The Annals of Statistics
1980, Vol. 8, No. 1, 147–164

ASYMPTOTICALLY EFFICIENT SELECTION OF THE ORDER OF THE MODEL FOR ESTIMATING PARAMETERS OF A LINEAR PROCESS

BY RITEI SHIBATA

Tokyo Institute of Technology

Let $\{x_t\}$ be a linear stationary process of the form $x_t + \Sigma_{1 \leq i < \infty} a_i x_{t-i} = e_t$, where $\{e_t\}$ is a sequence of i.i.d. normal random variables with mean 0 and variance σ^2. Given observations $x_1, \cdots, x_n$, least squares estimates $\hat{a}(k)$ of $a' = (a_1, a_2, \cdots)$, and $\hat{\sigma}_k^2$ of σ^2 are obtained if the kth order autoregressive model is assumed. By using $\hat{a}(k)$, we can also estimate coefficients of the best predictor based on k successive realizations. An asymptotic lower bound is obtained for the mean squared error of the estimated predictor when k is selected from the data. If k is selected so as to minimize $S_n(k) = (n + 2k)\hat{\sigma}_k^2$, then the bound is attained in the limit. The key assumption is that the order of the autoregression of $\{x_t\}$ is infinite.

1. Introduction. Methods of estimating parameters of time series have been developed by Hannan (1969), Box and Jenkins (1970), Parzen (1974), Anderson (1977) and others. These methods are based on the assumption that the data come from an autoregressive or moving average or autoregressive moving average process of known order, but it would be rare that such assumption can be justified. A more reasonable assumption would be that the data belong to a linear stationary process as defined in Section 2, that is, an infinite order autoregressive process. The estimation of parameters and spectral density of these processes has been investigated by Parzen (1974, 1975), Berk (1974), Huzii (1977), Shibata (1977) and Bhansali (1978). In these papers, the estimates of parameters are the least squares estimates obtained by fitting a kth order autoregressive model, where unestimated parameters are set at 0.

In Section 2, we will show that the above estimation is also that of coefficients of the best predictor based on k past observations. We can then obtain a predictor by using the estimated coefficients if the parameters are unknown. In order to reduce the mean squared error, we have to select the order k of the model.

Several selection methods have been proposed for finite autoregressive or autoregressive moving average process, for example, the final prediction error (FPE) method proposed by Akaike (1970), Akaike's information criterion (AIC) method (Akaike (1973a, b, 1974)) and the criterion autoregressive transfer function (CAT) method proposed by Parzen (1974). Although some properties of these methods have been investigated by Akaike (1937a), Shibata (1976), Gersch and Sharpe (1973), Tong (1975) and others, the statistical optimality has not been made so clear.

Received February 1978; revised August 1978.

AMS 1970 *subject classifications.* Primary 62M10; secondary 62M20, 62E20.

Key words and phrases. Autoregression, time series models, prediction, efficiency, model selection.

In Section 3, we obtain, assuming an infinite order autoregressive process, an asymptotic lower bound for the mean squared error of prediction when the order of the model is selected from the data. Furthermore an asymptotically efficient selection is proposed in Section 4, which attains the lower bound in the limit. It is verified that if FPE or AIC method is applied to our process they are also asymptotically efficient.

2. Estimation of parameters for prediction. Consider a Gaussian process $\{x_t\}$;

$$x_t + a_1 x_{t-1} + \cdots = e_t, \quad t = \cdots, -1, 0, 1, \cdots, \tag{2.1}$$

where $a_1, a_2, \cdots$ are real numbers and $\{\cdots, e_{-1}, e_0, e_1, \cdots\}$ is a sequence of independent, normally distributed random variables with means 0 and variances $\sigma^2 > 0$.

Assume the associated power series

$$A(z) = 1 + a_1 z + a_2 z^2 + \cdots$$

converges and is not zero for $|z| \leqslant 1$. The process $\{x_t\}$ is stationary and has the moving average representation

$$x_t = e_t + b_1 e_{t-1} + \cdots, \tag{2.2}$$

where $B(z) = 1/A(z) = 1 + b_1 z + b_2 z^2 + \cdots$.

Let us denote the autocovariance by $r_l = E(x_t x_{t+l})$ and the $k \times k$ covariance matrix by

$$R(k) = (r_{ij}, 1 \leqslant i, j \leqslant k),$$

where $r_{ij} = r_{|i-j|}$.

$$V = \{\alpha; \alpha' = (\alpha_1, \alpha_2, \cdots), \|\alpha\|_R < \infty\}$$

is the vector space with norm

$$\|\alpha\|_R = (\Sigma_{1 \leqslant i,j < \infty} \alpha_i \alpha_j r_{|i-j|})^{\frac{1}{2}}.$$

Consider the projection

$$a(h, k)' = (0, \cdots, 0, a_h(h, k), \cdots, a_{h+k-1}(h, k), 0, \cdots)$$

of the parameter $a' = (a_1, a_2, \cdots)$ on the $h + k - 1$ dimensional subspace

$$V(h, k) = \{\alpha; \alpha' = (0, \cdots, 0, \alpha_h, \alpha_{h+1}, \cdots, \alpha_{h+k-1}, 0, \cdots)\}$$

of V. Then the best predictor of x_{t+h} from $\{x_{t-k+1}, \cdots, x_t\}$ is given by

$$\hat{x}_{t+h} = E(x_{t+h} | x_{t-k+1}, \cdots, x_t)$$
$$= -\Sigma_{1 \leqslant i \leqslant h+k-1}\, a_i(h, k) x_{t+h-i}.$$

The vector $a(h, k)$ is specified by the equations

$$\Sigma_{h \leqslant j \leqslant h+k-1}\, r_{|i-j|} a_j(h, k) = -r_i,$$
$$i = h, h+1, \cdots, h+k-1.$$

Given observations $x_1, \cdots, x_n$, an estimate of $a(h, k)$ is a solution

$$\hat{a}(h, k)' = (0, \cdots, 0, \hat{a}_h(h, k), \cdots, \hat{a}_{h+k-1}(h, k), 0, \cdots)$$

of a set of equations

(2.3) $$\Sigma_{h \leqslant j \leqslant h+k-1} \hat{r}_{|i-j|} \hat{a}_j(h, k) = -\hat{r}_i, \qquad i = h, h+1, \cdots, h+k-1,$$

where $n \geqslant h + k$ and

$$\hat{r}_l = \Sigma_{1 \leqslant t \leqslant n-l} x_t x_{t+l} / (n - l), \qquad l = 0, 1, \cdots, n-1.$$

From (2.3), the well-known Yule-Walker equations, $\hat{a}(h, k)$ may be also thought of as an estimate of the parameter

$$\alpha' = (0, \cdots, 0, \alpha_h, \cdots, \alpha_{h+k-1}, 0, \cdots),$$

when a finite order autoregressive model

(2.4) $$x_{t+h} + \alpha_h x_t + \cdots + \alpha_{h+k-1} x_{t-k+1} = \varepsilon_{t+h}$$

was fitted to the observations $x_1, \cdots, x_n$. Here $\{\varepsilon_t\}$ is a sequence of i.i.d. random variables with means 0 and finite variances. Consequently an estimate $\hat{a}(h, k)$ of the coefficients of the predictor based on k observations is also that of the parameters obtained by fitting the model (2.4). When $h = 1$, $\hat{a}(1, k)$ is a common estimate of a obtained by the kth order autoregressive model fitting, which is originally an estimate of the one-step ahead predictor based on k past realizations.

The goodness of $\hat{a}(h, k)$ is evaluated by the mean squared error of prediction as defined below. By using $\hat{a}(h, k)$, we obtain an h-step ahead predictor

$$\hat{y}_{t+h} = -\Sigma_{1 \leqslant i \leqslant h+k-1} \hat{a}_i(h, k) y_{t+h-i}$$

from a realization $\{y_{t-k+1}, \cdots, y_t\}$, which is independent of $\{x_t\}$ but has the same probabilistic structure.

