MODEL SELECTION WITH DATA-ORIENTED PENALTY

Z.D. Bai, C.R. Rao[1] and Y. Wu

Technical Report 97-07

April 1997

Center for Multivariate Analysis
417 Thomas Building
Penn State University
University Park, PA 16802

[1]The research work of the author supported by the Army Research Office under Grant DAAHO4-96-1-0082. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 1997	Technical - 97-07

4. TITLE AND SUBTITLE
Model Selection with Data-Oriented Penalty

5. FUNDING NUMBERS
DAAH04-96-1-0082

6. AUTHOR(S)
Z.D. Bai, C.R. Rao and Y. Wu

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Center for Multivariate Analysis
Department of Statistics
Penn State University
417 Thomas Bldg.
University Park, PA 16802

8. PERFORMING ORGANIZATION REPORT NUMBER
97-07

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

We consider the model selection or variables selection in the classical regression problem. In the literature, there are two types of criteria for model selection, one based on prediction error (FPE) and another on information theoretic considerations (GIC). Each of these criteria uses a certain penalty function which is the product of the number of variables j in a submodel and a function C_n depending on n and satisfying some conditions to guarantee consistency in model selection. One of the important problems in such a procedure is the actual choice of C_n in a given situation. In this paper we show that a particular choice of C_n based on observed data, which makes it random, preserves the consistency property and shows improved performance over a fixed choice of C_n.

14. SUBJECT TERMS	15. NUMBER OF PAGES
AIC, FPE, GIC, Linear regression, Model selection Variables selection	22
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Model Selection with Data-Oriented Penalty

By Z. D. Bai

Department of Applied Mathematics, National Sun Yat-sen University, Taiwan

C. R. Rao[1]

Department of Statistics, Penn State University, U.S.A.

Y. Wu[2]

Department of Mathematics and Statistics, York University, Canada

ABSTRACT

We consider the model selection or variables selection in the classical regression problem. In the literature, there are two types of criteria for model selection, one based on prediction error (FPE) and another on information theoretic considerations (GIC). Each of these criteria uses a certain penalty function which is the product of the number of variables j in a submodel and a function C_n depending on n and satisfying some conditions to guarantee consistency in model selection. One of the important problems in such a procedure is the actual choice of C_n in a given situation. In this paper we show that a particular choice of C_n based on observed data, which makes it random, preserves the consistency property and shows improved performance over a fixed choice of C_n.

AMS 1991 Subject Classification: 62J05, 62F10.

Key Words and Phrases: AIC, FPE, GIC, Linear regression, Model selection, Variables selection.

[1]Research supported by Army Research Office under the grant no.DAAH04-96-1-0082

[2]Research supported by the Natural Sciences and Engineering Research Council of Canada

1 Introduction

Consider the multiple regression model

$$y_n = X_n\beta + e_n \tag{1.1}$$

where X_n is an $n \times p$ matrix, β is a p-vector of unknown regression parameters and e_n is a random error vector. Each component of β may be zero or nonzero. Each subset $\mathcal{M}$ of $\{1, 2, \cdots, p\}$ is called a sub-model. It is obvious that there are 2^p possible sub-models for the multiple regression problem. A sub-model is called a true model if $\beta_i = 0$ for all $i \notin \mathcal{M}$. The problem is to find the smallest true model which is defined to be the one whose all proper sub-models are not true models.

Many model selection rules have been proposed in the literature for choosing the smallest true model of the multiple regression problem. Cross-validation is a popular method for selecting the true model, which selects the sub-model such that it gives the best average prediction error for the observations. Reference may be made, among others, to Stone (1974, 1977a,b), Geisser (1975), Efron (1983, 1986), Picard and Cook (1984) and Rao (1987). When the number k of predictors is fixed, the cross-validation is equivalent to Akaike's AIC which does not provide a consistent procedure. Shao (1993) showed that $k/n \to 1$ as $n \to \infty$ is needed to guarantee the selected model to be asymptotically correct. When k is large, the amount of computation required for the cross-validation approach is in fact impractical. For reducing the computations with cross-validation for large k, several approaches have been proposed in Shao (1993) and their performances are examined by simulation studies.

Based on the prediction errors, the FPE_α criterion is suggested. For references, see Akaike (1970, 1974), Atkinson (1980), Shibata (1986), and others.

An alternative procedure of model selection is the so-called general information criterion (GIC), dating back to Akaike's AIC (1970, 1973). Further work in this direction can be

found in Mallows (1973), Schwartz (1978), Hanna and Quinn (1979), Shibata (1984) and Zhao *et al.* (1986).

Regarding the relation between FPE_α and GIC, it seems that GIC is more general than FPE_α. For example, the criterion proposed in Rao and Wu (1989) is an FPE_α, but it can also be viewed as a case of GIC. For the performance of the criterion, it is shown in Rao and Wu (1989) that if α is chosen such that $\alpha/n \to 0$, and $\alpha/\log\log n \to \infty$, then the criterion selects the smallest true model with probability one under some mild conditions. In this paper, the restriction on $\boldsymbol{e}_n$ will be relaxed to allow for the components of $\boldsymbol{e}_n$ to be nonidentically distributed. Accordingly, some adjustments will be made in the criterion. It will be shown that the new procedure is also strongly consistent.

