Ann. Inst. Statist. Math.
38 (1986), Part A, 459-474

SELECTION OF THE NUMBER OF REGRESSION VARIABLES; A MINIMAX CHOICE OF GENERALIZED FPE

RITEI SHIBATA

(Received Oct. 31, 1984; revised Jan. 16, 1985)

Summary

A generalized Final Prediction Error (FPE_α) criterion is considered. Based on n observations, the number k of regression variables is selected from a given range $0 \leqq k \leqq K$, so as to minimize $\text{FPE}_\alpha(k) = n\hat{\sigma}^2(k) + \alpha k\{n\hat{\sigma}^2(k)/(n-K)\}$. It is shown that if α tends to infinity with n, the selection is consistent but the maximum of the mean squared error of estimates of parameters diverges to infinity with the same order of divergence as that of α. A meaningful minimax choice of α exists for a regret type mean squared error, while for simple mean squared error it is trivially 0. The minimax regret choice of α converges to a constant, approximately 3.5 for $K \geqq 8$ if $n-K$ increases simultaneously with n, otherwise it diverges to infinity with n.

1. Introduction

Consider a regression model

$$y = X\beta(k) + \varepsilon , \qquad \varepsilon \sim N(0, \sigma^2 I) ,$$

where $\beta(k)' = (\beta_1, \beta_2, \cdots, \beta_k, 0, \cdots, 0)$ is the vector of regression parameters, X is the $n \times K$ design matrix and $0 \leqq k \leqq K$. We define $\beta(0) = 0$. We call the above model "model k", as k regression variables are included. Assume that σ^2 is unknown. Our problem is how to select a model, k, from a given range $0 \leqq k \leqq K$, $K \geqq 1$, based on an observed sample $y' = (y_1, \cdots, y_n)$.

Assume that X is of full rank K. Let $\hat{\beta}(k)' = (\hat{\beta}_1(k), \cdots, \hat{\beta}_k(k), 0, \cdots, 0)$ be the least squares estimate of $\beta(k)$ under the model k. Define $\hat{\beta}(0) = 0$. The residual sum of squares is then

$$n\hat{\sigma}^2(k) = \|y - X\hat{\beta}(k)\|^2 ,$$

Key words and phrases: Selection of regression variables, nested models, FPE, AIC.

where $\|\cdot\|$ denotes the Euclidean norm.

The selection procedure considered here is the minimum FPE_α criterion (Akaike [1], Bhansali and Downham [3], Atkinson [2], Shibata [9]). A model $\hat{k}_\alpha$ is selected so as to minimize

$$\mathrm{FPE}_\alpha(k) = n\hat{\sigma}^2(k) + \alpha k \tilde{\sigma}^2(K) ,$$

where $\tilde{\sigma}^2(K) = n\hat{\sigma}^2(K)/(n-K)$. As a loss function, we adopt the sum of squares,

$$L(\hat{k}_\alpha) = \| X\hat{\beta}(\hat{k}_\alpha) - X\beta \|^2 , \tag{1.1}$$

where $\beta' = (\beta_1, \beta_2, \cdots, \beta_K)$ is the vector of regression parameters to be estimated. The loss function (1.1) is connected with the prediction error of future observations at the same sampling points as those of y (see Shibata [9], [10], Stone [13]). As possible values of α, it is enough to consider only nonnegative α's, since $\hat{k}_\alpha \equiv K$ for $\alpha \leqq 0$.

It is already shown (Shibata [10], [11]) that if y is generated from a model with infinitely many regression variables, 2 is an optimal choice of α under the loss function (1.1). However, the situation is different if y is generated from a fixed model k_0 which is small. For such case, Shibata [12] analyzed the behavior of the risk

$$R(\hat{k}_\alpha) = \mathrm{E}\{L(\hat{k}_\alpha)\} ,$$

by using theorems of random walk. One of his analyses suggests that the larger α the better in this case. A controversial point is the key assumption that the underfitting risk is negligible. Such an assumption is justified if the last nonzero coordinate of β is significantly large, or if the size n of the observation is large for a fixed parameter β and a fixed α. The present paper aims to find a theoretical guide of how to choose α from the view point of the minimax principle, when the model k_0 is small and the underfitting risk is not negligible. We do not intend to propose any specific choice of α.

There are many papers in which an α greater than 2 is suggested. Some of the authors recommend the use of an α divergent with n, like in BIC (Schwarz [8]). On the other hand, Bhansali and Downham [3] suggests the use of a constant α, for example, 3 or 4, for any size n. The result in Atkinson [2] also supports such a choice. Thus a question arises, a divergent α or a constant α?

In Section 2, to simplify our problem, a canonical representation of $R(\hat{k}_\alpha)$ and that of $\mathrm{FPE}_\alpha(k)$ are derived. In Section 3, the behavior of $R(\hat{k}_\alpha)$ is analyzed. A minimax choice of α is unique but trivial $\alpha=0$, so that $R(\hat{k}_\alpha)$ does not give any meaningful choice of α. Such difficulty can be overcome by introducing a concept of "regret". The regret here is defined as $\delta R(\hat{k}_\alpha)=R(\hat{k}_\alpha)-R(k_0)$, since $R(k_0)$ signifies the risk for a selection $\hat{k}\equiv k_0$ when the model k_0 is supposed true. An approximation to the minimax solution in Section 5 suggests that the choice 3 or 4 can be justified if $n-K$ is increased with n. Otherwise the use of a divergent sequence is desirable. It is also shown in Theorem 3.1 that both the maximum risk and the maximum regret diverge to infinity with $O(\alpha)$.