The mean squared error of $\hat{y}_{t+h}$ is

(2.5) $$\begin{aligned} E^y(\hat{y}_{t+h} - y_{t+h})^2 &= \|\hat{a}(h, k) - a\|_R^2 + \sigma^2 \\ &= \|\hat{a}(h, k) - a(h, k)\|_R^2 + \|a(h, k) - a(h, \infty)\|_R^2 \\ &\quad + \|a(h, \infty) - a\|_R^2 + \sigma^2 \end{aligned}$$

where E^y denotes the expectation with respect to $\{y_t\}$, and $a(h, \infty)$ is the projection of a on the subspace

$$V(h, \infty) = \{\alpha; \alpha' = (0, \cdots, 0, \alpha_h, \alpha_{h+1}, \cdots)\}$$

(see Akaike (1970), Shibata (1976, 1977) and Bhansali (1978)).

If h and k are fixed, then the nonzero first k-coordinates of $n^{1/2}(\hat{a}(h, k) - a(h, k))$ are asymptotically normally distributed with mean 0 and covariance $\Sigma(k) = \sigma^2(h, k) R(k)^{-1}$ where $\sigma^2(h, k) = r_0 - \|a(h, k)\|_R^2$. It is seen that $\Sigma(k)^{-1}$ is identical to the Fisher information matrix of fitted model (2.4) assumed Gaussian with parameters $\alpha = a(h, k)$ and $E\varepsilon_t^2 = \sigma^2(h, k)$. Then $\hat{a}(h, k)$ has a high efficiency if the model (2.4) is a close approximation to the process (2.1), as was shown by Huzii (1977).

The first term of the right-hand side of (2.5) signifies the variance normalized by $R(k)^{-1}$ of the estimate. The second term is the bias of the estimate and the remaining terms are the prediction errors independent of n and k. If k is fixed, the first term converges to zero with order of magnitude $1/n$, but the second term is independent of n and not zero unless $\{x_t\}$ is an autoregressive process with the order lower than $h + k - 1$. Therefore, we have to select k so as to balance the first and second terms for each n.

3. Asymptotic efficiency of a selection of the order of the model. For simplicity we consider only the case $h = 1$ and denote $a(1, k)$ as $a(k)' = (a_1(k), a_2(k), \cdots, a_k(k), 0, \cdots)$. Clearly $a = a(1, \infty)$. Suppose that the order k is selected from a given range $1 \leqslant k \leqslant K_n (K_n < n)$. Given $x_1, \cdots, x_n$, the sample autocovariance vector and the matrix are defined by

$$\hat{r}(k)' = (\hat{r}_{1\,0}, \hat{r}_{2\,0}, \cdots, \hat{r}_{k\,0})$$

and

$$\hat{R}(k) = (\hat{r}_{lm}, 1 \leqslant l, m \leqslant k),$$

$$\hat{r}_{lm} = \Sigma_{K_n \leqslant t \leqslant n-1}\, x_{t+1-l} x_{t+1-m}/N \qquad 0 \leqslant l, m \leqslant K_n,$$

where $N = n - K_n$.

If the kth order model is applied, the least squares estimate

$$\hat{a}(k)' = (\hat{a}_1(k), \hat{a}_2(k), \cdots, \hat{a}_k(k))$$

of the regression parameters of the model is a solution of the equation

$$\hat{R}(k)\hat{a}(k) = -\hat{r}(k).$$

Since $\hat{a}(k)$ is asymptotically equivalent to $\hat{a}(1, k)$, for the convenience of evaluations, $\hat{a}(k)$ will be used as an estimate of $a(k)$, $k = 1, \cdots, K_n$, which are sometimes regarded as K_n-dimensional or infinite dimensional random vectors with undefined entries 0.

Define

$$e_{t+1,k} = x_{t+1} + a_1(k)x_t + \cdots + a_k(k)x_{t+1-k}$$

and

$$s_k^2 = \Sigma_{K_n \leqslant t \leqslant n-1}\, e_{t+1,k}^2/N.$$

Then an estimate of

$$\sigma_k^2 = \min_{c_1, \cdots, c_k} E(x_{t+1} + c_1 x_t + \cdots + c_k x_{t+1-k})^2$$

is given by

$$\hat{\sigma}_k^2 = \Sigma_{K_n \leqslant t \leqslant n-1}\, (x_{t+1} + \hat{a}_1(k)x_t + \cdots + \hat{a}_k(k)x_{t+1-k})^2/N. \tag{3.1}$$

The norm

$$\|\alpha\|_A = (\alpha' A\alpha)^{1/2}$$

is defined for any positive definite matrix A and the norm of A itself is defined by

$$\|A\| = \sup_{\|\alpha\| \leqslant 1}\|A\alpha\|,$$

where $\|\alpha\|$ is the Euclidean norm of the vector α.

Assumptions.

(A.1) $\{x_t\}$ is a stationary Gaussian process which satisfies the equation (2.1) and $\Sigma_{1\leqslant j<\infty}|a_j| < \infty$.

(A.2) $A(z)$ is nonzero for $|z| \leqslant 1$.

(A.3) $\{K_n\}$ is a sequence of positive integers such that $K_n \to \infty$ and $K_n/n^{1/2} \to 0$ as $n \to \infty$.

(A.4) $\{x_t\}$ is not degenerate to a finite order autoregressive process.

Under assumptions (A.1) and (A.2), the spectral density of $\{x_t\}$ is bounded and bounded away from zero and the norms of the covariance matrices are

$$0 < r_0 = \|R(1)\| \leqslant \|R(2)\| \leqslant \cdots \leqslant \|R\| < \infty,$$

where $R = (r_{|i-j|}, 1 \leqslant i, j < \infty)$ is the infinite dimensional covariance matrix with the norm

$$\|R\| = \sup_{\|\alpha\|\leqslant 1}\left(\Sigma_{1\leqslant i<\infty}(\Sigma_{1\leqslant j<\infty} r_{ij}\alpha_j)^2\right)^{1/2}.$$

By Wiener's theorem (Zygmund (1959, page 245)), the coefficients $b_1, b_2, \cdots$ of the moving average (2.2) are absolutely convergent. Then we have

$$\Sigma_{0\leqslant j<\infty}|r_j| < \infty.$$

We need the following lemmas for obtaining the asymptotic behaviour of

$$\hat{a}(k) - a(k) = -\hat{R}(k)^{-1}(\Sigma_{K_n\leqslant t\leqslant n-1} X_t(k)e_{t+1,k}/N),$$

where

$$X_t(k)' = (x_t, \cdots, x_{t+1-k}).$$

LEMMA 3.1. *If $\{x_t\}$ is a Gaussian stationary process, then for any $1 \leqslant k \leqslant K_n$,*

$$NE\|\Sigma_{K_n\leqslant t\leqslant n-1} X_t(k)(e_{t+1,k} - e_{t+1})/N\|^2 \leqslant k\|a - a(k)\|^2\|R\|(\Sigma_{-\infty<j<\infty}|r_j| + \|R\|),$$

where $r_{-j} = r_j$ $(j = 1, 2, \cdots)$.

PROOF. Putting $\delta_m = a_m(k) - a_m$, we have

(3.2) $$E\|\Sigma_{K_n\leqslant t\leqslant n-1} X_t(k)(e_{t+1,k} - e_{t+1})\|^2 = \Sigma_{1\leqslant l\leqslant k}\Sigma_{K_n\leqslant t_1,t_2\leqslant n-1}\Big\{(\Sigma_{1\leqslant m<\infty} r_{lm}\delta_m)^2 + r_{t_1t_2}\Sigma_{1\leqslant m_1,m_2<\infty}\delta_{m_1}r_{t_1-m_1,t_2-m_2}\delta_{m_2} + \left(\Sigma_{1\leqslant m<\infty} r_{t_1-l,t_2-m}\delta_m\right)^2\Big\}.$$

Since $a(k)$ is the projection on a, the first summand of the right-hand side of (3.2) vanishes. We obtain the desired result from the evaluations;

$$|\Sigma_{1 \leqslant m_1, m_2 < \infty} \delta_{m_1} r_{t_1 - m_1, t_2 - m_2} \delta_{m_2}| \leqslant \|R\| \, \|\delta\|^2$$

and

$$\Sigma_{K_n \leqslant t_1 \leqslant n-1} \big(\Sigma_{1 \leqslant m < \infty} r_{t_1 - l_1, t_2 - m} \delta_m\big)^2 \leqslant \|R\|^2 \|\delta\|^2.$$

LEMMA 3.2. *Assume* (A.1) *and* (A.2). *Then*

$$E\big(N \|\Sigma_{K_n \leqslant t \leqslant n-1} X_t(k) e_{t+1}/N\|^2_{R(k)^{-1}} - k\sigma^2\big)^2 = 2k\sigma^4 + O(1/N)k^2.$$

PROOF. First we evaluate

$$\Sigma_{t_1, \cdots, t_4} E(x_{t_1+1-l_1} \cdots x_{t_4+1-l_4} e_{t_1+1} \cdots e_{t_4+1}). \tag{3.3}$$

From the Gaussian property, each summand of (3.3) is the sum of the products of moments of the pairs. Let ρ and τ be permutations on (1, 2, 3, 4). We evaluate (3.3) by dividing into the following cases.

