Statistica Sinica **7**(1997), 395-406

SELECTION OF A MULTISTEP LINEAR PREDICTOR FOR SHORT TIME SERIES

Clifford M. Hurvich and Chih-Ling Tsai

New York University and University of California, Davis

Abstract: We develop a version of the Corrected Akaike Information Criterion (AIC_C) suitable for selection of an h-step-ahead linear predictor for a weakly stationary time series in discrete time. A motivation for this criterion is provided in terms of a generalized Kullback-Leibler information which is minimized at the optimal h-step predictor, and which is equivalent to the ordinary Kullback-Leibler information when $h = 1$. In a simulation study, we find that if the sample size is small and the predictor coefficients are estimated by Burg's method, then AIC_C typically outperforms both the ordinary Akaike Information Criterion (AIC) and the Final Prediction Error (FPE) for h-step prediction, and we present evidence to indicate that Burg estimation can produce much better selected predictors than Yule-Walker estimation.

Key words and phrases: AIC_C, Burg's estimator, Kullback-Leibler information.

1. Introduction

Given data $x_1, \ldots, x_n$ from a weakly stationary time series $\{x_t\}_{t=-\infty}^{\infty}$ with zero mean, suppose we wish to predict x_{n+h} where $h > 0$ is the lead time. We restrict attention to the case where the true autocovariance sequence $\{r_j\}$ is unknown and the predictor is a linear combination of the k present and past values $x_{n-k+1}, \ldots, x_n$, with coefficients estimated from the entire available data set using either the least squares, Burg's method, or the Yule-Walker method. Although in practice estimation and prediction will be done on the same time series $\{x_t\}$, we, as well as many other authors, find it convenient to follow the lead of Akaike (1970) and measure forecasting error by $E_y[y_{n+h} - \hat{y}_{n+h}]^2$, the mean squared error incurred in applying the predictor with coefficients obtained from $\{x_t\}$ to an independent realization $\{y_t\}$ which has the same probabilistic structure as $\{x_t\}$, where E_y denotes the expectation with respect to $\{y_t\}$.

An important question is how to choose k on the basis of the available data in such a way as to keep the mean squared error of the resulting fitted predictor as small as possible. For the case of one-step prediction ($h = 1$), Shibata (1980) proved, under certain conditions, that if k is selected to minimize Akaike's Information Criterion (AIC; Akaike (1973)) then the resulting predictor is asymptotically efficient in the sense of minimizing the one-step mean squared prediction error as defined above, within a class of candidate one-step predictors estimated by least squares, assuming that the true model is not a finite order autoregression.

For the case of multistep prediction ($h > 1$), Shibata (1980) asserted that if the one-step residual sum of squares in AIC is replaced by the h-step residual sum of squares, then the resulting selection criterion is asymptotically efficient in terms of h-step mean squared prediction error, within a class of h-step predictors fitted by least squares. This claim was subsequently proved by Bhansali (1996). Consequently, if the true generating mechanism is not a finite order autoregression, then the asymptotically optimal selection criterion depends on h, and no single choice of k is necessarily good for all values of h simultaneously. Therefore, it may be appropriate to consider using a different value of k, and in effect to fit a different model, for forecasting at each lead time h. This notion has previously been advocated by Findley (1983) and by Hurvich (1987). In a related idea which does not involve model selection, Tiao and Xu (1993) and Tiao and Tsay (1994) considered using different parameter values for different values of h.

In this paper, we will focus on the small-sample performance of methods of selecting k for an h-step predictor. It has been demonstrated in Hurvich and Tsai (1991) that the Corrected Akaike Information Criterion $\mathrm{AIC_C}$ can produce better one-step predictors than AIC in small samples. Here we introduce a multistep generalization of $\mathrm{AIC_C}$ and compare it in a simulation study with the multistep versions of AIC and the Final Prediction Error (FPE; Akaike (1970)). In the multistep version of $\mathrm{AIC_C}$, the estimated one-step innovation variance in $\mathrm{AIC_C}$ is replaced by the estimated h-step innovation variance. We provide a motivation for the new criterion by introducing a generalized version of the Kullback-Leibler information which is minimized at the optimal h-step predictor, and which is equivalent to the ordinary Kullback-Leibler information when $h = 1$. In a simulation study, we find that if the Burg estimate is used, $\mathrm{AIC_C}$ typically outperforms the other criteria for h-step prediction, and we present evidence to indicate that Burg estimation can produce much better selected predictors than Yule-Walker estimation.

