Biometrika (1991), 78, 3, *pp.* 499-509
Printed in Great Britain

Bias of the corrected AIC criterion for underfitted regression and time series models

BY CLIFFORD M. HURVICH

Department of Statistics and Operations Research, New York University, Tisch Hall, 40 West Fourth Street, New York, New York 10003, U.S.A.

AND CHIH-LING TSAI

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

SUMMARY

The Akaike Information Criterion, AIC (Akaike, 1973), and a bias-corrected version, AIC_C (Sugiura, 1978; Hurvich & Tsai, 1989) are two methods for selection of regression and autoregressive models. Both criteria may be viewed as estimators of the expected Kullback-Leibler information. The bias of AIC and AIC_C is studied in the underfitting case, where none of the candidate models includes the true model (Shibata, 1980, 1981; Parzen, 1978). Both normal linear regression and autoregressive candidate models are considered. The bias of AIC_C is typically smaller, often dramatically smaller, than that of AIC. A simulation study in which the true model is an infinite-order autoregression shows that, even in moderate sample sizes, AIC_C provides substantially better model selections than AIC.

Some key words: AIC; Autoregression; Kullback-Leibler information; Model selection.

1. INTRODUCTION

In a seminal paper, Akaike (1973) proposed that the expected Kullback-Leibler information be used as a means of discriminating between competing statistical models, even if the models have different dimensions. He proposed the Akaike Information Criterion, AIC, as an asymptotically unbiased estimator of this information. Since the underlying target criterion is sound, it may be hoped that minimization of an unbiased estimate of it will provide good model selections. The idea has been put in a general framework by Linhart & Zucchini (1986), who view model selection as the construction of approximately unbiased estimators of an underlying criterion function.

It is possible to prove independently that AIC produces good model selections in large samples (Shibata, 1980). Nevertheless (Findley, 1985) the bias itself seems to be a basic property worthy of study. Furthermore, one may hope that by improving the bias properties, one will also improve the quality of the selected models. This is indeed the case for the corrected AIC criterion, AIC_C, originally proposed by Sugiura (1978) with a view towards bias reduction, and found by Hurvich & Tsai (1989) to produce not only dramatic bias reduction but also greatly improved model selections in small samples.

For a normal linear regression, or autoregressive, model with p regression, or autoregressive, parameters, the AIC and AIC_C criteria are respectively defined by

$$AIC = n\{\log(2\pi\hat{\sigma}^2)+1\}+2(p+1),$$

$$AIC_C = n\log(2\pi\hat{\sigma}^2)+n\frac{1+p/n}{1-(p+2)/n},$$

where $\hat{\sigma}^2$ is the estimated error or innovations variance for the fitted pth order candidate model.

In previous work, the derivation of AIC_C and the study of its bias properties were limited to the case where the true model is of finite dimension and is either correctly specified or overfitted. In practice, however, since a variety of candidate models will be considered, it will often happen that the model is underfitted. We will say that a true model is correctly specified or overfitted if some configuration of parameter values in the candidate model, perhaps including some zero values, yields the true model. Otherwise, the true model is said to be underfitted, and the candidate model is referred to as an approximating model. If the true model is of infinite dimension, which we feel will be the typical situation in practice, then none of the candidate models will be capable of exactly producing the true model, and therefore the model will always be underfitted.

In this paper, we study the bias properties and model selection quality of AIC and AIC_C for the underfitting case. We consider both linear regression and autoregressive time series models. In the normal linear regression case, we derive exact expressions for the expectations of AIC, AIC_C and the Kullback-Leibler information. The bias of AIC and AIC_C depends on the true regression function, and on the form and dimension of the candidate model. We numerically evaluate the bias for a class of trigonometric candidate models, assuming a variety of true regression functions. We find that, although AIC_C is not uniformly less biased than AIC, the minimizers over a set of candidate model orders of the expected AIC_C and Kullback-Leibler information are similar to each other, and often quite different from the minimizer of the expected AIC. Furthermore, as the ratio of the model dimension to the sample size increases, AIC becomes strongly negatively biased, while the bias of AIC_C is often dramatically smaller than that of AIC. For the autoregressive case, exact finite-sample results are not available. Findley (1985) has given a rigorous derivation of the asymptotic bias of AIC for any correct or approximating ARMA model. We study the finite-sample bias properties of AIC and AIC_C, viewed as functions of the order of the approximating AR models, using a combination of theory and Monte Carlo. Once again, we find that AIC_C can be substantially less biased than AIC. We also find that AIC_C significantly outperforms AIC in terms of quality of the selected approximating model. These findings strengthen the case for using AIC_C in place of AIC, as was originally recommended by Hurvich & Tsai (1989).