(i) $\Sigma_{t_1, \cdots, t_4} E(x_{t_1+1-l_1} e_{t_{\rho(1)}+1}) \cdots E(x_{t_4+1-l_4} e_{t_{\rho(4)}+1})$.
Since $t_i + 1 - l_i < t_{\rho(i)} + 1$ for some i, all terms of (i) are zero.

(ii) $\Sigma_{t_1, \cdots, t_4} E(x_{t_1+1-l_1} x_{t_2+1-l_2}) E(e_{t_{\rho(1)}+1} e_{t_{\rho(2)}+1}) E(x_{t_3+1-l_3} e_{t_{\rho(3)}+1})$
$E(x_{t_4+1-l_4} e_{t_{\rho(4)}+1})$. If $(\rho(1), \rho(2)) = (1, 2)$ or $(2, 1)$, all terms of (ii) are zero. Otherwise, as the same evaluations follow, we may consider the ρ such that $\rho(1) = 1$, $\rho(2) = 3$, $\rho(3) = 4$ and $\rho(4) = 2$. For such ρ, (ii) is rewritten

$$\sigma^6 \Sigma_{t_1, \cdots, t_4} r_{t_1 - l_1, t_2 - l_2} \delta_{t_1 t_3} b_{t_3 - l_3 - t_4} b_{t_4 - l_4 - t_2}$$

where $\delta_{t_1 t_3}$ is Kronecker's delta, $b_0 = 1$ and $b_i = 0$ for $i < 0$. By simple evaluation, the above is bounded by

$$N\sigma^6 \Sigma_{0 \leqslant i < \infty} |b_i| \big(\Sigma_{-\infty < i < \infty} r_i^2 \cdot \Sigma_{0 \leqslant i < \infty} b_i^2\big)^{1/2}.$$

(ii) The same evaluation holds even when $t_1, \cdots, t_4$ are arbitrarily permutated in summands of (ii).

(iii) $\Sigma_{t_1, \cdots, t_4} E(x_{t_{\rho(1)}+1-l_{\rho(1)}} x_{t_{\rho(2)}+1-l_{\rho(2)}}) E(x_{t_{\rho(3)}+1-l_{\rho(3)}} x_{t_{\rho(4)}+1-l_{\rho(4)}}) E(e_{t_{\tau(1)}} e_{t_{\tau(2)}}) E(e_{t_{\tau(3)}} e_{t_{\tau(4)}})$, where $\rho(1) < \rho(2)$, $\rho(3) < \rho(4)$, $\tau(1) < \tau(2)$ and $\tau(3) < \tau(4)$.
If $\rho(i) = \tau(i)$, $i = 1, \cdots, 4$, then (iii) reduces to $N^2\sigma^4 r_{l_{\rho(1)} l_{\rho(2)}} r_{l_{\rho(3)} l_{\rho(4)}}$, where $\rho(1) < \rho(2)$ and $\rho(3) < \rho(4)$. It is bounded by $N\sigma^4 \Sigma_{-\infty < i < \infty} r_i^2$, otherwise.

Next, for $R(k)^{-1} = (r^{lm}, 1 \leqslant l, m \leqslant k)$, the following identity holds

$$r^{lm} = \Sigma_{1 \leqslant p \leqslant k} a_{p-l}(p-1) a_{p-m}(p-1)/\sigma^2_{p-1}$$

where $\sigma_0^2 = r_0$, $a_0(p) = 1$ and $a_i(p) = 0$ for $p \geqslant 0$ and $i < 0$. As was shown in Lemma 4 of Berk (1974),

$$\Sigma_{1 \leqslant l \leqslant k} |a_l(p)|, \qquad k = 1, 2, \cdots$$

are bounded uniformly in p. Then there exists $C > 0$ such that

$$\Sigma_{1 \leqslant l, m \leqslant k} |r^{lm}| \leqslant Ck.$$

Combining this result and evaluations (i) $\sim$ (iii), we obtain

$$E \|\Sigma_{K_n \leqslant t \leqslant n-1} X_t(k) e_{t+1}\|^4_{R(k)^{-1}}$$
$$= N^2 \sigma^4 \Sigma_\rho \Sigma_{1 \leqslant l_1, \cdots, l_4 \leqslant k} r^{l_1 l_2} r^{l_3 l_4} r_{l_{\rho(1)} l_{\rho(2)}} r_{l_{\rho(3)} l_{\rho(4)}} + O(N) k^2$$
$$= N^2 \sigma^4 (k^2 + 2k) + O(N) k^2,$$

where the summation Σ_ρ extends over all permutations such that $\rho(1) < \rho(2)$ and $\rho(3) < \rho(4)$. Noting

$$E \|\Sigma_{K_n \leqslant t \leqslant n-1} X_t(k) e_{t+1}\|^2_{R(k)^{-1}} = Nk\sigma^2,$$

we have

$$E\left(\|\Sigma_{K_n \leqslant t \leqslant n-1} X_t(k) e_{t+1}\|^2_{R(k)^{-1}} - Nk\sigma^2\right)^2 = 2N^2 k \sigma^4 + O(N) k^2.$$

The proof is complete.

LEMMA 3.3. *Under assumptions* (A.1) $\sim$ (A.3), *it holds that*

$$\text{p-lim}_{n \to \infty} \left(\max_{1 \leqslant k \leqslant K_n} \|\hat{R}(k) - R(k)\|\right) = 0$$

and

$$\text{p-lim}_{n \to \infty} \left(\max_{1 \leqslant k \leqslant K_n} \|\hat{R}(k)^{-1} - R(k)^{-1}\|\right) = 0,$$

where p-lim *means the limit in probability.*

PROOF. It is easy to verify that

$$\max_{1 \leqslant k \leqslant K_n} \|\hat{R}(k) - R(k)\|^2 \leqslant \Sigma_{1 \leqslant i, j \leqslant K_n} (\hat{r}_{ij} - r_{ij})^2$$

and

$$\Sigma_{1 \leqslant i, j \leqslant K_n} E(\hat{r}_{ij} - r_{ij})^2 \leqslant \text{const } K_n^2 / N.$$

The first assertion of the lemma follows from Assumption (A.3), and the last assertion is proved in the same way as in the proof of Lemma 3 of Berk (1974).

PROPOSITION 3.1. *Let* $\{k_n\}$ *be a sequence of integers such that* $1 \leqslant k_n \leqslant K_n$ *and*

(3.4) $$\lim_{n \to \infty} k_n = \infty.$$

Assume (A.1) $\sim$ (A.3). *Then*

$$\text{p-lim}_{n \to \infty} (N / k_n) \|\hat{a}(k_n) - a(k_n)\|^2_R = \sigma^2.$$

PROOF. (3.4) implies that

$$\lim_{n\to\infty}\|a - a(k_n)\| = 0.$$

Applying Lemmas 3.1 $\sim$ 3.3 and Chebyshev's inequality, we have the desired result.

As was seen in (2.5), the relation

$$\|\hat{a}(k) - a\|_R^2 = \|\hat{a}(k) - a(k)\|_R^2 + \|a(k) - a\|_R^2$$

holds, which implies the following corollary, where

$$L_n(k) = k\sigma^2/N + \|a(k) - a\|_R^2.$$

COROLLARY 3.1.

$$\text{p-lim}_{n\to\infty}\big(\|\hat{a}(k_n) - a\|_R^2/L_n(k_n)\big) = 1.$$

Corollary 3.1 shows that the behaviour of $\|\hat{a}(k_n) - a\|_R^2$ is asymptotically equal to that of $L_n(k_n)$. The first term of $L_n(k)$ corresponds to the variance of $\hat{a}(k)$ and the second term to the bias.