The paper is organized as follows: The proposed criteria will be stated and investigated in Section 2, by establishing some general theorems on the strong consistency. Section 3 is devoted to the development of sample-dependent penalty functions. Some applications to the general case will be discussed in Section 4. The simulation results are presented in Section 5. Discussions and comments are given in Section 6. Some technical lemmas are presented in the Appendix.

2 General Model Selection Criteria

Consider the regression model (1.1). Denote $X_n = (\boldsymbol{x}_{1n} \cdots \boldsymbol{x}_{pn}) = (\boldsymbol{x}^{(1)} \cdots \boldsymbol{x}^{(n)})'$. Throughout this paper, P_i stands for the orthogonal projection operator onto the space spanned by $\boldsymbol{x}_{1n}, \ldots, \boldsymbol{x}_{in}$. The following assumptions are needed for establishing our main results.

ASSUMPTION 1. There are constants a_1 and a_2 such that

$$0 < a_1 n \le \lambda_p(X_n'X_n) \le \lambda_1(X_n'X_n) \le a_2 n \tag{2.1}$$

where $\lambda_i(X_n'X_n)$ is the i-th eigen value of $X_n'X_n$.

ASSUMPTION 2. There is a constant $\delta > 0$ such that for each $1 \leq i \leq p$,

$$\sum_{j=1}^{n}(x_{in}^{j})^3 = O[(\boldsymbol{x}_{in}'\boldsymbol{x}_{in})^{3/2}/\log^{1+\delta}(\boldsymbol{x}_{in}'\boldsymbol{x}_{in})] \tag{2.2}$$

where x_{in}^{j} is the jth component of $\boldsymbol{x}_{in} = (x_{in}^{1}, \cdots, x_{in}^{n})'$.

ASSUMPTION 3. The components of $\boldsymbol{e}_n = (e_1, \ldots, e_n)'$ are independently distributed with zero mean and satisfy the moment conditions

$$0 < \nu^2 \leq E(e_i^2), \quad E(|e_i|^3) \leq \tau^3 < \infty \tag{2.3}$$

for all $1 \leq i \leq n$.

We first consider the p consecutive sub-models $\{M_1, \ldots, M_p\}$, where M_k denotes the model $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k \neq 0, 0, \ldots, 0)'$. Let S_k be the residual sum of squares under the model M_k. Define the following criterion functions:

(1) $G_n^{(1)}(k) = S_k + kC_nS_p/(n-p), \quad k = 1, \ldots, p;$

(2) $G_n^{(2)}(k) = S_k + kC_n, \quad k = 1, \ldots, p.$

(3) $G_n^{(3)}(k) = n\log S_k + kC_n, \quad k = 1, \ldots, p;$

where C_n is a function of n satisfying the conditions

$$\frac{C_n}{n} \to 0, \quad \frac{C_n}{\log\log n} \to \infty. \tag{2.4}$$

We propose the following selection rules based on the criteria $G_n^{(i)}$'s; the selected model is defined by $M_{\hat{k}_n}$ for which

$$G_n^{(i)}(\hat{k}_n) = \min_{1\leq k\leq p} G_n^{(i)}(k).$$

In the sequel, we shall call the so-defined selection procedure the Criterion (i).

We first establish the following theorem of the strong consistency of the above criteria.

THEOREM 2.1. *Suppose that the assumptions 1-3 hold for* $n = 1, 2, \ldots$ *and* M_{k_0} *is the smallest true model. If* C_n *satisfies (2.4), then with probability one, for all large* n*, the criterion (1) chooses the smallest true model. The same is true for the criterion (2).*

In order to prove this theorem, we need the following lemma.

LEMMA 2.1. *Suppose that the assumptions 1-3 hold for* $n = 1, 2, \ldots$*, then*

(L1) $a_2 n \geq \boldsymbol{x}_{in}' \boldsymbol{x}_{in} \geq a_1 n, \quad as\ n \to \infty, \quad 1 \leq i \leq p;$

(L2) $a_2 n \geq \boldsymbol{x}_{in}'(I - P_{i-1})\boldsymbol{x}_{in} \geq a_1 n > 0, \quad 1 \leq i \leq p;$

(L3) $\boldsymbol{x}_{in}' \boldsymbol{e}_n = O((n \log\log n)^{1/2}), \ a.s. \quad 1 \leq i \leq p;$

(L4) $\boldsymbol{e}_n' P_i \boldsymbol{e}_n = O(\log\log n), \ a.s. \quad 1 \leq i \leq p;$

(L5) $\sum_{i=1}^n e_i^2/n =$ *is bounded away from 0 and* ∞ *almost surely.*

(L6) $S_p/(n-p)$ *is bounded away from 0 and* ∞ *almost surely.*

PROOF. Using (2.1), (L1) and (L2) have been proved in Lemma A.1. The assertions (L3) and (L4)follow from Assumptions 2-3 and Lemmas A.2-A.3. Noting that

$$\frac{1}{n}\sum_{i=1}^n e_i^2 = \frac{1}{n}\sum_{i=1}^n (e_i^2 - Ee_i^2) + \frac{1}{n}\sum_{i=1}^n Ee_i^2,$$

by Assumption 3, (L5) is a consequence of Lemma A.4. Finally, one can derive (L6) from (L4) and (L5).

