The Annals of Statistics
1993, Vol. 21, No. 1, 299–313

MODEL SELECTION VIA MULTIFOLD CROSS VALIDATION[1]

BY PING ZHANG

University of Pennsylvania

A natural extension of the simple leave-one-out cross validation (CV) method is to allow the deletion of more than one observations. In this article, several notions of the multifold cross validation (MCV) method have been discussed. In the context of variable selection under a linear regression model, we show that the delete-d MCV criterion is asymptotically equivalent to the well known FPE criterion. Two computationally more feasible methods, the r-fold cross validation and the repeated learning-testing criterion, are also studied. The performance of these criteria are compared with the simple leave-one-out cross validation method. Simulation results are obtained to gain some understanding on the small sample properties of these methods.

1. Introduction. One of the most useful methods in selection problems is the cross validation (CV) method. During the past decade, the CV method has been developed quite extensively in the literature, especially in the area of nonparametric curve estimation. One of the appealing characteristics of CV is that it is applicable to a wide variety of problems, thus giving rise to applications in many areas. Examples include, but are not limited to, the choice of smoothing parameters in nonparametric smoothing and variable selection in regression. A considerable amount has been written on both the theoretical and practical aspects of this method. The idea is simply splitting the data into two parts, using one part to derive a prediction rule and then judge the goodness of the prediction by matching its outputs with the rest of the data, hence the name cross validation. One should, however, notice that in the literature, unless indicated explicitly, CV is usually referred to as the simple leave-one-out cross validation. This version of CV is unsatisfactory in several respects. Efron (1986) showed that the simple CV is a poor candidate for estimating the prediction error and suggested that some version of bootstrap would be better off. When selecting the correct model is the concern, it is well known that the model selected by CV criterion is apt to overfit.

The idea of multifold cross validation (MCV) first appeared in Geisser (1975) where instead of deleting one observation as in the simple CV, $d > 1$ observations are deleted. Some recent development in this area can be found in Breiman, Friedman, Olshen and Stone (1984) and Burman (1989). For general variance estimation problem, Shao and Wu (1989) introduced multifold jackknife and successfully remedied a problem encountered by the simple leave-one-out jackknife. For model selection, Breiman and Spector (1989) and

Received October 1990; revised May 1992.
[1]Partially supported by ONR N00014-89-J-1562.
AMS 1991 *subject classifications*. Primary 62J05; secondary 62E20, 65C05.
Key words and phrases. Bootstrap, FPE criterion, model selection, multifold cross validation.

Herzberg and Tsukanov (1986) have provided simulation evidence that MCV does better than simple CV. For choosing the number of knots in spline method, Burman (1990) shows that the multifold cross validation method is asymptotically optimal. This paper treats only the case of linear regression and our goal is to investigate, from a theoretical point of view, the performance of various notions of MCV model selection criteria. In particular, we show that the delete-d MCV is asymptotically equivalent to the well known FPE criterion.

Let $Y = (y_1, \ldots, y_n)^t$ be the response vector and $X = (x_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, K$, be the design matrix for the full model defined as

$$Y = X\beta + \varepsilon,$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$ is a vector of iid random variables. Suppose that the true model has k_0 covariates, or the true parameters β has exactly k_0 nonzero components. Throughout this paper, it is assumed that $\beta = (\beta_1, \ldots, \beta_{k_0}, 0, \ldots, 0)^t$. This corresponds to the situation where the K covariates are preordered according to their importance so that only the number of covariates needs to be determined. Let s denote a subset of $\{1, \ldots, n\}$. For $k \leq K$, we define

$$X_{s,k} = (x_{ij}), \qquad i \in s,\ j = 1, \ldots, k,$$

$$X_k = (x_{ij}), \qquad i = 1, \ldots, n,\ j = 1, \ldots, k,$$

$$H_{s,k} = X_{s,k}\left(X_k^t X_k\right)^{-1} X_{s,k}^t, \qquad Y_s = (y_i, i \in s)^t.$$

Denote by $\mathcal{M}_k$ the regression model with k covariates, and X_k the corresponding design matrix. We define the deleting-d multifold cross validation criterion as

$$\text{MCV}_k = \left[d\binom{n}{d}\right]^{-1} \sum_s \|Y_s - X_{s,k}\hat{\beta}_{(-s),k}\|^2, \tag{1.1}$$

where $\hat{\beta}_{(-s),k}$ is the OLS estimate of β under $\mathcal{M}_k$ using the cases not in s. The summation runs over all possible subsets of size d.

This notion of MCV has an obvious disadvantage, namely that a considerable amount of computation is involved. However, the performance of many useful alternative methods are closely related to the performance of the criterion (1.1). We consider in this paper two such methods: The r-fold cross validation of Breiman, Friedman, Olshen and Stone (1984) and the repeated learning-testing method of Burman (1989). Suppose that the sample size n can be written as $n = rd$, where r and d are integers. Instead of summing over all possible subsets of size d as in (1.1), let us divide $\{1, \ldots, n\}$ into r subgroups $s_1, \ldots, s_r$ which are mutually exclusive. Without losing generality, suppose that the division is as follows:

$$\overbrace{1, \ldots, d}^{s_1}, \overbrace{d+1, \ldots, 2d}^{s_2}, \ldots, \overbrace{(r-1)d, \ldots, rd}^{s_r}.$$

Breiman, Friedman, Olshen and Stone (1984) define their r-fold cross validation as

$$\text{MCV}_k^* = \frac{1}{n}\sum_{i=1}^{r} \|Y_{s_i} - X_{s_i,k}\hat{\beta}_{(-s_i),k}\|^2. \tag{1.2}$$

Actually, the above authors suggested that the partition should be made randomly to avoid possible biases.