DEFINITION 3.1. $\{k_n^*\}$ is a sequence of positive integers which attain the minimum of $L_n(k)$ for each n;

$$L_n(k_n^*) = \min_{1\leqslant k\leqslant K_n} L_n(k).$$

Here, if $K_n \to \infty$ and $K_n/N \to 0$ as $n \to \infty$, then $L_n(K_n)$ converges to zero. Thus $L_n(k_n^*)$ also converges to zero and k_n^* diverges to infinity as $n \to \infty$.

THEOREM 3.1. *Assume* (A.1) $\sim$ (A.4). *Then for any sequence* $\{k_n\}$ *such that* $1 \leqslant k_n \leqslant K_n$, *and for any* $\varepsilon > 0$,

$$\lim_{n\to\infty} P\big(\|\hat{a}(k_n) - a\|_R^2/L_n(k_n^*) \geqslant 1 - \varepsilon\big) = 1.$$

PROOF. We can choose a divergent sequence of integers $\{k_n^{**}\}$, $1 \leqslant k_n^{**} \leqslant K_n$ such that $L_n(k_n)/L_n(k_n^*) \to \infty$ as $n \to \infty$ for any $k_n < k_n^{**}$. If $k_n < k_n^{**}$, then $\|\hat{a}(k_n) - a\|_R^2/L_n(k_n^*)$ diverges in probability, for $\|\hat{a}(k_n) - a\|_R^2/L_n(k_n)$ is bounded in probability. Otherwise the result is clear from Corollary 3.1.

Corollary 3.1 shows that the sequence $\{k_n^*\}$ asymptotically minimizes $\|\hat{a}(k) - a\|_R^2$. However, as $\{k_n^*\}$ is a function of the parameters σ^2 and a, k_n^* must be estimated from observations. In the remainder of this section, we extend Theorem 3.1 to the case where k_n is a random variable depending on the observations $x_1, \cdots, x_n$.

LEMMA 3.4 (*A strong version of Lemma* 3.2). *Assume* (A.1) $\sim$ (A.2). *Then*

$$E\big(N\|\Sigma_{K_n\leqslant t\leqslant n-1} X_t(k)e_{t+1}/N\|_{R(k)^{-1}}^2 - k\sigma^2\big)^4 = (48k + 12k^2)\sigma^8 + O(1/N)k^4.$$

PROOF. As in the proof of Lemma 3.2, we have

$$E\|\Sigma_{K_n\leqslant t\leqslant n-1} X_t(k)e_{t+1}\|_{R(k)^{-1}}^8 = N^4\sigma^8(k^4 + 12k^3 + 44k^2 + 48k) + O(N^3)k^4$$

and

$$E\|\Sigma_{K_n \leq t \leq n-1} X_t(k)e_{t+1}\|^6_{R(k)^{-1}} = N^3\sigma^6(k^3 + 6k^2 + 8k) + O(N^2)k^3.$$

The lemma follows from the evaluations that

$$E\|\Sigma_{K_n \leq t \leq n-1} X_t(k)e_{t+1}\|^4_{R(k)^{-1}} = N^2\sigma^4(k^2 + 2k) + O(N)k^2$$

and

$$E\|\Sigma_{K_n \leq t \leq n-1} X_t(k)e_{t+1}\|^2_{R(k)^{-1}} = N\sigma^2 k.$$

PROPOSITION 3.2. *Assume* (A.1) $\sim$ (A.4). *Then*

$$\text{p-lim}_{n\to\infty}\left(\max_{1 \leq k \leq K_n}|\, \|\hat{a}(k) - a\|^2_R / L_n(k) - 1|\right) = 0.$$

PROOF. From Lemma 3.4 and Definition 3.1, there exists $C > 0$,

$$\begin{aligned}
&\Sigma_{1 \leq k \leq K_n} E\left\{\left(\|\Sigma_{K_n \leq t \leq n-1} X_t(k)e_{t+1}/N\|^2_{R(k)^{-1}} - k\sigma^2/N\right)/L_n(k)\right\}^4 \\
(3.5)\qquad &\leq \sigma^8 \Sigma_{1 \leq k \leq K_n}\left\{(48k + 12k^2 + Ck^4/N)/(NL_n(k))^4\right\} \\
&\leq 12\sigma^8\Sigma_{1 \leq k \leq K_n}\left\{(4k + k^2)/(NL_n(k))^4\right\} + CK_n/N \\
&\leq 60/k_n^* + 60\Sigma_{k_n^* < k \leq K_n}(1/k^2) + CK_n/N.
\end{aligned}$$

Then the left-hand side of (3.5) converges to zero as $n \to \infty$, for k_n^* diverges and $K_n = o(N^{\frac{1}{2}})$. Furthermore Lemma 3.1 implies that

$$\begin{aligned}
&\Sigma_{1 \leq k \leq K_n} E\Big[\big\{\|\Sigma_{K_n \leq t \leq n-1} X_t(k)e_{t+1,k}/N\|_{R(k)^{-1}} \\
(3.6)\qquad &\qquad - \|\Sigma_{K_n \leq t \leq n-1} X_t(k)e_{t+1}/N\|_{R(k)^{-1}}\big\}/L_n(k)\Big]^2 \\
&\leq \text{const } K_n^2/n.
\end{aligned}$$

Then the left-hand side of (3.6) converges to zero as $n \to \infty$. Accordingly from Lemma 3.3 and the definition of $\hat{a}(k)$, we have

$$\text{p-lim}_{n\to\infty}\left\{\max_{1 \leq k \leq K_n}\left(|\, \|\hat{a}(k) - a(k)\|^2_R - k\sigma^2/N|/L_n(k)\right)\right\} = 0.$$

The desired result follows from the identity

$$\|\hat{a}(k) - a(k)\|^2_R - k\sigma^2/N = \|\hat{a}(k) - a\|^2_R - L_n(k).$$

THEOREM 3.2 (*An extension of Theorem* 3.1). *Assume* (A.1) $\sim$ (A.4). *Then for any random variable* $\tilde{k}$ *possibly depending on* $x_1, \cdots, x_n$, *and for any* $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\|\hat{a}(\tilde{k}) - a\|^2_R / L_n(k_n^*) \geq 1 - \varepsilon\right) = 1.$$

PROOF. Applying Proposition 3.2, we have

$$\text{p-lim}_{n\to\infty}\left(\|\hat{a}(\tilde{k}) - a\|^2_R / L_n(\tilde{k})\right) = 1,$$

and the theorem is clear from Definition 3.1.

The above theorem shows that the loss $\|\hat{a}(\tilde{k}) - a\|^2_R$ of the estimate $\hat{a}(\tilde{k})$ is asymptotically never below $L_n(k_n^*)$ in probability for any order selection $\tilde{k}$. We call an order selection $\tilde{k}$ asymptotically efficient if

$$\text{p-lim}_{n\to\infty}\left(\|\hat{a}(\tilde{k}) - a\|^2_R / L_n(k_n^*)\right) = 1.$$

4. Asymptotically efficient selection of the order of the model. We propose an order selection $\hat{k}$ which is asymptotically efficient. Let $\hat{\sigma}_k^2$ be the sum of residuals given by (3.1). Then $\hat{k}$ is defined as the k which minimizes

$$S_n(k) = (N + 2k)\hat{\sigma}_k^2, \qquad 1 \leqslant k \leqslant K_n.$$

This order selection is a version of the final prediction error (FPE) method proposed by Akaike (1970) (see Example 4.2), and has a close relation to C_p method proposed by Mallows (1973). As will be seen later, small changes in $S_n(k)$ do not change the asymptotic efficiency.

