Stat Comput (2011) 21: 137–146
DOI 10.1007/s11222-009-9153-8

Estimation of prediction error by using K-fold cross-validation

Tadayoshi Fushiki

Received: 14 March 2009 / Accepted: 30 September 2009 / Published online: 10 October 2009

Abstract Estimation of prediction accuracy is important when our aim is prediction. The training error is an easy estimate of prediction error, but it has a downward bias. On the other hand, K-fold cross-validation has an upward bias. The upward bias may be negligible in leave-one-out cross-validation, but it sometimes cannot be neglected in 5-fold or 10-fold cross-validation, which are favored from a computational standpoint. Since the training error has a downward bias and K-fold cross-validation has an upward bias, there will be an appropriate estimate in a family that connects the two estimates. In this paper, we investigate two families that connect the training error and K-fold cross-validation.

Keywords Bias correction · K-fold cross-validation · Large dataset · Prediction error

1 Introduction

Estimation of prediction error is required to evaluate the performance of fitted models. Cross-validation is widely used to estimate the prediction error. Although leave-one-out cross-validation is studied in many researches (for example, Stone 1974, 1977; Efron 2004), K-fold cross-validation may be preferred from a computational standpoint. However, the bias of K-fold cross-validation may become a problem in real data analysis when K is small (Davison and Hinkley 1997). Bias-corrected versions of cross-validation have been proposed by Burman (1989) and Yanagihara et al. (2006). Burman (1989) considered bias correction of K-fold cross-validation, but Yanagihara et al. (2006) only considered bias correction of leave-one-out cross-validation. In this paper, two bias-corrected versions of K-fold cross-validation are derived. The results can be seen as a generalization of the results of Yanagihara et al. (2006).

Cross-validation has also been used for model selection (for example, Li 1987; Shao 1993; Yang 2007), but it should be noted that a better estimate of prediction error does not necessarily lead to a better model selection criterion.

This paper is organized as follows. In Sect. 2, we formulate the problem, and define two families that connect the training error and K-fold cross-validation. In Sect. 3, we consider bias-corrected versions of K-fold cross-validation. In Sect. 4, our methods are tested on some examples. Discussion is given in Sect. 5.

2 Problem formulation

Observations $\mathcal{D} = \{z_1, \ldots, z_N\}$ are independent and identically obtained from an unknown distribution F. If our aim is to construct a prediction model based on $\mathcal{D}$, it is important to estimate its prediction accuracy. We consider an example. Let $Z = (X, Y)$. To explain Y by X, a family $\{h(x;\theta)|\ \theta \in \Theta\}$ is used. Prediction of Y at X is given by $h(X;\hat{\theta})$, where $\hat{\theta}$ is obtained by minimizing $N^{-1}\sum_i (y_i - h(x_i;\theta))^2$. When we use this prediction model, it is important to know how large the prediction error $\mathrm{E}_{(X,Y)}\{(Y - h(X;\hat{\theta}))^2\}$ is.

By generalizing the above, we consider the following problem. Let F_N be the empirical distribution of $\mathcal{D}$. We con-

T. Fushiki (✉)
The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
e-mail: fushiki@ism.ac.jp

sider the following estimator:

$$\hat{\theta}(F_N) = \operatorname*{argmin}_{\theta\in\Theta}\left\{\int \Psi(z;\theta)\mathrm{d}F_N(z)\right\} = \operatorname*{argmin}_{\theta\in\Theta}\left\{\frac{1}{N}\sum_{i=1}^{N}\Psi(z_i;\theta)\right\}.$$

The prediction error is written as

$$s_N = \int \Psi(z;\hat{\theta}(F_N))\mathrm{d}F(z).$$

In the above example, $\Psi(z;\theta) = (y - h(x;\theta))^2$.

An easy estimate of the prediction error is the training error

$$\mathrm{TR}_N = \int \Psi(z;\hat{\theta}(F_N))\mathrm{d}F_N(z) = \frac{1}{N}\sum_{i=1}^{N}\Psi(z_i;\hat{\theta}(F_N)),$$

but it has a downward bias because observations used to estimate θ are again used as future observations.

To estimate the prediction error, K-fold cross-validation is widely used. The data is split into K roughly equal-sized parts $\mathcal{D}^{(1)},\ldots,\mathcal{D}^{(K)}$. Let $m_\alpha = |\mathcal{D}^{(\alpha)}|$; then $\sum_\alpha m_\alpha = N$. The subset obtained by removing $\mathcal{D}^{(\alpha)}$ from $\mathcal{D}$ is denoted by $\mathcal{D}^{(-\alpha)} = \mathcal{D}\setminus\mathcal{D}^{(\alpha)}$. Let $F_{N,K}^{(\alpha)}$ and $F_{N,K}^{(-\alpha)}$ be the empirical distributions of $\mathcal{D}^{(\alpha)}$ and $\mathcal{D}^{(-\alpha)}$, respectively. Then, the K-fold cross-validation estimate of the prediction error is

$$\mathrm{CV}_{N,K} = \sum_{\alpha=1}^{K} p_\alpha \int \Psi(z;\hat{\theta}(F_{N,K}^{(-\alpha)}))\mathrm{d}F_{N,K}^{(\alpha)}(z),$$

where $p_\alpha = m_\alpha/N$. When $K = N$, it is called leave-one-out cross-validation. Although leave-one-out cross-validation is approximately unbiased, the computational cost is considerable except for certain specific problems. For example, when $N = 10000$ and one hour is required to obtain one estimate, it is unrealistic to use leave-one-out cross-validation. In such situations, 5-fold or 10-fold cross-validation will be encouraged (Hastie et al. 2001). In K-fold cross-validation, each estimate of θ is calculated based on a part of $\mathcal{D}$. Therefore, the K-fold cross-validation estimate of the prediction error has an upward bias. Although this upward bias may be negligible in leave-one-out cross-validation, it sometimes cannot be neglected in 5-fold or 10-fold cross-validation.