The repeated learning-testing method is essentially a bootstrap method. Here, instead of summing over all possible subsets of size d, we resample without replacement d elements from the observed sample and repeat the procedure many times. Let $s_1^*, \ldots, s_N^*$ be the resampled subsets of size d. The repeated learning-testing criterion is defined by

$$\text{RLT}_k = \frac{1}{Nd}\sum_{i=1}^{N} \|Y_{s_i^*} - X_{s_i^*,k}\hat{\beta}_{(-s_i^*),k}\|^2. \tag{1.3}$$

The rest of the paper is organized the following way: In Section 2, some basic results are given. As an implication of these results, we show an interesting relationship between criterion (1.1) and the well known FPE criterion. Sections 3 and 4 treat the r-fold cross validation and the repeated learning-testing method, respectively. Finally in Section 5, brief discussion on some of the practical issues are presented along with some simulation results.

2. Basic results. Let $d = \#\{i\colon i \in s\}$, $f = X\beta$ and $P_k^{\perp} = I - P_k$, where $P_k = X_k(X_k^t X_k)^{-1}X_k^t$. We introduce the following assumptions:

ASSUMPTION A. $d \to \infty$, and $d/n = \lambda + o(1)$, where $\lambda > 0$.

ASSUMPTION B. $\sup_{d\to\infty}\sup_s \|d^{-1}X_{s,k}^t X_{x,k} - V_k\| = o(1)$, where V_k, $k \le K$ is a sequence of positive definite matrices.

ASSUMPTION C. For $k < k_0$,

$$\liminf_{n\to\infty} n^{-1}f^t P_k^{\perp} f = b_k > 0 \quad \text{and} \quad n^{-1}f^t P_k f \to 0.$$

ASSUMPTION D. For $k \le K$, $\max_{i\le n} h_{ii}^{(k)} \to 0$, where $h_{ii}^{(k)}$, $i = 1, \ldots, n$, are the diagonal elements of P_k.

Except for the first one, these assumptions are rather mild for asymptotic results. Actually, Assumption A is essential for all the proceeding results. Our conjecture is that the first order asymptotic structure of MCV with $d/n \to 0$ would be equivalent to the simple CV. For Assumption C, notice that b_k is decreasing and when $k \ge k_0$, $b_k = 0$.

The final prediction error (FPE) criterion can be written as

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

where $\mathrm{RSS}(k)$ is the residual sum of squares under model $\mathcal{M}_k$ and $\hat{\sigma}^2(K) = \mathrm{RSS}(K)/(n - K)$. The main conclusion of this section is to show that the MCV criterion (1.1) is asymptotically equivalent to the FPE criterion with $\alpha = (2 - \lambda)/(1 - \lambda)$. In order to achieve this, we prove three useful lemmas below.

Taken literally, the formula given by (1.1) requires the computation of a least squares estimator $\hat{\beta}_{(-s),k}$ for all subsets of size d. This amounts to solving $\binom{n}{d}$ linear equations of dimension $n - d$. The following result gives the relationship between $\hat{\beta}_{(-s),k}$ and $\hat{\beta}_k$ which in turn causes tremendous reduction in computation. More importantly, this relationship also provides us with a theoretically more illuminating representation of MCV.

LEMMA 1. *Under Assumptions A and B, we have for large d that*

$$Y_s - X_{s,k}\hat{\beta}_{(-s),k} = (I - H_{s,k})^{-1}\left(Y_s - X_{s,k}\hat{\beta}_k\right).$$

PROOF. For any matrices $A_{p\times p}$ and $U_{p\times n}$, $p < n$, it is straightforward to verify that

$$(A - U^tU)^{-1} = A^{-1} + A^{-1}U^t(I - UA^{-1}U^t)^{-1}UA^{-1}, \tag{2.1}$$

provided that all the inverses exist. This is often referred to as the Sherman–Morrison–Woodbury formula. Take $A = X_k^tX_k$ and $U = X_{s,k}$. It is easy to see from the assumptions that the inverses all exist when d is large. Hence by (2.1),

$$\begin{aligned}
&\left(X_{(-s),k}^tX_{(-s),k}\right)^{-1} \\
&\quad = \left(X_k^tX_k - X_{s,k}^tX_{s,k}\right)^{-1} \\
&\quad = \left(X_k^tX_k\right)^{-1} + \left(X_k^tX_k\right)^{-1}X_{s,k}^t(I - H_{s,k})^{-1}X_{s,k}\left(X_k^tX_k\right)^{-1},
\end{aligned}$$

where $X_{(-s),k} = (x_{ij})$, $i \notin s$. Observe that

$$X_{(-s),k}^tY_{(-s)} = X_k^tY - X_{s,k}^tY_s$$

and

$$\hat{\beta}_{(-s),k} = \left(X_{(-s),k}^tX_{(-s),k}\right)^{-1}X_{(-s),k}^tY_{(-s)}.$$

Some algebra will show that

$$\begin{aligned}
X_{s,k}\hat{\beta}_{(-s),k} = {} & X_{s,k}\hat{\beta}_k + H_{s,k}(I - H_{s,k})^{-1}X_{s,k}\hat{\beta}_k \\
& - H_{s,k}Y_s - H_{s,k}(I - H_{s,k})^{-1}H_{s,k}Y_s.
\end{aligned}$$