The statistic $S_n(k)$ can be rewritten as

$$(4.1) \qquad S_n(k) = NL_n(k) + 2k(\hat{\sigma}_k^2 - \sigma^2) + \left(k\sigma^2 - N\|\hat{a}(k) - a(k)\|^2_{\hat{R}(k)}\right) + N\sigma^2 + N(s_k^2 - \sigma_k^2).$$

LEMMA 4.1. *Assume* (A.1) $\sim$ (A.4). *Then*

$$\text{p-lim}_{n\to\infty}\max_{1\leqslant k\leqslant K_n}\left(k|\hat{\sigma}_k^2 - \sigma^2|/NL_n(k)\right) = 0.$$

PROOF. By the definition of $\hat{\sigma}_k^2$,

$$(4.2) \qquad |\hat{\sigma}_k^2 - \sigma^2| \leqslant |\hat{\sigma}_k^2 - s_k^2| + |s_k^2 - \sigma_k^2| + |\sigma_k^2 - \sigma^2| = \|\hat{a}(k) - a(k)\|^2_{\hat{R}(k)} + |s_k^2 - \sigma_k^2| + \|a(k) - a\|^2_R.$$

Proposition 3.2 and Lemma 3.3 imply that

$$\max_{1\leqslant k\leqslant K_n}\left\{k\|\hat{a}(k) - a(k)\|^2_{\hat{R}(k)}/NL_n(k)\right\}$$

converges to zero in probability as $n \to \infty$. We have also

$$\Sigma_{1\leqslant k\leqslant K_n}E(s_k^2 - \sigma_k^2)^2 \leqslant \text{const } K_n/N.$$

Then

$$\max_{1\leqslant k\leqslant K_n} k|s_k^2 - \sigma_k^2|/NL_n(k)$$

converges to zero in probability as $n \to \infty$. The desired result follows from Assumption (A.3) and

$$\|a(k) - a\|^2_R/L_n(k) < 1.$$

Lemma 4.1 and Proposition 3.2 show that compared with the first term $NL_n(k)$, the second and third terms on the right-hand side of (4.1) are negligible uniformly in $1 \leqslant k \leqslant K_n$. Although the last two terms are not negligible, it is sufficient to show that

$$N\left\{(s_k^2 - \sigma_k^2) - (s_{k_n^*}^2 - \sigma_{k_n^*}^2)\right\}, \qquad k = 1, \cdots, K_n$$

are uniformly negligible, as the behaviour of $\hat{k}$ is determined only by the differences of $S_n(k)$.

LEMMA 4.2. *Assume* (A.1) $\sim$ (A.2). *Then there exist constants* $C_1, C_2, \cdots, C_5 > 0$ *depending only on the autocorrelations* $r_0, r_1, \cdots,$ *and such that for any real vectors* $\delta' = (\delta_0, \delta_1, \cdots, \delta_{K_n})$ *and* $\eta' = (\eta_0, \eta_1, \cdots, \eta_{K_n})$,

$$\begin{aligned} E\big(\Sigma_{0 \leqslant i,j \leqslant K_n} \delta_i(\hat{r}_{ij} - r_{ij})\eta_j\big)^4 \\ \leqslant \big(C_1\|\delta\|^3|\delta|\ \|\eta\|\ |\eta|^3 + C_2\|\delta\|^4|\eta|^4\big)/N^2 \\ + \big(C_3\|\delta\|^4\|\eta\|^4 + C_4\|\delta\|^4\|\eta\|^2|\eta|^2 + C_5\|\delta\|^2|\delta|^2|\eta|^4\big)/N^3 \end{aligned} \tag{4.3}$$

where $|\delta| = \Sigma_{i=0}^{K_n}|\delta_i|$ *and* $|\eta| = \Sigma_{i=0}^{K_n}|\eta_i|$.

PROOF. $\hat{r}_{ij}$ is an unbiased estimate of r_{ij} and the cross moment is

$$E(\hat{r}_{i_1j_1}\hat{r}_{i_2j_2}) = r_{i_1j_1}r_{i_2j_2} + \xi_2(1, 2)/N^2,$$

where

$$\begin{aligned} \xi_2(\alpha, \beta) = \Sigma_{K_n \leqslant t_\alpha, t_\beta \leqslant n-1}\big(&r_{t_\alpha - i_\alpha, t_\beta - j_\beta} r_{t_\beta - i_\beta, t_\alpha - j_\alpha} \\ &+ r_{t_\alpha - i_\alpha, t_\beta - i_\beta} r_{t_\alpha - j_\alpha, t_\beta - j_\beta}\big) \end{aligned}$$

for any $1 \leqslant \alpha, \beta \leqslant 4$ (see Berk (1974)). The higher moments of $\hat{r}_{ij}$ have been calculated by Leonov and Shiryaev (1959). The third cross moment is

$$\begin{aligned} &E(\hat{r}_{i_1j_1} \cdots \hat{r}_{i_3j_3}) \\ &\quad = r_{i_1j_1} \cdots r_{i_3j_3} + \Sigma_\rho r_{i_{\rho(1)}j_{\rho(1)}}\xi_2(\rho(2), \rho(3))/2N^2 + \xi_3(1, 2, 3)/N^3, \end{aligned}$$

where Σ_ρ extends over all permutations on (1, 2, 3), and

$$\begin{aligned} &\xi_3(\alpha, \beta, \gamma) \\ &\quad = \Sigma_{K_n \leqslant t_\alpha, t_\beta, t_\gamma \leqslant n-1}\big(\Sigma_\rho r_{t_{\rho(\alpha)} - i_{\rho(\alpha)}, t_{\rho(\beta)} - i_{\rho(\beta)}} r_{t_{\rho(\gamma)} - i_{\rho(\gamma)}, t_{\rho(\alpha)} - j_{\rho(\alpha)}} \\ &\qquad \times r_{t_{\rho(\gamma)} - j_{\rho(\gamma)}, t_{\rho(\beta)} - j_{\rho(\beta)}} \\ &\qquad + r_{t_\alpha - i_\alpha, t_\beta - j_\beta} r_{t_\beta - i_\beta, t_\gamma - j_\gamma} r_{t_\gamma - j_\gamma, t_\alpha - j_\alpha} \\ &\qquad + r_{t_\alpha - i_\alpha, t_\gamma - j_\gamma} r_{t_\beta - i_\beta, t_\alpha - j_\alpha} r_{t_\gamma - i_\gamma, t_\beta - j_\beta}\big), \\ &\hspace{20em} 1 \leqslant \alpha, \beta, \gamma \leqslant 4. \end{aligned}$$

Define

$$\begin{aligned} \xi_4 = \Sigma_{K_n \leqslant t_1, \cdots, t_4 \leqslant n-1}\Big\{&\big(\Sigma_\rho r_{t_{\rho(1)} - i_{\rho(1)}, t_{\rho(2)} - i_{\rho(2)}} r_{t_{\rho(3)} - i_{\rho(3)}, t_{\rho(4)} - i_{\rho(4)}}\big) \\ &\times \big(\Sigma_\rho r_{t_{\rho(1)} - j_{\rho(1)}, t_{\rho(2)} - j_{\rho(2)}} r_{t_{\rho(3)} - j_{\rho(3)}, t_{\rho(4)} - j_{\rho(4)}}\big)/16 \\ &+ \Sigma^*_{\rho,\tau}\big(r_{t_{\rho(1)} - i_{\rho(1)}, t_{\rho(2)} - i_{\rho(2)}} r_{t_{\rho(3)} - i_{\rho(3)}, t_{\tau(3)} - j_{\tau(3)}} \\ &\times r_{t_{\rho(4)} - i_{\rho(4)}, t_{\tau(4)} - j_{\tau(4)}} r_{t_{\tau(1)} - j_{\tau(1)}, t_{\tau(2)} - j_{\tau(2)}}\big) \\ &+ \Sigma^{**}_\rho\big(r_{t_1 - i_1, t_{\rho(1)} - j_{\rho(1)}} r_{t_2 - i_2, t_{\rho(2)} - j_{\rho(2)}} r_{t_3 - i_3, t_{\rho(3)} - j_{\rho(3)}} \\ &\times r_{t_4 - j_4, t_{\rho(4)} - j_{\rho(4)}}\big)\Big\}, \end{aligned}$$

where $\Sigma^*_{\rho,\tau}$ extends over all permutations ρ and τ on (1, 2, 3, 4) such that $\rho(1) < \rho(2)$, $\rho(3) < \rho(4)$, $\rho(3) \neq \tau(3)$ and $\rho(4) \neq \tau(4)$, and Σ^{**}_ρ extends over all permutations such that $\rho(i) \neq i$ $(i = 1, \cdots, 4)$. Then the fourth cross moment is

$$E(\hat{r}_{i_1j_1} \cdots \hat{r}_{i_4j_4}) = r_{i_1j_1} \cdots r_{i_4j_4} + \Sigma_\rho r_{i_{\rho(1)}j_{\rho(1)}} r_{i_{\rho(2)}j_{\rho(2)}} \xi_2(\rho(3), \rho(4))/4N^2 + \Sigma_{1 \leqslant l \leqslant 4} r_{i_lj_l} E(\Pi_{m \neq l;\, 1 \leqslant m \leqslant 4} \hat{r}_{i_mj_m})/N^3 + \xi_4/N^4.$$

