

Repeated Half Sampling Criterion for Model Selection

B. Hafidi and A. Mkhadri
Cadi-Ayyad University, Marrakech, Morocco

Abstract

In this paper, the asymptotic property and the performance of the repeated half sampling (RHS) criterion are investigated. In the context of variable selection under a linear regression model, we show that RHS is asymptotically equivalent to the multifold cross-validated (MCV) criterion. While in the case where the candidate family of models doesn't include the true model, we establish that RHS and also MCV are asymptotically equivalent to a criterion similar to Takeuchi information criterion (TIC). The performance of RHS criterion is compared with CV, Akaike, corrected Akaike and BIC criteria. The results of a simulation study show that RHS improve upon the performance of some criteria in two important areas of application: multiple linear regression and multivariate regression.

AMS (2000) subject classification. Primary 62J05; secondary 62E20.

Keywords and phrases. Model selection, repeated half sampling, regression, cross-validation, TIC criterion.

1 Introduction

Cross-validation (CV) criterion is one of the most popular method in selection problems, classification and density estimation. The idea of traditional leave-one-out CV approach is to assess the predictive performance of a model by an average of certain measures for the ability of predicting one observation by model fit after deleting just this observation from training sample. This version of CV as an estimate of the mean squared error of prediction (MSEP) is unsatisfactory (cf. Efron, 1986, Bunke and Droge, 1984). Indeed, Bunke and Droge (1984) showed that some version of bootstrap criterion outperforms CV, which is in accordance with the finding of Efron (1983,1986). When selecting the correct model is the concern, it is well known that the model selected by CV criterion is apt to overfit. Thus, Shibata (1989) showed that the CV criterion is asymptotically equivalent to the Takeuchi information criterion (TIC, Takeuchi, 1976), an extension of the AIC criterion based on the Kulbback-Leibler information without the assumption that the candidate family of models includes the true model.

An alternative approach to CV is the multifold CV (MCV) where instead of deleting one observation as in the CV criterion, d observations are deleted (cf. Zhang, 1993, and references therein). For model selection in regression, Burman (1989), Shao (1993) and Zhang (1993) have each investigated a particular CV procedure where N sample sets are generated independently with a fixed fraction λ being used as test samples, and $1 - \lambda$ being used as training samples. In the later context, Zhang (1993) showed that the MCV criterion is asymptotically equivalent to the well known Final Prediction Error (FPE) criterion of Shibata (1984).

Smyth (2000) propose, for the first time, the use of CV likelihood as a tool for assessing the number K of components in the mixture model. The interest of such an approach is to circumvent the theoretical difficulties concerning the estimation of the honest number of components. His experiments results show that the choice of $\lambda = 0.5$ for Monte Carlo CV approach appears to be reasonably robust across a variety of problems. Nason (1996) proposed a half sampling CV method for selecting the wavelet threshold. Relatively little theory is currently available on the method's performance, although the method was founded to outperform universal thresholding (Donoho and Johnstone, 1994) in simulations studies. Furthermore, Celeux (2001) advocates the use of the repeated half sampling (RHS) criterion as an alternative approach to CV criterion. The main difference with Monte Carlo CV is that all observations of each generated sample are used in the test sample.

The main aim of this note is to study the asymptotic properties of RHS and also its performance in regression context. In particular, under the linear regression model, we show that RHS is asymptotically equivalent to the MCV criterion. On the other hand, when the candidate family of models doesn't include the true model, we establish that RHS and also MCV are asymptotically equivalent to some similar TIC criteria.

The rest of the paper is organized as follows. The general framework is described in Section 2. There, we introduce MCV and RHS criteria. Asymptotical study of RHS in regression context and general case is described in Section 3. In Section 4, we present some experiment results of comparisons of RHS criterion with other classical criteria in the context of regression models. We end the paper with a small discussion.

2 Repeated half sampling criterion

Assessing the number of components in mixture models encounters theoretical difficulties. A way to bypass those difficulties is to make use of resampling procedures. For instance, McLachlan (1987) proposed a parame-

terized bootstrap procedure to the assessment of the P-value of the likelihood ratio test in testing $K = K_0$ versus $K = K_1$. More recently Smyth (2000) proposed to choose the number of components in a mixture by maximizing the cross-validated likelihood. But, in this note we restrict our attention to half sampling criterion proposed by Celeux (2001) which is a particular limit version of cross-validation, and is of practical interest by its very simplicity. Estimation of the expected deviance of the models in competition by half-sampling is as follows. Draw at random two non-overlapping sub-samples \mathbf{s} and $\bar{\mathbf{s}}$ of size $n/2$ (assuming an even n) from the whole sample \mathbf{x} . Then compute the maximum likelihood (ml) estimate $\hat{\theta}_{\mathbf{s}}$ (resp. $\hat{\theta}_{\bar{\mathbf{s}}}$) from \mathbf{s} (resp. $\bar{\mathbf{s}}$). Then, minimizing the penalized deviance leads to choose the model minimizing the following half sampling criterion

