Regression and Time Series Model Selection in Small Samples
Author(s): Clifford M. Hurvich and Chih-Ling Tsai
Source: *Biometrika*, Vol. 76, No. 2 (Jun., 1989), pp. 297-307
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: https://www.jstor.org/stable/2336663
Accessed: 18-02-2020 03:08 UTC

Biometrika (1989), **76**, 2, *pp.* 297–307
Printed in Great Britain

Regression and time series model selection in small samples

BY CLIFFORD M. HURVICH
Department of Statistics and Operations Research, New York University, New York NY 10003, U.S.A.

AND CHIH-LING TSAI
Division of Statistics, University of California, Davis, California 95616, U.S.A.

SUMMARY

A bias correction to the Akaike information criterion, AIC, is derived for regression and autoregressive time series models. The correction is of particular use when the sample size is small, or when the number of fitted parameters is a moderate to large fraction of the sample size. The corrected method, called AIC_C, is asymptotically efficient if the true model is infinite dimensional. Furthermore, when the true model is of finite dimension, AIC_C is found to provide better model order choices than any other asymptotically efficient method. Applications to nonstationary autoregressive and mixed autoregressive moving average time series models are also discussed.

Some key words: AIC; Asymptotic efficiency; Kullback–Leibler information.

1. INTRODUCTION

The problems of regression and autoregressive model selection are closely related. Indeed, many of the proposed solutions can be applied equally well to both problems. One of the leading selection methods, and the primary focus of this paper, is the Akaike information criterion, AIC (Akaike, 1973). This was designed to be an approximately unbiased estimator of the expected Kullback–Leibler information of a fitted model. The minimum-AIC criterion produces a selected model which is, hopefully, close to the best possible choice.

If the true model is infinite dimensional, a case which seems most realistic in practice, AIC provides an asymptotically efficient selection of a finite dimensional approximating model. If the true model is finite dimensional, however, the asymptotically efficient methods, e.g., Akaike's FPE (Akaike, 1970), AIC, and Parzen's CAT (Parzen, 1977), do not provide consistent model order selections. Consistency can be obtained (Hannan & Quinn, 1979; Schwarz, 1978) only at the cost of asymptotic efficiency. We feel that of the two properties, asymptotic efficiency is the more desirable. Nevertheless, the existing efficient methods suffer some severe shortcomings, which become particularly evident in the finite dimensional case. The methods tend to overfit severely unless strong restrictions are placed on the maximum allowable dimension of the candidate models. The imposition of such cut-offs, moreover, seems arbitrary and is especially problematic when the sample size is small.

In the case of AIC, the cause of the overfitting problem becomes evident when one examines plots of AIC and the actual Kullback–Leibler information for the various candidate models. As m, the dimension of the candidate model, increases in comparison

to n, the sample size, AIC becomes a strongly negatively biased estimate of the information. This bias can lead to overfitting, even if a maximum cut-off is imposed. The bias of AIC may be attributed to the progressive deterioration, as m/n is increased, in the accuracy of certain Taylor series expansions for the information used in the derivation of AIC.

In this paper, we will obtain a bias-corrected version of AIC for nonlinear regression and autoregressive time series models. We achieve this by extending the applicability of the corrected AIC, AIC_C, method originally proposed for linear regression models by Sugiura (1978); AIC_C is asymptotically efficient, in both regression and time series. For linear regression, AIC_C is exactly unbiased, assuming that the candidate family of models includes the true model. For nonlinear regression and time series models, the unbiasedness of AIC_C is only approximate, since the motivation for AIC_C in these cases is based on asymptotic theory. In all cases, the reduction in bias is achieved without any increase in variance, since AIC_C may be written as the sum of AIC and a nonstochastic term. We explore the performance of AIC_C in small samples, by means of simulations in which the true model is finite dimensional. We find that the bias reduction of AIC_C compared to AIC is quite dramatic, as is the improvement in the selected model orders. Furthermore, a maximum model order cut-off is not needed for AIC_C. Among the efficient methods studied AIC_C is found to perform best. For small samples, AIC_C is able to out-perform even the consistent methods. In view of the theoretical and simulation results, we argue that AIC_C should be used routinely in place of AIC for regression and autoregressive model selection. In addition, we present simulation results demonstrating the effectiveness of AIC_C for selection of nonstationary autoregressive and mixed autoregressive-moving average time series models.

