Statistica Sinica 3(1993), 481-500

ORDER DETERMINATION FOR AUTOREGRESSIVE PROCESSES USING RESAMPLING METHODS

Changhua Chen, Richard A. Davis, Peter J. Brockwell and Zhi Dong Bai*

*Colorado State University and Temple University**

Abstract: Let $X_1, \dots, X_n$ be observations from an AR(p) model with unknown order p. A resampling procedure is proposed for estimating the order p. The classical criteria, such as AIC and BIC, estimate the order p as the minimizer of the function

$$\delta(k) = \ln(\hat{\sigma}_k^2) + kC_n$$

where n is the sample size, k is the order of the fitted model, $\hat{\sigma}_k^2$ is an estimate of the white noise variance, and C_n is a sequence of specified constants (for AIC, $C_n = 2/n$, for Hannan and Quinn's modification of BIC, $C_n = 2(\ln \ln n)/n$). Often, the traditional order selectors overfit or underfit the model for a given realization. To overcome this defect, a resampling scheme is proposed to estimate a suitable penalty factor C_n. Conditional on the data, this procedure produces a consistent estimate of p. Simulation results support the effectiveness of this procedure when compared with some of the traditional order selection criteria for both Gaussian and a range of non-Gaussian processes. A discussion of the merits of Yule-Walker estimation relative to Burg and maximum likelihood estimation for order determination is also given.

Key words and phrases: Autoregressive processes, order determination, AIC, Yule-Walker estimation, resampling.

1. Introduction

In this paper, we propose a resampling scheme for model selection of an autoregressive time series. Let $\{X_t\}$ be the causal AR(p) process satisfying the difference equations

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t, \tag{1.1}$$

where $\{Z_t\} \sim \text{IID}(0, \sigma^2)$ (i.e. $\{Z_t\}$ is an independent and identically distributed sequence of random variables with mean zero and variance σ^2), and where $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$ satisfies the causality condition $\phi(z) \neq 0$ for $|z| \leq 1$. The unknown parameters for this model are then $p, \phi_1, \dots, \phi_p$, and σ^2.

Most order selection criteria for autoregressive processes estimate the order p by minimizing an objective function of the form

$$\delta(k) = \log \hat{\sigma}_k^2 + C(n, k), \tag{1.2}$$

where n is the sample size, k is the order of the candidate AR model, $\hat{\sigma}_k^2$ is an estimate of the innovation variance for the fitted candidate model, and $C(n,k)$ is a sequence of constants depending on n and k only. Typical values for the penalty factor in (1.2) are:

$$\begin{aligned} C(n,k) &= \frac{2k}{n}, && \text{AIC (Akaike (1969))} \\ C(n,k) &= \frac{2(k+1)}{n-(k+2)}, && \text{AICC (Hurvich and Tsai (1989))} \\ C(n,k) &= \frac{(\log n)k}{n}, && \text{(Schwartz (1978))} \\ C(n,k) &= \frac{2c(\log\log n)k}{n}, && \text{(Hannan and Quinn (1979)).} \end{aligned}$$

The latter two choices lead to a consistent order selection procedure in the sense that the minimizer of $\delta(k)$ converges to p with probability one. Neither the AIC nor the AICC share this property. The AIC and AICC were designed to be approximately unbiased estimates of the Kullback-Leibler index when the innovation distribution is Gaussian; and are asymptotically efficient in the sense of Shibata (1980) (see also Brockwell and Davis (1991, Section 9.3)). (For both the AIC and AICC, one typically replaces the term $\log\hat{\sigma}^2$ in (1.2) by $-\frac{2}{n}\log L_X(\hat{\phi}_1,\ldots,\hat{\phi}_k,\hat{\sigma}^2)$ where $L_X(\cdot)$ is the Gaussian likelihood and $\hat{\phi}_1,\ldots,\hat{\phi}_k,\hat{\sigma}^2$ are the maximum likelihood estimators of the parameters when fitting an AR(k) model to the data.)

A shortcoming of all these order selection criteria is that the sequence of constants $C(n,k)$ completely ignores other potentially important information contained in the distribution of the process. In our procedure, we obtain such information from a resampling scheme based on the original data and use it to obtain a better penalty factor. The procedure may be described briefly as follows: The innovation distribution is estimated using the residuals from fitting a high order AR model to the original data, $X_1,\ldots,X_n$. Resampling from the estimated innovations distribution, we generate sample realizations, $Y_1^{(j)},\ldots,Y_n^{(j)}$ of AR(j) processes with $j=1,\ldots,K_1$ where K_1 is some specified constant. (The parameters used to generate $\{Y_t^{(j)}\}$ may be, but are not necessarily, those obtained by fitting an AR(j) model to the original data, $X_1,\ldots,X_n$.) The constants $C(n,k)$ are then chosen in such a way that the objective function $\delta(k)$ has a unique minimum at $k=j$ when applied to each of the generated series $\{Y_t^{(j)}\}$, $j=1,\ldots,K_1$.

In Section 2, we give a more detailed account of how to choose $C(n,k)$. It is also shown that conditional on the observed data, our method produces a consistent estimate of the order. In Section 3, we discuss implementation of our method and compare its performance with other order selection criteria. In Section 4, we discuss the merits of Yule-Walker estimation for order selection and demonstrate, with an example, the extreme variability of the coefficient estimators and the underestimation of the white noise variance which occur when highly overparameterized models are fitted by Burg (Burg (1967)) or maximum likelihood estimation.

2. Choosing a Penalty Factor

Let $\{X_t\}$ be a zero-mean AR(p) process satisfying equations (1.1). For each $k = 1, 2, \ldots$, let $\hat{X}_{k+1} = \phi_{k1} X_k + \cdots + \phi_{kk} X_1$ denote the best linear predictor of X_{k+1} in terms of $X_k, \ldots, X_1$. If $\gamma(h)$ is the autocovariance function of the process, then the coefficient vector $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kk})'$ and mean square error of prediction $\sigma_k^2 = E(X_{k+1} - \hat{X}_{k+1})^2$ are given by the Yule-Walker equations

$$\boldsymbol{\phi}_k = \Gamma_k^{-1} \boldsymbol{\gamma}_k \tag{2.1}$$

$$\sigma_k^2 = \gamma(0) - \boldsymbol{\phi}_k' \boldsymbol{\gamma}_k = \gamma(0) \prod_{i=1}^{k} (1 - \phi_{ii}^2) \tag{2.2}$$

where $\Gamma_k = [\gamma(i-j)]_{i,j=1}^{k}$ and $\boldsymbol{\gamma}_k = (\gamma(1), \ldots, \gamma(k))'$ (see Brockwell and Davis (1991, p.239)). The coefficient ϕ_{ii} is also the partial autocorrelation of $\{X_t\}$ at lag i. Since $\{X_t\}$ is assumed to be an AR(p) process ($\phi_p \neq 0$ in (1.1)), we have the elementary properties:

$$\boldsymbol{\phi}_k = (\phi_1, \ldots, \phi_p, 0, \ldots, 0)', \quad \text{for } k \geq p, \tag{2.3}$$

$$\sigma_k^2 = \gamma(0) \prod_{i=1}^{p} (1 - \phi_{ii}^2) = \sigma^2, \quad \text{for } k \geq p, \tag{2.4}$$

and

$$\sigma_k^2 > \sigma^2, \quad \text{for } k < p. \tag{2.5}$$

The Yule-Walker estimates $\hat{\boldsymbol{\phi}}_k$ and $\hat{\sigma}_k^2$ are obtained by replacing $\gamma(h)$ with the sample autocovariance function $\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} X_t X_{t+|h|}$ on the right hand side of equations (2.1) and (2.2).

