The Annals of Statistics
1989, Vol. 17, No. 2, 824–840

CONSISTENCY OF AKAIKE'S INFORMATION CRITERION FOR INFINITE VARIANCE AUTOREGRESSIVE PROCESSES[1]

BY KEITH KNIGHT

University of British Columbia

Suppose $\{X_n\}$ is a pth order autoregressive process with innovations in the domain of attraction of a stable law and the true order p unknown. The estimate $\hat{p}$ of p is chosen to minimize Akaike's information criterion over the integers $0, 1, \ldots, K$. It is shown that $\hat{p}$ is weakly consistent and the consistency is retained if $K \to \infty$ as $N \to \infty$ at a certain rate depending on the index of the stable law.

0. Introduction. Consider a stationary pth order autoregressive [AR(p)] process $\{X_n\}$,

$$X_n = \beta_1 X_{n-1} + \beta_2 X_{n-2} + \cdots + \beta_p X_{n-p} + \varepsilon_n,$$

where $\{\varepsilon_n\}$ are independent, identically distributed (i.i.d.) random variables. The parameters $\beta_1, \ldots, \beta_p$ satisfy the usual stationarity constraints, namely all zeros of the polynomial

$$z^p - \sum_{j=1}^{p} \beta_j z^{p-j}$$

have modulus less than 1.

Now assume that the true order p is unknown but bounded by some finite constant $K(N)$. Our main purpose here will be to estimate p by $\hat{p}$ where $\hat{p}$ will be obtained by minimizing a particular version of Akaike's information criterion (AIC) [Akaike (1973)] over the integers $\{0, 1, \ldots, K(N)\}$. Because we should be willing to examine a greater range of possible orders for our estimate as the number of observations increases, it makes sense to allow $K(N)$ to increase with N. In the finite variance case with $K(N) \equiv K$, AIC does not give a consistent estimate of p. In fact, there exists a nondegenerate limit distribution of $\hat{p}$ concentrated on the integers $p, p+1, \ldots, K$ [see Shibata (1976)].

The term AIC is used somewhat incorrectly in this context. Strictly speaking, use of AIC presumes that the distribution of the data is known up to a finite number of parameters. For a given statistical model Ω_b with k-dimensional parameter vector $\mathbf{b}$, AIC is defined as

$$\phi(\Omega_b) = -2\Lambda(\hat{\mathbf{b}}) + 2k,$$

where $\Lambda(\hat{\mathbf{b}})$ is the maximized log-likelihood for the model Ω_b. However, in the

Received April 1987; revised June 1988.

[1]This research was supported by ONR Contract N00014-84-C-0169 and by NSF Grant MCS-83-01807 and was part of the author's Ph.D. dissertation completed at the University of Washington.

AMS 1980 *subject classifications*. Primary 62M10; secondary 62F12, 60G10.

Key words and phrases. Akaike's information criterion, autoregressive processes, infinite variance, stable law.

time series literature, AIC is usually defined in terms of a Gaussian likelihood irrespective of the "true" distribution of the data, so for a kth order autoregressive model, we will define AIC as

$$\phi(k) = N \ln \hat{\sigma}^2(k) + 2k,$$

where $\hat{\sigma}^2(k)$ is the usual estimate of the innovations variance obtained from the Yule–Walker estimates defined in the next section. We will choose as our estimate of p the order which minimizes $\phi(k)$ for k between 0 and K, that is,

$$\hat{p} = \arg \min_{0 \le k \le K} \phi(k).$$

In the case where two or more orders achieve the minimum, we will take the smallest as our estimate.

1. Infinite variance autoregressions. We will be interested in the case where the innovations $\{\varepsilon_n\}$ are in the domain of attraction of a stable law with index $\alpha \in (0, 2)$. This means that both

$$P(|\varepsilon_n| > x) = x^{-\alpha} L(x),$$

$$\lim_{x \to \infty} \frac{P(\varepsilon_n > x)}{P(|\varepsilon_n| > x)} = \lambda \in [0, 1],$$

where $L(\cdot)$ is a slowly varying function [see Feller (1971) or Davis and Resnick (1985, 1986) for more background]. If $E(|\varepsilon_n|) < \infty$, then we will assume that $E(\varepsilon_n) = 0$.

Given observations $X_1, \ldots, X_N$ and order l (which may or may not equal the true order p), we will consider two estimates of the AR parameters, the least squares (LS) $[\tilde{\beta}(l)]$ and Yule–Walker (YW) $[\hat{\beta}(l)]$ estimates which satisfy the matrix equations

$$\tilde{C}_l \tilde{\beta}(l) = \tilde{r}_l \quad \text{and} \quad \hat{C}_l \hat{\beta}(l) = \hat{r}_l,$$

where

$$\tilde{C}_l(i, j) = \sum_{n=l+1}^{N} X_{n-i} X_{n-j},$$

$$\hat{C}_l(i, j) = \sum_{n=|i-j|+1}^{N} X_n X_{n-|i-j|},$$

$$\tilde{r}_l(i) = \sum_{n=l+1}^{N} X_n X_{n-i},$$

$$\hat{r}_l(i) = \sum_{n=i+1}^{N} X_n X_{n-i}.$$

For the LS estimates, $\tilde{\beta}_1(l), \ldots, \tilde{\beta}_l(l)$, where $l \ge p$, we have for $\delta > \alpha$,

$$N^{1/\delta}\left(\tilde{\beta}_k(l) - \beta_k\right) \xrightarrow{\text{a.s.}} 0,$$

where $\beta_k = 0$ for $k > p$ [Hannan and Kanter (1977)]. For YW estimates, $\hat{\beta}_1(l), \ldots, \hat{\beta}_l(l)$, a slightly weaker result holds: Convergence to 0 is in probability rather than almost sure. Davis and Resnick (1985, 1986) give results concerning the asymptotic distributions of these estimates.