PROOF OF THEOREM 2.1. Consider the case that $k \leq k_0$. By (L1)-(L4) of Lemma 2.1 and Cauchy-Schwarz inequality, we have

$$\begin{aligned} G_n^{(1)}(k) - G_n^{(1)}(k_0) &= S_k - S_{k_0} + (k - k_0) C_n S_p/(n-p) \\ &\geq \beta_{k_0}^2 a_1 n + \beta_{k_0} O((n \log\log n)^{1/2}) - (k_0 - k) C_n S_p/(n-p) \quad \text{a.s.} \end{aligned} \tag{2.5}$$

By the condition that $n^{-1}C_n \to 0$ of (2.4) and using (L6) of Lemma 3.1, one shows that

$$G_n^{(1)}(k) - G_n^{(1)}(k_0) > 0 \quad \text{a.s.}$$

Hence

$$\liminf \hat{k}_n \geq k_0 \quad \text{a.s.} \tag{2.6}$$

Then, consider the case $k > k_0$. By (L4) of Lemma 2.1, with probability one, for all large n, we have

$$\begin{aligned} & G_n^{(1)}(k) - G_n^{(1)}(k_0) \\ = \; & (k - k_0)C_n S_p/(n-p) + O(\log\log n) \end{aligned} \tag{2.7}$$

This, together with the condition $C_n/\log\log n \to \infty$ of (2.4)and (L6) of Lemma 3.1, implies that

$$G_n^{(1)}(k) - G_n^{(1)}(k_0) < 0.$$

This proves

$$\limsup \hat{k}_n \leq k_0, \text{ a.s.} \tag{2.8}$$

Combining (2.6) and (2.8), we ultimately obtain

$$\hat{k}_n \to k, \text{ a.s.}$$

Similarly, the second assertion of the theorem can be proved. The proof of Theorem 2.1 is complete.

The following theorem is concerned with the strong consistency of the third criterion. Although its statement is similar to those of the previous criteria, there are some differences in the proof and thus we state and prove it separately.

THEOREM 2.2. *Suppose that the assumptions 1-3 hold for $n = 1, 2, \ldots$ and M_{k_0} is the smallest true model. If C_n satisfies (2.4), then the criterion (3) is strongly consistent.*

PROOF. Note that

$$S_j = \begin{cases} \boldsymbol{\beta}' X_n'(I_n - P_j)X_n\boldsymbol{\beta} + 2\boldsymbol{\beta}' X_n'(I - P_j)\boldsymbol{e}_n + \boldsymbol{e}_n'(I - P_j)\boldsymbol{e}_n, & \text{if } j < k_0, \\ \\ \boldsymbol{e}_n'(I - P_j)\boldsymbol{e}_n, & \text{if } j \geq k_0. \end{cases} \tag{2.9}$$

By (L4)-(L5) of Lemma 2.1, we have, for $1 \leq j \leq p$,

$$\nu^2 + o(1) < S_j/n < a_2|\boldsymbol{\beta}|^2 + \nu^2 + o(1) \quad a.s. \tag{2.10}$$

and

$$\frac{S_j - S_{k_0}}{S_{k_0}} = \begin{cases} > \eta + o_{a.s.}(1), & \text{if } j < k_0, \\ \\ O_{a.s.}(n^{-1}\log\log n), & \text{if } j \geq k_0, \end{cases} \tag{2.11}$$

where $\eta = a_1\beta_{k_0}^2/(a_2|\boldsymbol{\beta}|^2 + \nu^2)$ is a positive constant.

Let $k > k_0$. Then, by (2.10), (2.4) and (2.11), we conclude

$$\begin{aligned} G_n^{(3)}(k) - G_n^{(3)}(k_0) &= n\log\frac{S_k}{S_{k_0}} + (k - k_0)C_n \\ &= n\left[\frac{S_k - S_{k_0}}{S_{k_0}} + o\left(\frac{S_k - S_{k_0}}{S_{k_0}}\right)\right] + (k - k_0)C_n \\ &= O(\log\log n) + (k - k_0)C_n > 0 \quad \text{a.s.} \end{aligned}$$

which implies that

$$\limsup \hat{k}_n \leq k_0 \quad \text{a.s.} \tag{2.12}$$

Next let $k < k_0$. Since $\log(1 + x)$ is an increasing function of x, by (2.11) and (2.4) we have

$$\begin{aligned} G_n^{(3)}(k) - G_n^{(3)}(k_0) &= n\log\frac{S_k}{S_{k_0}} - (k_0 - k)C_n \\ &\geq n\log(1 + \eta + o_{a.s.}(1)) - (k_0 - k)C_n > 0, \text{ a.s.} \end{aligned}$$

which implies that

$$\liminf \hat{k}_n \geq k_0 \quad \text{a.s.} \tag{2.13}$$

The results (2.12) and (2.13) establish the theorem.

