Statistica Sinica **7**(1997), 425-449

DIRECT AUTOREGRESSIVE PREDICTORS FOR MULTISTEP PREDICTION: ORDER SELECTION AND PERFORMANCE RELATIVE TO THE PLUG IN PREDICTORS

R. J. Bhansali

University of Liverpool

Abstract: A direct method for multistep prediction of a stationary time series consists of fitting a new autoregression for each lead time, h, by a linear regression procedure and to select the order to be fitted from the data. By contrast, a more usual 'plug in' method involves the least-squares fitting of an initial kth order autoregression; the multistep forecasts are then obtained from the model equation, but with the unknown future values replaced by their own forecasts. The asymptotic distributions of the direct and plug in estimates of the h-step prediction constants and their respective mean squared errors of prediction are derived for a finite autoregressive process; explicit asymptotic expressions for comparing the loss in predictive and parameter estimation efficiency due to using the direct method instead of the plug in method in this situation are also given. The finite sample behaviour of the prediction errors with these two methods is investigated by a simulation study.

Key words and phrases: AIC, FPE, order determination, time series.

1. Introduction

Consider a discrete-time autoregressive process of order m,

$$\sum_{u=0}^{m} a_m(u)x_{t-u} = \varepsilon_t, \quad a_m(0) = 1, \tag{1.1}$$

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

$$A_m(z) = \sum_{j=0}^{m} a_m(j)z^j \tag{1.2}$$

is bounded away from zero, $|z| \leq 1$ and $\sum |a_m(j)| < \infty$.

In practice, having only observed $x_1, \ldots, x_T$, m is invariably unknown. If m is finite, the Akaike information criterion, AIC, or the Final Prediction Error, FPE, criterion (see Akaike (1970, 1973)), do not provide a consistent estimator

of m, but m may be estimated consistently by using, for example, a criterion of Schwarz (1978) (see also Hannan and Quinn (1979)).

If, on the other hand, m is infinite, the orders selected by the AIC and FPE criteria and that by a criterion introduced by Shibata (1980, 1981) are asymptotically efficient for one-step prediction and spectral estimation, in the sense defined by this author; furthermore, now the consistent criteria referred to above are not asymptotically efficient in this sense.

Bhansali (1996) has recently shown that even when m in (1.1) is infinite, AIC and the related criteria are not asymptotically efficient in the Shibata sense for h-step prediction, $h > 1$. The notion of an asymptotically efficient model selection for h-step prediction is also introduced there: A direct procedure involving a linear least-squares regression of x_{t+h} on $x_t, \ldots, x_{t-k+1}$ is used for estimating the prediction constants, with $k = \tilde{k}_h$, say, treated as a random variable whose value is selected anew for each h by an order selection criterion. An asymptotic lower bound for the resulting mean squared error of prediction is derived and it is shown that the order selection by suitable h-step generalizations of the AIC and FPE criteria and that by an h-step criterion of Shibata (1980) are asymptotically efficient for h-step prediction as the bound is attained in the limit if $\tilde{k}_h$ is selected by any of these criteria.

The direct method may be viewed as an alternative to the widely-used 'plug in' method in which the multistep forecasts are obtained from an initial autoregression fitted to the data, but by repeatedly iterating the model and replacing the unknown future values by their own forecasts. If m is finite and the fitted order, k, of the autoregression is such that $k \geq m$, then for a Gaussian process the plug in method provides maximum likelihood estimates of the prediction constants. This result does not hold, however, if m is infinite; for this situation, Bhansali (1996) has also derived a lower bound for the h-step mean squared error of prediction of the plug in method, but with the initial order selected by AIC or a related criterion. The results given there point to a two-fold advantage of the direct method for multistep prediction: first, the asymptotic lower bound on its mean squared error of prediction is smaller than that for the plug in method; secondly, whereas the former bound is attainable asymptotically, that for the plug in method is not even asymptotically attainable.

The analysis described above is, however, based on the assumption of an infinite m. In practice, an analyst may not know whether this is so. Thus, a pertinent question and one addressed here is: How does the direct method behave for finite m?

In Section 3, the asymptotic distributions of the direct and plug in estimates of the h-step prediction constants, and asymptotic expressions for their respective

mean squared errors of prediction, are derived. For a finite m, unlike the plug in method, the direct method is shown not to provide asymptotically efficient estimates of the prediction constants. Explicit asymptotic expressions for evaluating and comparing the loss in parameter estimation efficiency and predictive efficiency due to using the direct method instead of the plug in method are also given in Section 3. In Section 4, these results are illustrated for some specific values of m and h. In Section 5, the h-step AIC and FPE criteria are shown not to be consistent for m even though they are asymptotically efficient in the sense described above. Simulation results are presented in Section 6.

Earlier references advocating lead-time dependent model selection and/or parameter estimation for multistep forecasting include Findley (1983), Tiao and Xu (1993) and Lin and Granger (1994).

2. Preliminaries

Suppose that the observed time series is a realization of a discrete-time process, $\{x_t\}$, satisfying the following assumption:

Assumption 1. $\{x_t\}$ has representation (1.1) where $m \geq 0$ is finite, $A_m(z)$ is bounded away from zero, $|z| \leq 1$, and $\{\varepsilon_t\}$ is a sequence of independent identically distributed random variables each with mean 0, variance σ^2 and finite fourth cumulant τ_4.

Having observed $x_1, \ldots, x_T$, $T > 1$, the kth order least- squares estimates, $\hat{\boldsymbol{a}}(k) = [\hat{a}_k(1), \ldots, \hat{a}_k(k)]'$ and $\hat{\sigma}^2(k)$ of the autoregressive parameters, conditional on a knowledge of $x_1, \ldots, x_k$, are given by:

$$\hat{\boldsymbol{a}}(k) = -\hat{\boldsymbol{R}}(k)^{-1}\hat{\boldsymbol{r}}(k), \tag{2.1}$$

$$\hat{\sigma}^2(k) = C_T(0,0) + \sum_{j=1}^{k} \hat{a}_k(j) C_T(0,1), \tag{2.2}$$

where $\hat{\boldsymbol{R}}(k) = [C_T(u,v)](u,v = 1, \ldots, k)$, $\hat{\boldsymbol{r}}(k) = [C_T(0,1), \ldots, C_T(0,k)]'$ and

$$C_T(u,v) = (T-K)^{-1} \sum_{t=K}^{T-1} x_{t-u+1} x_{t-v+1} \qquad (u,v = 0,1,\ldots,k).$$

Here, $k \geq 1$, is an arbitrary integer, that is, it does not necessarily equal m, K denotes the maximum autoregressive order to be fitted.

If the ε_t are Normally distributed and $k = m$, the estimates (2.1) - (2.2) are well known (Anderson (1971)) to also provide approximate maximum likelihood estimates of the autoregressive parameters. In what follows we call them approximate Gaussian maximum likelihood estimates even though we do not explicitly require that the ε_t be Gaussian.

We make the following assumption about K:

Assumption 2. $K \geq m$ is a sufficiently large known integer which remains fixed as $T \to \infty$.

The Akaike information criterion, AIC, for autoregressive model selection is a special case with $\alpha = 2$ of the criterion,

$$\mathrm{AIC}_\alpha(k) = T \ln \hat{\sigma}^2(k) + \alpha k \qquad (k = 0, 1, \ldots, K), \tag{2.3}$$

where $\alpha > 0$ is a real number; the selected order, $\hat{m}_\alpha$, say, is determined by minimizing this criterion, that is,

$$\mathrm{AIC}_\alpha(\hat{m}_\alpha) = \inf_k \mathrm{AIC}_\alpha(k). \tag{2.4}$$

A precursor of AIC is the Final Prediction Error, FPE, criterion of Akaike (1970), which, on ignoring $O(T^{-2})$ terms, may be written as a special case with $\alpha = 2$ of an extended criterion (see Bhansali and Downham (1977)),

$$\mathrm{FPE}_\alpha(k) = \hat{\sigma}^2(k)(1 + \alpha k/T), \tag{2.5}$$

and the order is once again selected by minimizing this criterion.

A third criterion we consider is due to Shibata (1980) and it may also be written as a special case, with $\alpha = 2$, of the criterion

$$S_\alpha(k) = \hat{\sigma}^2(k)(\tilde{N} + \alpha k) \tag{2.6}$$

in which $\tilde{N} = T - K$.

As discussed in Section 1, the above three criteria do not provide a consistent estimator of m with a fixed α although they are asymptotically efficient in the Shibata (1980) sense if m is infinite and certain additional conditions hold. If, however, $\alpha = \alpha(T)$, a function of T, satisfying $\alpha(T) \to \infty$, $\alpha(T)/T \to 0$, as $T \to \infty$, the selected order is consistent for m. Schwarz (1978), Akaike (1977) and Rissanen (1978) justify the choice $\alpha(T) = \log T$ from a variety of different perspectives.

Assumption 1 ensures that $\{x_t\}$ has a representation

$$x_t = \sum_{j=0}^{\infty} b(j)\varepsilon_{t-j}, \qquad b(0) = 1,$$

in which the $b(j)$ are absolutely summable and satisfy (1.2) but with the $b(j)$ replacing the $a_m(j)$ and $[A_m(z)]^{-1} = 1 + b(1)z + \cdots$.

The covariance function of $\{x_t\}$ is denoted by $R(s) = E(x_t x_{t+s})$ $(t, s = 0, \pm 1, \pm 2, \ldots)$, its spectral density function by $f(\mu)$ and the autoregressive transfer function by $A(\mu) = A\{\exp(-i\mu)\}$.

Under Assumption 1, for any integer n and all $h \geq 1$, we have

$$x_{n+h} = -\sum_{j=1}^{m} \varphi_h(j) x_{n+1-j} + z_n(h), \tag{2.7}$$

where $-\varphi_h(j)$ is the coefficient of x_{n+1-j} $(j = 1, \ldots, m)$ in the linear least-squares predictor, $\bar{x}_n(h)$, say, of x_{n+h} based on the infinite past, $\{x_t, t \leq n\}$, and

$$z_n(h) = \sum_{j=0}^{h-1} b(j) \varepsilon_{n+h-j} \tag{2.8}$$

is the h-step prediction error. The corresponding h-step mean squared error of prediction is given by

$$V(h) = E[\{z_n(h)\}^2] = \sigma^2 \sum_{j=0}^{h-1} b^2(j) \qquad (h \geq 1). \tag{2.9}$$

It is readily verified that $\varphi_1(j) = a_m(j)$, $j = 1, \ldots, m$ and for $h > 1$, the $\varphi_h(j)$ satisfy the recursive relations:

$$\varphi_h(j) = b(h-1)\varphi_1(j) + \varphi_{h-1}(j+1), \tag{2.10}$$

where $b(h) = -\varphi_h(1)$ and $\varphi_h(j) = 0$, $j > m$.