We may also wish to consider AR models of the form

$$X_n - \mu = \beta_1(X_{n-1} - \mu) + \cdots + \beta_p(X_{n-p} - \mu) + \varepsilon_n,$$

where μ is unknown and we retain the same assumptions on the β_k's and $\{\varepsilon_n\}$. It can be shown [Knight (1987)] that if we center the observed series by subtracting the sample mean $\overline{X}$ (i.e., $X_n' = X_n - \overline{X}$) and estimate $\beta_1, \ldots, \beta_p$ using X_n', $n = 1, \ldots, N$, we will still have $N^{1/\delta}(\hat{\beta}_k - \beta_k) \to_p 0$ for $\delta > \max(1, \alpha)$ for YW estimates and the convergence is almost sure for LS estimates. More generally, we can center the observed series by subtracting any reasonable (say $\sqrt{N}$-consistent) location estimate $\hat{\mu}$ and estimate the β's using the centered series. Depending on the precise convergence properties of $\hat{\mu}$ we may be able to obtain the full rate of convergence for the estimates of the AR parameters [Knight (1987)].

We will consider a triangular array of random variables $\{X_n^{(N)}\}_{n \le N}$,

$$\begin{aligned} &X_1^{(1)} \\ &X_1^{(2)}, X_2^{(2)} \\ &\vdots \\ &X_1^{(N)}, X_2^{(N)}, \ldots, X_N^{(N)}, \end{aligned}$$

where each row is a finite realization of an AR(p) process

$$X_n^{(N)} = \sum_{j=1}^{p} \beta_j^{(N)} X_{n-j}^{(N)} + \varepsilon_n^{(N)}.$$

The corresponding triangular array of innovations $\{\varepsilon_n^{(N)}\}_{n \le N}$ consists of row-rise i.i.d. random variables which are in the domain of attraction of a stable law. Given a single i.i.d. sequence $\{\varepsilon_n\}$, we could construct each element of the triangular array as

$$X_n^{(N)} = \sum_{j=0}^{\infty} c_j(\boldsymbol{\beta}^{(N)})\varepsilon_{n-j},$$

where the $c_j(\cdot)$'s are the coefficients in the linear process representation of $\{X_n^{(N)}\}$. We will require that $\boldsymbol{\beta}^{(N)} = (\beta_1^{(N)}, \ldots, \beta_p^{(N)})$ are contained in a closed (and hence compact) subset of the parameter space for all N. We can now shrink $\beta_p^{(N)}$ to zero as N goes to infinity and try to consistently estimate p at the same time. Intuitively, it would seem that the smaller $|\beta_p^{(N)}|$ is, the more difficult it should be to distinguish between a pth order and a lower order AR model. From simulations, this does seem to be the case. This is the real motivation for allowing the parameters to vary with N. Consider the following example. Suppose we observe a pth order AR process which has β_p very close to zero (say $\beta_p = 0.1$). To estimate the order of the process, we use a procedure which we

know to be consistent. So for N large enough, we will select the true order with arbitrarily high probability. However, for moderate sized N, the probability of underestimating p may be very high. Conversely, if $|\beta_p|$ is close to 1, then even for small N there will be a high probability of selecting the true order. So by allowing β_p to shrink to zero with N, we may get some idea of the relative sample sizes needed to get the same probability of correct order selection for two different sets of AR parameters. This approach is similar in spirit to considering a sequence of contiguous alternative hypotheses to a null hypothesis as is done in calculating the Pitman efficiency of hypothesis tests.

If we include unknown location μ in the model, we will assume that it does not vary with N. To have μ vary with N may make sense in some situations but we will not consider it here since μ is essentially a nuisance parameter in this situation.

We will provide an answer to the following question: Under what conditions (if any) on $K(N)$ and $(\beta_1^{(N)}, \ldots, \beta_p^{(N)})$ will AIC provide a consistent estimate $\hat{p}$ of p? If $K(N)$ is allowed to grow too fast then we may wind up severely overfitting much of the time; for example, $\hat{p}$ could equal $K(N)$ with high probability. The heuristic argument and simulations given in Bhansali (1984) indicate that AIC will consistently estimate the order if $K(N)$ varies slowly with N. Bhansali (1988) shows that the FPE_α criterion of Bhansali and Downham (1977) consistently estimates the true order of an AR process for fixed K; FPE_2 is asymptotically equivalent to AIC. In the next section, we give some consideration to the rate at which $K(N)$ may go to infinity and still preserve consistency of AIC.

2. Theoretical results. The main result of this paper is contained in Theorem 7; the first six results provide the necessary machinery for Theorem 7. We begin by stating two results dealing with rth moments of martingales and submartingales.

THEOREM 1 [Esseen and von Bahr (1965)]. *Let* $S_n = \sum_{k=1}^n X_k$. *If* $E(X_n|S_{n-1}) = 0$ *for* $2 \le n \le N$ *and* $E(|X_n|^r) < \infty$ *for* $1 \le r \le 2$, *then*

$$E(|S_n|^r) \le 2 \sum_{n=1}^{N} E(|X_n|^r).$$

(Note that $\{S_n;\ n \ge 1\}$ is a martingale.)