Since the training error has a downward bias and K-fold cross-validation has an upward bias, there will be an appropriate estimate in a family that connects the two estimates. In this paper, we investigate two such families. The first one is

$$\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda) = (1-\lambda)\mathrm{CV}_{N,K} + \lambda\mathrm{TR}_N. \tag{1}$$

The second one is

$$\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda) = \sum_{\alpha=1}^{K} p_\alpha \int \Psi(z;\hat{\theta}(F_{N,K}^{\lambda(-\alpha)}))\mathrm{d}F_{N,K}^{(\alpha)}(z), \tag{2}$$

where

$$F_{N,K}^{\lambda(-\alpha)} = \frac{N-m_\alpha}{N-m_\alpha+m_\alpha\lambda}F_{N,K}^{(-\alpha)} + \frac{m_\alpha\lambda}{N-m_\alpha+m_\alpha\lambda}F_{N,K}^{(\alpha)}.$$

They coincide with K-fold cross-validation when $\lambda = 0$, and the training error when $\lambda = 1$. Yanagihara et al. (2006) considered bias correction of leave-one-out cross-validation when $\Psi(z;\theta) = -\log p(z;\theta)$. They showed that (1) and (2) are bias-corrected leave-one-out cross-validation estimates of the prediction error when $\lambda = (2N)^{-1} + O(N^{-2})$. However, as discussed above, the bias of K-fold cross-validation will become a problem not when $K = N$, but when K is small. In the next section, we investigate which λ is appropriate for estimation of the prediction error.

3 Bias correction of K-fold cross-validation

According to Burman (1989), the bias of K-fold cross-validation $\mathrm{CV}_{N,K}$ is given by

$$\mathrm{E}(\mathrm{CV}_{N,K} - s_N) = c_0(K-1)^{-1}N^{-1} + o((K-1)^{-1}N^{-1}),$$

where c_0 is a constant that depends on F and Ψ. Burman (1989) proposed the following bias-corrected K-fold cross-validation:

$$\mathrm{CV}^{\mathrm{B}}_{N,K} = \mathrm{CV}_{N,K} + \int \Psi(z;\hat{\theta}(F_N))\mathrm{d}F_N(z) - \sum_{\alpha=1}^{K} p_\alpha \int \Psi(z;\hat{\theta}(F_{N,K}^{(-\alpha)}))\mathrm{d}F_N(z),$$

which satisfies

$$\mathrm{E}(\mathrm{CV}^{\mathrm{B}}_{N,K} - s_N) = o((K-1)^{-1}N^{-1}).$$

Davison and Hinkley (1997) reported in real data analysis that $\mathrm{CV}^{\mathrm{B}}_{N,K}$ significantly improved $\mathrm{CV}_{N,K}$ when K is small.

As described in the previous section, we consider two families $\{\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda)|0 \le \lambda \le 1\}$ and $\{\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda)|0 \le \lambda \le 1\}$, which connect the training error and K-fold cross-validation. In the following, we assume that K is fixed, thus $m = N/K$ is considered as $O(N)$, and $m_\alpha = m$. Since our interest is bias correction of 5-fold or 10-fold cross-validation for large-size datasets, this is a natural assumption. However, the obtained results hold when $K = N$.

The following theorem proved in Appendix provides the appropriate λ.

Theorem 1 *We assume that regularity conditions* (C.1)–(C.6) *described in Appendix hold. Then, we can prove the following two facts.*

1. *Let* $\lambda^{\mathrm{M}} = (2K-1)^{-1}$. *Then,* $\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda^{\mathrm{M}})$ *is asymptotically bias-corrected*:

$$\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda^{\mathrm{M}}) - s_N) = o((K-1)^{-1}N^{-1}).$$

2. *Let* $\lambda^{\mathrm{E}} = (K-1)\{(1-K^{-2})^{-1/2} - 1\}$. *Then,* $\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda^{\mathrm{E}})$ *is asymptotically bias-corrected*:

$$\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda^{\mathrm{E}}) - s_N) = o((K-1)^{-1}N^{-1}).$$

From the theorem, λ^{M} and λ^{E} do not depend on Ψ and F, but depends only on K. Thus, we can use $\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda^{\mathrm{M}})$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda^{\mathrm{E}})$ in data analysis without further estimation. The result that λ^{M} depends only on K is based on the fact that the asymptotic biases of TR_N and $\mathrm{CV}_{N,K}$ are equal up to a multiplicative factor depending only on K. When $K = N$, $\lambda^{\mathrm{M}} = (2N)^{-1} + O(N^{-2})$ and $\lambda^{\mathrm{E}} = (2N)^{-1} + O(N^{-2})$. Thus, the above results are consistent with the results of Yanagihara et al. (2006).

In the following, $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}$ mean $\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda^{\mathrm{M}})$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda^{\mathrm{E}})$, respectively.

4 Examples

Example 1 (Mean) Let $Z_1, \ldots, Z_N$ be independent and identically distributed observations from an unknown distribution F whose mean and variance are μ and σ^2, respectively. To predict a future observation, $\bar{Z} = N^{-1}\sum_i Z_i$ is used. The prediction error is

$$s_N = \mathrm{E}_Z\{(Z-\bar{Z})^2\} = \sigma^2 + (\bar{Z}-\mu)^2.$$

The K-fold cross-validation estimate is

$$\mathrm{CV}_{N,K} = N^{-1}\sum_{\alpha=1}^{K}\sum_{z\in\mathcal{D}^{(\alpha)}}(z-\bar{Z}^{(-\alpha)})^2,$$

where $\bar{Z}^{(-\alpha)} = (N-m)^{-1}\sum_{z\in\mathcal{D}^{(-\alpha)}} z$. Since

$$\mathrm{E}(s_N) = \sigma^2 + \frac{\sigma^2}{N}$$

and

$$\mathrm{E}(\mathrm{CV}_{N,K}) = \sigma^2 + \frac{K\sigma^2}{(K-1)N},$$

the bias of K-fold cross-validation is

$$\mathrm{E}(\mathrm{CV}_{N,K} - s_N) = \frac{\sigma^2}{(K-1)N}.$$

The bias-corrected versions of K-fold cross-validation are

$$\begin{aligned}\mathrm{CV}^{\mathrm{M}}_{N,K} &= \mathrm{CV}_{N,K} - \lambda^{\mathrm{M}}\frac{2K-1}{K^3}\sum_{\alpha=1}^{K}(\bar{Z}^{(\alpha)}-\bar{Z}^{(-\alpha)})^2\\ &= \mathrm{CV}_{N,K} - \frac{1}{K^3}\sum_{\alpha=1}^{K}(\bar{Z}^{(\alpha)}-\bar{Z}^{(-\alpha)})^2\end{aligned}$$