2. APPROXIMATING REGRESSION MODELS

2·1. *Theoretical derivation*

Given data $y=(y_1,\ldots,y_n)'$ generated from the operating model, i.e. true model, $y=\mu+\varepsilon$ where μ is the true mean of y and $\varepsilon\sim N(0,\sigma_0^2 I_n)$, we consider the candidate family of models approximating family $y=X\theta+u$, where X is a nonstochastic $n\times p$ matrix, θ is a $p\times 1$ parameter vector, and $u\sim N(0,\sigma^2 I_n)$. The parameters (θ,σ^2) are estimated by least squares, that is

$$\hat{\theta}=(X'X)^{-1}X'y,\quad \hat{\sigma}^2=(y-X\hat{\theta})'(y-X\hat{\theta})/n.$$

If $g_{\theta,\sigma^2}(y)$ denotes the likelihood for (θ, σ^2), and E_0 denotes the expectation with respect to the operating model, we define the discrepancy function

$$\begin{aligned} d(\theta, \sigma^2) &= E_0\{-2 \log g_{\theta,\sigma^2}(y)\} \\ &= n \log (2\pi\sigma^2) + E_0\{(\mu + \varepsilon - X\theta)'(\mu + \varepsilon - X\theta)/\sigma^2\} \\ &= n \log (2\pi\sigma^2) + n\sigma_0^2/\sigma^2 + (\mu - X\theta)'(\mu - X\theta)/\sigma^2. \end{aligned}$$

Thus

$$d(\hat{\theta}, \hat{\sigma}^2) = n \log (2\pi\hat{\sigma}^2) + n\sigma_0^2/\hat{\sigma}^2 + (\mu - X\hat{\theta})'(\mu - X\hat{\theta})/\hat{\sigma}^2.$$

Define the $n \times n$ projection matrix $H = X(X'X)^{-1}X'$. Note that $H^2 = H$ and $X\hat{\theta} = Hy$. Let $\lambda = \mu'(I - H)\mu/\sigma_0^2$, and let $\chi_k^2(\lambda)$ denote a noncentral χ_k^2 distribution with noncentrality parameter λ.

LEMMA. *The random variables* $(\mu - X\hat{\theta})'(\mu - X\hat{\theta})$ *and* $\hat{\sigma}^2$ *are independently distributed. Further,*

$$\{(\mu - X\hat{\theta})'(\mu - X\hat{\theta})/\sigma_0^2 - \lambda\} \sim \chi_p^2, \quad n\hat{\sigma}^2/\sigma_0^2 \sim \chi_{n-p}^2(\lambda).$$

A proof follows from the arguments of Rao (1973, pp. 186, 187, 209).

From Rao (1973, p. 182), if $X \sim \chi_k^2(\lambda)$ then X has density

$$g(x) = e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!} (\tfrac{1}{2}\lambda)^r f_{2r+k}(x),$$

where $f_{2r+k}(x)$ is the density of a central χ_{2r+k}^2 random variable. Since the logarithm of a χ_{2r+k}^2 random variable has expectation $\log 2 + \psi(r + \frac{1}{2}k)$, where $\psi(.)$ denotes the digamma function, it follows that

$$E_0\{n \log (2\pi\hat{\sigma}^2)\} = n \log (2\pi) + n\left[\log (\sigma_0^2/n) + \log 2 + e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} (\tfrac{1}{2}\lambda)^r \psi\{\tfrac{1}{2}(2r + n - p)\}\right]. \tag{1}$$

Since the inverse of a χ_{2r+k}^2 random variable has expectation $(2r + k - 2)^{-1}$, it follows that

$$E_0(n\sigma_0^2/\hat{\sigma}^2) = n^2 E_0(n\hat{\sigma}^2/\sigma_0^2)^{-1} = n^2 e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!} (\tfrac{1}{2}\lambda)^r \frac{1}{2r + n - p - 2}. \tag{2}$$

Thus, combining (1) and (2), the expected Kullback-Leibler discrepancy is

$$\begin{aligned} \Delta(\hat{\theta}, \hat{\sigma}^2) &= E_0\{d(\hat{\theta}, \hat{\sigma}^2)\} \\ &= E_0\{n \log (2\pi\hat{\sigma}^2)\} + E_0\left[n\left(\frac{\sigma_0^2}{\hat{\sigma}^2}\right)\left\{1 + (\mu - X\hat{\theta})'\left(\frac{\mu - X\hat{\theta}}{n\sigma_0^2}\right)\right\}\right] \\ &= n \log (4\pi\sigma_0^2/n) + ne^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} (\tfrac{1}{2}\lambda)^r \psi\{\tfrac{1}{2}(2r + n - p)\} \\ &\quad + \left\{n^2 e^{-\frac{1}{2}\lambda} \sum_{r=0}^{\infty} \frac{1}{r!} (\tfrac{1}{2}\lambda)^r \frac{1}{2r + n - p - 2}\right\}\{1 + (\lambda + p)/n\}. \end{aligned} \tag{3}$$

