ON THE SELECTION OF REGRESSION VARIABLES*

AN HONGZHI (安鸿志) and GU LAN (顾 岚)
(Institute of Applied Mathematics, Academia Sinica)

Abstract

The methods to minimize AIC or BIC criterion function for selection of regression variables are considered. The main calculations of some of these methods are completed economically and recursively. The methods are shown to be of strong consistency or overconsistency to the true model.

§ 1. Introduction

In regression problem one may consider the dependent variables $y(t)$ at time t as a linear function of P possible independent variables $x_1(t)$, $x_2(t)$, $\cdots$, $x_P(t)$. So the model can be written as

$$y(t)=\alpha_1 x_1(t)+\alpha_1 x_2(t)+\cdots+\alpha_P x_P(t)+\varepsilon(t), \quad t=1, 2, \cdots, T, \tag{1.1}$$

where $\varepsilon(t)$ is i.i.d. series with zero mean and variance σ^2. From (1.1) we have the following model

$$Y=X\alpha+\varepsilon, \tag{1.2}$$

where

$$Y=\begin{Bmatrix} y(1) \\ y(2) \\ \vdots \\ y(T) \end{Bmatrix}, \quad \alpha=\begin{Bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_P \end{Bmatrix}, \quad \varepsilon=\begin{Bmatrix} \varepsilon(1) \\ \varepsilon(2) \\ \vdots \\ \varepsilon(T) \end{Bmatrix},$$

$$X=\begin{Bmatrix} x_1(1) & x_2(1) & \cdots & x_P(1) \\ x_1(2) & x_2(2) & \cdots & x_P(2) \\ \vdots & \vdots & & \vdots \\ x_1(T) & x_2(T) & \cdots & x_P(T) \end{Bmatrix}=(x_1, x_2, \cdots, x_P).$$

Here x_i is ith column vector of matrix X. If the true model is such that $\alpha_i\neq 0$ for $i=i_1, i_2, \cdots, i_p$ but $\alpha_i=0$ for other i's, then, to fit a regression model means to choose regression variables from all the possible variables $x_1(t)$, $x_2(t)$, $\cdots$, $x_P(t)$ in model (1.1), that is, to estimate p and $i_1, i_2, \cdots, i_p$ based on the observations of Y and X, and to estimate the regression coefficients $\alpha_{i_1}, \alpha_{i_2}, \cdots, \alpha_{i_p}$.

If $i_1, i_2, \cdots, i_p$ are given, then (1.1) can be replaced by

$$y(t)=\alpha_{i_1}x_{i_1}(t)+\alpha_{i_2}x_{i_2}(t)+\cdots+\alpha_{i_p}x_{i_p}(t)+\varepsilon(t), \quad t=1, 2, \cdots T. \tag{1.3}$$

So we can write it as

* Received November 19, 1983.

$$Y=X(I_p)\alpha(I_p)+\varepsilon, \tag{1.4}$$

where $I_p=\{i_1, i_2, \cdots, i_p\}$ is the index set which corresponds to the variables in (1.3), and

$$X(I_p)=\begin{Bmatrix} x_{i_1}(1) & x_{i_2}(1) & \cdots & x_{i_p}(1) \\ x_{i_1}(2) & x_{i_2}(2) & \cdots & x_{i_p}(2) \\ \vdots & \vdots & & \vdots \\ x_{i_1}(T) & x_{i_2}(T) & \cdots & x_{i_p}(T) \end{Bmatrix}=(x_{i_1}, x_{i_2}, \cdots, x_{i_p}),$$

$$\alpha(I_p)=(\alpha_{i_1}, \alpha_{i_2}, \cdots, \alpha_{i_p})'.$$

Suppose $X'(I_p)X(I_p)$ is of full rank, then the least-squares estimate of $\alpha(I_p)$ is given by

$$\hat{\alpha}(I_p)=\{X'(I_p)X(I_p)\}^{-1}X'(I_p)Y. \tag{1.5}$$

The residual sum of squares in fitting model (1.4) by least-square is equal to

$$\begin{aligned} S(I_p)&=\{Y-X(I_p)\hat{\alpha}(I_p)\}'\{Y-X(I_p)\hat{\alpha}(I_p)\} \\ &=\|Y\|^2-Y'X(I_p)\{X'(I_p)X(I_p)\}^{-1}X'(I_p)Y \end{aligned} \tag{1.6}$$

where (and as well as in what follows) $\|Y\|^2$ is the square norm of vector Y.