3 Data-Oriented Penalty Criteria

In the criteria proposed in Section 2, the selection of C_n is essential. When $C_n = 2$, Criterion (1) reduces to the well known AIC, which has been proved to be inconsistent. Furthermore, the choice $C_n = \log n$, known as the BIC, is a special case of Theorem 2.1., which is strongly consistent. Hannan and Quinn (1979) argued that the minimal choice of C_n to guarantee strong consistency is $c \log\log n$ for some positive constant c. Although this result is not a special case of Theorem 2.1. or 2.2, by using the upper bound in our proofs, results similar to Hannan and Quinn can be obtained. However, this does not suggest an "optimal choice" of C_n in any particular case. In Bai, Krishnaiah and Zhao (1989), it is proved that higher the rate of the order of C_n the better is the performance of the criterion. However, this is only an asymptotic result. Choice of a large C_n usually gives serious underestimation of the order of the model. From the theorems in Section 2, the constant C_n needs only to satisfy the conditions $C_n/n \to 0$ and $C_n/\log\log n \to \infty$ to guarantee strong consistency. However, these conditions do not give any range of the choice of C_n for a given n. In other words, except for the AIC and BIC, the selection of the penalty is not clearly specified. Noting that the AIC is inconsistent and the BIC does not give the best convergence rate of the probability of wrong determination of the model, the problem of optimal selection of the penalty function C_n remains unsolved. Rao and Wu (1989) proposed a data-oriented penalty for model selection in linear models. Later, Chen *et al* (1992) used a data-oriented penalty to select models for AR time series. In this section, we shall further investigate the model selection with data-oriented penalty.

As an example, we consider the Criterion (1). Similar results are true for the other two criteria and the details are omitted. Let a sequence of experimental measurements $\{(y_1, \boldsymbol{x}^{(1)}), \ldots, (y_n, \boldsymbol{x}^{(n)})\}$ be available. Define, for a given integer q with $1 \leq q \leq p$,

$$X_n(q) = (\boldsymbol{x}_{1n} \cdots \boldsymbol{x}_{qn}), \quad \boldsymbol{\beta}(q) = (\beta_1, \ldots, \beta_q)'.$$

If the model M_q is true, it can be written as

$$\boldsymbol{y}_n = X_n(q)\boldsymbol{\beta}(q) + \boldsymbol{e}_n.$$

We shall use the following steps to choose the penalty C_n.

1. Compute any consistent estimate $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_{1n}, \ldots, \tilde{\beta}_{pn})'$ of $\boldsymbol{\beta}$. For example, let $\tilde{\boldsymbol{\beta}}_n$ be the least square estimate of $\boldsymbol{\beta}$ in the model M_p.

2. Compute $\tilde{\sigma}_p^2 = S_p/(n-p)$. Let $\overline{\boldsymbol{\beta}}_n = (\overline{\beta}_{1n}, \ldots, \overline{\beta}_{pn})'$ be defined as follows:

$$\overline{\beta}_{in} = \begin{cases} \tilde{\beta}_{in}, & \text{if } |\tilde{\beta}_{in}| \geq \kappa, \\ \kappa \text{sign}(\tilde{\beta}_{in}), & \text{if } |\tilde{\beta}_{in}| < \kappa, \end{cases} \quad \text{for } i = 1, \ldots, p,$$

   where κ is a constant.

3. Compute $\hat{\boldsymbol{\varepsilon}}_n = \boldsymbol{y}_n - X_n\tilde{\boldsymbol{\beta}}_n$.

4. Let

$$\boldsymbol{u}_n(h) = X_n(h)\overline{\boldsymbol{\beta}}_n(h) + \hat{\boldsymbol{\varepsilon}}_n,$$

   for $h = 1, \ldots, p$. Denote

$$D_n(q,h) = \overline{S}_q(h) - \overline{S}_h(h),$$

   where $S_q(h) = (\boldsymbol{u}_n(h))'(I - P_q)\boldsymbol{u}_n(h)$. It can be shown that $\overline{S}_p(h) = S_p$ if $\overline{\boldsymbol{\beta}}_n = \tilde{\boldsymbol{\beta}}_n$. Define

$$\Delta_{1h} = \min_{q<h}\left\{\frac{D_n(q,h)}{(h-q)\tilde{\sigma}_p^2}\right\},$$
$$\Delta_{2h} = \max_{q>h}\left\{\frac{D_n(q,h)}{(h-q)\tilde{\sigma}_p^2}\right\}.$$

Let $\Delta_h = (\Delta_{1h} + \Delta_{2h})/2$.

5. Define

$$C_n^{(R)} = \frac{\text{average of } \{\Delta_h,\ h = 1, \ldots, p\}}{1 + \sqrt{\lfloor 0.01n \rfloor}},$$

where $\lfloor b \rfloor$ denotes the integer part of b.

Then, C_n is set to be $C_n^{(R)}$.

REMARK. The constant κ used in the definition is determined by the practical requirement on the distinguishability of the regression coefficients from zero. Intuitively, a small choice of it will over estimate the model and vice versa.

We establish the following theorem to show that the procedure is asymptotically consistent.