Throughout the remainder of this paper we take the penalty factor in (1.2) to be linear in k ,i.e. $C(n,k) = C_n k$ where C_n is a sequence of numbers to be specified. To emphasize the dependence of $\delta(\cdot)$ on the penalty factor, we write

$$\delta(k, \alpha) = \log \hat{\sigma}_k^2 + k\alpha \tag{2.6}$$

so that $\delta(k) = \delta(k, C_n)$. The following proposition plays a key role in determining the penalty factor for our proposed scheme.

The result is motivated by the following observation. If we define

$$\mu_{jk} = \frac{\log\sigma_j^2 - \log\sigma_k^2}{k-j}, \quad k > j, \tag{2.7}$$

then $\mu_{jk} = 0$ if $j > p$. Also, from (2.2) we can write

$$\mu_{jk} = \frac{1}{j-k}\sum_{i=j+1}^{k}\log(1-\phi_{ii}^2), \tag{2.8}$$

so that $\mu_{jk} > 0$ if $\phi_{ii}^2 > 0$ for some $i \in (j,k]$. In Proposition 2.1 we replace μ_{jk} by a corresponding estimator and make use of the strong consistency of the Yule-Walker estimators.

Proposition 2.1. *Suppose $\{X_t\}$ is the AR(p) process defined by* (1.1) *and let $K \geq p$ be a fixed constant. Define, for observations $X_1,\ldots,X_n$,*

$$\alpha_n = \begin{cases} 0, & \text{if } p = K, \\ \max_{p<j\leq K} \frac{\log\hat\sigma_p^2 - \log\hat\sigma_j^2}{j-p}, & \text{if } p < K, \end{cases}$$

$$\beta_n = \begin{cases} \infty, & \text{if } p = 0, \\ \min_{0\leq j<p} \frac{\log\hat\sigma_j^2 - \log\hat\sigma_p^2}{p-j}, & \text{if } p > 0, \end{cases}$$

where $\hat\sigma_0^2 = \hat\gamma(0)$. Then, as $n \to \infty$,
(a) *$\alpha_n \to 0$ and $\beta_n \to b \geq 0$* a.s. *(If $\min_{1\leq j\leq p}|\phi_{jj}| > 0$, then $b > 0$.)*
(b) *If $\alpha_n \leq \beta_n$, then for any $C_n \in [\alpha_n, \beta_n]$,*

$$\delta(p, C_n) = \min_{0\leq j\leq K}\{\delta(j, C_n)\}.$$

Proof. (a) Since the Yule-Walker estimates are strongly consistent, (2.3)–(2.5) imply that for $p < K$

$$\alpha_n \xrightarrow{\text{a.s.}} \max_{p<j\leq K}\frac{\log\sigma_p^2 - \log\sigma_j^2}{j-p} = 0$$

and for $p > 0$

$$\beta_n \xrightarrow{\text{a.s.}} b = \min_{0\leq j<p}\frac{\log\sigma_j^2 - \log\sigma_p^2}{p-j} \geq 0.$$

The assertions for $p = K$ and $p = 0$ are trivial.

(b) For any j ($p < j \leq K$) and all $C_n \geq \alpha_n$, we have

$$C_n \geq \alpha_n \geq \frac{\log\hat\sigma_p^2 - \log\hat\sigma_j^2}{j-p},$$

so that

$$\log \hat{\sigma}_j^2 + jC_n \geq \log \hat{\sigma}_p^2 + pC_n. \tag{2.9}$$

On the other hand, for $j < p$ and $C_n \leq \beta_n$, we have

$$C_n \leq \beta_n \leq \frac{\log \hat{\sigma}_j^2 - \log \hat{\sigma}_p^2}{p - j}$$

and hence (2.9) is also valid for $j < p$. This proves (b).

We call C_n a *correct* penalty factor if the estimated order, defined as the minimizer of $\delta(\cdot, C_n)$, is equal to the true order. If the order of the model is known, then the above proposition tells us how to choose a correct penalty factor – in fact, *any* value of C_n in the interval $[\alpha_n, \beta_n]$ will result in selection of the correct order. Our task now is to determine a suitable penalty factor when the order is unknown and to show that, at least for n large enough, the set of suitable penalty factors is not empty.

For any fixed k, we shall refer to the AR(k) model,

$$X_t - \hat{\phi}_{k1}X_{t-1} - \cdots - \hat{\phi}_{kk}X_{t-k} = Z_t, \quad \{Z_t\} \sim \text{IID}(0, \hat{\sigma}_k^2), \tag{2.10}$$

where $\hat{\phi}_k$ and $\hat{\sigma}_k^2$ are the Yule-Walker estimators defined above, as the candidate AR(k) model for the data $X_1, \ldots, X_n$. We can compute the Yule-Walker estimators from the data and resample from the estimated noise distribution to generate a *test series* $Y_1^{(k)}, \ldots, Y_n^{(k)}$ from the model (2.10) and then generate the interval $[\alpha_n^{(k)}, \beta_n^{(k)}]$ for the test series by applying Proposition 2.1. This procedure is carried out for $k = 0, \ldots, K_1$ where K_1 is a specified constant whose value is discussed in Remarks 2 and 3 at the end of this section. In this way we obtain intervals $I_n^{(k)} = [\alpha_n^{(k)}, \beta_n^{(k)}]$, $k = 0, \ldots, K_1$. If the intersection of these intervals, $I_n = \cap_{k=0}^{K_1} I_n^{(k)}$, is nonempty, a value of C_n in I_n is chosen, which is a correct penalty factor for the $K_1 + 1$ test series. The asymptotic properties of I_n are examined in the following proposition.

