The Annals of Statistics
1987, Vol. 15, No. 1, 46–58

ASYMPTOTIC DISTRIBUTIONS OF PREDICTION ERRORS AND RELATED TESTS OF FIT FOR NONSTATIONARY PROCESSES

BY I. V. BASAWA

La Trobe University

Limit distributions of prediction errors when the parameters are estimated are obtained for a general class of nonstationary processes that includes supercritical branching processes and explosive autoregressive processes as typical examples. The estimated prediction errors are used to construct new tests of fit whose limit distributions are also derived.

1. Introduction. Properties of prediction errors where the predictors are obtained by minimizing the mean square error are studied extensively in the time series literature. This work is predominantly concerned with stationary type processes. The effect of parameter estimation on the asymptotic prediction error variance is known to be negligible for a large class of stationary time series models (see, e.g., Box and Jenkins (1976), page 267). More recently, Fuller and Hasza (1981) (also see Fuller (1976), page 382) have obtained large sample approximations for the prediction mean square error for *nonstationary* autoregressive processes. It is clear from their work that the estimation error cannot be ignored asymptotically in the case of the nonstationary processes.

Basawa and Koul (1979) introduced a class of regular nonergodic processes that includes the nonstationary autoregressive processes and the explosive branching processes as special cases. This class is a generalization of Le Cam's (1960) locally asymptotically normal family. The monograph by Basawa and Scott (1983) summarises the main results on asymptotic inference for such processes and also contains several references. One of the topics not considered previously in the literature on inference for nonergodic processes is that of the behaviour of prediction errors when the parameters are estimated. Another topic of considerable interest is the problem of testing goodness of fit of a model when the primary aim is to use the proposed model for prediction. In this paper we consider both these problems and present some preliminary results.

In Section 2 we derive, under some broad assumptions, the limit distribution of the estimated prediction error vector of one-through-p-step ahead predictors for the nonergodic processes. Section 3 is concerned with a test of fit, which is based on the estimated prediction errors. Finally, the results are applied to explosive autoregressive processes and supercritical branching processes, both of which are typical members of the nonergodic family. Further applications, mainly to general linear and log-linear models, are discussed elsewhere (Basawa (1986)).

Received August 1984; revised September 1985.

AMS 1980 *subject classifications.* Primary 62M07, 62M09; secondary 62M10.

Key words and phrases. Estimated prediction errors, branching processes, explosive autoregressive processes, predictive tests of fit.

2. General formulation and limit distributions of prediction errors. Let $X(n) = (X_1, \ldots, X_n)$ denote a vector of sample observations from a stochastic process $X = \{X_1, X_2, \ldots\}$ defined on a probability space $(R^\infty, \mathscr{B}^\infty, P_\theta)$, $\theta \in \Omega \subset R^k$. Let P_θ^n denote a restriction of P_θ to $(R^n, \mathscr{B}^n)$, and $p_n(\cdot\,; \theta)$ the corresponding density with respect to some product measure μ^n on $(R^n, \mathscr{B}^n)$. We suppose that the family of densities $\{p_n(\cdot\,; \theta)\}$ is regular nonergodic in the sense of Basawa and Koul (1979). More specifically, it will be assumed that conditions (2.5)–(2.7) of Basawa and Brockwell (1984) are satisfied. See also Basawa and Scott (1983) for a discussion of asymptotic inference problems for such processes.

If $\hat{\theta}_n$ denotes the maximum likelihood estimator of θ, one can establish the asymptotic normality of $\hat{\theta}_n$ using an appropriate norming. For strictly nonergodic families this norming needs to be a nondegenerate random variable for the limit distribution to be normal. Let $\{\xi_{nj}(\theta),\ j = 1, \ldots, k\}$, $n = 1, 2, \ldots$, be a sequence of positive random variables, $\xi_{nj}(\theta) \uparrow \infty$ almost surely as $n \to \infty$, and denote the $(k \times k)$ diagonal matrix with diagonal elements $\{\xi_{nj}(\theta),\ j = 1, \ldots, k\}$ by $\xi_n(\theta)$. Suppose $\xi_n(\theta)$ is chosen such that

$$\xi_n^{1/2}(\theta)\big(\hat{\theta}_n - \theta\big) \to_d N_k(0, A(\theta)). \tag{2.1a}$$

If $\tilde{\theta}_n$ is some other estimator such that

$$\xi_n^{1/2}(\theta)\big(\tilde{\theta}_n - \theta\big) \to_d N_k(0, A^*(\theta)) \tag{2.1b}$$

one can show, under regularity conditions, that the matrix $A^*(\theta) - A(\theta)$ is nonnegative definite. See Heyde (1978) for a proof of the latter result for $k = 1$.

Let $Y(p) = (X_{n+1}, X_{n+2}, \ldots, X_{n+p})^T$ denote the vector of p future observations we wish to predict using the sample $X(n)$. It is well known and is easily verified that the choice $\hat{Y}_\theta(p)$, where