$$\text{HS}(\mathbf{x}) = -2 \log \mathbf{f}(\bar{\mathbf{s}}|K, \hat{\theta}_{\mathbf{s}}) - 2 \log \mathbf{f}(\mathbf{s}|K, \hat{\theta}_{\bar{\mathbf{s}}}),$$

where $\mathbf{f}(\mathbf{x}|K, \theta)$ is the distribution under the mixture model parameterized with θ , and $\hat{\theta}$ is the ml estimate of θ . Interest of half sampling is its low time consuming especially for large sample sizes. For small sample sizes, half sampling can be expected to provide unreliable estimate of the expected deviance. In such case, ordinary cross-validation (*leaving-one-out*) procedure must be preferred. But, half sampling produces a bias estimate of the penalization term. Our impression is that this bias is not expected to be high. The high variability of half sampling seems to be a more serious problem. A simple way to attenuate this variability is to make use of a repeated half sampling procedure proposed by Celeux (2001). It consists of repeating half sampling N times as in Smyth (2000). At replication r , we get sub-samples \mathbf{s}_r and $\bar{\mathbf{s}}_r$ which lead to ml estimates $\hat{\theta}_{\mathbf{s}_r}$ and $\hat{\theta}_{\bar{\mathbf{s}}_r}$ and the resulting criterion takes the form

$$\text{RHS}(\mathbf{x}) = -\frac{1}{N} \sum_{r=1}^N [2 \log \mathbf{f}(\bar{\mathbf{s}}_r|K, \hat{\theta}_{\mathbf{s}_r}) + 2 \log \mathbf{f}(\mathbf{s}_r|K, \hat{\theta}_{\bar{\mathbf{s}}_r})]. \quad (1)$$

Now in regression setting, CV criterion is one of the most popular methods for the selection of regression models. But, it is well known that the model selected by CV criterion is apt to overfit. Then, based on results of Efron (1986), some version of bootstrap criteria (called Multifold-CV or MC-CV) are proposed and have provided simulation evidence to do better than simple CV (cf. Bunke and Droge, 1984, Burman, 1989, Breiman and Spector 1989 and Herzberg and Tsukanov, 1986). The asymptotic study of MCV criteria is described in Zhang (1993), there it is shown that the MCV criterion is asymptotically equivalent to the well known Final Prediction Error (FPE) criterion.

3 Asymptotic property of RHS

In this section, we describe RHS criterion as another MCV criterion for model selection in linear regression where we used the same notations as in Zhang (1993). Its asymptotic property is detailed in the same section. On the other hand, we establish that RHS and MCV are asymptotically equivalent to some similar Takeuchi information criteria (TIC, Takeuchi, 1976).

3.1. Equivalence between RHS and MCV criterion. Let $Y = (y_1, \dots, y_n)^t$ be the response vector and $X = (x_{ij}), i = 1, \dots, n, j = 1, \dots, K$, be the design matrix for the full model defined as

$$Y = X\beta + \epsilon$$

where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^t$ is a vector of iid random variables with $E\epsilon_i = 0$ and $E\epsilon_i^2 = \sigma^2$. We suppose that the true parameter vector β has exactly k_0 non zero components, and can be written as $\beta = (\beta_1, \dots, \beta_{k_0}, 0, \dots, 0)^t$. Let s denote a subset of $\{1, \dots, n\}$. For $k \leq K$, we define $X_{s,k} = (x_{ij}), i \in s, j = 1, \dots, k$, $X_k = (x_{ij}), i = 1, \dots, n, j = 1, \dots, k$, $H_{s,k} = X_{s,k}(X_k^t X_k)^{-1} X_{s,k}^t$, and $Y_s = (y_i, i \in s)^t$. Denote by \mathcal{M}_k the regression model with k covariates, and X_k the corresponding design matrix. The deleting-d multifold cross-validation (MCV) criterion can be defined as

$$MCV_k = \left[d \binom{n}{d} \right]^{-1} \sum_s \|Y_s - X_{s,k} \hat{\beta}_{(-s),k}\|^2, \tag{2}$$

where $\hat{\beta}_{(-s),k}$ is the OLS estimate of β under \mathcal{M}_k using the cases not in s . The summation is run over all possible subsets of size d which needs a considerable amount of computation. An alternative is to use the repeated learning-testing approach of Burman (1989) which is essentially a bootstrap method. It consists to resample without replacements d elements from the observed sample and repeat the procedure until obtaining N resampled subsets, s_1^*, \dots, s_N^* , of size d . Then repeated half sampling for model selection in linear regression can be defined as

$$RHS_k = \frac{1}{Nd} \sum_{i=1}^N \{ \|Y_{s_i^*} - X_{s_i^*,k} \hat{\beta}_{\bar{s}_i^*,k}\|^2 + \|Y_{\bar{s}_i^*} - X_{\bar{s}_i^*,k} \hat{\beta}_{s_i^*,k}\|^2 \}, \tag{3}$$

where \bar{s}_i^* denotes the complement of s_i^* in the observed sample (Y, \mathbf{x}) (i.e $s_i^* \cap \bar{s}_i^* = \emptyset$ and $s_i^* \cup \bar{s}_i^* = (Y, \mathbf{x})$). In the sequel, an asymptotic property of this criterion is derived.

