The Annals of Statistics
1991, Vol. 19, No. 3, 1155–1162

CONVERGENCE OF MOMENTS OF LEAST SQUARES ESTIMATORS FOR THE COEFFICIENTS OF AN AUTOREGRESSIVE PROCESS OF UNKNOWN ORDER

BY R. J. BHANSALI AND F. PAPANGELOU

University of Liverpool and University of Manchester

Given a realization of T consecutive observations of a stationary autoregressive process of unknown, possibly infinite, order m, it is assumed that a process of arbitrary finite order p is fitted by least squares. Under appropriate conditions it is known that the estimators of the autoregressive coefficients are asymptotically normal. The question considered here is whether the moments of the (scaled) estimators converge, as $T \to \infty$, to the moments of their asymptotic distribution. We establish a general result for stationary processes (valid, in particular, in the Gaussian case) which is sufficient to imply this convergence.

1. Introduction. Let $\{X_t\}$, $t = \dots, -1, 0, 1, \dots$, be an autoregressive process of order m,

$$\sum_{j=0}^{m} a_m(j) X_{t-j} = \varepsilon_t, \qquad a_m(0) = 1, \tag{1.1}$$

where $m \geq 0$ is finite or infinite, $\{\varepsilon_t\}$ is a sequence of independent identically distributed random variables, each with mean 0 and variance σ^2, and the $a_m(j)$ are real coefficients such that

$$\sum_{j=0}^{m} |a_m(j)| < \infty, \qquad \sum_{j=0}^{m} a_m(j) z^j \neq 0, \qquad |z| \leq 1.$$

The order m is seldom (if ever) known a priori. Having observed $X_1, \dots, X_T$, suppose that pth-order least squares estimators $\hat{a}(p) = [\hat{a}_p(1), \dots, \hat{a}_p(p)]'$ of the autoregressive coefficients are obtained by solving the following equation:

$$\hat{\mathbf{a}}(p) = -\hat{\mathbf{R}}(p)^{-1} \hat{\mathbf{r}}(p), \tag{1.2}$$

where $p \geq 1$ is arbitrary, that is, it does not necessarily coincide with m, $\hat{\mathbf{R}}(p) = [D^{(T)}(u, v)]$, $u, v = 1, \dots, p$, $\hat{\mathbf{r}}(p) = [D^{(T)}(0, 1), \dots, D^{(T)}(0, p)]'$ and

$$D^{(T)}(u, v) = (T - p)^{-1} \sum_{t=p+1}^{T} X_{t-u} X_{t-v}, \qquad u, v = 0, 1, \dots, p. \tag{1.3}$$

Received January 1990; revised June 1990.

AMS 1980 *subject classifications*. Primary 62M20; secondary 62M10, 62M15, 60G10, 60G15, 60G25.

Key words and phrases. Stationary process, time series, prediction, uniform integrability, convergence of moments.

Also, let

$$(1.4)\qquad \hat{\sigma}^2(p) = \sum_{u=0}^{p} \hat{a}_p(u) D^{(T)}(0,u), \qquad \hat{a}_p(0) = 1,$$

be the corresponding pth-order estimate of the residual error variance.

The theoretical parameters estimated by $\hat{a}_p(u)$ and $\hat{\sigma}^2(p)$ are

$$(1.5)\qquad \mathbf{a}(p) = -\mathbf{R}(p)^{-1}\mathbf{r}(p),$$

$$(1.6)\qquad \sigma^2(p) = \sum_{j=0}^{p} a_p(j) R(j), \qquad a_p(0) = 1,$$

where $\mathbf{R}(p) = [R(u-v)]$, $u, v = 1, \dots, p$, $\mathbf{r}(p) = [R(1), \dots, R(p)]'$, $\mathbf{a}(p) = [a_p(1), \dots, a_p(p)]'$ and $R(u) = E(X_t X_{t+u})$, $t, u = 0, \pm 1, \dots$, denotes the covariance function of $\{X_t\}$.

The asymptotic normality of the $\hat{a}_p(u)$ when m is finite and $p = m$ was established in a classic paper by Mann and Wald (1943). The case when m is unknown and therefore not necessarily equal to the chosen p has been considered by Kromer (1969), Ogata (1980), Bhansali (1981), Kunitomo and Yamamoto (1985) and Lewis and Reinsel (1988). In the present paper we study the question of boundedness of moments of the standardised $\hat{a}_p(u)$, that is, we examine whether as $T \to \infty$ the moments of the scaled estimators converge to the corresponding moments of their asymptotic distribution [Loève (1963), page 182].