Consequently,

$$\begin{aligned} Y_s - X_{s,k}\hat{\beta}_{(-s),k} &= Y_s - X_{s,k}\hat{\beta}_k - H_{s,k}(I - H_{s,k})^{-1}X_{s,k}\hat{\beta}_k \\ &\quad + H_{s,k}Y_s + H_{s,k}(I - H_{s,k})^{-1}H_{s,k}Y_s \\ &= \left[I + H_{s,k}(I - H_{s,k})^{-1}\right]\left(Y_s - X_{s,k}\hat{\beta}_k\right) \\ &= (I - H_{s,k})^{-1}\left(Y_s - X_{s,k}\hat{\beta}_k\right). \end{aligned}$$ □

The above lemma shows that

$$\text{MCV}_k = \left[d\binom{n}{d}\right]^{-1} \sum_s \left\|(I - H_{s,k})^{-1}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2. \tag{2.2}$$

In other words, for each model $\mathcal{M}_k$, $\hat{\beta}_k$ only needs to be calculated once. When $d = 1$, this reduces to the ordinary cross validation or PRESS. Likewise, we have

$$\text{MCV}_k^* = \frac{1}{n}\sum_{i=1}^{r} \left\|(I - H_{s_i,k})^{-1}\left(Y_{s_i} - X_{s_i,k}\hat{\beta}_k\right)\right\|^2, \tag{2.3}$$

and

$$\text{RLT}_k = \frac{1}{Nd}\sum_{i=1}^{N} \left\|(I - H_{s_i^*,k})^{-1}\left(Y_{s_i^*} - X_{s_i^*,k}\hat{\beta}_k\right\|^2. \tag{2.4}$$

The following two lemmas are key to our main result. It is essential to assume Assumption A, that is, one has to delete a fixed proportion of the whole sample.

LEMMA 2. *Let $P_{s,k}$ and P_k be the projection matrices corresponding the $X_{s,k}$ and X_k, respectively. Suppose that $\mathbf{E}\varepsilon_i = 0$, $\mathbf{E}\varepsilon_i^2 = \sigma^2$, $i = 1, \ldots, n$. Then under Assumptions A, B and D,*

$$\left[d\binom{n}{d}\right]^{-1} \sum_s \varepsilon_s^t P_{s,k}\varepsilon_s = \frac{1}{n}\left(\varepsilon^t P_k \varepsilon + \frac{1-\lambda}{\lambda}k\sigma^2\right) + o_p(n^{-1}).$$

PROOF. By definition and Assumption B, it is easy to see that

$$\begin{aligned} P_{s,k} &= X_{s,k}\left(X_{s,k}^t X_{s,k}\right)^{-1}X_{s,k}^t \\ &= \frac{n}{d}X_{s,k}\left(X_k^t X_k\right)^{-1}X_{s,k}^t + o_p(1) \\ &= \lambda^{-1}H_{s,k} + o_p(1). \end{aligned}$$

Next, let $\tilde{H}_{s,k} = (\tilde{h}_{ij})$ be the $n \times n$ matrix with $\tilde{h}_{ij}$ equaling to the corresponding element in $H_{s,k}$ if $i, j \in s$ and $\tilde{h}_{ij} = 0$ otherwise. Notice that $H_{s,k}$ is actually a diagonal submatrix of P_k. Simple combinatorics will show that $\sum_s \tilde{H}_{s,k}$, while summing over all possible subsets, will accumulate the diagonal

elements of P_k $\binom{n-1}{d-1}$ times, and the off diagonal elements $\binom{n-2}{d-2}$ times. Consequently,

$$\begin{aligned}\left[d\binom{n}{d}\right]^{-1}\sum_s \varepsilon_s^t H_{s,k}\varepsilon_s &= \left[d\binom{n}{d}\right]^{-1}\sum_s \varepsilon^t \tilde{H}_{s,k}\varepsilon \\ &= \left[d\binom{n}{d}\right]^{-1}\cdot\left[\binom{n-2}{d-2}\varepsilon^t P_k\varepsilon\right. \\ &\qquad \left.+\left(\binom{n-1}{d-1}-\binom{n-2}{d-2}\right)\varepsilon^t \operatorname{diag}(P_k)\varepsilon\right] \\ &= \frac{\lambda}{n}\left(\varepsilon^t P_k\varepsilon + \frac{1-\lambda}{\lambda}k\sigma^2\right)+o_p(n^{-1}).\end{aligned}$$

The last equation is due to Assumption D, which implies that $\varepsilon^t \operatorname{diag}(P_k)\varepsilon = k\sigma^2 + o_p(1)$. The proof is completed by noting the relationship between $H_{s,k}$ and $P_{s,k}$ shown above. □

LEMMA 3. *Under the same assumptions of Lemma 2, if $k \geq k_0$, then*

$$\left[d\binom{n}{d}\right]^{-1}\sum_s \left\| P_{s,k}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2 = \frac{1-\lambda}{\lambda}\cdot\frac{k\sigma^2}{n}+o_p(n^{-1}).$$

PROOF. Let $Y_s = f_s + \varepsilon_s$ and $Y = f + \varepsilon$. When $k \geq k_0$, it is easy to verify that $P_{s,k}f_s = X_{s,k}(X_k^tX_k)^{-1}X_k^tf$. Thus

$$\begin{aligned}\left\| P_{s,k}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2 &= \left\| P_{s,k}\left[f_s+\varepsilon_s - X_{s,k}\left(X_k^tX_k\right)^{-1}X_k^t(f+\varepsilon)\right]\right\|^2 \\ &= \left\| P_{s,k}\varepsilon_s - X_{s,k}\left(X_k^tX_k\right)^{-1}X_k^t\varepsilon\right\|^2 \\ &= \varepsilon_s^tP_{s,k}\varepsilon_s - 2\varepsilon_s^tX_{s,k}\left(X_k^tX_k\right)^{-1}X_k^t\varepsilon \\ &\quad + \varepsilon^tX_k\left(X_k^tX_k\right)^{-1}X_{s,k}^tX_{s,k}\left(X_k^tX_k\right)^{-1}X_k\varepsilon \\ &= \varepsilon_s^tP_{s,k}\varepsilon_s - 2\varepsilon_s^tX_{s,k}\left(X_k^tX_k\right)^{-1}X_k^t\varepsilon + \lambda\varepsilon^tP_k\varepsilon + o_p(1).\end{aligned}$$