and

$$\begin{aligned}\mathrm{CV}^{\mathrm{E}}_{N,K} &= \mathrm{CV}_{N,K}\\ &\quad - \frac{1}{K}\left\{\frac{2\lambda^{\mathrm{E}}}{K-1+\lambda^{\mathrm{E}}} - \left(\frac{\lambda^{\mathrm{E}}}{K-1+\lambda^{\mathrm{E}}}\right)^2\right\}\\ &\quad \times\sum_{\alpha=1}^{K}(\bar{Z}^{(\alpha)}-\bar{Z}^{(-\alpha)})^2\\ &= \mathrm{CV}_{N,K} - \frac{1}{K^3}\sum_{\alpha=1}^{K}(\bar{Z}^{(\alpha)}-\bar{Z}^{(-\alpha)})^2.\end{aligned}$$

The expectation of the bias correction term is

$$\frac{1}{K^3}\sum_{\alpha=1}^{K}\mathrm{E}\left\{(\bar{Z}^{(\alpha)}-\bar{Z}^{(-\alpha)})^2\right\} = \frac{\sigma^2}{(K-1)N}.$$

Hence, $\mathrm{CV}^{\mathrm{M}}_{N,K} = \mathrm{CV}^{\mathrm{E}}_{N,K}$ are unbiased estimators of the prediction error in this example, but the calculation to obtain them are quite different. We can prove that $\mathrm{CV}^{\mathrm{B}}_{N,K}$ is also equal to $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}$, in this example. They differ in the following examples.

Example 2 (Linear regression) Let $X = (X_1, \ldots, X_{d-1})^T$ be covariates and Y be a response variable. We have N independent and identically distributed observations of the pair $Z = (X, Y)$. A linear model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{d-1}X_{d-1} + \varepsilon$ is assumed, and $\theta = (\beta_0, \beta_1, \ldots, \beta_{d-1})^T$ is estimated by the least squares method. The true distribution is

$$\begin{aligned}&Y = r_0 + \sum_{k=1}^{d-1} r_k X_k + \varepsilon,\\ &\varepsilon \sim N(0, \sigma_0^2),\ X \sim U(-1,1)^{d-1}.\end{aligned}$$

We set $d = 250$ and $\sigma_0 = 1$, and $r = (r_0, r_1, \ldots, r_{d-1})^T$ was determined by a random number from $U(-1,1)^d$. The normalized prediction error $s_N = \mathrm{E}_{(X,Y)}\{(Y - \hat{\theta}^T X)^2\}/(2\sigma_0^2)$ was estimated by $\mathrm{CV}_{N,K}$, $\mathrm{CV}^{\mathrm{B}}_{N,K}$, $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}$. The

Table 1 Summary results for estimates of prediction error in linear regression

N	$\mathrm{E}(s_N)$		$K=5$	$K=10$	$K=N$
1000	0.667	$\mathrm{E}(\mathrm{CV}_{N,K}-s_N)$	0.061	0.026	0.001
		$\mathrm{E}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	0.015	0.007	**0.000**
		$\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	0.022	0.009	**0.000**
		$\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**−0.004**	**−0.001**	**0.000**
		$\mathrm{s.d.}(\mathrm{CV}_{N,K}-s_N)$	0.046	0.042	**0.039**
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	0.043	**0.040**	**0.039**
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	0.043	**0.040**	**0.039**
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**0.041**	**0.040**	**0.039**
2000	0.571	$\mathrm{E}(\mathrm{CV}_{N,K}-s_N)$	0.021	0.009	**0.000**
		$\mathrm{E}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	0.003	0.001	**0.000**
		$\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	0.004	0.001	**0.000**
		$\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**−0.001**	**0.000**	**0.000**
		$\mathrm{s.d.}(\mathrm{CV}_{N,K}-s_N)$	0.023	**0.021**	**0.021**
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	**0.022**	**0.021**	**0.021**
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	**0.022**	**0.021**	**0.021**
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**0.022**	**0.021**	**0.021**

prediction error s_N converges almost surely to 0.5 when $N\to\infty$, and $\mathrm{CV}^{\mathrm{B}}_{N,K}$, $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}$ are different in this example.

Table 1 shows the results based on 10,000 repeats. For each K, the minimum absolute value of the bias and the minimum value of the variance are indicated by bold type. When $N=1000$ and K is small, bias-corrected versions of cross-validation significantly improved ordinary cross-validation. In leave-one-out cross-validation, we can see little improvement. In this example, $\mathrm{CV}^{\mathrm{E}}_{N,K}$ provided the best estimate.

Example 3 (Nonlinear regression) To explain Y based on X, we consider the following nonlinear regression model

$$h(X;\theta)=\beta_0+\sum_{i=1}^{H}\beta_i\varsigma(w_{i0}+w_i^TX), \tag{3}$$

where $\varsigma(x)=1/\{1+\exp(-x)\}$ is the sigmoid function and $\theta=(\beta_0,\ldots,\beta_H,w_{10},w_1^T,\ldots,w_{H0},w_H^T)^T$. The parameter θ is estimated by finding an appropriate local minimum of $N^{-1}\sum_i(Y_i-h(X_i;\theta))^2$ based on ten different starting values. The true distribution is

$$Y=f_0(X)+\varepsilon,\quad \varepsilon\sim N(0,\sigma_0^2),\ X\sim N(0,1)^{d-1},$$

$$f_0(X)=1-3\varsigma(1+X_1+3X_2-X_3-2X_4+5X_5)+5\varsigma(-2+2X_1-3X_2+X_3+2X_4),$$

Table 2 Summary results for estimates of prediction error in nonlinear regression