Denote the kth order (direct) linear least-squares predictor of x_{n+h} based on the finite past $\{x_n, x_{n-1}, \ldots, x_{n-k+1}\}$, $k \geq 1$, by

$$\bar{x}_{Dhk}(n) = -\sum_{j=1}^{k} \varphi_{Dhk}(j) x_{n+1-j},$$

the corresponding prediction error by $z_{Dhk}(n) = x_{n+h} - \bar{x}_{Dhk}(n)$ and let $\boldsymbol{\Phi}_{Dh}(k) = [\varphi_{Dhk}(1), \ldots, \varphi_{Dhk}(k)]'$, $\boldsymbol{r}_h(k) = [R(h), R(h+1), \ldots, R(h+k-1)]'$ and $\boldsymbol{R}(k) = [R(u-v)](u, v, = 1, \ldots, k)$. We have,

$$\boldsymbol{\Phi}_{Dh}(k) = -\boldsymbol{R}(k)^{-1} \boldsymbol{r}_h(k), \tag{2.11}$$

$$V_D(h, k) = E[\{z_{Dhk}(n)\}^2] = R(0) + \boldsymbol{r}_h(k)' \boldsymbol{\Phi}_h(k). \tag{2.12}$$

As in Shibata (1980) and Bhansali (1996), let $\hat{\boldsymbol{\Phi}}_{Dh}(k) = [\hat{\varphi}_{Dhk}(1), \ldots, \hat{\varphi}_{Dhk}(k)]'$ be the kth order direct estimate of the h-step prediction constant, $\boldsymbol{\Phi}_{Dh}(k)$,

obtained by regressing x_{t+h} on $x_t, x_{t-1}, \ldots, x_{t-k+1}$, $t = K, K+1, \ldots, T-h$, and let $\hat{V}_{Dh}(k)$ denote the residual error variance in this regression. We have,

$$\hat{\boldsymbol{\Phi}}_{Dh}(k) = -\hat{\boldsymbol{R}}_h(k)^{-1}\hat{\boldsymbol{r}}_h(k), \tag{2.13}$$

$$\hat{V}_{Dh}(k) = \hat{d}_h(0) + \hat{\boldsymbol{r}}_h(k)'\hat{\boldsymbol{\Phi}}_{Dh}(k), \tag{2.14}$$

where $\hat{\boldsymbol{R}}_h(k) = [C_{hT}(u,v)](u,v=1,\ldots,k)$, $\hat{\boldsymbol{r}}_h(k) = [C_{hT}(-h+1,1),\ldots,C_{hT}(-h+1,k)]'$, $\hat{d}_h(0) = C_{hT}(-h+1,-h+1)$, the subscript D stands for the direct method and, with $N = T-h-K+1$,

$$C_{hT}(u,v) = N^{-1}\sum_{t=K}^{T-h} x_{t-u+1}x_{t-v+1}. \tag{2.15}$$

Let $R^{(T)}(s)$ denote the positive definite estimator of $R(s)$ and $I^{(T)}(\mu)$ the periodogram function. We have,

$$\begin{aligned} I^{(T)}(\mu) &= (2\pi T)^{-1}\Big|\sum_{t=1}^{T} x_t \exp(-it\mu)\Big|^2, \\ R^{(T)}(s) &= T^{-1}\sum_{t=1}^{T-|s|} x_t x_{t+|s|} \qquad (s = 0, \pm 1, \ldots, \pm T-1) \\ &= \int_{-\pi}^{\pi} I^{(T)}(\mu)\exp(is\mu)d\mu. \end{aligned} \tag{2.16}$$

Moreover, for $u - v \geq 0$ and $h + u \geq 1$ we may write

$$R^{(T)}(u-v) - NT^{-1}C_{hT}(u,v) = T^{-1}\Big[\sum_{t=1}^{K-u} x_t x_{t+u-v} + \sum_{t=T-h-u+2}^{T-u+v} x_t x_{t+u-v}\Big] \tag{2.17}$$

which still holds for $v > u$ by transposing u and v if $h + v \geq 1$.

For $h = 1$, we write $\boldsymbol{a}(k) \equiv \boldsymbol{\Phi}_{D1}(k), \sigma^2(k) \equiv V_D(1,k)$ and now, $\hat{\boldsymbol{\Phi}}_{Dh}(k) = \hat{\boldsymbol{a}}(k)$ and $\hat{V}_{Dh}(k) = \hat{\sigma}^2(k)$.

Bhansali (1996) has developed a bound for the mean squared error of h-step prediction of the direct method when $k = \tilde{k}_{DT}(h)$, say, is a random variable possibly dependent on $x_1, \ldots, x_T$ and $K = K_T$ is a function of T such that $K_T \to \infty$, $K_T^2/T \to 0$, as $T \to \infty$, m in (1.1) is infinite, that is, $\{x_t\}$ does not degenerate to a finite autoregression and certain additional regularity conditions hold. Moreover, this bound is attainable, as $T \to \infty$, if $\tilde{k}_{DT}(h)$ is determined by minimizing, with $\alpha = 2$, any of the following h-step generalizations of the criteria (2.3), (2.5) and (2.6) above:

$$\mathrm{AIC}h_\alpha(k) = T\ln\hat{V}_{Dh}(k) + \alpha k, \tag{2.18}$$

$$\mathrm{FPE}h_\alpha(k) = \hat{V}_{Dh}(k)(1 + \alpha k/T), \tag{2.19}$$

$$Sh_\alpha(k) = \hat{V}_{Dh}(k)(N + \alpha k), \tag{2.20}$$

and, also, if the $a_m(j)$ in (1.1) decrease to 0 exponentially as $j \to \infty$, with any fixed $\alpha > 1$. Thus, in this sense, when $\{x_t\}$ does not degenerate to a finite autoregression, the order selection by the criteria (2.18) - (2.20) is asymptotically optimal for h-step prediction. In Section 5, the order selected by these criteria is shown not to be consistent if m is finite.

Consider now the plug in method. As in Yamamoto (1976), the kth order plug in estimator, $\hat{\boldsymbol{\Phi}}_{Ph}(k) = [\hat{\varphi}_{_{Phk}}(1), \ldots, \hat{\varphi}_{_{Phk}}(k)]'$ of the h-step prediction constants is, with $e_k = [1, 0, \ldots, 0]'$, $\hat{\boldsymbol{\Phi}}_{Ph}(k) = -\boldsymbol{e}_k'\hat{\boldsymbol{\Gamma}}(k)^h$, and the corresponding theoretical parameter is $\boldsymbol{\Phi}_{Ph}(k) = -\boldsymbol{e}_k'\boldsymbol{\Gamma}(k)^h$. Here, $\boldsymbol{\Gamma}(k)$ is a $k \times k$ companion matrix for $\boldsymbol{a}(k)$ and it has $-\boldsymbol{a}(k)'$ in its first row, an identity matrix of dimension $k-1$ in its bottom left hand $(k-1) \times (k-1)$ corner and a vector of zeroes of dimension $k-1$ in its last column and $\hat{\boldsymbol{\Gamma}}(k)$ is defined analogously but with $\hat{\boldsymbol{a}}(k)$ replacing $\boldsymbol{a}(k)$.

For each $h \leq 0$ and $k \geq 1$, put $\varphi_{_h}(j) = -1$, $j = -h+1$, $\varphi_{_h}(j) = 0$, $j \neq -h+1$ and $\boldsymbol{\Phi}_{Dh}(k) = [\varphi_{_h}(1), \ldots, \varphi_{_h}(k)]'$. It is readily verified that $\boldsymbol{\Gamma}(m)^h$ has $-\varphi_{_{h-u+1}}(v)$ in its (u,v)th position, $u, v = 1, \ldots, m$; moreover, as, on setting $z_n(h) = 0$, $h \leq 0$, (2.7) still holds, $\boldsymbol{\Phi}_{Dh}(k)$ also satisfies (2.11) with $\boldsymbol{r}_h(k)$ as defined there but with $h \leq 0$. Note also that the $\varphi_{_h}(j)$ satisfy the recursive relations (2.10) for all $h \leq 0$, but with $b(j) = 0$, $j < 0$ and $b(0) = 1$.

From (2.9), a kth order plug in estimate of $V(h)$ is given by

$$\hat{V}_{Ph}(k) = \hat{\sigma}^2(k) \sum_{j=0}^{h-1} \{\hat{b}_{Pk}(j)\}^2, \tag{2.21}$$

where $\hat{b}_{Pk}(0) = 1$, and if $\hat{\boldsymbol{b}}_{Pk,h-1} = [\hat{b}_{Pk}(1), \ldots, \hat{b}_{Pk}(h-1)]'$,

$$\begin{aligned} \hat{\boldsymbol{H}}_k(h-1) &= [\hat{a}_k(u-v)](u, v = 1, \ldots, h-1), \quad \hat{a}_k(j) = 0, \quad j < 0, \\ \hat{\boldsymbol{a}}_{k,h-1} &= [\hat{a}_k(1), \ldots, \hat{a}_k(h-1)]', \\ \hat{\boldsymbol{b}}_{Pk,h-1} &= -\hat{\boldsymbol{H}}_k(h-1)^{-1}\hat{\boldsymbol{a}}_{k,h-1}, \end{aligned} \tag{2.22}$$

(see Bhansali (1989, 1993)).

Under Assumption 1, $\sigma^2\boldsymbol{R}(m)^{-1}$ may be decomposed as a difference of products of lower triangular times upper triangular Toeplitz matrices as follows :

$$\sigma^2\boldsymbol{R}(m)^{-1} = \boldsymbol{H}(m)\boldsymbol{H}(m)' - \boldsymbol{L}(m)\boldsymbol{L}(m)', \tag{2.23}$$

where, with $a_m(j) = 0$, $j < 0$ or $j > m$, $\boldsymbol{H}(m) = [a_m(u-v)]$, $\boldsymbol{L}(m) = [a_m(m+v-u)]$ $(u, v = 1, \ldots, m)$ (see Galbraith and Galbraith (1974)). Also, now $\boldsymbol{\Phi}_{Dh}(m) = \boldsymbol{\Phi}_{Ph}(m) = [\varphi_{_h}(1), \ldots, \varphi_{_h}(m)]' = \boldsymbol{\Phi}_h(m)$, say.