The main result of Zhang (1993), based on some assumptions, is that the MCV criterion is asymptotically equivalent to the final prediction error (FPE)

$$\text{RSS}(k) + \alpha k \hat{\sigma}^2(K),$$

where $\alpha = (2 - \lambda)/(1 - \lambda)$, $\lambda > 1$ and where $\text{RSS}(k)$ is the residual sum of squares under model \mathcal{M}_k and $\hat{\sigma}^2(K) = \text{RSS}(K)/(n - K)$. Using similar arguments, the following theorem shows that RHS_k is asymptotically equivalent to MCV_k .

THEOREM 3.1 *We suppose that $\mathbb{E}\epsilon_i^4 \leq \infty$ and the assumptions A to D as in Theorem 1 of Zhang (1993) are satisfied. If $N/n^2 \rightarrow \infty$, then*

$$\text{RHS}_k = 2\text{MCV}_k + o_p(n^{-1}).$$

PROOF. Let $a(s) = d^{-1} \|Y_s - X_{s,k} \hat{\beta}_{(-s),k}\|^2$ and \mathcal{F}_n be the σ -field generated by Y_1, \dots, Y_n . Then conditional on \mathcal{F}_n

$$\text{RHS}_k = \text{RLT}_k^{(1)} + \text{RLT}_k^{(2)}$$

where $\text{RLT}_k^{(1)} = N^{-1} \sum_{i=1}^N a(s_i^*)$ and $\text{RLT}_k^{(2)} = N^{-1} \sum_{i=1}^N a(\bar{s}_i^*)$ are the mean of N iid random variables. Thus, it is easy to show as in Zhang (1993) that

$$\mathbb{E}(\text{RHS}_k | \mathcal{F}_n) = 2\text{MCV}_k.$$

Now

$$\begin{aligned} \text{var}(\text{RHS}_k | \mathcal{F}_n) &= \text{var}(\text{RLT}_k^{(1)} | \mathcal{F}_n) + \text{var}(\text{RLT}_k^{(2)} | \mathcal{F}_n) \\ &\quad + 2\text{cov}(\text{RLT}_k^{(1)}, \text{RLT}_k^{(2)} | \mathcal{F}_n). \end{aligned}$$

From the proof of theorem 4 of Zhang (1993), we have

$$\mathbb{E}(\text{var}(\text{RLT}_k^{(1)} | \mathcal{F}_n) + \mathbb{E}(\text{var}(\text{RLT}_k^{(2)} | \mathcal{F}_n)) = O(N^{-1}).$$

Moreover, using Schwarz's inequality, it follows that

$$\begin{aligned} \text{cov}(\text{RLT}_k^{(1)}, \text{RLT}_k^{(2)} | \mathcal{F}_n) &\leq \sqrt{\text{var}(\text{RLT}_k^{(1)} | \mathcal{F}_n) \text{var}(\text{RLT}_k^{(2)} | \mathcal{F}_n)} \\ &\leq \frac{1}{N} \sqrt{\mathbb{E}(a^2(s_1^*) | \mathcal{F}_n) \mathbb{E}(a^2(\bar{s}_1^*) | \mathcal{F}_n)} \\ &= \left[Nd^2 \binom{n}{d} \right]^{-1} \sum_s \|Y_s - X_{s,k} \beta_{(-s),k}\|^4. \quad (4) \end{aligned}$$

Using the same argument as in the proof of theorem 4 of Zhang (1993), we have

$$\mathbb{E}(\text{cov}(\text{RLT}_k^{(1)}, \text{RLT}_k^{(2)} | \mathcal{F}_n)) = O(N^{-1}).$$

Hence, we can conclude that

$$\mathbb{E}(\text{var}(\text{RHS}_k | \mathcal{F}_n)) = O(N^{-1}),$$

which, as in Zhang(1993), further implies that

$$\text{RHS}_k = \mathbb{E}(\text{RHS}_k | \mathcal{F}_n) + O_p(N^{-1/2}) = 2\text{MCV}_k + o_p(n^{-1}),$$

which ends the proof. □

REMARK. The proof of theorem 3.1 is similar to that of Zhang (1993), but it is necessary to show that

$$\mathbb{E}(\text{cov}(\text{RLT}_k^{(1)}, \text{RLT}_k^{(2)} | \mathcal{F}_n)) = O(N^{-1}).$$