N	$\mathrm{E}(s_N)$		$K=5$	$K=10$	$K=N$
80	0.681	$\mathrm{E}(\mathrm{CV}_{N,K}-s_N)$	0.112	0.045	0.006
		$\mathrm{E}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	0.056	0.024	0.004
		$\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	0.069	0.028	0.004
		$\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**−0.010**	**−0.002**	**−0.001**
		$\mathrm{s.d.}(\mathrm{CV}_{N,K}-s_N)$	0.243	0.268	0.195
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	0.215	0.251	0.191
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	0.225	0.260	0.194
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**0.194**	**0.218**	**0.187**

where $d=6$ and $\sigma_0=0.5$. The normalized prediction error $s_N=\mathrm{E}_{(X,Y)}\{(Y-h(X;\hat\theta))^2\}/(2\sigma_0^2)$ was estimated by $\mathrm{CV}_{N,K}$, $\mathrm{CV}^{\mathrm{B}}_{N,K}$, $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}$.

We set $H=2$. Table 2 shows the results based on 10,000 repeats. In the simulation, s_N is estimated by a Monte Carlo integration based on 100,000 draws from the true distribution. When K is small, bias-corrected versions of cross-validation significantly improved ordinary cross-validation. In this example, $\mathrm{CV}^{\mathrm{E}}_{N,K}$ provided the best estimate.

Next, we analyze two real data, which can be obtained from the UCI machine learning repository. The first one is the concrete compressive strength data set. It has eight quantitative input variables, and the concrete compressive strength is predicted based on the input variables. The second one is the abalone data set. It has one qualitative variable and seven quantitative input variables, and the age of abalone is predicted based on the input variables. As a preprocessing, quantitative input variables are linearly converted to standardized variables with mean 0 and variance 1. The qualitative variable in abalone data takes three values {F, M, I}. By using two binary variables, it is converted as follows:

$$\mathrm{F}=(0,1),\qquad \mathrm{M}=(0,1),\qquad \mathrm{I}=(0,0).$$

Then, the nonlinear regression model (3) is used for prediction.

The sample size of the concrete compressive strength data is 1030; 824 observations are randomly selected for learning and 206 observations are used for estimation of the true prediction error. The sample size of the abalone data is 4177; 3133 observations are randomly selected for learning and 1044 observations are used for estimation of the true prediction error. For estimation of the parameter in (3), the weight decay method is used:

$$\sum_i(Y_i-h(X_i;\theta))^2+\tau\|\theta\|_2^2$$

Table 3 Summary results for estimates of prediction error in concrete compressive strength data

N	$\mathrm{E}(s_N)$			
824	32.8 (±0.4)			
	$\mathrm{E}(\mathrm{CV}_{N,K}-s_N)$	$\mathrm{E}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	$\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	$\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$
	3.7	1.4	2.2	**−0.6**
	$\mathrm{s.d.}(\mathrm{CV}_{N,K}-s_N)$	$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$
	8.6	7.5	8.0	**6.0**

Table 4 Summary results for estimates of prediction error in abalone data

N	$\mathrm{E}(s_N)$			
3133	4.67 (±0.06)			
1044	9.96 (±0.56)			
	$\mathrm{E}(\mathrm{CV}_{N,K}-s_N)$	$\mathrm{E}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	$\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	$\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$
3133	0.09	−0.03	**−0.01**	−0.15
1044	**−1.00**	−1.96	−1.63	−4.96
	$\mathrm{s.d.}(\mathrm{CV}_{N,K}-s_N)$	$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$
3133	0.94	0.93	0.93	**0.91**
1044	8.78	8.43	8.58	**7.95**

is minimized. We set $H=6$ and $\tau=0.0001$. To estimate the prediction error, 5-fold cross-validation is used. Such a procedure is repeated 200 times.

The results are shown in Tables 3 and 4. In the concrete compressive strength data, $\mathrm{CV}^{\mathrm{E}}_{N,K}$ provided the best result. In the abalone data, $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{B}}_{N,K}$ were good estimates from the standpoint of the bias. When the learning sample size is 1044, the estimation of the parameter was not stable and $\mathrm{CV}^{\mathrm{E}}_{N,K}$ had a large downward bias.

Example 4 (Polynomial regression) In this example, we consider a polynomial regression model

$$h_d(X;\theta)=\beta_0+\beta_1X+\beta_2X^2+\cdots+\beta_{d-1}X^{d-1},$$

where $\theta=(\beta_0,\ldots,\beta_{d-1})^T$. The true distribution is given by

$$Y=f_0(X)+\varepsilon,\quad X\sim U(0,3),\varepsilon\sim N(0,\sigma_0^2),$$
$$f_0(X)=1+5X-7X^2+2X^3,$$

where $\sigma_0=1$. The least squares method is used to estimate θ. When $d\geq 4$, the model contains the true (mean) function. The normalized prediction error $s_N=\mathrm{E}_{(X,Y)}\times\{(Y-h_d(X;\hat{\theta}))^2\}/(2\sigma_0^2)$ was estimated by $\mathrm{CV}_{N,K}$, $\mathrm{CV}^{\mathrm{B}}_{N,K}$, $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}$.

Tables 5 and 6 show the results for $d=2,3,4,5,6$ based on 1 million repeats. When $N=30$, $\mathrm{CV}^{\mathrm{E}}_{N,K}$ provided the best estimates for $d=2,3,4$. However, $\mathrm{CV}^{\mathrm{E}}_{N,K}$ had a large downward bias for $d=6$. In contrast, $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{B}}_{N,K}$ had large upward biases for $d=6$. In Table 6, we can see a tendency that bias-corrected versions of cross-validation select larger models than ordinary cross-validation. The following average prediction error based on the model selected by the cross-validation criteria is also shown in Table 6,

$$\mathrm{E}_{\mathcal{D},(X,Y)}\{(Y-h_{\hat{d}}(X;\hat{\theta}))^2\}/(2\sigma_0^2),$$

where $\hat{d}$ was determined by the cross-validation criteria.

Figure 1 shows the bias and standard deviation of $\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda)$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda)$ when λ varies from 0 to 1.