Accordingly we have

$$\begin{aligned} E(\Pi_{1 \leqslant l \leqslant 4} (\hat{r}_{i_lj_l} - r_{i_lj_l})) &= 4(\Pi_{1 \leqslant l \leqslant 4} r_{i_lj_l})/N^3 \\ &\quad + (\Sigma_\rho r_{i_{\rho(1)}j_{\rho(1)}} \xi_3(\rho(2), \rho(3), \rho(4)))(1/N^3 - 1)/6N^3 \\ &\quad + \xi_4/N^4 + (\Sigma_\rho r_{i_{\rho(1)}j_{\rho(1)}} r_{i_{\rho(2)}j_{\rho(2)}} \xi_2(\rho(3), \rho(4)))/2N^5. \end{aligned} \tag{4.4}$$

Here,

(i) $|\Pi_{1 \leqslant l \leqslant 4}(\Sigma_{0 \leqslant i_l, j_l \leqslant K_n} \delta_{i_l} r_{i_lj_l} \eta_{j_l})| \leqslant (\|\delta\|\ \|\eta\|\ \|R\|)^4$,

(ii) $\Sigma_{0 \leqslant i_1, i_2, i_3, j_1, j_2, j_3 \leqslant K_n} (\Pi_{1 \leqslant l \leqslant 3} \delta_{i_l} \eta_{j_l}) \xi_3(1, 2, 3)| \leqslant \text{const } N\|\delta\|^2|\delta|\ |\eta|^3$.

To see this, we first have

$$\begin{aligned} &|\Sigma_{i_1, i_2, i_3, j_1, j_2, j_3} \Sigma_{t_1, t_2, t_3} (\Pi_{1 \leqslant l \leqslant 3} \delta_{i_l} \eta_{j_l}) r_{t_1 - i_1, t_2 - i_2} r_{t_3 - i_3, t_1 - j_1} r_{t_3 - j_3, t_2 - j_2}| \\ &\quad = |\Sigma_{j_1, j_2, j_3} (\Pi_{1 \leqslant l \leqslant 3} \eta_{j_l}) \Sigma_{t_1, t_2, t_3} r_{t_1 - j_1, t_3 - i_3} (\Sigma_{i_1} r_{t_1 - i_1, t_2 - j_2} \delta_{i_1}) \\ &\qquad \times (\Sigma_{i_2, i_3} \delta_{i_2} r_{t_2 - i_2, t_3 - i_3} \delta_{i_3})| \\ &\quad \leqslant N\|\delta\|^2|\delta|\ |\eta|^3 (\Sigma_{-\infty < i < \infty} r_i^2) \|R\|. \end{aligned}$$

Next,

$$\begin{aligned} &|\Sigma_{i_1, i_2, i_3, j_1, j_2, j_3} \Sigma_{t_1, t_2, t_3} (\Pi_{1 \leqslant l \leqslant 3} \delta_{i_l} \eta_{j_l}) r_{t_1 - i_1, t_2 - j_2} r_{t_2 - i_2, t_3 - j_3} r_{t_3 - i_3, t_1 - j_1}| \\ &\quad = |\Sigma_{j_1, j_2, j_3} (\Pi_{1 \leqslant l \leqslant 3} \eta_{j_l}) \Sigma_{t_2, t_3} (\Sigma_{i_2} \delta_{i_2} r_{t_2 - i_2, t_3 - j_3}) \\ &\qquad \times \Sigma_{t_1} (\Sigma_{i_1} \delta_{i_1} r_{t_2 + i_1, t_1 + j_2}) (\Sigma_{i_3} \delta_{i_3} r_{t_3 - i_3, t_1 - j_1})| \\ &\quad \leqslant |\eta|^3 \Sigma_{t_2, t_3} |\Sigma_{i_2} \delta_{i_2} r_{t_2 - i_2, t_3 - j_3}|\ \|\delta\|^2 \|R\|^2 \\ &\quad \leqslant N\|\delta\|^2|\delta|\ |\eta|^3 (\Sigma_{-\infty < i < \infty} |r_i|) \|R\|^2. \end{aligned}$$

For the other terms in $\xi_3(1, 2, 3)$, the same evaluations hold. Therefore (ii) follows from Assumptions (A.1) and (A.2).

For the rest of the terms in (4.4), we also have

(iii) $|\Sigma_{0 \leqslant i_1, \cdots, i_4, j_1, \cdots, j_4 \leqslant K_n} (\Pi_{1 \leqslant l \leqslant 4} \delta_{i_l} \eta_{j_l}) \xi_4| \leqslant N\|\delta\|^2 |\eta|^4 (C_2 N \|\delta\|^2 + C_5 |\delta|^2)$ for some constants $C_2 > 0$ and $C_5 > 0$.

(iv) $|\Sigma_{0 \leqslant i_1, i_2, j_1, j_2 \leqslant K_n} (\Pi_{1 \leqslant l \leqslant 2} \delta_{i_l} \eta_{j_l}) \xi_2(1, 2)| \leqslant N\|\delta\|^2 |\eta|^2 \|R\| (\|R\| + \Sigma_{-\infty < i < \infty} |r_i|)$.

The conclusion then follows from the above evaluations (i) $\sim$ (iv).

LEMMA 4.3. *Assume* (A.1) $\sim$ (A.2). *If* k_n^* *diverges to infinity, then*

$$\text{p-lim}_{n\to\infty}\{\max_{1\leqslant k\leqslant K_n}|(a(k_n^*) - a(k))'(\hat{r}(K_n) - r(K_n))/L_n(k)|\} = 0,$$

where $a(k_n^*) - a(k)$ *is considered as* K_n*-dimensional vector with undefined entries* 0.

PROOF. In Lemma 4.2, taking $\delta_0 = 0$, $\eta_0 = 1$, $\eta_i = 0$ and $\delta_i = a_i(k_n^*) - a_i(k)$ for $i = 1, 2, \cdots, K_n$, and noting

$$|\delta|^2 \leqslant \max(k, k_n^*)\|\delta\|^2,$$

we find that the right-hand side of (4.3) is dominated by some constant times

$$(\max(k, k_n^*))^{1/2}\|\delta\|^4 N^{-2}.$$

As the norm of $R(K_n)^{-1}$ is bounded away from zero, it is sufficient to show that

$$\Sigma_{1\leqslant k\leqslant K_n}(\max(k, k_n^*))^{1/2}\|\delta\|_R^4 N^{-2} L_n(k)^{-4} \tag{4.5}$$

converges to zero as $n \to \infty$. Here

$$\begin{aligned}&\Sigma_{1\leqslant k\leqslant k_n^*}\left(k_n^{*1/2}\|\delta\|_R^4 N^{-2} L_n(k)^{-4}\right)\\ &\qquad\leqslant \Sigma_{1\leqslant k\leqslant k_n^*}\left(k_n^{*1/2} N^{-2} L_n(k)^{-2}\right) \leqslant k_n^{*3/2} N^{-2} L_n(k_n^*)^{-2}\\ &\qquad\leqslant k_n^{*-1/2}\sigma^{-4}\end{aligned}$$

and

$$\begin{aligned}&\Sigma_{k_n^*<k\leqslant K_n}\left(k^{1/2}\|\delta\|_R^4 N^{-2} L_n(k)^{-4}\right)\\ &\qquad\leqslant \Sigma_{k_n^*<k\leqslant K_n}\left(k^{1/2} N^{-2} L_n(k)^{-2}\right) \leqslant \left(\Sigma_{k_n^*<k\leqslant K_n} k^{-3/2}\right)\sigma^{-4}\end{aligned}$$

Then, (4.5) converges to zero as $n \to \infty$, and the proof is complete.