THEOREM 2 [Cf. Chung (1974), page 346]. *If* $\{X_n;\ n \ge 1\}$ *is an* L^r-*submartingale for some* $r > 1$, *then*

$$E\left[\max_{1 \le n \le N} |X_n|^r\right] \le \left(\frac{r}{r-1}\right)^r E(|X_N|^r).$$

The following lemma will allow us to ignore the dependence on N of the moments of $\{X_n^{(N)}\}$ by virtue of being able to bound the moments over any sequence of admissible parameters within a compact set.

LEMMA 3. *Let* $\{X_n(\beta)\}$ *be a stationary* AR(p) *process with parameter* β *and innovations* $\{\varepsilon_n\}$ *in the domain of attraction of a stable law with index* α. *Let* C *be a compact set of the parameter space. Then for all* $0 < \delta < \alpha$,

$$\sup_{\beta \in C} E\left[|X_n(\beta)|^{\delta}\right] < \infty.$$

PROOF. $X_n(\beta) = \sum_{j=0}^{\infty} c_j(\beta)\varepsilon_{n-j}$ where $c_j(\beta)$ is a continuous function of β for all j. Now

$$\begin{aligned} |X_n(\beta)| &\leq \sum_{j=0}^{\infty} |c_j(\beta)|\,|\varepsilon_{n-j}| \\ &\leq \sum_{j=0}^{\infty} a_j|\varepsilon_{n-j}|, \end{aligned}$$

where $a_j = \sup_{\beta \in C}|c_j(\beta)|$. However, it can be shown that $|a_j| \leq \text{const.}\ j^p|x|^j$ where $|x| < 1$ and so $\sum_{j=0}^{\infty}|a_j|^{\gamma} < \infty$ for all $\gamma > 0$. Under this summability condition, it follows from Cline (1983) that the random variable

$$X = \sum_{j=0}^{\infty} a_j|\varepsilon_j|$$

is finite almost surely with

$$\lim_{x \to \infty} \frac{P[X > x]}{P[|\varepsilon_1| > x]} = \sum_{j=0}^{\infty} a_j^{\alpha} < \infty.$$

This implies that $E(X^{\delta})$ is finite for all $0 < \delta < \alpha$ and the result follows. □

The following lemma will allow us to treat moments of ΣX_n the same as the moments of $\Sigma\varepsilon_n$ when $\alpha > 1$.

LEMMA 4. *Let* $\{X_n\}$ *be a zero mean stationary* AR(p) *process with innovations* $\{\varepsilon_n\}$ *in the domain of attraction of a stable law with index* $\alpha > 1$. *Then for any* $1 < r < \alpha$,

(a) $$E\left[\left|\sum_{n=1}^{N} X_n\right|^r\right] = O(N),$$

(b) $$E\left[\max_{1 \leq m \leq N}\left|\sum_{n=1}^{m} X_n\right|^r\right] = O(N).$$

PROOF.

$$\begin{aligned} \sum_{n=1}^{N} \varepsilon_n &= \sum_{n=1}^{N}\left(X_n - \sum_{k=1}^{p} \beta_k X_{n-k}\right) \\ &= \left(1 - \sum_{k=1}^{p} \beta_k\right)\sum_{n=1}^{N} X_n + R_N, \end{aligned}$$

where $|R_N| \le (\max_{1 \le k \le p}|\beta_k|)p(p+1)(\max_{1-p \le k \le N}|X_k|)$. Thus

$$\sum_{n=1}^{N} X_n = C\left(\sum_{n=1}^{N} \varepsilon_n - R_N \right),$$

where $C = (1 - \Sigma_{k=1}^{p}\beta_k)^{-1}$. Thus by Minkowski's inequality,

$$E\left[\left| \sum_{n=1}^{N} X_n \right|^r\right]^{1/r} \le CE\left[\left| \sum_{n=1}^{N} \varepsilon_n \right|^r\right]^{1/r} + CE\left[|R_N|^r\right]^{1/r}.$$

Now note that

$$\begin{aligned} \frac{1}{N}E\left[\max_{1-p \le k \le N}|X_k|^r\right] &\le \frac{t}{N} + \frac{1}{N}\int_t^\infty P\left[\max_{1-p \le k \le N}|X_k|^r > x\right] dx \\ &\le \frac{t}{N} + \frac{N-p}{N}\int_t^\infty P\left[|X_1|^r > x\right] dx \\ &\to 0 \quad \text{as } N \to \infty \text{ and } t \to \infty \end{aligned}$$

and so $E[|R_N|^r] = o(N)$ and part (a) follows by applying Theorem 1.

Part (b) follows similarly from Theorem 2 by noting that

$$\max_{1 \le m \le N}\left| \sum_{n=1}^{m} X_n \right| \le C \max_{1 \le m \le N}\left| \sum_{n=1}^{m} \varepsilon_n \right| + CR_N$$

and using Minkowski's inequality. □

The following theorem deals with uniform convergence of both LS and YW autoregressive parameter estimates in the case where location is known.

THEOREM 5. *Assume known location* μ. *Let* $K(N) = O(N^\delta)$ *for* $\delta < 1 - \alpha/2$ *and let* $\|\mathbf{v}\|$ *denote the Euclidean norm of the vector* $\mathbf{v}$. *Then*

(a) $$\sqrt{N} \max_{p \le l \le K(N)} \|\tilde{\boldsymbol{\beta}}(l) - \boldsymbol{\beta}^{(N)}\| \to_p 0,$$

(b) $$\sqrt{N} \max_{1 \le l \le K(N)} \|\hat{\boldsymbol{\beta}}(l) - \tilde{\boldsymbol{\beta}}(l)\| \to_p 0.$$

Note that the vectors are not fixed length but may vary with N.