As the autoregressive models play an important role in the statistical literature, this question is of interest in its own right. It, however, also arises in other contexts: for example, in studies dealing with the question of bias in estimating the autoregressive parameters [Shaman and Stine (1988)]; for evaluating the increase in the mean squared error of prediction due to employing the estimated autoregressive model for prediction [Kunitomo and Yamamoto (1985)]; and for giving a rigorous derivation of the Akaike information criterion, AIC, for autoregressive order selection [Bhansali (1986) and Findley (1985)].

Previous work on this question is by Fuller and Hasza (1981), Findley and Wei (1988) and Maliukevičius (1988).

Fuller and Hasza use a conditioning argument to demonstrate that if m is finite, $p = m$ and the ε_t are normally distributed then the powers of the (scaled) estimated least squares autoregressive coefficients are uniformly integrable.

Findley and Wei examine the use of AIC for nonnested model comparisons and in this context provide an extension of the result of Fuller and Hasza to the case of vector autoregression. Moreover, they consider the situation when m is finite and p does not necessarily equal m, and some special cases of the situation with m infinite.

Maliukevičius considers the estimation of a vector of parameters, $\Phi = [\varphi_1, \dots, \varphi_s]'$, say, of the spectral density function of a discrete-time stationary Gaussian process, based on a realization of T consecutive observations of this process. Let $\hat{\Phi}$ denote the maximum likelihood estimator of Φ. By appealing to results of Ibragimov and Has'minskii (1981), Maliukevičius establishes that, under a set of technical assumptions, the moments of $\sqrt{T}(\hat{\Phi} - \Phi)$ converge, as $T \to \infty$, to the corresponding moments of the asymptotic Gaussian distribution of this vector. His result applies, in particular, to the maximum likelihood estimators of the parameters of a Gaussian autoregressive moving average process of known order. However, the more plausible case in which the fitted model only provides a rational transfer function approximation [Hannan (1987)] to the spectral density function of the observed process has not been considered.

For the situation considered in the present paper, Bhansali (1981) and Kunitomo and Yamamoto (1985) obtain uniform integrability of the scaled powers of $\hat{a}_p(u)$ by making a "boundedness" assumption concerning $\{\hat{\mathbf{R}}(p) - \mathbf{R}(p)\}$ or $\hat{\mathbf{R}}(p)^{-1}$. This assumption is, however, more restrictive than necessary for establishing the result and, in any case, does not hold when $\{X_t\}$ is Gaussian.

It is nevertheless clear from an expansion for $\hat{\mathbf{R}}(p)^{-1}$ given by Bhansali (1981) that it is enough to assume a condition of the form

$$\limsup_{T\to\infty} E\left\{\left\|\hat{\mathbf{R}}(p)^{-1}\right\|^q\right\} < \infty, \tag{1.7}$$

where $\|\mathbf{C}\|$ denotes the operator norm of the matrix $\mathbf{C}$. Thus, for example, if (1.7) holds with $q = 3$ then it follows from Bhansali (1981) that for each fixed u,

$$E\{\hat{a}_p(u)\} = a_p(u) + O(T^{-1}), \tag{1.8}$$

provided also that the fourth-order moment of ε_t exists.

Other references where a condition analogous to (1.7) is used are Shaman and Stine (1988), Lewis and Reinsel (1988) and Findley and Wei (1988).

The result of the next section implies that if, for instance, $\{X_t\}$ is any stationary nondegenerate Gaussian process, then (1.7) is true for arbitrary p and $q > 0$.

2. Upper bounds for the norms. For our main result we do not assume that

$$\dots X_{-2}, X_{-1}, X_0, X_1, X_2, \dots \tag{2.1}$$

is an autoregressive process. Instead we assume the following:

(I) The process (2.1) is strictly stationary.

(II) The X_i's have finite moments of all orders.

(III) For any finite set $\{j, l_1, l_2, \dots, l_k\}$ of distinct integers the joint distribution of $X_j, X_{l_1}, X_{l_2}, \dots, X_{l_k}$ is absolutely continuous and there exist a constant $K > 0$ and a version $f_{jl_1\dots l_k}(\cdot|x_1, \dots, x_k)$ of the conditional probability density function of X_j given $X_{l_1} = x_1, \dots, X_{l_k} = x_k$ such that

$$f_{jl_1\dots l_k}(x|x_1, \dots, x_k) \le Kt$$

for all $x, x_1, \dots, x_k$.

