Linear Model Selection by Cross-Validation

JUN SHAO*

We consider the problem of selecting a model having the best predictive ability among a class of linear models. The popular leave-one-out cross-validation method, which is asymptotically equivalent to many other model selection methods such as the Akaike information criterion (AIC), the C_p, and the bootstrap, is asymptotically inconsistent in the sense that the probability of selecting the model with the best predictive ability does not converge to 1 as the total number of observations $n \to \infty$. We show that the inconsistency of the leave-one-out cross-validation can be rectified by using a leave-n_v-out cross-validation with n_v, the number of observations reserved for validation, satisfying $n_v/n \to 1$ as $n \to \infty$. This is a somewhat shocking discovery, because $n_v/n \to 1$ is totally opposite to the popular leave-one-out recipe in cross-validation. Motivations, justifications, and discussions of some practical aspects of the use of the leave-n_v-out cross-validation method are provided, and results from a simulation study are presented.

KEY WORDS: Balanced incomplete; Consistency; Data splitting; Model assessment; Monte Carlo; Prediction.

1. INTRODUCTION

Cross-validation is a method for model selection according to the predictive ability of the models. Suppose that n data points are available for selecting a model from a class of models. The data set is split into two parts. The first part contains n_c data points used for fitting a model (model construction), whereas the second part contains $n_v = n - n_c$ data points reserved for assessing the predictive ability of the model (model validation). Strictly speaking, model validation is carried out using not just n_v, but all the $n = n_v + n_c$ data. There are $\binom{n}{n_v}$ different ways to split the data set. Cross-validation, as its name indicates, selects the model with the best average predictive ability calculated based on all (or some) different ways of data splitting.

Clearly, the computational complexity of this method increases as n_v increases. That is why the simplest cross-validation with $n_v \equiv 1$ has been the main focus of researchers' attention over the past 30 years. Discussions and theoretical studies about the cross-validation method with $n_v \equiv 1$ under various situations can be found, for example, in Allen (1974), Stone (1974, 1977a,b), Geisser (1975), Wahba and Wold (1975), Efron (1983, 1986), Picard and Cook (1984), Herzberg and Tsukanov (1986), and Li (1987).

Throughout this article I assume that the number of predictors in each model under consideration does not change as n increases. In this case, it is known to many statisticians (although a rigorous statement has probably not been given in the literature) that the cross-validation with $n_v \equiv 1$ is asymptotically incorrect (inconsistent) and is too conservative in the sense that it tends to select an unnecessarily large model.

There are other methods for model selection, such as the Akaike information criterion (AIC) (Akaike 1974; Shibata 1981), the C_p (Mallows 1973), the jackknife, and the bootstrap (Efron 1983, 1986). All these methods are asymptotically equivalent to the cross-validation with $n_v \equiv 1$ (Stone 1977a; Efron 1983), however, and thus they share the same deficiency; that is, they are inconsistent.

In this article I show that in the problem of selecting linear models, this deficiency of the cross-validation with $n_v \equiv 1$ can be rectified by using a cross-validation with a large n_v (depending on n). Our result is somewhat surprising; to have an asymptotically correct cross-validation procedure, we need to select n_v having the same rate of divergence to infinity as n; that is, $n_v/n \to 1$ as $n \to \infty$. The reason why such a large n_v is needed is explored, after taking a close look at the asymptotic behavior of the cross-validation procedures.

When n_v is large, the amount of computation required to use the cross-validation may be impractical. We consider a "balanced incomplete" cross-validation; that is, only a much smaller part of $\binom{n}{n_v}$ splits are made according to a systematic manner. Two other alternatives—a Monte Carlo approximation and an analytic approximation to the leave-n_v-out cross-validation—are also considered. Their performances are examined in a simulation study.

The issue of using more than one observation at a time in validation against leave-one-out was also raised by other researchers. Herzberg and Tsukanov (1986) did some simulation comparisons between the cross-validation procedures with $n_v \equiv 1$ and $n_v \equiv 2$. They found that the leave-two-out cross-validation is sometimes better than the leave-one-out cross-validation, although the two procedures are asymptotically equivalent in theory. See also Geisser (1975), Burman (1989), and Zhang (1991). In the context of jackknife variance estimation for nonsmooth statistics (such as the sample quantiles), Shao and Wu (1989) showed that the inconsistency of the leave-one-out jackknife variance estimator can be rectified by using a leave-n_v-out jackknife. The difference is that here we require that $n_v/n \to 1$, whereas in Shao and Wu (1989) the rate of n_v diverging to infinity was related to the smoothness of the given statistic.

It should be noted that the story is quite different in the cases where the number of predictors in one of the models under consideration increases as n increases. In such cases, Li (1987) showed that under some conditions, the leave-one-out cross-validation is consistent and is asymptotically optimal in some sense.

2. MODEL SELECTION AND PREDICTION ERROR

Consider a linear model

$$y = \mathbf{x}'\beta + e, \tag{2.1}$$

* Jun Shao is Associate Professor, Department of Mathematics, University of Ottawa, Ottawa K1N 6N5, Canada. The research was supported by Natural Science and Engineering Research Council of Canada. The author thanks the referees for thoughtful comments and helpful suggestions.

Journal of the American Statistical Association
June 1993, Vol. 88, No. 422, Theory and Methods

where y is a response variable, $\mathbf{x}$ is a p vector of covariates (predictors), $\mathbf{x}'$ denotes the transpose of $\mathbf{x}$, β is a p vector of unknown parameters, and e is a random error with mean 0 and variance σ^2. Because some of the components of β may be 0, a more compact model might be

$$y = \mathbf{x}'_\alpha\beta_\alpha + e, \tag{2.2}$$

where α is a subset of d_α distinct positive integers that are less or equal to p and β_α (or $\mathbf{x}_\alpha$) is the d_α vector containing the components of β (or $\mathbf{x}$) that are indexed by the integers in α. There are $2^p - 1$ possible different models of the form (2.2), each of which corresponds to a subset α and is denoted by $\mathcal{M}_\alpha$. The dimension (or size) of $\mathcal{M}_\alpha$ is defined to be d_α, the number of predictors in $\mathcal{M}_\alpha$. Let $\mathcal{A}$ denote all nonempty subsets of $\{1, \ldots, p\}$. If we know whether each component of β is 0 or not, then the models $\mathcal{M}_\alpha$ can be classified into two categories:

- Category I: At least one nonzero component of β is not in β_α.
- Category II: β_α contains all nonzero components of β.

Clearly, the models in Category I are incorrect models, and the models in Category II may be inefficient because of their unnecessarily large sizes. The optimal model, denoted by $\mathcal{M}_*$, is the model in Category II with the smallest dimension. Note that model selection under this framework is the same as variable (predictor) selection. Selecting a model from Category I means missing at least one important predictor, whereas selecting the most compact model from Category II means eliminating all the variables that are unrelated to the response variable.

In statistical analysis, model selection is carried out by using data pairs $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, satisfying

$$y_i = \mathbf{x}'_i\beta + e_i,$$

with iid errors $e_1, \ldots, e_n$. Under model $\mathcal{M}_\alpha$, the least squares estimator of β_α is

$$\hat{\beta}_\alpha = (\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}'_\alpha\mathbf{y},$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$ is an $n \times 1$ response vector, $\mathbf{X}_\alpha = (\mathbf{x}_{1\alpha}, \ldots, \mathbf{x}_{n\alpha})'$ is an $n \times d_\alpha$ matrix assumed of full rank for any $\alpha \in \mathcal{A}$, and $\mathbf{x}_{i\alpha}$ is the d_α vector containing the components of $\mathbf{x}_i$ that are indexed by the integers in α. Denote $\mathbf{X}_\alpha$ with $\alpha = \{1, \ldots, p\}$ by $\mathbf{X}$.

We mainly consider the case of deterministic predictors. When $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are random, the results are still valid almost surely for given sequences $\mathbf{x}_1, \mathbf{x}_2, \ldots$, provided that (a) for given $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $e_1, \ldots, e_n$ are iid with mean 0 and variance σ^2; and (b) all the conditions on $\mathbf{x}_1, \ldots, \mathbf{x}_n$, stated in Theorems 1 and 2 in Section 3, hold almost surely for given sequences $\mathbf{x}_1, \mathbf{x}_2, \ldots$.