2. Optimal and Estimated Multistep Predictors

Suppose $\{x_t\}$ is weakly stationary with mean zero and autocovariance sequence $\{r_j\}$. The linear predictor of x_{t+h} based on $\{x_{t-k+1}, \ldots, x_t\}$ which is best in the sense of minimizing the h-step mean squared error is given by

$$\hat{x}_{t+h} = -\sum_{j=h}^{h+k-1} a_j(h,k) x_{t+h-j},$$

where the predictor coefficients $\{-a_j(h,k)\}_{j=h}^{h+k-1}$ are determined by Equation (1) below. This optimal predictor cannot be constructed in practice, since $\{r_j\}$ will be unknown. Define the $h+k$ dimensional vector $a(h,k) = [a_0(h,k), a_1(h,k), \ldots, a_{h+k-1}(h,k)]'$, where $a_0(h,k) = 1$ for all (h,k), and $a_1(h,k) = \cdots = a_{h-1}(h,k) = 0$ if $h > 1$. We refer to $a(h,k)$ as the optimal prediction error filter since the h-step prediction error is given by

$$x_{t+h} - \hat{x}_{t+h} = \sum_{j=0}^{h+k-1} a_j(h,k) x_{t+h-j} = a(h,k)'(x_{t+h}, \ldots, x_{t-k+1})'.$$

The h-step mean squared prediction error for this predictor is $\sigma^2(h,k) = E[x_{t+h} - \hat{x}_{t+h}]^2 = a(h,k)' R_{h+k} a(h,k)$, where R_m is the $m \times m$ covariance matrix of $(x_1, \ldots, x_m)'$. We assume that $\sigma^2(h,k) > 0$.

From the optimality of $\hat{x}_{t+h}$ it follows that $x_{t+h} - \hat{x}_{t+h}$ must be uncorrelated with each of $x_{t-k+1}, \ldots, x_t$, and hence the predictor coefficients must satisfy the equations $E[(x_{t+h} - \hat{x}_{t+h}) x_{t+h-i}] = 0$ for $i = h, \ldots, h+k-1$, or equivalently

$$\sum_{j=h}^{h+k-1} r_{|i-j|} a_j(h,k) = -r_i, \qquad (i = h, \ldots, h+k-1). \tag{1}$$

Given observations $x_1, \ldots, x_n$ with $n \geq h+k$, we can construct estimates $\{\hat{r}_j\}_{j=0}^{h+k-1}$ of $\{r_j\}_{j=0}^{h+k-1}$ and use these to form estimates $\{\hat{a}_j(h,k)\}_{j=h}^{h+k-1}$ of the parameters $\{a_j(h,k)\}_{j=h}^{h+k-1}$ by solving the sample analog of Equation (1),

$$\sum_{j=h}^{h+k-1} \hat{r}_{|i-j|} \hat{a}_j(h,k) = -\hat{r}_i \qquad (i = h, \ldots, h+k-1). \tag{2}$$

The corresponding estimated h-step innovation variance is $\hat{\sigma}^2(h,k) = \sum_{j=0}^{h+k-1} \hat{a}_j(h,k) \hat{r}_j$.

Here, we will focus on two methods of estimating the $\{r_j\}_{j=0}^{h+k-1}$. They are: (1) the Burg Method Burg (1978) (see also Hainz (1995), Haykin (1983), Chapter 2), yielding the Burg estimates of the predictor coefficients; (2) the averaged lagged products $\hat{r}_j = \frac{1}{n} \sum_{t=j+1}^{n} x_{t-j} x_t$, yielding the Yule-Walker estimates of the predictor coefficients.

For completeness, we note that for one step ahead prediction Shibata (1980) estimates the (l,m) entry of R_k by $\hat{r}_{lm} = \frac{1}{n-K} \sum_{t=K+1}^{n} x_{t-l} x_{t-m}$ where K is the largest value of k under consideration. This yields the least squares estimates of the predictor coefficients as the solution of the system $\hat{R}(k)\hat{a}(1,k) = -\hat{r}(k)$ where $\hat{R}(k) = (\hat{r}_{lm}\ 1 \leq l, m \leq k)$ and $\hat{r}(k) = (\hat{r}_{10}, \ldots, \hat{r}_{k0})'$. It can be shown that the least squares estimates $\hat{a}(1,k)$ are the value $a_1, \ldots, a_k$ which minimize the residual sum of squares $\sum_{t=K+1}^{n} (x_t - a_1 x_{t-1} \cdots - a_k x_{t-k})^2$.

Of the three estimators mentioned above, we prefer Burg's method for a variety of reasons. To simplify the discussion here, suppose we have n observations from a k'th order autoregression. The Burg and least squares autoregressive parameter estimates have identical biases, to terms of order $1/n$. (See Hainz (1995)). This bias, in turn, is less than that of the Yule-Walker estimate. (See Tjoestheim and Paulsen (1983), Shaman and Stine (1988).) On the other hand, the Burg estimate of the innovation variance σ^2 has an asymptotic bias of $-k\sigma^2/n$, compared to $-2k\sigma^2/n$ for $1/n$ times the least squares residual sum of squares, assuming $K = k$ (Shaman (1983)). Unlike the least squares estimates, the Burg parameter estimates are guaranteed to correspond to a stationary model, can be computed recursively as the candidate autoregressive order is increased, and do not suffer from the reduction of effective sample size which is inherent in least squares, where the response variable is the $n-K$ dimensional vector $(x_n, \ldots, x_{K+1})'$. Finally, the Burg estimates are not sensitive to the choice of the largest candidate autoregressive order, K.