A related work is by Hosoya [7]. He considered a similar problem under the constraint $k_0-1\leqq\hat{k}_\alpha\leqq k_0$ for a fixed k_0. This constraint does not allow any overfitting. We should note that the overfitting risk is inevitable in selecting a model and indispensable for realizing the principle of parsimony.

2. Canonical representation of the risk

Let Q be an orthogonal $n\times n$ matrix which transforms the design matrix X into an $n\times K$ matrix

$$\begin{bmatrix} S \\ 0 \end{bmatrix},$$

where S is a $K\times K$ upper triangular matrix. An example of such Q is given by the Householder transformation (Golub [5]). Let $S(k)$ be the $k\times k$ principal submatrix of S. Then Gauss-Markov's equation under the model k is transformed to

$$S(k)\begin{bmatrix} \hat{\beta}_1(k) \\ \vdots \\ \hat{\beta}_k(k) \end{bmatrix}=\begin{bmatrix} (Qy)_1 \\ \vdots \\ (Qy)_k \end{bmatrix}.$$

Putting $W_k=(Qy)_k/\sigma$ and $\mu_k=(QX\beta)_k/\sigma$ for $1\leqq k\leqq n$, we can write

$$\begin{aligned} L(k)&=\|QX\hat{\beta}(k)-QX\beta\|^2 \\ &=\sum_{l=1}^{k}\{(Qy)_l-(QX\beta)_l\}^2+\sum_{l=k+1}^{K}(QX\beta)_l^2 \\ &=\sigma^2\left\{\sum_{l=1}^{k}(W_l-\mu_l)^2+\sum_{l=k+1}^{K}\mu_l^2\right\}. \end{aligned} \tag{2.1}$$

On the other hand, the residual sum of squares is written as

$$n\hat{\sigma}^2(k)=\|Qy-QX\hat{\beta}(k)\|^2=\sigma^2\sum_{l=k+1}^{n}W_l^2 .$$

Since W_l, $l=1,\cdots,n$ are independent normally distributed random variables with mean μ_l and variance 1, $n\hat{\sigma}^2(k)/\sigma^2$ is distributed as noncentral

χ^2 with degree of freedom $n-k$ and with noncentrality $\sum_{l=k+1}^{K} \mu_l^2$, since $\mu_{K+1}=\cdots=\mu_n=0$.

We should note that $\hat{k}_\alpha$ is determined by whether the following differences are positive or not;

$$\text{FPE}_\alpha(k)-\text{FPE}_\alpha(l)=\{n\hat{\sigma}^2(k)-n\hat{\sigma}^2(l)\}+\alpha(k-l)\tilde{\sigma}^2(K)\,, \quad 0\leqq k<l\leqq K\,. \tag{2.2}$$

On the right hand side of (2.2), the first term

$$n\hat{\sigma}^2(k)-n\hat{\sigma}^2(l)=\left(\sum_{m=k+1}^{l} W_m^2\right)\sigma^2$$

and the second term

$$(n-K)\tilde{\sigma}^2(K)=\left(\sum_{m=K+1}^{n} W_m^2\right)\sigma^2\,,$$

are independent. If $\tilde{\sigma}^2(k)$ is used in place of $\tilde{\sigma}^2(K)$, the above two terms are no longer independent. The use of $\tilde{\sigma}^2(K)$ is advantageous not only for mathematical analysis but also for stability of $\hat{k}_\alpha$ (Shibata [12]). Another advantage is that $\tilde{\sigma}^2(K)$ is an unbiased, as well as consistent, estimate of σ^2 if y is generated from a model k_0 in $0\leqq k_0\leqq K$.

From (2.2), we can easily see that the minimization of $\text{FPE}_\alpha(k)$ is equivalent to the maximization of

$$S_k=\sum_{m=1}^{k}(W_m^2-\alpha U) \qquad \text{in } 0\leqq k\leqq K\,, \tag{2.3}$$

where $S_0=0$ and $U=\tilde{\sigma}^2(K)/\sigma^2$. Therefore, it suffices to analyze the behavior of the risk

$$R(\hat{k}_\alpha)=\sigma^2\,\text{E}\left\{\sum_{l=1}^{\hat{k}_\alpha}(W_l-\mu_l)^2+\sum_{l=\hat{k}_\alpha+1}^{K}\mu_l^2\right\}\,,$$

for $\hat{k}_\alpha$ which maximizes S_k in (2.3).

We hereafter consider the transformed vector $\mu'=(\mu_1,\cdots,\mu_K)$ instead of β. Then μ runs over $\boldsymbol{R}^K$ and $\mu_{k+1}=\cdots=\mu_K=0$ is equivalent to $\beta_{k+1}=\cdots=\beta_K=0$. We may assume that $\sigma^2=1$, since $\hat{k}_\alpha$ is invariant under changes of σ^2. For mathematical convenience, we sometimes assume that y is generated from a model k_0 in $1\leqq k_0\leqq K$, where a notation $\mu(k_0)$ is used in place of μ to signify that $\mu(k_0)$ is restricted on $\boldsymbol{R}^{k_0}$.

3. Behavior of $R(\hat{k}_\alpha)$ and that of minimax solution

We first analyze the behavior of $R(\hat{k}_\alpha)$ for the case when $\mu(k_0)$ tends to infinity for fixed n and fixed k_0 in $1\leqq k_0\leqq K$.

Let $\tilde{S}_{k_0}=0$ and $\tilde{S}_k=\sum_{l=k_0+1}^{k}(W_l^2-\alpha U)$ for $k\geqq k_0+1$. Then, $\tilde{S}_k$ is a random walk conditionally for given U.

