

Model selection with data-oriented penalty

Z.D. Bai^a, C.R. Rao^{b,*}, Y. Wu^{c,2}

^aDepartment of Applied Mathematics, National Sun Yat-sen University, Taiwan

^bDepartment of Statistics, Penn State University, 326 Thomas Building, University Park,
PA 16802-2111, USA

^cDepartment of Mathematics and Statistics, York University, Canada

Received 15 April 1997; accepted 8 July 1998

Abstract

We consider the problem of model (or variable) selection in the classical regression model using the GIC (general information criterion). In this method the maximum likelihood is used with a penalty function denoted by C_n , depending on the sample size n and chosen to ensure consistency in the selection of the true model. There are various choices of C_n suggested in the literature on model selection. 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 provides improved performance over a fixed choice of C_n . © 1999 Elsevier Science B.V. All rights reserved.

AMS classifications: 62J05; 62F10

Keywords: AIC; GIC; Linear regression; Model selection; Variables selection

1. Introduction

Consider the multiple regression model

$$y_n = X_n \beta + e_n, \quad (1.1)$$

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. The components of e_n are independently distributed but not required to have the same distribution. Each component of β may be zero or nonzero. Each subset \mathcal{M} of $\{1, 2, \dots, p\}$ 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.

* Corresponding author.

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.

Many model selection procedures have been proposed in the statistical literature for choosing the smallest true model for multiple regression. References to earlier work can be found in papers by Akaike (1970,1973,1974), Schwarz (1978), Hannan and Quinn (1979), Atkinson (1980), Shibata (1984,1986), Rao and Wu (1989), Shao (1993) and others. Some recent contributions by Bozdogan (1988) and Zheng and Loh (1995) provide new approaches to the problem, somewhat different from the earlier methods based on cross validation, prediction error and information criterion such as AIC, BIC and GIC.

Our object in the present paper is to pursue the investigation started in Rao and Wu (1989) and make some refinements, by withdrawing the restrictions placed on the error terms and by providing a data based penalty term in the GIC (general information criterion). We believe that our proposal will provide an objective procedure for estimating the model to be used in regression prediction problems.

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

Assumption 1. There are constants a_1 and a_2 such that

$$0 < a_1 n \leq \lambda_p(X_n' X_n) \leq \lambda_1(X_n' X_n) \leq a_2 n, \quad (2.1)$$

where $\lambda_i(X_n' X_n)$ is the i th eigenvalue of $X_n' X_n$.

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

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

where x_{in}^j is the j th component of $\mathbf{x}_{in} = (x_{in}^1, \dots, x_{in}^n)'$.

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

$$0 < v^2 \leq E(e_i^2), \quad E(|e_i|^3) \leq \tau^3 < \infty \quad (2.3)$$

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

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

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

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

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

where C_n is a function of n satisfying the conditions

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

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

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

Lemma 2.1. *Suppose that assumptions 1–3 hold for $n = 1, 2, \dots$, then*

$$(L1) \ a_2 n \geq \mathbf{x}'_m \mathbf{x}_m \geq a_1 n, \text{ as } n \rightarrow \infty, \quad 1 \leq i \leq p;$$

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

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

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

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

$$(L6) \ S_p / (n - p) \text{ is bounded away from 0 and } \infty \text{ almost surely.}$$

Proof. Using Eq. (2.1), (L1) and (L2) have been proved in Lemma A.1. The assertions (L3) and (L4) follow from assumptions 2 and 3 and Lemmas A.2 and 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 $n^{-1}C_n \rightarrow 0$ of Eq. (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

$$G_n^{(1)}(k) - G_n^{(1)}(k_0) = (k - k_0)C_n S_p / (n - p) + O(\log \log n). \tag{2.7}$$

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

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

This proves

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

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

$$\hat{k}_n \rightarrow k \quad \text{a.s.}$$

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 assumptions 1–3 hold for $n = 1, 2, \dots$ and M_{k_0} is the smallest true model. If C_n satisfies Eq. (2.4), then criterion (3) is strongly consistent.*

Proof. Note that

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

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

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

and

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

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

Let $k > k_0$. Then, by Eqs. (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 Eqs. (2.11) and (2.4) we have

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

which implies that

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

Results (2.12) and (2.13) establish the theorem.

