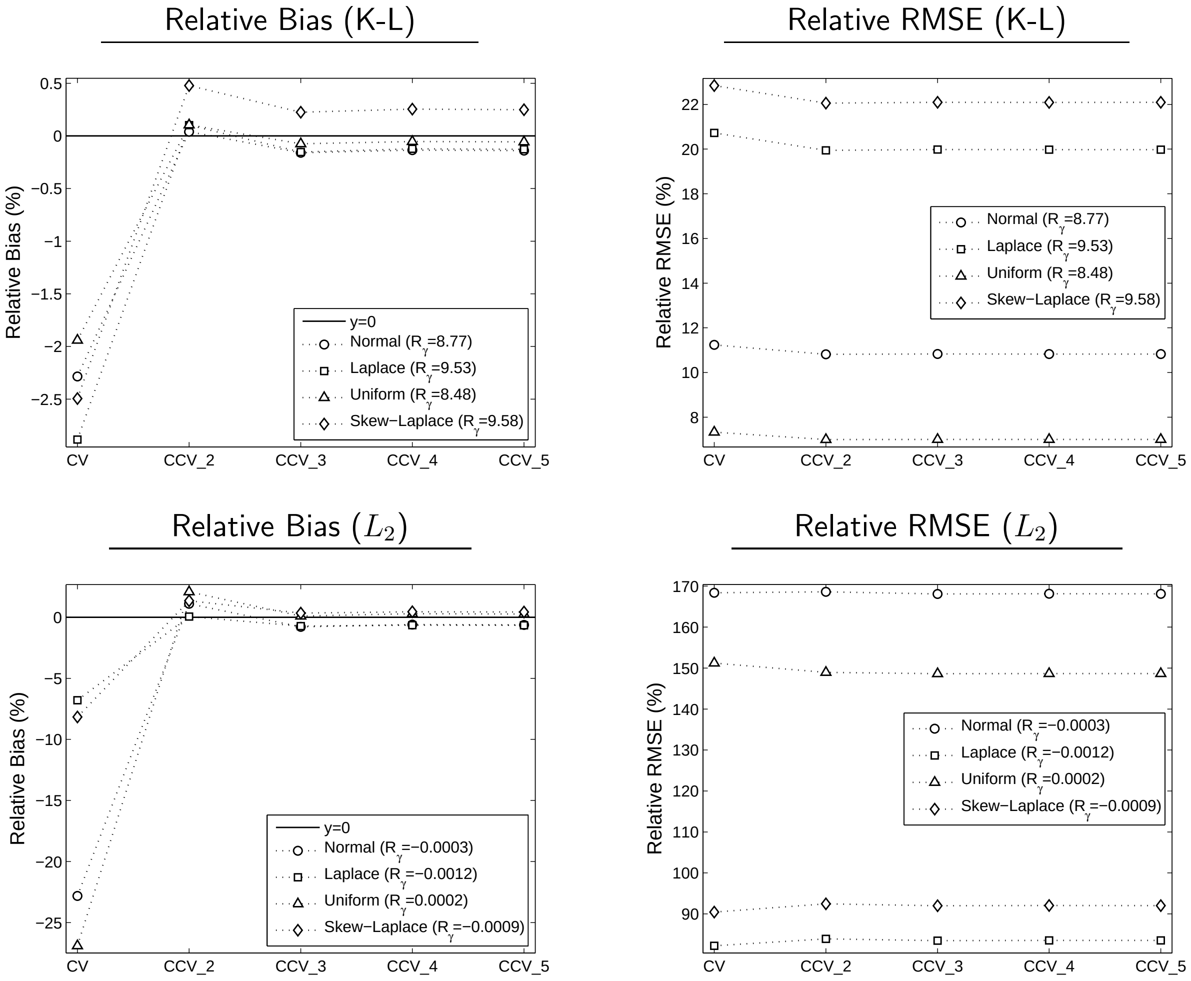


FIGURE 1. Relative biases and RMSEs of K-L and L_2 discrepancy

