Model Selection via Multi-fold Cross Validation

By

Ping Zhang

Technical Report No. 257
June 1990

Work partially supported by ONR N00014-89-J-1562.

Department of Statistics
University of California
Berkeley, California

Model Selection via Multi-fold Cross Validation*

Ping Zhang†
Department of Statistics, University of California
Berkeley, CA 94720

Abstract

In model selection, it is known that the simple *leave one out* cross validation method is apt to select overfitted models. In an attempt to remedy this problem, we consider two notions of multi-fold cross validation (MCV and MCV*) criteria. In the case of linear regression models, their performance is studied and compared with the simple CV method. As expected, it turns out that MCV indeed reduces the chance of overfitting. The intent of MCV* is rather different from that of MCV. The differences between these two notions of MCV are also discussed. Our result explains the phenomena observed by Breiman & Spector.

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 non-parametric 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 not limited to, the choice of smoothing parameters in nonparametric smoothing and variable selection in regression. A considerable

*KEY WORDS: Multi-fold cross-validation; Prediction; Model selection

†Work partially supported by ONR N00014-89-J-1562

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.

Despite its gaining popularity, researchers do find that the simple cross validation method suffers from some serious defects. B. Efron [2] pointed out that CV is a poor candidate for estimating the prediction error and suggested that some version of bootstrap would be better off. In Härdle, Hall & Marron [3], they show in the context of bandwidth selection that the cross validation selection tends to undersmooth, namely choosing small bandwidth. Furthermore, the selected bandwidth has extremely large variation (with an asymptotic order of $O(n^{-1/10})$ as opposed to $O(n^{-1/2})$ in conventional problems). In view of model selection, it is well known that the model selected by CV criterion is apt to overfit.

This kind of inconsistency is a common problem shared by many other criteria based on the rule of thumb of minimizing the prediction error. Some consistent criteria have been proposed, but none of them is based on prediction. In this paper, we attempt to rectify this defect by a natural generalization of CV called multi-fold cross validation (MCV) or delete-d cross validation, where $d > 1$ is the number of observations deleted. It is intuitively reasonable to expect that MCV would improve the simple CV. The issue has been raised by several authors. For general variance estimation problem, Shao & Wu [5] introduced multi-fold jackknife and successfully remedied a deficiency of the simple leave-one-out jackknife. Breiman & Spector [1] have considered MCV as a model selection criterion and provided evidence based on a simulation study that MCV does surpass the simple CV. See also Herzberg & Tsukanov [4].

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

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

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t$ 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 non-zero 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

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

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

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

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

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

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

This notion of MCV has an obvious disadvantage, namely that a considerable amount of computation is involved. Other feasible alternatives exist. The one we discuss below is due to Breiman & Spector [1]. Computational efficiency is the main concern here. Suppose that the sample size n can be written as $n = rd$, where r and d are integers. We still consider deleting-d cross validation. However, instead of summing over all possible subsets of size d, 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.

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

Breiman & Spector define their d-fold cross validation as

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

A major difference between MCV_k and MCV_k^* is that the former is intended to refine the simple CV while the latter aims at reducing the amount of computation. Acknowledging this fact, it is thus not surprising that for MCV_k^*, it is a matter of not to lose too much efficiency rather than anticipating any improvement over the simple CV. We will discuss this further in section 4.

Section 2 gives the main results of this paper, namely the asymptotic structures of MCV_k and MCV_k^*. Following in section 3, the asymptotic properties of the selected model order are discussed. It turns out that MCV, although not consistent, indeed reduces the chance of overfitting. However, this is true only when d, the number of observations deleted, is a proportion of the whole sample. Similar results are obtained for MCV^*. More discussions can be found in section 4 where comparisons of the two notions of MCV are made.