In fact we do not know the values of p, i_1, i_2, $\cdots$, i_p, and we just want to estimate them. When we use index set $J_k=\{j_1, j_2, \cdots, j_k\}$ instead of I_p in(1.4), (1.5) and (1.6),we get the values $\hat{\alpha}(J_k)$ and $S(J_k)$. The total number of set J_k for $k=1, 2, \cdots, P$ and $1\leqslant j_1<j_2<j_3<\cdots<j_k\leqslant P$ is equal to

$$C_P^1+C_P^2+\cdots+C_P^P=2^P-1. \tag{1.7}$$

Most of methods for selection of regression variables are concerned with the values $S(J_k)$. Examples of such methods are given by the minimum FPE or AIC method (see [1], [2]), by the minimum BIC method (see [7]), by the C_p method (see [6]) and some other methods (see [5], [9]).

When we use AIC (or BIC) method, we have to calculate the whole values of Akaike Information Criterion function, i.e.

$$\mathrm{AIC}(J_k)=\log S(J_k)+2kT^{-1}, \quad k=0, 1, \cdots, P;\ 1\leqslant j_1<j_2<\cdots<j_k\leqslant P, \tag{1.8}$$

where T is the number of sample size, $J_0=\emptyset$ (empty set) and $\mathrm{AIC}(\emptyset)=\log\|Y\|^2$, and we have to find out the minimum value of $\mathrm{AIC}(J_k)$, say $\mathrm{AIC}(J)$, and then use J to estimate I_p. For BIC method, instead of (1.8) we use

$$\mathrm{BIC}(J_k)=\log S(J_k)+k(\log T)T^{-1}. \tag{1.9}$$

From (1.7) we see that the calculations involved in AIC and BIC methods are too much for practical use. This is the main disadvantage of these methods.

Now we propose two steps:

Step 1. For each k from 1 to P find out index set R_k satisfying

$$S(R_k)=\inf_{J_k} S(J_k), \tag{1.10}$$

where "$\inf_{J_k}$" means to take the minimum value of $S(J_k)$ over all J_k having k elements belonging to the complete set $J_P=\{1, 2, \cdots, P\}$.

Step 2. Let $R_0=\emptyset$, $S(\emptyset)=\log\|Y\|^2$ and

$$\mathrm{BIC}(k)=\log S(R_k)+k(\log T)T^{-1}, \quad k=0, 1, \cdots, P \tag{1.11}$$

and find out r such that

$$\mathrm{BIC}(r)=\inf_{0<k<P}\mathrm{BIC}(k) \tag{1.12}$$

and then take R_r as the estimate of I_p.

It is easy to see that

$$\mathrm{BIC}(r)=\mathrm{BIC}(R_r)=\inf \mathrm{BIC}(J_k), \tag{1.13}$$

where J_k runs over all possible set showed in (1.8). (1.13) shows that to minimize $\mathrm{BIC}(k)$ in (1.11) and to minimize $\mathrm{BIC}(J_k)$ in (1.8) are equivalent. The same description can also be made for AIC method.

Now we are going to give two methods to replace R_k in Step 1, one of which is called "Forward order", and the other "Backward order".

"Forward order": We call sequence M_k as forward order index sets, if $M_0=\emptyset$ and define $M_k=\{m_1, m_2, \cdots, m_k\}$ inductively by

$$S(M_k)=\inf_{j\in M_{k-1}^c} S(M_{k-1}\cup\{j\}), \quad k=1, 2, \cdots, P, \tag{1.14}$$

where $M^c=J_P\backslash M$ is the complement set of M.

"Backward order": We call sequence N_k as backward order index sets, if $N_P=J_P$ and define $N_k=\{n_1, n_2, \cdots, n_k\}$ inductively by

$$S(N_{k-1})=\inf_{j\in N_k} S(N_k\backslash\{j\}), \quad k=P, P-1, \cdots, 3, 2 \tag{1.15}$$

with $N_0=\emptyset$.