LEMMA 4.4. *Assume* (A.1) $\sim$ (A.2). *If* k_n^* *diverges to infinity, then*

$$\text{p-lim}_{n\to\infty}\max_{1\leqslant k\leqslant K_n}|(a(k_n^*) - a(k))'(\hat{R}(K_n) - R(K_n))(a(k_n^*) + a(k))/L_n(k)| = 0.$$

PROOF. Put $\eta_0 = 0$ and $\eta_i = a_i(k_n^*) + a_i(k)$ for $i = 1, 2, \cdots, K_n$. As in the proof of Lemma 3.2, $|\eta|$ and $\|\eta\|$ are bounded. Applying Lemma 4.2 for this η and δ defined in Lemma 4.3, we have the desired result by the same way as in Lemma 4.3.

PROPOSITION 4.1. *Assume* (A.1) $\sim$ (A.2). *If* k_n^* *diverges to infinity, then*

$$\text{p-lim}_{n\to\infty}\max_{1\leqslant k\leqslant K_n}\left\{\left|\left(s_{k_n^*} - \sigma_{k_n^*}^2\right) - \left(s_k - \sigma_k^2\right)\right|/L_n(k)\right\} = 0.$$

PROOF. Using the identity

$$\begin{aligned}&\left(s_{k_n^*}^2 - \sigma_{k_n^*}^2\right) - \left(s_k - \sigma_k^2\right)\\ &\qquad= 2(a(k_n^*) - a(k))'(\hat{r}(K_n) - r(K_n))\\ &\qquad\quad+ (a(k_n^*) - a(k))'(\hat{R}(K_n) - R(K_n))(a(k_n^*) + a(k))\end{aligned}$$

and applying Lemmas 4.3 and 4.4, we obtain the result.

THEOREM 4.1. *(Asymptotic efficiency of $\hat{k}$). Assume* (A.1) $\sim$ (A.4). *Then*

$$\text{p-lim}_{n\to\infty}\{\|\hat{a}(\hat{k}) - a\|_R^2/L_n(k_n^*)\} = 1.$$

That is, $\hat{k}$ is an asymptotically efficient selection of the order of the model.

PROOF. Lemma 4.1 and Proposition 4.1 yield that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\big(L_n(\hat{k})/L_n(k_n^*) \leqslant 1 + \varepsilon\big) = 1,$$

because $S_n(\hat{k}) \leqslant S_n(k_n^*)$. On the other hand, from the definition of k_n^*,

$$L_n(\hat{k})/L_n(k_n^*) \geqslant 1.$$

Then

$$\text{p-lim}_{n\to\infty}\big(L_n(\hat{k})/L_n(k_n^*)\big) = 1.$$

Applying Proposition 3.2, we complete the proof.

Put, for any $\varepsilon > 0$,

$$\underline{k}_n^*(\varepsilon) = \min\{k;\ L_n(k)/L_n(k_n^*) \leqslant 1 + \varepsilon,\ 1 \leqslant k \leqslant K_n\}$$

and

$$\bar{k}_n^*(\varepsilon) = \max\{k;\ L_n(k)/L_n(k_n^*) \leqslant 1 + \varepsilon,\ 1 \leqslant k \leqslant K_n\}.$$

Clearly

$$\underline{k}_n^*(\varepsilon) \leqslant k_n^* \leqslant \bar{k}_n^*(\varepsilon).$$

In the following corollary, the behaviour of $\hat{k}$ itself is obtained.

COROLLARY 4.1. *For any $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\big(\underline{k}_n^*(\varepsilon) \leqslant \hat{k} \leqslant \bar{k}_n^*(\varepsilon)\big) = 1.$$

EXAMPLE 4.1. As is well known (Box and Jenkins (1970)), if $\{x_t\}$ is a finite order moving average process, its parameters $a_1, a_2, \cdots$, are exponentially decreasing. In such case, $\|a(k) - a\|_R^2$ also decreases exponentially. Therefore $k_n^* \sim \log n$. Applying Corollary 4.1 we have, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\big(\hat{k} \geqslant k_n^* - \varepsilon\big) = 1,$$

and

$$\text{p-lim}_{n\to\infty}\big(\hat{k}/k_n^*\big) = 1.$$

We now try to change $S_n(k)$ into

$$S_n^\circ(k) = (N + \delta_n(k) + 2k)\hat{\sigma}_k^2,$$

where $\delta_n(k)$ is a real-valued random or nonrandom function of k, $1 \leqslant k \leqslant K_n$. Then another selection $\hat{k}^\circ$ is obtained, which minimizes $S_n^\circ(k)$. The following theorem gives a sufficient condition for $\hat{k}^\circ$ to be asymptotically efficient.

THEOREM 4.2. *Assume* (A.1) $\sim$ (A.4). *If*

(4.6) $$\text{p-lim}_{n\to\infty}\max_{1\leqslant k\leqslant K_n}|\delta_n(k)|/N = 0$$

and

(4.7) $$\text{p-lim}_{n\to\infty}\max_{1\leqslant k\leqslant K_n}|(\delta_n(k)-\delta_n(k_n^*))/NL_n(k)|=0,$$

then the selection $\hat{k}^\circ$ *is also asymptotically efficient.*

PROOF. By simple calculation we have

(4.8) $$\begin{aligned} S_n^\circ(k) = {} & S_n(k) + \delta_n(k)\sigma^2(1-2k/N) \\ & + \delta_n(k)\{L_n(k) + (k\sigma^2 - \|\hat{a}(k)-a(k)\|_{\hat{R}}^2)/N\} \\ & + \delta_n(k)(s_k^2 - \sigma_k^2). \end{aligned}$$

The third and fourth terms on the right-hand side of (4.8) are negligible uniformly in k, compared with $NL_n(k)$, from (4.6) and the proof of Lemma 4.1. Then the condition (4.7) assures that

$$\max_{1\leqslant k\leqslant K_n}|(S_n^\circ(k)-S_n^\circ(k_n^*))-(S_n(k)-S_n(k_n^*))|/NL_n(k)$$

converges to zero in probability. Therefore

$$S_n^\circ(\hat{k}^\circ)\leqslant S_n^\circ(k_n^*),$$

implies that for any $\varepsilon>0$,

$$\lim_{n\to\infty}P(L_n(k_n^*)/L_n(\hat{k}^\circ)\geqslant 1-\varepsilon)=1.$$

The desired result follows from the definition of k_n^* and Proposition 3.2.

EXAMPLE 4.2. $S_n^\circ(k)=(n+2k)\hat{\sigma}_k^2$ satisfies (4.5) and (4.6), so that the corresponding $\hat{k}^\circ$ is also asymptotically efficient. Another application of Theorem 4.2 to $S_n^\circ(k)=\{n(n+k)/(n-k)\}\hat{\sigma}_k^2$ or $S_n^\circ(k)=n\exp(2k/n)\hat{\sigma}_k^2$, gives the asymptotic efficiency of the FPE or AIC method.

As was shown by the present author (1976) in connection with the FPE method, applied to an autoregressive process with finite order k_0, the asymptotic distribution of $\hat{k}$ is biased to higher order than k_0. This means that the method is apt to overestimate the order k_0. The defect can be overcome by changing the term $2k$ in $S_n(k)$ to αkN^β for some $\alpha>2$, $\beta>0$, or $k\log N$ (Akaike (1970), Parzen (1974), Schwarz (1978) and Bhansali and Downham (1977)) only at the cost of the properties (4.6) and (4.7). But by such modification the method loses asymptotic efficiency. We will exemplify the point.