3. Data-oriented penalty criteria

In the criteria proposed in Section 2, the penalty function C_n is required to satisfy the conditions $C_n/n \rightarrow 0$ and $C_n/\log \log n \rightarrow \infty$. The actual choice of C_n is not specified. In the statistical literature, some fixed choices of C_n have been suggested such as $C_n = 2$ by Akaike (1970,1973,1974), and $C_n = c \log \log n$ for some $c > 2$ by Hannan and Quinn (1979). Some comments on the choice of C_n have been made by Bai et al. (1989) and Zhao et al. (1986). The first attempt to provide a data-oriented penalty function is made in Rao and Wu (1989), which was applied to model selection problems in linear models. Later, Chen et al. (1993) used a data-oriented penalty to select models for the

AR time series. In this section we make some further investigations in the selection of a 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.

For the regression model (1.1), let a sequence of experimental measurements $\{(y_1, \mathbf{x}^{(1)}), \dots, (y_n, \mathbf{x}^{(n)})\}$ be available. Define, for a given integer q with $1 \leq q \leq p$,

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

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

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

In the following algorithm, we propose a data-oriented procedure to select the penalty C_n and then prove that the proposed procedure asymptotically satisfies the conditions of Theorems 3.1–3.2, while it works well for moderate sample sizes.

1. Compute any consistent estimate $\tilde{\boldsymbol{\beta}}_n = (\tilde{\beta}_{1n}, \dots, \tilde{\beta}_{pn})'$ of $\boldsymbol{\beta}$ and $\tilde{\sigma}_p^2 = S_p/(n-p)$, where S_p is the residual sum of squares. For instance, $\tilde{\boldsymbol{\beta}}_n$ can be chosen to be the least square estimate of $\boldsymbol{\beta}$.
2. Compute $\hat{\mathbf{e}}_n = \mathbf{y}_n - X_n \tilde{\boldsymbol{\beta}}_n$.
3. Let $\bar{\boldsymbol{\beta}}_n = (\bar{\beta}_{1n}, \dots, \bar{\beta}_{pn})'$ be defined as follows:

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

where the constant κ is a suitably chosen threshold value.

4. Let

$$\mathbf{u}_n(h) = X_n(h)\bar{\boldsymbol{\beta}}_n(h) + \hat{\mathbf{e}}_n, \quad h = 1, \dots, p,$$

where $\bar{\boldsymbol{\beta}}_n(h)$ is a vector of the first h components of $\bar{\boldsymbol{\beta}}_n$. Compute

$$D_n(q, h) = \bar{S}_q(h) - \bar{S}_h(h), \quad q = 0, 1, \dots, p,$$

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

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

$$A_{2h} = \max_{q > h} \left\{ \frac{D_n(q, h)}{(h - q)\tilde{\sigma}_p^2} \right\},$$

where the maximum is 0 if the selection set is empty. Let $\Delta_h = (A_{1h} + A_{2h})/2$.

5. Define

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

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

Then choose $C_n^{(R)}$ as the penalty C_n .

We comment here that the proposed data-oriented model selection procedure is asymptotically equivalent to criterion (1) and its performance for small sample sizes is better than procedure (1). The former is shown in the following Theorem 3.1 and the latter is demonstrated by our Monte Carlo study given in Section 5.

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

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

By definition, we have

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

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

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

which implies that

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

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, Eq. (3.2) turns out to be

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

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

Case 2: $q < h$.

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

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

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 Eq. (3.1) is true and hence completes the proof of Theorem 3.1.

For 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 $\{M_1, \dots, M_p\}$ 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 $\{M_1, \dots, M_p\}$. 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 $p + 1$ sums of residual squares.

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

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

and

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

Consider the model

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

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 condition (2.4), or as the random penalty $C_n^{(R)}$ defined in last section.