When we use sequence M_k instead of R_k in Step 2 of BIC method, let $\mathrm{BIC}_1(m)$ be the minimum value of

$$\mathrm{BIC}_1(k)=\log S(M_k)+kT^{-1}\log T, \quad k=0, 1, \cdots, P, \tag{1.16}$$

then we get a new estimate, M_m, of I_P. We call this method as BIC_1 method. Similarly when we use sequence N_k in Step 2, let $\mathrm{BIC}_2(n)$ be the minimum value of

$$\mathrm{BIC}_2(k)=\log S(N_k)+kT^{-1}\log T, \quad k=0, 1, \cdots, P, \tag{1.17}$$

we get another estimate, N_n, of I_p. We call this method as BIC_2 method. In (1.16), (1.17) and below we put $\mathrm{BIC}_j(0)=\log\|Y\|^2$. The same arguments can be applied to AIC_1 and AIC_2 methods too.

It is easy to see that the whole number of $S(J_k)$, which is calculated in BIC_j (or AIC_j), is only $P(P+1)/2$ which is extremely smaller than (2^P-1) especially for big P. Moreover we can greatly reduce the calculations for BIC_j (or AIC_j) methods again by using the stepwise regression procedure (see [10]), i.e. the calculation to add a new regression variable to model (1.3), or to remove an old regression variable from model (1.3) is always conducted in the same recursive form. To determine forward order sets M_k is equivalent to add a new variable in (1.3) one by one according to restriction (1.14). To determine backward order sets N_k is equivalent to remove an old variable from (1.1) one by one according to restriction (1.15). Therefore we can apply the same recursive form used in the stepwise regression.

In § 2 we shall discuss the consistency of these methods. In § 3 we give some simulation results.

§ 2. The Consistency about BIC and AIC Methods

As we see in § 1, I_p is the true index set by which $\alpha_{i_s} \neq 0$, $s=1, 2, \cdots, p$, $\alpha_i = 0$ for others i's in model (1.2). Now we call $\hat{I}_p$ an estimate of I_p, if $\hat{I}_p$ is a random index set depending on the observations of Y and X. For example, we take $\hat{I}_p = R_r$ when we use BIC method, take $\hat{I}_p = M_m$ for BIC_1 and take $\hat{I}_p = N_n$ for BIC_2. Notice that $\hat{I}_p$ and also R_r, M_m, N_n depend on T. Sometimes we use $\hat{I}_p(T)$, $R_r(T)$, $p(T)$ and so on instead of $\hat{I}_p$, R_r, p and so on respectively to indicate the dependence of $\hat{I}_p$, R_r, p on T.

Let V_T be a sequence of index sets, and V be a set of integers belonging to J_P. We say

$$\lim_{T\to\infty} V_T = V$$

if $V_T = V$ for all $T > T_0$ where T_0 is some positive integer, and say

$$\liminf_{T\to\infty} V_T \supseteq V$$

if $V_T \supseteq V$ for all $T > T_0$, and say

$$\limsup_{T\to\infty} V_T \subseteq V$$

if $V_T \subseteq V$ for infinitely many T.

Definition. If an estimate $\hat{I}_p(T)$ satisfies

$$\lim_{T\to\infty} \hat{I}_p(T) = I_p \quad \text{a.s.} \tag{2.1}$$

we say that $\hat{I}_p(T)$ is consistent estimate to I_p, or consistent for short, and the method to obtain $\hat{I}_p(T)$ is consistent too.

If $\hat{I}_p(T)$ satisfies

$$\liminf_{T\to\infty} \hat{I}_p(T) \supseteq I_p \quad \text{a.s.} \tag{2.2}$$

we say that $\hat{I}_p(T)$ and the method to obtain $\hat{I}_p(T)$ is overconsistent.

In the following theorems we need some conditions on $\varepsilon(t)$ and $x_i(t)$ $(i=1, 2, \cdots, P)$. Suppose

$$\lim_{T\to\infty} A^{-1}(X'X)A^{-1} = R > 0, \tag{2.3}$$

where $A = \operatorname{diag}(\|x_1\|, \|x_2\|, \cdots, \|x_P\|)$ and R is a positive definite matrix,

$$\lim_{T\to\infty} (\log T)^{-1} \min_{1\le i\le P} \|x_i\|^2 = \infty, \tag{2.4}$$