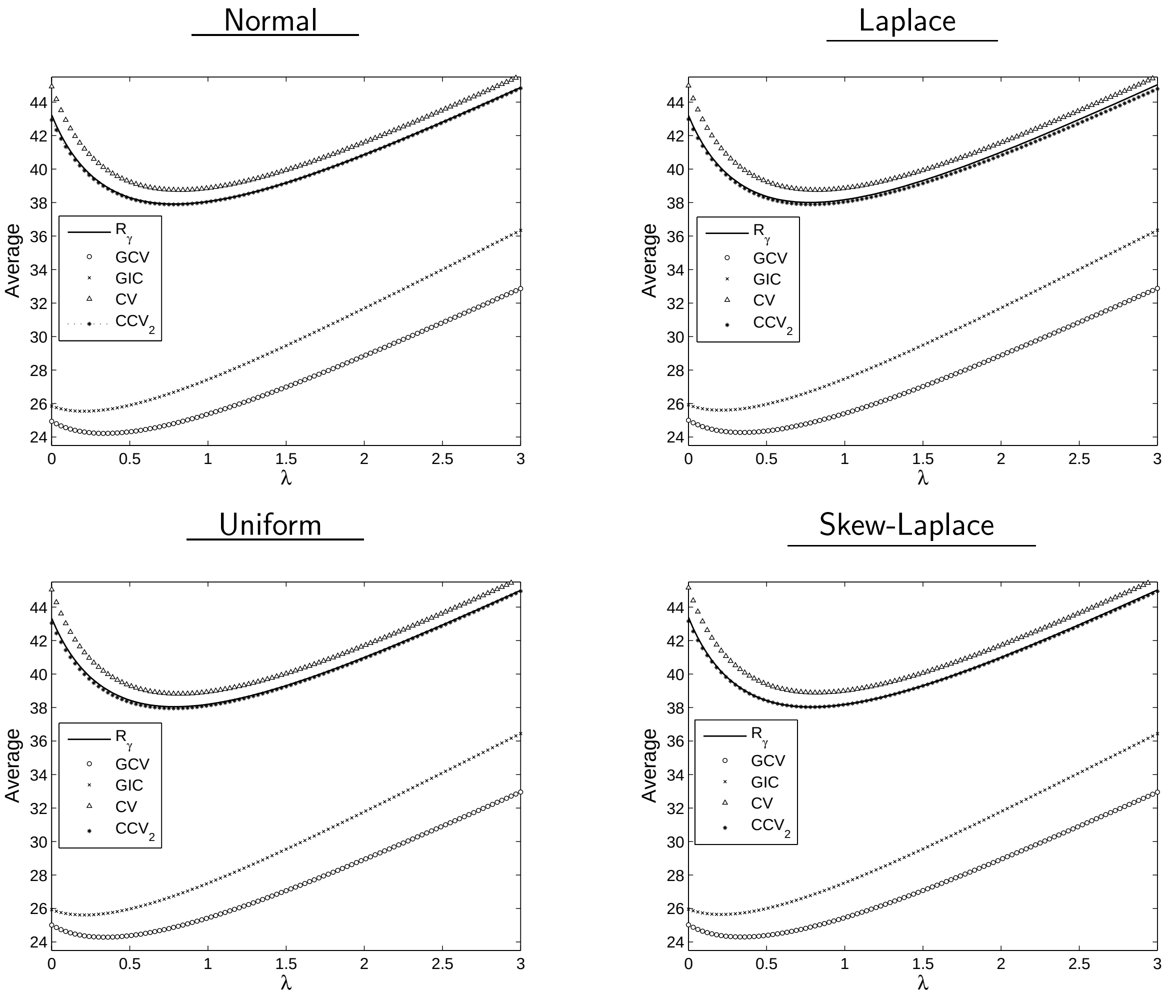


FIGURE 2. Average of GCV, GIC, CV and CCV_2.