Proposition 2.2. *Suppose the conditions of Proposition 2.1 are satisfied and that $EZ_t^4 < \infty$. Define*

$$\hat{Z}_t = X_t - \hat{\phi}_{K1}X_{t-1} - \cdots - \hat{\phi}_{KK}X_{t-K}, \quad t = K + 1, \ldots, n.$$

Let $K_1 \leq p$ be a non-negative integer, and for each $k = 0, \ldots, K_1$, let $\{Y_1^{(k)}, \ldots, Y_n^{(k)}\}$ be a sequence of observations from the AR(k) model

$$Y_t^{(k)} - \hat{\phi}_{k1}Y_{t-1}^{(k)} - \cdots - \hat{\phi}_{kk}Y_{t-k}^{(k)} = Z_t^*$$

where $\{Z_t^*\}$ is an iid sequence whose distribution is the empirical distribution (corrected to have mean 0) of $\{\hat{Z}_t\}$. (For $k = 0, Y_t^{(0)} = Z_t^*$.) For each $k = 0, \ldots, K_1$, let $I_n^{(k)} = [\alpha_n^{(k)}, \beta_n^{(k)}]$ denote the interval obtained when Proposition 2.1 (with $p = k$) is applied to the $AR(k)$ series $\{Y_t^{(k)}\}$. Then for almost all sample sequences of $\{X_t\}$

(a) $\alpha_n = \max_{0\le k\le K_1}\{\alpha_n^{(k)}\} \xrightarrow{P_n} 0$,

(b) $\beta_n = \min_{0\le k\le K_1}\{\beta_n^{(k)}\} \xrightarrow{P_n} b \ge 0$,

where $\xrightarrow{P_n}$ denotes convergence in probability conditional on $X_1, \ldots, X_n$. In particular, $I_n = \cap_{k=0}^{K_1} I_n^{(k)}$ $(= [\alpha_n, \beta_n]$ if $\alpha_n \le \beta_n)$ converges in conditional probability to a nonempty set.

Proof. It suffices to show that

$$\alpha_n^{(k)} \xrightarrow{P_n} 0 \text{ and } \beta_n^{(k)} \xrightarrow{P_n} b_k \ge 0. \tag{2.11}$$

We first show that the conditional variance of the sample acf $\hat{\gamma}^*(h) = \frac{1}{n}\sum_{j=1}^n Y_t^{(k)} Y_{t+h}^{(k)}$, denoted by $\mathrm{Var}_n(\hat{\gamma}^*(h))$, converges to 0 a.s. Since the Yule-Walker estimates always produce a causal model (see Brockwell and Davis (1991, Problem 8.3)), $\{Y_t^{(k)}\}$ has the representation

$$Y_t^{(k)} = \sum_{j=0}^{\infty} \hat{\psi}_j Z_{t-j}^*$$

where $\hat{\psi}_0, \hat{\psi}_1, \ldots$, are the coefficients in the power series expansion of $(1 - \hat{\phi}_{k1}z - \cdots - \hat{\phi}_{kk}z^k)^{-1}$ on $|z| \le 1$. The conditional mean of $\hat{\gamma}^*(h)$ is given by

$$\gamma^*(h) = E_n\left(Y_t^{(k)} Y_{t+h}^{(k)}\right) = \sum_{j=0}^{\infty} \hat{\psi}_j \hat{\psi}_{j+h} \sigma^{*2} \tag{2.12}$$

where $E_n(\cdot)$ denotes expectation relative to P_n and σ^{*2} is the sample variance of the $\hat{Z}_t$'s. The strong consistency of the Yule-Walker estimates implies that for almost all sample paths, $\hat{\phi}_k \to \phi_k$ and $\hat{\sigma}_k^2 \to \sigma_k^2$. It follows that $\hat{\psi}_j \to \psi_j$ as $n \to \infty$ ($\psi_0, \psi_1, \ldots$ are the coefficients in the expansion of $(1 - \phi_{k1}z - \cdots - \phi_{kk}z^k)^{-1}$), and in particular, that there exist constants $C > 0$ and $r < 1$ depending on the sample path such that

$$|\hat{\psi}_j| \le Cr^j \tag{2.13}$$

for $j = 0, 1, \ldots$ and all n large. Similarly, one can show that $\sigma^{*2} \to \sigma_k^2$ which, combined with (2.12) and (2.13), yields

$$\gamma^*(h) \to \sigma_k^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \tag{2.14}$$

as $n \to \infty$. Now, using the above relations and Equation (7.3.5) in Brockwell and Davis (1991) with $\gamma(\cdot)$ and ψ_j replaced by $\gamma^*(\cdot)$ and $\hat{\psi}_j$ respectively, we find that

$$\limsup_{n\to\infty} n\mathrm{Var}_n(\hat{\gamma}^*(h)) < \infty.$$

Thus,

$$\mathrm{Var}_n(\hat{\gamma}^*(h)) \to 0$$

and hence, by (2.14)

$$\hat{\gamma}^*(h) \xrightarrow{P_n} \sigma_k^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h}.$$

The weak consistency of the sample acf implies that the Yule-Walker estimates, in fitting an AR(j) model to the $Y_t^{(k)}$ data, are also weakly consistent relative to P_n, from which (2.11) follows as in the proof of Proposition 2.1 (a). Consequently,

$$\max_{0\le k\le K_1} \{\alpha_n^{(k)}\} \xrightarrow{P_n} 0$$

and

$$\min_{0\le k\le K_1} \{\beta_n^{(k)}\} \xrightarrow{P_n} \min_{0\le k\le K_1} \{b_k\} \ge 0,$$

as required.

Theorem 2.1. *Suppose that the conditions of Proposition 2.2 are satisfied and* $\min_{1\le j\le K_1} |\phi_{jj}| > 0$. *Define*

$$C_n = \begin{cases} \alpha_n + \frac{c(\log n)\beta_n}{n}, & \text{if } \alpha_n < \beta_n, \\ \frac{(\log n)\beta_n}{n}, & \text{otherwise}, \end{cases} \tag{2.15}$$

where $c > 0$ is a constant such that $C_n \in I_n$ if $\alpha_n < \beta_n$. If $\hat{p}$ is the minimizer of $\delta(k, C_n)$ for $0 \le k \le K$, then for almost all sample sequences of $\{X_t\}$,

$$\hat{p} \xrightarrow{P_n} p$$

as $n \to \infty$.

Proof. From Proposition 2.2, we see that for almost all sample paths,

$$C_n \xrightarrow{P_n} 0 \quad \text{and} \quad \frac{nC_n}{\log\log n} \xrightarrow{P_n} \infty.$$

The consistency of $\hat{p}$ now follows directly from the argument given in Hannan and Quinn (1979, p.191).

Remark 1. Of course any sequence of constants $\{C_n\}$ satisfying $C_n \to 0$ and $\frac{nC_n}{\log\log n} \to const > 2$ will produce a consistent estimate of p. The point of the resampling procedure is to produce a data-dependent sequence $\{C_n\}$ which gives consistent estimates of p and at the same time performs optimally on each of the test sequences.

Remark 2. Ideally, we would like to optimize over as large a class of test models as possible in order to ensure that a model close to the true model is included in the test set. This requires a large K_1. However, Theorem 2.1 requires that $K_1 \leq p$ and $\min_{1\leq j\leq K_1} |\phi_{jj}| > 0$. Since p and ϕ are unknown we cannot be certain that these restrictions hold in practice. If we start with a K_1 which happens to be bigger than p, then it is likely that the set I_n will be empty. In this case, we reduce the value of K_1 by 1, continuing to do so until a nonempty I_n is obtained.

Remark 3. In the statement of Proposition 2.2, we have assumed that the test series $\{Y_t^{(k)}\}$ has been generated from the candidate model. The conclusion of the proposition remains unchanged, if instead, $\{Y_1^{(k)},\ldots,Y_n^{(k)}\}$ is generated from the model

$$Y_t^{(k)} - a_{k1}Y_{t-1}^{(k)} - \cdots - a_{kk}Y_{t-k}^{(k)} = Z_t^*,$$

where the coefficient vector $(a_{k1},\ldots,a_{kk})'$ is nonrandom. The advantages of this modification of the procedure are two-fold: (i) K_1 can be any non-negative integer less than or equal to K and (ii) the parameter vectors for generating the test series can be chosen in such a way as to increase the probability of the event $\alpha_n < \beta_n$.