2 Basic Results

Let $d = \#\{i : i \in s\}$, $f = X\beta$ and $P_k^\perp = I - P_k$. We introduce the following assumptions:

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

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

C $\underline{\lim}_{n\to\infty} n^{-1} f^t P_k^\perp f = b_k > 0$, and $n^{-1} f^t P_k f \to 0$, $k < k_0$.

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 of 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 MCV with $d/n \to 0$ would be equivalent to the simple CV. For assumption (C), notice that b_k is decreasing and when $k \geq k_0$, $b_k = 0$.

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*

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

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 equation (2.1),

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

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

$$\hat{\beta}_{(-s),k} = (X_{(-s),k}^tX_{(-s),k})^{-1}X_{(-s),k}^tY_s$$

Some algebra will show that

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

Consequently,

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

□

The above lemma shows that

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

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

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

The following two lemmas are the key to our main result. It is essential to assume condition (A), i.e., 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 $\mathbf{E}\epsilon_i = 0$, $\mathbf{E}\epsilon_i^2 = \sigma^2$, $i = 1, \ldots, n$. Then under assumptions (A),(B) and (D),*

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

Proof. By definition and assumption (B), it is easy to see that

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

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,

$$\begin{aligned}
& [d\tbinom{n}{d}]^{-1} \sum_s \epsilon_s^t H_{s,k} \epsilon_s \\
= \; & [d\tbinom{n}{d}]^{-1} \sum_s \epsilon^t \tilde{H}_{s,k} \epsilon \\
= \; & [d\tbinom{n}{d}]^{-1} \cdot \left[\tbinom{n-2}{d-2} \epsilon^t P_k \epsilon + \left(\tbinom{n-1}{d-1} - \tbinom{n-2}{d-2} \right) \epsilon^t \mathrm{diag}(P_k) \epsilon \right] \\
= \; & \frac{\lambda}{n} \left(\epsilon^t P_k \epsilon + \frac{1-\lambda}{\lambda} k \sigma^2 \right) + o_p(n^{-1})
\end{aligned}$$

The last equation above is due to assumption (D), which implies that $\epsilon^t \mathrm{diag}(P_k)\epsilon = k\sigma^2 + o_p(1)$. 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*

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

Proof. Let $Y_s = f_s + \epsilon_s$ and $Y = f + \epsilon$. When $k \geq k_0$, it is easy to verify that $P_{s,k} f_s = X_{s,k}(X_k^t X_k)^{-1} X_k^t f$. Thus

$$\begin{aligned}
& \|P_{s,k}(Y_s - X_{s,k}\hat{\beta}_k)\|^2 \\
= \; & \|P_{s,k}[f_s + \epsilon_s - X_{s,k}(X_k^t X_k)^{-1} X_k^t (f + \epsilon)]\|^2 \\
= \; & \|P_{s,k}\epsilon_s - X_{s,k}(X_k^t X_k)^{-1} X_k^t \epsilon\|^2 \\
= \; & \epsilon_s^t P_{s,k} \epsilon_s - 2\epsilon_s^t X_{s,k}(X_k^t X_k)^{-1} X_k^t + \epsilon^t X_k (X_k^t X_k)^{-1} X_{s,k}^t X_{s,k} (X_k^t X_k)^{-1} X_k \epsilon \\
= \; & \epsilon_s^t P_{s,k} \epsilon_s - 2\epsilon_s^t X_{s,k}(X_k^t X_k)^{-1} X_k^t + \lambda \epsilon^t P_k \epsilon + o_p(1)
\end{aligned}$$

Observe that

$$[d\binom{n}{d}]^{-1}\sum_s \epsilon_s^t X_{s,k} = n^{-1}\epsilon^t X_k$$

Thus, it follows from Lemma 2 that

$$\begin{aligned} & [d\binom{n}{d}]^{-1}\sum_s \|P_{s,k}(Y_s - X_{s,k}\hat{\beta}_k)\|^2 \\ = \;& [d\binom{n}{d}]^{-1}\sum_s \epsilon_s^t P_{s,k}\epsilon_s - \frac{1}{n}\epsilon^t P_k \epsilon + o_p(n^{-1}) \\ = \;& \frac{1-\lambda}{\lambda}\cdot\frac{k\sigma^2}{n} + o_p(n^{-1}) \end{aligned}$$