2·2. *Numerical results*

Here, we consider the operating model

$$y_t = \mu_t + \varepsilon_t \quad (t = 0, \ldots, n - 1)$$

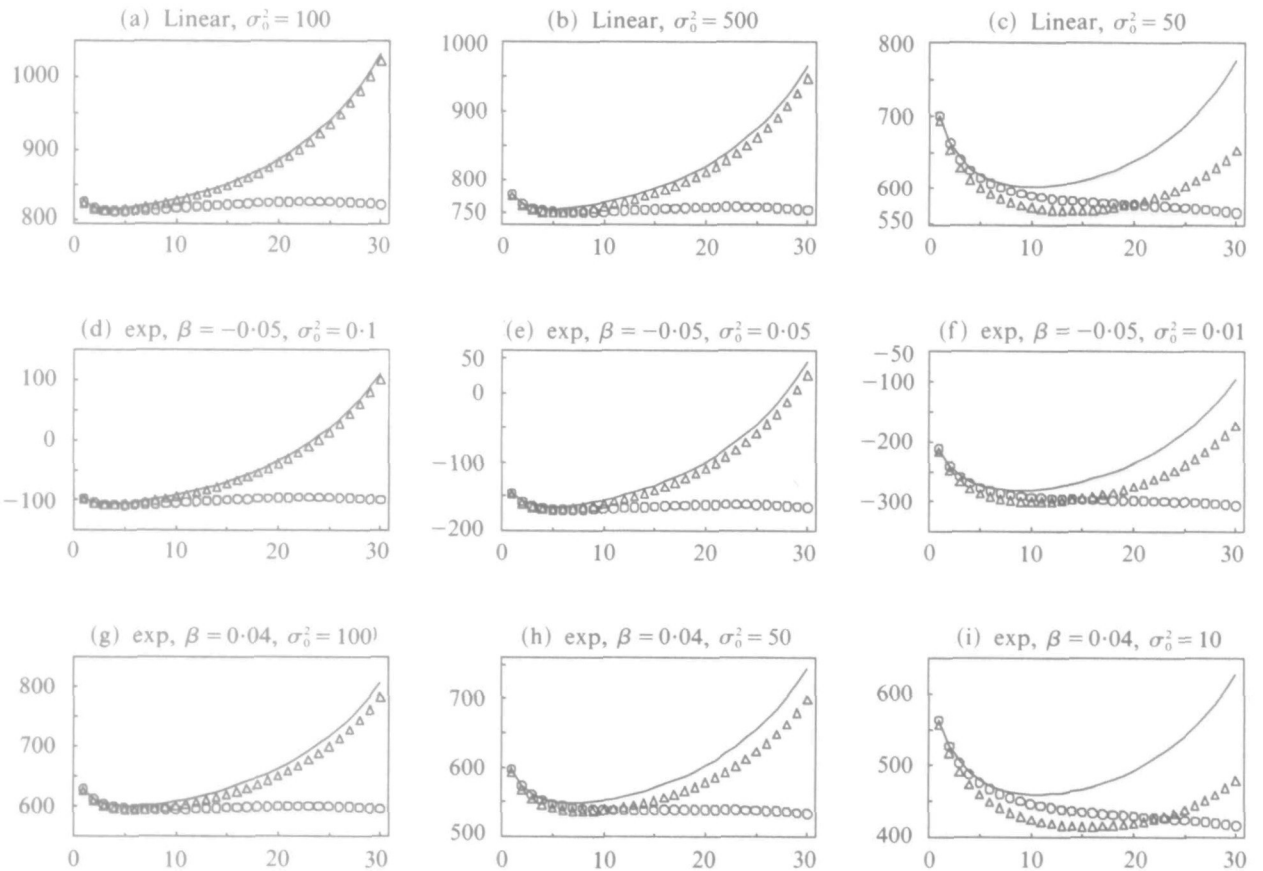

Fig. 1. Expected AIC_C, shown by lines, Δ^*, triangles, and AIC, circles, versus p. Trigonometric regression candidates, linear and exponential operating models; $n = 100$.

where ε_t are independent identically distributed normal random variables with mean zero and variance σ_0^2, together with the trigonometric approximating models

$$y_t = A_0 + \sum_{j=1}^{p} \{A_j \cos(\omega_j t) + B_j \sin(\omega_j t)\} + u_t,$$

where $A_0, A_1, B_1, \ldots, A_p, B_p$ are real-valued parameters, $\omega_j = 2\pi j/n$, and u_t are independent identically distributed normal random variables with mean zero and variance σ^2.