THEOREM 3.1. *For any fixed n and K, a finite boundary value exists,*

$$\lim_{\mu(k_0)\to\infty} R(\hat{k}_\alpha)=k_0+\sum_{m=1}^{K-k_0} \mathrm{P}\left(F_{m+2,n-K}\geqq\frac{\alpha m}{m+2}\right), \qquad \text{for } 1\leqq k_0\leqq K,$$

and

$$R(\hat{k}_\alpha)\equiv\sum_{m=1}^{K} \mathrm{P}\left(F_{m+2,n-K}\geqq\frac{\alpha m}{m+2}\right), \qquad \text{for } k_0=0. \tag{3.1}$$

PROOF. We prove (3.1) only for the case when $k_0=1$. The other proofs are similar. Let us define $M=\max_{1\leqq k\leqq K}\tilde{S}_k$, $W=-(W_1^2-\alpha U)$ and $T=\max(\hat{k}_\alpha-1, 0)$. Then

$$\begin{aligned} R(\hat{k}_\alpha)-1&=R(\hat{k}_\alpha)-\mathrm{E}(W_1-\mu_1)^2 \\ &=\mathrm{E}\left[\{\mu_1^2-(W_1-\mu_1)^2\}I_{(M<W)}\right]+\mathrm{E}\{(M+\alpha T)I_{(M>W)}\}, \end{aligned} \tag{3.2}$$

where I_A is the indicator function of a measurable set A. We first show that the first term on the right hand side of (3.2), which defines the risk when $\hat{k}_\alpha=0$, converges to zero as μ_1 tends to infinity. Noting that $M\geqq 0$ is independent of μ_1, we have, for the first term on the right hand side of (3.2),

$$\begin{aligned} &\mathrm{E}\{|\mu_1^2-(W_1-\mu_1)^2|I_{(M<W)}\} \\ &\quad\leqq\mu_1^2\,\mathrm{P}(M<W)+\mathrm{E}\{(W_1-\mu_1)^2I_{(M<W)}\} \\ &\quad\leqq\mu_1^2\,\mathrm{P}(W>0)+\mathrm{E}\{(W_1-\mu_1)^2I_{(W>0)}\}. \end{aligned} \tag{3.3}$$

Since W_1 in $W=-(W_1^2-\alpha U)$ is normally distributed with mean μ_1 and variance 1, the probability $\mathrm{P}(W>0)$ exponentially goes to zero as μ_1 tends to infinity. Therefore the right hand side of (3.3) converges to 0. By the same reason, the second term on the right hand side of (3.2) converges to

$$\mathrm{E}(M+\alpha T). \tag{3.4}$$

As it is proved in Shibata [11] that

$$\mathrm{E}(M+\alpha T|U)=\sum_{m=1}^{K-1}\mathrm{P}(\chi_{m+2}^2>\alpha mU|U), \tag{3.5}$$

we have the desired result by taking expectations of both sides of (3.5) with respect to U.

From the above theorem, noting that $R(\hat{k}_\alpha)$ is an even continuous function of $\mu(k_0)$, we see that $\max_{\mu(k_0)} R(\hat{k}_\alpha)$ exists and is finite. We therefore find a minimax solution $\alpha=0$ from the inequality

$$\max_{\mu} R(\hat{k}_\alpha)=\max_{k_0} \max_{\mu(k_0)} R(\hat{k}_\alpha)\geqq K=R(\hat{R}_0) .$$

There still remains a possibility of other nontrivial minimax solutions existing. To prove the uniqueness we need the following lemma. The proof is placed in Appendix.

LEMMA 3.1. *The function*

$$f_\mu(a)=\int_{-a}^{a} \{\mu^2-(x-\mu)^2\}\phi(x-\mu)dx$$

is an increasing function of a, on $0\leqq a\leqq a^*(\mu)$ for any μ such that $\mu^2>1/2$, and is positive for any $a>0$ provided that $|\mu|>1$. Furthermore,

$$a^2\{\Phi(2a)-1/2\}-1/2\leqq\max_{\mu} f_\mu(a)\leqq a^2 . \tag{3.6}$$

Here $\phi(x)$ and $\Phi(x)$ are the standard normal density and the distribution, respectively, and $a^*(\mu)$ is the solution of the equation

$$(2|\mu|-a)=(2|\mu|+a)\exp(-2|\mu|a) .$$

THEOREM 3.2. *For the risk $R(\hat{k}_\alpha)$, the minimax solution $\alpha=0$ is unique.*

PROOF. As is shown in Theorem 3.1, $R(\hat{k}_\alpha)$ converges to

$$K+\mathrm{E}[\{\mu_K^2-(W_K-\mu_K)^2\}I_{(W_K^2<\alpha U)}] ,$$

when $\mu_1,\cdots$, and μ_{K-1} tend to infinity but μ_K is fixed. Put $a=(\alpha U)^{1/2}$ and $\mu=\mu_K$ in Lemma 3.1, then

$$\mathrm{E}[\{\mu_K^2-(W_K-\mu_K)^2\}I_{(W_K^2<\alpha U)}|U]>0$$

a.s. for any $|\mu_K|>1$ and for $\alpha>0$. This implies that

$$\max_{\mu} R(\hat{k}_\alpha)>K$$

for any $\alpha>0$, and the theorem is proved.

The theorem indicates that some kind of modification is necessary for the risk $R(\hat{k}_\alpha)$, to obtain a meaningful minimax solution. From the principle of parsimony, let us consider a "regret"

$$\delta R(\hat{k}_\alpha)=R(\hat{k}_\alpha)-R(k_0) .$$

in place of $R(\hat{k}_\alpha)$. This regret measures how much the risk increases by using $\hat{k}_\alpha$ rather than using the true k_0, for which the risk $R(k_0)$ is constant k_0. Such a concept of the regret was introduced earlier in Shibata [9] by the name of "increase in risk". It is called "opportunity risk" in Hosoya [7]. In the next section, we will analyze the behavior of $\delta R(\hat{k}_\alpha)$ and the existence of a nontrivial minimax solution.