Suppose that z_i is the future value of the response variable to be predicted when the prediction variable is equal to $\mathbf{x}_i$. Using model $\mathcal{M}_\alpha$ fitted based on the data $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, the average squared prediction error is

$$\frac{1}{n}\sum_i (z_i - \mathbf{x}'_{i\alpha}\hat{\beta}_\alpha)^2.$$

Given $\mathbf{y}$, the conditional expected squared prediction error is

$$\sigma^2 + \frac{1}{n}\sum_i (\mathbf{x}'_i\beta - \mathbf{x}'_{i\alpha}\hat{\beta}_\alpha)^2.$$

The overall unconditional expected squared prediction error (conditional on $\mathbf{x}_1, \ldots, \mathbf{x}_n$ for random predictor case) is

$$\Gamma_{\alpha,n} = \sigma^2 + n^{-1}d_\alpha\sigma^2 + \Delta_{\alpha,n}, \tag{2.3}$$

where

$$\Delta_{\alpha,n} = n^{-1}\beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_\alpha)\mathbf{X}\beta,$$
$$\mathbf{P}_\alpha = \mathbf{X}_\alpha(\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}'_\alpha$$

is the projection matrix under model $\mathcal{M}_\alpha$ and $\mathbf{I}_k$ is the identity matrix of order k. Note that $\Gamma_{\alpha,n}$ consists of two components: the variability of the future observations and $n^{-1}d_\alpha\sigma^2 + \Delta_{\alpha,n}$, which reflects the error in model selection and estimation.

When $\mathcal{M}_\alpha$ is in Category II, $\mathbf{X}\beta = \mathbf{X}_\alpha\beta_\alpha$ and hence

$$\Gamma_{\alpha,n} = \sigma^2 + n^{-1}d_\alpha\sigma^2. \tag{2.4}$$

Because $\mathbf{P}_\alpha$ is the projection matrix of a submatrix $\mathbf{X}_\alpha$ of $\mathbf{X}$, $\Delta_{\alpha,n} > 0$ for any fixed n if $\mathcal{M}_\alpha$ is in Category I.

Assuming that p is the same for every n and that n is large, some asymptotic results can be established under the following condition:

$$\liminf_{n\to\infty} \Delta_{\alpha,n} > 0 \quad \text{for} \quad \mathcal{M}_\alpha \text{ in Category I.} \tag{2.5}$$

For any $\mathcal{M}_\alpha$ in Category I and $\mathcal{M}_\gamma$ in Category II with $d_\alpha = d_\gamma$, the ratio $\Gamma_{\alpha,n}/\Gamma_{\gamma,n}$ may be arbitrarily close to 1, although $\Gamma_{\alpha,n}/\Gamma_{\gamma,n} > 1$ for all n. If $\lim_{n\to\infty}\Gamma_{\alpha,n}/\Gamma_{\gamma,n} = 1$, then, asymptotically, models $\mathcal{M}_\alpha$ and $\mathcal{M}_\gamma$ have no difference in terms of their predictive ability. Because $\liminf_{n\to\infty}\Gamma_{\alpha,n}/\Gamma_{\gamma,n} > 1$ if and only if $\liminf_{n\to\infty}\Delta_{\alpha,n} > 0$, (2.5) is a type of asymptotic model identifiability condition and is very minimal for asymptotic analysis.

3. THE CROSS-VALIDATION METHOD: MOTIVATION AND THEORY

Similar to other model selection methods, the cross-validation method selects a model by minimizing estimated $\Gamma_{\alpha,n}$ over all α. Suppose that we split the data set into two parts: $\{(y_i, \mathbf{x}_i), i \in s\}$ and $\{(y_i, \mathbf{x}_i), i \in s^c\}$, where s is a subset of $\{1, \ldots, n\}$ containing n_v integers and s^c is its complement containing n_c integers, $n_v + n_c = n$. The model $\mathcal{M}_\alpha$ is fitted using the construction data $\{(y_i, \mathbf{x}_i), i \in s^c\}$ and the prediction error is assessed using the validation data $\{(y_i, \mathbf{x}_i), i \in s\}$, treated as if they were future values. The average squared prediction error is

$$n_v^{-1}\|\mathbf{y}_s - \hat{\mathbf{y}}_{\alpha,s^c}\|^2 = n_v^{-1}\|(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}(\mathbf{y}_s - \mathbf{X}_{\alpha,s}\hat{\beta}_\alpha)\|^2, \tag{3.1}$$

where $\|\mathbf{a}\| = (\mathbf{a}'\mathbf{a})^{1/2}$ for a vector $\mathbf{a}$; $\mathbf{y}_s$ is the n_v vector containing the components of $\mathbf{y}$ indexed by $i \in s$; $\mathbf{X}_{\alpha,s}$ is the $n_v \times d_\alpha$ matrix containing the rows of $\mathbf{X}_\alpha$ indexed by $i \in s$; $\hat{\mathbf{y}}_{\alpha,s^c}$ is the prediction of $\mathbf{y}_s$ using the construction data and the least squares method under model $\mathcal{M}_\alpha$, $\mathbf{Q}_{\alpha,s} = \mathbf{X}_{\alpha,s}(\mathbf{X}'_\alpha\mathbf{X}_\alpha)^{-1}\mathbf{X}'_{\alpha,s}$; $\hat{\beta}_\alpha$ is the least squares estimator of β_α

using all n observations; and the equality follows from a straightforward matrix algebra.

There are $\binom{n}{n_v}$ different subsets s of size n_v. For each model $\mathcal{M}_\alpha$, the cross-validation estimate of $\Gamma_{\alpha,n}$ is obtained by averaging the quantities in (3.1) over all or some different subsets s of size n_v. The model selected by cross-validation is then the model that minimizes the cross-validation estimates over all $\alpha \in \mathcal{A}$. I shall call this method the leave-n_v-out cross-validation, abbreviated as CV(n_v). The error rate of using the CV(n_v) for selecting the optimal model $\mathcal{M}_*$ is

$$P(\text{the selected model is not } \mathcal{M}_*). \tag{3.2}$$

3.1 The CV(1) Method

From the computational point of view, the simplest CV(n_v) is the one with $n_v \equiv 1$; that is, the CV(1). Letting $s = \{i\}$ and averaging the squared prediction errors over all i, I conclude from (3.1) that the CV(1) estimate of $\Gamma_{\alpha,n}$ is

$$\hat{\Gamma}^{\mathrm{CV}}_{\alpha,n} = \frac{1}{n}\sum_i [(1 - w_{i\alpha})^{-1}(y_i - \mathbf{x}'_{i\alpha}\hat{\boldsymbol{\beta}}_\alpha)]^2,$$

where $w_{i\alpha}$ is the ith diagonal element of the projection matrix $\mathbf{P}_\alpha$. Under the conditions

$$\mathbf{X}'\mathbf{X} = O(n) \quad \text{and} \quad (\mathbf{X}'\mathbf{X})^{-1} = O(n^{-1}), \tag{3.3}$$

and

$$\lim_{n\to\infty} \max_{i\le n} w_{i\alpha} = 0 \quad \text{for any} \quad \alpha \in \mathcal{A}, \tag{3.4}$$

it is shown in the Appendix that if $\mathcal{M}_\alpha$ is in Category I, then

$$\hat{\Gamma}^{\mathrm{CV}}_{\alpha,n} = \Gamma_{\alpha,n} + o_p(1), \tag{3.5}$$

and if $\mathcal{M}_\alpha$ is in Category II, then

$$\hat{\Gamma}^{\mathrm{CV}}_{\alpha,n} = n^{-1}\mathbf{e}'\mathbf{e} + 2n^{-1}d_\alpha\sigma^2 - n^{-1}\mathbf{e}'\mathbf{P}_\alpha\mathbf{e} + o_p(n^{-1}), \tag{3.6}$$

where $\mathbf{e} = (e_1, \ldots, e_n)'$. Because $n^{-1}\mathbf{e}'\mathbf{e}$ converges to σ^2 almost surely, $\hat{\Gamma}^{\mathrm{CV}}_{\alpha,n}$ is consistent for $\Gamma_{\alpha,n}$. But this does not ensure that the error rate given in (3.2) vanishes as $n \to \infty$. As pointed out by Stone (1977b), this type of consistency is not of great interest, because if $\mathcal{M}_\alpha$ is in Category II, then $\Gamma_{\alpha,n} \to \sigma^2$, which is independent of α. In fact, when (2.5), (3.3), and (3.4) hold, the model selected by using CV(1), denoted by $\mathcal{M}_{\mathrm{CV}}$, satisfies

$$\lim_{n\to\infty} P(\mathcal{M}_{\mathrm{CV}} \text{ is in Category I}) = 0. \tag{3.7}$$

But if $\mathcal{M}_*$ is not of size p, then

$$\lim_{n\to\infty} P(\mathcal{M}_{\mathrm{CV}} = \mathcal{M}_*) \neq 1. \tag{3.8}$$

If $\mathcal{M}_\alpha$ is in Category II but $\mathcal{M}_\alpha \neq \mathcal{M}_*$, then, by (3.6),

$$P(\mathcal{M}_\alpha \text{ is preferable to } \mathcal{M}_* \text{ by the CV(1)})$$
$$= P(2(d_\alpha - d_{\alpha_*})\sigma^2 < \mathbf{e}'(\mathbf{P}_\alpha - \mathbf{P}_{\alpha_*})\mathbf{e}) + o(1), \tag{3.9}$$

where α_* is the subset corresponding to $\mathcal{M}_*$ and $d_\alpha > d_{\alpha_*}$. If $\mathbf{e}$ is distributed as $N(0, \sigma^2\mathbf{I}_n)$, then the probability in (3.9) equals

$$P(2k < \chi^2(k)) + o(1),$$

where $k = d_\alpha - d_{\alpha_*}$ and $\chi^2(k)$ is the chi-square random variable with k degrees of freedom. Clearly, $P(2k < \chi^2(k)) \neq 0$ for any $k \geq 1$.