PROOF. (a) The style of proof will mimic Hannan and Kanter (1977). For convenience we suppress the notation indicating the dependence of $\{X_n\}$, $\{\varepsilon_n\}$ and $\boldsymbol{\beta}$ on N. For $l \ge p$ the LS estimating equations can be expressed as

$$\tilde{C}_l(\tilde{\boldsymbol{\beta}}(l) - \boldsymbol{\beta}) = \mathbf{r}_l^*,$$

where

$$\mathbf{r}_l^*(i) = \sum_{n=l+1}^{N} \varepsilon_n X_{n-i}.$$

Fix $\delta < 1 - \alpha/2$ and set $K = K(N) = O(N^\delta)$. For each l, $\tilde{C}_l$ is nonnegative

definite and so it suffices to show that for some $\kappa < 2/\alpha$,

(i) $$\max_{p \le l \le K} N^{1/2-\kappa} \|\mathbf{r}_l^*\| \to_p 0,$$

(ii) $$\min_{p \le l \le K(N)} \min_{\|\mathbf{v}\|=1} N^{-\kappa} \mathbf{v}' \tilde{C}_l \mathbf{v} \to_p \infty.$$

If (i) and (ii) hold, then clearly

$$\sqrt{N} \max_{p \le l \le K} \|\tilde{\boldsymbol{\beta}}(l) - \boldsymbol{\beta}\| \to_p 0.$$

To prove (i), it suffices to show that

$$E_N = E\left[\max_{p \le l \le K} \left(N^{1-2\kappa} \sum_{j=1}^{l} \left| \sum_{n=l+1}^{N} \varepsilon_n X_{n-j} \right|^2 \right)^{\gamma}\right] \to 0$$

for some $\gamma < \alpha/2$.

Now

$$\begin{aligned}
E_N &\le N^{(1-2\kappa)\gamma} \sum_{j=1}^{K} E\left[\max_{1 \le l \le K} \left| \sum_{n=l+1}^{N} \varepsilon_n X_{n-j} \right|^{2\gamma}\right] \\
&\le N^{(1-2\kappa)\gamma} \sum_{j=1}^{K} E\left[\left\{ \max_{1 \le l \le K} \left| \sum_{n=1}^{l} \varepsilon_n X_{n-j} \right|^{\gamma} + \left| \sum_{n=1}^{N} \varepsilon_n X_{n-j} \right|^{\gamma} \right\}^2\right] \\
&\le N^{(1-2\kappa)\gamma} \sum_{j=1}^{K} \left\{ E\left[\max_{1 \le l \le K} \left| \sum_{n=1}^{l} \varepsilon_n X_{n-j} \right|^{2\gamma}\right] + E\left[\left| \sum_{n=1}^{N} \varepsilon_n X_{n-j} \right|^{2\gamma}\right] \right\} \\
&\quad + 2N^{(1-2\kappa)\gamma} \sum_{j=1}^{K} E\left[\max_{1 \le l \le K} \left| \sum_{n=1}^{l} \varepsilon_n X_{n-j} \right|^{2\gamma}\right]^{1/2} E\left[\left| \sum_{n=1}^{N} \varepsilon_n X_{n-j} \right|^{2\gamma}\right]^{1/2} \\
&= \sum_{j=1}^{K} N^{(1-2\kappa)\gamma} \left(V_{N_j} W_{N_j} + 2V_{N_j}^{1/2} W_{N_j}^{1/2} \right).
\end{aligned}$$

If $2\gamma < 1$, then by the so-called c_r-inequality

$$V_{N_j} \le E\left[\sum_{n=1}^{K} |\varepsilon_n X_{n-j}|^{2\gamma}\right] = O(K(N))$$

uniformly over j between 1 and $K(N)$.

If $2\gamma \ge 1$, then $\alpha > 1$ and so $S_{k,j} = \sum_{n=1}^{k} \varepsilon_n X_{n-j}$ is a martingale for each j. Hence $|S_{k,j}|$ is an $L^{2\gamma}$-submartingale and so by Theorems 1 and 2,

$$V_{N_j} \le CE\left[|S_{K,j}|^{2\gamma}\right] \le 2C \sum_{n=1}^{K} E\left[|\varepsilon_n X_{n-j}|^{2\gamma}\right] = O(K(N))$$

uniformly over j.

Similarly it can be shown that for all permissible values of γ, $W_{N_j} = O(N)$ uniformly over j between 1 and $K(N)$. Thus for a given sequence $K(N) = O(N^\delta)$ by taking κ sufficiently close to $2/\alpha$ and γ sufficiently close to $\alpha/2$, we will have

$$E\left[\max_{p\le l\le K}\left(N^{1-2\kappa}\sum_{j=1}^{l}\left|\sum_{n=l+1}^{N}\varepsilon_n X_{n-j}\right|^2\right)^\gamma\right] = o(1)$$

as desired.