$$\hat{Y}_\theta(p) = E_\theta(Y(p)|\mathscr{B}^n), \tag{2.2}$$

minimizes the mean squared error of prediction when θ is known. When the parameter θ is unknown, it is a usual practice to replace θ by a suitable estimator $\tilde{\theta}_n$ in the optimal predictor $\hat{Y}_\theta(p)$. The question as to how well $\hat{Y}_{\tilde{\theta}_n}(p)$ approximates $\hat{Y}_\theta(p)$ in the sense of mean squared error of prediction leads to an important practical as well as theoretical problem. If $e_n(\theta)$ denotes the prediction error when the optimal predictor $\hat{Y}_\theta(p)$ is used, i.e., $e_n(\theta) = \hat{Y}_\theta(p) - Y(p)$, we have $e_n(\tilde{\theta}_n) = e_n(\theta) + (\hat{Y}_{\tilde{\theta}_n}(p) - \hat{Y}_\theta(p))$. In some nonergodic cases (see, e.g., the branching process application in Section 4) the optimal prediction error $e_n(\theta)$ does not remain bounded in probability for any fixed p as $n \to \infty$. It is therefore necessary to use an appropriate norming for asymptotic considerations. A natural norming is the diagonal matrix $\eta_n(\theta)$ whose diagonal elements are given by

$$\eta_{nj}(\theta) = \operatorname{var}_\theta\big(X_{n+j}|\mathscr{B}_n\big), \qquad j = 1, \ldots, p. \tag{2.3}$$

We therefore consider the standardised prediction errors,

$$\begin{aligned} \eta_n^{-1/2}(\theta)e_n\big(\tilde{\theta}_n\big) &= \eta_n^{-1/2}(\theta)e_n(\theta) + \eta_n^{-1/2}(\theta)\big(\hat{Y}_{\tilde{\theta}_n}(p) - \hat{Y}_\theta(p)\big) \\ &= U_n(\theta) + V_n(\theta). \end{aligned} \tag{2.4}$$

It follows from the properties of the conditional expectations that $U_n(\theta)$ and $V_n(\theta)$ are uncorrelated for every n. Whereas $U_n(\theta)$ represents the prediction error when θ is known, $V_n(\theta)$ is a contribution to the prediction error due to the estimation of θ. Consider the following condition:

$$\text{(C1)} \qquad \begin{pmatrix} U_n(\theta) \\ V_n(\theta) \end{pmatrix} \to_d N_{2p}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1(\theta) & 0 \\ 0 & \Sigma_2(\theta) \end{pmatrix}\right) \quad \text{under } (P_\theta),$$

where $\Sigma_1(\theta)$ is a nonsingular matrix. The matrix $\Sigma_2(\theta)$ is permitted to be singular.

One may use a linearization technique to simplify $V_n(\theta)$ further as follows:

$$V_n(\theta) = \eta_n^{-1/2}(\theta)\left(\frac{d\hat{Y}_\theta(p)}{d\theta}\right)_{\theta_n^*} \xi_n^{-1/2}(\theta)\xi_n^{1/2}(\theta)(\tilde{\theta}_n - \theta), \tag{2.5}$$

where the norming matrix $\xi_n(\theta)$ is such that (2.1b) holds, and θ_n^* is a $(p \times 1)$ random vector such that for some $0 \le u \le 1$, $\theta_n^* = \theta + u(\tilde{\theta}_n - \theta)$. Note that $(d\hat{Y}_\theta(p))/d\theta$ is a $(p \times k)$ matrix. Now consider the condition

(C2) There exists a $(p \times k)$ nonrandom matrix $G(\theta)$ such that, under P_θ,

$$\eta_n^{-1/2}(\theta)\left(\frac{d\hat{Y}(\theta)}{d\theta}\right)_{\theta_n^*} \xi_n^{-1/2}(\theta) = G(\theta) + o_p(1).$$

Under condition (C2) and the assumption (2.1b) we have

$$V_n(\theta) \to_d N_p(0, G(\theta)A^*(\theta)G^T(\theta)), \tag{2.6}$$

and hence $\Sigma_2(\theta)$ in (C1) is determined by

$$\Sigma_2(\theta) = G(\theta)A^*(\theta)G^T(\theta).$$

We can now state

THEOREM 2.1. *Under* (C1), (C2), *and* (2.1b) *we have*

$$\eta_n^{-1/2}(\theta)e_n(\tilde{\theta}_n) \to_d N_p(0, \Sigma(\theta)), \quad \textit{where}$$

$$\Sigma(\theta) = \Sigma_1(\theta) + G(\theta)A^*(\theta)G^T(\theta). \tag{2.7}$$

REMARKS. As one would expect from the decomposition in (2.4), the asymptotic prediction error variance $\Sigma(\theta)$ given in (2.7) is the sum of two terms, the first one being the asymptotic prediction error variance corresponding to the optimal predictor $\hat{Y}_\theta(p)$ when θ is known, and the second term representing the effect of parameter estimation. As remarked earlier (after (2.1b)) we may minimize $\Sigma(\theta)$ by choosing the maximum likelihood estimator $\hat{\theta}_n$ in place of $\tilde{\theta}_n$ in Theorem 2.1. Thus, we can state the corollary:

COROLLARY 2.1. *For regular nonergodic processes, and under the conditions of Theorem* 2.1, *we have*

$$\text{as. var}\left(\eta_n^{-1/2}(\theta)e_n(\tilde{\theta}_n)\right) - \text{as. var}\left(\eta_n^{-1/2}(\theta)e_n(\hat{\theta}_n)\right) \ge 0.$$

PROOF. If $\Sigma_0(\theta) = \Sigma_1(\theta) + G(\theta)A(\theta)G^T(\theta)$, where $A(\theta)$ is defined in (2.1a), and if $\Sigma(\theta)$ is given by (2.7), we have $\Sigma(\theta) - \Sigma_0(\theta) \geq 0$ in the sense that the matrix difference is nonnegative definite. This result essentially follows from Heyde's (1978) result mentioned after (2.1b), viz., $A^*(\theta) - A(\theta) \geq 0$. □