□

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

Theorem 1 *Under assumptions (A) to (D), we have*

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

Proof. From assumption (B), it is easy to verify that

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

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 above equation implies that

$$(I - H_{s,k})^{-2} = I + \mu P_{s,k} + o(P_{s,k}) \qquad (2.4)$$

Therefore,

$$\begin{aligned} & \|(I - H_{s,k})^{-1}(Y_s - X_{s,k}\hat{\beta}_k)\|^2 \qquad (2.5) \\ = \;& (Y_s - X_{s,k}\hat{\beta}_k)^t[I + \mu P_{s,k} + o(P_{s,k})](Y_s - X_{s,k}\hat{\beta}_k) \\ = \;& \|Y_s - X_{s,k}\hat{\beta}_k\|^2 + \mu\|P_{s,k}(Y_s - X_{s,k}\hat{\beta}_k)\|^2 + o(\|P_{s,k}(Y_s - X_{s,k}\hat{\beta}_k)\|^2) \end{aligned}$$

Substitue this into (2.2). By lemma 3, for $k \geq k_0$,

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

Moreover, when $k \geq k_0$,

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

Consequently, for $k \geq k_0$,

$$\mathrm{MCV}_k = n^{-1}\epsilon^t P_k^{\perp}\epsilon + \frac{2-\lambda}{1-\lambda} \cdot \frac{k\sigma^2}{n} + o_p(n^{-1}) \tag{2.6}$$

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

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

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

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

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

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

The conclusion follows immediately.

□

Remark: When $\lambda \to 0$, it is seen that as expected, the MCV criterion is equivalent to the well known C_p and AIC criteria. In particular, this justifies a long time conception that the cross validation criterion is asymptotically equivalent to C_p and AIC.

As for MCV*, we have the following.

Theorem 2 *Suppose that $r > 1$ is fixed integer. Under assumptions (A) to (D), we have*

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

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

Proof. Since equation (2.5) is still valid, the proof is essentially the same as that of Theorem 1 except that the term

$$\sum_{i=1}^r \epsilon_{s_i}^t P_{s_i,k}\epsilon_{s_i}$$

can not be further reduced, so we leave it in the final expression.

□

3 Limiting Distributions

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

$$p_k = \sum^* \left\{ \prod_{i=1}^k \frac{1}{r_i!} \left(\frac{\alpha_i}{i} \right)^{r_i} \right\} \tag{3.1}$$

and

$$q_k = \sum^* \left\{ \prod_{i=1}^k \frac{1}{r_i!} \left(\frac{1-\alpha_i}{i} \right)^{r_i} \right\} \tag{3.2}$$

where $\alpha_i = \mathbf{P}(S_i > 0)$ and the sum $\sum^*$ is over all k-tuples $(r_1, \ldots, r_k)$ such that $r_1 + 2r_2 +, \ldots, + kr_k = k$. We have

Lemma 4 *Under the above notations, $\mathbf{P}(\hat{k} = k) = p_k q_{K-k}$.*

Proof. See [6].

□

Theorem 3 *Suppose that $\hat{k} = \arg\min_{k \leq K} MCV_k$. Then under the assumptions of Theorem 1, $\hat{k}$ converges weakly to a random variable $\hat{k}_\lambda$ having the following distribution.*

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

where p_k abd q_k are defined by (3.1) and (3.2) with $\alpha_i = \mathbf{P}(\chi_i^2 > i(2-\lambda)/(1-\lambda))$.