5 Discussion

Cross-validation is widely used to estimate the prediction error. Even if using leave-one-out cross-validation is desirable, 5-fold or 10-fold cross-validation may be used for large-size datasets because the computational burden of leave-one-out cross-validation is too heavy. However, K-fold cross-validation has an upward bias, and the bias may not be neglected when K is small. We considered two families that connect the training error and K-fold cross-validation. Each family has a parameter λ, and we investigated which λ is appropriate for estimation of the prediction error. The obtained λ depends only on K, thus we can use it without further estimation.

In numerical experiments, the bias-corrected versions of K-fold cross-validation significantly improved ordinary

Table 5 Summary results for estimates of prediction error in polynomial regression

N	K		$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
30	10	$\mathrm{E}(s_N)$	2.092	1.195	0.586	0.641	0.818
		$\mathrm{E}(\mathrm{CV}_{N,K}-s_N)$	0.021	0.028	0.014	0.039	0.186
		$\mathrm{E}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	0.002	0.008	0.004	**0.020**	**0.135**
		$\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	0.002	0.010	0.005	0.025	0.154
		$\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**0.000**	**0.001**	**−0.002**	−0.023	−0.161
		$\mathrm{s.d.}(\mathrm{CV}_{N,K}-s_N)$	0.718	0.500	0.213	1.156	8.242
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	**0.711**	0.491	0.208	1.121	7.740
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	0.712	0.492	0.210	1.137	7.975
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**0.711**	**0.483**	**0.200**	**0.959**	**5.072**
60	10	$\mathrm{E}(s_N)$	2.002	1.094	0.537	0.549	0.564
		$\mathrm{E}(\mathrm{CV}_{N,K}-s_N)$	0.010	0.009	0.004	0.006	0.011
		$\mathrm{E}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	0.001	0.001	**0.000**	**0.001**	0.003
		$\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	0.001	0.001	**0.000**	**0.001**	0.004
		$\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**0.000**	**0.000**	**0.000**	**−0.001**	**−0.002**
		$\mathrm{s.d.}(\mathrm{CV}_{N,K}-s_N)$	0.442	0.234	0.108	0.117	0.152
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{B}}_{N,K}-s_N)$	**0.440**	**0.232**	**0.107**	0.116	0.148
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{M}}_{N,K}-s_N)$	**0.440**	**0.232**	**0.107**	0.116	0.150
		$\mathrm{s.d.}(\mathrm{CV}^{\mathrm{E}}_{N,K}-s_N)$	**0.440**	**0.232**	**0.107**	**0.114**	**0.138**

Table 6 Selected models by the cross-validation criteria and average prediction error (a. p. e.) based on the model selected by the cross-validation criteria in polynomial regression

N	K		$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	a. p. e.
30	10	$\mathrm{CV}_{N,K}$	0.1%	0.4%	72.6%	17.1%	9.8%	0.670
		$\mathrm{CV}^{\mathrm{B}}_{N,K}$	0.1%	0.4%	71.7%	17.5%	10.4%	0.673
		$\mathrm{CV}^{\mathrm{M}}_{N,K}$	0.1%	0.4%	71.9%	17.4%	10.2%	0.672
		$\mathrm{CV}^{\mathrm{E}}_{N,K}$	0.1%	0.3%	68.3%	18.2%	13.1%	0.688
60	10	$\mathrm{CV}_{N,K}$	0.0%	0.0%	74.1%	15.9%	10.0%	0.549
		$\mathrm{CV}^{\mathrm{B}}_{N,K}$	0.0%	0.0%	72.8%	16.4%	10.8%	0.550
		$\mathrm{CV}^{\mathrm{M}}_{N,K}$	0.0%	0.0%	73.0%	16.4%	10.7%	0.550
		$\mathrm{CV}^{\mathrm{E}}_{N,K}$	0.0%	0.0%	72.0%	16.6%	11.3%	0.550

cross-validation when K is small. In our examples, $\mathrm{CV}^{\mathrm{E}}_{N,K}$ provided the minimum variance. This is because $\mathcal{D}^{(\alpha)}$ is used both in estimation and evaluation of the prediction error. When the sample size is not enough and the estimate varies greatly by adding a small amount of $\mathcal{D}^{(\alpha)}$, $\mathrm{CV}^{\mathrm{E}}_{N,K}$ may provide too small prediction error. In contrast, $\mathrm{CV}^{\mathrm{M}}_{N,K}$ and $\mathrm{CV}^{\mathrm{B}}_{N,K}$ may provide too large prediction error in such a case. A higher-order bias-correction may resolve this problem. From the numerical results, $\mathrm{CV}^{\mathrm{E}}_{N,K}$ seems to be a better estimate when $\mathrm{CV}_{N,K}-\mathrm{CV}^{\mathrm{E}}_{N,K}$ and $\mathrm{CV}_{N,K}-\mathrm{CV}^{\mathrm{M}}_{N,K}$ are not quite different, and the variances of bias-corrected versions of K-fold cross-validation were smaller than the variance of ordinary cross-validation, but theoretically these have not been confirmed yet in a general setting. To investigate the basic properties of the bias-corrected cross-validation, simple models were used in this paper, but they can be applied in various problems.

In Tables 1 and 2, the bias of cross-validation estimate of the prediction error was greatly reduced by using $\mathrm{CV}^{\mathrm{E}}_{N,K}$ when $K=5$ and $K=10$, but it seems to be better to use large K even if bias-corrected versions of K-fold cross-validation are used. Hence, there is also the trade-off between the bias and the computational cost if bias-corrected versions of K-fold cross-validation are used.