REMARKS. (i) The predictor $Y_{\hat{\theta}_n}(p)$, where $\hat{\theta}_n$ is the maximum likelihood estimator (or its equivalent), can be said to be *asymptotically efficient* in the sense that the asymptotic mean squared error (i.e., the mean squared error corresponding to the limiting distribution) is minimized. For the ergodic type processes such as the stationary autoregressive processes discussed by Box and Jenkins ((1976), page 267) it turns out that $G(\theta) \equiv 0$. Therefore, the predictor $Y_{\tilde{\theta}_n}(p)$, where $\tilde{\theta}_n$ is any estimator (not necessarily the efficient estimator such as the maximum likelihood estimator) satisfying (2.1b), would be asymptotically efficient in the above sense since $\Sigma(\theta) \equiv \Sigma_1(\theta)$, indicating that the effect of estimating θ is asymptotically negligible. For the explosive processes, however, $G(\theta)$ does not vanish and it is necessary to use an efficient estimator of θ in $Y_\theta(p)$ in order to obtain an efficient predictor.

(ii) We have used the random normings $\xi_n(\theta)$ and $\eta_n(\theta)$ in (2.1a), (2.1b) and (2.4) in order to obtain the asymptotic normality of the estimated prediction errors in Theorem 2.1. If we use appropriate nonrandom normings instead it can be shown that the limit distribution in Theorem 2.1 is a variance mixture of normals and hence *nonnormal*. Also, in general, the use of random norming usually results in a loss of information. However, in the type of applications considered in this paper, the random normings used can be shown to contain asymptotically negligible information as compared with $\hat{\theta}_n$ and $e_n(\hat{\theta}_n)$. One of the reviewers and the associate editor have noted that this phenomenon can be related to the concept of asymptotic ancillarity.

(iii) As mentioned above, if a nonrandom norming is used for $(\hat{\theta}_n - \theta)$ in (2.1a), we obtain a nonnormal distribution. However, the (unconditional) asymptotic optimality of $\hat{\theta}_n$ (with a nonrandom norming) can still be established as stated in Theorem 3 of Basawa and Scott ((1983), Chapter 2). The latter result replaces the usual asymptotic variance criterion used in limiting normal experiments by a more general criterion in terms of the limiting probability of concentration or a limiting risk function. The asymptotic (unconditional) optimality of $e_n(\hat{\theta}_n)$, using a nonrandom norming can similarly be established by replacing the asymptotic variance criterion in Corollary 2.1 by a more general criterion such as the probability of concentration.

3. A goodness of fit statistic for prediction. In this section we propose a new goodness of fit statistic based on the prediction errors. The test is useful mainly for the situations where prediction is the primary aim of fitting a model to data. We suppose that we have a sample of $N = n + p$ observations $X(N) = (X_1, \ldots, X_n, X_{n+1}, \ldots, X_{n+p}) = (X(n), Y(p))$, with n large and p relatively small, and fixed. First, pretend that the observations $Y(p)$ are not available. We can then estimate (or predict) $Y(p)$ from the sample $X(n)$ by $\hat{Y}_{\theta_0}(p)$, assuming for the moment that the parameter value θ_0 is known under the null hypothesis.

Since the vector $Y(p)$ has in fact been observed we can now directly compare the estimates $\hat{Y}_{\theta_0}(p)$ with the observed $Y(p)$ and compute the standardized prediction errors $U_n(\theta_0) = \eta_n^{-1/2}(\theta_0)(\hat{Y}_{\theta_0}(p) - Y(p))$. Consider the null hypothesis

$$H_0\text{: } X(N) \text{ has the joint density } p_N^{\circ}(\cdot\,; \theta_0). \tag{3.1}$$

The statistic proposed for testing H_0 is the quadratic form

$$Q_n^{\circ} = U_n^T(\theta_0)\Sigma_1^{-1}(\theta_0)U_n(\theta_0). \tag{3.2}$$

If we assume that (cf. (C1) in Section 2)

$$U_n(\theta_0) \to_d N_p(0, \Sigma_1(\theta_0)) \quad \text{under } (P_{\theta_0}^{\circ}), \tag{3.3}$$

where $\Sigma_1(\theta_0)$ is a nonsingular matrix we can deduce directly the null limit distribution of Q_n° as

$$Q_n^{\circ} \to_d \chi^2(p). \tag{3.4}$$

In order to study the asymptotic power properties of Q_n° one can consider a sequence of alternative hypotheses of the type

$$K_n^{\circ}\text{: } X(N) \text{ has the joint density } p^{\circ}(\cdot\,; \theta_n(h)), \tag{3.5}$$

where $\theta_n(h) = \theta_0 + I_n^{-1/2}(\theta_0)h$, $I_n(\theta_0)$ is a $(k \times k)$ nonrandom diagonal matrix with diagonal elements $0 < I_{nj}(\theta_0) \uparrow \infty$ as $n \to \infty$. Assume that the following conditions (C3) are satisfied:

$$\text{(C3)(i)} \qquad \eta_n^{-1/2}(\theta_0)\left(\frac{d\hat{Y}_{\theta}(p)}{d\theta}\right)_{\theta_n^*(u)} \xi_n(\theta_0) = G(\theta_0) + o_p(1),$$

under $P_{\theta_n}(h)$ for all u, $0 \le u \le 1$, where $\theta_n^*(u) = \theta_0 + uI_n^{-1/2}(\theta_0)h$, and $\xi_n(\theta)$ is a certain random matrix; see, e.g., (C2) in Section 2.

$$\text{(C3)(ii)} \qquad \begin{pmatrix} \eta_n^{-1/2}(\theta_0)\left(\hat{Y}_{\theta_n(h)}(p) - Y(p)\right) \\ \xi_n(\theta_0)I_n^{-1}(\theta_0) \end{pmatrix} \to_d \begin{pmatrix} \Sigma_1^{1/2}(\theta_0)Z \\ W(\theta_0) \end{pmatrix} \quad \text{under } \left(P_{\theta_n(h)}\right),$$

where $\Sigma_1(\theta)$ is a $(p \times p)$ nonrandom nonsingular matrix, Z is a $(p \times 1)$ vector of independent $N(0,1)$ random variables, $W(\theta_0)$ is a $(k \times k)$ diagonal matrix whose diagonal elements are a.s. positive random variables independent of Z.