4. Behavior of the regret $\delta R(\hat{k}_\alpha)$ and that of minimax solution

The following theorem shows an asymptotic behavior of $\delta R(\hat{k}_\alpha)$. The result plays an important role in the next section for obtaining an approximation to the minimax solution by computer simulations.

THEOREM 4.1. *The maximum regret* $\max_{\mu(k_0)} \delta R(\hat{k}_\alpha)$ *diverges to infinity with* $O(\alpha)$ *as* α *tends to infinity. For* $\delta R(\hat{k}_\alpha)$, *the minimax solution of* α *exists and is finite.*

PROOF. From Theorem 3.1, the maximum regret

$$\max_{\mu} \delta R(\hat{k}_\alpha) = \max_{0 \leq k_0 \leq K} \max_{\mu(k_0)} \delta R(\hat{k}_\alpha)$$

always exists and is finite. The latter part of the theorem follows from the first part, since the above maximum regret is a continuous function of $\alpha \geqq 0$. We prove the first part only for the case when $k_0 = 1$. Notations are the same as in the proof of Theorem 3.1. Putting $a = (\alpha U - M)^{1/2}$ and $\mu = \mu_1$ in (3.6), we have

$$\alpha \geqq \max_{\mu_1} \delta R(\hat{k}_\alpha) \geqq \max_{\mu_1} \mathrm{E}\,[\{\mu_1^2 - (W_1 - \mu_1)^2\} I_{(M<W)}] \geqq \mathrm{E}\,\{\xi(\alpha U - M)\}\ , \tag{4.1}$$

where $\xi(x) = x\{\Phi(2x^{1/2}) - 1/2\} - 1/2$ if $x > 0$, otherwise 0. As α tends to infinity, the random walk $\tilde{S}_k$, $1 \leqq k \leqq K$ drifts to $-\infty$. The maximum M then a.s. converges to 0. This proves the desired result.

An important implication of Theorem 4.1 is the following. If α is chosen as a divergent sequence in n, the $\max_{\mu(k_0)} R(\hat{k}_\alpha)$ as well as the $\max_{\mu(k_0)} \delta R(\hat{k}_\alpha)$ diverges to infinity with n. Some of known consistent procedures have such a divergent sequence. For example, BIC by Schwarz [8] has $\alpha = \log n$, and φ by Hannan and Quinn [6] has $\alpha = c \log\log n$ for some $c > 2$. Furthermore, in the context of time series models, it is proved that the FPE_α procedure is strongly consistent if and only if $\alpha \geqq 2 \log\log n$ (Hannan and Quinn [6]). However, in view of Theorem 4.1 such consistency is obtained at the cost of uniform boundedness of $R(\hat{k}_\alpha)$ or $\delta R(\hat{k}_\alpha)$. The consistency of $\hat{k}_\alpha$ and the uniform boundedness of the mean squared error $R(\hat{k}_\alpha)$ may not be compatible.

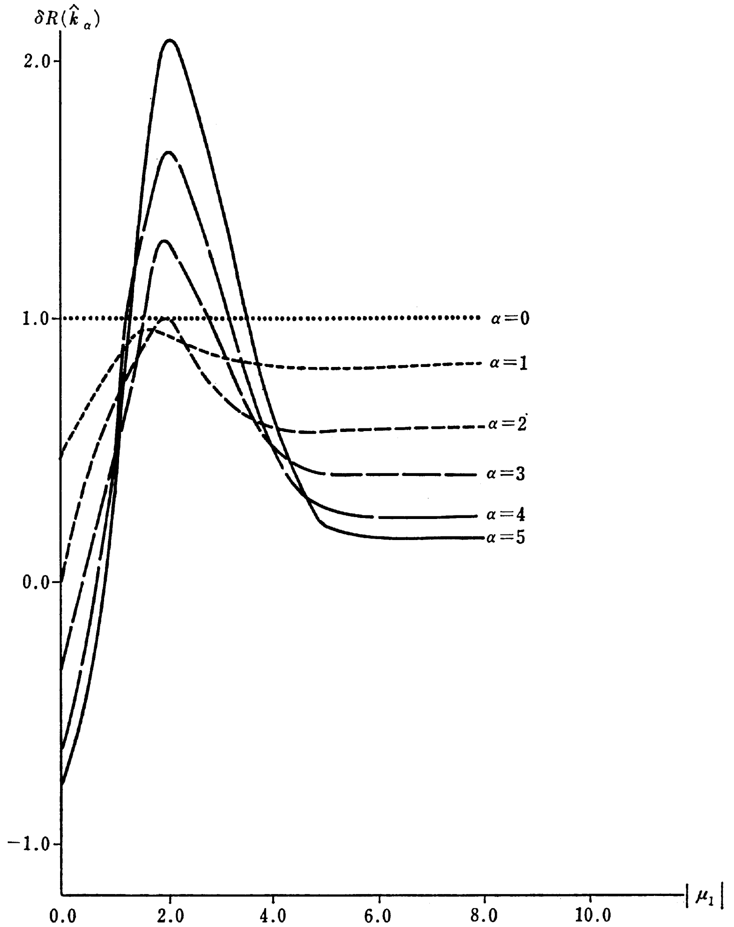

Fig. 1. Regret $\delta R(\hat{k}_\alpha)$ for large n, $0 \leqq \hat{k}_\alpha \leqq K=2$ and $k_0=1$.