3. Generalized Kullback-Leibler Information

Consider a candidate linear predictor $\tilde{x}_{t+h}$ of x_{t+h} based on $x_{t-k+1}, \ldots, x_t$,

$$\tilde{x}_{t+h} = -\sum_{j=h}^{h+k-1} b_j(h,k) x_{t+h-j},$$

where the coefficients $\{b_j(h,k)\}_{j=h}^{h+k-1}$ are arbitrary real constants. Define the $h+k$-dimensional vector $b(h,k) = [b_0(h,k), b_1(h,k), \ldots, b_{h+k-1}(h,k)]'$, where $b_0(h,k) = 1$ for all (h,k), and $b_1(h,k) = \cdots = b_{h-1}(h,k) = 0$ if $h > 1$. Then the prediction error of $\tilde{x}_{t+h}$ is $x_{t+h} - \tilde{x}_{t+h} = b(h,k)'(x_{t+h}, \ldots, x_{t-k+1})'$, so that $b(h,k)$ is the candidate prediction error filter. The mean squared error of $\tilde{x}_{n+h}$ is given by

$$b(h,k)' R_{h+k} b(h,k) = \sigma^2(h,k) + [b(h,k) - a(h,k)]' R_{h+k} [b(h,k) - a(h,k)]. \quad (3)$$

From the point of view of linear prediction of x_{t+h} based on $x_{t-k+1}, \ldots, x_t$, a characterization of the process $\{x_t\}$ is provided by the optimal prediction error filter $a(h,k)$, together with the optimal prediction error variance $\sigma^2(h,k)$. We consider $a(h,k)$ and $\sigma^2(h,k)$ as unknown parameters. Let $\tau^2(h,k) > 0$ be an arbitrary candidate for the prediction error variance $E[x_{t+h} - \tilde{x}_{t+h}]^2$, not necessarily equal to the value given in Equation (3).

Our proposed generalization of the Kullback-Leibler information provides a discrepancy measure between the candidate parameters $b(h,k)$, $\tau^2(h,k)$ and the true parameters $a(h,k), \sigma^2(h,k)$. It is defined by

$$d_{h,k,a,\sigma^2}(b, \tau^2) = \log \tau^2(h,k) + b(h,k)' R_{h+k} b(h,k) / \tau^2(h,k), \quad (4)$$

where for ease of readability we have partially suppressed the (h,k) notation. From Equation (3), if τ^2 is held fixed, it is seen that $d_{h,k,a,\sigma^2}(b,\tau^2)$ is minimized over the set of all candidate prediction error filters by $b=a(h,k)$, resulting in the discrepancy

$$d_{h,k,a,\sigma^2}(a,\tau^2)=\log\tau^2(h,k)+\sigma^2(h,k)/\tau^2(h,k). \tag{5}$$

Treated as a function of τ^2, the quantity $d_{h,k,a,\sigma^2}(a,\tau^2)$ is minimized by taking $\tau^2=\sigma^2(h,k)$. Thus, we have shown that $d_{h,k,a,\sigma^2}(\cdot,\cdot)$ is indeed a discrepancy function in the sense of Linhart and Zucchini (1986), since it is minimized over the set of all candidate parameters b, τ^2 by the true parameters $a(h,k)$ and $\sigma^2(h,k)$. In the case $h=1$, if Whittle's approximation to the log likelihood is used, then it is easily shown that our proposed d reduces to the ordinary Kullback-Leibler discrepancy.

4. The Proposed Selection Criterion

A reasonable criterion for judging the adequacy of the candidate h-step predictor $(\hat{a}(h,k),\hat{\sigma}^2(h,k))$ in the light of the observed data is obtained by replacing (b,τ^2) in Equation (4) by $(\hat{a},\hat{\sigma}^2)$, multiplying by n and taking expectations, yielding $\Delta_h(k)=E[nd_{h,k,a,\sigma^2}(\hat{a},\hat{\sigma}^2)]$. Therefore, for selection of a linear predictor for h-step prediction, we propose to choose k to minimize an approximately unbiased estimate of

$$\Delta_h(k)=E[n\log\hat{\sigma}^2(h,k)+n\hat{a}(h,k)'R_{h+k}\hat{a}(h,k)/\hat{\sigma}^2(h,k)].$$