Figure 1 gives plots of $\Delta(\hat{\theta}, \hat{\sigma}^2)$, denoted by Δ^*, together with the expectations of AIC and AIC_C, all functions of p, where $p = 1, \ldots, 30$, for a sample of size $n = 100$ under nine different combinations of choices for μ_t and σ_0^2. Figures 1(a), (b), (c) assume the linear operating model $\mu_t = t$, denoted by 'linear'. Figures 1(d)-(i) assume the exponential operating model $\mu_t = e^{\beta t}$, denoted by 'exp', using $\beta = -0{\cdot}05$ and $\beta = 0{\cdot}04$.

Although AIC_C is not uniformly less biased than AIC, the expected value of AIC_C outperforms that of AIC in capturing the overall shape of the Δ^* curves, viewed as functions of p. In particular, the values of p which minimize $E(\text{AIC}_C)$ and Δ^* are similar, while $E(\text{AIC})$ is often minimized at very large, and clearly suboptimal, values of p. Also AIC_C becomes positively biased as p is increased, a tendency which becomes more pronounced as σ_0^2 is decreased. Finally, the patterns observed here depend more strongly on the operating variance σ_0^2 than on the form of the operating mean μ_t.

3. Approximating autoregressive time series models

3·1. *Theoretical derivation*

Suppose we have data $x = (x_0, \ldots, x_{n-1})'$ from a zero-mean stationary Gaussian series $\{x_t\}_{t=-\infty}^{\infty}$ having an infinite order autoregressive, AR(∞), representation

$$\sum_{k=0}^{\infty} b_k x_{t-k} = \eta_t,$$

where $\{b_k\}$ are constants, $b_0 = 1$ and $\{\eta_t\}$ is a zero-mean Gaussian white noise series with variance σ_0^2. Note that $\{x_t\}$ need not have a finite-order autoregressive representation. The candidate model is a pth order autoregression, AR(p), of form

$$\sum_{k=0}^{p} a_k x_{t-k} = \varepsilon_t,$$

where $a_0 = 1$ and $\{\varepsilon_t\}$ is a zero-mean Gaussian white noise series with variance σ^2. The parameters are estimated by maximum likelihood, least squares, Burg's (1978) method, or any other asymptotically equivalent method. Findley (1985) has examined the bias of AIC for this case and has shown that, as $n \to \infty$ and $p \to \infty$, AIC is asymptotically unbiased for the expected Kullback-Leibler information. In a Monte Carlo study, in which the operating model is MA(1), we compare the bias properties of AIC and AIC_C for fixed values of n and p. Further, we compare the performance of AIC and AIC_C in terms of quality of the selected models. Before presenting the Monte Carlo results, we obtain a rough asymptotic approximation to the expected Kullback-Leibler information. This approximation indicates that AIC, although asymptotically unbiased to first order, may in fact be strongly negatively biased for a given n and p.

Let $\theta = (\sigma^2, a_1, \ldots, a_p)'$ and $\theta_0 = (\sigma_0^2, b_1, b_2, \ldots)'$ denote the candidate AR(p) and true AR(∞) parameter vectors, respectively. The Kullback-Leibler discrepancy is

$$d(\theta, \theta_0) = E_0\{-2 \log g_\theta(x)\},$$

where $g_\theta(x)$ is the likelihood function for the candidate model parameters, and E_0 denotes expectation under the true model. Let Σ_θ and Σ_{θ_0} denote the $n \times n$ covariance matrices of x under the models with θ and θ_0, respectively. Since

$$-2 \log g_\theta(x) = n \log (2\pi) + \log |\Sigma_\theta| + x' \Sigma_\theta^{-1} x,$$

one obtains

$$d(\theta, \theta_0) = n \log (2\pi) + \log |\Sigma_\theta| + \operatorname{tr} (\Sigma_{\theta_0} \Sigma_\theta^{-1}).$$

Let $\hat{\theta} = (\hat{\sigma}^2, \hat{a}_1, \ldots, \hat{a}_p)'$ denote the estimated parameters in the candidate model. Note that $\hat{\theta}$ need not be the maximum likelihood estimator. The selection methods AIC and AIC_C may be viewed as estimators of the expected Kullback-Leibler information,

$$\Delta(\hat{\theta}) = E_0\{d(\hat{\theta}, \theta_0)\} = n \log (2\pi) + E_0(\log |\Sigma_{\hat{\theta}}|) + E_0\{\operatorname{tr} (\Sigma_{\theta_0} \Sigma_{\hat{\theta}}^{-1})\}. \tag{4}$$