3. Implementation and Simulation

Our order selection procedure may be implemented as follows. Assume that $X_1,\ldots,X_n$ are observations from an AR(p) process defined in (1.1) and that $K \geq p$ is a fixed constant.

Step 1. Compute the Yule-Walker estimates $\hat\phi_{K1},\ldots,\hat\phi_{KK},\hat\sigma_K^2$ for the observed data, $\{X_t\}_{t=1}^n$.

Step 2. Compute the residuals,

$$\hat Z_t = X_t - \hat\phi_{K1}X_{t-1} - \cdots - \hat\phi_{KK}X_{t-K}$$

for $t = K+1,\ldots,n$. Center the residuals by subtracting the sample mean $\frac{1}{n-K}\sum_{t=K+1}^n \hat Z_t$. Let $\hat F_n$ denote the empirical distribution function of the centered residuals.

Step 3. Set $\hat\sigma_0^2 = \hat\gamma(0)$ and for $k = 1,\ldots,K$ compute the Yule-Walker estimates $\hat\phi_k, \hat\sigma_k^2$ from the observed data, $\{X_t\}_{t=1}^n$.

Step 4. Choose a positive integer $K_1 \le K$. For $k = 0, \ldots, K_1$ generate observations $Y_1^{(k)}, \ldots, Y_n^{(k)}$ from the model

$$Y_t^{(k)} - \hat{\phi}_{k1} Y_{t-1}^{(k)} - \cdots - \hat{\phi}_{kk} Y_{t-k}^{(k)} = Z_t^*$$

where $\{Z_t^*\}$ is an iid sequence with distribution function $\hat{F}_n$. The case $k = 0$ corresponds to $Y_t^{(0)} = Z_t^*$. Such processes may be generated by setting k consecutive observations in the distant past equal to 0, i.e. for m a large negative integer, put

$$Y_t^{(k)} = \begin{cases} 0, & \text{for } t < m, \\ \hat{\phi}_{k1} Y_{t-1}^{(k)} + \cdots + \hat{\phi}_{kk} Y_{t-k}^{(k)} + Z_t^*, & \text{for } t = m, \ldots, n. \end{cases}$$

Step 5. For $k = 0, \ldots, K_1$, compute the Yule-Walker estimate of the innovation variance in fitting an AR(j) model to $\{Y_t^{(k)}\}_{t=1}^n$ for $j = 0, \ldots, K$. Denote this estimate by $\hat{\sigma}_j^2(k)$.

Step 6. For $k = 0, \ldots, K_1$ compute

$$\alpha_n^{(k)} = \begin{cases} 0, & \text{if } k = K, \\ \max_{k<j\le K} \frac{\log \hat{\sigma}_k^2(k) - \log \hat{\sigma}_j^2(k)}{j-k}, & \text{if } k < K, \end{cases}$$

and

$$\beta_n^{(k)} = \begin{cases} \infty, & \text{if } k = 0, \\ \min_{0\le j<k} \frac{\log \hat{\sigma}_j^2(k) - \log \hat{\sigma}_k^2(k)}{k-j}, & \text{if } k > 0. \end{cases}$$

Step 7. Compute

$$\alpha_n = \max_{0\le k\le K_1} \{\alpha_n^{(k)}\}$$

and

$$\beta_n = \min_{0\le k\le K_1} \{\beta_n^{(k)}\}.$$

Step 8. If $\alpha_n < \beta_n$ set

$$C_n = \alpha_n + \frac{c(\log n)\beta_n}{n}$$

where $c > 0$ is such that $C_n \le \beta_n$. If $\alpha_n \ge \beta_n$, then reduce the value of K_1 by 1 and return to Step 7.

Step 9. The estimated order $\hat{p}$ is defined to be the minimizer of

$$\delta(k, C_n) = \log \hat{\sigma}_k^2 + kC_n$$

for $0 \leq k \leq K$.

Due to sampling error, it is often necessary to obtain many replicates of the test series $\{Y_t^{(k)}\}_{t=1}^n$ in Step 4. The computed values of $\alpha_n^{(k)}$ and $\beta_n^{(k)}$ in Step 5 are then replaced by their respective averages over the replications.

We also considered the modification of this procedure alluded to earlier in Remark 3. The only difference in the modification is that the test series $\{Y_t^{(k)}\}$ in Step 4 are generated with $\hat{\phi}_k$ replaced by a prespecified sequence of parameter vectors $(a_{k1}, \ldots, a_{kk})'$, $k = 1, \ldots, K_1$.

We compared the above procedure and its modification with 4 other well known order selection criteria defined by minimization of the following objective functions:

$$\begin{array}{ll}
\text{AIC} & -\frac{2}{n} \log L_X(\hat{\phi}, \hat{\sigma}_k^2) + \frac{2k}{n} \\
\text{AICC} & -\frac{2}{n} \log L_X(\hat{\phi}, \hat{\sigma}_k^2) + \frac{2(k+1)}{n-(k+2)} \\
\text{H\&Q} & \log \hat{\sigma}_k^2 + \frac{(2 \log \log n)k}{n} \\
\text{BIC} & (n-k)\log[n\hat{\sigma}_k^2/(n-k)] + k \log\left[\left(\sum_{t=1}^{n} X_t^2 - n\hat{\sigma}_k^2\right)\Big/k\right]
\end{array}$$

where L_X is the Gaussian likelihood based on the observed values. (The expression for BIC is taken from Akaike (1978, p.18), with $p+q$ replaced by k.)

We generated 100 sample paths of various lengths from each of the following AR models:

$$X_t = .40X_{t-1} + Z_t \tag{3.1}$$
$$X_t = 1.4X_{t-1} - .49X_{t-2} + Z_t \tag{3.2}$$
$$X_t = .48X_{t-1} - .34X_{t-2} + .38X_{t-3} - .48X_{t-4} + .42X_{t-5} + Z_t \tag{3.3}$$
$$\begin{aligned} X_t = {} & .48X_{t-1} + .30X_{t-2} - .30X_{t-3} - .38X_{t-4} + .32X_{t-5} \\ & -.51X_{t-6} - .30X_{t-7} + .38X_{t-8} + .43X_{t-9} - .56X_{t-10} + Z_t \end{aligned} \tag{3.4}$$

where $\{Z_t\}$ is an iid sequence of $N(0,1)$ random variables. The frequencies of the estimated orders for each of the 6 criteria are summarized in Tables 1–4 (the method described above and its modification are listed as DC and MDC respectively). In all of our simulations we took $K = 20$, $K_1 = 2$, and $c = .6$. The $\alpha_n^{(k)}$ and $\beta_n^{(k)}$ were computed as an average based on 50 replicates of the test series. For the modified procedure (MDC), the parameter vectors were $a_{11} = .8$ and $(a_{21}, a_{22}) = (.5, -.96)$. From the defining equations for $\alpha_n^{(k)}$ and $\beta_n^{(k)}$, we see that a large partial autocorrelation at lag k increases the likelihood that

$\alpha_n^{(k)} < \beta_n^{(k)}$ for $k = 0, \ldots, K_1$, and hence that we obtain a well defined set I_n from which to choose C_n. We found that in terms of the frequency of correct order selection, MDC is relatively insensitive to the values of a_{11}, a_{21}, a_{22} provided $|a_{11}|$ and $|a_{22}|$ are not too close to zero, say greater than .6. The frequency distribution of the selected order is more dependent on the choice of coefficients with larger values of $|a_{11}|$ and $|a_{22}|$ giving more low order models.