3. Asymptotic Properties of the Direct Estimates

For each $h \geq 1$, let $R_{h-1}(u) = E\{z_n(h)z_{n+u}(h)\}$ $(n, u = 0, \pm 1, \ldots)$ denote the covariance function of $\{z_n(h)\}$, where

$$R_{h-1}(u) = \sigma^2 \sum_{j=0}^{h-1-|u|} b(j)b(j+|u|) \qquad (|u| \leq h-1), \tag{3.1}$$

$R_{h-1}(u) = 0$, $|u| \geq h$ and let $f_{h-1}(\mu)$ denote the spectral density function of $\{z_n(h)\}$, where

$$f_{h-1}(\mu) = (\sigma^2/2\pi)\Big|\sum_{j=0}^{h-1} b(j)\exp(-ij\mu)\Big|^2. \tag{3.2}$$

For all $h \leq 0$, we set $R_{h-1}(u) = 0$, all u, as now $z_n(h) = 0$.

As discussed in Section 2, for a Gaussian process, $\hat{\boldsymbol{a}}(m)$ provides a maximum likelihood estimator of $\boldsymbol{a}(m)$ and this estimator is asymptotically efficient in the sense that its asymptotic covariance matrix attains the Cramer-Rao lower bound applicable to this situation. Furthermore, if $u\{\boldsymbol{a}(m)\}$ denotes a differentiable function of $\boldsymbol{a}(m)$ then the corresponding estimator, $u\{\hat{\boldsymbol{a}}(m)\}$, based on $\hat{\boldsymbol{a}}(m)$ of this quantity is also asymptotically efficient in this sense. In the sequel, we use the term asymptotically Gaussian efficient for the plug in estimates of $\boldsymbol{\Phi}(m)$ and $V(h)$ and related parameters in accordance with this definition.

The asymptotic distribution of $\hat{\boldsymbol{\Phi}}_{Dh}(k)$ when $\{x_t\}$ satisfies Assumption 1 and $k = m$ is given in the following theorem:

Theorem 3.1. *Let $\{x_t\}$ satisfy Assumption 1. Then, as $T \to \infty$, $T^{1/2}\{\hat{\boldsymbol{\Phi}}_{Dh}(m) - \boldsymbol{\Phi}_{Dh}(m)\}$ is asymptotically distributed as Normal with a $\mathbf{0}$ mean vector and covariance matrix $\boldsymbol{U}(m) = \boldsymbol{R}(m)^{-1}\boldsymbol{W}(m)\boldsymbol{R}(m)^{-1}$, where $\boldsymbol{W}(m) = [w(u,v)]$ $(u, v = 1, \ldots, m)$ and*

$$\begin{aligned} w(u,v) &= \sum_{s=-h+1}^{h-1} R_{h-1}(s)R(u-v-s) \\ &= 2\pi \int_{-\pi}^{\pi} f_{h-1}(\mu)f(\mu)\exp\{i(u-v)\mu\}d\mu. \end{aligned} \tag{3.3}$$

Remark. The term involving the fourth order cumulant, τ_4, of $\{\varepsilon_t\}$ disappears algebraically from (3.3).

Proof. We may write

$$\sqrt{T}\{\hat{\boldsymbol{\Phi}}_{Dh}(m) - \boldsymbol{\Phi}_h(m)\} = -\{\hat{\boldsymbol{R}}_h(m)^{-1}\boldsymbol{p}_h(m)\}\sqrt{T}, \tag{3.4}$$

where

$$\boldsymbol{p}_h(m) = \hat{\boldsymbol{r}}_h(m) + \hat{\boldsymbol{R}}_h(m)\boldsymbol{\Phi}_h(m) \tag{3.5}$$

and its typical element is given by,

$$p_{hu}(m) = \sqrt{T}\Big\{C_{hT}(-h+1,u) + \sum_{j=1}^{m} \varphi_h(j) C_{hT}(u,j)\Big\} \tag{3.6}$$

with $u = 1, \ldots, m$. As $T \to \infty$, each element of $\hat{\boldsymbol{R}}_h(m)^{-1}$ converges in probability to that of $\boldsymbol{R}(m)^{-1}$. The theorem now follows from the results of Anderson (1971), Whittle (1963), p. 32 (see also Bhansali (1981)), by demonstrating that $\boldsymbol{W}(m)$ is the covariance matrix of $\boldsymbol{p}(m)$.

We note that our Theorem 3.1 accords with the earlier results of Hosoya and Taniguchi (1982) and Kabaila (1981) on the asymptotic distribution of the $\hat{\varphi}_{Dh}(j)$; these authors do not, however, present the asymptotic covariance matrix of the direct estimates in the explicit form given in (3.3).

Thus, in view of the representation (2.7), the actual spectral density function of $\{x_t\}$ may be written as

$$f_{true}(\mu) = f_{h-1}(\mu) \Big/ \Big|1 + \sum_{j=1}^{m} \varphi_h(j) \exp\{-i(h+j-1)\mu\}\Big|^2. \tag{3.7}$$

On the other hand, as the direct method of estimating the $\varphi_h(j)$ is equivalent to fitting a (scale-free) spectral density of the form given in Remark 3.2 of Hosoya and Taniguchi (1982), it follows from this remark that our Theorem 3.1 agrees with their Theorem 3.2.

Kabaila (1981) earlier established that the estimates of the parameters of a non-linear autoregression obtained by minimising the sum of squares of h-step prediction errors are, $T \to \infty$, asymptotically normal with a 0 mean vector, and gave an expression for evaluating the asymptotic covariance matrix of the estimates. For the direct method considered here, result (1.6) of this author coincides exactly with our Theorem 3.1. Thus, treating $\boldsymbol{\Phi}_h(m)$ as the parameter of interest, this author has proved that, as $T \to \infty$, $T^{1/2}\{\hat{\boldsymbol{\Phi}}_{Dh}(m) - \boldsymbol{\Phi}_h(m)\}$ is asymptotically normal with a 0 mean vector and covariance matrix $\boldsymbol{C}^{-1}\boldsymbol{Z}\boldsymbol{C}^{-1}$, say, where $\boldsymbol{C} = [C(u,v)]$, $\boldsymbol{Z} = [Z(u,v)]$ $(u,v = 1, \ldots, m)$ and from (2.7) and (2.8)

$$\begin{aligned} C(u,v) &= 2E([\{\partial/\partial\varphi_h(u)\}z_n(h)][\{\partial/\partial\varphi_h(v)\}z_n(h)]) = 2R(u-v), \\ Z(u,v) &= 4\sum_{j=-h+1}^{h-1} E\{z_n(h)x_{n+1-v}z_{n+j}(h)x_{n+j+1-u}\} \\ &= 4\sum_{j=-h+1}^{h-1} R_{h-1}(j)R(u-v-j); \end{aligned}$$

and this result is readily seen to agree with our Theorem 3.1.

The asymptotic distribution of $\hat{V}_{Dh}(m)$, $h = 1, \ldots, J$, with $J \geq 1$ denoting the maximum prediction lead time, is given below:

Theorem 3.2. *Let $\{x_t\}$ satisfy Assumption 1. Then, as $T \to \infty$,*

$$\sqrt{T}\{\hat{V}_{D1}(m) - V(1)\}, \ldots, \sqrt{T}\{\hat{V}_{DJ}(m) - V(J)\}$$

are jointly asymptotically normal with a $\mathbf{0}$ *mean vector and covariance matrix,* $\boldsymbol{Q}$*, whose typical element is given by*

$$q(h,n) = 2 \sum_{s=-\min(n,h)+1}^{\min(n,h)-1} R_{h-1}(s) R_{n-1}(s) + (\tau_4/\sigma^4) R_{h-1}(0) R_{n-1}(0) \ (h, n = 1, \ldots, J),$$

where $R_{n-1}(u)$ is given by (3.1) *but with h replaced by n.*

Proof. For a fixed $h \geq 1$, let,

$$g_1(h) = \left\{\hat{d}_h(0) - R(0)\right\} + 2\sum_{j=1}^{m} \varphi_h(j) \sqrt{T}\{C_{hT}(-h+1, j) - R(h-1+j)\Big\}$$
$$+ \sum_{u=1}^{m} \sum_{j=1}^{m} \varphi_h(u) \varphi_h(j) \sqrt{T}\{C_{hT}(u,j) - R(u-j)\}. \tag{3.8}$$

From (2.7) and (2.8), on using (3.4), the difference between $T^{1/2}\{\hat{V}_{Dh}(m) - V(h)\}$ and $g_1(h)$ tends to 0 in probability, as $T \to \infty$, for each fixed $h \geq 1$. Also, by (2.17), with $K = m$, the difference between $T^{1/2}\{C_{hT}(u,j) - R^{(T)}(u-j)\}$ converges to 0 in probability for each fixed (u, j) such that $h+u \geq 1$ and $h+j \geq 1$. Hence, replacing $C_{hT}(u,j)$ by $R^{(T)}(u-j)$ on the right of (3.8) and using (2.16) and (3.7), the difference between $g_1(h)$ and $g_2(h)$ also tends to 0 in probability as $T \to \infty$, where

$$g_2(h) = \int_{-\pi}^{\pi} f_{h-1}(\mu)\{f(\mu)\}^{-1} \sqrt{T}\{I^{(T)}(\mu) - f(\mu)\} d\mu. \tag{3.9}$$

The asymptotic normality of the $\hat{V}_{Dh}(m)$ now follows from (3.8) and the asymptotic covariance structure from (3.9) (see Anderson (1971), pp. 463-467 and Brillinger (1975) pp. 254-255).

Note that the asymptotic covariance matrix of the $\hat{V}_{Dh}(m)$ derived above in Theorem 3.2 under the assumption of a finite and known m coincides exactly with that derived by Bhansali (1993), who earlier obtained the joint asymptotic distribution of a finite collection of $T^{1/2}\{\hat{V}_{Dh}(k) - V(h)\}$ on the assumption that k converges to infinity simultaneously but sufficiently slowly with T. Hence, Theorem 3.2 has a similar interpretation to that given there, and, even for a

finite m, the asymptotic distribution of $\hat{V}_{Dh}(m)$ and that of a 'nonparametric' estimator, $R_{h-1}^{(T)}(0)$, given by (2.16) but with the z's replacing the x's, of $V(h)$ are the same.