In view of (3.7) and (3.8), the CV(1) is asymptotically incorrect and is too conservative in the sense that it may select a model of excessive size, unless the optimal model is the one with size p.

Let me explain why the CV(1) is asymptotically incorrect. From (2.4), the difference between two models in Category II appears in the second-order term $n^{-1}d_\alpha\sigma^2$, a term of order n^{-1}. From (3.6), the term in $\hat{\Gamma}^{\mathrm{CV}}_{\alpha,n}$ affected by the model difference is

$$n^{-1}d_\alpha\sigma^2 + \delta_{\alpha,n},$$

where

$$\delta_{\alpha,n} = n^{-1}d_\alpha\sigma^2 - n^{-1}\mathbf{e}'\mathbf{P}_\alpha\mathbf{e} \tag{3.10}$$

is the error in assessing the differences of the models in Category II by using CV(1). Note that the error $\delta_{\alpha,n}$ has mean 0 but has the same order of magnitude as $n^{-1}d_\alpha\sigma^2$. Hence the CV(1) fails to distinguish the models in Category II. The story is different for the models in Category I. From (2.3), the term in $\Gamma_{\alpha,n}$ that distinguishes the models is $\Delta_{\alpha,n}$, a term that does not vanish as $n \to \infty$. From (3.5), the error in assessing the models in Category I by CV(1) is $\hat{\Gamma}^{\mathrm{CV}}_{\alpha,n} - \Gamma_{\alpha,n} = o_p(1)$, a term of lower order than $\Delta_{\alpha,n}$, and hence the result (3.7) holds.

3.2 The Balanced Incomplete CV(n_v) Method

In this section I show and explain why the deficiency of CV(1) can be rectified by the CV(n_v) with a large n_v. It is impractical and also unnecessary to carry out the validation for all different splits when $n_v > 1$. Let $\mathcal{B}$ be a collection of b subsets of $\{1, \ldots, n\}$ that have size n_v. $\mathcal{B}$ is selected according to the following "balance" conditions: (a) every i, $1 \leq i \leq n$, appears in the same number of subsets in $\mathcal{B}$; and (b) every pair (i, j), $1 \leq i < j \leq n$, appears in the same number of subsets in $\mathcal{B}$.

The cross-validation estimate of $\Gamma_{\alpha,n}$ is then obtained by averaging the quantities in (3.1) over the subsets $s \in \mathcal{B}$. This method will be called the balanced incomplete CV(n_v), abbreviated as BICV(n_v), because $\mathcal{B}$ is in fact a balanced incomplete block design (BIBD) if each subset is treated as a "block" and each i as a "treatment." Examples of BIBD can be found in John (1971). The repetition size $b \geq n$ is usually a linear function of n; that is, $b = O(n)$. The BICV(n_v) selects a model by minimizing

$$\hat{\Gamma}^{\mathrm{BICV}}_{\alpha,n} = \frac{1}{n_v b}\sum_{s\in\mathcal{B}} \|\mathbf{y}_s - \hat{\mathbf{y}}_{\alpha,s^c}\|^2$$

over all $\alpha \in \mathcal{A}$.

The following result shows that the BICV(n_v) is asymptotically correct if $n_c \to \infty$ and $n_v/n \to 1$.

Theorem 1. Suppose that (2.5), (3.3), and (3.4) hold and

$$\lim_{n\to\infty} \max_{s\in\mathcal{B}} \left\| \frac{1}{n_v}\sum_{i\in s} \mathbf{x}_i\mathbf{x}'_i - \frac{1}{n_c}\sum_{i\in s^c} \mathbf{x}_i\mathbf{x}'_i \right\| = 0. \tag{3.11}$$

Suppose also that n_v is selected so that

$$n_v/n \to 1 \quad \text{and} \quad n_c = n - n_v \to \infty. \tag{3.12}$$

Then we have the following conclusions:

(a) If $\mathcal{M}_\alpha$ is in Category I, then there exists $R_n \geq 0$ such that

$$\hat{\Gamma}_{\alpha,n}^{\text{BICV}} = n^{-1}\mathbf{e}'\mathbf{e} + \Delta_{\alpha,n} + o_p(1) + R_n. \tag{3.13}$$

(b) If $\mathcal{M}_\alpha$ is in Category II, then

$$\hat{\Gamma}_{\alpha,n}^{\text{BICV}} = n^{-1}\mathbf{e}'\mathbf{e} + n_c^{-1}d_\alpha\sigma^2 + o_p(n_c^{-1}). \tag{3.14}$$

(c) Consequently,

$$\lim_{n\to\infty} P(\text{the selected model is } \mathcal{M}_*) = 1. \tag{3.15}$$

The proof of Theorem 1 is given in the Appendix. The following is an explanation of why the BICV(n_v) improves over the CV(1) and why n_v should be chosen according to (3.12).

Efron (1986) pointed out that the CV(1) estimates the expected squared prediction error based on a sample of size $n-1$, rather than n; that is, $\hat{\Gamma}_{\alpha,n}^{\text{CV}}$ estimates $\Gamma_{\alpha,n-1}$, not $\Gamma_{\alpha,n}$. This has been neglected by most researchers, because the difference between $\Gamma_{\alpha,n-1}$ and $\Gamma_{\alpha,n}$ is asymptotically inappreciable. The difference between Γ_{α,n_c} and $\Gamma_{\alpha,n}$ is not negligible if and only if n_c/n does not tend to 1. Recall that the cross-validation selects a model in two steps: (1) fitting a model using n_c data, not n data, and (2) validating the fitted model using n_v data. Naturally, the BICV(n_v) estimates Γ_{α,n_c}. This is also justified by (3.14).

We do not necessarily need a very accurate model fitting in step (1) of the cross-validation, but we do need an accurate assessment of the prediction error in step (2), because the overall purpose of cross-validation is to select a model and the selected model will then be refitted using the full data set for the prediction purpose. From this standpoint, we do not need to use an n_c close to n in step (1). It is actually wise to use a relatively small n_c, because for the models in Category II,

$$\Gamma_{\alpha,n_c} = \sigma^2 + n_c^{-1}d_\alpha\sigma^2$$

is flat (as a function of α) if n_c is large. Therefore, it is difficult to find the minimum of Γ_{α,n_c} with a large n_c, using a small n_v for validation. Using the CV(1) method can be compared to using a telescope to see some objects 10,000 meters away, whereas using the BICV(n_v) method is more like using the same telescope to see the same objects only 100 meters away. Of course, the latter method can see the differences among these objects more clearly. But $n_c \to \infty$ is still needed to ensure the consistency of the model fitting in step (1).

The previous argument shows heuristically the necessity of having a large n_v and a relatively small n_c. Hence the result in Theorem 1 is not so surprising as it seems at first glance. But why is $n_v/n \to 1$ needed?

By considering the following special case, I show that if n_v/n does not converge to 1, the same problem occurs as when CV(1) is used.

Suppose that a particular $\mathcal{B}$ can be selected such that

$$\frac{1}{n_v}\sum_{i\in s}\mathbf{x}_i\mathbf{x}_i' = \frac{1}{n_c}\sum_{i\in s^c}\mathbf{x}_i\mathbf{x}_i' \quad \text{for all} \quad s \in \mathcal{B}. \tag{3.16}$$

An example is when

$$\mathbf{X} = \text{block diagonal } (\mathbf{1}, \ldots, \mathbf{1}),$$

where $\mathbf{1}$ is an m vector of 1s. Under (3.16), it can be shown, by some algebraic calculations, that

$$\begin{aligned}\hat{\Gamma}_{\alpha,n}^{\text{BICV}} &= \frac{1}{n_v b}\sum_{s\in\mathcal{B}}\|\mathbf{y}_s - \mathbf{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha\|^2 \\ &\quad + \frac{n+n_c}{n_c^2 b}\sum_{s\in\mathcal{B}}\left\|\frac{n}{n_v}\mathbf{Q}_{\alpha,s}(\mathbf{y}_s - \mathbf{X}_{\alpha,s}\hat{\boldsymbol{\beta}}_\alpha)\right\|^2 \\ &= \frac{1}{n}\|\mathbf{y} - \mathbf{X}_\alpha\hat{\boldsymbol{\beta}}_\alpha\|^2 \\ &\quad + \frac{n+n_c}{n_c(n-1)}\sum_i w_{i\alpha}(y_i - \mathbf{x}_{i\alpha}'\hat{\boldsymbol{\beta}}_\alpha)^2. \end{aligned}\tag{3.17}$$