The remainder of this paper is organized as follows. Section 2 develops AIC_C for general regression models, and presents Monte Carlo results for linear regression model selection. Section 3 develops AIC_C and presents simulation results for autoregressive model selection. The criteria for regression and autoregressive models have exactly the same form. Section 4 gives concluding remarks. An appendix outlines the derivation of AIC_C for autoregressive models.

2. Model selection for regression

Here, we follow essentially the notation of Linhart & Zucchini (1986). Suppose data are generated by the operating model, i.e. true model,

$$y = \mu + \varepsilon, \tag{1}$$

where

$$y = (y_1, \ldots, y_n)^\text{T}, \quad \mu = (\mu_1, \ldots, \mu_n)^\text{T}, \quad \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\text{T},$$

and the ε_i are independent identically distributed normal random variables with mean zero and variance σ_0^2. Additional assumptions about the form of the operating model will be made below. Consider the approximating, or candidate, family of models

$$y = h(\theta) + u, \tag{2}$$

where θ is an $m \times 1$ vector,

$$u = (u_1, \ldots, u_n)^\text{T}, \quad h(\theta) = (h_1(\theta), \ldots, h_n(\theta))^\text{T},$$

h is assumed to be twice continuously differentiable in θ, and the u_i are independent identically distributed normal with mean zero and variance σ^2. We refer to (2) as a

model, or alternatively as a family of models, one model for each particular value of (θ, σ^2). In the special case that the approximating family and operating model are both linear, we have $h(\theta) = X\theta$, $\mu = X_0\theta_0$, where X and X_0 are respectively $n \times m$ and $n \times m_0$ matrices of full rank, and θ_0 is an $m_0 \times 1$ parameter vector. A useful measure of the discrepancy between the operating and approximating models is the Kullback–Leibler information

$$\Delta(\theta, \sigma^2) = E_F\{-2 \log g_{\theta,\sigma^2}(y)\},$$

where F denotes the operating model and $g_{\theta,\sigma^2}(y)$ denotes the likelihood function under the approximating model. We have

$$\begin{aligned}\Delta(\theta, \sigma^2) &= -2E_F \log \{(2\pi\sigma^2)^{-\frac{1}{2}n} \exp [-\{y - h(\theta)\}^{\mathrm{T}}\{y - h(\theta)\}/(2\sigma^2)]\} \\ &= n \log (2\pi\sigma^2) + E_F\{\mu + \varepsilon - h(\theta)\}^{\mathrm{T}}\{\mu + \varepsilon - h(\theta)\}/\sigma^2 \\ &= n \log (2\pi\sigma^2) + n\sigma_0^2/\sigma^2 + \{\mu - h(\theta)\}^{\mathrm{T}}\{\mu - h(\theta)\}/\sigma^2.\end{aligned}$$

A reasonable criterion for judging the quality of the approximating family in the light of the data is $E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$, where $\hat{\theta}$ and $\hat{\sigma}^2$ are the maximum likelihood estimates of θ and σ^2 in the approximating family: $\hat{\theta}$ minimizes $\{y - h(\theta)\}^{\mathrm{T}}\{y - h(\theta)\}$, and

$$\hat{\sigma}^2 = \{y - h(\hat{\theta})\}^{\mathrm{T}}\{y - h(\hat{\theta})\}/n.$$

Ignoring the constant $n \log (2\pi)$, we have

$$\Delta(\hat{\theta}, \hat{\sigma}^2) = n \log \hat{\sigma}^2 + n\sigma_0^2/\hat{\sigma}^2 + \{\mu - h(\hat{\theta})\}^{\mathrm{T}}\{\mu - h(\hat{\theta})\}/\hat{\sigma}^2.$$

Given a collection of competing approximating families, then, the one which minimizes $E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$ is, in a sense, closest to the truth, and is to be preferred. Of course, $E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$ is unknown, but it can be estimated if certain additional assumptions are made. The Akaike information criterion

$$\text{AIC} = n(\log \hat{\sigma}^2 + 1) + 2(m + 1), \tag{3}$$

where m is the dimensionality of the approximating model, was designed to provide an approximately unbiased estimate of $E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$.