Denote the true spectral density by $f(\omega)$ for $\omega \in [-\pi, \pi]$, and denote the AR(p) spectral estimate by

$$\hat{f}_p(\omega) = \frac{\hat{\sigma}^2/(2\pi)}{|1 + \sum \hat{a}_k \exp(i\omega k)|^2},$$

where the sum is over the range $k = 1, \ldots, p$. From Parzen (1983, p. 235), the eigenvectors and corresponding eigenvalues of Σ_θ may be approximated by

$$n^{-\frac{1}{2}}\{\exp(-i\omega_j t)\} \quad (t = 0, \ldots, n-1), \quad 2\pi f(\omega_j) \quad (j = 0, \ldots, n-1).$$

It follows that the expected Kullback-Leibler information $\Delta(\hat{\theta})$ may be approximated by

$$\delta(\hat{\theta}) = E_0\{n \log (2\pi\hat{\sigma}^2)\} + \{n/(2\pi)\} E_0 \int_{-\pi}^{\pi} \{f(\omega)/\hat{f}_p(\omega)\}\, d\omega.$$

To obtain an approximation for the second term we use the result of Berk (1974). If $p \to \infty$, $n \to \infty$ with $p^3/n \to 0$ then $\hat{f}_p(\omega)$ is asymptotically equivalent to the truncated periodogram estimator

$$f^*(\omega) = \frac{1}{2\pi} \sum_{|r| < p} \hat{c}_r \exp (ir\omega),$$

where

$$\hat{c}_r = \frac{1}{n} \sum_{t=0}^{n-|r|} x_t x_{t-|r|}$$

is the sample autocovariance. From Bloomfield (1976, p. 191), we obtain the approximations

$$E\{f^*(\omega)\} \doteq f(\omega), \quad \operatorname{var}\{f^*(\omega)\} \doteq 2pn^{-1} f^2(\omega).$$

From Bloomfield (1976, p. 196), if we define $\nu = n/p$, then the distribution of $\nu f^*(\omega)/f(\omega)$ may be approximated by χ^2_ν. If we treat all the above approximations as exact and assume that $\hat{f}_p(\omega) = f^*(\omega)$ then we obtain

$$E_0\left\{\frac{f(\omega)}{\hat{f}_p(\omega)}\right\} = E_0\left(\frac{\nu}{\chi^2_\nu}\right) = \frac{\nu}{\nu - 2} = \frac{1}{1 - 2p/n}.$$

Thus

$$\Delta(\hat{\theta}) \doteq E_0\{n \log (2\pi\hat{\sigma}^2)\} + n \frac{1}{1 - 2p/n}. \tag{5}$$

Note that, to first order, the approximation (5) to the expected Kullback-Leibler information has penalty term $2p/n$, in agreement with that of AIC. Nevertheless, the full penalty term in (5) is $(1-2p/n)^{-1}$, which is always larger, and potentially much larger, than the penalty term of AIC, $2p/n$. Thus, AIC may be strongly negatively biased. The Monte Carlo results given below, which do not rely on the approximations used in the above derivation, indicate that the exact penalty term of the expected Kullback-Leibler information is in fact quite close to the penalty term of AIC_C, that is,

$$n\frac{1+p/n}{1-(p+2)/n},$$

and that AIC_C is much less biased than AIC.

3·2. *Monte Carlo results*

Here we present Monte Carlo results on the performance of AIC and AIC_C for autoregressive time series model selection, when the operating model is Gaussian AR(∞). We study the finite-sample bias properties of AIC and AIC_C, viewed as estimators of $\Delta(\hat{\theta})$. We also study the quality of the models selected by AIC and AIC_C. The true model used throughout is the first-order moving average process $x_t = \varepsilon_t + 0{\cdot}99\varepsilon_{t-1}$, where $\{\varepsilon_t\}$ are independent and identically distributed standard normal. Note that $\{x_t\}$ has an AR(∞) representation, and cannot be written as a finite-order AR. For each of the sample sizes $n = 23$, 30, 40, 50, 75 and 100, we generated 100 independent realizations $x_0, \ldots, x_{n-1}$ of the moving average process. For each realization, autoregressive models of orders $p = 1, \ldots, 20$ were fitted by the Burg method, and the criteria AIC, AIC_C and SIC (Schwarz, 1978) were computed. The SIC criterion is given by

$$\text{SIC} = n \log(2\pi\hat{\sigma}^2) + p \log n.$$

Also computed was $d(\hat{\theta}_p, \theta_0)$, where the subscript in $\hat{\theta}_p$ has been added for clarity to explicitly indicate model order. Averages of the criterion functions as well as $d(\hat{\theta}_p, \theta_0)$ were computed over the 100 realizations. All these are functions of the candidate model order p. We denote the average of the 100 values of $d(\hat{\theta}_p, \theta_0)$ by $\tilde{\Delta}(p)$, or simply $\tilde{\Delta}$. Note that $\tilde{\Delta}$ serves as an approximation to the expected Kullback-Leibler information $\Delta(\hat{\theta}_p)$ defined in (4). Figure 2 shows that, almost without exception, AIC_C exhibits less bias than AIC in estimating $\tilde{\Delta}$. Furthermore, the magnitude of the bias of AIC increases with model order, while AIC_C remains nearly unbiased for all model orders. These results parallel those found for the overfitting case in Hurvich & Tsai (1989).