If $\mathcal{M}_\alpha$ is in Category II, then, by (3.17),

$$\begin{aligned}\hat{\Gamma}_{\alpha,n}^{\text{BICV}} &= n^{-1}\mathbf{e}'\mathbf{e} - n^{-1}\mathbf{e}'\mathbf{P}_\alpha\mathbf{e} \\ &\quad + n_c^{-1}(n-1)^{-1}(n+n_c)[d_\alpha\sigma^2 + o_p(1)] \\ &= n^{-1}\mathbf{e}'\mathbf{e} + n_c^{-1}d_\alpha\sigma^2 + \varepsilon_{\alpha,n},\end{aligned}$$

where

$$\begin{aligned}\varepsilon_{\alpha,n} &= n_c^{-1}(n-1)^{-1}(1+n_c)d_\alpha\sigma^2 \\ &\quad - n^{-1}\mathbf{e}'\mathbf{P}_\alpha\mathbf{e} + o_p(n_c^{-1}).\end{aligned}\tag{3.18}$$

Similar to the $\delta_{\alpha,n}$ in (3.10), the $\varepsilon_{\alpha,n}$ in (3.18) is the error in assessing the model difference by using the BICV(n_v), because the term in Γ_{α,n_c} that distinguishes the models in Category II is $n_c^{-1}d_\alpha\sigma^2$. If n_v/n does not converge to 1, then $\varepsilon_{\alpha,n}$ has the same order of magnitude as $n_c^{-1}d_\alpha\sigma^2$, because

$$\varepsilon_{\alpha,n}/(n_c^{-1}d_\alpha\sigma^2) = \frac{n_c}{n}[1 - (d_\alpha\sigma^2)^{-1}\mathbf{e}'\mathbf{P}_\alpha\mathbf{e}] + o_p(1). \tag{3.19}$$

Therefore, like the CV(1), the BICV(n_v) with n_v/n not converging to 1 cannot distinguish the models in Category II and thus is inconsistent. From (3.19), the only situation where $\varepsilon_{\alpha,n}$ is of a lower order than $n_c^{-1}d_\alpha\sigma^2$ is when $n_c/n \to 0$; that is, $n_v/n \to 1$.

This example shows that the condition $n_v/n \to 1$ is necessary for the consistency of the BICV(n_v).

3.3 Other CV(n_v) Methods

Using the BICV(n_v) requires a "balanced" collection $\mathcal{B}$ of subsets. If such a $\mathcal{B}$ is not available or is hard to obtain, the following two alternatives may be used.

3.3.1 A Monte Carlo CV(n_v). A simple and easy method is to use Monte Carlo. Randomly draw (with or without replacement) a collection $\mathcal{R}$ of b subsets of $\{1, \ldots, n\}$ that have size n_v and select a model by minimizing

$$\hat{\Gamma}_{\alpha,n}^{\text{MCCV}} = \frac{1}{n_v b}\sum_{s\in\mathcal{R}}\|\mathbf{y}_s - \hat{\mathbf{y}}_{\alpha,s^c}\|^2. \tag{3.20}$$

This method will be called the Monte Carlo CV(n_v), abbreviated as MCCV(n_v). The Monte Carlo cross-validation was also considered in Picard and Cook (1984), because (3.20) is obtained by randomly splitting the data b times and averaging the squared prediction errors over the splits.

This yields the following result, which is similar to Theorem 1. The proof is given in the Appendix. The probability statements in Theorem 2 are with respect to the joint probability corresponding to y and the Monte Carlo selection of the subsets.

Theorem 2. Suppose that (2.5), (3.3), (3.4), and (3.12) hold and

$$\max_{s\in\mathcal{R}} \left\| \frac{1}{n_v}\sum_{i\in s}\mathbf{x}_i\mathbf{x}_i' - \frac{1}{n_c}\sum_{i\in s^c}\mathbf{x}_i\mathbf{x}_i' \right\| = o_p(1), \quad (3.21)$$

where $\mathcal{R}$ contains b subsets selected randomly with b satisfying

$$b^{-1}n_c^{-2}n^2 \to 0. \quad (3.22)$$

This yields the following conclusions:

(a) If $\mathcal{M}_\alpha$ is in Category I, then there exists $R_n \geq 0$ such that

$$\hat{\Gamma}_{\alpha,n}^{\mathrm{MCCV}} = \frac{1}{n_v b}\sum_{s\in\mathcal{R}}\mathbf{e}_s'\mathbf{e}_s + \Delta_{\alpha,n} + o_p(1) + R_n, \quad (3.23)$$

where $\mathbf{e}_s = \mathbf{y}_s - \mathbf{X}_s\beta$.

(b) If $\mathcal{M}_\alpha$ is in Category II, then

$$\hat{\Gamma}_{\alpha,n}^{\mathrm{MCCV}} = \frac{1}{n_v b}\sum_{s\in\mathcal{R}}\mathbf{e}_s'\mathbf{e}_s + n_c^{-1}d_\alpha\sigma^2 + o_p(n_c^{-1}). \quad (3.24)$$

(c) Consequently, (3.15) holds.

Condition (3.22) imposes some restrictions on b and n_c. The fewer data used in model construction, the more splits are needed. (3.12) and (3.22) imply that $b \to \infty$ as $n \to \infty$. If n_c is selected such that $n_c^{-2}n \to 0$, then $b \geq n$ is enough for (3.22).

3.3.2 An analytic approximate CV(n_v). Another alternative to the BICV(n_v) is the leading term in $\hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}}$:

$$\hat{\Gamma}_{\alpha,n}^{\mathrm{APCV}} = \frac{1}{n}\|\mathbf{y} - \mathbf{X}_\alpha\hat{\beta}_\alpha\|^2 + \frac{n+n_c}{n_c(n-1)}\sum_i w_{i\alpha}(y_i - \mathbf{x}_{i\alpha}'\hat{\beta}_\alpha)^2. \quad (3.25)$$

This method will be called the approximate CV(n_v), abbreviated as APCV(n_v). From (3.17), $\hat{\Gamma}_{\alpha,n}^{\mathrm{APCV}} = \hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}}$ exactly in the special case where (3.16) holds. Under (2.5), (3.3), (3.4), and (3.12), results (3.13)–(3.15) hold with $\hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}}$ replaced by $\hat{\Gamma}_{\alpha,n}^{\mathrm{APCV}}$. In fact, from the proof of Theorem 1 in the Appendix, (3.13) holds for $\hat{\Gamma}_{\alpha,n}^{\mathrm{APCV}}$, with R_n being the second term on the right side of (3.25); and when $\mathcal{M}_\alpha$ is in Category II,

$$\hat{\Gamma}_{\alpha,n}^{\mathrm{APCV}} = \frac{1}{n}\mathbf{e}'(\mathbf{I}_n - \mathbf{P}_\alpha)\mathbf{e} + \frac{n+n_c}{n_c(n-1)}[d_\alpha\sigma^2 + o_p(1)].$$

The APCV(n_v) is consistent and requires less computation than does either the BICV(n_v) or the MCCV(n_v). But unlike the BICV(n_v) and the MCCV(n_v), the APCV(n_v) depends on the special nature of the linear models, and its extension to other models is not straightforward. Also, from the simulation study in Section 5, it seems that the performance of the APCV(n_v) is not as good as that of the MCCV(n_v), which indicates that to have a good performance, the APCV(n_v) requires a larger n than the MCCV(n_v).

4. FURTHER DISCUSSIONS

4.1 Model Selection From a Given Class

In the previous sections I considered the selection of a model from all possible models

$$\mathcal{C} = \{\mathcal{M}_\alpha, \alpha \in \mathcal{A}\}.$$

$\mathcal{C}$ is a very large class if p is large. From computational and other practical considerations, sometimes we may restrict our attention to a smaller class of models $\mathcal{C}_1 \subset \mathcal{C}$. For example, $\mathcal{C}_1$ may contain only two models. It is clear that the CV(n_v)—for example, BICV(n_v), MCCV(n_v) and APCV(n_v)—can be used to select the best model within $\mathcal{C}_1$. That is, if $\mathcal{C}_1$ contains some models in Category II, then the probability that the CV(n_v) selects the model in $\mathcal{C}_1$ and in Category II with the smallest size tends to 1 as $n \to \infty$. If all the models in $\mathcal{C}_1$ are in Category I, then the CV(n_v) selects the model that minimizes $\Delta_{\alpha,n}$, provided that the R_n in (3.13) or (3.23) satisfies

$$R_n = o_p(1). \quad (4.1)$$

Condition (4.1) will be discussed later.