Proof. By the result of Theorem 1, it is obvious that the minimum of MCV$_k$ can not be less than k_0. So we only need to consider the case $k \geq k_0$. It is clear that minimizing MCV$_k$ is equivalent to minimizing $S_k = n\text{MCV}_k - \epsilon^t\epsilon$. Furthermore, Theorem 1 implies that for $k \geq k_0$,

$$S_k = \frac{2-\lambda}{1-\lambda}k\sigma^2 - \epsilon^t P_k \epsilon + o_p(1)$$

Define $W_1 = P_1$ and $W_k = P_k - P_{k-1}$, $k > 1$. Then the above equation can be written as

$$S_k = \sum_{i=1}^{k}\left[\frac{2-\lambda}{1-\lambda}\sigma^2 - \epsilon^t W_i \epsilon\right] + o_p(1)$$

It is easy to verify that W_k, $k = 1, \ldots, K$ are perpendicular idenpotent matrices, i.e., $W_i W_j = \delta(i-j)W_i$. Thus S_k is approximately a random walk. The conclusion follows from Lemma 4 by noticing that

$$\begin{aligned}\alpha_i &= \mathbf{P}(S_i > 0)\\ &= \mathbf{P}\left(\frac{\epsilon^t P_i \epsilon}{\sigma^2} < \frac{2-\lambda}{1-\lambda}i + o_p(1)\right)\\ &= \mathbf{P}(\chi_i^2 < i(2-\lambda)/(1-\lambda)) + o(1)\end{aligned}$$

□

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 $K - k_0$, i.e., the number of extra variables.

Theorem 4 *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.*

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

where p_k abd q_k are defined by (3.1) and (3.2) with $\alpha_i = \mathbf{P}(F_{i,i(r-1)} < (2r-1)/(r-1))$.

Proof. As in Theorem 3, we only need to consider the case when $k \geq k_0$. It is clear that minimizing MCV$^*_k$ is equivalent to minimizing $S_k = n\mathrm{MCV}^*_k - \epsilon^t\epsilon$. Theorem 2 thus implies that for $k \geq k_0$,

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

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

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

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 above equation can be written as

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

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

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

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

$$\epsilon^t W_i \epsilon = a\xi_i - \eta_i$$

where ξ_i is independent of η_i and (ξ_i, η_i) are iid with distribution (χ^2_{r-1}, χ^2_1). Therefore, the conclusion follows from Lemma 4 by noting that $a = (2r-1)/(r-1)^2$ and

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

□

4 Discussions

Although the simple cross validation is an unbiased estimate of the prediction error, estimating the prediction error is like estimating the mean of a χ^2 distribution. The heavy tail of the underlying distribution thus causes the selection to be biased. Theorem 1 shows that the delete-d MCV actually overestimates the prediction error, which in turn reduces the bias in model selection. In order to enhance the simple CV, we have to allow d, the number of observations deleted, to be a fixed proportion of the whole sample. The larger the proportion, the less the chance of overfitting. To be precise, we have the following

Theorem 5 *Under assumptions (A) to (D), let $\hat{k}_\lambda$ be as in Theorem 1. Then* $\mathbf{P}(\hat{k}_{\lambda'} \leq \hat{k}_\lambda) = 1$ *for* $\lambda' > \lambda$.

Proof. Without losing generality, we could assume that $\hat{k}_\lambda \geq k_0$. Theorem 3 shows that the selection problem is equivalent to that of minimizing

$$\tilde{S}_{k,\lambda} = \sum_{i=1}^{k} \left[\frac{2-\lambda}{1-\lambda}\sigma^2 - \epsilon^t W_i \epsilon\right]$$

where W_i is as defined before. Now for any $\lambda' > \lambda$,

$$\tilde{S}_{k,\lambda'} = \frac{\lambda' - \lambda}{(1-\lambda')(1-\lambda)} k\sigma^2 + \tilde{S}_{k,\lambda} \tag{4.1}$$