3.2. *Equivalence between RHS and a TIC criterion.* Takeuchi information criterion, (TIC, Takeuchi 1976), is an extension of the AIC criterion based on the Kulbback-Leibler information without the assumption that the candidate family of models includes the true model. Let $X_n^t = (x_1, \dots, x_n)$ independent observations but not necessarily identically distributed. In this section, \mathbb{E} denotes the expectation with respect to the vector of random variable X_n . Since the observations are independent, the log-likelihood can be written as

$$\ell(\theta) = \sum_{i=1}^n \ell_i(\theta), \quad \text{where } \ell_i(\theta) = \log f_i(x_i, \theta).$$

The TIC criterion is defined by

$$\text{TIC} = -2\ell(\hat{\theta}) + 2\text{trace}(\hat{I}\hat{J}^{-1}),$$

where

$$\hat{I} = \sum_i \frac{\partial}{\partial \theta} \ell_i(\hat{\theta}) \frac{\partial}{\partial \theta^t} \ell_i(\hat{\theta}) \quad \text{and} \quad \hat{J} = - \sum_i \frac{\partial^2}{\partial \theta \partial \theta^t} \ell_i(\hat{\theta}).$$

Now, we consider the same assumptions A1 to A3 as in Shibata (1989), and assuming the assumptions A4 and A5 for each replication r :

- A1. The parameter space Θ is a Euclidean p -dimensional space \mathbb{R}^p or an open sub-space of it. Both the Gradient vector

$$g_n(\theta) = \frac{\partial}{\partial \theta_i} \ell(\theta), i = 1, \dots, p$$

and the Hessian matrix

$$H_n = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta), 1 \leq i, j \leq p$$

of the log-likelihood function, are well defined with probability 1, and both continuous with respect to θ .

- A2. $\mathbb{E}|g_n(\theta)| < \infty$ and $\mathbb{E}|H_n(\theta)| < \infty$, where $|\cdot|$ denotes the absolute value of each component of a vector or of a matrix.

- A3. There exist a unique θ^* in Θ , which is the solution of $\mathbb{E}g_n(\theta^*) = 0$. For any $\epsilon > 0$,

$$\sup_{\|\theta - \theta^*\| > \epsilon} (\ell(\theta) - \ell(\theta^*))$$

diverges to $-\infty$ a.s..

- A4. For any $\epsilon > 0$,

$$\max_{\bar{s}_r} \sup_{\|\theta - \theta^*\| > \epsilon} (\ell_{\bar{s}_r}(\theta) - \ell_{\bar{s}_r}(\theta^*))$$

diverges to $-\infty$ a.s. as n tends to infinity, where $\ell_{\bar{s}_r}(\theta) = \ell(\theta) - \ell_{s_r}(\theta)$ for $r = 1, \dots, N$. This implies that $\hat{\theta}_{\bar{s}_r}$ is the solution of

$$\frac{\partial}{\partial \theta} \ell_{\bar{s}_r}(\hat{\theta}_{\bar{s}_r}) = 0, r = 1, \dots, N.$$

- A5. For any $\epsilon > 0$,

$$\max_{\bar{s}_r} \sup_{\|\theta - \theta^*\| > \epsilon} (\ell_{s_r}(\theta) - \ell_{s_r}(\theta^*))$$

diverges to $-\infty$ a.s. as n tends to infinity. This implies that $\hat{\theta}_{s_r}$ is the solution of

$$\frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}_{s_r}) = 0, r = 1, \dots, N.$$

The following theorem establishes the asymptotic equivalence between RHS and a similar TIC criterion.

THEOREM 3.2 *Under assumptions A1 to A5, we have*

$$RHS(x) = -2\ell(\hat{\theta}) + \frac{4}{N} \sum_{r=1}^N \text{trace}(\hat{I}_{s_r} \hat{J}^{-1})(1 + o_p(1)).$$

PROOF. From the definition of $\hat{\theta}_{\bar{s}_r}$, we have

$$\begin{aligned} \frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}_{\bar{s}_r}) &= \frac{\partial}{\partial \theta} \ell(\hat{\theta}_{\bar{s}_r}) \\ &= \frac{\partial}{\partial \theta} \ell(\hat{\theta}) + \frac{\partial^2}{\partial \theta \partial \theta^t} \ell(\hat{\theta})(\hat{\theta}_{\bar{s}_r} - \hat{\theta})(1 + o_p(1)) \\ &= -\hat{J}(\hat{\theta}_{\bar{s}_r} - \hat{\theta})(1 + o_p(1)). \end{aligned}$$