Put $\boldsymbol{\delta}(h-1) = [R_{h-1}(1), \ldots, R_{h-1}(h-1)]'$, $\boldsymbol{G}(m) = \boldsymbol{H}(m)^{-1} = [b(u-v)]$ $(u, v = 1, \ldots, m)$, and let $\boldsymbol{E}_{h-1}(m) = [\boldsymbol{I}_{h-1}, \boldsymbol{0}_{h-1,m-h+1}]$ and $\boldsymbol{E}_m^*(h-1) = [\boldsymbol{I}_m, \boldsymbol{0}_{m,h-1-m}]'$, in which $\boldsymbol{0}_{i,j}$ denotes an $i \times j$ matrix of 0's, be two auxiliary matrices, each of dimension $(h-1) \times m$.

The asymptotic distributions of the corresponding plug in estimators, $\hat{\boldsymbol{\Phi}}_{Ph}(m)$ and $\hat{V}_{Ph}(m)$ is given in the following theorem:

Theorem 3.3. *Let $\{x_t\}$ satisfy Assumption 1. Then, as $T \to \infty$,* (a) $T^{1/2}\{\hat{\boldsymbol{\Phi}}_{Ph}(m) - \boldsymbol{\Phi}_h(m)\}$ *is asymptotically distributed as Normal with a $\boldsymbol{0}$ mean vector and covariance matrix $\boldsymbol{F}(m)$, where*

$$\boldsymbol{F}(m) = \sigma^2\Big[\sum_{j=0}^{h-1} b(j)\{\boldsymbol{\Gamma}(m)'\}^{h-1-j}\Big][\boldsymbol{R}(m)^{-1}]\Big[\sum_{k=0}^{h-1} b(k)\{\boldsymbol{\Gamma}(m)\}^{h-1-k}\Big]. \quad (3.10)$$

(b) $T^{1/2}\{\hat{V}_{P1}(m) - V(1)\}, \ldots, T^{1/2}\{\hat{V}_{PJ}(m) - V(J)\}$ *are jointly asymptotically normal with a 0 mean vector and covariance matrix $\boldsymbol{\Omega} = [\Omega(h,n)]$ $(h, n = 1, \ldots, J)$, where*

$$\Omega(h,n) = \boldsymbol{q}(h,n) - \boldsymbol{d}(h,n), \qquad 1 \le h, n \le m+1 \quad (3.11a)$$

$$= q(h,n) - d(h,n) = \Omega(n,h), \quad n \ge m+2,\ h = 1, \ldots, J, \quad (3.11b)$$

and where

$$\boldsymbol{d}(h,n) = 4\boldsymbol{\delta}(h-1)'\boldsymbol{G}(h-1)\boldsymbol{E}_{h-1}(m)\boldsymbol{L}(m)\boldsymbol{L}(m)'E_{n-1}(m)'\boldsymbol{G}(n-1)\boldsymbol{\delta}(n-1),$$

$$q(h,n) = 4\boldsymbol{\delta}(h-1)'\boldsymbol{G}(h-1)'\boldsymbol{E}_m^*(h-1)\boldsymbol{H}(m)\boldsymbol{H}(m)'\boldsymbol{E}_m^*(n-1)'\boldsymbol{G}(n-1)'\boldsymbol{\delta}(n-1)$$
$$+2R_{h-1}(0)R_{n-1}(0) + (\tau_4/\sigma^4)R_{h-1}(0)R_{n-1}(0),$$

and $\tilde{d}(h,n)$ is obtained from $d(h,n)$ by replacing $\boldsymbol{E}_{h-1}(m)$ and $\boldsymbol{E}_{n-1}(m)$ by $\boldsymbol{E}_m^*(h-1)$ and $\boldsymbol{E}_m^*(n-1)$, respectively.

Proof. (a) Since $\hat{\boldsymbol{\Gamma}}(m)' - \boldsymbol{\Gamma}(m)' = \{\hat{\boldsymbol{\Gamma}}(m)' - \boldsymbol{\Gamma}(m)'\}\boldsymbol{e}_m\boldsymbol{e}_m'$, the result follows by noting that $\hat{b}_m(j) = \boldsymbol{e}_m'\hat{\boldsymbol{\Gamma}}(m)^j\boldsymbol{e}_m$ converges in probability to $b(j)$ for each $j \ge 1$, as $T \to \infty$, and

$$\sqrt{T}\{\hat{\boldsymbol{\Phi}}_{Ph}(m) - \boldsymbol{\Phi}_h(m)\}$$
$$= \sqrt{T}\Big[\sum_{j=0}^{h-1}\{\boldsymbol{\Gamma}(m)'\}^{h-1-j}\{\hat{\boldsymbol{\Gamma}}(m)' - \boldsymbol{\Gamma}(m)'\}\{\hat{\boldsymbol{\Gamma}}(m)'\}^j\Big]\boldsymbol{e}_m$$
$$= \sum_{j=0}^{h-1}\{\boldsymbol{\Gamma}(m)'\}^{h-1-j}\sqrt{T}\{\hat{\boldsymbol{a}}(m) - \boldsymbol{a}(m)\}\hat{b}_m(j).$$

(b) Let $\boldsymbol{b}_{m,h-1} = [b(1), \ldots, b(h-1)]'$. For each fixed $h \geq 1$, we may write

$$\begin{aligned}&\sqrt{T}\{\hat{V}_{Ph}(m) - V(h)\}\\ &= \sigma^{-2}V(h)\sqrt{T}\{\hat{\sigma}^2(m) - \sigma^2\} + 2\sigma^2\boldsymbol{b}'_{m,h-1}\sqrt{T}\{\hat{\boldsymbol{b}}_{Pm,h-1} - \boldsymbol{b}_{m,h-1}\} + o_p(1), \quad (3.12)\end{aligned}$$

where $o_p(1)$ denotes a term tending to 0 in probability as $T \to \infty$. Now, if $h \leq m+1$, $\boldsymbol{b}_{m,h-1} = -\boldsymbol{H}(h-1)^{-1}\boldsymbol{E}_{h-1}(m)\boldsymbol{a}(m)$, then from (2.22), we may write (see also Bhansali (1989)).

$$\sqrt{T}\{\hat{\boldsymbol{b}}_{Pm,h-1} - \boldsymbol{b}_{m,h-1}\} = -\boldsymbol{G}(h-1)^2\boldsymbol{E}_{h-1}(m)\sqrt{T}\{\hat{\boldsymbol{a}}(m) - \boldsymbol{a}(m)\} + o_p(1), \quad (3.13)$$

and a similar result holds also for $h > m+1$, provided on the right hand side of (3.13) $\boldsymbol{E}_{h-1}(m)$ is replaced by $\boldsymbol{E}^*_m(h-1)$. The result (b) now follows from the representation (2.23) for $\sigma^2\boldsymbol{R}(m)^{-1}$ by observing that $\boldsymbol{\delta}(h-1)' = \sigma^2\boldsymbol{b}'_{m,h-1}\boldsymbol{G}(h-1)$ and by noting that, as $T \to \infty$, the two terms occurring on the right of (3.12) are asymptotically independent.

Yamamoto (1976) and Stine (1987) also derive the asymptotic distributions of $T^{1/2}\{\hat{\boldsymbol{\Phi}}_{Ph}(m) - \boldsymbol{\Phi}_{Ph}(m)\}$ and $T^{1/2}\{\hat{V}_{Ph}(m) - V(h)\}$, respectively; these authors, however, do not present the asymptotic covariance structure of these quantities in the explicit form given in Theorem 3.3.

Bhansali (1993) shows that if as $T \to \infty$, $k \to \infty$ and additional regularity conditions hold, $\hat{V}_{Dh}(k)$ and $\hat{V}_{Ph}(k)$ have the same asymptotic distributions. This result may be gleaned from Theorem 3.3 as an iterated limit since, as $m \to \infty$, $d(h,n) \to 0$ for each fixed (h,n). For a finite and known m, however, as discussed earlier, unlike $\hat{V}_{Dh}(m), \hat{V}_{Ph}(m)$ provides a Gaussian maximum likelihood estimator of $V(h)$ and in this situation it is asymptotically Gaussian efficient; for $h \leq m+1$, $d(h,n)$ now provides an asymptotic measure of the extent to which $\hat{V}_{Dh}(m)$ is asymptotically inefficient, relative to $\hat{V}_{Ph}(m)$.

Note that although $\{z_n(h)\}$ is a moving average process of order $h-1$, the problem of estimating its variance is different from that of constructing an asymptotically efficient estimator (see Anderson(1975)) of the variance of a moving average process based only on a partial realization of this process; thus, under the hypothesis of Assumption 1, $\hat{V}_{Ph}(m)$ provides a Gaussian maximum likelihood estimator of $V(h) = \mathrm{Var}\,(z_t)$ only because the $b(j)$ are functionally related to the $a_m(j)$.

Next, we compare the asymptotic covariance matrices of the direct and plug in estimates of the h-step prediction constants. We have the following proposition:

Proposition 3.1. *Let $\{x_t\}$ satisfy Assumption* 1. *Then the following results hold:*

(a) *The asymptotic covariance matrix,* $\boldsymbol{U}(m)$*, of* $T^{1/2}\{\hat{\boldsymbol{\Phi}}_{Dh}(m) - \boldsymbol{\Phi}_h(m)\}$ *may also be written as*

$$\boldsymbol{U}(m) = R_{h-1}(0)\boldsymbol{R}(m)^{-1} + \sum_{s=1}^{h-1} R_{h-1}(s)[\boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m)^s + \boldsymbol{\Gamma}(m)'^s\boldsymbol{R}(m)^{-1}]. \tag{3.14}$$

(b) *If, for each* $h > 1$*,*

$$[\partial\boldsymbol{\Phi}(h)/\partial\boldsymbol{\Phi}(1)] = [\partial\varphi_{_h}(j)/\partial\varphi_{_1}(k)] \qquad (j,k = 1,\ldots,m)$$

denotes the matrix of partial derivatives of the $\varphi_{_h}(j)$ *with respect to the* $\varphi_{_1}(k)$*, then*

$$[\partial\boldsymbol{\Phi}(h)/\partial\boldsymbol{\Phi}(1)] = \sum_{j=0}^{h-1} b(j)\{\boldsymbol{\Gamma}(m)'\}^{h-1-j}$$

and the estimator $\hat{\boldsymbol{\Phi}}_{Ph}(m)$ *of* $\boldsymbol{\Phi}_h(m)$ *is asymptotically Gaussian efficient.*

(c) *For each* $h > 1$*, we may write*

$$\begin{aligned}\sigma^{-2}[\boldsymbol{U}(m) - \boldsymbol{F}(m)] = &\sum_{s=1}^{h-2}\sum_{j=0}^{h-2-s} b(j)b(j+s)[\boldsymbol{D}(s,j+s,j) + \boldsymbol{D}(s,j+s,j)'] \\ &+ \sum_{j=0}^{h-2} b^2(j)\boldsymbol{D}(0,j,j),\end{aligned} \tag{3.15}$$

where $\boldsymbol{D}(i,j,k) = \boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m)^i - \boldsymbol{\Gamma}(m)'^{h-1-j}\boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m)^{h-1-k}$.