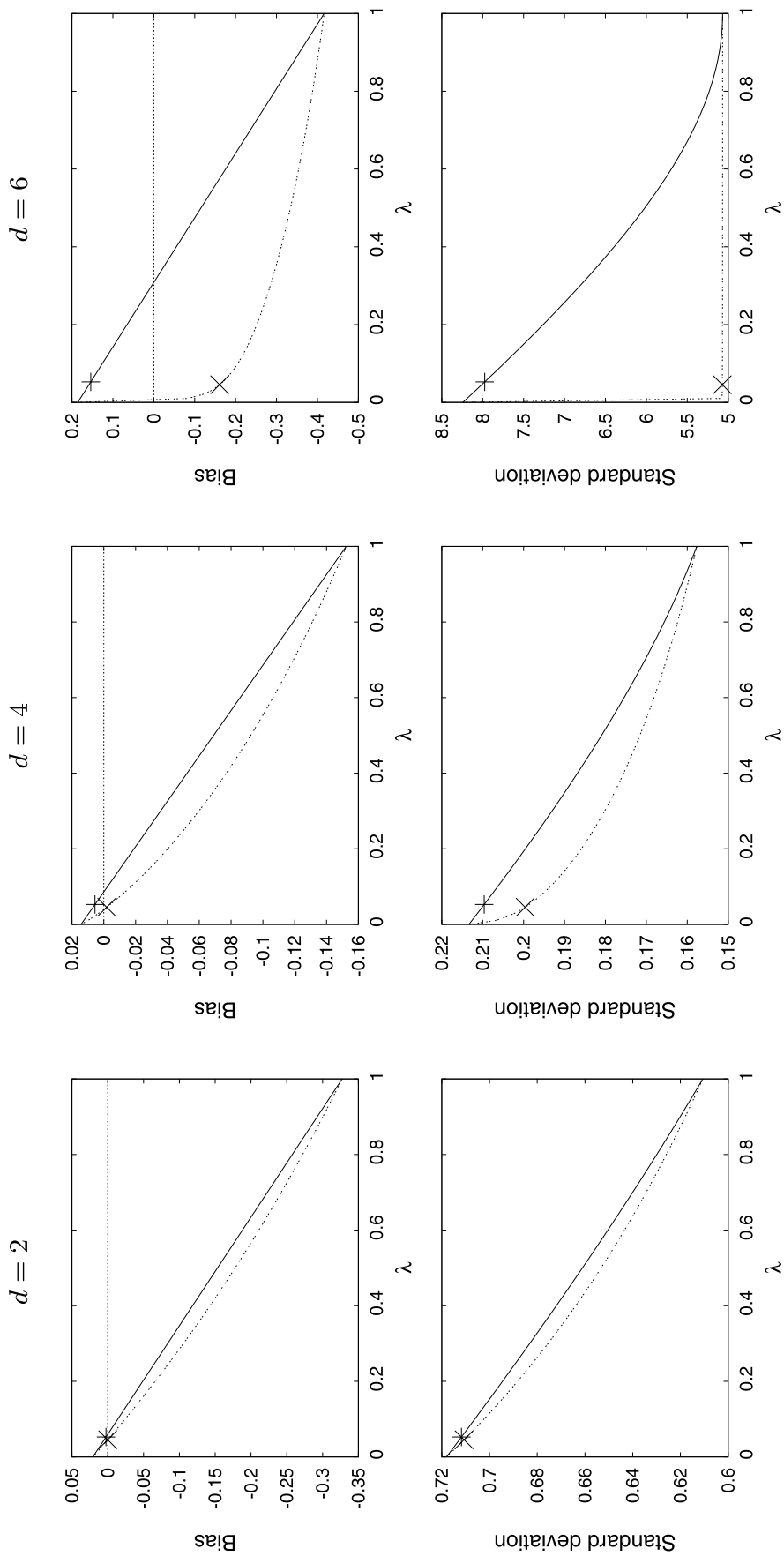

Fig. 1 The bias and standard deviation of $\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda)$ and $\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda)$ when λ varies from 0 to 1 in polynomial regression ($N = 30$, $K = 10$). *The upper three panels* show $\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda) - s_N)$ (*solid line*) and $\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda) - s_N)$ (*dotted line*). *The lower three panels* show s.d.$(\mathrm{CV}^{\mathrm{M}}_{N,K}(\lambda) - s_N)$ (*solid line*) and s.d.$(\mathrm{CV}^{\mathrm{E}}_{N,K}(\lambda) - s_N)$ (*dotted line*). *The left panels* are the results of $d = 2$, *the middle panels* are the results of $d = 4$ and *the right panels* are the results of $d = 6$. The values at λ^{M} and λ^{E} are shown by $+$ and $\times$, respectively

We can consider other families that connect the training error and K-fold cross-validation. For example,

$$\sum_{\alpha=1}^{K} p_\alpha \int \Psi(z; (1-\lambda)\hat{\theta}(F_{N,K}^{(-\alpha)}) + \lambda\hat{\theta}(F_{N,K}))\mathrm{d}F_{N,K}^{(\alpha)}(z),$$

is one candidate. This approach will theoretically work well, but we do not recommend using it. When there are many local minima in $\int \Psi(z;\theta)\mathrm{d}F(z)$, it is difficult to find the global minimum of $N^{-1}\sum_i \Psi(z_i;\theta)$. In such a model, the estimate of θ is obtained by finding an appropriate local minimum of $N^{-1}\sum_i \Psi(z_i;\theta)$ as in Example 3. When $\hat{\theta}(F_{N,K})$ and $\hat{\theta}(F_{N,K}^{(-\alpha)})$ are close to different local minima of $\int \Psi(z;\theta)\mathrm{d}F(z)$, $(1-\lambda)\hat{\theta}(F_{N,K}^{(-\alpha)}) + \lambda\hat{\theta}(F_{N,K})$ may become a poor estimate. Therefore, this approach may not work well in application.

Acknowledgements The author would like to thank two anonymous referees for their helpful comments. This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Japan, Grant-in-Aid for Young Scientists (B), 20700260, 2008–2009.