Moreover,

$$\begin{aligned} \ell_{s_r}(\hat{\theta}_{\bar{s}_r}) &= \ell_{s_r}(\hat{\theta}) + (\hat{\theta}_{\bar{s}_r} - \hat{\theta}) \frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}) + (\hat{\theta}_{\bar{s}_r} - \hat{\theta})^t \frac{\partial^2}{\partial \theta \partial \theta^t} \ell_{s_r}(\theta^{**})(\hat{\theta}_{\bar{s}_r} - \hat{\theta}). \\ &= \ell_{s_r}(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta^t} \ell_{s_r}(\hat{\theta}_{\bar{s}_r}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}) \right\} (1 + o_p(1)), \end{aligned} \tag{5}$$

where θ^{**} is a value between $\hat{\theta}$ and $\hat{\theta}_{\bar{s}_r}$. As Shibata (1989, p.224), we assume that $\frac{\partial}{\partial \theta} \ell_i(\theta)$, for $i = 1, \dots, n$, is well defined and continuous with respect to θ . We can assume for large n that

$$\frac{\partial}{\partial \theta^t} \ell_i(\hat{\theta}_{-i}) \approx \frac{\partial}{\partial \theta^t} \ell_i(\hat{\theta}).$$

where $\hat{\theta}_{-i}$ denotes the maximum likelihood estimate of θ based on the sample X_n when the i th observation x_i is deleted. Then for large n , the application this later equation d times leads to

$$\frac{\partial}{\partial \theta^t} \ell_{s_r}(\hat{\theta}_{\bar{s}_r}) = \frac{\partial}{\partial \theta^t} \ell_{s_r}(\hat{\theta}).$$

Substituting this equation into (5), we get

$$\ell_{s_r}(\hat{\theta}_{\bar{s}_r}) = \ell_{s_r}(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta^t} \ell_{s_r}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}) \right\} (1 + o_p(1)). \tag{6}$$

Similarly, we have

$$\frac{\partial}{\partial \theta} \ell_{\bar{s}_r}(\hat{\theta}_{s_r}) = -\hat{J}(\hat{\theta}_{s_r} - \hat{\theta})(1 + o_p(1)),$$

and

$$\ell_{\bar{s}_r}(\hat{\theta}_{s_r}) = \ell_{\bar{s}_r}(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta^t} \ell_{\bar{s}_r}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_{\bar{s}_r}(\hat{\theta}) \right\} (1 + o_p(1)). \quad (7)$$

Now, substituting (6) and (7) into (1) we obtain

$$\begin{aligned} RHS(x) &= -\frac{2}{N} \sum_{r=1}^N \left\{ \ell_{\bar{s}_r}(\hat{\theta}_{s_r}) + \ell_{s_r}(\hat{\theta}_{\bar{s}_r}) \right\} \\ &= -\frac{2}{N} \sum_{r=1}^N \left\{ \ell_{s_r}(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta^t} \ell_{s_r}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}) \right\} (1 + o_p(1)) \right\} \\ &\quad - \frac{2}{N} \sum_{r=1}^N \left\{ \ell_{\bar{s}_r}(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta^t} \ell_{\bar{s}_r}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_{\bar{s}_r}(\hat{\theta}) \right\} (1 + o_p(1)) \right\} \\ &= -2\ell(\hat{\theta}) + \frac{2}{N} \sum_{r=1}^N \left\{ \frac{\partial}{\partial \theta^t} \ell_{s_r}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}) \right. \\ &\quad \left. + \frac{\partial}{\partial \theta^t} \ell_{\bar{s}_r}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_{\bar{s}_r}(\hat{\theta}) \right\} (1 + o_p(1)). \end{aligned} \quad (8)$$

Or, we have $\ell_{\bar{s}_r}(\theta) = \ell(\theta) - \ell_{s_r}(\theta)$ and $\frac{\partial}{\partial \theta} \ell(\hat{\theta}) = 0$. Then

$$\frac{\partial}{\partial \theta} \ell_{\bar{s}_r}(\hat{\theta}) = -\frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}). \quad (9)$$

Substituting (9) into (8), it leads to

$$\begin{aligned} RHS(x) &= -2\ell(\hat{\theta}) + \frac{4}{N} \sum_{r=1}^N \left\{ \frac{\partial}{\partial \theta^t} \ell_{s_r}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_{s_r}(\hat{\theta}) \right\} (1 + o_p(1)) \\ &= -2\ell(\hat{\theta}) + \frac{4}{N} \sum_{r=1}^N \left\{ \sum_{i=1}^d \frac{\partial}{\partial \theta_i^t} \ell_{s_{r_i}}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta_i} \ell_{s_{r_i}}(\hat{\theta}) \right\} (1 + o_p(1)) \\ &= -2\ell(\hat{\theta}) + \frac{4}{N} \sum_{r=1}^N \text{trace}(\hat{I}_{s_r} \hat{J}^{-1}) (1 + o_p(1)). \quad \square \end{aligned}$$

REMARK. If $N = 1$ and $d = n$, we can write

$$\begin{aligned} RHS(x) &= -2\ell(\hat{\theta}) + 4\text{trace}(\hat{I} \hat{J}^{-1}) (1 + o_p(1)) \\ &= TIC + 2\text{trace}(\hat{I} \hat{J}^{-1}) (1 + o_p(1)). \end{aligned}$$

So, we can say that RHS is asymptotically equivalent to a TIC type criterion.