A similar situation is where some predictors, which should be included in the model, are overlooked by the data analyst. Then the CV(n_v) selects the best model within $\mathcal{C}$.

4.2 Computation Algorithms

It is possible that using a good algorithm may preclude the need to compute $\hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}}$ (or $\hat{\Gamma}_{\alpha,n}^{\mathrm{MCCV}}$ and $\hat{\Gamma}_{\alpha,n}^{\mathrm{APCV}}$) for all $2^p - 1$ subsets α when selecting a model from $\mathcal{C}$. For example, a backward selection may be used: Suppose that $\hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}}$ is first computed for the subset $\gamma = \{1, \ldots, p\}$ and all the subsets α with $d_\alpha = p - 1$. If

$$\min_{\alpha : d_\alpha = p-1} \hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}} > \hat{\Gamma}_{\gamma,n}^{\mathrm{BICV}},$$

then the computation may be stopped and model $\mathcal{M}_\gamma$ is selected. This is because if $\mathcal{M}_\gamma$ is not the optimal model, then

$$P\Big(\min_{\alpha : d_\alpha = p-1} \hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}} > \hat{\Gamma}_{\gamma,n}^{\mathrm{BICV}}\Big) \to 0.$$

A forward selection can be used similarly. Further discussion of computation algorithms is beyond the scope of this article.

4.3 Extensions

One advantage of the cross-validation method over other methods is that its extension to more complicated models, such as nonlinear regression and generalized linear models, is straightforward. One simply uses

$$\frac{1}{n_v b}\sum_{s\in\mathcal{L}} Q(\mathbf{y}_s, \hat{\mathbf{y}}_{\alpha,s^c}),$$

where $Q(\cdot, \cdot)$ is a loss function, $\hat{\mathbf{y}}_{\alpha,s^c}$ is the prediction of $\mathbf{y}_s$ based on the construction data under model $\mathcal{M}_\alpha$, $\mathcal{L} = \mathcal{B}$ for BICV(n_v), and $\mathcal{L} = \mathcal{R}$ for MCCV(n_v).

4.4 Conditions (3.11) and (3.21)

Condition (3.11) is a technical condition required for the consistency of BICV(n_v). I illustrate here that it is a reasonably weak condition.

This condition requires some degree of resemblance between the validation data $\{(y_i, \mathbf{x}_i), i \in s\}$ and the construction data $\{(y_i, \mathbf{x}_i), i \in s^c\}$. Note that $\mathbf{x}_i\mathbf{x}_i'$ is the Fisher information matrix about β contained in the pair $(y_i, \mathbf{x}_i)$. Then condition (3.11) requires that

$$\frac{1}{n_v}\sum_{i\in s}\mathbf{x}_i\mathbf{x}_i' - \frac{1}{n_c}\sum_{i\in s^c}\mathbf{x}_i\mathbf{x}_i', \tag{4.2}$$

which is the difference between the average Fisher information matrices based on the validation and construction data, vanishes as $n \to \infty$ uniformly over all splits used in cross-validation; that is, all $s \in \mathcal{B}$.

Clearly, (3.11) is implied by

$$\lim_{n\to\infty}\max_{\text{all } s}\left\|\frac{1}{n_v}\sum_{i\in s}\mathbf{x}_i\mathbf{x}_i' - \frac{1}{n_c}\sum_{i\in s^c}\mathbf{x}_i\mathbf{x}_i'\right\| = 0. \tag{4.3}$$

But (4.3) is much stronger than (3.11) because it requires that the differences of the form (4.2) be small uniformly over all subsets s and that $\mathcal{B}$ contains much fewer subsets.

As an example, suppose that the (u, v)th element of $\mathbf{x}_i\mathbf{x}_i'$ is $a_i = a_{iuv}$ and that for any (u, v), $\{a_i, i = 1, 2, \ldots\}$ is a sequence of nonincreasing numbers satisfying

$$\frac{1}{m}\sum_{i=1}^{m} a_i = \xi + O(m^{-\delta})$$

for some $\delta > 0$. For any subset s,

$$\frac{1}{n_c}\sum_{i=n_v+1}^{n} a_i \le \frac{1}{n_c}\sum_{i\in s^c} a_i \le \frac{1}{n_c}\sum_{i=1}^{n_c} a_i = \xi + O(n_c^{-\delta})$$

and

$$\frac{1}{n_c}\sum_{i=n_v+1}^{n} a_i - \xi = \frac{n}{n_c}\left(\frac{1}{n}\sum_{i=1}^{n} a_i - \xi\right) - \frac{n_v}{n_c}\left(\frac{1}{n_v}\sum_{i=1}^{n_v} a_i - \xi\right)$$

$$= \frac{n}{n_c}O(n^{-\delta}) + \frac{n_v}{n_c}O(n_v^{-\delta}) = O(n_c^{-1}n^{1-\delta}),$$

if $n_v/n \to 1$. It is easy to select an n_c satisfying (3.12) and $n^{1-\delta}/n_c \to 0$. Hence

$$\lim_{n\to\infty}\max_{\text{all } s}\left|\frac{1}{n_c}\sum_{i\in s^c} a_i - \xi\right| = 0.$$

Similarly, it can be shown that

$$\lim_{n\to\infty}\max_{\text{all } s}\left|\frac{1}{n_v}\sum_{i\in s} a_i - \xi\right| = 0,$$

and thus (4.3) holds.

If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are random, then (3.11) holds in probability under weak conditions. Suppose that $\{\mathbf{x}_i\}$ is a sequence of iid random vectors satisfying $E(\mathbf{x}_i'\mathbf{x}_i)^{2+\tau} < \infty$ with a $\tau > 0$. Then, using a Berry–Esseen inequality (e.g., Shorack and Wellner 1986, thm. 3, p. 849), it can be shown that for any s,

$$P\left(\left\|\frac{1}{n_v}\sum_{i\in s}\mathbf{x}_i\mathbf{x}_i' - \frac{1}{n_c}\sum_{i\in s^c}\mathbf{x}_i\mathbf{x}_i'\right\| > \varepsilon\right) = O(n_c^{-(1+\tau)}).$$

Then

$$P\left(\max_{s\in\mathcal{B}}\left\|\frac{1}{n_v}\sum_{i\in s}\mathbf{x}_i\mathbf{x}_i' - \frac{1}{n_c}\sum_{i\in s^c}\mathbf{x}_i\mathbf{x}_i'\right\| > \varepsilon\right) = O(bn_c^{-(1+\tau)}).$$

Hence (3.11) holds for almost all $\mathbf{x}_1, \mathbf{x}_2, \ldots$, if for some $\delta > 0$,

$$bn_c^{-(1+\tau)} \le n^{-(1+\delta)}. \tag{4.4}$$

For example, n_c can be chosen to be the integer part of $n^{3/4}$. Then (3.12), (3.22), and (4.4) hold for $b = O(n)$ and any $\tau > \frac{5}{3}$. This choice of n_c is used in the simulation study in Section 5.

The discussion for condition (3.21) is similar.

Table 1. The Values of x_{ki}

x_{2i}	x_{3i}	x_{4i}	x_{5i}
.3600	.5300	1.0600	.5326
1.3200	2.5200	5.7400	3.6183
.0600	.0900	.2700	.2594
.1600	.4100	.8300	1.0346
.0100	.0200	.0700	.0381
.0200	.0700	.0700	.3440
.5600	.6200	2.1200	1.4559
.9800	1.0600	2.8900	4.0182
.3200	.2000	.7600	.4600
.0100	.0000	.0700	.1540
.1500	.2500	.5000	.6516
.2400	.2800	.5900	.0611
.1100	.3500	.4000	.1922
.0800	.1300	.2800	.0931
.6100	.8500	.4900	.0538
.0300	.0300	.2300	.0199
.0600	.1100	.5000	.0419
.0200	.0800	.2500	.1093
.0400	.2400	.0800	.0328
.0000	.0200	.0400	.0797
.0900	.1800	.5900	.1855
.0200	.1600	.2400	.1572
.0200	.1100	.2100	.0998
.0500	.2400	.4300	.2804
.1100	.3900	.2900	.2879
.1800	.1100	.4300	.6810
.0400	.0900	.2300	.3242
.8500	1.3300	2.7000	2.6013
.1700	.3200	.6600	.4469
.0800	.1200	.4900	.2436
.3800	.1800	.4900	.4400
.1100	.1300	.1800	.3351
.3900	.3800	.9900	1.3979
.4300	.4600	1.4700	2.0138
.5700	1.1600	1.8200	1.9356
.1300	.0300	.0800	.1050
.0400	.0500	.1400	.2207
.1300	.1800	.2800	.0180
.2000	.9500	.4100	.1017
.0700	.0600	.1800	.0962