THEOREM 3.1. *Under conditions* (C3) *we have, under* $(P_{\theta_n(h)})$,

$$Q_n^{\circ} \to_d \chi_*^2(p, \Lambda),$$

where $\Lambda = h^T W^{1/2}(\theta_0)G^T(\theta_0)\Sigma_1^{-1}(\theta_0)G(\theta_0)W^{1/2}(\theta_0)h$. *The random variable* χ_*^2 *is such that, conditional on* $\Lambda = \lambda$, *it has a noncentral chi-square distribution with p degrees of freedom and the noncentrality parameter* λ.

PROOF. We give an outline of the proof only.

$$
\begin{aligned}
U_n(\theta_0) &= \eta_n^{-1/2}(\theta_0)\big(\hat{Y}_{\theta_0}(p) - Y(p)\big) \\
&= \eta_n^{-1/2}(\theta_0)\big(\hat{Y}_{\theta_n(h)}(p) - Y(p)\big) - \eta_n^{-1/2}(\theta_0)\big(\hat{Y}_{\theta_n(h)} - \hat{Y}_{\theta_0}(p)\big) \\
&= \eta_n^{-1/2}(\theta_0)\big(\hat{Y}_{\theta_n(h)}(p) - Y(p)\big) - \eta_n^{-1/2}(\theta_0)\left(\frac{d\hat{Y}_\theta(p)}{d\theta}\right)_{\theta_n^*(u)}(\theta_n(h) - \theta_0) \\
&= \eta_n^{-1/2}(\theta_0)\big(\hat{Y}_{\theta_n(h)}(p) - Y(p)\big) \\
&\quad - \left\{\eta_n^{-1/2}(\theta_0)\left(\frac{d\hat{Y}_\theta(p)}{d\theta}\right)_{\theta_n^*(u)}\xi_n^{-1/2}(\theta_0)\right\}\xi_n^{1/2}(\theta_0)I_n^{-1/2}(\theta_0)h.
\end{aligned}
$$

Condition (C3) then gives

$$
U_n(\theta_0) \to_d N_p^*\big(-G(\theta_0)W^{1/2}(\theta_0)h,\ \Sigma_1(\theta_0)\big) \quad \text{under } P_{\theta_n(h)}, \tag{3.6}
$$

where *conditional* on $W(\theta_0)$, N_p^* is a p-variate normal distribution. The result in Theorem 3.1 now follows readily from (3.6). □

REMARKS. For $h \equiv 0$ note that Theorem 3.1 gives the null distribution in (3.4). For nonergodic type processes, the nonnull limit distribution is a mixture of noncentral chi-square, with the distribution of $W(\theta_0)$ acting as the mixing distribution. We assume $G \not\equiv 0$ in the preceding theorem to avoid triviality.

Consider now the composite null hypothesis when θ is unspecified:

$$
H_\theta\text{: } X(N) \text{ has the joint density } p_N^\circ(\cdot\,;\theta),\ \theta \in \Omega. \tag{3.7}
$$

To test H_θ we propose the statistic

$$
Q_n = e_n^T(\tilde{\theta}_n)\eta_n^{-1/2}(\tilde{\theta}_n)\Sigma^{-1}(\tilde{\theta}_n)\eta_n^{-1/2}(\tilde{\theta}_n)e_n(\tilde{\theta}_n), \tag{3.8}
$$

where it is assumed that Σ is nonsingular. From Theorem 2.1 and standard asymptotics it follows readily that

$$
Q_n \to_d \chi^2(p) \quad \text{under } (P_\theta^\circ), \text{ for any } \theta \in \Omega. \tag{3.9}
$$

The limit distribution of Q_n under H_θ is thus given by (3.9).

4. Applications. In this section we present two applications, both of which are nonergodic and explosive. We give the main arguments and the results, only omitting some details.

EXAMPLE 1. *Galton–Watson branching process.* Let $X = \{X_1, X_2, \dots\}$ be a Galton–Watson branching process with $X_0 = 1$, and μ and σ^2 denote the mean and the variance of the offspring distribution, $\theta = (\mu, \sigma^2)$. We assume that the process is supercritical, i.e., $\mu > 1$. Let $E^c = \{\omega\text{: } X_n(\omega) > 0, \text{ for all } n\}$ denote the

set of nonextinction. Conditional on E^c, it is well known that $X_n \to \infty$ almost surely as $n \to \infty$, and hence X is an *explosive* process in this sense. In what follows all the limit results are conditional on E^c.