As pointed out by Shibata (1980), $\hat{a}(h,k)$ may be thought of as an estimate of the parameter $\alpha=(1,0,\ldots,0,\alpha_h,\ldots,\alpha_{h+k-1})'$ when an autoregressive model

$$x_{t+h}+\alpha_h x_t+\cdots+\alpha_{h+k-1}x_{t-k+1}=\varepsilon_{t+h} \tag{6}$$

is fitted to observations $x_1,\ldots,x_n$. Note that in the model (6), the true coefficients of $x_{t+h-1},\ldots,x_{t+1}$ are all zero, so that the best linear predictor of x_{t+h} is $\hat{x}_{t+h}=-\sum_{j=h}^{h+k-1}\alpha_j x_{t+h-j}$. If follows that if model (6) is correct, then $\alpha=a(h,k)$, and $\operatorname{Var}[\varepsilon_t]=\sigma^2(h,k)$. Furthermore, it is shown in the Appendix for a multistep version of the linear regression estimators given in Brockwell and Davis (1991), Eq. (8.10.3) (chosen here for the sake of mathematical tractability, even though we will use the Burg estimators in our simulation study) that if model (6) holds and $\{\varepsilon_t\}$ is Gaussian white noise, then the final k entries of $\sqrt{n}(\hat{a}(h,k)-a(h,k))$ are asymptotically distributed as $N(0,\sigma^2(h,k)R_k^{-1})$, and $n\hat{\sigma}^2(h,k)/\sigma^2(h,k)$ is asymptotically distributed as χ^2_{n-k}, independently of $\hat{a}(h,k)$. In order to facilitate the derivation of a reasonably simple selection criterion, we will, for the remainder of this section, make the strong assumption that this asymptotic distribution theory holds exactly for the given sample size. The resulting criterion will later be evaluated for its practical usefulness.

Equation (3) with $b = \hat{a}$ yields

$$n\hat{a}(h,k)'R_{h+k}\hat{a}(h,k) = n\sigma^2(h,k) + n[\hat{a}(h,k) - a(h,k)]'R_{h+k}[\hat{a}(h,k) - a(h,k)].$$

Since the initial h entries of $\hat{a}(h,k) - a(h,k)$ are zero, we conclude under the above assumptions that $n\hat{a}(h,k)'R_{h+k}\hat{a}(h,k)/\hat{\sigma}^2(h,k)$ is distributed as $n^2/\chi^2_{n-k} + n\chi^2_k/\chi^2_{n-k}$, where χ^2_k and χ^2_{n-k} are independent, so that $E[\Delta_h(k)] = E[n\log\hat{\sigma}^2(h,k)] + n(n+k)/(n-k-2)$. It follows that $\text{AIC}_\text{C}(h,k) = n\log\hat{\sigma}^2(h,k) + n(n+k)/(n-k-2)$ is an approximately unbiased estimator of $\Delta_h(k)$. An equivalent expression which is more readily compared with the h-step version of AIC is

$$\text{AIC}_\text{C}(h,k) = n[\log\hat{\sigma}^2(h,k) + 1] + 2(k+1)\Big[\frac{n}{n-k-2}\Big]. \qquad (7)$$

If we define

$$\text{AIC}(h,k) = n[\log\hat{\sigma}^2(h,k) + 1] + 2(k+1), \qquad (8)$$

then

$$\text{AIC}_\text{C}(h,k) = \text{AIC}(h,k) + \frac{2(k+1)(k+2)}{n-k-2},$$

so that $\text{AIC}_\text{C}(h,k)$ is asymptotically efficient for selection of an h-step linear predictor.

5. Monte Carlo Results

For each of three sample sizes, $n = 30, 50, 75$, we generated one hundred realizations from each of three time series models. The models were a noninvertible second order moving average (MA(2)) $x_t = \varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}$, a fourth order autoregression (AR(4)) $x_t = 2.7607x_{t-1} - 3.8106x_{t-2} + 2.6535x_{t-3} - .9238x_{t-4} + \varepsilon_t$, and a second order autoregression (AR(2)) $x_t = .99x_{t-1} - .8x_{t-2} + \varepsilon_t$, where in each case the ε_t are independent identically distributed standard normal. For the AR(2) model with $n = 30$ we also used errors with a long-tailed distribution, $(1/\sqrt{3})t_3$. The AR(4) model was used in the simulation study of Beamish and Priestley (1981) and the AR(2) model was used in the simulation study of Hurvich and Tsai (1989).

For each realization from each process and sample size, we computed h-step predictors with $k = 0, \ldots, 20$ and $h = 1, 2, 5$ by solving Equation (2) using the Burg estimates of the autocovariances $\{\hat{r}_j\}_{j=0}^{h+k-1}$, and we evaluated the corresponding mean squared errors $\hat{a}(h,k)'R_{h+k}\hat{a}(h,k)$, as well as the selection criteria $\text{AIC}_\text{C}(h,k)$ (Eq. (7)), $\text{AIC}(h,k)$ (Eq. (8)), and $\text{FPE}(h,k) = \hat{\sigma}^2(h,k)(n+k)/(n-k)$. Note that the predictor with $k = 0$ is the process mean, zero. For each process, sample size and lead time, we also estimated the expected mean squared error, $\text{MSE}\,(h,k) = E_x[\hat{a}(h,k)' R_{h+k}\hat{a}(h,k)]$ for $k = 0, \ldots, 20$, by averaging $\hat{a}(h,k)' R_{h+k}\hat{a}(h,k)$ over the one hundred realizations.