3.3. *Asymptotic study of MCV.* Since RHS is similar to MCV, we deduce the similar asymptotic result for MCV in this section.

THEOREM 3.3 *Under assumptions A1 to A4, the criterion MCV can be written as*

$$MCV_k = -\binom{n}{d}^{-1} \sum_s \left\{ 2\ell_s(\hat{\theta}) - 2\text{trace}(\hat{I}_s \hat{J}^{-1})(1 + o_p(1)) \right\}.$$

PROOF. Indeed, the criterion MCV is defined by

$$MCV_k = -\binom{n}{d}^{-1} \sum_s 2\ell_s(\hat{\theta}_s). \tag{10}$$

Similarly to RHS we have

$$\ell_s(\hat{\theta}_s) = \ell_s(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta^t} \ell_s(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_s(\hat{\theta}) \right\} (1 + o_p(1)). \tag{11}$$

Substituting this equation into (10), we get

$$\begin{aligned} MCV_k &= -2\binom{n}{d}^{-1} \sum_s \left\{ \ell_s(\hat{\theta}) - \left\{ \frac{\partial}{\partial \theta^t} \ell_s(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta} \ell_s(\hat{\theta}) \right\} (1 + o_p(1)) \right\} \\ &= -2\binom{n}{d}^{-1} \sum_s \left\{ \ell_s(\hat{\theta}) - \left\{ \sum_{i=1}^d \frac{\partial}{\partial \theta_i^t} \ell_{s_i}(\hat{\theta}) \hat{J}^{-1} \frac{\partial}{\partial \theta_i} \ell_{s_i}(\hat{\theta}) \right\} (1 + o_p(1)) \right\} \\ &= -\binom{n}{d}^{-1} \sum_s \left\{ 2\ell_s(\hat{\theta}) - 2\text{trace}(\hat{I}_s \hat{J}^{-1})(1 + o_p(1)) \right\}. \quad \square \end{aligned}$$

REMARK. If $d = n$, we can write

$$MCV_k = -2\ell(\hat{\theta}) + 2\text{trace}(\hat{I} \hat{J}^{-1})(1 + o_p(1)).$$

So, we can say that MCV_k is asymptotically equivalent to a TIC type criterion.

4 Numerical Experiments

We carried out a fairly large simulation study of the performance of the RHS against AIC (Akaike, 1973), corrected AIC (AICc, Hurvich and Tsai, 1989, and Bedrick and Tsai, 1994), Bayesian Information Criterion (BIC,

Schwarz, 1978), RLT Burman (1989) and CV method. This simulation study focuses on two important modeling frameworks: the multiple regression and multivariate regression. We consider small to large sample-size setting.

4.1. Multiple regression. Consider the ordinary linear model

$$Y = \mathbf{X}\beta + \epsilon,$$

where Y is an $n \times 1$ observation vector, \mathbf{X} is a known $n \times K$ design matrix, β is a $K \times 1$ vector and ϵ is a gaussian random vector with zero mean and variance-covariance matrix equal $\sigma^2 I$. The goal is to determine which potential independent variables should be included in \mathbf{X} in order to adequately describe the response variable Y . We assume that the candidate models (for $K = 2$ to $K = 8$) are nested. This corresponds to practical settings where the predictor variables can be listed in some order of importance.

We compare the behaviour of the RHS criterion against the other five criteria by simulating a setting where one must decide among seven candidate models M_2, M_3, \dots, M_8 corresponding to nested models of dimension 2 to dimension 8 respectively. In all of our simulations, the values of the first column of X are fixed equal to 1. The values of the other columns are generated from uniform distribution on the interval $(0, 5)$. The variance of the random error is fixed equal to one. One thousand sets of data are generated from the true model, with dimension 3 and 5, in the class candidate. For every data set, the seven models in the candidate class are fit to the data. The criteria AIC, AICc, BIC, RHS, RLT_d with $d = n/2$ and CV are evaluated and the favoured model for each criterion is recorded. Over the one thousand data sets, the selections are summarized in Table 1. The simulations are run using two true models and three sample size: small ($n = 20$), moderate ($n = 50$) and large ($n = 100$). For the RHS, three number of replications are chosen: small ($N = 20$ RHS_S), moderate ($N = 50$ RHS_M) and large ($N = 100$ RHS_L). For small sample size, and for both dimension 3 and 5 of true model, RHS greatly outperforms CV and RLT: RHS correctly chooses the true model 85 – 93% of the time, compared to 68 – 71% and 85 – 88% correct selection rate for CV and RLT, respectively. While the correct selection rate of AIC, AICc, and BIC is, respectively, 60%, 85 – 90% and 78%. For moderate and large sample sizes, the correct selection rate of CV method increases slightly by 0.2 – 0.4% but those of RHS decreases moderately by 0.3 – 16%. Note that the correct selection rate of CV and RHS_L is approximately the same. But the correct selection rate of BIC still better than all other criteria.