THEOREM 3.1. *Under the assumptions of Theorem 2.1, with probability one, the Criterion (1) eventually selects the smallest true model if C_n is chosen as $C_n^{(R)}$.*

PROOF. By Theorem 2.1, we need to show that

$$\frac{C_n^{(R)}}{n} \to 0, \quad \text{and}, \quad \frac{C_n^{(R)}}{\log\log n} \to \infty. \tag{3.1}$$

By definition, we have

$$\begin{aligned} D_n(q,h) &= (\boldsymbol{u}_n(h))'(P_h - P_q)\boldsymbol{u}_n(h) \\ &= (X_n(h)\overline{\boldsymbol{\beta}}_n(h) + X_n(k_0)\boldsymbol{\beta}(k_0) - X_n\tilde{\boldsymbol{\beta}}_n + \boldsymbol{e}_n)'(P_h - P_q) \\ &\quad (X_n(h)\overline{\boldsymbol{\beta}}_n(h) + X_n(k_0)\boldsymbol{\beta}(k_0) - X_n\tilde{\boldsymbol{\beta}}_n + \boldsymbol{e}_n). \end{aligned} \tag{3.2}$$

Note that $X_n(k_0)\boldsymbol{\beta}(k_0) = X_n\boldsymbol{\beta}$ and by Lemma 2.1,

$$\tilde{\boldsymbol{\beta}}_n = (\boldsymbol{\beta}(k_0)'\ \mathbf{0}')' + (X_n'X_n)^{-1}X_n'\boldsymbol{e}_n = (\boldsymbol{\beta}(k_0)'\ \mathbf{0}')' + O_{a.s.}(\sqrt{n^{-1}\log\log n}),$$

which implies that

$$X_n(k_0)\beta(k_0) - X_n\tilde{\beta}_n = O_{a.s.}(\sqrt{\log\log n}).$$

Consider the following two cases for each fixed h.

Case 1. $q > h$.

In this case, $(P_h - P_q)X_n(h) = 0$. Then, (3.2) turns out to be

$$\begin{aligned} D_n(q,h) &= -(X_n(k_0)\beta(k_0) - X_n\tilde{\beta}_n + e_n)'(P_q - P_h)(X_n(k_0)\beta(k_0) - X_n\tilde{\beta}_n + e_n) \\ &= -O_{a.s.}(\log\log n). \end{aligned}$$

Note that $D_n(q,h)$ is a negative number of order $O_{a.s.}(\log\log n)$. Thus, Δ_{2h} is a positive number of order $O_{a.s.}(\log\log n)$.

Case 2. $q < h$.

Note that $\overline{\beta}_n(h) = \overline{\beta}(h) + O_{a.s.}(\sqrt{n^{-1}\log\log n})$, where $\overline{\beta}$ is the p-vector whose ith element is $\text{sign}(\beta_i)\max(|\beta_i|, \kappa)$. By Lemma A.1, $n^{-1}\overline{\beta}(h)'X_n(h)'(P_h - P_q)X_n(h)\overline{\beta}(h)$ is bounded away from both zero and infinity. Therefore,

$$D_n(q,h) = \overline{\beta}(h)'X_n(h)'(P_h - P_q)X_n(h)\overline{\beta}(h)(1 + o(1)) \quad \text{a.s.} \tag{3.3}$$

which is positive and has the exact order as n. Combining the both cases, we conclude that $C_n^{(R)}$ has the exact order as $\sqrt{n}$. This shows that (3.1) is true and hence completes the proof of Theorem 3.1.

For the Criteria (2) and (3), similarly defining the data-oriented penalty $C_n^{(R)}$, we can establish results similar to those stated for Criterion (1) in Theorem 3.1.

The small sample behavior of the proposed procedures is studied by Monte Carlo simulation in Section 5.

4 Extensions of the Model Selection Criteria

In Section 2, we discussed the model selection from the p consecutive sub-models $\{M_1, \ldots, M_p\}$ associated with the multiple regression model (1.1). As mentioned there, we actually have 2^p sub-models since each component of β may be zero or not. In this section, we shall extend the model selection for all these possible sub-models. For any true β, rearranging the components of β and the columns of the design matrix X_n, we can get an equivalent regression model whose smallest true model is one of the sub-models $\{M_1, \ldots, M_p\}$. Then, we can apply the criteria introduced in Section 2. Since the assumptions do not change under the rearrangement, the estimated model is still consistent. Select the smallest $\hat{k}$ among the model selections for all rearrangements. However, this approach involves a huge amount of computation if p is large. In fact, there are 2^p residual sum of squares to be computed. Here, we suggest leave one approach (see Rao and Wu (1989)) to select the smallest true model which needs only the computation of $p+1$ residual sum of squares.

For each $1 \leq i \leq p$, denote

$$\beta_{-i} = (\beta_1, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p)'$$

and

$$X_{n,-i} = (\boldsymbol{x}_{1n} \cdots \boldsymbol{x}_{i-1,n} \boldsymbol{x}_{i+1,n} \cdots \boldsymbol{x}_{pn}).$$

Consider the model

$$\boldsymbol{y}_n = X_{n,-i}\beta_{-i} + \boldsymbol{e}_n.$$

Write the corresponding usual residual sum of squares by S_{-i}. Define, for $1 \leq i \leq p$,

$$G_n^{(1)}(-i) = S_{-i} - S_p - C_n S_p/(n-p) \tag{4.1}$$

where C_n may be chosen in accordance with the condition (2.4), or as the random penalty $C_n^{(R)}$ defined in last section.

Then, choose the model as

$$\beta_i = 0 \quad \text{if } G_n^{(1)}(-i) \le 0 \quad \text{and} \quad \beta_i \neq 0 \quad \text{if } G_n^{(1)}(-i) > 0$$
$$i = 1, \ldots, p. \tag{4.2}$$

We now establish the following theorem.