By an abuse of notation, since $\hat{k}_\lambda$ is the minimizer, $\tilde{S}_{k,\lambda} > \tilde{S}_{\hat{k}_\lambda,\lambda}$ for any $k > \hat{k}_\lambda$. Combining with (4.1), this implies that

$$\tilde{S}_{k,\lambda'} > \frac{\lambda' - \lambda}{(1-\lambda')(1-\lambda)} k\sigma^2 + \tilde{S}_{\hat{k}_\lambda,\lambda} > \tilde{S}_{\hat{k}_\lambda,\lambda'}$$

Consequently, any minimizer of $\tilde{S}_{k,\lambda'}$ can not exceed $\hat{k}_\lambda$. The proof is completed.

□

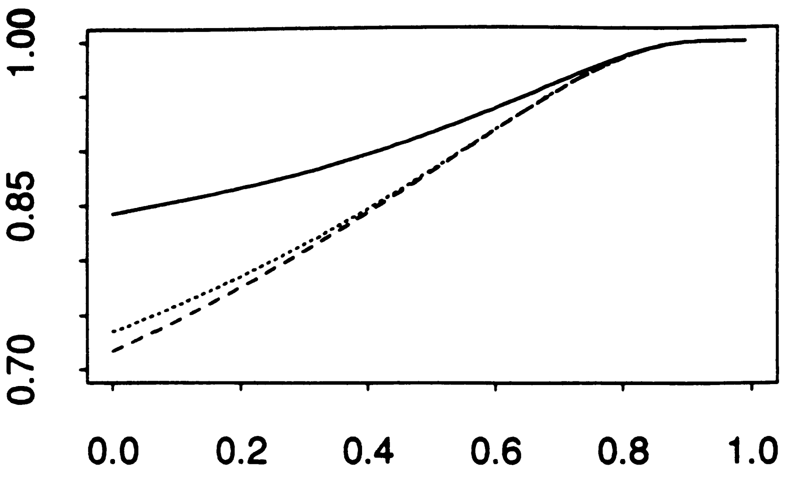

Figure 1: $f(\lambda) = \mathbf{P}(\hat{k}_\lambda = k_0)$; The smooth, dotted and dashed curves correspond to the case $K - k_0 = 1, 5$ and 10 respectively

Question arises how to choose d (or λ), the number (or the proportion) of observations deleted. It would be desirable if one could provide some guideline for this choice, such as suggesting a threshold value which sort of 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. After all, there is no free lunch.

For MCV_k^*, when r decreases, the criterion is actually getting cruder in the sense that less information is being used. Also, it is not clear as to how to divide the sample into r groups. Assuming that this is done a priori, we now consider the impact of different r on the model selected. In the appendix, we provide tables for the distribution of $\hat{k}_r$ when $r = 2, 5, 10, 20$ and ∞ respectively. The last case corresponds to AIC and thus coincides with the table given by Shibata [6]. As we can see, in the extreme case when $r = 2$, the selection can be

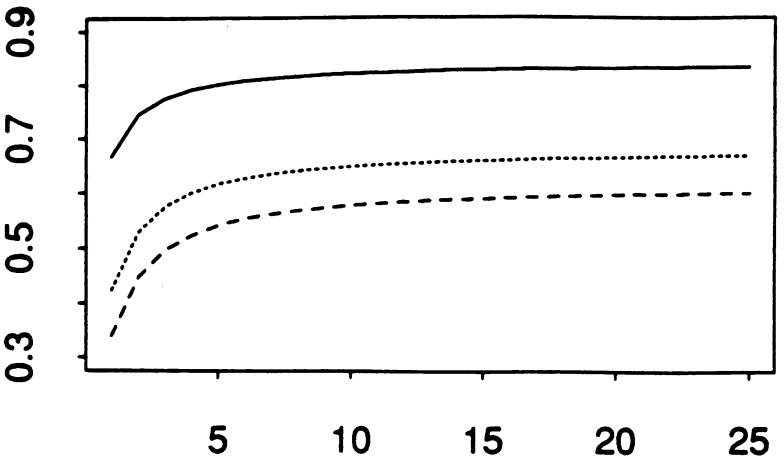

Figure 2: $g(r) = \mathbf{P}(\hat{k}_r = k_0)$; The smooth, dotted and dashed curves correspond to the case $K - k_0 = 1, 5$ and 10 respectively