Note that condition (III) is satisfied by any Gaussian process with nondegenerate finite-dimensional distributions, since for such a process the conditional variance of X_j given $X_{l_1} = x_1, \dots, X_{l_k} = x_k$ is the same for all values $x_1, \dots, x_k$. If a Gaussian process is autoregressive, nondegeneracy is guaranteed by the errors $\{\varepsilon_t\}$ in (1.1). For non-Gaussian autoregressive processes it would be of interest to have broad conditions on the distribution of ε_t guaranteeing (III) [or (2.2) below], but we do not discuss this matter in the present article.

Let p be a positive integer and for each $j = 0, 1, 2, \dots$ consider the positive-definite symmetric $p \times p$ matrix:

$$A_j = \begin{bmatrix} X_{p+j}X_{p+j} & \cdots & X_{p+j}X_{1+j} \\ \vdots & & \vdots \\ X_{1+j}X_{p+j} & \cdots & X_{1+j}X_{1+j} \end{bmatrix}.$$

Note that the operator norm of any positive-definite symmetric matrix is equal to its largest eigenvalue. We will prove the following theorem.

THEOREM. *If* (2.1) *is any stochastic process satisfying* (I), (II) *and* (III), *then for each* $q > 0$ *there is a nonnegative random variable* Λ_0 *and a number* $r > 0$ *such that* $E(\Lambda_0^q) < \infty$ *and for all* $N \geq r$,

$$\left\| \left[\frac{1}{N} \sum_{j=0}^{N-1} A_j \right]^{-1} \right\| \leq \Lambda_0 \quad \textit{almost surely}.$$

PROOF. Given $q > 0$, fix a positive integer k greater than $4p + 2q - 2$. With v denoting a vector $\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_p \end{bmatrix}$, consider the quadratic form

$$Q(v) = v^T \left[\sum_{j=0}^{(k-1)p} A_j \right] v = \sum_{t=p+1}^{kp+1} \left(X_{t-1}v_1 + X_{t-2}v_2 + \cdots + X_{t-p}v_p \right)^2.$$

Let v' be a fixed vector such that $\|v'\| = 1$. For $\beta > 0$,

$$\begin{aligned} &P(Q(v') < \beta) \\ &\quad \leq P\Bigg(-\sqrt{\beta} < \sum_{i=1}^{p} X_{p+1-i}v_i' < \sqrt{\beta}, -\sqrt{\beta} < \sum_{i=1}^{p} X_{2p+1-i}v_i' < \sqrt{\beta}, \\ &\qquad\qquad \dots, -\sqrt{\beta} < \sum_{i=1}^{p} X_{kp+1-i}v_i' < \sqrt{\beta} \Bigg). \end{aligned}$$

For at least one i, $|v_i'|$ is greater than or equal to $1/\sqrt{p}$. If for this i we apply condition (III) to the conditional pdf of X_{p+1-i} given the other X_ν's, we see that the last probability is

$$\leq \text{const}\sqrt{\beta}\, P\Bigg(-\sqrt{\beta} < \sum_{i=1}^{p} X_{2p+1-i} v_i' < \sqrt{\beta}, \dots,$$
$$-\sqrt{\beta} < \sum_{i=1}^{p} X_{kp+1-i} v_i' < \sqrt{\beta} \Bigg).$$

Iterating, we obtain

$$P(Q(v') < \beta) \leq \text{const}\ \beta^{k/2}. \tag{2.2}$$

Now let v, v' be on the unit sphere S of $\mathbb{R}^p$ and note that the inequality

$$\left| \sum_{i=1}^{p} X_{t-i} v_i' \right| \leq \left| \sum_{i=1}^{p} X_{t-i} v_i \right| + \left[\sum_{i=1}^{p} X_{t-i}^2 \right]^{1/2} \|v' - v\|$$

implies that if $\|v' - v\| < \varepsilon$ and $\sum_{i=1}^{kp} X_i^2 < 1/\varepsilon$, then

$$Q(v') \leq 2Q(v) + 2(kp - p + 1)\varepsilon. \tag{2.3}$$

It is easy to see that, given $\varepsilon \in (0, 1)$, there is a subset S' of S, with fewer than $2[4p^2/\varepsilon^2]^{p-1}$ elements (a crude upper bound), such that given any $v \in S$ there exists $v' \in S'$ with $\|v' - v\| < \varepsilon$. From this and (2.3) we deduce that

$$\begin{aligned} &P(Q(v) < \varepsilon \text{ for some } v \in S) \\ &\quad \leq P\left(\sum_{i=1}^{kp} X_i^2 \geq \frac{1}{\varepsilon} \right) + P(Q(v') < 2\varepsilon + 2(kp - p + 1)\varepsilon \text{ for some } v' \in S') \\ &\quad \leq P\left(\sum_{i=1}^{kp} X_i^2 \geq \frac{1}{\varepsilon} \right) + 2\frac{(4p^2)^{p-1}}{\varepsilon^{2(p-1)}} \text{const}[2(kp + 1)\varepsilon]^{k/2} \\ &\quad \leq \text{const}\ \varepsilon^{k/2-2p+2}, \end{aligned} \tag{2.4}$$

since $\sum_{i=1}^{kp} X_i^2$ has moments of all orders.