THEOREM 4.1. *Under the conditions of Theorem 2.1, the estimated model by the rule (4.2) is strongly consistent for the smallest true model.*

PROOF. Suppose that in the true model $\beta_i \neq 0$. By (2.5) with $k_0 = p$ and $k = p-1$, (L6) of Lemma 2.1 and (2.4), we have $G_n^{(1)}(-i) > 0$ almost surely. Therefore, with probability one, β_i is taken to be non-zero in the selected model. Conversely, suppose that in the true model $\beta_i = 0$. By (2.7) with $k_0 = p - 1$ and $k = p$, (L4) and (L6) of Lemma 2.1 and (2.4), we have $G_n^{(1)}(-i) < 0$ almost surely, which implies that with probability one, β_i is excluded in the selected model. This completes the proof of the theorem.

Similar to (4.1), one may define for each $1 \le i \le p$,

$$G_n^{(2)}(-i) = S_{-i} - S_p - C_n,$$

or

$$G_n^{(3)}(-i) = n(\log S_{-i} - \log S_p) - C_n,$$

respectively. Then choose the model by letting

$$\beta_i = 0 \quad \text{if } G_n^{(j)}(-i) \le 0 \quad \text{and} \quad \beta_i \neq 0 \quad \text{if } G_n^{(j)}(-i) > 0$$
$$i = 1, \ldots, p,$$

$j = 2$ or 3.

Under the conditions of Theorem 2.1, one can show that with probability one these

criteria will eventually select the smallest true model. The proofs are similar to those of Theorems 2.2 and 4.1, and thus are omitted.

5 Monte Carlo Study

In this section, by computer simulations, we verify the small-sample performance of the model selection rules proposed in this paper. The regression model is assumed to be:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + e_i, \quad i = 1, \ldots, n,$$

where $x_{1i}, \ldots, x_{5i}$, $i = 1, \ldots, n$, are iid $N(0,1)$ random variables. In the simulations, κ is set to be 0.01. In Tables 4.1 and 4.2, $e_1, \ldots, e_n$ are chosen to be independently distributed as $N(0, u^2)$ where u is a discrete random variable uniformly distributed within $\{1, \ldots, 5\}$. In Tables 4.3, 4.4 and 4.5, $e_1, \ldots, e_n$ are chosen to be independent and identically distributed as $N(0,1)$ random variables. In the tables, RC(1) denotes C(1) with the use of $C_n^{(R)}$ of Section 3 as the choice of C_n and the numbers shown in the tables are the counts of the correct selection of the smallest true model based on 1,000 replications. In simulation, IMSL subroutines DRNNOF and RNUND were used to generate the random numbers.

From the Table 4.1, it is seen that with the same C_n, the criterion C(1) is superior to the others and that the RC(1) is comparable with C(1). The criteria AIC, SW and HQ based on Akaike (1970), Schwarz (1978) and Hannan & Quinn(1979) respectively, do not perform as well as $C(1)$ and $RC(1)$. Table 4.2 shows that for the general multiple regression model, the performance of RC(1) is very good, absolutely superior to all the others. Comparing Tables 4.1 and 4.2, one finds that the criterion C(1) with $C_n = 5(\log n)^3$ performs for the two models quite differently but the performance of RC(1) is very stable for different models. Comparing Table 4.3 with Table 4.4, it can be seen that in either case, RC(1) shows a very good performance. From Tables 4.3 and 4.5, it can be seen that for different signal-to-noise

ratios, the performance of C(1) depends on the choice of C_n but RC(1) automatically adapts to the optimal choice of C_ns for different signal-to-noise ratios.

Table 4.1 $C_n = 5(\log n)^3$ and $\beta = (6\ 3\ 7\ 0\ 0)'$

Sample size	C(1)	C(2)	C(3)	RC(1)	AIC	SW	HQ
15	993	973	876	923	683	714	592
20	998	975	917	951	705	743	630
25	1,000	975	924	969	752	786	683
30	1,000	975	918	982	726	769	667
35	1,000	981	935	992	747	792	678
40	1,000	978	935	995	769	804	698
45	1,000	985	940	1,000	767	819	713
50	1,000	976	926	997	741	779	682

Table 4.2 $\beta = (6\ 3\ 0\ 0\ 7)'$

Sample size	15	20	25	30	35	40	45	50
C(1) $5(\log n)^3$	23	11	23	48	80	95	130	251
C(1) $4(\log n)^2$	293	287	464	625	754	824	876	955
C(1) $(\log n)^3$	322	182	191	212	221	186	181	273
RC(1)	801	792	897	955	967	960	946	986

Table 4.3 $\beta = (6\ 3\ 0\ 0\ 7)'$

Sample size	15	20	25	30	35	40	45	50
C(1) $5(\log n)^3$	946	995	1,000	1,000	1,000	1,000	1,000	1,000
C(1) $\log n$	752	737	747	740	731	773	742	733
C(2) $5(\log n)^3$	999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
C(2) $\log n$	786	773	764	763	770	782	753	734
RC(1)	998	999	1,000	1,000	1,000	999	1,000	1,000

Table 4.4 $\beta = (6\ 0\ 3\ 7\ 0)'$

Sample size	15	20	25	30	35	40	45	50
C(1) $5(\log n)^3$	557	983	995	1,000	1,000	1,000	1,000	1,000
C(1) $\log n$	699	718	708	720	736	770	729	763
C(2) $5(\log n)^3$	524	1,000	1,000	1,000	1,000	1,000	1,000	1,000
C(2) $\log n$	743	748	739	747	754	778	747	767
RC(1)	470	963	975	1,000	1,000	999	1,000	1,000