We now assume that the approximating family includes the operating model. This is a strong assumption, but it is also used in the derivation of AIC (Linhart & Zucchini, 1986, p. 245). In this case, the mean response function μ of the operating model can be written as $\mu = h(\theta^*)$, where θ^* is an $m \times 1$ unknown vector. The linear expansion of $h(\hat{\theta})$ at $\theta = \theta^*$ is given by

$$h(\hat{\theta}) \doteq h(\theta^*) + V(\hat{\theta} - \theta^*),$$

where $V = \partial h/\partial\theta$ evaluated at $\theta = \theta^*$. Then under the operating model, $\hat{\theta} - \theta^*$ is approximately multivariate normal, $N\{0, \sigma_0^2(V^{\mathrm{T}}V)^{-1}\}$, the quantity $n\hat{\sigma}^2/\sigma_0^2$ is approximately distributed as χ^2_{n-m} independently of $\hat{\theta}$ (Gallant, 1986, p. 17), and

$$\begin{aligned}\left(\frac{n-m}{nm}\right)\frac{1}{\hat{\sigma}^2}\{\mu - h(\hat{\theta})\}^{\mathrm{T}}\{\mu - h(\hat{\theta})\} &= \left(\frac{n-m}{nm}\right)\frac{1}{\hat{\sigma}^2}\{h(\theta^*) - h(\hat{\theta})\}^{\mathrm{T}}\{h(\theta^*) - h(\hat{\theta})\} \\ &\doteq \left(\frac{n-m}{nm}\right)\frac{1}{\hat{\sigma}^2}(\hat{\theta} - \theta^*)^{\mathrm{T}}V^{\mathrm{T}}V(\hat{\theta} - \theta^*)\end{aligned}$$

is approximately distributed as $F(m, n-m)$. Thus,

$$E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\} \doteq E_F(n \log \hat{\sigma}^2) + n^2/(n-m-2) + nm/(n-m-2).$$

Consequently,

$$\text{AIC}_\text{C} = n \log \hat{\sigma}^2 + n\frac{1+m/n}{1-(m+2)/n}$$

is an approximately unbiased estimator of $E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$. An equivalent form is

$$\text{AIC}_\text{C} = \text{AIC} + \frac{2(m+1)(m+2)}{n-m-2}. \tag{4}$$

Thus, AIC_C is the sum of AIC and an additional nonstochastic penalty term,

$$2(m+1)(m+2)/(n-m-2).$$

If the approximating models are linear, it follows from Shibata (1981, p. 53) that AIC_C is asymptotically efficient.

For the remainder of this section, we assume for simplicity that the approximating family and operating model are both linear; $h(\theta) = X\theta$, $\mu = X_0\theta_0$. If the approximating family includes the operating model, then $V = X$ and $\mu = X\theta^*$. In this case, AIC_C is an exactly unbiased estimator of $E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$, as originally given for the linear regression case by Sugiura (1978, eqn (3.5)). Curiously Sugiura (1978) did not explore the small-sample performance of AIC_C, and indeed for the two data sets he examined, AIC and AIC_C produced identical selections.

To compare the small-sample performance of various selection criteria in the linear regression case, 100 realizations were generated from model (1) with $\mu = X_0\theta_0$, $m_0 = 3$, $\theta_0 = (1, 2, 3)^\text{T}$ and $\sigma_0^2 = 1$. Two sample sizes were used: $n = 10$ and $n = 20$. There were seven candidate variables, stored in an $n \times 7$ matrix X of independent identically distributed normal random variables. The candidate models were linear, and included the columns of X in a sequentially nested fashion; i.e. the candidate model of dimension m consisted of columns $1, \ldots, m$ of X. The true model consisted of X_0, the first 3 columns of X.