Since $E(X_{n+t}|\mathscr{B}^n) = \mu^t X_n$, $t = 1, 2, \ldots$, we have

$$\hat{Y}_\theta(p) = \left(\mu X_n, \mu^2 X_n, \ldots, \mu^p X_n\right)^T. \tag{4.1}$$

Since the predictor in (4.1) depends on θ only through μ we may assume initially that σ^2 is known. However, the various norms used to obtain the limit distributions do depend on σ^2, and a consistent estimate of σ^2 will be needed for the construction of the test statistic when σ^2 is unknown. We shall return to this point later.

We propose to use the Harris estimate of μ (see Harris (1948)) given by $\tilde{\mu}_n = \Sigma_k^n X_j / \Sigma_1^n X_{j-1}$. Harris derived this as a nonparametric "maximum likelihood" estimate without assuming any specific offspring distribution. The estimate can alternatively be derived as a weighted conditional least square estimate (see Basawa and Prakasa Rao (1980), Chapter 2) by minimizing $D_n(\mu)$ with respect to μ, where

$$\begin{aligned} D_n(\mu) &= \sum \left[\left\{ X_j - E\left(X_j|\mathscr{B}^{j-1}\right) \right\}^2 / \operatorname{var}\left(X_j|\mathscr{B}^{j-1}\right) \right] \\ &= \sum \left\{ \left(X_j - \mu X_{j-1}\right)^2 / \left(X_{j-1}\sigma^2\right) \right\}. \end{aligned} \tag{4.2}$$

Taking $\xi_n(\theta) = \sigma^{-2}\Sigma_1^n X_{j-1}$, it is well known in the literature (see, e.g., Dion (1974), Theorem 3.1) that

$$\xi_n^{1/2}(\theta)(\tilde{\mu}_n - \mu) \to_d N(0,1) \quad \text{as } n \to \infty. \tag{4.3}$$

Thus, (2.1b) is satisfied with $A^*(\theta) = 1$.

If we assume that the offspring distribution is given by the power series distribution, viz.,

$$P(X_1 = j) = a_j \lambda^j / f(\lambda), \qquad j = 0, 1, 2, \ldots, \ a_j \geq 0, \ \lambda > 0,$$

and $f(\lambda) = \Sigma_0^\infty a_j \lambda^j$, it follows that (see, e.g., Heyde (1975)) the maximum likelihood estimator $\hat{\mu}_n$ of μ, where $\mu = EX_1 = \lambda\{d/d\lambda(\ln f(\lambda))\}$, is identical with the Harris estimate $\tilde{\mu}_n$ considered earlier. Consequently, at least for the power series offspring distribution, $\hat{\mu}_n$ and $\tilde{\mu}_n$ in (2.1a) and (2.1b) are identical, and $A(\theta) = A^*(\theta) = 1$.

Returning to the general case we can verify that

$$\eta_{nj}(\theta) = \operatorname{var}\left(X_{n+j}|\mathscr{B}^n\right) = \sigma^2 X_n \left\{ \frac{\mu^{j-1}(\mu^j - 1)}{(\mu - 1)} \right\}. \tag{4.4}$$

Consider now the j-step ahead prediction error

$$\begin{aligned} U_{nj}(\theta) &= \eta_{nj}^{-1/2}(\theta)\left(\mu^j X_n - X_{n+j}\right) \\ &= \eta_{nj}^{-1/2}(\theta) \sum_{i=1}^{X_n} \left\{ \mu^j - Z_i(n, j) \right\}, \qquad j = 1, \ldots, p, \end{aligned} \tag{4.5}$$

where $Z_i(n, j)$, $i = 1, \ldots, X_n$, are independent and identically distributed random variables each distributed as X_j. Note that $Z_i(n, j)$ denotes the jth generation size of a new branching process started by the ith individual of the nth generation of the original branching process. We have

$$E\{Z_i(n, j)\} = \mu^j \quad \text{and} \quad \operatorname{var}\{Z_i(n, j)\} = \sigma^2\mu^{j-1}(\mu^j - 1)/(\mu - 1). \tag{4.6}$$

Using similar arguments to those in Dion (1974) we find, via (4.4)–(4.6), that

$$U_{nj}(\theta) \to_d N(0,1), \qquad j = 1, \ldots, p. \tag{4.7}$$

The Cramér–Wold device then gives the limit distribution of the vector $U_n(\theta) = (U_{nj}(\theta), j = 1, \ldots, p)^T$;

$$U_n(\theta) \to_d N_p(0, \Sigma_1(\theta)), \tag{4.8}$$

where routine calculation yields $\Sigma_1(\theta) = ((\sigma_1(i, j)))$ with

$$\sigma_1(i, j) = \mu^{(i-j)/2}\left(\frac{\mu^j - 1}{\mu^i - 1}\right)^{1/2}, \qquad i \geq j. \tag{4.9}$$

Now consider

$$\begin{aligned} V_{nj}(\theta) &= \eta_{nj}^{-1/2}(\theta)\left(\tilde{\mu}_n^j - \mu^j\right)X_n \\ &= \left(\frac{X_n}{\Sigma_1^n X_{j-1}}\right)^{1/2}\left(\frac{(\mu - 1)}{\mu^{j-1}(\mu^j - 1)}\right)^{1/2} \xi_n^{1/2}(\theta)\left(\tilde{\mu}_n^j - \mu^j\right). \end{aligned} \tag{4.10}$$

It is easily verified that

$$\left(\frac{X_n}{\Sigma X_{j-1}}\right) \to (\mu - 1) \quad \text{a.s.} \tag{4.11}$$

Now, using (4.11), (4.3), and a well known convergence theorem (see Rao (1973), page 385) it follows from (4.10) that, for $j = 1, \ldots, p$,

$$V_{nj}(\theta) \to_d N\left(0, j^2\mu^{j-1}(\mu - 1)^2/(\mu^j - 1)\right). \tag{4.12}$$