Table 4.5 $\beta = (1.2\ 1.5\ 0\ 0\ 1.3)'$

Sample size	15	20	25	30	35	40	45	50
C(1) $5(\log n)^3$	34	24	86	184	310	388	515	596
C(1) $\log n$	752	737	747	740	731	773	742	733
C(2) $5(\log n)^3$	0	0	3	44	209	320	478	581
C(2) $\log n$	786	773	764	763	770	782	753	734
RC(1)	912	923	980	994	992	996	993	994

6 Discussions and Conclusions

To remedy the inconsistency of AIC, various criteria were proposed in the literature. The cross-validation has been proved to be equivalent to the AIC. Most other criteria use a fixed choice of the penalty function C_n such that $c \log\log n \leq C_n = o(n)$, for some constant $c > 0$, to guarantee strong consistency. However, a fixed choice may be good in some situations and bad in some other situations. As shown in our simulation, the criterion with a data-oriented penalty has some advantages.

7 Appendix. Preliminary Lemmas

Denote the eigenvalues of a symmetric matrix A of order k by $\lambda_1(A) \geq \ldots \geq \lambda_k(A)$. The following lemmas are used in the proofs of the main results.

LEMMA A.1. *Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$ be n-vectors and denote $W_i = B_i' B_i$ where*

$$B_i = (\boldsymbol{b}_1 \cdots \boldsymbol{b}_i), \quad i = 1, \ldots, p.$$

If there exist constants η_1 and η_2 such that

$$0 < \eta_1 \leq \lambda_p(W_p) \leq \lambda_1(W_p) \leq \eta_2,$$

then

$$\begin{aligned}
&(1) \quad \eta_1 \leq \boldsymbol{b}_i' \boldsymbol{b}_i \leq \eta_2, \quad 1 \leq i \leq p, \\
&(2) \quad \eta_1 \leq \boldsymbol{b}_i' Q_{i-1} \boldsymbol{b}_i \leq \eta_2, \quad 1 \leq i \leq p, \\
&(3) \quad \eta_1 < \lambda_{i-j}(B_i'(P_i - P_j)B_i) \leq \lambda_1(B_i'(P_i - P_j)B_i) \leq \eta_2, \quad j < i,
\end{aligned} \tag{A.1}$$

where P_i is the projection matrix onto the space spanned by $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_i$ and $Q_i = I - P_i$.

PROOF. For any vector $\boldsymbol{x}$ such that $\boldsymbol{x}'\boldsymbol{x} = 1$, we have

$$\eta_1 \leq \lambda_p(W_p) \leq \boldsymbol{x}'W_p\boldsymbol{x} \leq \lambda_1(W_p) \leq \eta_2.$$

Then the result (i) follows by choosing $\boldsymbol{x}' = (0, \ldots, 0, 1, 0, \ldots, 0)$ where the number 1 is in the i-th position.

By the interlace theorem (see Sturmian Separation Theorem in Rao (1973, page 64)),

$$\lambda_j(W_i) \geq \lambda_j(W_{i-1}) \geq \lambda_{j+1}(W_i), \quad j = 1, \ldots, i-1. \tag{A.2}$$

Note that

$$\boldsymbol{b}_i'Q_{i-1}\boldsymbol{b}_i = \frac{|W_i|}{|W_{i-1}|} = \frac{\lambda_1(W_i)\cdots\lambda_i(W_i)}{\lambda_1(W_{i-1})\cdots\lambda_{i-1}(W_{i-1})}$$

so that by (A.2)

$$\lambda_i(W_i) \leq \boldsymbol{b}_i'Q_{i-1}\boldsymbol{b}_i \leq \lambda_1(W_i).$$

The assertion (2) then follows, since, using (A.2) once again

$$\lambda_p(W_p) \leq \lambda_i(W_i) \quad \text{and} \quad \lambda_1(W_i) \leq \lambda_1(W_p) \quad \text{for } i \leq p.$$

Since $\lambda_k((I - P_j)B_iB_i') = \lambda_k(B_i'(I - P_j)B_i)$ and $\lambda_k(B_iB_i') = \lambda_k(B_i'B_i)$, for $k = 1, \ldots, i$, by the interlace theorem, it follows that

$$\lambda_i(B_i'B_i) \leq \lambda_{i-j}(B_i'(P_i - P_j)B_i) \leq \lambda_1(B_i'(P_i - P_j)B_i) \leq \lambda_1(B_i'B_i)$$

which, together with (A.2), implies the conclusion (3).

LEMMA A.2. *Let $X_n = (\boldsymbol{x}_{1n}\cdots\boldsymbol{x}_{kn})$, where $\boldsymbol{x}_{in}$'s are n-vectors. Assume that $\boldsymbol{e}_n$'s are n-dimensional random vectors, $n = 1, 2, \ldots$, such that*

$$\boldsymbol{x}_{in}'\boldsymbol{e}_n = O(n \log\log n)^{1/2}, \quad a.s., \quad 1 \leq i \leq k \tag{A.3}$$

and

$$0 < cn \leq \lambda_k(X_n'X_n). \tag{A.4}$$

Then

$$e_n'P_ne_n = O(\log\log n), \quad a.s.$$

where $P_n = X_n(X_n'X_n)^{-1}X_n'$.