Table 1. Frequencies of estimated order in 100 replications from the AR(1) model given by (3.1) with sample sizes 50 and 100 (in parentheses)

	estimated order							
Criterion	0	1	2	3	4	5	6 – 10	11 – 20
AIC	33 (23)	48 (57)	9 (9)	3 (1)	2 (5)	2 (1)	1 (4)	2 (0)
AICC	34 (23)	52 (60)	9 (9)	2 (3)	1 (1)	0 (0)	1 (4)	1 (0)
H&Q	19 (1)	64 (87)	11 (6)	2 (5)	3 (1)	0 (0)	0 (0)	1 (0)
BIC*	0 (0)	51 (65)	7 (5)	1 (1)	2 (1)	3 (1)	10 (9)	16 (11)
DC	15 (1)	59 (78)	12 (8)	3 (4)	4 (3)	3 (2)	2 (4)	2 (0)
MDC	33 (6)	64 (93)	3 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)

(* row totals are less than 100 since for some realizations, BIC is not calculable due to a negative value of $\sum_{t=1}^{n} X_t^2 - n\tilde{\sigma}_k^2$.)

Table 2. Frequencies of estimated order in 100 replications from the AR(2) model given by (3.2) with sample sizes 50 and 100 (in parentheses)

	estimated order							
Criterion	0	1	2	3	4	5	6 – 10	11 – 20
AIC	0 (0)	13 (0)	74 (77)	7 (7)	3 (7)	1 (2)	1 (5)	1 (2)
AICC	0 (0)	13 (0)	76 (81)	8 (7)	3 (7)	0 (3)	0 (2)	0 (0)
H&Q	0 (0)	9 (0)	84 (95)	5 (3)	2 (2)	0 (0)	0 (0)	0 (0)
BIC	0 (0)	25 (1)	73 (96)	2 (2)	0 (1)	0 (0)	0 (0)	0 (0)
DC	0 (0)	8 (0)	80 (94)	7 (3)	4 (1)	1 (2)	0 (0)	0 (0)
MDC	0 (0)	16 (1)	82 (98)	2 (1)	0 (0)	0 (0)	0 (0)	0 (0)

Although DC and MDC are computationally intensive the cpu time required to run these procedures is certainly not a limiting factor in most applications. For example in a sample of size 100, DC and MDC take approximately 1.31 and 1.28 seconds, respectively, to run on a SPARC IPC workstation. This includes generating the 100 data values and generating 50 replicates of the test series in Step 4 of the procedure. In contrast, AIC takes only .1 seconds to identify the order of the model from a sample of size 100.

Table 3. Frequencies of estimated order in 100 replications from the AR(5) model given by (3.3) with sample sizes 100 and 200 (in parentheses)

	estimated order						
Criterion	0 – 1	2 – 3	4	5	6	7 – 10	11 – 20
AIC	1 (0)	1 (0)	0 (0)	77 (68)	8 (10)	11 (21)	2 (1)
AICC	1 (0)	1 (0)	1 (0)	81 (75)	8 (11)	7 (14)	1 (0)
H&Q	1 (0)	1 (0)	1 (0)	87 (88)	8 (7)	2 (5)	0 (0)
BIC	2 (0)	0 (0)	1 (0)	77 (83)	8 (6)	7 (9)	5 (2)
DC	1 (0)	0 (0)	1 (0)	82 (85)	8 (7)	5 (8)	3 (0)
MDC	7 (0)	2 (0)	5 (0)	85 (97)	1 (3)	0 (0)	0 (0)

Table 4. Frequencies of estimated order in 100 replications from the AR(10) model given by (3.4) with sample sizes 100 and 200 (in parentheses)

	estimated order						
Criterion	< 8	8	9	10	11	12	> 12
AIC	0 (0)	0 (0)	0 (0)	80 (74)	11 (10)	3 (3)	6 (13)
AICC	1 (0)	0 (0)	0 (0)	90 (82)	6 (11)	2 (2)	1 (5)
H&Q	4 (0)	3 (0)	0 (0)	82 (89)	8 (5)	3 (4)	0 (2)
BIC	11 (0)	0 (0)	0 (0)	86 (97)	2 (2)	1 (1)	0 (0)
DC	5 (0)	0 (0)	0 (0)	79 (85)	11 (6)	3 (3)	2 (6)
MDC	7 (0)	2 (0)	1 (0)	89 (99)	0 (1)	1 (0)	0 (0)

Table 5. Frequencies of correct order selection in 100 replications from the models (3.1)-(3.4) using MDC.

	sample size				
Model	30	50	100	200	400
AR(1)	46	64	93	99	100
AR(2)	59	82	98	98	99
AR(5)	26	45	85	97	98
AR(10)	2	33	89	99	100

The frequencies reported in these tables are highly dependent on the choice of model parameters. For example, in the AR(1) case with sample size 30, a change in the AR parameter from .4 to $-.8$ increases the correct order selection frequency for MDC from 46 to 94 (see also Table 1 in Hannan and Quinn (1979)).

In terms of the frequency of correct order selection, MDC generally performed best. As can be seen in Table 5, MDC estimates the true order well even for moderate sample sizes. All of the procedures did surprisingly well in identifying the AR(10) model (Table 4) with MDC achieving success rates 99% and 100% for sample sizes 200 and 400 respectively. At a sample size of 30, DC and MDC were the only criteria to correctly estimate the order of the AR(10) model at least once. For smaller sample sizes, AIC and AICC were often competitive with the consistent criteria especially when the true order is large. This is due in part to the small penalty factors in AIC and AICC which tend to favor higher order models.

The superior performance of MDC over DC is somewhat counter-intuitive in model (3.1). However model (3.1) has small partial autocorrelations at all lags and so the test series generated from the estimated model will be difficult to distinguish from white noise. On the other hand, this is not an issue with MDC since the test series are generated from models chosen to have large partial autocorrelations. In model (3.2), the partial autocorrelations at lags 1 and 2 are large and both DC and MDC give good results. In models (3.3) and (3.4), the fitted models of orders 0, 1, and 2 bear little relation to the true models so there is no reason to expect DC to perform as well as MDC. Also, the partial autocorrelations at lags 1 and 2 for these two models are relatively small.

Table 6. Frequencies of correct order selection in 100 replications from model (3.2) with different noise distributions. Sample sizes are 50 and 100 (in parentheses).