TABLE 1. SELECTING DIMENSIONS BY DIFFERENT CRITERIA FOR MULTIPLE REGRESSION

dimen- sion	size	order	Criterion							
			AIC	AIC _c	BIC	CV	RHS _S	RHS _M	RHS _L	RLT _{n/2}
3	20	< 3	0	0	0	0	0	0	0	0
		= 3	595	851	798	681	857	888	888	854
		> 3	405	149	202	319	143	112	112	146
3	50	< 3	0	0	0	0	0	0	0	0
		= 3	672	766	924	700	784	844	864	744
		> 3	328	234	76	300	216	156	136	256
3	100	< 3	0	0	0	0	0	0	0	0
		= 3	698	746	955	712	759	845	855	672
		> 3	302	254	45	288	241	155	145	328
5	20	< 5	0	0	0	0	5	2	3	0
		= 5	613	903	782	713	893	932	910	882
		> 5	384	97	218	287	104	66	87	116
5	50	< 5	0	0	0	0	0	0	0	0
		= 5	708	817	921	744	766	837	865	776
		> 5	292	183	79	256	234	163	135	224
5	100	< 5	0	0	0	0	0	0	0	0
		= 5	735	774	960	751	727	813	845	693
		> 5	265	226	40	249	273	187	155	307

4.2. *Multivariate regression.* Now, as in Cavanaugh and Shumway (1998), we consider the multivariate regression model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U},$$

where the rows of $\mathbf{Y}_{n \times p}$ correspond to p response variables on each of n individuals. $\mathbf{X}_{n \times m}$ is a known matrix of covariate values, and $\beta_{m \times p}$ is a matrix of unknown regression parameters. The rows of the error matrix $\mathbf{U}_{n \times p}$ are assumed to be independents, with identical $\mathcal{N}_p(0, \Sigma)$ distribution. The number of unknown parameters in the approximating multivariate regression is $pm + 0.5p(p + 1)$.

We consider a setting where $p = 2$, so that the rows of \mathbf{Y} represent bi-variate data pair. There were eight candidate models stored in an $n \times 8$ matrix \mathbf{X} , with a column of ones, followed by seven columns of independent measurement on a random variable having a uniform distribution on the interval (0,5). The candidate models included the columns of \mathbf{X} in a sequentially nested fashion; that is, columns 1 to m define the design matrix for the candidate model with m covariates. The design matrix for the true model was the first m_0 columns of \mathbf{X} . We compare the behaviour of the RHS criterion against the other four criteria by simulating a setting where one must choose among eight candidate models (The RLT criterion was deleted

since it was equivalent to the CV criterion). One thousand sets of data are generated from a model having the forme

$$\mathbf{Y}_0 = \mathbf{X}_0\beta_0 + \mathbf{U}, \quad \text{where } \Sigma = \begin{pmatrix} 4 & 7 \\ 7 & 16 \end{pmatrix},$$

with $m_0 = 3$ and $m_0 = 5$; i.e. dimension 9 and 13. The $m_0 \times 2$ parameter matrices β_0 have all elements set equal to 1. For every data set, the eight models in the candidate class are fit to the data. The criteria AIC, AIC_c, BIC, RHS and CV are evaluated and the favoured model for each criterion is recorded. Over the one thousand data sets, the selections are summarized in Table 2. For small sample size, and for both dimension 9 and 13, RHS and CV have a strong tendency to underfit and correctly choose the true model only 22 – 41% and 47% of times, respectively. This may be due to the great dimension of candidate models compared to the small sample size considered. However, all other criteria don't exhibit the same tendency of underfitting as RHS and CV. Note that AIC_c is much better than other criteria with 66 – 81% of correct selection rate. For moderate and large sample sizes, the tendency of RHS and CV to underestimate the correct dimension decreases and is approximatively zero for other criteria. Here RHS was outperformed by BIC and AIC_c, except when $n = 100$ and $N = 100$ where RHS outperforms AIC_c.

TABLE 2. SELECTING DIMENSIONS BY DIFFERENT CRITERIA FOR MULTIVARIATE REGRESSION

dimen- sion	size	order	Criterion						
			AIC	AIC _c	BIC	CV	RHS _S	RHS _M	RHS _L
9	20	< 9	22	137	88	267	540	555	543
		= 9	543	813	747	463	394	393	404
		> 9	435	50	165	270	66	52	53
9	50	< 9	0	0	1	26	75	81	82
		= 9	707	855	963	670	708	757	800
		> 9	293	145	36	304	217	162	118
9	100	< 9	0	0	0	0	0	1	0
		= 9	778	832	987	773	737	847	870
		> 9	222	168	13	257	263	152	130
13	20	< 13	27	314	117	301	759	775	366
		= 13	510	666	665	479	222	209	416
		> 13	463	20	217	220	19	16	218
13	50	< 13	0	0	1	29	123	109	81
		= 13	737	864	959	731	700	744	815
		> 13	263	136	40	240	207	147	104
13	100	< 13	0	0	0	1	2	4	0
		= 13	792	857	988	765	732	808	885
		> 13	208	143	12	233	266	188	115