Appendix: Outline of proof

We define some notation used in this appendix:

$$\Psi(\mathcal{D};\theta) = \sum_{i=1}^{N} \Psi(z_i;\theta),$$

$$\Psi'(\mathcal{D};\theta) = \frac{\partial}{\partial\theta}\Psi(\mathcal{D};\theta),$$

$$\Psi''(\mathcal{D};\theta) = \frac{\partial^2}{\partial\theta\partial\theta^T}\Psi(\mathcal{D};\theta),$$

and

$$\theta_0 = \hat{\theta}(F),$$

$$\hat{\theta} = \hat{\theta}(F_N),$$

$$\hat{\theta}^{(-\alpha)} = \hat{\theta}(F_{N,K}^{(-\alpha)}),$$

$$\hat{\theta}^{[-\alpha]} = \hat{\theta}(F_{N,K}^{\lambda^{\mathrm{E}}(-\alpha)}).$$

(C.1) The parameter space Θ is an open subset of the Euclidean space $\boldsymbol{R}^p$, and $\theta_0 \in \Theta$ is unique.

(C.2) For each z, $\Psi(z;\theta)$ is three times continuously differentiable with respect to θ. There exist a neighborhood B of θ_0 and integrable functions M_1 and M_2 and a square-integrable function M_3 with respect to F such that for each θ in B,

$$\left|\frac{\partial}{\partial\theta_i}\Psi(z;\theta)\right| < M_1(z), \qquad \left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\Psi(z;\theta)\right| < M_2(z),$$

$$\left|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\Psi(z;\theta)\right| < M_3(z), \quad i,j,k = 1,\ldots,p.$$

(C.3) For each θ in B, $J(\theta) = \int \Psi''(z;\theta)\mathrm{d}F(z)$ is nonsingular and

$$\int |\Psi(z;\theta)|^2 \mathrm{d}F(z) < \infty,$$

$$\int \|\Psi''(z;\theta)J(\theta)^{-1}\Psi'(z;\theta)\|^2 \mathrm{d}F(z) < \infty,$$

$$\int \|\Psi'(z;\theta)\Psi'(z;\theta)^T\|_{\mathrm{F}}^2 \mathrm{d}F(z) < \infty,$$

$$\int \|\Psi''(z;\theta)\|_{\mathrm{F}}^2 \mathrm{d}F(z) < \infty,$$

$$\int |\Psi'(z;\theta)^T J(\theta)^{-1}\Psi''(z;\theta)J(\theta)^{-1}\Psi'(z;\theta)|^2 \mathrm{d}F(z) < \infty,$$

where $\|\cdot\|$ means the Euclidean norm and $\|\cdot\|_{\mathrm{F}}$ means the Frobenius norm.

Under these conditions, estimators can be written as the following forms (for example, van der Vaart 1998):

$$\begin{aligned}
\hat{\theta} - \theta_0 &= -N^{-1}J(\theta_0)^{-1}\Psi'(\mathcal{D};\theta_0) + N^{-1/2}R_N,\\
\hat{\theta}^{(-\alpha)} - \theta_0 &= -(N-m)^{-1}J(\theta_0)^{-1}\Psi'(\mathcal{D}^{(-\alpha)};\theta_0)\\
&\quad + N^{-1/2}R_N^{(-\alpha)},\\
\hat{\theta}^{[-\alpha]} - \theta_0 &= -(N-m+m\lambda^{\mathrm{E}})^{-1}J(\theta_0)^{-1}\\
&\quad \times \left\{\Psi'(\mathcal{D}^{(-\alpha)};\theta_0) + \lambda^{\mathrm{E}}\Psi'(\mathcal{D}^{(\alpha)};\theta_0)\right\}\\
&\quad + N^{-1/2}R_N^{[-\alpha]},
\end{aligned}$$

where

$$R_N = o_p(1), \qquad R_N^{(-\alpha)} = o_p(1), \qquad R_N^{[-\alpha]} = o_p(1).$$

Thus,

$$\begin{aligned}
&\hat{\theta} - \hat{\theta}^{(-\alpha)}\\
&\quad = -K^{-1}J(\theta_0)^{-1}\\
&\qquad \times \left\{m^{-1}\Psi'(\mathcal{D}^{(\alpha)};\theta_0) - (N-m)^{-1}\Psi'(\mathcal{D}^{(-\alpha)};\theta_0)\right\}\\
&\qquad + o_p(N^{-1/2}),
\end{aligned}$$

$$\begin{aligned}
&\hat{\theta}^{[-\alpha]} - \hat{\theta}^{(-\alpha)}\\
&\quad = -\frac{\lambda^{\mathrm{E}}}{K-1+\lambda^{\mathrm{E}}}J(\theta_0)^{-1}\\
&\qquad \times \left\{m^{-1}\Psi'(\mathcal{D}^{(\alpha)};\theta_0) - (N-m)^{-1}\Psi'(\mathcal{D}^{(-\alpha)};\theta_0)\right\}\\
&\qquad + o_p(N^{-1/2}).
\end{aligned}$$

Furthermore, we assume the following conditions.

(C.4) The prediction error and the estimators of the prediction error are square-integrable:

$$\mathrm{E}({s_N}^2) < \infty, \qquad \mathrm{E}({\mathrm{TR}_N}^2) < \infty,$$
$$\mathrm{E}({\mathrm{CV}_{N,K}}^2) < \infty, \qquad \mathrm{E}({\mathrm{CV}^{\mathrm{E}}_{N,K}}^2) < \infty.$$

(C.5) Let

$$\begin{aligned} A_N = \max\{&\|R_N\|, \|R_N^{(-1)}\|, \ldots, \|R_N^{(-K)}\|, \\ &\|R_N^{[-1]}\|, \ldots, \|R_N^{[-K]}\|, \\ &\|\hat{\theta} - \theta_0\|, \|\hat{\theta}^{(-1)} - \theta_0\|, \ldots, \\ &\|\hat{\theta}^{(-K)} - \theta_0\|, \|\hat{\theta}^{[-1]} - \theta_0\|, \ldots, \\ &\|\hat{\theta}^{[-K]} - \theta_0\|\}, \end{aligned}$$

then

$$\Pr(A_N \geq N^{-\beta}) = O(N^{-\gamma}),$$

where $\beta \in (0, 1/2)$ and $\gamma > 2$.

(C.6) Let $\tilde{\theta}$ be any estimator appeared in this paper, then

$$\mathrm{E}(\|\tilde{\theta} - \theta_0\|^l) = O(N^{-l/2})$$

for any positive integer l.