Distribution	Exp	Laplace	$t_{2.5}$	$U[-1,1]$
AIC	76 (71)	73 (72)	79 (75)	72 (72)
AICC	83 (79)	79 (77)	84 (81)	77 (77)
H&Q	82 (89)	82 (88)	88 (89)	79 (88)
BIC	79 (94)	74 (96)	78 (97)	72 (97)
DC	80 (84)	82 (85)	82 (83)	78 (86)
MDC	83 (99)	80 (97)	90 (96)	80 (96)

We also compared the performance of the order selection criteria over a wide range of noise distributions: exponential, two-sided exponential, t-distribution with 2.5 degrees of freedom and uniform on $[-1,1]$. MDC continued to perform the best across these distributions. A summary of these results, using the AR(2) model given in (3.2), is reported in Table 6.

Finally we considered the AR(3) model

$$X_t = .7X_{t-3} + Z_t, \quad \{Z_t\} \sim \mathrm{IID}(0,1)$$

which does not satisfy the condition $\min_{1\le j\le K_1} |\phi_{jj}| > 0$ of Theorem 2.1 if $K_1 \geq 1$. As a result the DC method does not produce a consistent estimate of p unless $K_1 = 0$. Nevertheless, for simulated series of length 50, DC still correctly identified the true order 76 out of 100 times (MDC managed 80 out of 100). It appears that the performance of DC is not adversely affected if some of the partial autocorrelations are zero. Of course, this is not an issue with MDC since it produces consistent estimates regardless of the values of the parameters in the true model.

The appealing feature of DC and MDC is that they attempt to choose, in a data driven fashion, an optimal penalty factor. Here, optimal is in the sense of correctly identifying the true order. Since, in practice, there is rarely such a thing as the *true order* of the model, one might compare order selection criteria relative to a different notion of optimality. For instance, if the modelling objective is to obtain h-step-ahead forecasts, then one might choose an order selection procedure which gives a model with minimum mean square error of the h-step forecast. In future work, we will explore extensions of our resampling scheme designed to obtain suitable penalty factors for a given optimality criterion.

4. Yule-Walker vs. MLE and Burg

While the three estimation procedures for finite order autoregressive processes, maximum likelihood, Burg, and Yule-Walker are asymptotically equivalent, it is generally accepted that for small to moderate sample sizes with p known, maximum likelihood and Burg are the preferred estimation procedures. Yule-Walker estimates tend to have larger biases which are even more pronounced when the autoregressive polynomial has one or more roots near the unit circle. However, when they are used to fit an overparameterized model to the data, as in any order selection procedure, maximum likelihood and Burg can produce extremely poor parameter estimates. On the other hand, Yule-Walker estimates tend to be well behaved even when the order of the fitted model approaches the sample size.

To illustrate this point, we generated 10,000 replicates of time series with length 23 from the AR(2) model,

$$X_t = .99X_{t-1} - .8X_{t-2} + Z_t, \quad \{Z_t\} \sim \mathrm{IID}(0,1).$$

This model and sample size were used in the simulation study of Hurvich and Tsai (1989). Figures 1–3 contain boxplots of the parameter estimates obtained in fitting an AR(15) model to the data using Yule-Walker, Burg, and maximum likelihood respectively. In this case the true coefficient vector is $(.99, -.8, 0, \ldots, 0)'$. As is clearly evident, Yule-Walker outperforms Burg and maximum likelihood by a wide margin. The 'box' for the mle of ϕ_{15} ranges from $-.45$ to $.45$, an incredibly wide range when one considers that ϕ_{15} is constrained to the region $(-1,1)$. The sample standard deviations for the Yule-Walker, Burg and maximum likelihood estimates of ϕ_{15} are .107, .392, and .544 respectively.

Figure 4 contains boxplots of the estimates of σ^2 using Yule-Walker, Burg and maximum likelihood. As can be seen in Figure 4, the mle estimate of σ^2 can be much too small when a high order model is fitted. In contrast, the Yule-Walker estimate of σ^2 is more stable as the order of the model increases and has much smaller bias than either the mle or Burg estimator. This partially explains why order selection criteria using mle or Burg estimation, rather than Yule-Walker estimation, tend to select more overparameterized models. Table 7, which compares the results of order selection using AIC for the AR(2) example above, bears this out.

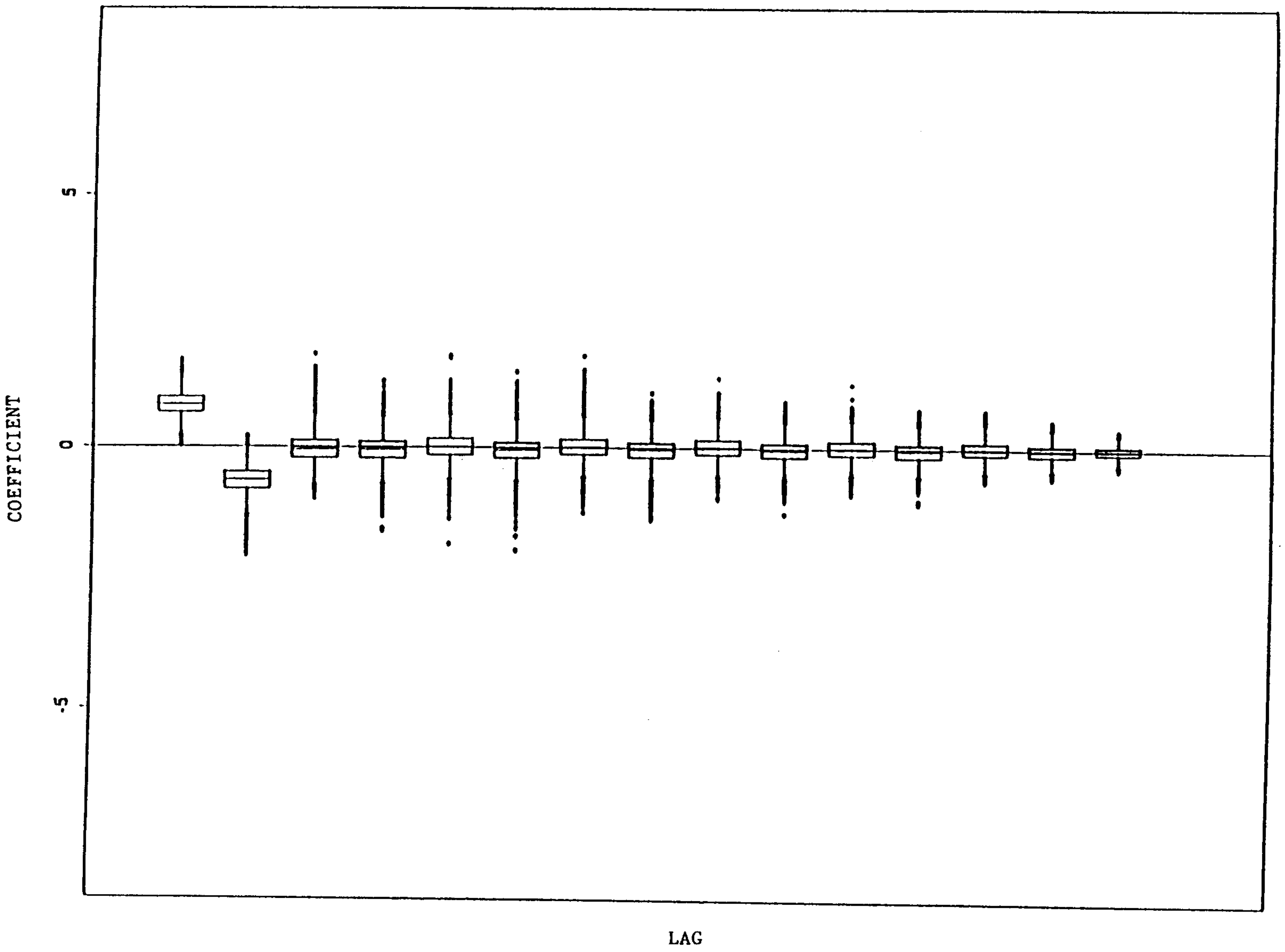
Table 7. Frequencies of estimated order in 100 replications for the AR(2) model given above and using AIC (sample size is 23)