As a measure of the performance of a selection criterion for a given simulated realization and lead time, we used $\text{MSE}\,(h,\hat{k})$, where $\hat{k}$ is the value of k chosen by the criterion from the candidates $k = 0, \ldots, 20$. Averages (over the one hundred realizations) of $\text{MSE}\,(h,\hat{k})$ for AIC, AIC_C and FPE are given in Tables 1-3. We denote these averages by $\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}})\}$, $\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}_\text{C}})\}$, and $\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{FPE}})\}$, respectively. To provide some context for these average mean squared errors, Tables 1-3 also provide values of $\text{Ave}\{\text{MSE}\,(h,k^*)\}$, where k^* is the minimizer of $\text{Ave}\{\text{MSE}(h,\cdot)\}$. In all cases studied except for the MA(2) process with $n = 75$ and $h = 1$, $\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}_\text{C}}\}$ was less than both $\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}})\}$ and $\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{FPE}})\}$. In all cases, $\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}_\text{C}})\}$ exceeded $\text{Ave}\{\text{MSE}(h,k^*)\}$ by less than 9%. By contrast, for $n = 30$, with only two exceptions, the excesses for AIC and FPE were all greater than 50%. For $n = 50$ and $n = 75$, the excesses for AIC and FPE were smaller, but the excess for AIC was greater that 9% in 6 of the 9 cases with $n = 50$.

Table 1. Average MSE of the selected h-step predictors for MA(2) process.

	$n=30$			$n=50$			$n=75$		
	$h{=}1$	$h{=}2$	$h{=}5$	$h{=}1$	$h{=}2$	$h{=}5$	$h{=}1$	$h{=}2$	$h{=}5$
$\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}_\text{C}})\}$	2.14	6.22	6.34	1.83	6.11	6.23	1.63	6.06	6.26
$\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}})\}$	3.06	11.00	11.26	1.89	6.63	6.62	1.63	6.25	6.47
$\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{FPE}})\}$	2.72	9.71	9.57	1.86	6.50	6.58	1.63	6.23	6.46
$\text{Ave}\{\text{MSE}(h,k^*)\}$	1.99	6.00	6.00	1.75	6.00	6.00	1.58	5.93	6.00

Note: k^* is the minimizer of $\text{Ave}\{\text{MSE}(h,\cdot)\}$. Averages based on one hundred simulated realizations.

Table 2. Average MSE of the selected h-step predictors for AR(4) process.

	$n=30$			$n=50$			$n=75$		
	$h{=}1$	$h{=}2$	$h{=}5$	$h{=}1$	$h{=}2$	$h{=}5$	$h{=}1$	$h{=}2$	$h{=}5$
$\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}_\text{C}})\}$	1.63	15.72	61.73	1.20	11.00	42.78	1.13	10.20	39.64
$\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{AIC}})\}$	3.39	33.95	146.84	1.29	11.90	47.13	1.14	10.39	40.17
$\text{Ave}\{\text{MSE}(h,\hat{k}_{\text{FPE}})\}$	2.52	26.22	124.27	1.27	11.80	45.81	1.14	10.36	40.00
$\text{Ave}\{\text{MSE}(h,k^*)\}$	1.62	15.46	59.27	1.18	10.76	41.43	1.13	10.14	38.30

Note: k^* is the minimizer of $\text{Ave}\{\text{MSE}(h,\cdot)\}$. Averages based on one hundred simulated realizations.

Table 3. Average MSE of the selected h-step predictors for AR(2) process.

	$n = 30$			$n = 30$, $(1/\sqrt{3})t_3$ Errors		
	$h=1$	$h=2$	$h=5$	$h=1$	$h=2$	$h=5$
$\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{AIC_C}})\}$	1.14	2.36	3.77	1.15	2.40	3.73
$\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{AIC}})\}$	2.34	4.80	7.10	1.87	4.14	6.40
$\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{FPE}})\}$	1.75	4.08	6.07	1.62	3.52	5.60
$\mathrm{Ave}\{\mathrm{MSE}(h, k^*)\}$	1.10	2.23	3.47	1.10	2.22	3.47
	$n = 50$			$n = 75$		
	$h=1$	$h=2$	$h=5$	$h=1$	$h=2$	$h=5$
$\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{AIC_C}})\}$	1.07	2.17	3.37	1.04	2.09	3.20
$\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{AIC}})\}$	1.12	2.32	3.68	1.05	2.14	3.27
$\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{FPE}})\}$	1.11	2.29	3.58	1.05	2.12	3.27
$\mathrm{Ave}\{\mathrm{MSE}(h, k^*)\}$	1.05	2.12	3.25	1.03	2.07	3.14