Finally, the Cramér–Wold argument gives

$$V_n(\theta) \to_d N(0, \Sigma_2(\theta)), \tag{4.13}$$

where $\Sigma_2(\theta) = ((\sigma_2(i, j)))$ with

$$\sigma_2(i, j) = ij\mu^{(i+j-2)/2}(\mu - 1)^2\left\{(\mu^i - 1)(\mu^j - 1)\right\}^{-1/2}, \qquad i \geq j. \tag{4.14}$$

Condition (C1) can then be verified using the Cramér–Wold device. It may be noted that (C2) can be verified with $G(\theta)$ being a $(p \times 1)$ vector with elements

$$i\mu^{(i-1)/2}(\mu - 1)(\mu^i - 1)^{-1/2}, \qquad i = 1, \ldots, p. \tag{4.15}$$

In any case, we have obtained $\Sigma_2(\theta)$ directly in (4.14). Thus, the conditions of Theorem 2.1 are verified for the branching process application. The predictor $Y_{\tilde{\theta}_n}(p)$ is efficient in the sense of Section 2 for the power series offspring distribution, since in this case $\tilde{\theta}_n \equiv \hat{\theta}_n$, the maximum likelihood estimator.

For testing the composite hypothesis H_θ, where H_θ: $X = (X_1, X_2, \dots)$ is a branching process with mean $\theta > 1$ (θ unknown), we can use Q_n given by (3.8) as a test statistic. When σ^2 is unknown it may be estimated by a consistent estimate such as

$$\tilde{\sigma}_n^2 = n^{-1} \sum_1^n \left\{ \left(X_j - \tilde{\mu}_n X_{j-1} \right)^2 / X_{j-1} \right\}. \tag{4.16}$$

The estimate (4.16) was suggested by Basawa and Prakasa Rao ((1980), Chapter 2) and is motivated by the expression $D_n(\mu)$ in (4.2) from which the estimate $\tilde{\mu}_n$ was derived.

Suppose now that μ_0 (> 1) is specified by the null hypothesis H_0 in (3.1). Assume in particular that the offspring distribution is geometric on the set $\{1, 2, \dots\}$ with mean μ_0. The conditions (C3) can be verified with $I_n(\theta_0) = E_{\theta_0}\{\xi_n(\theta_0)\}$ and $W(\theta_0)$ having an exponential density with mean 1. Theorem 3.1 then holds with $G(\theta_0)$ specified by (4.15) and $W(\theta_0)$ being a standard exponential random variable.

EXAMPLE 2. *Explosive autoregressive processes.* Consider a kth order autoregressive process $X = \{X_1, X_2, \dots\}$, defined by

$$X_n - (\theta_1 X_{n-1} + \theta_2 X_{n-2} + \cdots + \theta_k X_{n-k}) = Z_n, \qquad n = 1, 2, \dots, \tag{4.17}$$

where $\{Z_n\}$ are independent $N(0, 1)$ variates, with $X_n = 0$ for $1 - p \le n \le 0$. Consider the polynomial equation

$$m^k - \left(\theta_1 m^{k-1} + \theta_2 m^{k-2} + \cdots + \theta_k \right) = 0. \tag{4.18}$$

The process X will be stationary if the roots of the above polynomial equation are all less than unity in absolute value. The prediction problem for the stationary case has been studied extensively in the literature (see Box and Jenkins (1976)). We now consider the explosive case where we assume that there exists a root β of (4.18) such that it is the largest of all the k roots in absolute value, and $|\beta| > 1$. The remaining $(k - 1)$ roots are assumed to be less than unity in absolute value. See Basawa and Brockwell (1984), Basawa and Koul (1979) and Fuller and Hasza (1981) for discussions of various inference problems for explosive processes.

For the above model with $|\beta| > 1$, it can be shown as in Basawa and Brockwell (1984) that there exists a $N(0, 1)$ random variable $Z(\theta)$ such that $\sqrt{(\beta^2 - 1)}\, \beta^{-n} X_n$ converges to $Z(\theta)$ a.s. as $n \to \infty$, and consequently $|X_n| \to \infty$ a.s. Consider

$$\begin{aligned} \hat{Y}_\theta(p) &= \left(E\left(X_{n+j} | \mathscr{B}^n \right),\ j = 1, 2, \dots, p \right) \\ &= \left(Y_\theta^{(1)}, Y_\theta^{(2)}, \dots, Y_\theta^{(p)} \right), \end{aligned}$$

where $E(X_{n+j} | \mathscr{B}^n) = Y_\theta^{(j)}$ satisfies the recursive relation

$$Y_\theta^{(j)} = \sum_{i=1}^k \theta_i Y_\theta^{(j-1)}, \tag{4.19}$$

with $Y_\theta^{(\delta)} = X_\delta$ for $\delta \le n$. In particular, for $k = 1$, $Y_\theta^{(j)} = \theta_1^j X_n$. In the general case ($k \ge 1$) the prediction error can be written in terms of the innovations Z as

$$\left(Y_\theta^{(j)} - X_{n+j}\right) = -\sum_{i=0}^{j-1} a_i(\theta) Z_{n+j-i}(\theta), \qquad j = 1, \ldots, p, \tag{4.20}$$

where from (4.17),

$$Z_r(\theta) = X_r - (\theta_1 X_{r-1} + \theta_2 X_{r-2} + \cdots + \theta_k X_{r-k}).$$