very poor, especially when $K - k_0$ is large. When $K - k_0 = 10$, the chance of selecting the correct model is only 0.339! Furthermore, the larger the r, the less biased the selection is. The rate of this improvement, however, is affected by the value of $K - k_0$. For $r = 20$, the chance of getting the correct model is 0.8318 when $K - k_0 = 1$ as opposed to 0.5965 when $K - k_0 = 10$. These are about 99% and 83% of the best chance ($r = \infty$) respectively.

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. Remember that the intent of MCV_k^* is to reduce computation. This in some sense confirms the observation made by Breiman & Spector [1].

One final remark: Although MCV$_k$ improves the simple CV, the computation involved can be formidable, since all possible subsets of size d has to be considered. One feasible approach to get around this hassle is through some kind of bootstrap. Specifically, instead of summing over all possible subsets of size d, we could resample without replacement d elements from the observed sample and repeat the procedure many times. It is not clear yet how this will work out, and we plan to report work in this direction separately.

A Tables of Distributions

Note: In the following tables, $K = 11$. The 10 rows represent the probability distributions of $\hat{k}_r$ with k_0 ranges from 1 to 10.

Table I. Asymptotic Distribution of $\hat{k}_r$: $r = 2$

0.339	0.1169	0.0809	0.0657	0.0575	0.0527	0.0502	0.0496	0.0511	0.0568	0.0795
0.000	0.3507	0.1214	0.0844	0.0690	0.0609	0.0565	0.0548	0.0558	0.0614	0.0852
0.000	0.0000	0.3642	0.1267	0.0887	0.0730	0.0652	0.0616	0.0616	0.0669	0.0920
0.000	0.0000	0.0000	0.3800	0.1330	0.0939	0.0782	0.0711	0.0693	0.0740	0.1004
0.000	0.0000	0.0000	0.0000	0.3990	0.1408	0.1006	0.0854	0.0800	0.0832	0.1110
0.000	0.0000	0.0000	0.0000	0.0000	0.4225	0.1509	0.1097	0.0960	0.0960	0.1248
0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4527	0.1646	0.1235	0.1152	0.1440
0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4938	0.1852	0.1481	0.1728
0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5556	0.2222	0.2222
0.000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6667	0.3333

Table II. Asymptotic Distribution of $\hat{k}_r$: $r = 5$

0.5236	0.1112	0.0688	0.0520	0.0430	0.0374	0.0339	0.0317	0.0307	0.0313	0.0364
0.0000	0.5348	0.1139	0.0706	0.0536	0.0445	0.0391	0.0358	0.0341	0.0343	0.0395
0.0000	0.0000	0.5474	0.1169	0.0728	0.0555	0.0464	0.0412	0.0384	0.0381	0.0433
0.0000	0.0000	0.0000	0.5620	0.1205	0.0754	0.0579	0.0490	0.0443	0.0429	0.0480
0.0000	0.0000	0.0000	0.0000	0.5792	0.1248	0.0787	0.0611	0.0526	0.0494	0.0542
0.0000	0.0000	0.0000	0.0000	0.0000	0.6000	0.1302	0.0830	0.0656	0.0587	0.0624
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6261	0.1374	0.0892	0.0732	0.0742
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6604	0.1476	0.0995	0.0925
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7096	0.1647	0.1256
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7920	0.2080