To illustrate the behavior of $\delta R(\hat{k}_\alpha)$, a plot of $\delta R(\hat{k}_\alpha)$ is given in Fig. 1, where $K=2$ and $k_0=1$ with large enough n. We see that $\delta R(\hat{k}_\alpha)$ quickly converges to a constant as μ_1 increases. This is the fact proved in Theorem 3.1. The maximum of $\delta R(\hat{k}_\alpha)$ is attained at a relatively small μ_1, but the value of the maximum itself rapidly increases with α, as was already proved in Theorem 4.1. An interesting fact is that $\delta R(\hat{k}_\alpha)$ is negative for $\alpha \geqq 2$ around $\mu_1=0$. This shows that if μ_1 is very small but not zero, then the mean squared error can be reduced by fitting a smaller model k_0-1 rather than fitting the true model k_0.

More generally, Theorem 3.1 gives a necessary and sufficient condition for $\delta R(\hat{k}_\alpha)$ being negative at $\mu_1=\cdots=\mu_{k_0-1}=\infty$ and $\mu_{k_0}\approx 0$,

$$\sum_{m=1}^{K-k_0+1} \mathrm{P}\left(F_{m+2,n-k} > \frac{\alpha m}{m+2}\right) < 1 \,. \tag{4.2}$$

In the case of Fig. 1, the condition (4.2) becomes equivalent to $\alpha \geqq 2.0$. Similarly, $\alpha \geqq 2.4$, 2.6, 2.7 and 2.8 are equivalent to (4.2), respectively for $K-k_0=2$, 3, 4 and 5, provided that n is large enough. Therefore, if the condition (4.2) is required for any k_0 in $0 \leqq k_0 \leqq K$, then α should be greater than 2.0, 2.4, 2.6, 2.7 or 2.8, respectively for $K=1$ to 5. Such boundary values are worthy of consideration as a guide of how to choose α, together with an upper bound in the next section.

5. An approximate minimax regret solution under a constraint

In this section, we try to find an approximate value of the minimax solution of α for $\partial R(\hat{k}_\alpha)$ by computer simulations. To save computation time, we put a constraint $k_0-1 \leqq \hat{k}_\alpha \leqq K$ for each k_0 in $0 \leqq k_0 \leqq K$. This is equivalent to restricting our attention to the parameters of the form $\mu(k_0)$, in which $\mu_1, \cdots, \mu_{k_0-1}$ are large but μ_{k_0} is not. There may be an objection that such constraint is artificial. If $K=1$ or 2, the constraint is of no effect, otherwise it forces the solution to be greater, for less chance of underfitting. But, as will be seen later, our solution gives a good upper bound for the unconstraint minimax solution.

The following Theorem 5.1 holds true without any constraint, but for simplicity we give the theorem under the constraint.

The maximum of $\partial R(\hat{k}_\alpha)$ with respect to μ_{k_0} is independent of $k_0 \geqq 1$ itself, but depends on $K-k_0$ and α. We can then define

$$\partial R^*(\alpha, K-k_0) = \begin{cases} \max\limits_{\mu_{k_0}} \partial R(\hat{k}_\alpha) & \text{for } 1 \leqq k_0 \leqq K \\ \partial R(\hat{k}_\alpha) & \text{for } k_0 = 0 \,. \end{cases}$$

In previous sections, theorems are derived for the case when both K and n are fixed, but here we analyze the behavior of $\partial R^*(\alpha, K-k_0)$ is analyzed for the case when $K-k_0$ is increased to infinity.

THEOREM 5.1. *If $n-K$ diverges to infinity as K tends to infinity together with n, then for any fixed k_0 in $0 \leqq k_0 \leqq K$,*

$$\lim_{K\to\infty} \partial R^*(\alpha, K-k_0) = \begin{cases} \partial R^*(\alpha, \infty) & \textit{for } \alpha > 1 \\ \infty & \textit{for } \alpha \leqq 1 \,, \end{cases}$$

where $\partial R^*(\alpha, \infty)$ is defined in (5.5). *The minimax solution α in Theorem* 4.1 *converges to a constant $\alpha^* > 1$ as K tends to infinity.*

If $n-K$ is fixed but K tends to infiity together with n, then for any fixed $0\leqq k_0\leqq K$ and for any α,

$$\lim_{K\to\infty} \delta R^*(\alpha, K-k_0)=\infty . \tag{5.2}$$

The minimax solution α diverges to infinity as K tends to infinity.

PROOF. It is enough to show (5.1) only for the case when $k_0=1$. Define

$$\delta R(\hat{k}_\alpha | U)=\mathrm{E}\{L(\hat{k}_\alpha)-L(k_0)|U\} .$$

Then

$$\delta R(\hat{k}_\alpha | U)=\mathrm{E}[\{\mu_1^2-(W_1-\mu_1)^2\}I_{(M<W)}|U]+\mathrm{E}[(M+\alpha T)I_{(M>W)}|U] . \tag{5.3}$$

We hereafter investigate the behavior of M or T for given U. Since the random variables M and T are monotone increasing function of K, random variables $M_\infty=\lim_{K\to\infty} M$ and $T_\infty=\lim_{K\to\infty} T$ are a.s. well defined. The limit variables M_∞ and T_∞ have proper distributions if and only if

$$\lim_{K\to\infty} \sum_{l=1}^{K-1} \mathrm{P}(\tilde{S}_{l+1}>0|U)/l<\infty . \tag{5.4}$$

The condition (5.4) is equivalent to $\alpha U>1$. This equivalence follows from the fact that, as an increase of l the probability

$$\mathrm{P}(\tilde{S}_{l+1}>0|U)=\mathrm{P}\left(\frac{1}{l}\sum_{m=2}^{l+1} W_m^2>\alpha U|U\right)$$

exponentially goes to zero, or goes to 1/2, or goes to 1, whether $\alpha U=1$ or $\alpha U=1$ or $\alpha U<1$, respectively.