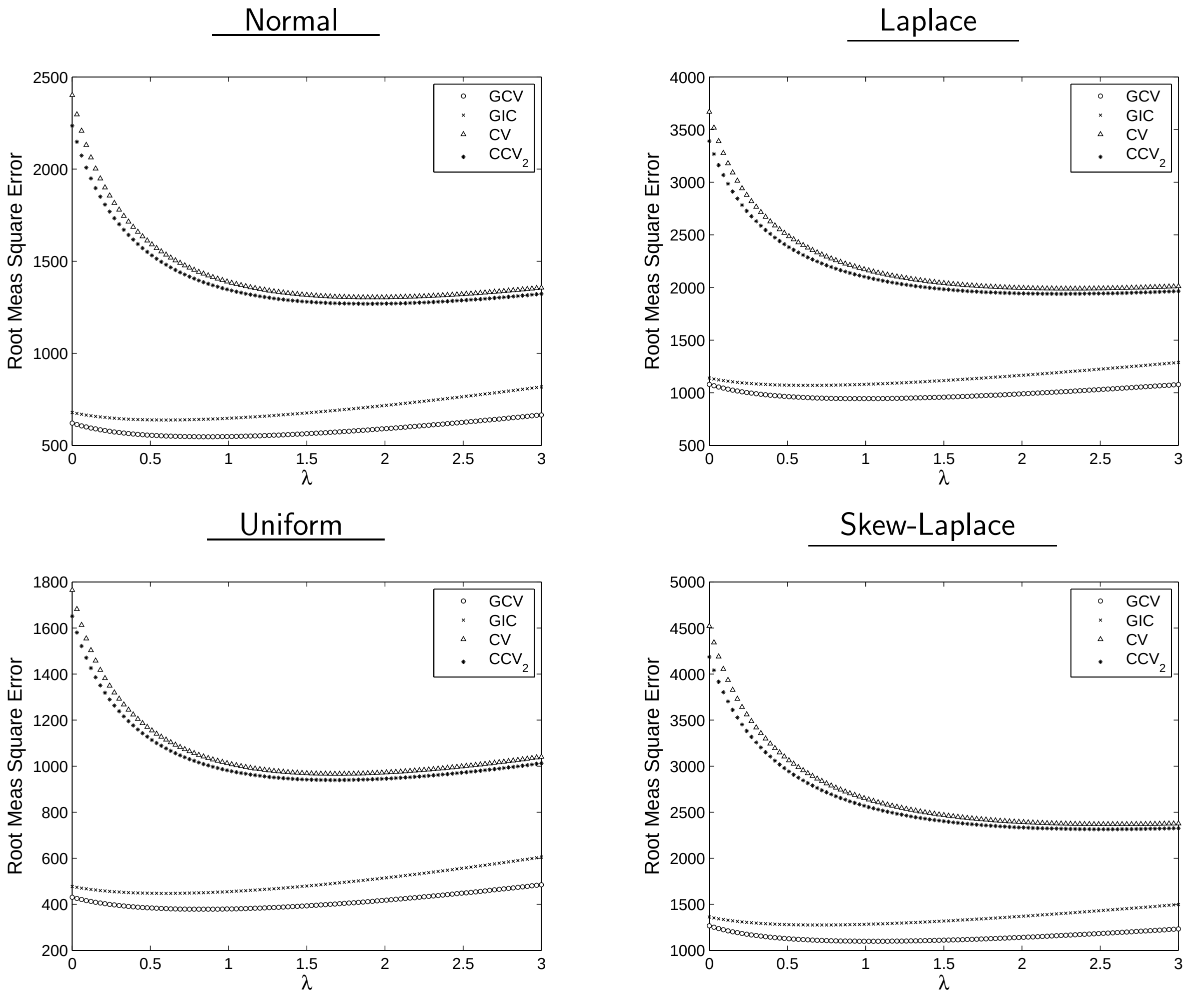
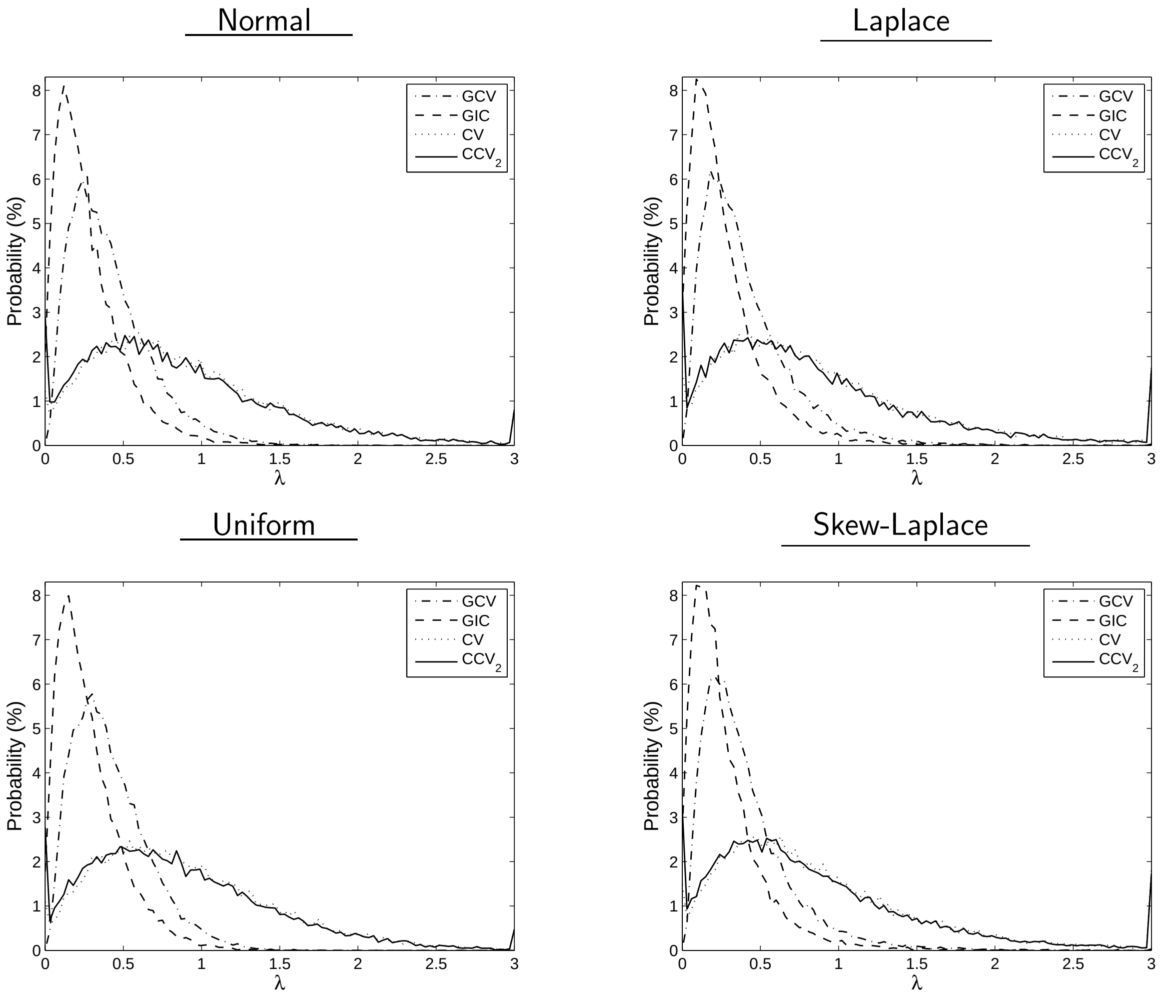


FIGURE 3. MSE of GCV, GIC, CV and CCV_2.



Normal

Laplace

Uniform

Skew-Laplace

FIGURE 4. Frequency of λ selected by GCV, GIC, CV and CCV$_2$.