Boxplots of $\mathrm{MSE}(h, \hat{k})(h = 1, 5)$ for AIC, $\mathrm{AIC_C}$ and FPE are given in Figure 1 for the AR(2) process with normal errors. The plots indicate that $\mathrm{AIC_C}$ typically performs better than the other two criteria. (Boxplots for the other two processes yielded similar findings, and are omitted to save space.) To provide a more formal assessment, for each process and lead time, we performed a one-sample Wilcoxon test on the set of values of $\mathrm{MSE}(h, \hat{k}_{\mathrm{AIC}}) - \mathrm{MSE}\,(h, \hat{k}_{\mathrm{AIC_C}})$ for the null hypothesis that the true median of the differences is zero against the alternative hypothesis that the median is positive, and we performed a similar Wilcoxon test on the values of $\mathrm{MSE}(h, \hat{k}_{\mathrm{FPE}}) - \mathrm{MSE}(h, \hat{k}_{\mathrm{AIC_C}})$. A significant result is taken as an indication that $\mathrm{AIC_C}$ performed better than the other criterion under consideration, for the given process, sample size and lead time. Of the 60 tests performed, 54 were significant at level .005. Of the remaining 6 cases, only two failed to be significant at level .05. They occurred for the MA(2) process with $n = 75$, and $h = 1$, in the comparisons with both FPE and AIC. These coincide with the only two cases where $\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{AIC_C}})\}$ exceeds either $\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{AIC}})\}$ or $\mathrm{Ave}\{\mathrm{MSE}(h, \hat{k}_{\mathrm{FPE}})\}$.

A crude description of the values of k selected by the various criteria is provided in Table 4 and 5, for the MA(2) and AR(4) processes, respectively. The tables present the average values of $\hat{k}_{\mathrm{AIC_C}}$, $\hat{k}_{\mathrm{AIC}}$ and $\hat{k}_{\mathrm{FPE}}$ together with k^* for lead times $h = 1, 2, 5$. In all cases, the average of $\hat{k}_{\mathrm{AIC_C}}$ is closer than the averages of $\hat{k}_{\mathrm{AIC}}$ or $\hat{k}_{\mathrm{FPE}}$ to the optimal value, k^*. It is notable that for the MA(2) process, k^* decreases quickly with h, suggesting the potential benefit of allowing the selection of k to depend on h. Correspondingly, the average values of $\hat{k}_{\mathrm{AIC_C}}$ are typically much less for $h > 1$ than for $h = 1$. Similar patterns hold for the average values of $\hat{k}_{\mathrm{AIC}}$ and $\hat{k}_{\mathrm{FPE}}$, although these are often much farther than the average value of $\hat{k}_{\mathrm{AIC_C}}$ from k^*. For the AR(4) process, k^* is, not surprisingly, almost always equal to 4, and once again the average values of $\hat{k}_{\mathrm{AIC_C}}$ are often substantially closer to k^* than the averages of $\hat{k}_{\mathrm{AIC}}$ or $\hat{k}_{\mathrm{FPE}}$.

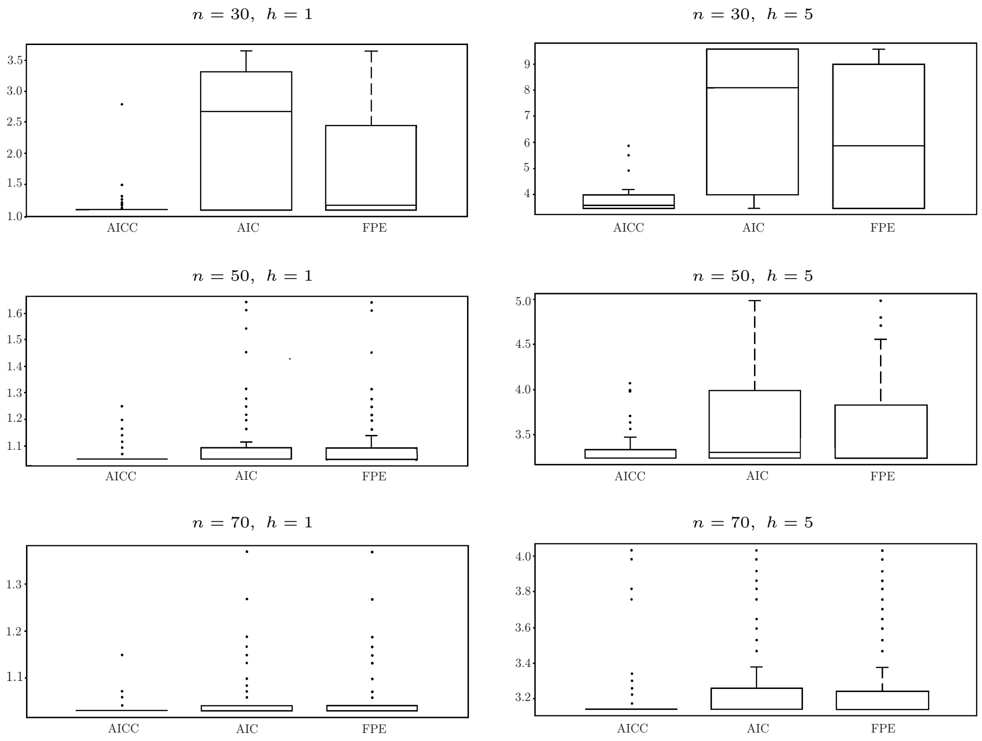

Figure 1. h-step prediction MSE for AR(2) process, normal errors.