We first consider the case when both $n-K$ and K are large enough. The condition (5.4) is then equivalent to $\alpha>1$. Therefore, if $\alpha>1$, $\delta R(\hat{k}_\alpha|U)$ converges a.s. to

$$\delta R_\infty(\hat{k}_\alpha)=\mathrm{E}[\{\mu_1^2-(W_1-\mu_1)^2 I_{(M_\infty<W_\infty)}]+\mathrm{E}[(M_\infty+\alpha T_\infty)I_{(M_\infty>W_\infty)}] ,$$

where $W_\infty=-(W_1^2-\alpha)$. We should note that the above convergence is uniform in μ_1. It is for this reason that

$$\max_{\mu_1} \mathrm{E}\{|\mu_1^2-(W_1-\mu_1)^2|^2 I_{(W>0)}|U\}$$

is a.s. bounded and is independent of K. Since $\delta R(\hat{k}_\alpha|U)=\mathrm{E}(M+\alpha T|U)$ when $k_0=0$, we have (5.1) by defining

$$\delta R^*(\alpha, \infty)=\begin{cases} \max_{\mu_1} \delta R_\infty(\hat{k}_\alpha) & \text{for } k_0\geqq 1 \\ \mathrm{E}(M_\infty+\alpha T_\infty) & \text{for } k_0=0 . \end{cases} \tag{5.5}$$

Therefore,

$$\max_{0 \leqq k_0 \leqq K} \delta R^*(\alpha, K-k_0)$$

remains unchanged for large enough K and for $\alpha>1$. For $\alpha \leqq 1$, the conditional regret $\delta R(\hat{k}_\alpha|U)$ as well as the regret $\delta R(\hat{k}_\alpha)$ diverges to infinity. Therefore, the minimax solution converges to a finite value as $n-K$ and K simultaneously tend to infinity.

Next consider the case when $n-K$ is fixed. For any U such that $\alpha U<1$, $\delta R(\hat{k}_\alpha|U)$ a.s. diverges to infinity with K. The probability $\mathrm{P}(\alpha U<1)$ remains unchanged and nonzero, thus (5.2) follows. The proof is complete if we show that the minimax solution diverges to infinity with K. From Lemma 3.1, we have

$$\delta R(\hat{k}_\alpha|U) \leqq \alpha U + \mathrm{E}(M+\alpha T|U)$$

for any μ_{k_0}. Therefore

$$\delta R^*(\alpha, K-k_0) \leqq \alpha + \mathrm{E}\left\{\sum_{m=1}^{K-k_0} \mathrm{P}(\chi^2_{m+2} > \alpha m U|U)\right\} \tag{5.6}$$

for $0 \leqq k_0 \leqq K$. Here, from the inequality used in Shibata [12],

$$\mathrm{P}(\chi^2_{m+2} > \alpha m U|U) \leqq \exp\left[-\frac{1}{12}\frac{\{(\alpha U-1)m-2\}^2}{m+2}\right]$$

for an $\alpha m U > m+2$, we have the boundedness of

$$\mathrm{E}\left[\sum_{m=1}^{K-k_0} \mathrm{P}(\chi^2_{m+2} > \alpha m U|U) I_{(\alpha U>3)}\right]$$

both in K and α. On the other hand,

$$\mathrm{E}\left[\sum_{m=1}^{K-k_0} \mathrm{P}(\chi^2_{m+2} > \alpha m U|U) I_{(\alpha U \leqq 3)}\right]$$

is bounded by $(K-k_0)\,\mathrm{P}(\alpha U \leqq 3)$. Combining these results, we see that the right hand side of (5.6) is bounded, so that we may find a divergent sequence α_K such that

$$\lim_{K\to\infty} \max_{0 \leqq k_0 \leqq K} \delta R^*(\alpha_K, K-k_0)/K = 0 \,. \tag{5.7}$$

Whereas, for fixed α

$$\liminf_{K\to\infty} \max_{0 \leqq k_0 \leqq K} \delta R^*(\alpha, K-k_0)/K \geqq \mathrm{P}(1>\alpha U) \,, \tag{5.8}$$

which follows from the inequality,

$$\delta R^*(\alpha, K-k_0) \geqq \sum_{m=1}^{K} \mathrm{E}\{\mathrm{P}(\chi^2_{m+2} > \alpha m U|U)\} \,,$$

since $P(\chi^2_{m+2}>\alpha m U|U)$ a.s. converges to $P(1>\alpha U|U)$ as m tends to infinity. Therefore, the minimax solution diverges to infinity with K as far as $n-K$ is fixed.

Theorem 5.1 is well illustrated by the results of computer simulations. For various μ_{k_0}'s, the values of $\delta R(\hat{k}_\alpha)$ were estimated by 1000 experiments based on generated normal random numbers for $W_1, \cdots, W_K$. Table 1 is a part of the results for the case when n is large enough. In this table, the column "Max" stands for $\delta R^*(\alpha, K-k_0)=\max_{\mu_{k_0}} \delta R(\hat{k}_\alpha)$ and the column "Limit" stands for $\lim_{\mu_{k_0}\to\infty} \delta R(\hat{k}_\alpha)$, which is the regret for the case of "Sharp fit" (see Shibata [12]). We can see how fast $\delta R^*(\alpha, K-k_0)$ converges to a constant as an increase of $K-k_0$. If $K-k_0\geqq 14$, these values are satisfactorily convergent. The value for $K-k_0=19$ is an exception, which is always equal to the corresponding value in "Limit", since it allows no underfitting. Although our experiments are limited, the minimax solution $\alpha=3.5$ obtained in Table 1 does not seem far from the limit α^* in Theorem 5.1.