Now consider the smallest eigenvalue $\lambda_0 = \inf_{v \in S} Q(v)$ of $\sum_{j=0}^{(k-1)p} A_j$. By (2.4), $P(\lambda_0 < \varepsilon) \leq \text{const}\ \varepsilon^{k/2-2p+2}$, hence

$$\begin{aligned} E\left(\lambda_0^{-(q+1)}\right) &= \int_0^\infty P\left(\lambda_0^{-(q+1)} > x\right) dx \\ &\leq 1 + \int_1^\infty P\left(\lambda_0 < x^{-1/(q+1)}\right) dx \\ &\leq 1 + \text{const} \int_1^\infty \left(x^{-1/(q+1)}\right)^{k/2-2p+2} dx, \end{aligned}$$

which is finite since k was chosen greater than $4p + 2q - 2$.

Next set $r = (k-1)p + 1$, let λ_ν be the smallest eigenvalue of $\sum_{j=\nu r}^{(\nu+1)r-1} A_j$, $\nu = 0, 1, 2, \dots$, and define

$$\Lambda = \sup_{n \geq 1} n \left[\sum_{\nu=0}^{n-1} \lambda_\nu \right]^{-1}.$$

For $\alpha > 0$, $\{\Lambda > \alpha\} = \{n > \alpha \sum_{\nu=0}^{n-1} \lambda_\nu$ for some $n \geq 1\} = \{\sup_{n \geq 1} \sum_{\nu=0}^{n-1} (1 - \alpha\lambda_\nu) > 0\}$. Since the random variables λ_ν are integrable, the maximal ergodic theorem [Doob (1953), page 466, Lemma 2.2] implies $\int_{\{\Lambda > \alpha\}} (1 - \alpha\lambda_0)\, dP \geq 0$, that is,

$$\int_{\{\Lambda > \alpha\}} \lambda_0 \, dP \leq \alpha^{-1} P(\Lambda > \alpha).$$

By Hölder's inequality

$$\begin{aligned} P(\Lambda > \alpha) &= E\left[1_{\{\Lambda > \alpha\}} \lambda_0^{(q+1)/(q+2)} \lambda_0^{-(q+1)/(q+2)}\right] \\ &\leq E\left[1_{\{\Lambda > \alpha\}} \lambda_0\right]^{(q+1)/(q+2)} E\left[\lambda_0^{-(q+1)}\right]^{1/(q+2)} \\ &\leq \left(\alpha^{-1} P(\Lambda > \alpha)\right)^{(q+1)/(q+2)} E\left[\lambda_0^{-(q+1)}\right]^{1/(q+2)}, \end{aligned}$$

which implies $P(\Lambda > \alpha) \leq \alpha^{-(q+1)} E(\lambda_0^{-(q+1)})$ for all $\alpha > 0$ and hence $E(\Lambda^q) < \infty$ [cf. Neveu (1965), page 212].

Now from the definition of Λ, $n \leq \Lambda \sum_{\nu=0}^{n-1} \lambda_\nu \leq \Lambda v^T [\sum_{j=0}^{nr-1} A_j] v$ for all $v \in S$ and all $n \geq 1$. This implies that the smallest eigenvalue of $\sum_{j=0}^{nr-1} A_j$ is almost surely greater than or equal to $n\Lambda^{-1}$ and hence the largest eigenvalue of $[\sum_{j=0}^{nr-1} A_j]^{-1}$ is less than or equal to $n^{-1}\Lambda$, that is,

$$n \left\| \left[\sum_{j=0}^{nr-1} A_j \right]^{-1} \right\| \leq \Lambda$$

almost surely. It follows that

$$\left\| \left[\frac{1}{N} \sum_{j=0}^{N-1} A_j \right]^{-1} \right\| \leq r\Lambda$$

for all N of the form nr, $n \geq 1$. If $N = nr + s$, $n \geq 1$, $0 < s < r$, an easy extension shows that

$$\left\| \left[\frac{1}{N} \sum_{j=0}^{N-1} A_j \right]^{-1} \right\| \leq 2r\Lambda,$$

which proves the theorem. Note that condition (III) was only used in deriving (2.2). □