Let $\hat{k}^{(\alpha)}$ be an order selection which attains the minimum of

$$S_n^{(\alpha)}(k)=(N+\alpha k)\hat{\sigma}_k^2, \qquad 1\leqslant k\leqslant K_n,$$

for $\alpha>0$. Putting

$$L_n^{(\alpha)}(k)=(\alpha-1)k\sigma^2/N+\|a-a(k)\|_R^2,$$

by the same way as in Theorem 4.2 we can show that

$$\text{p-lim}_{n\to\infty}L_n^{(\alpha)}(k^{*(\alpha)})/L_n^{(\alpha)}(\hat{k}^{(\alpha)})=1,$$

where $k_n^{*(\alpha)}$ is an integer so as to minimize $L_n^{(\alpha)}(k)$. Thus we find

$$\text{(4.9)} \quad \text{p-lim}_{n\to\infty}\|a(\hat{k}^{(\alpha)}) - a\|_R^2/L_n(k_n^*)$$
$$= \text{p-lim}_{n\to\infty}\big(NL_n(k_n^{*(\alpha)}) - (\alpha - 2)\big(\hat{k}^{(\alpha)} - k_n^{*(\alpha)}\big)\sigma^2\big)/NL_n(k_n^*).$$

Here we may assume $\alpha > 1$. Otherwise $k_n^{*(\alpha)} = K_n$ and at least in the following cases $\hat{k}^{(\alpha)}$ is not asymptotically efficient unless

$$\lim_{n\to\infty} L_n(K_n)/L_n(k_n^*) = 1.$$

Note that $\hat{k}^{(1)}$ is asymptotically equivalent to the CAT method proposed by Parzen (1974). This can be shown by the same arguments as in Theorem 4.2.

CASE I. *Parameters are decreasing as k to some power.* Put simply

$$\|a - a(k)\|_R^2 = Ck^{-\beta}$$

for some constants $C, \beta > 0$. Then $k_n^{*(\alpha)} = m_n$ or $m_n + 1$, where

$$m_n = \left[\left(C\beta N/\left((\alpha - 1)\sigma^2\right)\right)^{1/(\beta+1)}\right]$$

and $[x]$ denotes the integral part of x. Thus, for any $\gamma > 0$,

$$\lim_{n\to\infty} L_n^{(\alpha)}\left(\left[k_n^{*(\alpha)}\gamma\right]\right)/L_n^{(\alpha)}\left(k_n^{*(\alpha)}\right)$$
$$= \left(\gamma^{-\beta} + \gamma\beta\right)/(1 + \beta).$$

This implies that

$$\text{p-lim}_{n\to\infty}\left(\hat{k}^{(\alpha)}/k_n^{*(\alpha)}\right) = 1$$

as in Corollary 4.1. Noting $k_n^{*(\alpha)} = O(NL_n(k_n^*))$ and (4.9), we have

$$\text{p-lim}_{n\to\infty}\|\hat{a}(\hat{k}^{(\alpha)}) - a\|_R^2/L_n(k_n^*)$$
$$= \lim_{n\to\infty} L_n\left(k_n^{*(\alpha)}\right)/L_n(k_n^*)$$
$$= (\alpha - 1)^{\beta/(\beta+1)}(1 + \beta/(\alpha - 1))/(1 + \beta).$$

This is equal to 1, that is, $\hat{k}^{(\alpha)}$ attains the lower bound in the limit if and only if $\alpha = 2$.

CASE II. *Parameters are exponentially decreasing (Example* 4.1). If

$$\|a - a(k)\|_R^2 = Ce^{-\beta k}$$

for some constants $C, \beta > 0$, then $k_n^{*(\alpha)} = m_n$ or $m_n + 1$, where

$$m_n = \left[(1/\beta)\log\left(C\beta N/\left((\alpha - 1)\sigma^2\right)\right)\right].$$

Hence $k_n^{*(\alpha)} = O(NL_n(k_n^*))$ and $\text{p-lim}(\hat{k}^{(\alpha)}/k_n^{*(\alpha)}) = 1$. Accordingly we have

$$\text{p-lim}_{n\to\infty}\|\hat{a}(\hat{k}^{(\alpha)}) - a\|_R^2/L_n(k_n^*)$$
$$= \lim_{n\to\infty} L_n\left(k_n^{*(\alpha)}\right)/L_n(k_n^*)$$
$$= 1.$$

That is, $\hat{k}^{(\alpha)}$ always attains the lower bound in the limit as $n \to \infty$.

The above discussion shows that the situation is different whether parameters are decreasing as some power or exponentially. But the order of decreasing is usually unknown a priori, so that the choice $\alpha = 2$ is essential to our purpose. In other words, $S_n(k)$ is an appropriate estimate of $NL_n(k) + N\sigma^2$, which can be rewritten as $N\sigma_k^2 + k\sigma^2$. If σ_k^2 and σ^2 are simply replaced by $\hat{\sigma}_k^2$, the use of $S_n^{(1)}(k)$ might be suggested. Still, taking account of the bias we must add $k\hat{\sigma}_k^2$ to $S_n^{(1)}(k)$ in compensation for it. This compensation has played an important role in our analysis.

5. Remarks and generalizations. The reader might have an objection as to assumption (A.4). We can replace it by the assumption that the order of the process $\{x_t\}$ is finite and bounded away from a constant C_n which goes to infinity as n. Under this assumption k_n^* also diverges to infinity and we can show the asymptotic efficiency of $\hat{k}$.

For h-step ahead prediction, the same results will be obtained if

$$S_{n,h}(k) = (N + 2k)\hat{\sigma}_{h,k}^2$$

is used instead of $S_n(k)$, where

$$\hat{\sigma}_{h,k}^2 = \Sigma_{K_n+h-1 \leqslant t \leqslant n-1} \left(x_{t+1} + \hat{a}_1(h,k)x_t + \cdots + \hat{a}_{h+k-1}(h,k)x_{t-h-k+2}\right)^2 / N$$

and $N = n - K_n - h + 1$.

The assumption of normality of $\{e_t\}$ is posed only for the convenience of evaluation of higher order moments. Thus the same results will hold true if the moments of $\{x_t\}$ are close to that of a Gaussian process up to the sixteenth order.

Acknowledgments. I wish to thank Professor M. Huzii and Dr. R. Simizu for kindness in correcting the manuscript and Professor E. J. Hannan, whose comments led to simplifying assumptions.

REFERENCES

[1] Akaike, H. (1969a). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21** 243–247.

[2] Akaike, H. (1969b). Power spectrum estimation through autoregressive model fitting. *Ann. Inst. Statist. Math.* **21** 407–419.

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

[4] Akaike, H. (1973a). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60** 255–265.

[5] Akaike, H. (1973b). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*. (B. N. Petrov and F. Csáki, eds.), Akadémia Kiado, Budapest, 267–281.

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

[7] Anderson, T. W. (1977). Estimation for autoregressive moving average models in the time and frequency domains. *Ann. Statist.* **6** 842–865.

[8] Berk, K. N. (1974). Consistent autoregressive spectral estimates. *Ann. Statist.* **2** 489–502.

[9] Bhansali, R. J. (1978). Linear prediction by autoregressive model fitting in the time domain. *Ann. Statist.* **6** 224–231.

[10] BHANSALI, R. J. and DOWNHAM, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion. *Biometrika* **64** 547–551.

[11] BOX, G. E. P. and JENKINS, G. M. (1970). *Time Series Analysis Forcasting and Control*. Holden-Day, San Francisco.

[12] GERSCH, W. and SHARPE, R. (1973). Estimation of power spectra with finite-order autoregressive models. *IEEE Trans. Automatic Control* **AC-18** 367–369.

[13] HANNAN, E. J. (1969). The estimation of mixed moving average autoregressive systems. *Biometrika* **56** 579–594.

[14] HUZII, M. (1977). On a spectral estimate obtained by an autoregressive model fitting. *Ann. Inst. Statist. Math*; Part A **29** 415–431.

[15] LEONOV, V. P. and SHIRYAEV, A. N. (1959). On a method of calculation of semi-invariants. *Theor. Probability Appl.* **4** 319–329.

[16] MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **12** 591–612.

[17] PARZEN, E. (1974). Some recent advances in time series modelling. *IEEE Trans. Automatic Control* **AC-19** 723–730.

[18] PARZEN, E. (1975). Multiple time series: determining the order of approximating autoregressive schemes. Technical Report 23, Dept. Computer Science, State Univ. New York.

[19] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

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

[21] SHIBATA, R. (1977). Convergence of least squares estimates of autoregressive parameters. *Austral. J. Statist.* **19** 226–235.

[22] TONG, H. (1975). Autoregressive model fitting with noisy data by Akaike's information criterion. *IEEE Trans. Information Theory* **IT-21** 476–480.

[23] ZYGMUND, A. (1959). *Trigonometric Series (2nd ed.)*. Cambridge Press, Cambridge.

DEPARTMENT OF MATHEMATICS
TOKYO INSTITUTE OF TECHNOLOGY
OH-OKAYAMA, MEGURO
TOKYO, JAPAN