	estimated order								
Est. procedure	0	1	2	3	4	5	6 – 10	11 – 15	16 – 20
YW	0	0	79	14	5	2	0	0	0
Burg	0	0	63	17	3	2	7	5	3
MLE	0	0	60	18	3	2	7	5	5

For these reasons we have consistently used Yule-Walker estimation for order selection. While Yule-Walker may not provide the best estimates of the parameters when the order of the model is known, it is more reliable for fitting overparameterized models than maximum likelihood and Burg estimation.

Acknowledgement

The research of the second and third authors was supported by NSF Grant DMS 9100392.



True model is AR(2). Boxplots based on 10000 replicates

Figure 1. AR(15) coefficients fitted by Yule-Walker

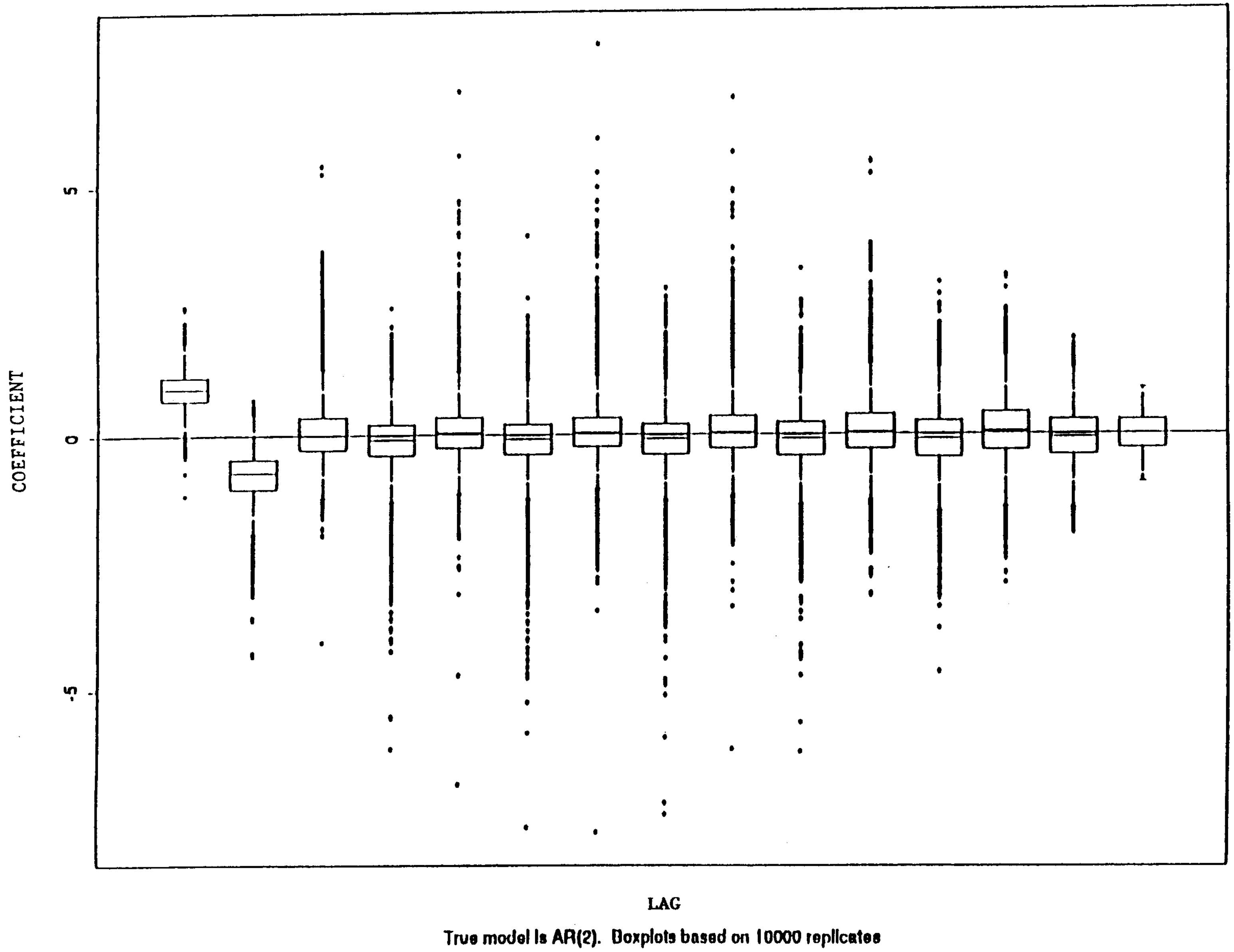
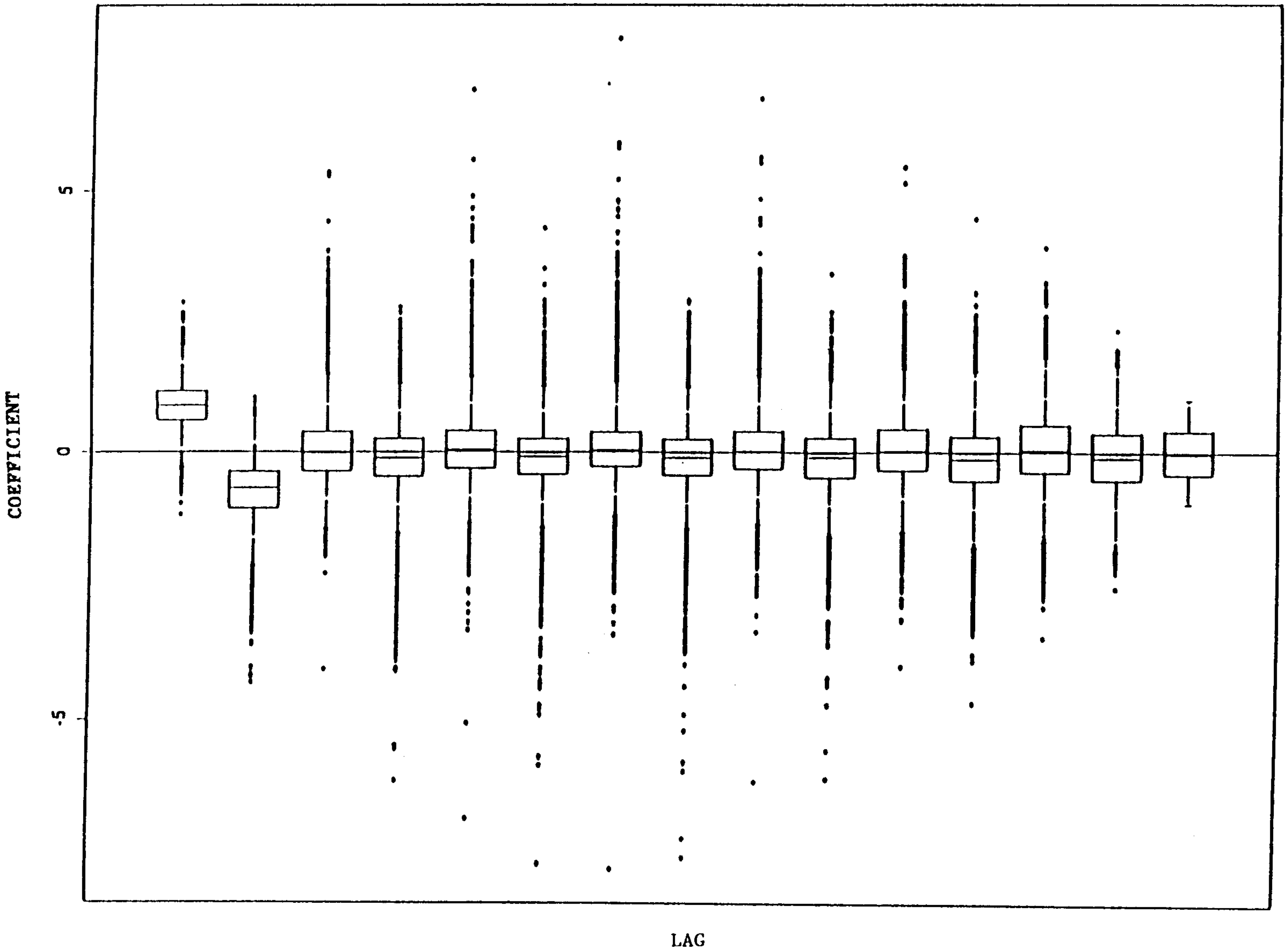