Next, we explore the quality of the models selected by AIC, AIC_C and SIC. Since there is no true finite autoregressive model order in the current study, we will measure quality here using the expected Kullback-Leibler discrepancy, instead of simply examining the selected model orders. Another reasonable measure of quality, prediction error, will be considered at the end of this section. For each realization, the criteria yielded selected model orders $\hat{p}(\text{AIC})$, $\hat{p}(\text{AIC}_C)$, $\hat{p}(\text{SIC})$, and corresponding expected Kullback-Leibler discrepancies $\Delta_{\text{AIC}} = \tilde{\Delta}\{\hat{p}(\text{AIC})\}$, $\Delta_{\text{AIC}_C} = \tilde{\Delta}\{\hat{p}(\text{AIC}_C)\}$, $\Delta_{\text{SIC}} = \tilde{\Delta}\{\hat{p}(\text{SIC})\}$. In order to allow these discrepancy values to be viewed relative to an absolute zero, the constant $d(\theta_0, \theta_0)$ was subtracted, yielding

$$D_{\text{AIC}} = \Delta_{\text{AIC}} - d(\theta_0, \theta_0), \quad D_{\text{AIC}_C} = \Delta_{\text{AIC}_C} - d(\theta_0, \theta_0), \quad D_{\text{SIC}} = \Delta_{\text{SIC}} - d(\theta_0, \theta_0).$$

The average values of D_{AIC}, D_{AIC_C} and D_{SIC} over the 100 realizations are given in Table 1. For all sample sizes studied the average value of D_{AIC_C} is less than those of D_{AIC} and