For each realization, the following criteria were used to select a value of m: AIC_C, equation (4); AIC, equation (2); FPE (Akaike, 1970, eqn (4.7)); FPE4 (Bhansali & Downham, 1977, p. 547); HQ (Hannan & Quinn, 1979, p. 191); SIC (Schwarz, 1978; Priestley, 1981, p. 376); CP (Mallows, 1973, eqn (3)); and PRESS (Allen, 1974, p. 126). Of these criteria, HQ and SIC are consistent (Shibata, 1986), and AIC_C, AIC, FPE, CP are asymptotically efficient (Shibata, 1981, p. 53).

For $n = 10$, the left-hand side of Table 1 gives the frequency of the order selected by the various criteria. Here, AIC_C clearly provides the best selection of m among all criteria

Table 1. *Frequency of order selected by various criteria in* 100 *realizations of regression model with* $m_0 = 3$, $n = 10$, 20

	Selected model order, m											
	2	3	4	5	6	7	2	3	4	5	6	7
Criterion	$n = 10$						$n = 20$					
AIC_C	2	96	2	0	0	0	0	88	9	2	1	0
AIC	0	36	8	6	16	34	0	64	13	9	7	7
FPE	0	46	12	12	9	21	0	68	13	8	6	5
FPE4	0	57	10	9	7	17	0	87	7	5	1	0
HQ	0	24	11	13	13	39	0	70	12	8	6	4
SIC	0	41	9	10	11	29	0	84	8	5	1	2
CP	0	61	8	8	6	17	0	77	11	7	3	2
PRESS	1	58	11	12	7	11	0	75	13	8	3	1

studied. The other criteria often show a tendency to overfit the model. We focus now on a comparison between AIC and AIC_C, both of which are designed to be estimates of the Kullback-Leibler discrepancy. Figure 1 plots the average values of AIC, AIC_C and $\Delta(\hat{\theta}, \hat{\sigma}^2)$, DELTA, as functions of m. For $m > m_0$, AIC is a strongly negatively biased estimator of $E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$. As m is increased beyond m_0, AIC first reaches a local maximum and then decreases, eventually falling below the value for $m = m_0$. In contrast, the shape of AIC_C tends to mirror that of $\Delta(\hat{\theta}, \hat{\sigma}^2)$, particularly for $m \geq m_0$, a region in which AIC_C is exactly unbiased for $E_F\{\Delta(\hat{\theta}, \hat{\sigma}^2)\}$. The average value of AIC_C attains a global minimum at the correct value, $m = 3$.

For $n = 20$ in Table 1, AIC_C still provides the best selection of m, although several of the other methods also performed well. Among the efficient criteria, AIC_C strongly outperformed its competitors.

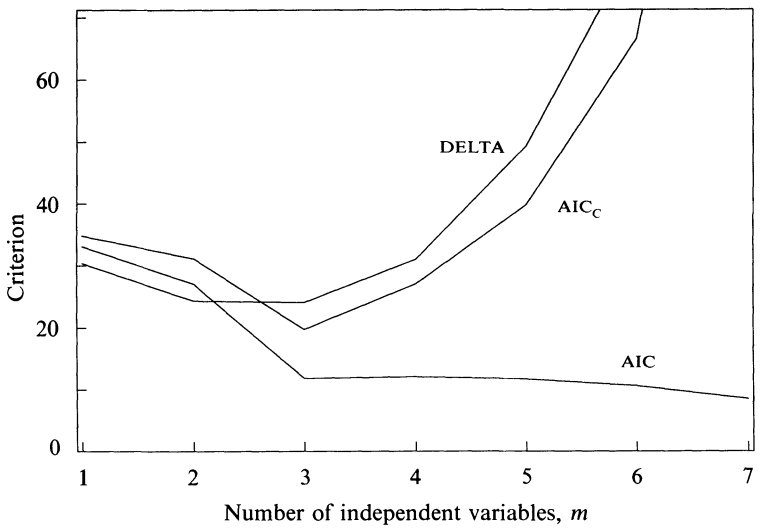

Fig. 1. Average criterion functions and Kullback-Leibler discrepancy in 100 realizations from a regression model with $m_0 = 3$ and $n = 10$.