Under the conditions (C.1)–(C.6), we can prove the theorem. According to Burman (1989),

$$\mathrm{E}(\mathrm{CV}_{N,K} - s_N) = \frac{\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\}}{2(K-1)N} + o((K-1)^{-1}N^{-1}),$$

where

$$I(\theta) = \int \Psi'(z;\theta)\Psi'(z;\theta)^T \mathrm{d}F(z).$$

As seen in the calculation of information criteria (Konishi and Kitagawa 2007), $\mathrm{TR}_N - s_N$ is decomposed into the following three terms:

$$\begin{aligned} \mathrm{TR}_N - s_N &= N^{-1}\Psi(\mathcal{D};\hat{\theta}) - \int \Psi(z;\hat{\theta})\mathrm{d}F(z) \\ &= \left\{N^{-1}\Psi(\mathcal{D};\hat{\theta}) - N^{-1}\Psi(\mathcal{D};\theta_0)\right\} \\ &\quad + \left\{N^{-1}\Psi(\mathcal{D};\theta_0) - \int \Psi(z;\theta_0)\mathrm{d}F(z)\right\} \\ &\quad + \left\{\int \Psi(z;\theta_0)\mathrm{d}F(z) - \int \Psi(z;\hat{\theta})\mathrm{d}F(z)\right\}. \end{aligned}$$

By calculating each term,

$$\mathrm{E}(\mathrm{TR}_N - s_N) = -N^{-1}\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\} + o(N^{-1}).$$

Thus,

$$\begin{aligned} &\mathrm{E}(\mathrm{CV}^{\mathrm{M}}_{N,K} - s_N) \\ &= \mathrm{E}(\mathrm{CV}_{N,K} - s_N) \\ &\quad + \lambda^{\mathrm{M}}\left\{\mathrm{E}(\mathrm{TR}_N - s_N) - \mathrm{E}(\mathrm{CV}_{N,K} - s_N)\right\} \\ &= o((K-1)^{-1}N^{-1}). \end{aligned}$$

By using the Taylor expansion,

$$\begin{aligned} &\mathrm{CV}^{\mathrm{E}}_{N,K} - \mathrm{CV}_{N,K} \\ &= N^{-1}\sum_{\alpha=1}^{K} \Psi'(\mathcal{D}^{(\alpha)};\hat{\theta}^{(-\alpha)})^T(\hat{\theta}^{[-\alpha]} - \hat{\theta}^{(-\alpha)}) \\ &\quad + \frac{1}{2}N^{-1}\sum_{\alpha=1}^{K}(\hat{\theta}^{[-\alpha]} - \hat{\theta}^{(-\alpha)})^T \\ &\quad \times \Psi''(\mathcal{D}^{(\alpha)};\hat{\theta}^{(-\alpha)})(\hat{\theta}^{[-\alpha]} - \hat{\theta}^{(-\alpha)}) \\ &\quad + o_p((K-1)^{-1}N^{-1}) \\ &= N^{-1}\sum_{\alpha=1}^{K} \Psi'(\mathcal{D}^{(\alpha)};\theta_0)^T(\hat{\theta}^{[-\alpha]} - \hat{\theta}^{(-\alpha)}) \\ &\quad + N^{-1}\sum_{\alpha=1}^{K}(\hat{\theta}^{(-\alpha)} - \theta_0)^T\Psi''(\mathcal{D}^{(\alpha)};\theta_0)(\hat{\theta}^{[-\alpha]} - \hat{\theta}^{(-\alpha)}) \\ &\quad + \frac{1}{2}N^{-1}\sum_{\alpha=1}^{K}(\hat{\theta}^{[-\alpha]} - \hat{\theta}^{(-\alpha)})^T \\ &\quad \times \Psi''(\mathcal{D}^{(\alpha)};\theta_0)(\hat{\theta}^{[-\alpha]} - \hat{\theta}^{(-\alpha)}) \\ &\quad + o_p((K-1)^{-1}N^{-1}). \end{aligned}$$

Thus,

$$\begin{aligned} &\mathrm{E}(\mathrm{CV}^{\mathrm{E}}_{N,K} - s_N) \\ &= \mathrm{E}(\mathrm{CV}_{N,K} - s_N) \\ &\quad - \frac{\lambda^{\mathrm{E}}}{K-1+\lambda^{\mathrm{E}}}\frac{K}{N}\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\} \\ &\quad - \frac{\lambda^{\mathrm{E}}}{K-1+\lambda^{\mathrm{E}}}\frac{K}{(K-1)N}\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\} \\ &\quad + \frac{1}{2}\left(\frac{\lambda^{\mathrm{E}}}{K-1+\lambda^{\mathrm{E}}}\right)^2\frac{K^2}{(K-1)N}\mathrm{tr}\{I(\theta_0)J(\theta_0)^{-1}\} \\ &\quad + o((K-1)^{-1}N^{-1}) \\ &= o((K-1)^{-1}N^{-1}). \end{aligned}$$

References

Burman, P.: A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. Biometrika **76**, 503–514 (1989)

Davison, A.C., Hinkley, D.V.: Bootstrap Methods and Their Application. Cambridge University Press, Cambridge (1997)

Efron, B.: The estimation of prediction error: covariance penalties and cross-validation (with discussion). J. Am. Stat. Assoc. **99**, 619–642 (2004)

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)

Konishi, S., Kitagawa, G.: Information Criteria and Statistical Modeling. Springer, New York (2007)

Li, K.-C.: Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann. Stat. **15**(3), 958–975 (1987)

Shao, J.: Linear model selection by cross-validation. J. Am. Stat. Assoc. **88**, 486–494 (1993)

Stone, M.: Cross-validatory choice and assessment of statistical predictions (with discussion). J. R. Stat. Soc., Ser. B **36**, 111–147 (1974)

Stone, M.: An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J. R. Stat. Soc., Ser. B **39**, 44–47 (1977)

van der Vaart, A.W.: Asymptotic Statistics. Cambridge University Press, Cambridge (1998)

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

Yang, Y.: Consistency of cross validation for comparing regression procedures. Ann. Stat. **35**, 2450–2473 (2007)