Observe that

$$\left[d\binom{n}{d}\right]^{-1}\sum_s \varepsilon_s^tX_{s,k} = n^{-1}\varepsilon^tX_k.$$

Thus from Lemma 2,

$$\begin{aligned}\left[d\binom{n}{d}\right]^{-1}\sum_s \left\| P_{s,k}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2 &= \left[d\binom{n}{d}\right]^{-1}\sum_s \varepsilon_s^tP_{s,k}\varepsilon_s \\ &\quad -\frac{1}{n}\varepsilon^tP_k\varepsilon+o_p(n^{-1}) \\ &= \frac{1-\lambda}{\lambda}\cdot\frac{k\sigma^2}{n}+o_p(n^{-1}).\end{aligned}$$

□

Regarding MCV_k as a stochastic function of k, it turns out that asymptotically, MCV_k has a rather simple structure which allows us to study in an elegant fashion the properties of the selected model. Our main result of this paper is the following.

THEOREM 1. *Under Assumptions A to D, suppose that* $\mathbf{E}\varepsilon_i = 0$, $\mathbf{E}\varepsilon_i^2 = \sigma^2$, $i = 1, \ldots, n$. *Then*

$$MCV_k = \begin{cases} n^{-1}\varepsilon^t P_k^{\perp}\varepsilon + \dfrac{2-\lambda}{1-\lambda}\cdot\dfrac{k\sigma^2}{n} + o_p(n^{-1}), & k \geq k_0, \\ n^{-1}\varepsilon^t\varepsilon + b_k + o_p(1), & k < k_0. \end{cases}$$

PROOF. From Assumption B, it is easy to verify that

$$H_{s,k} = X_{s,k}\left(X_k^t X_k\right)^{-1} X_{s,k}^t = (\lambda + o(1))P_{s,k}.$$

Thus

$$(I - H_{s,k})^2 = I - \lambda(2-\lambda)P_{s,k} + o(P_{s,k}).$$

Here by an abuse of notation, $o(P_{s,k})$ represents a symmetric matrix Γ such that $\Gamma \leq \gamma_n P_{s,k}$, $\gamma_n \to 0$. Let $\mu = \lambda(2-\lambda)/(1-\lambda)^2$. The previous equation implies that

$$(I - H_{s,k})^{-2} = I + \mu P_{s,k} + o(P_{s,k}). \tag{2.5}$$

Therefore,

$$\begin{aligned} &\left\|(I - H_{s,k})^{-1}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2 \\ &\quad = \left(Y_s - X_{s,k}\hat{\beta}_k\right)^t\left[I + \mu P_{s,k} + o(P_{s,k})\right]\left(Y_s - X_{s,k}\hat{\beta}_k\right) \\ &\quad = \left\|Y_s - X_{s,k}\hat{\beta}_k\right\|^2 + \mu\left\|P_{s,k}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2 \\ &\qquad + o\left(\left\|P_{s,k}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2\right). \end{aligned} \tag{2.6}$$

Substitute this into (2.2). By Lemma 3, for $k \geq k_0$,

$$\mathrm{MCV}_k = \left[d\binom{n}{d}\right]^{-1}\sum_s \left\|Y_s - X_{s,k}\hat{\beta}_k\right\|^2 + \frac{2-\lambda}{1-\lambda}\cdot\frac{k\sigma^2}{n} + o_p(n^{-1}).$$

Moreover, when $k \geq k_0$,

$$\left[d\binom{n}{d}\right]^{-1}\sum_s \|Y_s - X_{s,k}\hat{\beta}_k\|^2 = n^{-1}\|Y - X_k\hat{\beta}_k\|^2 = n^{-1}\varepsilon^t P_k^{\perp}\varepsilon.$$

Consequently, for $k \geq k_0$,

$$\mathrm{MCV}_k = n^{-1}\varepsilon^t P_k^{\perp}\varepsilon + \frac{2-\lambda}{1-\lambda}\cdot\frac{k\sigma^2}{n} + o_p(n^{-1}). \tag{2.7}$$

When $k < k_0$, however, we still have

$$\mathrm{MCV}_k = n^{-1}\|Y - X_k\hat{\beta}_k\|^2 + O\left(\left[d\binom{n}{d}\right]^{-1}\sum_s \left\|P_{s,k}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2\right).$$

For the first term on the right hand side, we have

$$\begin{aligned} n^{-1}\|Y - X_k\hat{\beta}_k\|^2 &= n^{-1}\|P_k^{\perp}f + P_k^{\perp}\varepsilon\|^2 \\ &= n^{-1}\varepsilon^t P_k^{\perp}\varepsilon + n^{-1}f^t P_k^{\perp}f + 2n^{-1}\varepsilon^t P_k^{\perp}f \\ &= n^{-1}\varepsilon^t\varepsilon + n^{-1}f^t P_k^{\perp}f + o_p(1). \end{aligned}$$

For the second term on the right-hand side, since $n^{-1}f^t P_k f \to 0$, an argument similar to that leading to Lemma 3 will show that

$$\sum_s \left\|P_{s,k}\left(Y_s - X_{s,k}\hat{\beta}_k\right)\right\|^2 = o_p(1).$$