Table 2. Probabilities (Based on 1,000 Simulations) of Selecting Each Model

	Model	Category	CV(1)	MCCV(n_v)	APCV(n_v)
$\beta = (2, 0, 0, 4, 0)'$	1, 4	Optimal	.484	.934	.501
	1, 2, 4	II	.133	.025	.116
	1, 3, 4	II	.127	.026	.085
	1, 4, 5	II	.138	.012	.172
	1, 2, 3, 4	II	.049	.000	.038
	1, 2, 4, 5	II	.029	.001	.039
	1, 3, 4, 5	II	.030	.002	.037
	1, 2, 3, 4, 5	II	.009	.000	.012
$\beta = (2, 0, 0, 4, 8)'$	1, 4, 5	Optimal	.641	.947	.651
	1, 2, 4, 5	II	.158	.032	.161
	1, 3, 4, 5	II	.138	.020	.131
	1, 2, 3, 4, 5	II	.063	.001	.057
$\beta = (2, 9, 0, 4, 8)'$	1, 4, 5	I	.005	.016	.000
	1, 2, 4, 5	Optimal	.801	.965	.818
	1, 3, 4, 5	I	.005	.002	.000
	1, 2, 3, 4, 5	II	.189	.017	.182
$\beta = (2, 9, 6, 4, 8)'$	1, 2, 3, 5	I	.000	.002	.000
	1, 2, 4, 5	I	.000	.005	.000
	1, 3, 4, 5	I	.015	.045	.001
	1, 2, 3, 4, 5	Optimal	.985	.948	.999

4.5 Condition (4.1)

From the proof of (3.13) in the Appendix, for the BICV(n_v),

$$R_n = \hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}} - \frac{1}{n}\|\mathbf{y} - \mathbf{X}_\alpha\hat{\beta}_\alpha\|^2. \tag{4.5}$$

Note that

$$\frac{n + n_c}{n_c(n-1)} \sum_i w_{i\alpha}(y_i - \mathbf{x}_{i\alpha}'\hat{\beta}_\alpha)^2 = O_p\left(\frac{n}{n_c}\max_{i\le n} w_{i\alpha}\right). \tag{4.6}$$

Then by (A.6), (A.7), and (A.9) in the Appendix, (4.1) holds if the right side of (4.6) is of the order $o_p(1)$, which is equivalent to [by (3.3)]

$$n_c^{-1}\max_{i\le n}\mathbf{x}_i'\mathbf{x}_i = o(1). \tag{4.7}$$

Note that (4.7) holds if $\{\|\mathbf{x}_i\|\}$ is bounded. For random iid $\mathbf{x}_i$, (4.7) holds almost surely if $E(\mathbf{x}_i'\mathbf{x}_i)^2 < \infty$ and $n_c = n^{3/4}$.

The discussion for the MCCV(n_v) or the APCV(n_v) is similar.

5. A SIMULATION STUDY

Now the finite sample performance of the cross-validation method is studied by simulation. The following model is considered:

$$y_i = \beta_i x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + e_i,$$

where $i = 1, \ldots, 40$, e_i are iid from $N(0, 1)$, x_{ki} is the ith value of the kth prediction variable x_k, $x_{1i} \equiv 1$, and the values of x_{ki}, $k = 2, \ldots, 5$, $i = 1, \ldots, 40$, are taken from an example in Gunst and Mason (1980) (see Table 1). Some of the β_k may be 0. Therefore, some prediction variables are selected from five possible variables $\{x_1, \ldots, x_5\}$, and the model with the best predictive ability is chosen. Note that there are 31 possible models, and each model will be denoted by a subset of $\{1, \ldots, 5\}$ that contains the indices of the variables x_k in the model.

Three cross-validation methods are considered: the CV(1), the MCCV(n_v), and the APCV(n_v) given in Section 3.3 with $n_v = 25$ ($n_c = 15 \approx n^{3/4}$). For the MCCV(n_v), $b = 2n$ is used. Table 2 gives the empirical probabilities (based on 1,000 simulations) of selecting each model in several different cases.

The results in Table 2 can be summarized as follows:

1. In terms of the probability of selecting the optimal model, the MCCV(n_v) has the best performance among the three methods under consideration, except for the case where the largest model (the model with all the $\beta_k \neq 0$) is the optimal model. The APCV(n_v) is slightly better than the CV(1) in all the cases.
2. The probability of selecting a model from Category I (incorrect model) is negligible for all three methods in all cases under consideration.
3. As expected, the CV(1) tends to select unnecessarily large models. The probability of selecting the optimal model by using the CV(1) can be very low (e.g., $\le .5$). The more zero components the β has, the worse performance the CV(1) has. On the other hand, the performance of the MCCV(n_v) is stable and is much better than the CV(1) in the cases where the optimal model is not the largest model.
4. The performance of the APCV(n_v) is only slightly better than the CV(1), although the APCV(n_v) is consistent and the CV(1) is inconsistent. This indicates that to have a good performance, the APCV(n_v) may require a larger sample size than the MCCV(n_v).

APPENDIX: PROOFS

A.1 Proof of (3.5) and (3.6)

From $(1 - w_{i\alpha})^{-2} = 1 + 2w_{i\alpha} + O(w_{i\alpha}^2)$,

$$\hat{\Gamma}_{\alpha,n}^{\mathrm{CV}} = \frac{1}{n}\sum_i r_{i\alpha}^2 + \frac{1}{n}\sum_i [2w_{i\alpha} + O(w_{i\alpha}^2)]r_{i\alpha}^2, \tag{A.1}$$

where $r_{i\alpha} = y_i - \mathbf{x}_{i\alpha}'\hat{\beta}_\alpha$. Let $\xi_{\alpha,n}$ and $\zeta_{\alpha,n}$ be the first term and the second term on the right side of (A.1). Then (3.5) follows from $\zeta_{\alpha,n} \le O(\max_i w_{i\alpha})\xi_{\alpha,n}$ and

$$\begin{aligned}\xi_{\alpha,n} &= n^{-1}\mathbf{e}'(\mathbf{I}_n - \mathbf{P}_\alpha)\mathbf{e} + \Delta_{\alpha,n} + 2n^{-1}\mathbf{e}'(\mathbf{I}_n - \mathbf{P}_\alpha)\mathbf{X}\beta \\ &= n^{-1}\mathbf{e}'\mathbf{e} + \Delta_{\alpha,n} + o_p(1),\end{aligned}$$

because $E(\mathbf{e}'\mathbf{P}_\alpha\mathbf{e}) = d_\alpha\sigma^2$ and $E[\mathbf{e}'(\mathbf{I}_n - \mathbf{P}_\alpha)\mathbf{X}\beta]^2 = \sigma^2\beta'\mathbf{X}'(\mathbf{I}_n - \mathbf{P}_\alpha)\mathbf{X}\beta = O(n)$. If $\mathcal{M}_\alpha$ is in Category II, then $\zeta_{\alpha,n} = 2n^{-1}d_\alpha\sigma^2 + o_p(n^{-1})$ and $\xi_{\alpha,n} = n^{-1}\mathbf{e}'(\mathbf{I}_n - \mathbf{P}_\alpha)\mathbf{e}$. Hence (3.6) holds.

A.2 Proof of Theorem 1

From (3.1) and the balance property of $\mathcal{B}$,

$$\hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}} \geq \frac{1}{n_v b}\sum_{s\in\mathcal{B}}\|\mathbf{y}_s - \mathbf{X}_{\alpha,s}\hat{\beta}_\alpha\|^2 = n^{-1}\|\mathbf{y} - \mathbf{X}_\alpha\hat{\beta}_\alpha\|^2$$
$$= \xi_{\alpha,n} = n^{-1}\mathbf{e}'\mathbf{e} + \Delta_{\alpha,n} + o_p(1),$$

where the last equality follows from the proof of (3.5). Hence (3.13) follows by letting R_n be as given by (4.5).