Then, choose the model as

$$\beta_i = 0 \text{ if } G_n^{(1)}(-i) \leq 0 \text{ and } \beta_i \neq 0 \text{ if } G_n^{(1)}(-i) > 0$$

$$i = 1, \dots, p. \quad (4.2)$$

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 Eq. (2.5) with $k_0 = p$ and $k = p - 1$, (L6) of Lemma 2.1 and Eq. (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 Eq. (2.7) with $k_0 = p - 1$ and $k = p$, (L4) and (L6) of Lemma 2.1 and Eq. (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. \square

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

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

or

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

respectively. Then choose the model by letting

$$\beta_i = 0 \text{ if } G_n^{(j)}(-i) \leq 0 \text{ and } \beta_i \neq 0 \text{ if } G_n^{(j)}(-i) > 0,$$

$$i = 1, \dots, p,$$

$j = 2$ 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

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

where x_{1i}, \dots, x_{5i} , $i = 1, \dots, n$, are iid $N(0, 1)$ random variables in the examples of Tables 1–5 and $(x_{1i}, \dots, x_{5i})'$ is distributed according to $N(\mathbf{0}, A)$, where $A = (a_{ij})$ with

Table 1

$$C_n = 5(\log n)^3 \text{ and } \beta = (6 \ 3 \ 7 \ 0 \ 0)'$$

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	1000	975	924	969	752	786	683
30	1000	975	918	982	726	769	667
35	1000	981	935	992	747	792	678
40	1000	978	935	995	769	804	698
45	1000	985	940	1000	767	819	713
50	1000	976	926	997	741	779	682

Table 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 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	1000	1000	1000	1000	1000	1000
C(1) $\log n$	752	737	747	740	731	773	742	733
C(2) $5(\log n)^3$	999	1000	1000	1000	1000	1000	1000	1000
C(2) $\log n$	786	773	764	763	770	782	753	734
RC(1)	998	999	1000	1000	1000	999	1000	1000

Table 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	1000	1000	1000	1000	1000
C(1) $\log n$	699	718	708	720	736	770	729	763
C(2) $5(\log n)^3$	524	1000	1000	1000	1000	1000	1000	1000
C(2) $\log n$	743	748	739	747	754	778	747	767
RC(1)	470	963	975	1000	1000	999	1000	1000

$a_{ii} = 1$, for $i = 1, \dots, 5$ and $A_{ij} = 0.25$ for $i \neq j$ in the example of Table 6. In the simulations, κ is set to be 0.01. For Tables 1 and 2, e_1, \dots, e_n are chosen to be independently distributed as $N(0, u^2)$ where u is a discrete random variable uniformly distributed within $\{1, \dots, 5\}$. For Tables 3–7, e_1, \dots, 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 1000 replications. In Table 5, the index is equal to -1 if the selected model is not a true

Table 5

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

Sample size	Index	15	20	25	30	35	40	45	50
C(1) $5(\log n)^3$	-1	966	976	914	816	690	612	485	404
	0	34	24	86	184	310	388	515	596
	1	0	0	0	0	0	0	0	0
C(1) $\log n$	-1	0	0	0	0	0	0	0	0
	0	752	737	747	740	731	773	742	733
	1	247	263	253	260	269	227	258	267
C(2) $5(\log n)^3$	-1	1000	1000	997	956	791	680	522	419
	0	0	0	3	44	209	320	478	581
	1	0	0	0	0	0	0	0	0
C(2) $\log n$	-1	0	0	0	0	0	0	0	0
	0	786	773	764	763	770	782	753	734
	1	214	227	236	237	230	218	247	266
RC(1)	-1	85	75	20	6	8	4	7	6
	0	912	923	980	994	992	996	993	994
	1	2	1	0	0	0	0	0	0