Proof. (a) Since, for each integer s, $\boldsymbol{\Phi}_s(m) = -\boldsymbol{R}(m)^{-1}\boldsymbol{r}_s(m)$, the result follows by writing

$$\boldsymbol{U}(m) = \boldsymbol{U}_0(m) + \boldsymbol{U}_1(m) + \boldsymbol{U}_2(m), \quad \text{say,}$$

where $\boldsymbol{U}_0(m) = R_{h-1}(0)\boldsymbol{R}(m)^{-1}$,

$$\begin{aligned}\boldsymbol{U}_1(m) &= \boldsymbol{R}(m)^{-1}\Big\{\sum_{s=1}^{h-1} R_{h-1}(s)[\boldsymbol{r}_s(m)\boldsymbol{r}_{s-1}(m)\cdots\boldsymbol{r}_{s-m+1}(m)]\boldsymbol{R}(m)^{-1}\Big\} \\ &= \Big\{\sum_{s=1}^{h-1} R_{h-1}(s)\boldsymbol{\Gamma}(m)'^s\Big\}\boldsymbol{R}(m)^{-1}, \\ \boldsymbol{U}_2(m) &= \boldsymbol{R}(m)^{-1}\Big\{\sum_{s=1}^{h-1} R_{h-1}(s)[\boldsymbol{r}_s(m)\boldsymbol{r}_{s-1}(m)\cdots\boldsymbol{r}_{s-m+1}(m)]'\boldsymbol{R}(m)^{-1}\Big\} \\ &= \sum_{s=1}^{h-1} R_{h-1}(s)\boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m)^s.\end{aligned}$$

It should be observed that the expression inside the square brackets on the right of $\boldsymbol{U}_j(m)$ $(j = 1, 2)$ is in fact a matrix.
(b) The proof is by induction. The stated result clearly holds for $h = 1$ and it is readily verified for $h = 2$. Now assume that the result holds for some integer $h = n$. To show that it also holds for $h = n + 1$, use (2.10) and deduce that for $1 \le j \le m$,

$$[\partial\varphi_{n+1}(j)/\partial\Phi_1(m)] = -\sum_{u=0}^{n} b(u)[\varphi_{n-u}(j)\varphi_{n-u-1}(j)\cdots\varphi_{n-u-m+1}(j)].$$

(c) The result follows readily from (3.10) and (a) above by using (3.1).

Proposition 3.1 (b) implies that, under Assumption 1, $\hat{\boldsymbol{\Phi}}_{Ph}(m)$ provides a Gaussian asymptotically efficient estimator of $\boldsymbol{\Phi}_h(m)$ and it asymptotically attains the Cramer-Rao lower bound for the variance of an unbiased estimator of this parameter. By contrast, the direct estimator, $\hat{\boldsymbol{\Phi}}_{Dh}(m)$ is not even asymptotically Gaussian efficient, and it follows by standard theory that the difference, $\boldsymbol{U}(m) - \boldsymbol{F}(m)$, between their asymptotic covariance matrices is nonnegative definite. Proposition 3.1 (c) provides an explicit expression for evaluating this difference and shows that its 'size' depends upon the magnitudes of the $b(j)$ and on the 'size' of $\boldsymbol{D}(s, j + s, j)$, $0 \le j$, $s \le h - 2$.

For two finite-dimensional matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, let $\|\boldsymbol{A}\|$, $\|\boldsymbol{B}\|$ denote their respective matrix norms. On using the inequalities, $\|\boldsymbol{A} - \boldsymbol{B}\| \ge \|\boldsymbol{A}\| - \|\boldsymbol{B}\|$, $\|\boldsymbol{AB}\| \le \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$ and noting that $\|\boldsymbol{\Gamma}(m)\| < 1$, it readily follows that for all $0 \le j, s \le h - 2$

$$\begin{aligned}\|\boldsymbol{D}(s, j + s, j)\| &\ge \|\boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m)^s\| - \|\boldsymbol{\Gamma}(m)'^{h-1-j-s}\boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m)^s\boldsymbol{\Gamma}(m)^{h-1-j-s}\| \\ &\ge \{1 - \|\boldsymbol{\Gamma}(m)\|\,\|\boldsymbol{\Gamma}(m)'\|\}\|\boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m)^s\| > 0.\end{aligned}$$

Unfortunately, (3.15) is a complex expression involving a double sum of the different $\boldsymbol{D}(s, j + s, j + s)$'s, and a useful bound valid for all values of m and h for the difference between $\boldsymbol{U}(m)$ and $\boldsymbol{F}(m)$ is not readily given; instead, the efficiency loss in using the direct estimates is illustrated in Section 4 for some specific values of m and/or h.

Next, consider the h-step mean squared error of prediction. As in Akaike (1970), among others, we assume that the process to be predicted, $\{y_t\}$, say, is independent of and has the same stochastic structure as $\{x_t\}$, and for some integer n and $J \ge 1$, the past $\{y_{n+1-j}, j = 1, \ldots, m\}$ of $\{y_t\}$ is known and the future $\{y_{n+h}, h = 1, \ldots, J\}$ is to be predicted by fitting an autoregression of order m by either the direct or the plug-in method from a realization of T consecutive

observations of $\{x_t\}$. Let $\hat{y}_{Dn}(h)$ and $\hat{y}_{Pn}(h)$ denote the respective h-step direct and plug in forecasts of y_{n+h}. We have

$$\hat{y}_{Dn}(h) = -\sum_{j=1}^{m} \hat{\varphi}_{Dh}(j) y_{n+1-j} \tag{3.16}$$

and $\hat{y}_{Pn}(h)$ is defined analogously, but with the $\hat{\varphi}_{Ph}(j)$ replacing the $\hat{\varphi}_{Dh}(j)$ in (3.16).

If $o_p(T^{-1/2})$ terms are ignored then as in Yamamoto (1976) the h-step mean squared error of prediction, $E[\{\hat{y}_{Pn}(h) - y_{n+h}\}^2]$, of the plug-in method may be approximated by $PMSE_P(h)$, where

$$PMSE_P(h) = V(h) + M_P(h) \qquad (h = 1, \ldots, J), \tag{3.17}$$

$$M_P(h) = T^{-1}\sigma^2 \sum_{j=0}^{h-1}\sum_{k=0}^{h-1} b(j)b(k)\,\mathrm{tr}[\boldsymbol{\Gamma}(m)'^{h-1-j}\boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m)^{h-1-k}\boldsymbol{R}(m)]. \tag{3.18}$$

By a similar argument, it follows from Theorem 3.1 that if $o_p(T^{-1/2})$ terms are ignored, the h-step mean squared error of the direct method may be approximated by $PMSE_D(h)$, where

$$PMSE_D(h) = V(h) + M_D(h), \tag{3.19}$$

$$M_D(h) = T^{-1}\,\mathrm{tr}\{\boldsymbol{R}(m)^{-1}\boldsymbol{W}(m)\}. \tag{3.20}$$

Bhansali (1981) has shown that (3.17) remains valid even when $o(T^{-1})$ terms are ignored; however, here we do not attempt to extend the results (3.19)-(3.20) for the direct method in this direction. Rather we examine the loss in predictive efficiency by using the direct method for a finite m and give an explicit expression for the difference, $M_D(h) - M_P(h)$, between $PMSE_D(h)$ and $PMSE_P(h)$. We first establish the following lemma:

Lemma 3.1. *Suppose that $\{x_t\}$ satisfies Assumption 1. Then for each $h \geq 1$ and $0 \leq j, k \leq h-1$,*

$$\boldsymbol{\Gamma}(m)^{h-1-k}\boldsymbol{R}(m)\boldsymbol{\Gamma}(m)'^{h-1-j} = \boldsymbol{\Theta}_1(j,k) - \boldsymbol{\Theta}_2(j,k), \tag{3.21}$$

where $\boldsymbol{\Theta}_1(j,k) = [R(v-u+j-k)]$ $(u,v = 1,\ldots,m)$ and the term, $[\Theta_2(j,k)]_{u,v}$, in the (u,v)th position of $\boldsymbol{\Theta}_2(j,k)$ is given by

$$[\Theta_2(j,k)]_{u,v} = \begin{cases} R_{h-1-j-v}(k+u-v-j), & \text{if } v-u \leq k-j, \\ R_{h-1-k-u}(j+v-k-u), & \text{if } v-u \geq k-j, \end{cases} \tag{3.22}$$

where, if $s \geq 0$, $R_s(u)$ is defined by (3.1), but with s replacing $h-1$, and, if $s < 0$, $R_s(u) \equiv 0$, all u.

Proof. As $\boldsymbol{\Gamma}(m)^{h-1-k} = [\varphi_{h-1-k-u+1}(v)]$ $(u, v = 1, \ldots, m)$, it follows from (2.11) that $\boldsymbol{\Gamma}(m)^{h-1-k}\boldsymbol{R}(m) = [R(h-k-u+v-1)]$ $(u, v = 1, \ldots, m)$. The proof is completed by demonstrating that $[\boldsymbol{\Gamma}(m)^{h-1-k}\boldsymbol{R}(m)\boldsymbol{\Gamma}(m)'^{h-1-j}]_{u,v}$, the term in the (u, v)th position of $\boldsymbol{\Gamma}(m)^{h-1-k}\boldsymbol{R}(m)\boldsymbol{\Gamma}(m)'^{h-1-j}$, is given by

$$[\boldsymbol{\Gamma}(m)^{h-1-k}\boldsymbol{R}(m)\boldsymbol{\Gamma}(m)'^{h-1-j}]_{u,v} = R(j+v-k-u) - E\{z_n(h-j-v)z_n(h-k-u)\},$$

where $z_n(h)$ is defined by (2.8), if $h > 0$, and $z_n(h) \equiv 0$, $h \leq 0$, and that $\boldsymbol{\Theta}_2(j,k) = [E\{z_n(h-j-v)z_n(h-k-u)\}]$ $(u, v = 1, \ldots, m)$.