From Table 1, we can also obtain minimax solutions for other K's.

Table 1. The maximum and the limit of the regret $\delta R(\hat{k}_\alpha)$ with respect to μ_{k_0} for the case when $K=19$ and $n-K$ is large

$K-k_0$	$\alpha=1.0$		$\alpha=2.0$		$\alpha=3.0$		$\alpha=3.5$		$\alpha=4.0$	
	Max	Limit	Max	Limit	Max	Limit	Max	Limit	Max	Limit
0	0.26	0.00	0.60	0.00	1.04	0.00	1.26	0.00	1.41	0.00
1	0.95	0.80	1.09	0.58	1.30	0.40	1.44	0.30	1.63	0.26
2	1.66	2.33	1.47	0.98	1.57	0.60	1.63	0.42	1.79	0.31
3	2.33	2.22	1.77	1.30	1.61	0.75	1.62	0.48	1.76	0.34
4	3.01	2.90	1.99	1.53	1.79	0.81	1.77	0.52	1.87	0.36
5	3.64	3.55	2.18	1.74	1.70	0.84	1.69	0.55	1.73	0.38
6	4.22	4.16	2.25	1.81	1.75	0.85	1.73	0.55	1.88	0.38
7	4.83	4.77	2.44	1.94	1.88*	0.85	1.84*	0.55	1.93*	0.38
8	5.48	5.40	2.40	2.05	1.70	0.85	1.66	0.55	1.78	0.38
9	6.12	6.07	2.50	2.08	1.77	0.85	1.77	0.55	1.83	0.38
10	6.75	6.71	2.48	2.12	1.81	0.89	1.73	0.55	1.80	0.38
11	7.60	7.58	2.67	2.22	1.84	0.89	1.83	0.55	1.86	0.38
12	8.27	8.22	2.75*	2.28	1.85	0.89	1.76	0.55	1.87	0.38
13	8.80	8.77	2.62	2.20	1.74	0.89	1.70	0.55	1.79	0.38
14	9.33	9.30	2.67	2.32	1.74	0.89	1.66	0.55	1.71	0.38
15	9.98	9.96	2.68	2.35	1.73	0.89	1.63	0.55	1.76	0.38
16	10.55	10.54	2.67	2.35	1.78	0.89	1.72	0.55	1.86	0.38
17	11.19	11.15	2.68	2.35	1.75	0.89	1.70	0.55	1.84	0.38
18	11.76	11.72	2.74	2.35	1.75	0.89	1.69	0.55	1.80	0.38
19	12.23*	12.23	2.39	2.39	0.89	0.89	0.55	0.55	0.38	0.38

* denotes the maximum regret for each α.

Since there is no underfitting when $k_0=0$, taking the maximum of the first $K-1$ values in the column "Max" and the K-th value in the column "Limit" in Table 1, we can obtain

$$\max_{0\leq k_0\leq K} \delta R^*(\alpha, K-k_0) ,$$

Table 2. The maximum regret, $\max \delta R(\hat{k}_\alpha)$ when $n-K$ is large

α	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$
1.0	0.80	2.33	2.22	2.90	3.55	4.16	4.77	5.40	6.07
1.3	0.73	1.34	1.90	2.42	2.88	3.29	3.72	4.14	4.50
1.5	0.69	1.23	1.72	2.18	2.51	2.79	3.07	3.43	3.68
1.8	0.63	1.10	1.52	1.86	2.17	2.37	2.56	2.79	2.79
2.0	0.60*	1.09*	1.47*	1.77	1.99	2.18	2.25	2.44	2.44
2.3	0.78	1.14	1.51	1.71	1.87	1.91	1.93	2.12	1.97
2.5	0.85	1.18	1.51	1.62	1.80	1.78	1.82	2.01	2.01
2.8	0.96	1.28	1.52	1.60*	1.74*	1.74*	1.75*	1.89	1.89
3.0	1.04	1.30	1.57	1.61	1.79	1.79	1.79	1.88	1.88
3.3	1.15	1.36	1.58	1.63	1.79	1.79	1.79	1.87	1.87
3.5	1.26	1.44	1.63	1.63	1.77	1.77	1.77	1.84*	1.84*
3.8	1.40	1.54	1.73	1.73	1.81	1.81	1.81	1.90	1.90
4.0	1.41	1.63	1.79	1.79	1.87	1.87	1.88	1.93	1.93
5.0	1.94	2.04	2.11	2.11	2.25	2.25	2.25	2.29	2.29

* denotes the minimax value for each K.

In Table 2, the α runs more densely than in Table 1, but the case $K>9$ is omitted to save the space. The obtained approximate solution is 2.0 for $1\leqq K\leqq 3$, 2.8 for $4\leqq K\leqq 7$, and 3.5 for $K\geqq 8$.

Table 3. The maximum regret, $\max \delta R(\hat{k}_\alpha)$ when $n-K=2$

α	$K=1$	$K=5$	$K=6$	$K=9$	$K=10$	$K=11$	$K=17$	$K=18$	$K=19$
1.0	0.82	3.71	4.38	6.50	7.16	7.94	11.65	12.26	12.91
2.0	0.65*	2.76*	4.63	4.93	5.38	7.61	8.15	8.49	8.86
3.0	1.01	2.79	3.07	4.11	4.34	4.66	6.63	6.85	7.15
4.0	1.38	2.83	3.03*	3.95*	4.09*	4.27	5.74	5.90	6.29
5.0	1.75	3.06	3.14	3.99	4.09*	4.23*	5.41	5.50	5.84
6.0	2.12	3.24	3.31	4.02	4.14	4.24	5.27*	5.38	5.59
7.0	2.47	3.55	3.62	4.19	4.28	4.34	5.28	5.36*	5.54*
8.0	2.84	3.82	3.91	4.37	4.47	4.56	5.43	5.55	5.65

* denotes the minimax value for each K.