The conclusion follows immediately. □

Suppose that $S_1, \ldots, S_K$ is a sequence of random walk. Let $\hat{k} = \arg\min_{k \le K} S_k$. We define

$$p_k = \sum^{*}\left\{\prod_{i=1}^{k}\frac{1}{r_i!}\left(\frac{\alpha_i}{i}\right)^{r_i}\right\} \tag{2.8}$$

and

$$q_k = \sum^{*}\left\{\prod_{i=1}^{k}\frac{1}{r_i!}\left(\frac{1-\alpha_i}{i}\right)^{r_i}\right\}, \tag{2.9}$$

where $\alpha_i = \mathbf{P}(S_i < 0)$ and the sum Σ^* is over all k-tuples $(r_1, \ldots, r_k)$ such that $r_1 + 2r_2 + \cdots + kr_k = k$. From standard random walk theory, we know that $\mathbf{P}(\hat{k} = k) = p_k q_{K-k}$. The following result follows immediately from Theorem 1. A proof can be found in Shibata (1984).

COROLLARY 1. *Suppose that* $\hat{k} = \arg\min_{k \le K} MCV_k$. *Then under the assumptions of Theorem* 1, $\hat{k}$ *converges weakly to a random variable* $\hat{k}_\lambda$ *having the following distribution*:

$$\mathbf{P}\left(\hat{k}_\lambda = k\right) = \begin{cases} p_{k-k_0}q_{K-k}, & k_0 \le k \le K, \\ 0, & \textit{otherwise}, \end{cases}$$

where p_k *and* q_k *are defined by* (2.8) *and* (2.9) *with* $\alpha_i = \mathbf{P}(\chi_i^2 > i(2-\lambda)/(1-\lambda))$.

It is interesting to notice that the asymptotic distribution does not depend on the design matrix or any other features of the underlying true model. In fact, it is totally determined by the value of $K - k_0$, that is, the number of superfluous variables observed.

3. The r-fold cross validation criterion. As defined in (2.3), the r-fold cross validation criterion is aimed at reducing the computation involved in the simple CV method. As a result, the performance of the MCV_k^* is expected to be not as good as that of CV. We present some useful theoretical results in this section.

As usual, regarding MCV_k^* as a stochastic process indexed by $k = 1, \ldots, K$, we have the following result.

THEOREM 2. *Suppose that $r > 1$ is a fixed integer. Under Assumptions A–D, we have*

$$MCV_k^* = \begin{cases} n^{-1}\varepsilon^t\varepsilon + an^{-1}\sum_{i=1}^r \varepsilon_{s_i}^t P_{s_i,k}\varepsilon_{s_i} - bn^{-1}\varepsilon^t P_k\varepsilon + o_p(n^{-1}), & k \geq k_0, \\ n^{-1}\varepsilon^t\varepsilon + b_k + o_p(1), & k < k_0, \end{cases}$$

where $a = (r/(r-1))^2 - 1$ and $b = a + 1$.

PROOF. A slight modification of Lemma 3 shows that for $k \geq k_0$,

$$\frac{1}{n}\sum_{i=1}^r \left\| P_{s_i,k}\left(Y_{s_i} - X_{s_i,k}\hat{\beta}_k\right)\right\|^2 = \frac{1}{n}\sum_{i=1}^r \varepsilon_{s_i}^t P_{s_i,k}\varepsilon_{s_i} - \frac{1}{n}\varepsilon^t P_k\varepsilon + o_p(n^{-1}).$$

By substituting (2.6) into (2.3), we have for $k \geq k_0$,

$$\begin{aligned} \mathrm{MCV}_k^* &= \frac{1}{n}\sum_{i=1}^r \|Y_{s_i} - X_{s_i,k}\hat{\beta}_k\|^2 + \frac{a}{n}\sum_{i=1}^r \varepsilon_{s_i}^t P_{s_i,k}\varepsilon_{s_i} - \frac{a}{n}\varepsilon^t P_k\varepsilon + o_p(n^{-1}) \\ &= n^{-1}\varepsilon^t\varepsilon + an^{-1}\sum_{i=1}^r \varepsilon_{s_i}^t P_{s_i,k}\varepsilon_{s_i} - bn^{-1}\varepsilon^t P_k\varepsilon + o_p(n^{-1}). \end{aligned}$$

The case $k < k_0$ follows the same argument as in the proof of Theorem 1. □

It is easy to relate MCV_k^* to a random walk sequence so that the argument used in the previous section can also be applied here. Specifically, we have the following theorem.

THEOREM 3. *Suppose that $\hat{k} = \arg\min_{k \leq K} MCV_k^*$. Then under the assumptions of Theorem 2, $\hat{k}$ converges weakly to a random variable $\hat{k}_r$ having the following distribution:*

$$\mathbf{P}(\hat{k}_r = k) = \begin{cases} p_{k-k_0}q_{K-k}, & k_0 \leq k \leq K, \\ 0, & \text{otherwise}, \end{cases}$$

where p_k and q_k are defined by (2.8) *and* (2.9) *with* $\alpha_i = \mathbf{P}(F_{i,i(r-1)} > (2r-1)/(r-1))$.