Proposition 3.2. *Let $\{x_t\}$ satisfy Assumption 1. Then, for each $h \geq 1$,*

$$\text{(a)} \quad TM_D(h) = mR_{h-1}(0) - 2\sum_{s=1}^{h-1}\sum_{j=1}^{\min(s,m)} R_{h-1}(s)\varphi_{s-j+1}(j), \tag{3.23}$$

$$\text{(b)} \quad TM_P(h) = T\{M_D(h) - M_\Theta(h)\},$$

where

$$M_\Theta(h) = \sigma^2 \sum_{j=0}^{h-2}\sum_{k=0}^{h-2} b(j)b(k)\,\mathrm{tr}\{\boldsymbol{R}(m)^{-1}\boldsymbol{\Theta}_2(j,k)\}. \tag{3.24}$$

Proof. (a) The result follows directly from Proposition 3.1 (a) on using (2.11) and by applying standard rules for evaluating the trace of a matrix.
(b) The result follows directly from definition (3.1) of $R_{h-1}(u)$ and Lemma 3.1.

Proposition 3.2 provides an explicit asymptotic expression for evaluating the increase in the h-step mean squared error of prediction, $M_D(h)$, due to estimating the prediction constants by the direct method. This proposition also gives an explicit expression for evaluating asymptotically the difference between the h-step mean squared errors of prediction of the direct and plug in methods, which is known to be positive by Proposition 3.1. It may be deduced from (3.24) that this difference depends upon the magnitudes of the $b(j)$ and those of $\mathrm{tr}\{\boldsymbol{R}(m)^{-1}\boldsymbol{\Theta}_2(j,k)\}$, $j, k = 0, \ldots, h-2$. As $\boldsymbol{\Theta}_2(j,k)$ defines the cross-covariance matrix of $z_n(h-j-v)$ and $z_n(h-k-u)$, $u, v = 1, \ldots, m$, we may in general expect the 'size' of $\mathrm{tr}\{\boldsymbol{R}(m)^{-1}\boldsymbol{\Theta}_2(j,k)\}$ to be dependent upon the extent to which the observed series is predictable from its own past. Thus, for example, if the observed series is not highly predictable from the past then $\boldsymbol{\Theta}_2(j,j)$ would be close to $\boldsymbol{R}(m)$ and the value of $\mathrm{tr}\{\boldsymbol{R}(m)^{-1}\boldsymbol{\Theta}_2(j,j)\}$ 'large'; the converse may be expected to hold for highly predictable series since now $\boldsymbol{\Theta}_2(j,j)$ would be closer to 0 and so would the value of this trace. In Section 5, we illustrate the behaviour of (3.24) for several specific values of m and/or h.

Lewis and Reinsel (1985) and Bhansali (1993) show that if an autoregressive model of order k is fitted, where $k \to \infty$ as $T \to \infty$ and additional regularity

conditions hold, $(T/k)M_P(h)$ and $(T/k)M_D(h)$ both converge to $R_{h-1}(0)$ and the difference between the asymptotic mean squared errors of prediction of the plug in and direct methods now vanishes. This result follows from (3.23) and (3.24) as an iterated limit since, when divided by m, the first term to the right of (3.23) converges to $R_{h-1}(0)$ and the second term to 0, as the number of terms occurring in this term remains fixed as m increases. It follows analogously that since $[\boldsymbol{\Theta}_2(j,k)]_{u,v} \equiv 0$, $0 \le j,k \le h-1$ and $u > h-1$ or $v > h-1$, $M_{\boldsymbol{\Theta}}(h)$ divided by m, converges to 0 as $m \to \infty$.

It should be observed that although the results of this section have been derived by assuming that the fitted order, k, equals the true order, m, of the process (1.1), the results continue to hold even when $k \ge m$ (see Bhansali (1981)).

4. Examples

We illustrate the results of Section 3 by evaluating (3.3), (3.10), (3.15), (3.23) and (3.24) explicitly for $m = 1, 2$ and $h = 2, 3$; the case of a general m with $h = 2$ and a general h with $m = 1$ is also discussed.

Thus, suppose that $m = 1$ and $x_t = ax_{t-1} + \varepsilon_t$, $|a| < 1$, where $\{\varepsilon_t\}$ is as in Assumption 1. Now, $\varphi_{h}(1) = -a^h$ and it follows from Theorems 3.1 and 3.3 that as $T \to \infty$, $T^{1/2}\{\hat{\varphi}_{Dh}(1) - \varphi_{h}(1)\}$ and $T^{1/2}\{\hat{\varphi}_{Ph}(1) - \varphi_{h}(1)\}$ are asymptotically normal with 0 means and variances $v_{Dh}(1)$ and $v_{Ph}(1)$, where

$$v_{Dh}(1) = (1-a^2)^{-1}[1 + a^2 - (2h+1)a^{2h} + (2h-1)a^{2h+2}],$$
$$v_{Ph}(1) = h^2 a^{2h-2}(1-a^2).$$

If, in particular, $h = 2$, $v_{D2}(1) = 1 + 2a^2 - 3a^4$ and $v_{P2}(1) = 4a^2(1-a^2)$; also, if $h = 3$, $v_{D3}(1) = (1-a^2)(1+3a^2+5a^4)$ and $v_{P3}(1) = 9a^4(1-a^2)$. On adopting $e_{PDh}(1) = v_{Ph}(1)/v_{Dh}(1)$ as a measure of the asymptotic relative efficiency of the plug-in method with respect to the direct method, we have, $e_{PD2}(1) = 4a^2(1+3a^2)^{-1}$, $e_{PD3}(1) = 9a^4(1+3a^2+5a^4)^{-1}$, depend only on the absolute value, $|a|$, of a, and, for $h = 2$ and 3, $e_{PDh}(1) \to 0$ as $|a| \to 0$ and $e_{PDh}(1) \to 1$ as $|a| \to 1$. The actual numerical values of $e_{PDh}(1)$ for some specific values of $|a|$ are shown below:

$\lvert a\rvert$	:	0.1	0.5	0.9
$e_{PD2}(1)$	:	0.04	0.57	0.94
$e_{PD3}(1)$	:	0.001	0.27	0.88

The loss in asymptotic efficiency by using the direct method is seen to be greater for $h = 3$ than for $h = 2$, moreover, this loss could be substantial for small values of $|a|$.

Consider now the loss in predictive efficiency by using the direct method. If $m = 1$, we get from (3.23) - (3.24), with $h \geq 1$,

$$\begin{aligned} T\sigma^{-2}M_D(h) &= (1-a^2)^{-1}[1+2a^2(1-a^{2h-2})(1-a^2)^{-1} - (2h-1)a^{2h}], \\ T\sigma^{-2}M_\Theta(h) &= T\sigma^{-2}\{M_D(h) - M_P(h)\} \\ &= (1-a^2)^{-1}[1-h^2a^{2h-2}+(h-1)^2a^{2h}+2a^2(1-a^{2h-2})(1-a^2)^{-1}], \quad (4.1) \\ T\sigma^{-2}M_P(h) &= h^2a^{2h-2}. \end{aligned}$$

If now $h = 2$, $T\sigma^{-2}M_D(h) = 1 + 3a^2$, $T\sigma^{-2}M_P(h) = 4a^2$, and $T\sigma^{-2}\{M_D(h) - M_P(h)\} = 1-a^2$; also, if $h = 3$, $T\sigma^{-2}M_D(h) = 1+3a^2+5a^4, T\sigma^{-2}M_P(h) = 9a^4$, and $T\sigma^{-2}\{M_D(h) - M_P(h)\} = (1 - a^2)(1 + 4a^2)$; thus, $\{M_D(h) - M_P(h)\}$ is a monotonic decreasing function of a^2 for $h = 2$; for $h = 3$, however, it is a quadratic function of a^2 and attains its maximum when $a^2 = 0.75$.

Next suppose that $m = 2$ and $x_t = a_1x_{t-1}+a_2x_{t-2}+\varepsilon_t$, where $1-a_1z-a_2z^2 \neq 0$, $|z| \leq 1$. We consider only $h = 2, 3$ as the case of a general h is awkward to illustrate.

We have, from (3.15), if $h = 2$ and $m = 2$,

$$\sigma^{-2}\{\boldsymbol{U}(m) - \boldsymbol{F}(m)\} = \begin{bmatrix} a_1^2(1+a_2)^2 & -a_1(1-a_2^2)(1+a_2) \\ -a_1(1+a_2)(1-a_2^2) & (1-a_2^2)^2 \end{bmatrix}.$$

Next, consider the loss in predictive efficiency with $m = 2$. From (3.23) and (3.24), we have, if $m = 2$, $h = 2$, $T\sigma^{-2}M_D(h) = 2 + 4a_1^2$, $T\sigma^{-2}M_P(h) = 1 + 4a_1^2 + a_2^2$,

$$T\sigma^{-2}\{M_D(h) - M_P(h)\} = 1 - a_2^2; \quad (4.2)$$

and, if $h = 3$, with $m = 2$, $T\sigma^{-2}M_D(h) = 2\{1+2a_1^2+3(a_1^2+a_2)^2\}$, $T\sigma^{-2}\{M_D(h)-M_P(h)\} = (1 + a_2)(1 - 2a_1^2a_2 - a_2)$. Thus, with $h = 2$ the loss in predictive efficiency is a monotonic decreasing function of only a_2 and the predictive efficiency of the direct method with respect to the plug in method increases as $|a_2|$ approaches one, the converse also holds and for $|a_2|$ close to zero the direct method would be least efficient in comparison with the plug in method. It may be observed, however, that for $h = 3$, the predictive efficiency loss of the direct method depends on the values of both a_1 and a_2.