Table 4. Averages of the selected values of k and the optimal value k^*, for MA(2) process.

	$n = 30$			$n = 50$			$n = 75$		
	h=1	h=2	h=5	h=1	h=2	h=5	h=1	h=2	h=5
Ave($\hat{k}_{\mathrm{AIC_C}}$)	5.37	1.38	1.19	8.07	2.15	1.39	11.16	4.24	2.17
Ave($\hat{k}_{\mathrm{AIC}}$)	15.72	12.04	11.56	13.58	6.30	3.58	14.52	7.24	4.14
Ave($\hat{k}_{\mathrm{FPE}}$)	13.29	9.56	8.62	12.37	5.42	3.34	14.36	7.10	4.02
k^*	8	0	0	8	0	0	11	3	0

Note: k^* is the minimizer of Ave$\{\mathrm{MSE}(h,\cdot)\}$.

Finally, we discuss the potential merits of Burg estimation over Yule-Walker estimation for selection of a linear predictor. Chen, Davis, Brockwell, and Bai (1993) have presented simulation evidence that for a finite-order autoregressive process, the Burg estimator can perform very poorly in small samples if the model order used in the estimator greatly exceeds the true order. They found that the Yule-Walker estimator performed much better at these high orders, and therefore suggested that it may be better to use Yule-Walker estimators than Burg estimators when the goal is model selection.

Table 5. Averages of the selected values of k and the optimal value k^*, for AR(4) process.

	$n=30$			$n=50$			$n=75$		
	$h=1$	$h=2$	$h=5$	$h=1$	$h=2$	$h=5$	$h=1$	$h=2$	$h=5$
Ave($\hat{k}_{\mathrm{AIC_C}}$)	4.24	4.34	3.92	4.48	4.84	4.57	4.44	4.42	4.88
Ave($\hat{k}_{\mathrm{AIC}}$)	10.25	10.99	13.14	6.85	7.34	8.05	4.91	5.51	5.96
Ave($\hat{k}_{\mathrm{FPE}}$)	7.80	8.77	10.82	6.52	7.13	7.11	4.81	5.35	5.74
k^*	4	4	3	4	4	4	4	4	4

Note: k^* is the minimizer of Ave$\{\mathrm{MSE}(h,\cdot)\}$.

If the $\mathrm{AIC_C}$ criterion is used, however, then it will be rare that a model order which is much too large will be selected. So given an autoregressive process for which, at the true order, the Burg estimators perform much better than the Yule-Walker estimators (an example is provided by our AR(4) model; See Tjϕstheim and Paulsen (1983) for a more complete discussion), we would expect that the predictors selected by $\mathrm{AIC_C}$ using the Burg method will perform well compared to predictors selected by $\mathrm{AIC_C}$ using the Yule-Walker method.

Using the same one hundred simulated realizations of the AR(4) model as above, and using Yule-Walker estimators, we obtained the following values for Ave$\{\mathrm{MSE}(h,\hat{k}_{\mathrm{AIC_C}})\}$ ($h = 1, 2, 5$): 26.7, 104.7, 247.1 (with $n = 30$), and 20.3, 80.0, 184.7 (with $n = 50$). Comparison with the corresponding values of Ave$\{\mathrm{MSE}(h,\hat{k}_{\mathrm{AIC_C}})\}$ in Table 2, which are based on Burg estimates, reveals that for the AR(4) model studied here, if $\mathrm{AIC_C}$ is used for selection of a multi-step linear predictor, Burg estimates produce vastly superior selected predictors than Yule-Walker estimates. This is made possible by the well-known superiority of Burg over Yule-Walker estimators for correctly-identified AR models with characteristic roots close to the unit cricle. For the given AR(4) model, the characteristic polynomial $P(z) = 1 + \sum_{j=1}^{4} \alpha_j z^j$ has roots at $z = 0.7862 \pm 0.65i$, and $z = 0.65 \pm 0.7859i$, where $i = \sqrt{-1}$.

Acknowledgement

The research of the first author was supported by NSF Grant DMS-92-03347, and the research of the second author was supported by NSF Grants DMS-95-10511 and DMS-92-02759. The authors thank the Referee and Associate Editor for helpful comments.