$$\log\log \max_{1\le i\le P} \|x_i\|^2 = O(\log T), \tag{2.5}$$

$$\max_{1\le i\le T} x_i^2(t) = O\{\|x_i\|^2 (\log\|x_i\|^2)^{-\rho}\}$$

$$\text{for any } \rho > 0,\ i = 1, 2, \cdots, P, \tag{2.6}$$

$$E\varepsilon^4(t) < \infty. \tag{2.7}$$

Theorem 1. Under the conditions (2.3) to (2.7) BIC method is consistent, that is

$$\lim_{T\to\infty} R_r(T) = I_p \quad \text{a.s.} \tag{2.8}$$

Proof. Notice that the order of $i_1, i_2, \cdots, i_p$ in (1.3) is not essential, since on changing the order of $i_1, i_2, \cdots\cdots, i_p$ in (1.3) we get equivalent models. Now we put $J_{k-1}=\{j_1, j_2, \cdots, j_{k-1}\}$ and $J_k=\{j_1, j_2, \cdots, j_{k-1}, j_k\}\supseteq I_p$ where $j_k\notin J_{k-1}$, no matter whether $j_k>j_{k-1}$ or not, and put

$$X(J_k)=(x_{j_1}, x_{j_2}, \cdots, x_{j_k}),$$
$$\Lambda(J_k)=\operatorname{diag}(\|x_{j_1}\|, \|x_{j_2}\|, \cdots, \|x_{j_k}\|),$$
$$C=\{X'(J_k)X(J_k)\}^{-1}=(c_{ij})$$

and

$$\hat{\alpha}(J_k)=\{X'(J_k)X(J_k)\}^{-1}X'(J_k)Y, \tag{2.9}$$
$$S(J_k)=Y'Y-Y'X(J_k)\{X'(J_k)X(J_k)\}^{-1}X'(J_k)Y. \tag{2.10}$$

It is known that (see [10])

$$S(J_{k-1})-S(J_k)=\{\hat{\alpha}(J_k)\}_k^2/c_{kk}, \tag{2.11}$$

where (as well as in what follows) we use $\{\cdot\}_k$ as kth element of vector.

Notice that when $J_k\supseteq I_p$ (1.3) can be replaced by

$$Y=X(J_k)\alpha(J_k)+\varepsilon. \tag{2.12}$$

Under the conditions of the theorem with model (2.12) we can use the result appearing in [4] which shows

$$\{\hat{\alpha}(J_k)-\alpha(J_k)\}_k=o\{(c_{kk}|\log\log c_{kk}|)^{1/2}\} \quad \text{a.s.} \tag{2.13}$$

By the condition (2.3)

$$\Lambda(J_k)\{X'(J_k)X(J_k)\}^{-1}\Lambda(J_k)\xrightarrow[T\to\infty]{}U(J_k)=\{u_{ij}(J_k)\}, \tag{2.14}$$

hence there exist two positive numbers u_1 and u_2 such that

$$0<u_1<u_{ij}(J_k)<u_2<\infty. \tag{2.15}$$

From this inequality and (2.14) we have

$$0<u_1\leqslant\lim_{T\to\infty}c_{kk}\|x_{j_k}\|^2=u_{kk}(J_k)\leqslant u_2\leqslant\infty, \tag{2.16}$$

consequently $c_{kk}\to 0$, $c_{kk}|\log\log c_{kk}|\to 0$. Under the same conditions with (2.12) we can use the result appearing in [11] which shows

$$\lim_{T\to\infty}\sigma^2(T)=\lim_{T\to\infty}S(J_k)/T=\sigma^2 \quad \text{a.s.} \tag{2.17}$$

If $j_k\in I_p$, say $j_k=i_\tau$, then $\{\alpha(J_k)\}_k=\alpha_{i_\tau}\neq 0$, and from (2.9), (2.10), (2.11), (2.13), (2.16) and (2.17) we have

$$\begin{aligned}\frac{S(J_{k-1})-S(J_k)}{S(J_k)}&=\frac{\{\hat{\alpha}(J_k)-\alpha(J_k)+\alpha(J_k)\}_k^2}{S(J_k)c_{kk}}\\&=\frac{T^{-1}(\{\hat{\alpha}(J_k)-\alpha(J_k)\}_k+\alpha_{i_\tau})^2\|x_{j_k}\|^2}{T^{-1}S(J_k)c_{kk}\|x_{j_k}\|^2}\\&\geqslant(Tu_2)^{-1}a_T^2\alpha^2\{1+o(c_{kk}|\log\log c_{kk}|)^{1/2}\}^2\\&\geqslant(Tu_2)^{-1}a_T^2\alpha^2\{1+o(1)\} \quad \text{a.s.}\end{aligned} \tag{2.18}$$

where $a_T^2=\min_{1\leqslant i\leqslant P}\|x_i\|^2/\sigma^2$ and $\alpha^2=\min_{1\leqslant s\leqslant p}\alpha_{i_s}^2$, and then (2.18) holds uniformly for $J_k\supseteq I_p$ and $J_k\backslash J_{k-1}\in I_p$. Consequently

$$\min_{J_k\supseteq I_p, j_k\in I_p}\frac{S(J_{k-1})-S(J_k)}{S(J_k)}\geqslant(Tu_2)^{-1}a_T^2\alpha^2\{1+o(1)\}\quad\text{a.s.}\tag{2.19}$$