Table 6

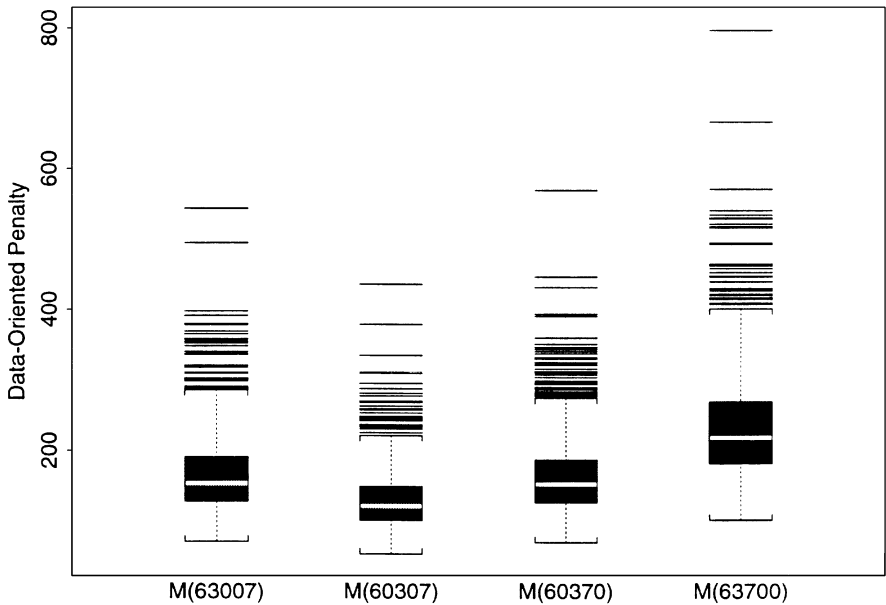
$$\beta = (6 \ 3 \ 0 \ 0 \ 7)', \quad \text{var}(x_{fi}) = 1, \quad \text{Cov}(x_{fi}, x_{fj}) = 0.25$$

Sample size	15	20	25	30	35	40	45	50
C(1) $5(\log n)^3$	312	489	601	848	934	996	996	1000
C(1) $\log n$	704	723	729	744	754	760	753	757
C(2) $5(\log n)^3$	128	411	613	936	980	1000	1000	1000
C(2) $\log n$	751	760	765	769	780	762	769	759
RC(1)	551	866	744	982	986	1000	1000	1000

model and has one parameter less than the smallest true model; the index is equal to zero if the smallest true model is selected; the index is equal 1 if the selected model is a true model and has one parameter more than the smallest true model. In Table 7, M(63007) stands for the model with $\beta = (6 \ 3 \ 0 \ 0 \ 7)'$; M(60307) stands for the model with $\beta = (6 \ 0 \ 3 \ 0 \ 7)'$; M(60370) stands for the model with $\beta = (6 \ 0 \ 3 \ 7 \ 0)'$; M(63700) stands for the model with $\beta = (6 \ 3 \ 7 \ 0 \ 0)'$ and *Data-Oriented Penalty* stands for $C_n^{(R)}$. In simulation, IMSL subroutines DRNNOF and RNUND were used to generate the random numbers.

From the Table 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 and Quinn (1979) respectively, do not perform as well as C(1) and RC(1). Table 2 shows that for the general multiple regression model, the performance of RC(1) seems to be good, absolutely superior to all the others. Comparing Tables 1 and 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 3 with Table 4, it can be seen that in either case, RC(1) shows good performance. From

Table 7
Boxplot



Tables 3 and 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_n 's for different signal-to-noise ratios. It can also be observed from Table 5 that using $C_n = 5(\log n)^3$ gives more underestimation while using $C_n = \log n$ gives more overestimation, but using $C_n^{(R)}$ the bias is at low level. Table 6 again shows the good performance of $C_n^{(R)}$. Table 7 gives the boxplot of $C_n^{(R)}$ for four models when the sample size is 30.

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 may not perform well in some other situations. As shown in our simulation, the criterion with a data-oriented penalty has some advantages.