The coefficients $a_i(\theta)$ satisfy the relations:

$$\begin{aligned} a_i(\theta) - \sum_{\delta=1}^{k} \theta_\delta a_{i-\delta}(\theta) &= 0, \qquad i = 1, 2, \ldots, j-1, \\ a_0(\theta) &= 1, \qquad a_i(\theta) = 0 \quad \text{for } i < 0. \end{aligned} \tag{4.21}$$

See Box and Jenkins ((1976), page 128) for the details. Also, we have

$$\eta_{nj}(\theta) = \operatorname{var}\left(X_{n+j} | \mathscr{B}^n\right) = \sum_{i-0}^{j-1} a_i^2(\theta), \tag{4.22}$$

where $a_i(\theta)$ are determined by (4.21). For the first-order autoregressive process, we have $k = 1$,

$$a_i(\theta_1) = \theta_1^i, \qquad 1 \le i \le j-1, \quad \text{and} \quad \eta_{nj}(\theta_1) = \sum_{i=0}^{j-1} \left(\theta_1^i\right)^2 = \frac{\theta_1^{2j} - 1}{\theta_1^2 - 1}.$$

Now, for any $k \ge 1$,

$$U_{nj}(\theta) = \eta_{nj}^{-1/2}(\theta)\left(Y_\theta^{(j)} - X_{n+j}\right), \qquad j = 1, 2, \ldots, p,$$

are correlated $N(0, 1)$ variables, and thus

$$U_n(\theta) \to_d N(0, \Sigma_1(\theta)), \tag{4.23}$$

where $\Sigma_1(\theta) = ((\sigma_1(i, j)))$ and

$$\sigma_1(i, j) = \left\{\sum_{l=0}^{i-1} a_l^2(\theta)\right\}^{-1/2} \left\{\sum_{l=0}^{j-1} a_l^2(\theta)\right\}^{-1/2} \left\{\sum_{l=0}^{i-1} a_l(\theta) a_{j-i+l}(\theta)\right\}, \tag{4.24}$$

$$i \le j, 1 \le i, j \le p.$$

For the special case $k = 1$, we have

$$\sigma_1(i, j) = \theta_1^{j-i}\left(\frac{\theta_1^{2i} - 1}{\theta_1^{2j} - 1}\right)^{1/2}, \qquad i \le j, 1 \le i, j \le p.$$

Let $\hat{\theta}_n$ be the maximum likelihood estimate of θ obtained as a solution of the equation

$$\sum_{r=1}^{n} X_{r-i}(X_r - \theta_1 X_{r-1} - \cdots - \theta_k X_{r-k}) = 0, \qquad 1 \le i \le k. \tag{4.25}$$

Taking $\xi_{nj}(\theta) = X_n^2$, $1 \le j \le k$, it can be shown using Theorem 5.1 of Basawa and Brockwell (1984) that the result in (2.1a) obtains with $A(\theta) = ((a_{ij}(\theta)))$,

$$a_{ij}(\theta) = \left\{ \frac{|\beta|(1-\beta^{-2})^{3/2}}{(1-\beta^{-2k})} \right\}^2 \beta^{2-i-j}, \qquad 1 \le i, j \le k. \tag{4.26}$$

Also, see Anderson (1959) for early work on this topic. For the special case $k = 1$, we have $a_{11}(\theta) = \theta_1^2 - 1$. It may be noted for $k > 1$ that the matrix $A(\theta)$ given by (4.25) is singular with rank unity.

From (4.20) we have

$$\begin{aligned} \frac{\partial Y_\theta^{(j)}}{\partial \theta_r} &= \frac{\partial}{\partial \theta_r}\left[-\sum_{i=0}^{j-1} a_i(\theta) Z_{n+j-i}(\theta) \right] \\ &= -\left[\sum_{i=0}^{j-1} \frac{\partial a_i(\theta)}{\partial \theta_r} Z_{n+j-i}(\theta) + \sum_{i=0}^{j-1} a_i(\theta) \frac{\partial Z_{n+j-i}(\theta)}{\partial \theta_r} \right] \qquad (4.27) \\ &= -\left[\sum_{i=0}^{j-1} \frac{\partial a_i(\theta)}{\partial \theta_r} Z_{n+j-i}(\theta) - \sum_{i=0}^{j-1} a_i(\theta) X_{n+j-i-r} \right], \qquad 1 \le r \le k, \end{aligned}$$

where $a_i(\theta)$ are determined by the difference equation (4.21).

Note that the first term on the right of (4.27) is a linear combination of a finite number of independent and identically distributed (innovation) random variables, and hence remains bounded in probability as $n \to \infty$. We can check (C2) (of Section 2) taking $\eta_{nj}(\theta) = \sum_{i=0}^{j-1} a_i^2(\theta)$ (free from n), $\xi_{nj}(\theta) = X_n^2$, and using the fact that $\sqrt{(\beta^2 - 1)}\,\beta^{-n} X_n$ converges in probability to a $N(0,1)$ random variable $Z(\theta)$. We finally get

$$g_{rj}(\theta) = \left\{ \sum_{i=0}^{j-1} a_i(\theta) \beta_i(r, j) \right\} \left\{ \sum_{i=0}^{j-1} a_i^2(\theta) \right\}^{-1/2}, \qquad 1 \le r \le k, 1 \le j \le p, \tag{4.28}$$

where

$$\beta_i(r, j) = \{\operatorname{sgn} Z(\theta)\} \lim_{n \to \infty} \beta^{n+j-i-r} |\beta|^{-n}, \tag{4.29}$$

sgn $Z(\theta)$ stands for 1 and -1 for $Z(\theta) > 0$ and < 0, respectively. Note that we have ignored the case $Z(\theta) = 0$ in obtaining the preceding limit (in probability) since $P(Z(\theta) = 0) = 0$. Thus, Theorem 2.1 holds with $\Sigma_1(\theta)$, $A(\theta)$, and $G(\theta)$ determined by (4.24), (4.26), and (4.28), respectively.