If $j_k\notin I_p$, then $\{\alpha(J_k)\}_k=0$ and by the same deductions for getting (2.18) and (2.19) we have

$$\begin{aligned}\frac{S(J_{k-1})-S(J_k)}{S(J_k)}&=\frac{T^{-1}(\{\hat\alpha(J_k)-\alpha(J_k)\}_k+0)^2\|x_{j_k}\|^2}{T^{-1}S(J_k)c_{kk}\|x_{j_k}\|^2}\\&=o(T^{-1}c_{kk}|\log\log c_{kk}|)\|x_{j_k}\|^2\\&=o(T^{-1}\log T)\quad\text{a.s.}\quad\text{(by (2.5))}\end{aligned}\tag{2.20}$$

and then (2.20) holds uniformly for $J_k\supseteq I_p$, $J_k\backslash J_{k-1}\notin I_p$. Consequently

$$\max_{J_k\supseteq I_p, j_k\notin I_p}\frac{S(J_{k-1})-S(J_k)}{S(J_k)}=o(T^{-1}\log T)\quad\text{a.s.}\tag{2.21}$$

Secondly we want to prove

$$\liminf_{T\to\infty}R_k\supseteq I_p\quad\text{a.s.}\tag{2.22}$$

Now put

$$\liminf_{T\to\infty}R_r=G=\{g_1,\ g_2,\ \cdots,\ g_Q\}\quad\text{a.s.}\tag{2.23}$$

and $I_p\backslash G=\{\tau_1,\ \tau_2,\ \cdots,\ \tau_s\}$ for $I_p\backslash G\neq\emptyset$, and say $s=0$ for $I_p\backslash G=\emptyset$. Because G is a random set, so s is a random variable taking values from $J_P=\{1,\ 2,\ \cdots,\ P\}$. According to (2.23) there exists T_0 such that

$$R_r=G\quad\text{for } T>T_0.$$

Put $$J_r=G,\ J_{r+j}=G\cup\{\tau_1,\ \tau_2,\ \cdots,\ \tau_j\},\quad j=1,\ 2,\ \cdots,\ s.$$

By the definition of least-square method we know

$$S(J_{r+j})-S(J_{r+j+1})\geqslant 0,\quad j=0,\ 1,\ \cdots,\ s-1.\tag{2.24}$$

By the definition of $S(R_k)$ (see (1.10)) we know

$$S(R_{s+r})\leqslant S(J_{s+r}).\tag{2.25}$$

By the definition of r (see (1.12)) we have

$$\begin{aligned}&\log S(R_r)+rT^{-1}\log T\leqslant\log S(R_{s+r})+(s+r)T^{-1}\log T\\&\qquad\text{for } T>T_0\text{ and } s>0.\end{aligned}\tag{2.26}$$

Consequently from the last three inequalities it follows that

$$\begin{aligned}\log S(J_{s+r-1})-\log S(J_{s+r})&\leqslant\sum_{j=0}^{s-1}\{\log S(J_{r+j})-\log S(J_{r+j+1})\}\quad\text{(by (2.24))}\\&=\log S(J_r)-\log S(J_{s+r})\\&=\log S(R_r)-\log S(J_{s+r})\quad\text{(by the definition of } J_r)\\&\leqslant\log S(R_r)-\log S(R_{s+r})\quad\text{(by (2.25))}\\&\leqslant sT^{-1}\log T\quad\text{(by (2.26))}\end{aligned}$$

which implies that for $T>T_0$ and $s>0$

$$T(\log T)^{-1}\{\log S(J_{s+r-1})-\log S(J_{s+r})\}\leqslant s.\tag{2.27}$$

Using Taylor's expansion of $\log(1+x)$ and (2.19) we know that for $s>0$

$$\begin{aligned} s &\geqslant T(\log T)^{-1}\{\log S(J_{s+r-1}) - \log S(J_{s+r})\} \\ &= T(\log T)^{-1}\log\{1+(S(J_{s+r-1})-S(J_{s+r}))S^{-1}(J_{s+r})\} \\ &\geqslant (u_2 \log T)^{-1}\alpha^2 a_T^2\{1+o(1)\} \xrightarrow[T\to\infty]{} \infty \quad \text{a.s.} \quad \text{(by (2.4))} \end{aligned} \tag{2.28}$$

(2.28) contradicts $P(s>0)=0$. therefore

$$s=0 \quad \text{a.s.}$$

which implies (2.22).