To prove (ii), we define $X_{n,v}$, $\varepsilon_{n,v}$ as

$$X_{n,v} = \sum_{k=1}^{K} v_k X_{n-k},$$

$$\varepsilon_{n,v} = \sum_{k=1}^{K} v_k \varepsilon_{n-k},$$

with $\|\mathbf{v}\|^2 = \sum_{k=1}^{K} v_k^2 = 1$. It suffices to show

$$\min_{\|\mathbf{v}\|=1}\left\{N^{-\kappa}\sum_{n=K+1}^{N} X_{n,v}^2\right\} \to_p \infty.$$

Now note that $X_{n,v} = \sum_{j=1}^{p}\beta_j X_{n-j,v} + \varepsilon_{n,v}$. By the triangle inequality,

$$\left\{N^{-\kappa}\sum_{n=K+1}^{N} X_{n,v}^2\right\}^{1/2} \ge \left|\left\{N^{-\kappa}\sum_{n=K+1}^{N}\left(\sum_{k=1}^{p}\beta_k X_{n-k,v}\right)^2\right\}^{1/2} - \left\{N^{-\kappa}\sum_{n=K+1}^{N}\varepsilon_{n-k,v}^2\right\}^{1/2}\right|.$$

Now

$$N^{-\kappa}\sum_{n=K+1}^{N}\left(\sum_{k=1}^{p}\beta_k X_{n-k,v}\right)^2 \le \sum_{j=1}^{p}\beta_j^2\sum_{k=1}^{p} N^{-\kappa}\sum_{n=K+1}^{N} X_{n-k,v}^2$$

$$= \sum_{j=1}^{p}\beta_j^2\sum_{k=1}^{p} N^{-\kappa}\sum_{n=K+1}^{N} X_{n,v}^2 + o_p(1).$$

It remains only to show that $N^{-\kappa}\Sigma\varepsilon_{n,v}^2 \to_p \infty$ uniformly. If this is true, then $N^{-\kappa}\Sigma X_{n,v}^2 \to_p \infty$ uniformly since the probability that this quantity stays bounded clearly must tend to zero:

$$N^{-\kappa}\sum_{n=K+1}^{N}\varepsilon_{n,v}^2 = N^{-\kappa}\sum_{n=K+1}^{N}\left\{\sum_{k=1}^{K} v_k^2\varepsilon_{n-k}^2 + 2\sum_{k=2}^{K}\sum_{j=1}^{k-1} v_j v_k \varepsilon_{n-j}\varepsilon_{n-k}\right\}.$$

Now

$$\sum_{n=K+1}^{N}\sum_{k=1}^{K} v_k^2\varepsilon_{n-k}^2 \ge \sum_{n=K+1}^{N-K}\varepsilon_n^2.$$

Thus

$$N^{-\kappa} \sum_{n=K+1}^{N} \sum_{k=1}^{K} v_k^2 \varepsilon_{n-k}^2 \to_p \infty$$

since

$$N^{-\kappa} \sum_{n=K+1}^{N-K} \varepsilon_n^2 \to_p \infty.$$

Thus we need only show that

$$N^{-\kappa} \sum_{n=K+1}^{N} \sum_{k=2}^{K} \sum_{j=1}^{k-1} v_j v_k \varepsilon_{n-j} \varepsilon_{n-k} \to_p 0.$$

Now

$$\left| \sum_{k=2}^{K} v_k \sum_{j=1}^{k-1} v_j \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right| \le \sum_{k=2}^{K} |v_k| \sum_{j=1}^{k-1} |v_j| \left| \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right|.$$

Now take $\gamma < \alpha$ and note that $j \neq k$. If $\gamma < 1$, then

$$E\left[\left| \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right|^{\gamma} \right] \le E\left[\sum_{n=K+1}^{N} |\varepsilon_{n-j} \varepsilon_{n-k}|^{\gamma} \right] = O(N).$$

If $\gamma \ge 1$, then necessarily $\alpha > 1$. Thus $S_l = \sum_{n=K+1}^{l} \varepsilon_{n-j} \varepsilon_{n-k}$ is an L^{γ}-martingale and hence

$$E\left[\left| \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right|^{\gamma} \right] = O(N)$$

uniformly over $j \neq k$ by Theorem 1.

Now

$$E\left[\left(N^{-\kappa} \sum_{k=2}^{N} |v_k| \sum_{j=1}^{k-1} |v_j| \left| \sum_{n=K+1}^{N} \varepsilon_{n-j} \varepsilon_{n-k} \right| \right)^{\gamma} \right] = O\left(N^{1-\kappa\gamma} K (N)^{2\gamma} \right) = o(1)$$

since $|v_k| \le 1$ for all k.

(b) From the definitions of $\hat{C}_l$, $\tilde{C}_l$, $\hat{\mathbf{r}}_l$ and $\tilde{\mathbf{r}}_l$, it is easy to see that

$$(1) \qquad T_N = \max_{1 \le l \le K} \max_{1 \le i, j \le l} \left| \hat{C}_l(i, j) - \tilde{C}_l(i, j) \right| \le \sum_{n=1}^{K} X_n^2 + \sum_{n=N-K+1}^{N} X_n^2$$

and

$$(2) \qquad S_N = \max_{1 \le l \le K} \max_{1 \le i \le l} \left| \hat{\mathbf{r}}_l(i) - \tilde{\mathbf{r}}_l(i) \right| \le \sum_{n=1}^{K} X_n^2.$$

Thus using (1) and (2) and noting that

$$\sum_{n=1}^{K} X_n^2 =_d \sum_{n=N-K+1}^{N} X_n^2,$$

we have

$$N^{-\kappa}T_N = o_p(1/N) \tag{3}$$

and

$$N^{-\kappa}S_N = o_p(1/N) \tag{4}$$

for $\kappa < 2/\alpha$. Now using some elementary facts about vector and matrix norms and (3) and (4), we get

$$\max_{1 \le l \le K} N^{-\kappa}\|\hat{C}_l - \tilde{C}_l\| = o_p(K(N)/N) = o_p(1) \tag{5}$$

and

$$\max_{1 \le l \le K} N^{-\kappa}\|\hat{\mathbf{r}}_l - \tilde{\mathbf{r}}_l\| = o_p\left(\sqrt{K(N)}/N\right) = o_p(1/\sqrt{N}), \tag{6}$$

where the matrix norm is that which corresponds to the Euclidean vector norm.