From condition (3.11), for $s \in \mathcal{B}$,

$$\frac{1}{n}\mathbf{X}_\alpha'\mathbf{X}_\alpha - \frac{1}{n_v}\mathbf{X}_{\alpha,s}'\mathbf{X}_{\alpha,s} = \frac{n_c}{n}\left[\frac{1}{n_c}\mathbf{X}_{\alpha,s^c}'\mathbf{X}_{\alpha,s^c} - \frac{1}{n_v}\mathbf{X}_{\alpha,s}'\mathbf{X}_{\alpha,s}\right]$$
$$= o\left(\frac{n_c}{n}\right),$$

which together with (3.3) implies that

$$(\mathbf{X}_{\alpha,s}'\mathbf{X}_{\alpha,s})^{-1} - \frac{n}{n_v}(\mathbf{X}_\alpha'\mathbf{X}_\alpha)^{-1} = o\left(\frac{n_c}{n}\right)(\mathbf{X}_{\alpha,s}'\mathbf{X}_{\alpha,s})^{-1},$$

and, therefore,

$$\mathbf{P}_{\alpha,s} = \frac{n}{n_v}\mathbf{Q}_{\alpha,s} + o\left(\frac{n_c}{n}\right)\mathbf{P}_{\alpha,s}, \qquad s \in \mathcal{B}, \tag{A.2}$$

where $\mathbf{P}_{\alpha,s} = \mathbf{X}_{\alpha,s}(\mathbf{X}_{\alpha,s}'\mathbf{X}_{\alpha,s})^{-1}\mathbf{X}_{\alpha,s}'$. From (A.2) and condition (3.12),

$$\mathbf{Q}_{\alpha,s} = \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right]\mathbf{P}_{\alpha,s}, \qquad s \in \mathcal{B}. \tag{A.3}$$

From the balance property of $\mathcal{B}$,

$$\frac{1}{n_v b}\sum_{s\in\mathcal{B}}\mathbf{r}_{\alpha,s}'\mathbf{Q}_{\alpha,s}\mathbf{r}_{\alpha,s} = \left[\frac{1}{n} - \frac{n_v - 1}{n(n-1)}\right]\sum_i w_{i\alpha}r_{i\alpha}^2,$$

where $\mathbf{r}_{\alpha,s} = \mathbf{y}_s - \mathbf{X}_{\alpha,s}\hat{\beta}_\alpha$. Then, by (A.3) and (3.12),

$$\frac{c_n}{n_v b}\sum_{s\in\mathcal{B}}\|\mathbf{P}_{\alpha,s}\mathbf{r}_{\alpha,s}\|^2 = \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right]^{-1}\frac{c_n}{n_v b}\sum_{s\in\mathcal{B}}\mathbf{r}_{\alpha,s}'\mathbf{Q}_{\alpha,s}\mathbf{r}_{\alpha,s}$$
$$= \left[1 + o\left(\frac{n_c}{n}\right)\right]\frac{n + n_c}{n_c(n-1)}\sum_i w_{i\alpha}r_{i\alpha}^2, \tag{A.4}$$

where

$$c_n = n_v(n + n_c)n_c^{-2}. \tag{A.5}$$

Define

$$\mathbf{U}_{\alpha,s} = (\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})(\mathbf{I}_{n_v} + c_n\mathbf{P}_{\alpha,s})(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s}),$$
$$A_\alpha = \frac{1}{n_v b}\sum_{s\in\mathcal{B}}\mathbf{r}_{\alpha,s}'(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}\mathbf{U}_{\alpha,s}(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}\mathbf{r}_{\alpha,s},$$

and

$$B_\alpha = \frac{1}{n_v b}\sum_{s\in\mathcal{B}}\mathbf{r}_{\alpha,s}'(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}(\mathbf{I}_{n_v} - \mathbf{U}_{\alpha,s})(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}\mathbf{r}_{\alpha,s}.$$

Then, by (3.1),

$$\hat{\Gamma}_{\alpha,n}^{\mathrm{BICV}} = A_\alpha + B_\alpha. \tag{A.6}$$

From the balance property of $\mathcal{B}$ and (A.4),

$$A_\alpha = \frac{1}{n_v b}\sum_{s\in\mathcal{B}}\|\mathbf{r}_{\alpha,s}\|^2 + \frac{c_n}{n_v b}\sum_{s\in\mathcal{B}}\|\mathbf{P}_{\alpha,s}\mathbf{r}_{\alpha,s}\|^2$$
$$= \frac{1}{n}\|\mathbf{y} - \mathbf{X}_\alpha\hat{\beta}_\alpha\|^2 + \left[1 + o\left(\frac{n_c}{n}\right)\right]\frac{n + n_c}{n_c(n-1)}\sum_i w_{i\alpha}r_{i\alpha}^2. \tag{A.7}$$

Assume that $\mathcal{M}_\alpha$ is in Category II. Then, by (A.7) and the fact that $\sum_i w_{i\alpha}r_{i\alpha}^2 = d_\alpha\sigma^2 + o_p(1)$, we have

$$A_\alpha = \frac{1}{n}\mathbf{e}'(\mathbf{I} - \mathbf{P}_\alpha)\mathbf{e} + \left[1 + o\left(\frac{n_c}{n}\right)\right]\frac{n + n_c}{n_c(n-1)}[d_\alpha\sigma^2 + o_p(1)]$$
$$= \frac{1}{n}\mathbf{e}'\mathbf{e} + \frac{d_\alpha\sigma^2}{n_c} + o_p\left(\frac{1}{n_c}\right).$$

It remains to show that $B_\alpha = o_p(n_c^{-1})$. From (A.3),

$$(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})\mathbf{P}_{\alpha,s}(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s}) = \left[1 - \frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right]\mathbf{P}_{\alpha,s}(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})$$
$$= \left[1 - \frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right]^2\mathbf{P}_{\alpha,s}$$
$$= \left[\frac{n_c}{n} + o\left(\frac{n_c}{n}\right)\right]^2\mathbf{P}_{\alpha,s}.$$

Hence

$$\left(\frac{n}{n_c}\right)^2(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})\mathbf{P}_{\alpha,s}(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s}) = [1 + o(1)]^2\mathbf{P}_{\alpha,s} \geq \frac{1}{2}\mathbf{P}_{\alpha,s}$$

for $s \in \mathcal{B}$ when n is sufficiently large. Then

$$(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}\mathbf{P}_{\alpha,s}(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1} \leq 2\left(\frac{n}{n_c}\right)^2\mathbf{P}_{\alpha,s}. \tag{A.8}$$

Also, by (A.3),

$$\mathbf{U}_{\alpha,s} = \left\{\mathbf{I}_{n_v} - \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right]\mathbf{P}_{\alpha,s}\right\}(\mathbf{I}_{n_v} + c_n\mathbf{P}_{\alpha,s})$$
$$\times\left\{\mathbf{I}_{n_v} - \left[\frac{n_v}{n} + o\left(\frac{n_c}{n}\right)\right]\mathbf{P}_{\alpha,s}\right\}$$
$$= \left(\mathbf{I}_{n_v} - \frac{n_v}{n}\mathbf{P}_{\alpha,s}\right)(\mathbf{I}_{n_v} + c_n\mathbf{P}_{\alpha,s})\left(\mathbf{I}_{n_v} - \frac{n_v}{n}\mathbf{P}_{\alpha,s}\right)$$
$$+ \left[o\left(\frac{n_c}{n}\right)\right]^2(1 + c_n)\mathbf{P}_{\alpha,s} + 2\left[o\left(\frac{n_c}{n}\right)\right]\left(1 - \frac{n_v}{n}\right)(1 + c_n)\mathbf{P}_{\alpha,s}$$
$$= \left(\mathbf{I}_{n_v} - \frac{n_v}{n}\mathbf{P}_{\alpha,s}\right)^2 + c_n\left(1 - \frac{n_v}{n}\right)^2\mathbf{P}_{\alpha,s} + \left[o\left(\frac{n_c}{n}\right)\right]^2(1 + c_n)\mathbf{P}_{\alpha,s}$$
$$= \mathbf{I}_{n_v} + \left[o\left(\frac{n_c}{n}\right)\right]^2(1 + c_n)\mathbf{P}_{\alpha,s},$$

because

$$c_n\left(1 - \frac{n_v}{n}\right)^2 = \frac{n_v}{n}\left(2 - \frac{n_v}{n}\right).$$

Then, by (A.8),

$$(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}(\mathbf{I}_{n_v} - \mathbf{U}_{\alpha,s})(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}$$
$$= \left[o\left(\frac{n_c}{n}\right)\right]^2(1 + c_n)(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}\mathbf{P}_{\alpha,s}(\mathbf{I}_{n_v} - \mathbf{Q}_{\alpha,s})^{-1}$$
$$\leq o(1)(1 + c_n)\mathbf{P}_{\alpha,s}.$$

Therefore,

$$B_\alpha \leq o(1)(1 + c_n)\left(\frac{1}{n_v b}\sum_{s\in\mathcal{B}}\|\mathbf{P}_{\alpha,s}\mathbf{r}_{\alpha,s}\|^2\right) = o_p\left(\frac{1}{n_c}\right), \tag{A.9}$$

because from the previous proof,

$$\frac{c_n}{n_v b}\sum_{s\in\mathcal{B}}\|\mathbf{P}_{\alpha,s}\mathbf{r}_{\alpha,s}\|^2 = O_p\left(\frac{1}{n_c}\right).$$

This proves (3.14).