Finally we are going to prove

$$\limsup_{T\to\infty} R_r \subseteq I_p \quad \text{a.s.} \tag{2.29}$$

From (2.22) we see that

$$\limsup_{T\to\infty} R_r = F = \{f_1, f_2, \cdots, f_d\} \supseteq I_p \quad \text{a.s.}$$

Put $F\backslash I_p = \{s_1, s_2, \cdots, s_e\}$ for $F\backslash I_p \neq \emptyset$, say $e=0$ for $F\backslash I_p = \emptyset$. By the definition of F there exists T_0 such that

$$R_r(T_k) = F \quad \text{for } T_k > T_0,$$

where T_k is a subsequence of T. If $e>0$ we put $K = F\backslash\{s_1\}$.

By an argument similar to that in getting (2.27) we have

$$\begin{aligned} \log S(R_r) - \log S(K) &\leqslant \log S(R_r) - \log S(R_{r-1}) \\ &\leqslant -T^{-1}\log T \quad \text{for } T = T_k > T_0. \end{aligned}$$

Consequently

$$\frac{T_k}{\log T_k}\{\log S(K) - \log S(R_r)\} \geqslant 1 \quad \text{for } T = T_k > T_0. \tag{2.30}$$

By the same deduction for getting (2.28) we know that for $e>0$ and $T=T_k>T_0$

$$\begin{aligned} 1 &\leqslant \frac{T_k}{-\log T_k}\{\log S(K) - \log S(R_r)\} = \frac{T_k}{\log T_k}\{\log S(K) - \log S(F)\} \\ &= \frac{T_k}{\log T_k}\left(\log\left\{1+\frac{S(K)-S(F)}{S(K)}\right\}\right) = \frac{T_k}{\log T_k}\, o\left(\frac{1}{T_k}\log T_k\right) \quad \text{(by (2.21))} \\ &= o(1) \quad \text{a.s.} \end{aligned} \tag{2.31}$$

(2.31) contradicts $P(e>0)>0$. Therefore

$$e=0 \quad \text{a.s.}$$

which implies (2.29), (2.22) and (2.29) complete the proof.

Theorem 2. Under the same conditions of Theorem 1, BIC_2 method is consistent too.

Proof. The main technics in the proof of this theorem is similar to those of last theorem. So we only give an outline of the proof of this theorem.

At first we can prove that

$$\liminf_{T\to\infty} N_{P-k} \supseteq I_p \quad \text{a.s.} \tag{2.32}$$

one by one for $k=0, 1, 2, \cdots, P-p$ according to the similar procedure of getting (2.22). Secondly we can prove

$$\liminf_{T\to\infty} n \geqslant p \quad \text{a.s.} \tag{2.33}$$

by the same argument, and then prove

$$\limsup_{T\to\infty} n \leqslant p \quad \text{a.s.} \tag{2.34}$$

by the same method in proving (2.29). Combining (2.32), (2.33) and (2.34) theorem 2 follows immediately.

Theorem 3. Under the conditions (2.3), (2.4) and (2.7) AIC, AIC_1, AIC_2 and BIC_1 are all overconsistent.

Proof. The proof of this theorem is the same as the first half of proof of Theorem 1, so we remove the restrictions (2.5), (2.6) which were only used in the final part of the proof of Theorem 1.

One can give some example to show AIC, AIC_1 and AIC_2 are real overconsistent. These happened in time series analysis (see [8]). Now we are interested in an example to show BIC_1 is real overconsistent too.