Now from the definitions of $\hat{\beta}(l)$ and $\tilde{\beta}(l)$, we get

$$N^{-\kappa}\hat{C}_l(\tilde{\beta}(l) - \hat{\beta}(l)) = o_p(1/\sqrt{N})$$

uniformly in l by (6). Finally we must show that the minimum eigenvalue of $N^{-\kappa}\hat{C}_l$ tends in probability to infinity uniformly in l since $\|(N^{-\kappa}\hat{C}_l)^{-1}\|$ is (in the case of symmetric positive definite matrices) merely the reciprocal of this minimum eigenvalue. Note that for unit vectors $\mathbf{v}$

$$\begin{aligned} N^{-\kappa}\mathbf{v}'\hat{C}_l\mathbf{v} &= N^{-\kappa}\mathbf{v}'\tilde{C}_l\mathbf{v} + N^{-\kappa}\mathbf{v}'(\hat{C}_l - \tilde{C}_l)\mathbf{v} \\ &\ge N^{-\kappa}\mathbf{v}'\tilde{C}_l\mathbf{v} - N^{-\kappa}\|\hat{C}_l - \tilde{C}_l\| \to_p \infty \end{aligned}$$

uniformly over l and unit vectors $\mathbf{v}$ by condition (ii) of the proof of part (a) of this theorem and (5) above. Therefore

$$\left\|\left(N^{-\kappa}\hat{C}_l\right)^{-1}\right\| \to_p 0$$

as required. □

In the case where we have an unknown location parameter and we estimate it with some location estimate $\hat{\mu}$, we can obtain the following corollary.

COROLLARY 6. (a) *If* $(\hat{\mu} - \mu)^2 = O_p(N^\gamma)$ *for* $\gamma \le \min[(2/\alpha - 5/2 + \alpha/2), 0]$ *uniformly over all compact subsets of the parameter space, then Theorem* 5 *still holds. For* $\alpha > 1$, *the sample mean,* $\bar{X}$, *satisfies this condition.*

(b) *If* $\alpha \le 1$ *and* $\hat{\mu} \equiv \bar{X}$ *and* $K(N) = O(N^\delta)$ *for* $\delta < \frac{1}{2}$, *then conclusions* (a) *and* (b) *of Theorem* 5 *hold.*

PROOF. (a)(i) Assume without loss of generality that $\mu = 0$. We can again reexpress the LS estimating equations as

$$\tilde{C}_l(\tilde{\beta} - \beta) = \mathbf{r}_l^*,$$

where now

$$\tilde{C}_l(i, j) = \sum_{n=l+1}^{N} (X_{n-i} - \hat{\mu})(X_{n-j} - \hat{\mu})$$

and

$$\mathbf{r}_l^*(j) = \sum_{n=l+1}^{N} \left(\varepsilon_n + \hat{\mu}\left(1 - \sum_{k=1}^{p} \beta_k\right)\right)(X_{n-j} - \hat{\mu})$$

$$= \mathbf{r}_l^{**}(j) + (N - l)\left(1 - \sum_{k=1}^{p} \beta_k\right)\hat{\mu}^2.$$

By similar methods to those used in the proof of Theorem 5, it is easy to show that for some $\kappa < 2/\alpha$,

$$\max_{p \le l \le K} N^{1/2-\kappa} \|\mathbf{r}_l^{**}\| \to_p 0.$$

(The term involving ΣX_{n-j} is killed using Lemma 4.)

In addition, using the conditions on $\hat{\mu}$,

$$N^{1/2-\kappa} K(N) N\hat{\mu}^2 \to_p 0.$$

Finally, it follows easily that

$$\min_{p \le l \le K} \min_{\|\mathbf{v}\|=1} N^{-\kappa} \mathbf{v}' \tilde{C}_l \mathbf{v} \to_p \infty.$$

(ii) Defining T_N and S_N analogously to the proof of Theorem 5, we again get that for some $\kappa < 2/\alpha$,

$$N^{-\kappa} T_N = o_p(1/N)$$

and

$$N^{-\kappa} S_N = o_p(1/N)$$

and the rest of the proof follows as in the proof of Theorem 5.

(b) Everything follows from the fact that for any $0 < \gamma < \alpha$,

$$E\left[\max_{1 \le l \le K} \left| \sum_{n=l+1}^{N} X_n \right|^{\gamma}\right] = O(N),$$

which implies that

$$\max_{1 \le l \le K} \left| \sum_{n=l+1}^{N} X_n \right| = O_p(N^{1/\gamma}).$$

So by taking γ close to α and κ close to $2/\alpha$, we get

$$N^{1/2-\kappa} K(N) \frac{1}{N} \max_{1 \le l \le K} \left| \sum_{n=l+1}^{N} X_n \right|^2 \to_p 0$$

and conclusions (a) and (b) of Theorem 5 follow directly from this. □

THEOREM 7. *If* $\liminf N|\beta_p^{(N)}|^2 > 2p$ *and conclusions* (a) *and* (b) *of Theorem* 5 *hold for some* $K(N)$ *and if* $\hat{p}$ *minimizes* AIC, *then*

$$\hat{p} \to_p p.$$

PROOF. First we note that since $\hat{p}$ is integer-valued, $\hat{p} \to_p p$ is equivalent to $P[\hat{p} = p] \to 1$ (as $N \to \infty$). From here on, we will refer to $K(N)$ as K and to $\beta_k^{(N)}$ as β_k, thus suppressing the dependence on N.