For the special case $k = 1$, we have $a_i(\theta) = \theta_1^i$, $\beta = \theta_1$, giving

$$g_{1j}(\theta_1) = \left\{ \sum_{i=0}^{j-1} \theta_1^i \beta_i(1, j) \right\} \left(\frac{\theta_1^2 - 1}{\theta_1^{2j} - 1} \right)^{1/2}, \qquad 1 \le j \le p, \tag{4.30}$$

where $\beta_i(1, j)$ is given by (4.29) with β replaced by θ_1. Thus, for $k = 1$,

$\Sigma(\theta) = ((\sigma(i, j)))$ in (2.7) is seen to be

$$\sigma(i, j) = \theta_1^{j-i}\left(\frac{\theta_1^{2i} - 1}{\theta_1^{2j} - 1}\right)^{1/2} + \left(\theta_1^2 - 1\right)^2\left(\theta_1^{2i} - 1\right)^{-1/2} \times\left(\theta_1^{2j} - 1\right)^{-1/2}(i\theta^{i-1})(j\theta^{j-1}), \qquad i \le j. \tag{4.31}$$

We now consider the stationary case. For simplicity take $k = 1$, and assume $|\theta_1| < 1$ for stationarity. Set $\xi_{nj}(\theta) = n$ and $\eta_{nj}(\theta) = (\theta_1^{2j} - 1)/(\theta_1^2 - 1)$. It is then easily verified that $V_n(\theta)$ in (2.5) converges in probability to zero since $G(\theta)$ in (C2) equals zero and $\sqrt{n}\,(\hat{\theta}_n - \theta)$ is bounded in probability as $n \to \infty$. Thus, as remarked in Section 2 for the nonexplosive ergodic case $|\theta_1| < 1$, the second term in (2.7) vanishes. Consequently, $\Sigma(\theta)$ is determined by the first term on the right of (4.31) when $|\theta_1| < 1$.

Returning to the nonstationary case, the results of Section 3 are directly applicable. Consider the case $k = 1$ and $|\theta_1| > 1$ (explosive). The composite null hypothesis (3.7) here is:

> H: X is a first order Gaussian explosive autoregressive process with the unknown parameter θ_1, $|\theta_1| > 1$.

The statistic Q_n with Σ determined by (4.31) will then have the limiting chi-square distribution under H_θ (see (3.9)).

The statistics Q_n° in (3.2) can be used to test the simple hypothesis H_0: X is a first order Gaussian autoregressive process with a specified parameter $\theta_1 = \theta_{10}$, $|\theta_{10}| > 1$. Conditions (C3) can be verified with W having a chi-square distribution with one degree of freedom if we take $\xi_{nj}(\theta) = X_n^2$ and $I_{nj}(\theta) = \theta_1^{2n}/(\theta_1^2 - 1)$. The nonnull limit distribution in Theorem 3.1 then obtains.

Acknowledgment. My thanks are to the referees for helpful comments and suggested improvements on an earlier version.

REFERENCES

ANDERSON, T. W. (1959). On asymptotic distributions of estimates of parameters of stochastic difference equations. *Ann. Math. Statist.* **30** 676–687.

BASAWA, I. V. (1986). Statistical forecasting for stochastic processes. To appear in *Ann. Oper. Res.* (under Proc. T.I.M.S. and O.R.S.A. meeting held at Williamsburgh).

BASAWA, I. V. and BROCKWELL, P. J. (1984). Asymptotic conditional inference for regular non-ergodic models with an application to autoregressive processes. *Ann. Statist.* **12** 161–171.

BASAWA, I. V. and KOUL, H. L. (1979). Asymptotic tests of composite hypotheses for non-ergodic type stochastic processes. *Stochastic Process. Appl.* **9** 291–305.

BASAWA, I. V. and PRAKASA RAO, B. L. S. (1980). *Statistical Inference for Stochastic Processes.* Academic, London.

BASAWA, I. V. and SCOTT, D. J. (1983). *Asymptotic Optimal Inference for Non-ergodic Models. Lecture Notes in Statist.* **17**. Springer, Berlin.

BOX, G. E. P. and JENKINS, G. M. (1976). *Time Series Analysis: forecasting and control*, 2nd ed. Holden-Day, San Francisco.

DION, J.-P. (1974). Estimation of the mean and the initial probabilities of a branching process. *J. Appl. Probab.* **11** 687–694.

FULLER, W. A. (1976). *Introduction to Statistical Time Series*. Wiley, New York.
FULLER, W. A. and HASZA, D. P. (1981). Properties of predictors for autoregressive time series. *J. Amer. Statist. Assoc.* **76** 155–161.
HARRIS, T. E. (1948). Branching processes. *Ann. Math. Statist.* **19** 474–494.
HEYDE, C. C. (1975). Remarks on efficiency in estimation for branching processes. *Biometrika* **62** 49–55.
HEYDE, C. C. (1978). An optimum property of the maximum likelihood estimator for stochastic processes. *Stochastic Process. Appl.* **8** 1–9.
LE CAM, L. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.* **3** 37–98.
RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF GEORGIA
ATHENS, GEORGIA 30602