Example. In model (1.3) we take $P=3$, $p=2$, $I_p=\{1, 2\}$, $\alpha=(1, 1, 0)'$, and

$$y(t)=x_1(t)+x_2(t)+\varepsilon(t), \quad t=1, 2, \cdots, T, \tag{2.35}$$

where $\varepsilon(t)$ is i.i.d. with zero mean and finite fourth moment, $x_1(t)$, $x_2(t)$ and $x_3(t)$ are periodical series with

$$x_1=(1, 1, 1, 1, 1, 1, \cdots)',$$
$$x_2=(-1+\varepsilon_1, -1+\varepsilon_2, -1+\varepsilon_3, -1+\varepsilon_1, -1+\varepsilon_2, -1+\varepsilon_3, \cdots)',$$
$$x_3=(0, 0, c, 0, 0, c, \cdots)'.$$

We will take ε_1, ε_2 and ε_3 small enough and $c\neq 0$.

It is easy to check that all the conditions of Theorem 1 are satisfied. Now we want to show that BIC_1 method is not consistent for some values of ε_1, ε_2, ε_3 and c.

At first we point out that

$$\frac{3}{T}\sum_1^T x_i(t)\to\|\xi_i\|^2, \quad T\to\infty,$$

where

$$\xi_1=(1, 1, 1)', \quad \xi_2=(-1+\varepsilon_1, -1+\varepsilon_2, -1+\varepsilon_3)', \quad \xi_3=(0, 0, c)',$$

and that

$$\frac{3}{T}\sum_1^T(x_1(t)+x_2(t), x_i(t))\to(\xi_1+\xi_2, \xi_i), \quad T\to\infty,$$

where (ξ, η) means inner product of vector ξ and η. It is obvious that

$$\|\xi_1\|^2=3, \quad \|\xi_2\|^2=\sum_{i=1}^3(1-\varepsilon_i)^2, \quad \|\xi_3\|^2=c^2,$$

$$(\xi_1+\xi_2, \xi_1)=\sum_{i=1}^3\varepsilon_i, \quad (\xi_1+\xi_2, \xi_2)=-\sum_{i=1}^3\varepsilon_i(1-\varepsilon_i), \quad (\xi_1+\xi_2, \xi_3)=c\varepsilon_3.$$

Secondly we use the result appearing in [3], we have

$$\sum_{i=1}^T\varepsilon(t)x_i(t)=o(\|x_i\|(\log\|x_i\|^2)^{(1+\lambda)/2})=o(T^{1/2}(\log T)^{(1+\lambda)/2}) \quad \text{a.s.}$$

for $i=1, 2, 3$ and any $\lambda>0$. Using (1.6) and above equalities, we have

$$S(\{i\}) = \|Y\|^2 - \|X_i\|^{-2}\{\sum_1^T y(t)x_i(t)\}^2$$

$$= \|Y\|^2 - \|X_i\|^{-2}\{\sum_1^T [x_1(t)+x_2(t)]x_i(t) + \sum_1^T \varepsilon(t)x_i(t)\}^2$$

$$= \|Y\|^2 - \frac{T}{3}\{(\xi_1+\xi_2,\ \xi_i)\|\xi_i\|^{-1} + o(T^{-1/2}(\log T)^{(1+\lambda)/2})\}^2$$

$$= \|Y\|^2 - \frac{T}{3}(\xi_1+\xi_2,\ \xi_i)^2\|\xi_i\|^{-2}\{1+o(T^{-1/2}(\log T)^{(1+\lambda)/2}\}^2 \quad \text{a.s.}$$

Now we take

$$\varepsilon_1 = 10^{-1},\ \varepsilon_2 = 10^{-2},\ \varepsilon_3 = 1/2,\ c = 1.$$

Thus

$$(\xi_1+\xi_2,\ \xi_1)^2\|\xi_1\|^{-2} = \frac{1}{3}\sum_1^3 \varepsilon_i < \frac{1}{4},$$

$$(\xi_1+\xi_2,\ \xi_2)^2\|\xi_2\|^{-2} = \Big[\sum_{i=1}^3 \varepsilon_i(1-\varepsilon_i)\Big]^2\Big[\sum_1^3 (1-\varepsilon_i)^2\Big]^{-1} < \frac{1}{4},$$

$$(\xi_1+\xi_2,\ \xi_3)^2\|\xi_3\|^{-2} = (c\varepsilon_3)^2/c^2 = \varepsilon_3^2 = \frac{1}{4}.$$

These inequalities show that when $T\to\infty$

$$S(\{3\}) = \min_{1\le i\le 3} S(\{i\}) \quad \text{a.s.}$$

which means

$$\lim_{T\to\infty} M_1 = \{3\} \quad \text{a.s.}$$

Finally using Theorem 3 we know

$$\liminf_{T\to\infty} M_m \supseteq I_p = \{1,\ 2\} \quad \text{a.s.}$$

Consequently

$$\liminf_{T\to\infty}\ m(T) \geqslant 2 \quad \text{a.s.}$$

By the definition of M_k (see (1.14))

$$M_1 \subseteq M_2,$$

thus

$$\liminf_{T\to\infty} M_m \supseteq \{3\} \quad \text{a.s.}$$

hence

$$\lim_{T\to\infty} M_m = \{1,\ 2,\ 3\} \neq \{1,\ 2\} = I_p \quad \text{a.s.}$$

This means BIC_1 is inconsistent, but it is overconsistent so we call it real overconsistent in this case.