A.3 Proof of Theorem 2

Assume that $\mathcal{M}_\alpha$ is in Category II. Let B_α be as given in the proof of Theorem 1, with $\mathcal{B}$ replaced by $\mathcal{R}$. Then (A.9) can be established using conditions (3.21) and (3.22) and the same argument as that in the proof of Theorem 1. Let c_n be as given by (A.5),

$$A_{\alpha 1} = \frac{c_n}{n_v b} \sum_{s \in \mathcal{R}} \mathbf{e}_s' \mathbf{P}_{\alpha,s} \mathbf{e}_s,$$

$$A_{\alpha 2} = \frac{2(1 + c_n)}{n_v b} \sum_{s \in \mathcal{R}} \mathbf{e}_s' \mathbf{X}_{\alpha,s} (\hat{\beta}_\alpha - \beta_\alpha),$$

and

$$A_{\alpha 3} = \frac{1 + c_n}{n_v b} \sum_{s \in \mathcal{R}} (\hat{\beta}_\alpha - \beta_\alpha)' \mathbf{X}_{\alpha,s}' \mathbf{X}_{\alpha,s} (\hat{\beta}_\alpha - \beta_\alpha).$$

Then

$$\hat{\Gamma}_{\alpha,n}^{\mathrm{MCCV}} = \frac{1}{n_v b} \sum_{s \in \mathcal{R}} \mathbf{e}_s' \mathbf{e}_s + A_{\alpha 1} - A_{\alpha 2} + A_{\alpha 3} + o_p(n_c^{-1}). \quad \text{(A.10)}$$

Using condition (3.21), (A.2) holds with $\mathcal{B}$ replaced by $\mathcal{R}$. Then

$$A_{\alpha 1} = \frac{c_n}{n_v b} \sum_{s \in \mathcal{R}} \frac{n}{n_v} \mathbf{e}_s' \mathbf{Q}_{\alpha,s} \mathbf{e}_s + o_p\left(\frac{n_c}{n}\right) \frac{c_n}{n_v b} \sum_{s \in \mathcal{R}} \mathbf{e}_s' \mathbf{P}_{\alpha,s} \mathbf{e}_s. \quad \text{(A.11)}$$

Denote the two terms on the right side of (A.11) by $B_{\alpha 1}$ and $B_{\alpha 2}$. Let $E_{\mathcal{R}}$ and $V_{\mathcal{R}}$ be the expectation and variance with respect to the random selection of $\mathcal{R}$. Using the equality

$$E_{\mathcal{R}}\left(\frac{1}{b} \sum_{s \in \mathcal{R}} a_s\right) = \binom{n}{n_v}^{-1} \sum_{\text{all } s} a_s,$$

we obtain that

$$\begin{aligned} &E_{\mathcal{R}}(B_{\alpha 1} - A_{\alpha 2} + A_{\alpha 3}) \\ &= \frac{c_n n}{n_v} \left[\frac{n_v - 1}{n(n-1)} \mathbf{e}' \mathbf{P}_\alpha \mathbf{e} + \frac{n_c}{n(n-1)} \sum_i w_{i\alpha} e_i^2 \right] \\ &\quad - \frac{2(1 + c_n)}{n} \mathbf{e}' \mathbf{P}_\alpha \mathbf{e} + \frac{1 + c_n}{n} \mathbf{e}' \mathbf{P}_\alpha \mathbf{e} \\ &= \frac{c_n n}{n_v} \left[\frac{n_v}{n^2} \mathbf{e}' \mathbf{P}_\alpha \mathbf{e} + \frac{d_\alpha \sigma^2 n_c}{n^2} + o_p\left(\frac{n_c}{n^2}\right) \right] - \frac{1 + c_n}{n} \mathbf{e}' \mathbf{P}_\alpha \mathbf{e} \\ &= \frac{d_\alpha \sigma^2}{n_c} + o_p(n_c^{-1}). \end{aligned} \quad \text{(A.12)}$$

Using the inequality

$$V_{\mathcal{R}}\left(\frac{1}{b} \sum_{s \in \mathcal{R}} a_s\right) \le \frac{1}{b} E_{\mathcal{R}} a_s^2$$

and letting $t_n = O[n^2/(n_c^2 b)]$, we obtain that

$$\begin{aligned} V_{\mathcal{R}}(n_c B_{\alpha 1}) &\le t_n E_{\mathcal{R}} (\mathbf{e}_s' \mathbf{Q}_{\alpha,s} \mathbf{e}_s)^2 \\ &\le 2 t_n E_{\mathcal{R}} \left[\left(\sum_{i \in s} w_{i\alpha} e_i^2 \right)^2 + \left(\sum_{i,j \in s, i \ne j} w_{ij\alpha} e_i e_j \right)^2 \right] \\ &\le 2 t_n \left[\left(\sum_i w_{i\alpha} e_i^2 \right)^2 + \binom{n}{n_v}^{-1} \sum_{\text{all } s} \left(\sum_{i,j \in s, i \ne j} w_{ij\alpha} e_i e_j \right)^2 \right] \\ &= 2 t_n [O_p(1) + O_p(1)] = O_p(t_n), \end{aligned} \quad \text{(A.13)}$$

$$\begin{aligned} V_{\mathcal{R}}(n_c A_{\alpha 2}) &\le t_n (\hat{\beta}_\alpha - \beta_\alpha)' E_{\mathcal{R}} (\mathbf{X}_{\alpha,s}' \mathbf{e}_s \mathbf{e}_s' \mathbf{X}_{\alpha,s}) (\hat{\beta}_\alpha - \beta_\alpha) \\ &= O_p(t_n) \end{aligned} \quad \text{(A.14)}$$

and

$$\begin{aligned} V_{\mathcal{R}}(n_c A_{\alpha 3}) &\le t_n E_{\mathcal{R}} [(\hat{\beta}_\alpha - \beta_\alpha)' \mathbf{X}_{\alpha,s}' \mathbf{X}_{\alpha,s} (\hat{\beta}_\alpha - \beta_\alpha)]^2 \\ &\le t_n [(\hat{\beta}_\alpha - \beta_\alpha)' \mathbf{X}_\alpha' \mathbf{X}_\alpha (\hat{\beta}_\alpha - \beta_\alpha)]^2 = O_p(t_n), \end{aligned}$$

where $w_{ij\alpha}$ is the (i,j)th element of $\mathbf{P}_\alpha$, (A.13) follows from the fact that for any s,

$$E\left(\sum_{i,j \in s, i \ne j} w_{ij\alpha} e_i e_j \right)^2 = \sum_{i,j \in s, i \ne j} w_{ij\alpha}^2 \sigma^4 \le p \sigma^4,$$

and (A.14) follows from $E E_{\mathcal{R}} (\mathbf{X}_{\alpha,s}' \mathbf{e}_s \mathbf{e}_s' \mathbf{X}_{\alpha,s}) \le \sigma^2 \mathbf{X}_\alpha' \mathbf{X}_\alpha$. Because $t_n \to 0$ under (3.22),

$$V_{\mathcal{R}}(B_{\alpha 1} - A_{\alpha 2} + A_{\alpha 3}) = o_p(n_c^{-1}). \quad \text{(A.15)}$$

Hence (3.24) follows from (A.10)–(A.12), (A.15), and the fact that $E \mathbf{e}_s' \mathbf{P}_{\alpha,s} \mathbf{e}_s = d_\alpha \sigma^2$ implies $B_{\alpha 2} = o_p(n_c^{-1})$.

The proof for (3.23) is similar.

[*Received May 1991. Revised November 1991.*]

REFERENCES

Akaike, H. (1974), "A New Look at Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Allen, D. M. (1974), "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," *Technometrics*, 16, 125–127.

Burman, P. (1989), "A Comparative Study of Ordinary Cross-Validation, *v*-Hold Cross-Validation, and the Repeated Learning-Testing Methods," *Biometrika*, 76, 503–514.

Efron, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331.

——— (1986), "How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association*, 81, 461–470.

Geisser, S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

Gunst, G. F., and Mason, R. L. (1980), *Regression Analysis and Its Application*, New York: Marcel Dekker.

Herzberg, G., and Tsukanov, S. (1986), "A Note on Modifications of the Jackknife Criterion on Model Selection," *Utilitas Mathematics*, 29, 209–216.

John, P. W. M. (1971), *Statistical Design and Analysis of Experiments*, New York: Macmillan.

Li, K.-C. (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation, and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958–975.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

Picard, R. R., and Cook, R. D. (1984), "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575–583.

Shao, J., and Wu, C. F. J. (1989), "A General Theory for Jackknife Variance Estimation," *The Annals of Statistics*, 17, 1176–1197.

Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45–54.

Shorack, G. R., and Wellner, J. A. (1986), *Empirical Process with Applications to Statistics*, New York: John Wiley.

Stone, M. (1974), "Cross-Validation Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

——— (1977a), "An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion," *Journal of the Royal Statistical Society*, Ser. B, 39, 44–47.

——— (1977b), "Asymptotics for and Against Cross-Validation," *Biometrika*, 64, 29–38.

Wahba, G., and Wold, S. (1975), "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation," *Communications in Statistics*, 4, 1–17.

Zhang, P. (1991), "Model Selection Via Multifold Cross-Validation," Preprint.