Table III. Asymptotic Distribution of $\hat{k}_r$: $r = 10$

0.5737	0.1053	0.0634	0.0471	0.0385	0.0332	0.0297	0.0275	0.0262	0.0262	0.0293
0.0000	0.5843	0.1074	0.0649	0.0484	0.0397	0.0344	0.0311	0.0292	0.0288	0.0319
0.0000	0.0000	0.5962	0.1099	0.0666	0.0499	0.0411	0.0360	0.0331	0.0321	0.0351
0.0000	0.0000	0.0000	0.6099	0.1128	0.0686	0.0517	0.0431	0.0383	0.0364	0.0392
0.0000	0.0000	0.0000	0.0000	0.6260	0.1163	0.0712	0.0542	0.0458	0.0421	0.0443
0.0000	0.0000	0.0000	0.0000	0.0000	0.6454	0.1206	0.0746	0.0576	0.0504	0.0514
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6696	0.1263	0.0793	0.0633	0.0614
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7011	0.1344	0.0872	0.0773
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7459	0.1477	0.1063
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8198	0.1802

Table IV. Asymptotic Distribution of $\hat{k}_r$: $r = 20$

0.5965	0.1020	0.0607	0.0448	0.0364	0.0312	0.0278	0.0256	0.0243	0.0241	0.0265
0.0000	0.6067	0.1040	0.0620	0.0459	0.0374	0.0323	0.0290	0.0271	0.0265	0.0289
0.0000	0.0000	0.6182	0.1062	0.0636	0.0473	0.0387	0.0337	0.0308	0.0296	0.0319
0.0000	0.0000	0.0000	0.6315	0.1088	0.0654	0.0489	0.0404	0.0357	0.0336	0.0356
0.0000	0.0000	0.0000	0.0000	0.6471	0.1120	0.0677	0.0511	0.0428	0.0390	0.0404
0.0000	0.0000	0.0000	0.0000	0.0000	0.6657	0.1159	0.0706	0.0541	0.0468	0.0469
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6889	0.1210	0.0748	0.0591	0.0562
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7191	0.1281	0.0817	0.0710
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7619	0.1399	0.0982
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8318	0.1682

Table V. Asymptotic Distribution of $\hat{k}_r$: $r = \infty$

0.7172	0.1131	0.0577	0.0351	0.0232	0.0162	0.0117	0.0087	0.0067	0.0054	0.0049
0.0000	0.7188	0.1134	0.0580	0.0353	0.0234	0.0164	0.0120	0.0090	0.0072	0.0064
0.0000	0.0000	0.7210	0.1139	0.0583	0.0356	0.0238	0.0167	0.0124	0.0097	0.0085
0.0000	0.0000	0.0000	0.7241	0.1146	0.0588	0.0361	0.0242	0.0173	0.0132	0.0115
0.0000	0.0000	0.0000	0.0000	0.7285	0.1156	0.0596	0.0369	0.0251	0.0186	0.0157
0.0000	0.0000	0.0000	0.0000	0.0000	0.7349	0.1171	0.0608	0.0382	0.0269	0.0220
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7446	0.1196	0.0630	0.0409	0.0319
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7602	0.1239	0.0674	0.0485
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7874	0.1326	0.0800
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.8427	0.1573

References

[1] BREIMAN, L. & SPECTOR, P. (1989). Submodel selection and evaluation in regression: The X-random case, *Tech.Report. No. 197*, Department of Statistics, University of California at Berkeley.

[2] EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *JASA*, **81**, 461-470.

[3] HÄRDLE, W., HALL, P. & MARRON, J. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *JASA*, **83**, 86-95

[4] HERZBERG, A.M. & TSUKANOV, A.V. (1986), A note on modifications of the jackknife criterion for model selection, *Utilitas Mathematica*, **29**, 209-216.

[5] SHAO, J. & WU, C.F.J. (1989), A general theory for jackknife variance estimation, *Ann. Statist.*, **17**, 1176-1197.

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