To apply this theorem to the estimators of the first section, assume that the process (2.1) is a Gaussian autoregressive process. Then (I), (II) and (III) hold and the theorem shows that

$$\|\hat{\mathbf{R}}(p)^{-1}\| \le \Lambda_0 \tag{2.5}$$

for $T - p \ge r$. From (1.2), (1.5), (2.5) and the results of Bhansali (1981), it then follows that

$$E\{\|\hat{\mathbf{a}}(p) - \mathbf{a}(p)\|^q\} = O(T^{-q/2})$$

as $T \to \infty$, for arbitrary q. Another consequence of (2.4) is that

$$E\left[\{\hat{\sigma}^2(p-1)\}^{-q}\right] \le E\left[\left(\mathrm{tr}\{\hat{\mathbf{R}}(p)^{-1}\}\right)^q\right] \le p^q E(\Lambda_0^q)$$

for sufficiently large T, since $\{\hat{\sigma}^2(p-1)\}^{-1}$ is the first diagonal element of $\hat{\mathbf{R}}(p)^{-1}$ [Bhansali (1990)]. The case $q > 1$ of this result is useful in the derivation of the Akaike information criterion [Findley and Wei (1988)]. The same bound holds for the other diagonal elements of $\hat{\mathbf{R}}(p)^{-1}$, a fact useful in the interpolation problem.

Acknowledgment. The authors wish to record their thanks to a referee for several valuable comments.

REFERENCES

BHANSALI, R. J. (1981). Effects of not knowing the order of an autoregressive process on the mean squared error of prediction. I. *J. Amer. Statist. Assoc.* **76** 588–597.

BHANSALI, R. J. (1986). A derivation of the information criteria for selecting autoregressive models. *Adv. in Appl. Probab.* **18** 360–387.

BHANSALI, R. J. (1990). On a relationship between the inverse of a stationary covariance matrix and the linear interpolator. *J. Appl. Probab.* **27** 156–170.

DOOB, J. L. (1953). *Stochastic Processes*. Wiley, New York.

FINDLEY, D. F. (1985). On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. *J. Time Series Anal.* **6** 229–252.

FINDLEY, D. F. and WEI, C. Z. (1988). An analysis of AIC for linear stochastic regression and control. Unpublished manuscript.

FULLER, W. A. and HASZA D. P. (1981). Properties of predictors for autoregressive time series. *J. Amer. Statist. Assoc.* **76** 155–161.

HANNAN, E. J. (1987). Rational transfer function approximation. *Statist. Sci.* **2** 135–161.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.

KROMER, R. E. (1969). Asymptotic properties of the autoregressive spectral estimator. Technical Report 13, Dept. Statistics, Stanford Univ.

KUNITOMO, N. and YAMAMOTO, T. (1985). Properties of predictors in misspecified autoregressive time series models. *J. Amer. Statist. Assoc.* **80** 941–950.

LEWIS, R. A. and REINSEL, G. C. (1988). Prediction error of multivariate time series with misspecified models. *J. Time Series Anal.* **9** 43–57.

LOÈVE, M. (1963). *Probability Theory*. Van Nostrand, Princeton, N.J.

MALIUKEVIČIUS, R. (1988). Maximum likelihood estimation of the spectral density parameter. *Lithuanian Math. J.* **28** 353–364.

MANN, H. B. and WALD, A. (1943). On the statistical treatment of stochastic difference equations. *Econometrica* **11** 173–220.

NEVEU, J. (1965). *Mathematical Foundations of the Calculus of Probability*. Holden-Day, San Francisco.

OGATA, Y. (1980). Maximum likelihood estimates of incorrect Markov models for time series and the derivation of AIC. *J. Appl. Probab.* **17** 59–72.

SHAMAN, P. and STINE, R. A. (1988). The bias of autoregressive coefficient estimators. *J. Amer. Statist. Assoc.* **83** 842–848.

DEPARTMENT OF STATISTICS AND
COMPUTATIONAL MATHEMATICS
THE UNIVERSITY OF LIVERPOOL
BROWNLOW HILL
P.O. BOX 147
LIVERPOOL L69 3BX
ENGLAND

DEPARTMENT OF MATHEMATICS
THE UNIVERSITY OF MANCHESTER
OXFORD ROAD
MANCHESTER M13 9PL
ENGLAND