Appendix. Preliminary lemmas

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

Lemma A.1. Let $\mathbf{b}_1, \dots, \mathbf{b}_p$ be n -vectors and denote $W_i = B'_i B_i$ where

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

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

- (1) $\eta_1 \leq \mathbf{b}'_i \mathbf{b}_i \leq \eta_2, \quad 1 \leq i \leq p,$
- (2) $\eta_1 \leq \mathbf{b}'_i Q_{i-1} \mathbf{b}_i \leq \eta_2, \quad 1 \leq i \leq p,$
- (3) $\eta_1 < \lambda_{i-j}(B'_i(P_i - P_j)B_i) \leq \lambda_1(B'_i(P_i - P_j)B_i) \leq \eta_2, \quad j < i,$ (A.1)

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

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

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

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

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

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

Note that

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

so that by Eq. (A.2)

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

Assertion (2) then follows, since, using Eq. (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 $\lambda_k((I - P_j)B_i B'_i) = \lambda_k(B'_i(I - P_j)B_i)$ and $\lambda_k(B_i B'_i) = \lambda_k(B'_i B_i)$, for $k = 1, \dots, i$, 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 Eq. (A.2), implies conclusion (3).

Lemma A.2. Let $X_n = (\mathbf{x}_{1n} \cdots \mathbf{x}_{kn})$, where \mathbf{x}_m 's are n -vectors. Assume that \mathbf{e}_n 's are n -dimensional random vectors, $n = 1, 2, \dots$, such that

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

and

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

Then

$$e'_n P_n e_n = O(\log \log n) \quad a.s.$$

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 $\Delta_n = \text{diag}(\lambda_1(X'_n X_n), \dots, \lambda_k(X'_n X_n))$. Then the (i, j) th element of $(X'_n X_n)^{-1}$ is

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

using condition (A.4).

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

$$e'_n P_n e_n = e'_n X_n (X'_n X_n)^{-1} X'_n e_n = O(\log \log n),$$

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

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

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

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 $0 < v^2 \leq \sigma_i^2$ and $E(|\varepsilon_i|^3) \leq \tau^3$ for $i = 1, 2, \dots$, by the Theorem 3 of Petrov (1975) (p. 111) and Assumption (II), we have, for some constant $M > 0$,

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

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

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

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

Lemma A.4. *Suppose that ξ_1, ξ_2, \dots are independently distributed random variables with zero means and bounded $(1 + \delta)$ th moments for some $\delta > 0$. Then*

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

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

References

- Akaike, H., 1970. Statistical predictor identification. *Ann. Inst. Statist. Math.* 22, 203–217.
- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: B.N. Petrov, F. Czàki (Eds.), 2nd Int. Symp. on Information Theory. Akademiai Kiadó, Budapest, pp. 267–281.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* AC-19, 716–723.
- Atkinson, A.C., 1980. A note on the generalized information criterion for choice of a model. *Biometrika* 67, 413–418.
- Bai, Z.D., Krishnaiah, P.R., 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.
- Bozdogan, H., 1988. ICOMP: A new model selection criterion. In: Bock, H.H. (Ed.), *Classification and Related Methods of Data Analysis*. North-Holland, Amsterdam, pp. 599–608.
- Chen, C.H., Davis, R.A., Brockwell, P.J., Bai, Z.D., 1993. Order determination for autoregressive processes using resampling methods. *Statistica Sinica* 3(2) 481–500.
- Chung, K.L., 1974. *A Course in Probability Theory*, 2nd ed. Academic Press Inc., London.
- Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. *J. Roy. Statist. Soc., Ser. B* 41, 190–195.
- Petrov, V.V., 1975. *Sum of Independent Random Variables*. Springer, Berlin.
- Rao, C.R., 1973. *Linear Statistical Inference and Its Applications*. Wiley, New York.
- Rao, C.R., Wu, Y., 1989. A strongly consistent procedure for model selection in a regression problem. *Biometrika* 76, 369–374.
- Schwarz, G., 1978. Estimating the dimension of a model. *Ann. Statist.* 6, 461–464.
- Shibata, R., 1984. Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* 71, 43–49.
- Shibata, R., 1986. Selection of the number of regression variables; a minimax choice of generalized FPE. *Ann. Inst. Statist. Math.* 38, 459–474.
- Shao, J., 1993. Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* 88, 486–494.
- Zhao, L.C., Krishnaiah, P.R., Bai, Z.D., 1986. On detection of the number of signals in presence of white noise. *J. Multivariate Anal.* 20, 1–25.
- Zheng, X., Loh, W., 1995. Consistent variable selection in linear models. *J. Am. Statist. Assoc.* 90, 151–156.