§ 3. Some Further Discussions and Simulations

Although in the last section we pointed out that AIC and BIC_1 methods are overconsistent, but some one prefer to use them in practice some times. Because in practice there is no true model for real data, so in some cases overconsistent selection of regression variables is not bad. However the following simulation

results show that BIC_2 is better than BIC_1 and AIC, and is close to BIC.

Both in practice and theory it is interesting to consider the case where P increases with sample size T. In this case we need some improved results of [4] and [10], but by now we have not got them yet.

Recently we got a new result by which the condition (2.6) can be removed from theorems in this paper.

Finally we put some simulation results of the example of § 2 in the table.

$T=$	100					200					400				
$i=$	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
BIC	9	10	11	0	0	18	18	2	0	1	19	17	5	0	0
BIC_2	15	16	5	0	0	18	18	2	0	1	19	17	5	0	0
BIC_1	4	4	20	0	1	12	4	20	0	1	19	11	20	0	0
AIC	15	15	9	4	3	20	18	3	4	5	20	16	5	2	2

The table of numbers of i appearing in 20 estimates $\hat{I}_p$ for each method and each T.

Besides variables x_1, x_2 and x_3 of the example in § 2 we add two more variables again which are

$$x_4(t)=\cos\frac{2\pi}{4}t, \quad x_5(t)=\sin\frac{2\pi}{4}t, \quad t=1, 2, \cdots, T,$$

then $P=5$ instead of 3, and assume that $\varepsilon(t)$ is normal distributed in (2.35). We show the results of our simulations of selection variables of (2.35) model by using BIC, BIC_1, BIC_2 and AIC method for estimating $I_p=\{1, 2\}$. For each $T(=100, 200, 400)$ we repeated 20 times of the same simulation independently and then gave 20 estimates $\hat{I}_p$ for every method. The table lists the numbers of each $i(=1, 2, 3, 4, 5)$ appearing among the 20 estimates $\hat{I}_p$ for each method and each T.

We thank Professor Wang Shouren for help in writting this paper.

References

[1] Akaike, H., Statistical Predictor Identification, *Ann. Inst. Statist.*, **22** (1970), 203—217.
[2] Akaike, H., Information Theory and an Extension of the Maximum Likelihood Principle, In 2nd International Symposium on Information Theory, Eds. B. N. Petrov and F. Csaki, 267—281, Budapest: Akademia Kiado, 1973.
[3] Chen G. J., Lai T. L. and Wei C. Z., Convergence System and Strong Consistency of Least Squares Estimates in Regression Models, *J. Multivariate Anal.*, **11** (1981), 319—333.
[4] Lai T. L. and Wei C. Z.,A Law of the Iterated Logarithm for Double Arrays of Independent Random Variables with Application to Regression and Time Series Models, *Ann. Prob.*, **10**: 2 (1982), 320—335.
[5] Liu C. W. and Wu G. F., Criteria of Variables Selection in Regression Analysis, *Mathematics in Practice and Theory*, **1** (1983), 61—70.
[6] Mallows, C. L., Some Comments on C_p, *Technomethics*, **12** (1973), 591—612.
[7] Schwarz, G., Estimating the Dimension of a Model, *Ann. Statist.*, **6** (1978), 461—464.
[8] Shibata, R., Selection of the Order of an Autoregressive Model by Akaike's Information Criterion, *Biometrika*, **63** (1976), 117—126.
[9] Shibata, R., An Optimal Selection of Regression Variables (to appear in *Biometrika*).
[10] Zhang R. T. and Fang K. T., Introduction of Multivariate Analysis, Science Press. (in Chinese)
[11] Zhao L. Z., The Necessary and Sufficient Condition for the Estimate of Variance of Error in Linear Model to be of Consistency, *Sci. Sinica*, **10** (1981), 1187—1191.