Finally, for a general $m \geq 1$ and $h = 2$ we get from (3.15),

$$\begin{aligned} \sigma^{-2}\{\boldsymbol{U}(m) - \boldsymbol{F}(m)\} &= \boldsymbol{R}(m)^{-1} - \boldsymbol{\Gamma}(m)'\boldsymbol{R}(m)^{-1}\boldsymbol{\Gamma}(m), \\ \|\sigma^{-2}\{\boldsymbol{U}(m) - \boldsymbol{F}(m)\}\| &\geq \|\boldsymbol{R}(m)^{-1}\|\{1 - \|\boldsymbol{\Gamma}(m)'\|\,\|\boldsymbol{\Gamma}(m)\|\} > 0. \end{aligned}$$

Also, we may generalize the results (4.1) (4.2) given above for $h = 2$ and $m = 1, 2$. We have, for all $m \geq 1$ and $h = 2$, $T\sigma^{-2}M_D(h) = \{m + (m + 2)a_1^2\}$, $T\sigma^{-2}M_P(h) = (m - 1) + (m + 2)a_1^2 + a_m^2$, and

$$T\sigma^{-2}\{M_D(h) - M_P(h)\} = 1 - a_m^2, \quad (4.3)$$

where $a_j = -a_m(j)$, $j = 1, \ldots, m$. Thus, for $h = 2$ and all $m \geq 1$, the loss in predictive efficiency of the direct method with respect to the plug in method is a monotonic decreasing function of $a_m(m)^2$ and it tends to 1 as $a_m(m) \to 0$ and it tends to 0 as $a_m(m)^2 \to 1$, where $-1 \leq a_m(m) \leq 1$.

It should be noted, however, that both $M_P(h)$ and $M_D(h)$ are $O(T^{-1})$ and even for moderately large values of T their respective contributions to the overall prediction mean squared errors would be dominated by $V(h)$, the leading term in (3.17) and (3.19); consequently, the overall increase in the prediction mean squared error due to using the direct method in preference to the plug in method may not be substantial, a point illustrated further in Section 6.

5. Order Selection by the Direct Method

We now show that the h-step criteria, (2.18) - (2.20), are not consistent for m if α remains fixed as $T \to \infty$, but they are consistent for each $h \geq 1$, if $\alpha = \alpha(T)$, a function of T, and $\alpha(T) \to \infty$, $\alpha(T)/T \to 0$. However, as we do not attempt to establish a law of the iterated logarithm for the direct estimates, an answer to the question of how fast should $\alpha(T)$ grow with T for obtaining a consistent order selection is not given. For $h = 1$, as is well-known, setting $\alpha(T) > 2 \log\log T$, and $\alpha(T) = \log T$, in particular, yields a consistent estimator of m.

An explicit expression for the difference between $\hat{V}_{Dh}(m)$ and $\hat{V}_{Dh}(m+s)$, $s > 0$ is given in the following lemma; the lemma generalizes to $h > 1$ the well-known result (Hannan (1970), p.337) relating the differences between $\hat{\sigma}^2(m)$ and $\hat{\sigma}^2(m+s)$ to the estimated partial correlations. To save space, a proof of this lemma, and that of Theorem 4.1 and Proposition 4.1 below, has been omitted; the methods used essentially generalise to $h > 1$ the arguments used by Anderson (1971), pp.13-16, Shibata (1976) and Hannan (1980) for establishing the corresponding results for $h = 1$.

Lemma 5.1. *Let $\{x_t\}$ satisfy Assumption 1. For each fixed $s \geq 1$,*

$$N\{\hat{V}_{Dh}(m) - \hat{V}_{Dh}(m+s)\} = \sigma^2 \sum_{i=1}^{s} N\{\hat{\varphi}_{Dhm+i}(m+i)\}^2 + o_p(1).$$

Let $\hat{m}_\alpha(h)$ denote the order selected by minimising any of the h-step criteria, (2.18) - (2.20). We have the following theorem:

Theorem 5.1. *Let $\{x_t\}$ satisfy Assumption 1 and K Assumption 2, and suppose that $\alpha > 0$ is a fixed constant. Then, for each $h \geq 1$,*
(a) $\lim_{T\to\infty} P\{\hat{m}_\alpha(h) < m\} = 0$;
(b) $\lim_{T\to\infty} P\{\hat{m}_\alpha(h) = m\} < 1$;

(c) $\lim_{T\to\infty} P\{\hat{m}_\alpha(h) > m\} > 0$.

Suppose now that $\alpha = \alpha(T)$, a function of T, such that as $T \to \infty$, $\alpha(T) \to \infty$ but $\alpha(T)/T \to 0$. We have the following proposition:

Proposition 5.1. *Let the assumptions of Theorem* 4.1 *hold but* $\alpha = \alpha(T)$ *be such that as* $T \to \infty$, $\alpha(T) \to \infty$, $\alpha(T)/T \to 0$. *Then,*

$$\lim_{T\to\infty} P\{\hat{m}_\alpha(h) = j\} = \begin{cases} 0, & j \neq m, \\ 1, & j = m. \end{cases}$$

6. Simulation Results

The direct and plug in methods were applied to several different (15 altogether) autoregressive-moving average, ARMA, models of order (2, 2) of the form

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \delta_1 \varepsilon_{t-1} + \delta_2 \varepsilon_{t-2} + \varepsilon_t. \tag{6.1}$$

However, to save space, the discussion here is restricted to 5 second-order autoregressive, AR(2), models, called Models 1 - 5, with parameters, (a_1, a_2), shown in Table 1, and to 4 ARMA models, called Models 6 - 9, with parameter values shown in Table 2 using the notation $(p, q; a_1, \ldots, a_p; \delta_1, \ldots, \delta_q)$, where (p, q) denotes the order of the simulated ARMA model. These 9 models form a subset of the models used by Bhansali (1989, 1993) in two related but different studies.

The ε_t were generated as independent standard normal deviates using the routine G05DDF of the NAG library. Four different values of T, namely $T = 50, 100, 200, 500$, were considered, and the number of simulations for each model and each T was 500.

For each (Model, Simulation) combination, the direct and plug in forecasts up to 10 steps ahead were obtained. For ensuring that the observations used for generating the forecasts are (approximately) independent of those used for estimating the prediction constants, a total of $T+100$ observations was generated at each simulation, of which only the first T observations were used for fitting autoregressions by the direct and plug in methods as described in Section 2; the remaining 100 observations were split in three blocks: the first block, consisting of 70 observations was ignored, the second block of 20 observations was used for actually computing the multistep forecasts and the final block of 10 observations was used for a comparison of the calculated forecasts with the actual (simulated) values and for computing the simulated mean squared errors of prediction for the direct and plug in methods; thus, for each simulation, the value of n in (2.7) was set equal to $T + 90$.

Table 1. Simulated h-step mean squared errors of prediction* for various AR(2) models.

		Correct order fitted				Order Selected			
		Plug-in Method		Direct Method		$\alpha = 2$		$\alpha = \ln T$	
Model	h	Simul.	Asymp.	Simul.	Asymp.	Plug-in	Direct	Plug-in	Direct
				$T = 100$					
1 (0.4,-0.15)	2	121.3	117.7	121.2	118.6	122.9	125.3	121.5	119.3
	4	106.1	116.4	108.3	119.0	108.2	109.0	106.5	106.1
	6	108.6	116.4	112.1	119.1	110.6	111.1	108.8	110.3
	10	121.6	116.4	120.4	119.1	121.1	122.3	121.7	120.8
2 (0.4,0.3)	2	120.7	117.7	122.5	118.6	123.5	126.9	124.4	123.2
	4	133.1	148.7	136.2	151.1	136.7	140.5	136.1	138.9
	6	148.3	158.8	151.0	162.7	150.5	156.2	149.4	153.7
	10	164.8	163.0	162.0	168.9	163.0	166.5	164.7	164.1
3 (1.1,-0.24)	2	241.3	226.9	245.4	227.8	245.7	256.4	244.3	248.4
	4	373.9	392.1	383.7	397.2	384.4	403.1	379.9	389.5
	6	417.2	463.2	428.6	474.4	428.2	439.9	419.8	440.6
	10	486.1	497.7	470.5	521.5	473.4	483.5	484.8	475.8
4 (0.95,-0.9)	2	207.7	195.7	208.3	195.9	210.5	212.1	207.5	209.5
	4	266.1	273.2	268.7	274.4	280.1	278.4	270.3	272.6
	6	328.9	343.2	333.5	346.9	336.6	331.8	328.8	332.0
	10	452.8	497.1	474.6	509.2	468.4	501.3	454.4	486.9
5(1.75,-0.96)	2	471.1	420.4	472.9	420.5	480.6	493.2	479.0	482.2
	4	1386.1	1318.6	1403.3	1320.8	1398.0	1490.0	1401.4	1415.1
	6	1514.9	1604.1	1560.9	1612.4	1534.4	1639.0	1525.2	1620.9
	10	2402.2	2456.8	2484.2	2482.7	2419.7	2580.4	2428.8	2552.3
				$T = 500$					
1 (0.4,-0.15)	2	117.4	116.3	116.7	116.5	117.6	118.9	117.5	118.4
	4	122.9	116.3	123.6	117.0	122.8	124.8	123.0	123.1
	6	113.4	116.4	113.9	117.0	112.9	114.4	113.4	113.3
	10	132.5	116.4	132.1	117.0	132.4	132.8	132.5	132.0
2 (0.4,0.3)	2	117.6	116.3	117.1	116.5	106.8	119.6	117.5	116.8
	4	149.6	146.9	150.8	147.3	150.2	152.8	149.5	151.1
	6	158.6	157.3	160.1	158.1	158.5	162.0	158.6	159.4
	10	177.2	162.4	175.6	163.6	177.4	179.8	177.2	177.5
3 (1.1,-0.24)	2	225.8	222.2	225.8	222.4	224.8	228.4	225.8	225.7
	4	386.6	382.1	390.3	383.1	388.2	397.2	386.8	394.4
	6	466.6	452.0	471.4	454.3	467.9	483.5	466.6	473.9
	10	502.2	491.4	498.0	496.2	501.8	513.5	502.1	504.0
4 (0.95,-0.9)	2	195.1	191.3	195.2	191.4	193.3	195.8	195.0	195.0
	4	272.5	265.0	272.7	265.2	270.0	275.2	272.5	273.2
	6	304.9	331.8	304.9	332.5	304.4	307.6	305.0	305.6
	10	486.9	474.2	490.5	476.6	488.8	498.4	487.0	489.5
5 (1.75,-0.96)	2	431.9	409.1	431.7	409.1	425.7	425.9	431.4	430.9
	4	1302.9	1262.2	1302.3	1262.6	1298.1	1300.4	1301.4	1301.6
	6	1602.4	1530.7	1606.5	1532.3	1601.3	1648.0	1601.8	1612.5
	10	2202.9	2288.3	2211.5	2293.5	2194.3	2212.2	2200.8	2202.4

* The mean squared errors have been multiplied by 100.