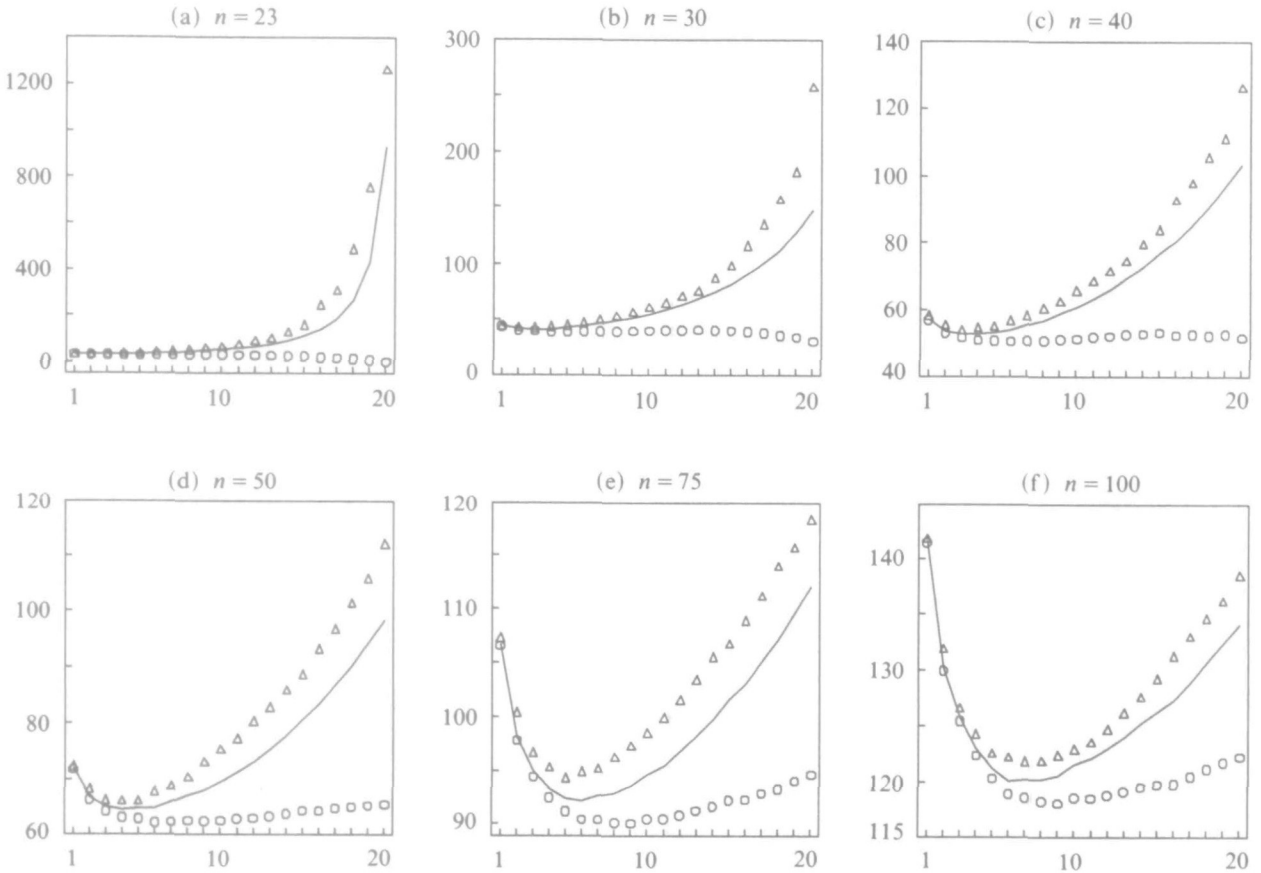

Fig. 2. Average AIC_C, shown by lines, $\tilde{\Delta}$, triangles, and AIC, circles, versus candidate autoregressive model order, based on 100 realizations of MA(1).

Table 1. *Averages of standardized discrepancies*

n	Ave (D_{AIC})	Ave (D_{AIC_C})	Ave (D_{SIC})
23	864·438	10·638	560·918
30	112·516	11·410	33·571
40	32·255	12·591	13·076
50	25·674	14·586	14·665
75	21·067	18·036	18·787
100	21·524	19·744	21·784

Based on 100 realizations from an MA(1) process.

D_{SIC}, suggesting that AIC_C provides the best model selections, on average. To test whether apparent differences in performance were significant, we performed pairwise comparisons of discrepancy values, at the level of individual realizations. Specifically, for pairs (L, R) of selection criteria, we computed the 100 values, one for each realization, of $D_L - D_R$. A one-sample Wilcoxon test was performed on the set of values of $D_L - D_R$ for the null hypothesis that the median of $D_L - D_R$ is zero. The p-values are given in Table 2 showing that AIC_C is in all cases strongly superior to AIC. Furthermore, AIC_C is superior, and in most cases strongly superior, to SIC.

In the comparisons of AIC and AIC_C the largest p-value occurred for $n = 100$. Since the maximum candidate model order was held fixed at 20, it is to be expected that as n is increased the behaviour of AIC and AIC_C will become increasingly similar, since the two criteria are asymptotically equivalent in this case. Also in the comparisons of SIC and AIC_C, the p-values do not decrease monotonically with n. The initial increase of the p-values, reaching a maximum of 0·200 at $n = 40$, may be attributed to the fact that the maximum ratio of model order to sample size is $20/n$, which decreases with n, and to the fact that SIC is strongly negatively biased when the model order is close to n but increases fairly sharply with model order when the model order is a moderate fraction of n. The eventual decrease of the p-values for $n \geq 40$ is to be expected since AIC_C is asymptotically efficient while SIC is not.

Table 2. *One-sided p-values for one-sample Wilcoxon test based on differences of standardized average discrepancies $D_L - D_R$ for selection criteria (L, R)*

n	L	R	No. neg.	No. zero	No. pos.	p-value
23	AIC	AIC_C	0	4	96	*
23	SIC	AIC_C	1	36	63	*
30	AIC	AIC_C	1	26	73	*
30	SIC	AIC_C	10	69	21	$7{\cdot}01 \times 10^{-4}$
40	AIC	AIC_C	3	34	63	*
40	SIC	AIC_C	15	68	17	0·200
50	AIC	AIC_C	5	37	58	*
50	SIC	AIC_C	19	56	25	0·131
75	AIC	AIC_C	2	58	40	*
75	SIC	AIC_C	18	30	52	0·0016
100	AIC	AIC_C	7	60	33	$1{\cdot}55 \times 10^{-5}$
100	SIC	AIC_C	9	33	58	*

Numbers neg., zero, pos. denote the number of negative, zero and positive differences in $D_L - D_R$ for the 100 realizations.
* p-value less than 10^{-5}.

Another criterion for assessing the quality of a fitted autoregressive model is mean squared prediction error. For simplicity, we use the normalized prediction error

$$\text{NPE}(p) = E(\hat{a}'R\hat{a}) - \sigma_0^2,$$

where $\hat{a} = (1, \hat{a}_1, \ldots, \hat{a}_p)'$ is the vector of fitted AR(p) coefficients, R is the true $(p+1) \times (p+1)$ covariance matrix of the process and σ_0^2 is the innovation variance. Note that $E(\hat{a}'R\hat{a})$ is the one-step mean squared error incurred in predicting an independent realization $\{y_t\}$ of the process $\{x_t\}$ using an AR(p) model fitted to $\{x_t\}$, while σ_0^2 is the minimum one-step mean squared prediction error attainable by any linear predictor. Since it would be difficult to derive the exact values of NPE(p) analytically, we will instead use the approximations to NPE(p) obtained by averaging the values of $\hat{a}'R\hat{a}$ over 100 simulated realizations.