Appendix

Suppose that model (6) holds with Gaussian innovations, i.e.,

$$x_{t+h} + a_h(h,k)x_t + \cdots + a_{h+k-1}(h,k)x_{t-k+1} = \varepsilon_{t+h}, \tag{A.1}$$

where $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2(h,k))$. We show here that if $\hat{a}(h,k)$ and $\hat{\sigma}^2(h,k)$ are the multistep linear regression estimators defined below, then the final k entries of $\sqrt{n}(\hat{a}(h,k) - a(h,k))$ are asymptotically distributed as $N(0, \sigma^2(h,k)R_k^{-1})$, and $n\hat{\sigma}^2(h,k)/\sigma^2(h,k)$ is asymptotically distributed as χ^2_{n-k}, independently of $\hat{a}(h,k)$.

The multistep linear regression estimators are defined as follows. Denote the final k entries of $a(h,k)$ by ϕ. Define $Y = (x_1, \ldots, x_n)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, and

$$X = \begin{bmatrix} x_{1-h} & x_{-h} & \cdots & x_{2-h-k} \\ x_{2-h} & x_{1-h} & \cdots & x_{3-h-k} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n-h} & x_{n-h-1} & \cdots & x_{n-h-k+1} \end{bmatrix}.$$

Then, the model (A.1) for the data Y can be expressed as

$$Y = -X\phi + \varepsilon, \tag{A.2}$$

where $\varepsilon \sim N(0, \sigma^2(h,k)I_{n\times n})$, and $I_{n\times n}$ is the $n \times n$ identity matrix. Based on Equation (A.2), the multistep linear regression estimators of ϕ and of the scale parameter $\sigma^2(h,k)$ are given by $\hat{\phi} = -(X'X)^{-1}X'Y$, and $\hat{\sigma}^2 = S(\hat{\phi})/n$, where $S(\hat{\phi}) = (Y + X\hat{\phi})'(Y + X\hat{\phi})$. The multistep linear regression estimators are a specific version of multistep least squares estimators, since they minimize the h-step residual sum of squares. By a straightforward modification of Brockwell and Davis (1991), p. 262 Proposition 8.10.1, we have $\sqrt{n}(\hat{\phi} - \phi) \to N(0, \sigma^2(h,k)R_k^{-1})$. From Wei (1990), p. 354 and Priestley (1981), p. 366, it follows that $n\hat{\sigma}^2$ is asymptotically distributed as $\sigma^2(h,k)\chi^2_{n-k}$, independently of $\hat{\phi}$.

References

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

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

Beamish, N. and Priestley, M. B. (1981). A study of autoregressive and window spectral estimation. *Appl. Statist.* **30**, 41-58.

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

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods,* Second Edition. Springer-Verlag, New York.

Burg, J. P. (1978.) A new analysis technique for time series data. In *Modern Spectrum Analysis* (Edited by D. G. Childers), 42-48. IEEE Press, New York.

Chen, C., Davis, R. A., Brockwell, P. J. and Bai, Z. D. (1993). Order determination for autoregressive processes using resampling methods. *Statist. Sinica* **3**, 481-500.

Findley, D. F. (1983). On using a different time series forecasting model for each forecast lead. Statistical Research Division Report No. CENSUS/SRD/RR/-83/06, Bureau of Census, Washington, D. C.

Hainz, G. (1995). The asymptotic properties of Burg estimators. Preprint Univ. of Heidelberg.

Haykin, S. (ed.) (1983). Nonlinear methods of spectral analysis. *Topics in Applied Physics,* Vol. 34, Second Ed., Springer-Verlag, New York.

Hurvich, C. M. (1987). Automatic selection of a linear predictor through frequency domain cross-validation. *Comm. Statist. Theory Mehtods* **16**, 3199-3234.

Hurich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

Hurvich, C. M. and Tsai, C.-L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrika* **78**, 499-509.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* John Wiley, New York.

Priestley, M. B. (1981). *Spectral Analysis And Time Series.* Academic Press, New York.

Shaman, P. (1983). Properties of estimates of the mean square error of prediction in autoregressive models. In *Studies in Econometrics, Time Series and Multivariate Statistics* (Edited by S. Karlin, T. Amemiya, L. A. Goodman), 331-342. Academic Press, New York.

Shaman, P. and Stine, R. A. (1988). The bias of autoregressive coefficient estimators. *J. Amer. Statist. Assoc.* **83**, 842-848.

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

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

Tiao, G. C. and Tsay, R. S. (1994). Some advances in non-linear and adaptive modelling in time series. *J. Forecasting* **13**, 109-140.

Tjϕstheim, D. and Paulsen, J. (1983). Bias of some commonly-used time series estimates. *Biometrika* **70**, 389-399.

Wei, W. S. (1990). *Time Series Analysis.* Addison-Wesley, New York.

Department of Statistics and Operations Research, New York University, 44 West Fourth Street, New York, NY 10012, U.S.A.

Graduate School of Management, University of California, Davis CA 95616, U.S.A.

(Received May 1994; accepted November 1995)