Moreover we will assume that the observations X_n are already centered, that is, we have subtracted out the location estimate $\hat{\mu}$ (if we are assuming unknown location).

We now use the fact that

$$\hat{\sigma}^2(k) = \hat{\sigma}^2(0) \prod_{l=1}^{k} \left(1 - \hat{\beta}_l^2(l)\right) \quad \text{for } k \geq 1,$$

where

$$\hat{\sigma}^2(0) = \frac{1}{N} \sum_{n=1}^{N} X_n^2.$$

Now

$$P[\hat{p} < p] \leq P\left[\min_{0 \leq k < p} \phi(k) \leq \phi(p)\right]$$

and since

$$\min_{0 \leq k < p} \phi(k) \geq N \sum_{l=1}^{p-1} \ln\left(1 - \hat{\beta}_l^2(l)\right) + N \ln \hat{\sigma}^2(0),$$

we can write

$$\begin{aligned} P[\hat{p} < p] &\leq P\left[\ln\left(1 - \hat{\beta}_p^2(p)\right) \geq -2p/N\right] \\ &= P\left[\left(1 - \hat{\beta}_p^2(p)\right) \geq \exp(-2p/N)\right] \\ &\leq P\left[N\hat{\beta}_p^2(p) \leq 2p\right]. \end{aligned}$$

However,

$$N\hat{\beta}_p^2(p) = \left(\sqrt{N}|\beta_p| + o_p(1)\right)^2$$

and so

$$\limsup_{N \to \infty} P\left[N\hat{\beta}_p^2(p) \leq 2p\right] = 0$$

since $\liminf \sqrt{N}\,|\beta_p^{(N)}| > \sqrt{2p}$. Thus $P[\hat{p} < p] \to 0$.

We also have that

$$\begin{aligned} P[\hat{p} > p] &\leq P[\phi(k) < \phi(k-1) \text{ for some } p < k \leq K] \\ &\leq P\left[N \min_{p < k \leq K} \ln\left(1 - \hat{\beta}_k^2(k)\right) < -2\right]. \end{aligned}$$

If the conclusions of Theorem 5 hold, it follows that

$$N \max_{p<k\leq K} \hat{\beta}_k^2(k) \to_p 0$$

and hence

$$N \min_{p<k\leq K} \ln\left(1 - \hat{\beta}_k^2(k)\right) \to_p 0.$$

Therefore, $P[\hat{p} > p] \to 0$.

Thus $P[\hat{p} \neq p] \to 0$ and so $P[\hat{p} = p] \to 1$ which implies that $\hat{p} \to_p p$. □

3. Simulation results. The "practical" implication of Theorem 7 is that if N is large, with high probability $\hat{p}$ will equal p provided that $|\beta_p|$ is not too

TABLE 1
Frequency of selected order for AR(1) *process.* $N = 100$, $\alpha = 0.5$.

Estimated order	AR parameter		
	0.1	0.5	0.9
0	89	0	1
1	4	91	87
2	4	3	2
3	0	1	1
4	1	0	0
5	0	3	2
6	1	1	2
7	0	0	3
8	0	0	1
9	1	1	0
10	0	0	1

TABLE 2
Frequency of selected order for AR(1) *process.* $N = 900$, $\alpha = 0.5$.

Estimated order	AR parameter		
	0.1	0.5	0.9
0	0	0	0
1	93	95	91
2	0	0	0
3	0	0	1
4	0	0	0
5	0	0	0
6	0	0	0
7	4	0	0
8	1	5	2
9	0	0	1
10–15	2	0	5

TABLE 3
Frequency of selected order for AR(1) *process.* $N = 100$, $\alpha = 1.2$.

Estimated order	AR parameter		
	0.1	0.5	0.9
0	70	0	0
1	15	86	86
2	7	7	4
3	3	3	3
4	1	1	3
5	0	1	1
6	0	0	0
7	2	0	2
8	2	1	1
9	0	1	0
10	0	0	0

TABLE 4
Frequency of selected order for AR(1) *process.* $N = 900$, $\alpha = 1.2$.

Estimated order	AR parameter		
	0.1	0.5	0.9
0	0	0	0
1	80	87	90
2	6	3	3
3	6	0	1
4	1	4	3
5	1	2	0
6	0	0	0
7	2	2	1
8	1	0	0
9	0	0	0
10–15	3	2	2

TABLE 5
Frequency of selected order for AR(1) *process.* $N = 100$, $\alpha = 1.9$.

Estimated order	AR parameter		
	0.1	0.5	0.9
0	57	0	0
1	25	76	71
2	5	8	10
3	2	5	9
4	3	6	6
5	2	1	2
6	3	1	0
7	1	1	1
8	1	0	0
9	0	0	1
10	1	2	0

TABLE 6
Frequency of selected order for AR(1) *process.* $N = 900$, $\alpha = 1.9$.