Figure 2. AR(15) coefficients fitted by Burg



True model is AR(2). Boxplots based on 10000 replicates

Figure 3. AR(15) coefficients fitted by MLE

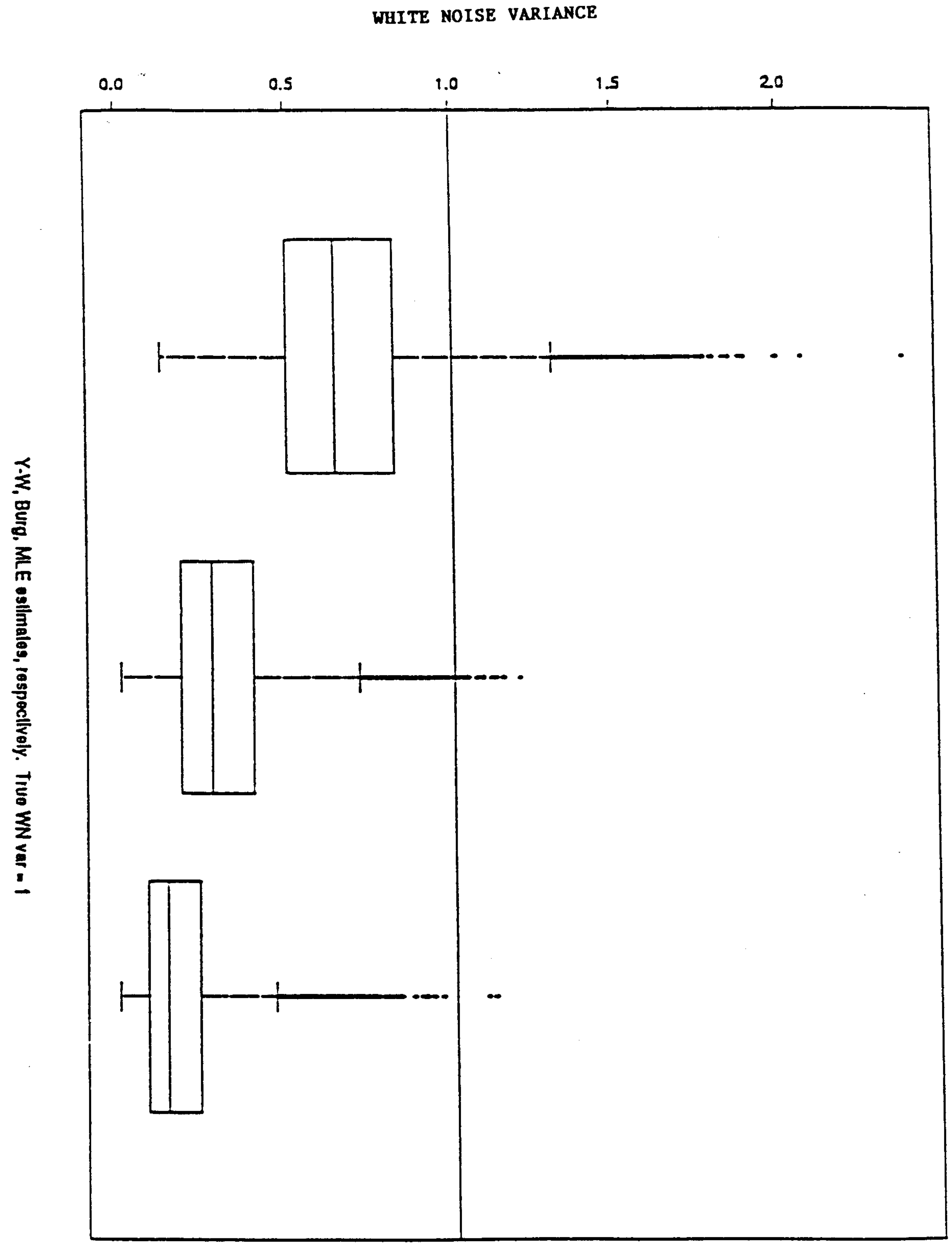


Figure 4. Estimates of WN variance from AR(15) fit

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21**, 243–247.

Akaike, H. (1978). Time series analysis and control through parametric models. *Applied Time Series Analysis* (Edited by D. F. Findley), Academic Press, New York.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*, 2nd edition. Springer-Verlag, New York.

Burg, J. P. (1967). Maximum entropy spectral analysis. 37th Annual International S.E.G. Meeting, Oklahoma City, Oklahoma.

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

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

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

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

Department of Statistics, Colorado State University, Fort Collins, CO 80523, U.S.A.
Department of Statistics, Temple University, Philadelphia, PA 19122, U.S.A.

(Received June 1992; accepted January 1993)