Table 2. Ratios* of the simulated h-step mean squared errors of prediction of the Plug in and direct methods when fitting AR models of fixed & selected orders.

		k							Order Selected	
Model	h	1	2	3	4	5	8	10	$\alpha = 2$	$\alpha = \ln T$
6 (0,1;0.5)	2	101.4	100.5	99.1	99.3	99.5	99.4	99.2	99.9	98.2
	4	100.8	99.9	99.3	99.2	99.5	100.1	99.6	99.6	100.5
	6	100.9	99.4	99.1	99.3	98.2	98.9	98.9	99.0	100.1
	10	100.4	100.4	99.5	99.0	99.2	98.2	96.8	99.4	100.4
7 (1,2;-0.7;-1.1,0.3)	2	103.4	102.3	101.3	101.1	99.4	99.9	99.8	99.0	99.3
	4	109.1	101.6	103.7	98.2	99.0	99.4	100.3	98.8	99.4
	6	111.9	104.9	101.2	99.0	100.1	98.9	99.6	99.3	100.3
	10	105.9	101.1	100.5	100.4	100.3	100.3	99.3	100.0	101.0
8 (2,1;-0.64,-0.7;0.8)	2	156.0	107.5	102.4	105.9	99.7	100.0	99.4	99.4	99.8
	4	101.0	106.1	101.0	102.8	100.7	99.6	99.9	100.3	102.2
	6	105.5	111.1	104.5	99.6	98.4	99.8	98.8	98.6	100.7
	10	101.6	98.4	97.8	102.2	102.2	97.4	97.2	98.4	99.5
9 (1,1;0.5;0.8)	2	104.2	102.8	99.8	99.4	99.7	100.0	99.4	100.9	97.4
	4	104.9	101.2	102.1	102.2	99.6	99.7	100.0	99.6	99.9
	6	101.8	100.3	99.3	98.9	100.6	97.7	98.2	101.7	101.1
	10	100.8	100.5	99.3	99.6	99.0	98.2	98.2	99.3	100.1
10 (2,0;0.95;-0.9)	2	99.7	108.6	106.2	126.8	101.6	99.8	100.4	98.8	101.9
	4	101.0	104.0	136.9	134.1	107.0	101.1	100.3	101.6	102.2
	6	167.5	166.1	151.4	153.6	111.1	98.6	100.6	102.1	102.4
	10	99.5	100.3	110.2	111.1	104.0	100.1	100.8	98.7	99.4

* The ratios have been multiplied by 100.

The simulated mean squared errors of prediction of the direct and plug in methods were computed for the five AR(2) models in two separate situations: first, when the order of the fitted model equalled the actual second order of the generated model and second when the fitted order was selected from the data by the criterion (2.20) but with $\alpha = 2$ and $\alpha = \ln T$, the latter corresponding to the use of a consistent criterion; the value of K, the maximum order fitted, was set to equal 20 with $T \geq 100$, but a smaller value was used with $T = 50$.

For the remaining four ARMA models, the simulated mean squared errors of prediction were again computed in two separate situations, first, when an AR(k) model was fitted by the direct and plug in methods, but with k taking each of ten different fixed values, namely $k = 1, 2, \ldots, 10$, for all h and in all 500 simulations, and, secondly, when k was selected by the criterion (2.20) but with $\alpha = 2$ and $\alpha = \ln T$.

Note that (6.1) is a linear model and even when $q > 0$, the generated process could be well approximated by an autoregressive model. For comparing the behaviour of the direct and plug in forecasts when the stochastic structure producing an observed time series is unknown but it could be non-linear, we also

generated a stretch of $NN = T + 500$ observations from the following model:

$$\text{Model 10:} \quad y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t,$$

where $\{\varepsilon_t\}$ was generated as a sequence of independent Exponential variates, each with mean 1.0. To save space, we only consider ($a_1 = 0.95$, $a_2 = -0.9$) and $T = 500$. The direct and plug in methods were however actually applied to observations $\tilde{x}_t = x_t - \bar{x}$, where $\bar{x}$ denotes the arithmetic mean of the x_t's and the x_t's were obtained by a non-linear transformation of y_t defined as follows:

$$\begin{aligned} x_t &= y_t^{*2}, \\ y_t^* &= y_{NN-t+1}, \qquad t = 1, \ldots, NN,\ NN = T + 500. \end{aligned}$$

The actual technique used for generating the multistep forecasts by the direct and plug in methods was however the same as that described above and the details are not repeated.

The simulated mean squared errors of prediction for the five AR(2) models are shown in Table 1 together with the 'asymptotic' mean squared errors. The latter were computed from (3.17), (3.19) and Proposition 3.2 and apply only when the correct second order is fitted. To save space, only the results for $h = 2, 4, 6, 10$ and $T = 100$ and 500 are shown: The simulated mean squared errors of the direct method are seen to be generally greater than those of the plug in method, both when the order is selected and when the order is treated as known. On the other hand, the difference between their respective mean squared errors is not large and it is more noticeable for Model 5 than for other models. As may be expected on intuitive grounds, the simulated mean squared errors are generally larger when the order is selected than when the correct order is fitted; also the former tend to be smaller when $\alpha = \ln T$ than when $\alpha = 2$.

For models 6 - 10, the ratio of the simulated mean squared errors of prediction of the plug in and direct methods, after multiplication by 100, is shown in Table 2; thus, if the displayed value of this ratio equals 100 then the two methods have identical mean squared errors, a value less than 100 indicates that the plug in method has a smaller mean squared error, and a value greater than 100 indicates that the converse holds and the direct method has a smaller mean squared error; moreover, in both these situations a quantification of the extent to which the relevant mean squared error is smaller is also given in Table 2. For all these five models, there is no 'true' autoregressive order and thus the results obtained with the various values of k demonstrate the effect of approximating the generated process by a possibly under-parametrized autoregressive model; by contrast, the results for order selected show the effectiveness of the criterion (2.20) in selecting an approximating autoregressive model for describing the generating structure of the process.

Unlike the results described above for pure autoregressive models, the plug in method is not necessarily superior to the direct method for Models 6 - 10, especially when judged by the magnitudes of their respective mean squared errors of prediction. It is seen that when an underparametrized autoregressive model is fitted, that is, especially with $k = 1$ and 2, the simulated mean squared errors of the plug in method are generally greater than those of the direct method and for Models 8 and 10 in particular the former could be greater by as much as 66%. On the other hand, however, if a sufficiently long autoregression is fitted, that is, especially with $k = 8$ and 10, the plug in method tends to have a smaller mean squared error, though the difference is not necessarily substantial. To gain an appreciation for this behaviour, we observe that, in the former situation, the main advantage of using the direct method, namely its smaller 'bias' is likely to dominate its disadvantage, namely its greater variability in estimating the prediction constants; but the converse may hold in the latter situation and the greater variability of the direct method could nullify its smaller bias.

For an observed time series, the order of an approximating autoregressive model would be selected by employing an order determining criterion. It is seen that in this situation, neither of these two methods has a clear advantage over the other. Thus, for $\alpha = \ln T$, the direct method tends to have a smaller mean squared error, and the converse holds when $\alpha = 2$; at the same time, the difference between their respective mean squared errors is not necessarily substantial.

Thus, in conclusion, the simulation results indicate that for a finite autoregressive process, the plug in method has a clear advantage over the direct method for multistep prediction, even when the order of the fitted model is selected by an order determining criterion. For other models, however, the position is less clear cut and the use of direct method would help in reducing the mean squared error of prediction if an under-parametrized model is used and that this possibility could arise if the autoregressive order is selected by the criterion (2.20) with $\alpha = \ln T$.

References

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

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2*nd International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csáki), 267-281. Akadémiai Kiadó, Budapest.

Akaike, H. (1977). On entropy maximization principle. In *Applications of Statistics* (Edited by P. Krishnaiah), 27-41. North Holland, Amsterdam.

Anderson, T. W. (1971). *The Statistical Analysis of Time Series.* Wiley, New York.

Anderson, T. W. (1975). Maximum likelihood estimation of parameters of autoregressive processes with moving average residuals and other covariance matrices with linear structure. *Ann. Statist.* **3**, 1283-1304.

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. (1989). Estimation of the moving-average representation of a stationary process by autoregressive model fitting. *J. Time Series Anal.* **10**, 215-232.

Bhansali, R. J. (1993). Estimation of the prediction error variance and an R^2 measure by autoregressive model fitting. *J. Time Series Anal.* **14**, 125-146.

Bhansali, R. J. (1996). Asymptotically efficient autoregressive model selection for multistep prediction. *Ann. Inst. Statist. Math.* **48**, 577-602.

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

Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory*. Holt, New York.

Findley, D. F. (1983). On the use of multiple models for multi- period forecasting. In *Proc. Bus. Econ. Statist. Sect.* pp. 528-531. Washington, D. C., Amer. Statist. Assoc.

Galbraith, R. F. and Galbraith, J. I. (1974). On the inverses of some patterned matrices arising in the theory of stationary time series. *J. Appl. Probab.* **11**, 63-71.

Hannan, E. J. (1970). *Multiple Time Series*. Wiley, New York.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8**, 1071-1081.

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

Hosoya, Y. and Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *Ann. Statist.* **10**, 132-153.

Kabaila, P. V. (1981). Estimation based on one step ahead prediction versus estimation based on multi-step ahead prediction. *Stochastics* **6**, 43-55.

Lewis, R. C. and Reinsel, G. C. (1985). Prediction of multivariate time series by autoregressive model fitting. *J. Multivariate Anal.* **16**, 393-411.

Lin, J.-L. and Granger, C. W. J. (1994). Forecasting from non-linear models in practice. *J. Forecasting* **13**, 1-9.

Rissanen, J. (1978). Modelling by shortest data description. *Automatica* **14**, 465-471.

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

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

Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147-164.

Shibata, R. (1981). An optimal autoregressive spectral estimate. *Ann. Statist.* **9**, 300-306.

Stine, R. A. (1987). Estimating properties of autoregressive forecasts. *J. Amer. Statist. Assoc.* **82**, 1072-1078.

Tiao, G. C. and Xu, D. (1993). Robustness of MLE for multi-step predictions: the exponential smoothing case. *Biometrika* **80**, 623-641.

Whittle, P. (1963). *Prediction and Regulation by Linear Least Squares Methods*. London, English University Press.

Yamamoto, T. (1976). Asymptotic mean square prediction error for an autoregressive model with estimated coefficients. *Appl. Statist.* **25**, 123-127.

Department of Mathematical Science, University of Liverpool, P. O. Box 147, Liverpool L69 3BX, United Kingdom.

(Received January 1995; accepted December 1996)