Estimated order	AR parameter		
	0.1	0.5	0.9
0	5	0	0
1	78	74	75
2	7	14	9
3	2	2	7
4	2	5	2
5	1	1	2
6	1	1	0
7	3	1	3
8	0	0	0
9	0	0	1
10–15	1	2	1

TABLE 7
Frequency of selected order for AR(1) *process.* $N = 100$, *normal distribution.*

Estimated order	AR parameter		
	0.1	0.5	0.9
0	63	0	0
1	25	75	75
2	4	3	12
3	1	6	2
4	0	7	5
5	2	2	2
6	2	4	3
7	1	0	0
8	1	1	1
9	0	2	0
10	1	0	0

TABLE 8
Frequency of selected order for AR(1) *process.* $N = 900$, *normal distribution.*

Estimated order	AR parameter		
	0.1	0.5	0.9
0	0	0	0
1	83	79	80
2	3	3	11
3	4	4	3
4	4	6	0
5	2	3	3
6	0	0	0
7	0	4	0
8	4	1	1
9	0	0	2
10–15	0	0	0

small with respect to N. In other words, for fixed (but large) N, the probability of selecting the correct order decreases as $|\beta_p|$ decreases.

For illustrative purposes, a small simulation study was carried out using four symmetric stable innovations distributions with $\alpha = 0.5$, 1.2, 1.9 and 2.0 (the latter being the normal distribution). The underlying processes were AR(1) processes with the AR parameter $\beta = 0.1$, 0.5 and 0.9. The stable random variables were generated using the algorithm of Chambers, Mallows and Stuck (1976); normal random variables were generated using an unpublished algorithm of Marsaglia. The sample sizes considered were 100 and 900. For $N = 100$, the maximum order K was taken to be 10 while for $N = 900$, K was taken to be 15. 100 replications were made for each of the 24 possible arrangements of α, β and N. The results of the study are given in Tables 1–8. The results are much as expected. We can see that for $N = 100$ and $\beta_1 = 0.1$, AIC underestimates the true order with high probability. For $N = 900$, the probabilities of selecting the true order increases over those for $N = 100$.

4. Comments. Bhansali and Downham (1977) propose a generalization of AIC which amounts to minimizing $\phi'(k) = N \ln \hat{\sigma}^2(k) + \gamma k$ where $\gamma \in (0, 4)$. It is easy to see from the proof of the above result that their criterion will also lead to consistent estimates of p under similar conditions on $K(N)$ and $\beta_p^{(N)}$. In fact, if $\gamma = \gamma(N) > 0$ satisfies $\gamma(N)/N \to 0$, then the criterion corresponding to $\phi''(k) = N \ln \hat{\sigma}^2(k) + \gamma(N)k$ will consistently estimate p. Specifically, with known location, the estimate will be consistent provided

$$\liminf_{N\to\infty} \frac{N}{\gamma(N)} |\beta_p^{(N)}|^2 > p$$

with $\gamma(N)$ bounded away from zero and with the same conditions on $K(N)$. With an appropriate choice of $\gamma(N)$, this criterion will also be consistent in the finite variance case. However, if $\gamma(N)$ grows too quickly with N then the criterion may seriously underestimate the true order p in small samples in both the finite and infinite variance cases. In an application such as autoregressive spectral density estimation (assuming now finite variance), underestimation is more serious than overestimation since, if the order is underestimated, the resulting spectral density estimate may be lacking important features which may indeed exist.

Throughout this paper, we have assumed that the innovations $\{\varepsilon_n\}$ are in the domain of attraction of a stable law. This assumption is somewhat stronger than necessary; all that is really needed is the condition that for $\kappa < 2/\alpha$,

$$N^{-\kappa} \sum_{n=1}^{N} \varepsilon_n^2 \to_p \infty,$$

where now

$$\alpha = \sup\{\delta: E(|\varepsilon_n|^\delta) < \infty\}. \tag{7}$$

A sufficient condition for this is that $\{\varepsilon_n\}$ are in the domain of attraction of a stable law. The conclusions of Lemmas 3 and 4 still hold if we define α as in (7).

Acknowledgment. The author would like to thank his supervisor R. Douglas Martin for his support and encouragement in preparing this paper.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (B. Petrov and F. Csáki, eds.) 267–281. Akedemiai Kiado, Budapest.

BARNDORFF-NIELSEN, O. and SCHOU, G. (1973). On the parametrization of autoregressive models by partial autocorrelations. *J. Multivariate Anal.* **3** 408–419.

BHANSALI, R. J. (1984). Order determination for processes with infinite variance. *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statist.* **26** 17–25. Springer, New York.

BHANSALI, R. J. (1988). Consistent order determination for processes with infinite variance. *J. Roy. Statist. Soc. Ser. B* **50** 46–60.

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

CHAMBERS, J. M., MALLOWS, C. L. and STUCK, B. W. (1976). A method for simulating stable random variables. *J. Amer. Statist. Assoc.* **71** 340–344.

CHUNG, K. L. (1974). *A Course in Probability Theory*, 2nd ed. Academic, New York.

CLINE, D. (1983). Infinite series of random variables with regularly varying tails. Technical Report No. 83-24, Institute of Applied Mathematics and Statistics, Univ. British Columbia.

DAVIS, R. and RESNICK, S. (1985). More limit theory for the sample correlation function of moving averages. *Stochastic Process. Appl.* **20** 257–279.

DAVIS, R. and RESNICK, S. (1986). Limit theory for the sample covariance and correlation functions of moving averages. *Ann. Statist.* **14** 533–558.

ESSEEN, C. and VON BAHR, B. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.* **36** 299–303.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

HANNAN, E. J. and KANTER, M. (1977). Autoregressive processes with infinite variance. *J. Appl. Probab.* **14** 411–415.

KNIGHT, K. (1987). Rate of convergence of centered estimates of autoregressive parameters for infinite variance autoregressions. *J. Time Ser. Anal.* **8** 51–60.

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

DEPARTMENT OF STATISTICS
UNIVERSITY OF TORONTO
TORONTO, ONTARIO
CANADA M5S 1A1