Table 3 gives average values of NPE($\hat{p}_{\text{AIC}}$), NPE($\hat{p}_{\text{AIC}_C}$) and NPE($\hat{p}_{\text{SIC}}$) using the same 100 realizations of the MA(1) process as reported earlier, with $\sigma_0^2 = 1$. The results are reasonably similar to those found in Table 1 for the Kullback-Leibler discrepancy, with AIC_C performing uniformly best. However, AIC and AIC_C are much closer in terms of average NPE than they were in terms of average Kullback-Leibler discrepancy. This is particularly true for the larger sample sizes, 75 and 100. An explanation, revealed by examining plots of NPE(p), not shown here, is that NPE places more penalty on small

Table 3. *Average normalized prediction errors,* NPE

n	Ave {NPE($\hat{p}_{\text{AIC}}$)}	Ave {NPE($\hat{p}_{\text{AIC}_C}$)}	Ave{NPE($\hat{p}_{\text{SIC}}$)}
23	5·154	0·491	3·492
30	1·469	0·418	0·665
40	0·555	0·354	0·390
50	0·411	0·322	0·361
75	0·266	0·254	0·299
100	0·219	0·214	0·258

Based on 100 realizations from an MA(1) process.

Table 4. *One-sided p-values for one-sample Wilcoxon test based on differences of normalized prediction errors,* NPE($\hat{p}_L$) − NPE($\hat{p}_R$) *for selection criteria* (L, R)

n	L	R	No. neg.	No. zero	No. pos.	p-value
23	AIC	AIC_C	1	4	95	*
23	SIC	AIC_C	0	36	64	*
30	AIC	AIC_C	4	26	70	*
30	SIC	AIC_C	1	69	30	*
40	AIC	AIC_C	13	34	53	*
40	SIC	AIC_C	2	68	30	*
50	AIC	AIC_C	12	37	51	*
50	SIC	AIC_C	5	56	39	*
75	AIC	AIC_C	12	58	30	0·0011
75	SIC	AIC_C	10	30	60	*
100	AIC	AIC_C	13	60	27	0·023
100	SIC	AIC_C	3	33	64	*

Numbers neg., zero, pos. denote the number of negative, zero and positive differences in NPE($\hat{p}_L$) − NPE($\hat{p}_R$) for the 100 realizations.
* p-value less than 10^{-5}.

model orders, and much less penalty on large model orders, than does the Kullback-Leibler discrepancy Δ.

Table 4 gives p-values for Wilcoxon tests on differences of the form $\text{NPE}(\hat{p}_L) - \text{NPE}(\hat{p}_R)$ for pairs (L, R) of selection criteria; AIC_C is strongly superior to both AIC and SIC in terms of normalized prediction error for all cases studied. Compared with the case of the Kullback-Leibler criterion, Table 2, the superiority of AIC_C over AIC is somewhat weaker here for $n = 75$ and $n = 100$, while the superiority of AIC_C over SIC is stronger here than before. Both phenomena can be explained as above, since AIC tends to overfit and SIC to underfit, compared with AIC_C.

ACKNOWLEDGEMENT

The authors are grateful to the referee for suggesting the consideration of prediction error in the simulation study.

REFERENCES

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, Ed. B. N. Petrov and F. Csaki, pp. 267-81. Budapest: Akademia Kiado.

BERK, K. (1974). Consistent autoregressive spectral estimates. *Ann. Statist.* **2**, 489-502.

BLOOMFIELD, P. (1976). *Fourier Analysis of Time Series: An Introduction.* New York: Wiley.

BURG, J. P. (1978). A new analysis technique for time series data. In *Modern Spectrum Analysis*, ed. D. G. Childers, pp. 42-8, New York: IEEE Press.

FINDLEY, D. (1985). On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. *J. Time Ser. Anal.* **6**, 229-52.

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

LINHART, H. & ZUCCHINI, W. (1986). *Model Selection.* New York: Wiley.

PARZEN, E. (1978). Some recent advances in time series modeling. In *Modern Spectrum Analysis*, Ed. D. G. Childers, pp. 226-33. New York: IEEE Press.

PARZEN, E. (1983). Autoregressive spectral estimation. In *Handbook of Statistics*, **3**, Ed. D. R. Brillinger and P. R. Krishnaiah, pp. 221-47. New York: Elsevier.

RAO, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. New York: Wiley.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-4.

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147-64.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

SUGIURA, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist.* A **7**, 13-26.

[*Received June* 1990. *Revised November* 1990]