Table 3 is a part of the results for the case when $n-K=2$. This is an extreme case, since $n-K$ has to be greater than 1 in view of degree of freedom. The obtained solution among $\alpha=1,\cdots,8$ is 2 for $1\leqq K\leqq 5$, 4 for $6\leqq K\leqq 10$, 5 for $K=11$, 6 for $12\leqq K\leqq 17$, and 7 for 18

$\leqq K \leqq 19$. As was proved in Theorem 3.1, the solution will diverge to infinity with K.

Table 4. The maximum regret, $\max \delta R(\hat{k}_\alpha)$ when $n-K=5$

α	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=10$	$K=11$	$K=19$
1.0	0.81	1.56	2.28	2.93	3.57	4.18	6.87	3.99	12.65
2.0	0.61*	1.16*	1.60*	2.06	2.41	2.74	3.86	3.99	6.30
3.0	1.01	1.43	1.71	2.01*	2.27*	2.42	3.05	3.13	4.25
4.0	1.43	1.75	1.90	2.09	2.33	2.33*	2.76*	2.81	3.41
5.0	1.85	2.10	2.21	2.29	2.50	2.41	2.77	2.79*	3.07*
6.0	2.28	2.49	2.56	2.66	2.79	2.68	3.03	2.89	3.11

* denotes the minimax value for each K.

The case when $n-K=5$ is also simulated. A part of the results are placed in Table 4. The solution is 2 for $1\leqq K\leqq 3$, 3 for $4\leqq K\leqq 5$, 4 for $6\leqq K\leqq 10$ and 5 for $11\leqq K\leqq 19$. The solutions are smaller than those in the case of $n-K=2$, but, still greater than those in case of large enough $n-K$. It is intuitively clear that the minimax solution becomes large as a decrease of $n-K$, to compensate such tendencies of overfitting, since the estimation error of $\tilde{\sigma}^2(K)$ leads the FPE procedure to select an overfitted model. As a final remark, if the range of selection is of the form $\underline{k}\leqq k\leqq \bar{k}$, the same solution is available only by replacing K by $\bar{k}-\underline{k}$.

Acknowledgements

This work was partly done during the author's stay in University of Pittsburgh and in Mathematical Sciences Research Institute, Berkeley. The author would like to express his sincere thanks to Professor P. R. Krishnaiah for his encouragement and for making this study possible. Suggestions by Dr. K. Subramanyam were helpful. Discussion with Professor C. J. Stone was helpful for improving the original manuscript.

This research was partly supported by the National Science Foundation, Grant MCS-812-0790.

Appendix: Proof of Lemma 3.1

Without loss of generality we may assume $\mu>0$. It is easy to show $(\partial/\partial a)f_\mu(a)\geqq 0$ on $0\leqq a\leqq a^*(\mu)$ for $\mu^2>1/2$. Therefore, for positiveness of $f_\mu(a)$ it is enough to show

$$f_\mu(a)>0 \qquad \text{for } a>\mu>1 .$$

This is because $a^*(\mu)\geqq\mu$ as long as $\mu>1$. If $\mu<a\leqq 2\mu$

$$f_\mu(a) > \int_0^\infty (\mu^2 - x^2)\phi(x)dx = (\mu^2 - 1)/2$$

and if $a \geqq 2\mu$

$$f_\mu(a) > \int_{-\infty}^\infty (\mu^2 - x^2)\phi(x)dx = \mu^2 - 1 .$$

We now prove (3.6). For $\mu \geqq a$,

$$\begin{aligned} f_\mu(a) &= \int_{-a}^a (-x^2 + 2x\mu)\phi(x-\mu)dx \\ &\leqq a\{2(\mu - a) + a\} \int_0^a x\phi(x-\mu)dx \\ &\leqq 2a^2(\mu - a)\phi(\mu - a) + a^2/2 \\ &\leqq a^2\{2\phi(1) + 1/2\} \leqq a^2 . \end{aligned}$$

For $\mu \leqq a$,

$$f_\mu(a) \leqq \mu^2 \int_0^a \phi(x-\mu)dx \leqq a^2 .$$

Therefore the right hand side of (3.6) follows. Putting $\mu = a$, we have

$$f_a(a) = \int_0^{2a} (a^2 - x^2)\phi(x)dx \geqq a^2\{\Phi(2a) - 1/2\} - 1/2 .$$

The left hand side of (3.6) then follows.

KEIO UNIVERSITY, DEPARTMENT OF MATHEMATICS

REFERENCES

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

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

[3] Bhansali, R. J. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion, *Biometrika*, **64**, 547-551.

[4] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications II*, John Wiley and Sons, New York.

[5] Golub, G. H. (1965). Numerical methods for solving linear least squares problems, *Numer. Math.*, **7**, 206-216.

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

[7] Hosoya, Y. (1983). Information criteria and tests for time-series models, *In Time Series Analysis: Theory and Practice 5*, pp. 39-52, (ed. O. D. Anderson), North-Holland, Amsterdam and New York.

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

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

[10] Shibata, R. (1981). An optimal selection of regression variables, *Biometrika*, **68**, 45-54; Correction, **69**, 492.
[11] Shibata, R. (1983). Asymptotic efficiency of a selection of regression variables, *Ann. Inst. Statist. Math.*, **35**, 415-423.
[12] Shibata, R. (1984). Approximate efficiency of a selection procedure for the number of regression variables, *Biometrika*, **71**, 43-49.
[13] Stone, C. (1981). Admissible selection of an accurate and parsimonious normal linear regression model, *Ann. Statist.*, **9**, 475-485.