PROOF. Let γ_{in} be the i-th eigenvector of $X_n'X_n$ and $\Delta_n = \mathrm{diag}(\lambda_1(X_n'X_n), \ldots, \lambda_k(X_n'X_n))$. Then the (i,j)-th element of $(X_n'X_n)^{-1}$ is

$$\gamma_{in}'\Delta_n^{-1}\gamma_{jn} = O(n^{-1})$$

using the condition (A.4).

Now by (A.3) and (A.4), it follows that

$$e_n'P_ne_n = e_n'X_n(X_n'X_n)^{-1}X_n'e_n = O(\log\log n)$$

since each component of $e_n'X_n$ is $O((n\log\log n)^{1/2})$ and each element of $(X_n'X_n)^{-1}$ is $O(n^{-1})$. The lemma is proved.

LEMMA A.3. *Let* $\varepsilon_1, \varepsilon_2, \ldots$ *be a sequence of independent variables with zero mean such that* $0 < \nu^2 \leq E(\varepsilon_i^2) = \sigma_i^2$ *and* $E(|\varepsilon_i|^3) \leq \tau^3 < \infty$ *for* $i = 1, 2, \ldots$. *If* $a_1, a_2, \ldots$ *is a sequence of constants such that*

(I) $\qquad A_n = \sum_{i=1}^n a_i^2 \to \infty, \quad as\ n \to \infty;$

(II) $\quad \sum_{i=1}^n |a_i|^3 = O(A_n^3(\log A_n^2)^{-(1+\delta)}), \quad for\ some\ \delta > 0,$

then, almost surely,

$$\sum_{i=1}^n a_i\varepsilon_i = O(A_n^2\log\log A_n^2)^{1/2}). \tag{A.5}$$

PROOF. Let $B_n^2 = \sum_{i=1}^n \sigma_i^2 a_i^2$ and let F_n and Φ denote the distributions of $B_n^{-1} \sum_{i=1}^n a_i \varepsilon_i$ and the standard normal random variable respectively. Since $0 < \nu^2 \leq \sigma_i^2$ and $E(|\varepsilon_i|^3) \leq \tau^3$ for $i = 1, 2, \ldots$, by the Theorem 3 of Petrov (1975, page 111) and Assumption (II), we have, for some constant $M > 0$,

$$\begin{aligned} \sup_x |F_n(x) - \Phi(x)| &\leq M B_n^{-3} \textstyle\sum_{i=1}^n |a_i|^3 E|\varepsilon_i|^3 \\ &= O(A_n^{-3} \textstyle\sum_{i=1}^n |a_i|^3) = O((\log A_n^2)^{-1-\delta}). \end{aligned} \tag{A.6}$$

Now from Assumptions (I) and (II), it follows that

$$\frac{A_{n-1}^2}{A_n^2} = 1 - \frac{\sigma_n^2 a_n^2}{A_n^2} \to 1. \tag{A.7}$$

By Assumption (I), (A.6) and (A.7), (A.5) follows from Theorem 3 of Petrov (1975, page 305).

LEMMA A.4. *Suppose that $\xi_1, \xi_2, \ldots$ are independently distributed random variables with zero means and bounded $(1+\delta)$th moments for some $\delta > 0$. Then*

$$\frac{1}{n} \sum_{i=1}^n \xi_i \to 0 \quad a.s.$$

A proof of this lemma can be found in Chung (1974).

References

[1] Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203-217.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Second International Symposium on Information Theory, B. N. Petrov and F. Czàki eds., Akademiai Kiadò, Budapest, 267-281.

[3] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control* AC-19, 716-723.

[4] Atkinson, A. C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika 67*, 413-418.

[5] Bai, Z. D., Krishnaiah, P. R. and Zhao, L. C. (1989). On rates of convergence of efficient detection criteria in signal processing with white noise. *IEEE Trans. Inform. Theory* **35**, 380-388.

[6] Chen, C. H., Davis, R. A., Brockwell, P. J. and Bai, Z. D. (1993). Order determination for autoregressive processes using resampling methods. *Statistica Sinica*, Vol. 3, No. 2, pp 481-500.

[7] Chung, K. L. (1974). *A Course in Probability Theory*, 2nd ed. Academic Press Inc., London.

[8] Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.

[9] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.

[10] Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70**, 320-328.

[11] Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc., Ser. B* **41**, 190-195.

[12] Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

[13] Petrov, V. V. (1975). *Sum of Independent Random Variables.* Springer-Verlag, Berlin.

[14] Picard and Cook (1984). Cross-validation of regression models. *J. Amer. Statist. Assoc.* **79**, 575-583.

[15] Rao, C.R. (1987). Prediction of future observations in growth curve type models. *J. Statistical Science* **2**, 437-471.

[16] Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76**, 369–374.

[17] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

[18] Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* **71**, 43-49.

[19] Shibata, R. (1986). Selection of the number of regression variables; a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.* **38**, 459-474.

[20] Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

[21] Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. *J. Roy. Statist. Soc. Ser. B.* **36**, 111-133.

[22] Stone, M. (1977a). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B.* **39**, 44-47.

[23] Stone, M. (1977b). Asymptotics for and against cross-validation. *Biometrika* **64**, 29-38.

[24] Zhao, L. C., Krishnaiah, P. R. and Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.* **20**, 1-25.