PROOF. As in Theorem 2, we only need to consider the case when $k \geq k_0$. It is clear that minimizing MCV_k^* is equivalent to minimizing $S_k = n\ \mathrm{MCV}_k^* - \varepsilon^t\varepsilon$. Theorem 2 thus implies that for $k \geq k_0$,

$$S_k = a\sum_{i=1}^{r} \varepsilon_{s_i}^t P_{s_i,k}\varepsilon_{s_i} - b\varepsilon^t P_k\varepsilon + o_p(1).$$

Let $\tilde{P}_k = \mathrm{diag}(P_{s_1,k}, \ldots, P_{s_r,k})$. Then

$$S_k = \varepsilon^t\big(a\tilde{P}_k - bP_k\big)\varepsilon + o_p(1).$$

Define $W_1 = a\tilde{P}_1 - bP_1$ and $W_k = a(\tilde{P}_k - \tilde{P}_{k-1}) - b(P_k - P_{k-1})$, $k > 1$. Then the preceding equation can be written as

$$S_k = \sum_{i=1}^{k} \varepsilon^t W_k\varepsilon + o_p(1).$$

It is easy to verify that W_k, $k = 1, \ldots, K$ are perpendicular to each other. Thus S_k is approximately a random walk.

Next, let $Z_i = d^{-1/2}\sum_{j\in s_i}\varepsilon_j$, $Z = (Z_1, \ldots, Z_r)^t$. Then

$$\begin{aligned}
\varepsilon^t W_1\varepsilon &= a\varepsilon^t\tilde{P}_1\varepsilon - b\varepsilon^t P_1\varepsilon \\
&= \frac{a}{d}\left[\left(\sum_{i\in s_1}\varepsilon_i\right)^2 + \cdots + \left(\sum_{i\in s_r}\varepsilon_i\right)^2\right] - \frac{b}{n}\left(\sum_{i}^{n}\varepsilon_i\right)^2 \\
&= aZ^tZ - bZ^tP_*Z \\
&= aZ^tP_*^{\perp}Z - Z^tP_*Z.
\end{aligned}$$

Here we use P_* to denote the r-dimensional projection matrix onto the space spanned by $(1, \ldots, 1)^t$. We have therefore shown that $\varepsilon^t W_i\varepsilon$ can be written as

$$\varepsilon^t W_i\varepsilon = a\xi_i - \eta_i,$$

where ξ_i is independent of η_i and (ξ_i, η_i) are iid with distribution (χ^2_{r-1}, χ^2_1). Let $a = (2r-1)/(r-1)^2$, the conclusion follows from standard random walk theory while noting that

$$\begin{aligned}
\alpha_i &= \mathbf{P}(S_i < 0) \\
&= \mathbf{P}\left(a\sum_{j=1}^{i}\xi_j < \sum_{j=1}^{i}\eta_j + o_p(1)\right) \\
&= \mathbf{P}\big(F_{i,i(r-1)} > (2r-1)/(r-1)\big) + o(1). \qquad \square
\end{aligned}$$

4. The repeated learning-testing criterion. We mentioned earlier that the MCV criterion defined by (1.1) is not a practical method due to the huge amount of computation required. Notice, literally, that $\binom{n}{d}$ regressions are to be carried out when implementing (1.1). Let $d \approx \lambda n$, we have by Stirling's

formula that

$$\binom{n}{d} \approx (2\pi n)^{-1/2}\left[\lambda^{\lambda}(1-\lambda)^{(1-\lambda)}\right]^{-n}.$$

Thus the computational complexity of MCV is exponential.

By means of bootstrap, we can reduce the amount of computation substantially while still perform as well as MCV_k in selecting models. The main result is the following theorem.

THEOREM 4. *In addition to the assumptions of Theorem* 1, *let* $\mathbf{E}\varepsilon_i^4 < \infty$. *If* $N/n^2 \to \infty$, *then*

$$RLT_k = MCV_k + o_p(n^{-1}).$$

PROOF. Define $a(s) = d^{-1}\|Y_s - X_{s,k}\hat{\beta}_{(-s),k}\|^2$. Let $\mathscr{F}_n$ be the σ-field generated by $Y_1, \ldots, Y_n$. Then conditional on $\mathscr{F}_n$, $\mathrm{RLT}_k = N^{-1}\sum_{i=1}^{N} a(s_i^*)$ is the mean of N iid random variables. Thus

$$\mathbf{E}(\mathrm{RLT}_k|\mathscr{F}_n) = \mathbf{E}(a(s_1^*)|\mathscr{F}_n) = \mathrm{MCV}_k$$

and

$$\begin{aligned}\operatorname{var}(\mathrm{RLT}_k|\mathscr{F}_n) &= \frac{1}{N}\operatorname{var}(a(s_1^*)|\mathscr{F}_n)\\ &\le \frac{1}{N}\mathbf{E}(a^2(s_1^*)|\mathscr{F}_n)\\ &= \frac{1}{Nd^2}\cdot\frac{1}{\binom{n}{d}}\sum_s \left\|Y_s - X_{s,k}\hat{\beta}_{(-s),k}\right\|^4.\end{aligned}$$

By (2.6) and Lemma 1, it is easy to show that

$$\begin{aligned}\mathbf{E}\left\|Y_s - X_{s,k}\hat{\beta}_{(-s),k}\right\|^4 &\le O\left(\mathbf{E}\|Y_s - X_{s,k}\hat{\beta}_k\|^4\right)\\ &\le O\left(\mathbf{E}\|Y - X_k\hat{\beta}_k\|^4\right) = O(n^2),\end{aligned}$$

where the last equality used the Assumption C. Consequently, we have

$$\mathbf{E}(\operatorname{var}(\mathrm{RLT}_k|\mathscr{F}_n)) = O(N^{-1}),$$

which further implies that

$$\mathrm{RLT}_k = \mathbf{E}(\mathrm{RLT}_k|\mathscr{F}_n) + O_p(N^{-1/2}) = \mathrm{MCV}_k + o_p(n^{-1}). \qquad \square$$

To conclude, by using this resampling scheme, we can reduce the computational complexity of the MCV method from exponential to just a bit more than second order polynomial. In particular, the result of Corollary 1 holds for RLT_k given the assumptions of Theorem 4.