5 Discussion

In this note, we have presented and studied the asymptotic properties of RHS criterion. In the context of variable selection under the linear model, we have established that RHS is equivalent to the FPE criterion. While in the general case, when the family of models candidate doesn't include the true model, we have shown that RHS and MCV criteria are asymptotically equivalent to some similar TIC criteria. Furthermore, our simulations in multiple regression indicate that, for small sample size and for moderate and large number of replication, RHS outperforms all other criteria. But for other sample size, RHS was outperformed only by BIC criterion. In multivariate regression, RHS has a strong tendency to underfit in the small sample size context. Moreover, RHS was outperformed by BIC and AIC_c for other cases, except when the number of replications and the sample size are large where RHS performs better than the AIC_c criterion.

From Nishii (1984) two notions of consistency of a selection procedure $\hat{k} \in \{1, \dots, K\}$ can be defined in linear model context with fixed dimension K . In general setting, a procedure \hat{k} is called consistent if its probability of selecting the true model tends to one. Now, let $M_0 = \{m \in M_n | m_0 \subset m\}$ be the set of models containing the true one. Then, the procedure \hat{k} is called M_0 -consistent if the probability of selecting a model not including the true one tends to zero. Zhang (1993) established that the RLT method is asymptotically equivalent to FPE criterion. Then, we conclude that Zhang's results imply that under his assumptions the RLT method is M_0 -consistent but not consistent. Another interesting conclusion of Zhang is that the probability of choosing the true model m_0 is an increasing function of λ . Consequently the same asymptotic conclusion are valid for RHS method.

Acknowledgements. The research of the second author was supported in part by TWAS and projet IS2 of INRIA Rhône-Alpes. He would like to thank Gilles Celeux for helpful discussion and for providing him financial support during his visit to IS2.

References

- AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, B.N. Petrov and F. Csaki, eds., Akadémiai Kiadó, Budapest, Hungary, 267-281.
- BEDRICK, E.J. and TSAI, C.L. (1994) Model selection for multivariate regression in small samples. *Biometrics* **50**, 226-231.

- BREIMAN, L. and SPECTOR, P. (1989). Submodel selection and evaluation in regression: The X-random case. *Technical Report 197*, Dept. Statistics, University of California, Berkeley.
- BUNKE, O. and DROGE, B. (1984). Bootstrap and cross-validation estimates of the prediction error for linear regression models. *Ann. Statist.* **12**, 1400-1424.
- BURMAN, L. (1989). A comparative study of ordinary cross-validation, v -fold cross-validation and repeated learning-testing methods. *Biometrika* **76**, 503-514.
- CAVANAUGH, J.E. and SHUMWAY, R.H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *J. Statist. Plann. Inference.* **67**, 45-65.
- CELEUX, G. (2001). Different points of view for choosing the number of components in a mixture model. *10th International Symposium on Applied Stochastic Models and Data Analysis*, G. Govaert, J. Janssen, and N. Limnios, eds., 21-28.
- DONOHO, D.L. and JOHNSTONE, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425-55.
- EFRON, B. (1983). Estimating the error rate of a prediction: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.
- EFRON, B. (1986). How biased is apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.
- HERZBERG, A.M. and TSUKANOV, A.V. (1986). A note on modifications of the jackknife criterion for model selection. *Utilitas Math.* **29**, 209-216.
- HURVICH, C.M. and TSAI, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.
- MCLACHLAN, G.J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.* **36**, 318-324.
- NASON, G.P. (1996). Wavelet shrinkage using cross-validation. *J. Roy. Statist. Soc.*, **58**, 463-479.
- NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758-765.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.
- SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, **88**, 486-494.
- SHIBATA, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika*, **71**, 43-49.
- SHIBATA, R. (1989). Statistical aspects of model selection. In *From Data to Model*, J.C. Willems, ed., Springer, Berlin, 215-240.
- SMYTH, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, **10**, 63-72.
- STONE, M. (1977). An asymptotic equivalence of choice model by cross-validation and Akaike criterion. *J. Roy. Statist. Soc.* **B 39**, 44-47.
- TAKEUCHI, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, **153**, 12-18 (in Japanese).

ZHANG, P. (1993). Model selection via multifold cross validation. *Ann. Statist.*, **21**, 299-313.

B. HAFIDI AND A. MKHADRI
CADI-AYYAD UNIVERSITY
FACULTY OF SCIENCES SEMLALIA
DEPARTMENT OF MATHEMATICS
PB.2390 MARRAKECH, MOROCCO
E-mail: b.hafidi@ucam.ac.ma
mkhadri@ucam.ac.ma

Paper received: January 2004; revised July 2004.