5. Discussion. Suppose that $\tilde{Y} = f + \tilde{\varepsilon}$ is a new observation from the *true* model. Let PE_k be the conditional prediction error based on model $\mathscr{M}_k$. It is then easy to show that

$$\begin{aligned}\mathrm{PE}_k &= n^{-1}\mathbf{E}\left(\|\tilde{Y} - X_k\hat{\beta}_k\|^2 | \mathscr{F}_n\right)\\ &= n^{-1}\mathbf{E}\left(\|\tilde{\varepsilon} + P_k^{\perp} f - P_k\varepsilon\|^2 | \mathscr{F}_n\right)\\ &= \sigma^2 + n^{-1}f^t P_k^{\perp} f + n^{-1}\varepsilon^t P_k \varepsilon.\end{aligned}$$

Thus asymptotically, the FPE criterion is unbiased in estimating PE_k if and only if $\alpha = 2$. In this sense, the simple CV criterion can be viewed as an unbiased estimator of PE_k. It is obvious from Theorem 1 that the delete-d MCV lacks this property since $(2 - \lambda)/(1 - \lambda) > 2$. Some bias correction methods are suggested by Burman (1989). The bias in estimating prediction error, however, is compensated by the enhancement in the chance of selecting the correct model. This can be seen by noting that $\mathbf{P}(\chi_k^2 < k(2 - \lambda)/(1 - \lambda))$ is an increasing function of λ. Thus by Corollary 1, $\mathbf{P}(\hat{k}_\lambda = k_0)$, or the probability of choosing the correct model, is also an increasing function of λ. When $\lambda \to 0$, the criterion (1.1) becomes equivalent to the CV criterion. In this sense, MCV_k is always better than CV in terms of the chance of overfitting.

A natural question arises how to choose d, the number of observations deleted. It would be desirable if one could provide some guideline for this choice. For instance, one might want to suggest a threshold value which characterizes a significant improvement. Unfortunately, this does not seem to be possible. Let $f(\lambda) = \mathbf{P}(\hat{k}_\lambda = k_0)$ be the probability of choosing the right model. Figure 1 shows the function when $K - k_0 = 1$, 5 and 10, respectively. As we can see, the curves are almost linear except when λ is very high, making it difficult to choose an appropriate λ. For further discussion on this matter, the reader is referred to Zhang (1992).

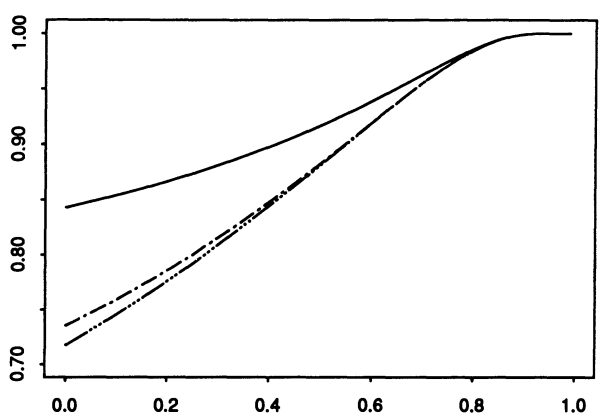

FIG. 1. $f(\lambda) = \mathbf{P}(\hat{k}_\lambda = k_0)$; *the smooth, dotted and dashed curves correspond to the case* $K - k_0 = 1$, 5 *and* 10, *respectively.*

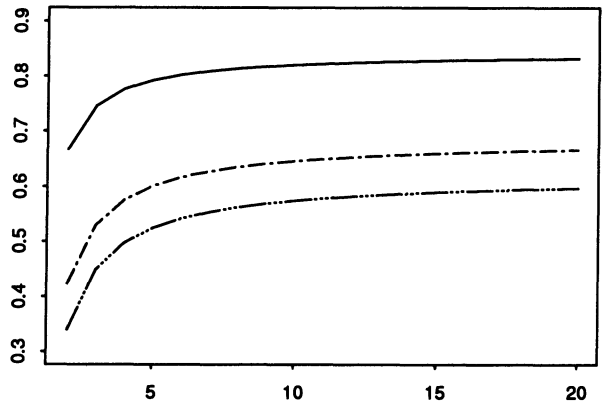

FIG. 2. $g(r) = \mathbf{P}(\hat{k}_r = k_0)$; *the smooth, dotted and dashed curves correspond to the case* $K - k_0 = 1, 5$ *and* 10, *respectively.*

For MCV_k^*, when r decreases, the criterion is actually getting cruder in the sense that less information is being used. The difficulty here is how to divide the sample into r groups. Some approaches have been suggested by Geisser (1975). Assuming that this partition is given a priori, we now consider the impact of different r on the model selected. Intuitively, $\mathbf{P}(\hat{k}_r = k_0)$ should be increasing with r. This has been confirmed by Figure 2. Notice that when $r \to \infty$, $\mathbf{P}(\hat{k}_r = k) \to \mathbf{P}(\hat{k}_{\text{CV}} = k)$. Thus MCV_k^* is always worse than CV in terms of the chance of overfitting. The worst case is $r = 2$. This point is also observed by Burman (1989). It is interesting to notice that deleting half of the data is also the worst strategy for the MCV_k criterion since the computational complexity of MCV_k is a monotone function of $[\lambda^\lambda(1-\lambda)^{(1-\lambda)}]^{-1}$ which achieves its maximum at $\lambda = 1/2$.

Let $g(r) = \mathbf{P}(\hat{k}_r = k_0)$. Figure 2 shows the function when $K - k_0 = 1$, 5 and 10, respectively. Unlike the case with MCV_k, a crude threshold for choosing r is available. It is interesting to notice that the most dramatic improvement occurs between $r = 2$ and $r = 10$. After that, the curves are rather flat. Thus while 5-fold or 10-fold MCV could be beneficial, 20-fold MCV might not be worth the trouble because the intent of MCV_k^* is to reduce computation. This in some sense confirms the observation made by Breiman and Spector (1989).

We close this paper by a simple simulation study. Pseudo observations are generated from the model

$$y_i = 0.6x_{1i} + \varepsilon_i, \qquad i = 1, \ldots, 20, \tag{5.1}$$

where $\varepsilon_i \sim N(0, 0.01)$. Competing models are the null model $y = \varepsilon$, the model (5.1) and a two-dimensional model $y = \alpha x_1 + \beta x_2 + \varepsilon$. Here $x_1, x_2 \in R^{20}$ are iid samples from $N(0, 1)$ and fixed before simulation starts. Random numbers are generated using the IMSL subroutine library.

The results for the MCV* criterion are summarized in Table 1. Notice that underfitting never occurs. As far as the chance of overfitting is concerned,

TABLE 1
Criterion MCV*; 500 *replications*

	Frequencies			Estimated PE
	0	**1**	**2**	**PE**
$r = 2$	0	68.8	31.2	1.080E − 2 ± 1.847E − 4
$r = 5$	0	82.4	17.6	1.046E − 2 ± 1.606E − 4
$r = 10$	0	82.6	17.4	1.057E − 2 ± 1.547E − 4
$r = 20$	0	82.8	17.2	1.058E − 2 ± 1.582E − 4

TABLE 2
Criterion RLT *with* $N = 20$; 500 *replications*

	Frequencies			Estimated PE
	0	**1**	**2**	**PE**
$d = 1$	0	75.6	24.4	1.042E − 2 ± 2.103E − 4
$d = 5$	0	79.6	20.4	1.065E − 2 ± 1.636E − 4
$d = 10$	0	90.0	10.0	1.121E − 2 ± 1.723E − 4
$d = 15$	0	96.6	3.4	1.300E − 2 ± 2.084E − 4

TABLE 3
Criterion RLT *with* $N = 100$; 500 *replications*

	Frequencies			Estimated PE
	0	**1**	**2**	**PE**
$d = 1$	0	79.8	20.2	1.043E − 2 ± 1.704E − 4
$d = 5$	0	84.2	15.8	1.072E − 2 ± 1.551E − 4
$d = 10$	0	91.4	8.6	1.109E − 2 ± 1.642E − 4
$d = 15$	0	98.6	1.4	1.355E − 2 ± 2.108E − 4

5-fold MCV performs as well as 20-fold MCV which by definition is simply the leave-one-out CV criterion. 2-fold MCV is clearly the poorest. By estimated PE, we mean the Monte Carlo average of $\min_k \mathrm{MCV}_k^*$. Since the *true* prediction error equals

$$(1 + k_0/n)\sigma^2 = (1 + 1/20).01 = 1.05 \times 10^{-2},$$

we see that for $r \geq 5$, MCV* estimates the prediction error pretty well. The observation fits very well with the speculation described above.

Tables 2 to 4 contain the results for the RLT criterion with $N = 20$, 100 and 400, respectively. In all cases, deleting 75% of the data leads to a 97% chance of choosing the correct model, much better than the CV criterion does. The gain, however, is at the expense of overestimating the prediction error. The most encouraging observation is that although Theorem 4 requires N to

TABLE 4
Criterion RLT *with* $N = 400$; 500 *replications*

	Frequencies			Estimated PE
	0	**1**	**2**	**PE**
$d = 1$	0	76.4	23.6	1.042E − 2 ± 1.601E − 4
$d = 5$	0	84.2	15.8	1.058E − 2 ± 1.559E − 4
$d = 10$	0	89.6	10.4	1.131E − 2 ± 1.655E − 4
$d = 15$	0	97.8	2.2	1.310E − 2 ± 1.817E − 4

be much larger than 400, no significant differences can be found between $N = 20$ and $N = 400$. When $N = 20$, the computation of RLT is about the same as that of CV. In this respect, it seems that the RLT method could be even more practical than what Theorem 4 asserts.

Acknowledgment. Thanks are due to two referees for making critical and constructive comments. One of them has provided some very useful references.

REFERENCES

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

BREIMAN, L. and SPECTOR, P. (1989). Submodel selection and evaluation in regression: The *X*-random case. Technical Report 197, Dept. Statistics, Univ. California, Berkeley.

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

BURMAN, P. (1990). Estimation of optimal transformations using *v*-fold cross validation and repeated learning-testing methods. *Sankhyā Ser. A* **52** 314–345.

EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461–470.

GEISSER, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320–328.

HERZBERG, A. M. and TSUKANOV, A. V. (1986). A note on modifications of the jackknife criterion for model selection. *Utilitas Math.* **29** 209–216.

SHAO, J. and WU, C. F. J. (1989). A general theory for jackknife variance estimation. *Ann. Statist.* **17** 1176–1197.

SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126.

SHIBATA, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71** 43–49.

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.

ZHANG, P. (1992). On the distributional properties of model selection criteria. *J. Amer. Statist. Assoc.* **87** 732–737.

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
3000 STEINBERG HALL–DIETRICH HALL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302