3. Model selection for autoregression

Suppose that time series data $x_0, \ldots, x_{n-1}$ are generated from a Gaussian zero-mean weakly stationary stochastic process. The approximating model is an order-m autoregressive model with parameters $\hat{a} = (1, \hat{a}_1, \ldots, \hat{a}_m)^T$ and white noise variance $\hat{P}_m$ fitted to the data by maximum likelihood or some other asymptotically equivalent method, e.g. least-squares or Burg's (1978) method. The AIC criterion for selecting an autoregressive model is given by

$$\text{AIC} = n(\log \hat{P}_m + 1) + 2(m+1).$$

In the Appendix it is shown that, if the approximating family includes the operating model, an approximately unbiased estimator of the Kullback-Leibler discrepancy is given by

$$AIC_C = n \log \hat{P}_m + n\frac{1 + m/n}{1 - (m+2)/n}.$$

This has exactly the same form as the version of AIC_C obtained earlier for regression. Also as in the regression case, AIC_C and AIC are related by (4).

To examine small-sample performance, 100 realizations were generated from the second-order autoregressive model

$$x_t = 0{\cdot}99x_{t-1} - 0{\cdot}8x_{t-2} + \varepsilon_t \quad (t = 0, \ldots, n-1),$$

with ε_t independent identically distributed standard normal. Two sample sizes were used: $n = 23$ and $n = 30$. For each realization, Burg's method was used to fit candidate autoregressive models of orders $1, \ldots, 20$, and various criteria were used to select from among the candidate models. Most of the criteria examined here are direct generalizations of the corresponding regression criteria with $\hat{P}_m$ used in place of $\hat{\sigma}^2$: AIC_C, AIC, FPE, HQ, SIC. Two additional criteria proposed specifically for time series were also examined: BIC (Akaike, 1978; Priestley, 1981, p. 375), and CAT (Parzen, 1977, eqn (2.9)). The efficient criteria were AIC_C, AIC, FPE (Shibata, 1980), and CAT (Bhansali, 1986). The consistent criteria were HQ, SIC, BIC.

For $n = 23$, Table 2 gives first the frequency of the model orders selected by the criteria. Two different maximum model order cut-offs were used: max = 10, max = 20. For $n = 23$, max = 20, AIC_C performed best, followed closely by BIC, while all other criteria performed poorly. When max was reduced to 10, AIC_C was slightly outperformed by BIC, but AIC_C was still the best of the efficient methods. Figure 2 plots the average values of the Kullback–Leibler discrepancy, AIC_C and AIC as functions of m. The patterns are quite similar to those observed in Fig. 1 for the linear regression case.

Table 2. *Frequency of order selected by various criteria in* 100 *realizations of second-order autoregressive model with* $n = 23$, *first value, and* $n = 30$, *second value*

	Selected model order, m								
	1	2	3–5	6–10	11–20	1	2	3–5	6–10
Criterion			max = 20					max = 10	
AIC_C	6, 1	80, 73	10, 22	4, 3	0, 1	6, 1	80, 74	10, 22	4, 3
AIC	1, 0	7, 31	2, 12	2, 6	88, 51	3, 0	52, 52	19, 28	26, 20
FPE	2, 0	19, 41	5, 17	7, 8	67, 34	3, 0	52, 52	19, 28	26, 20
HQ	1, 0	11, 50	3, 12	4, 7	81, 31	4, 1	56, 64	19, 22	21, 13
SIC	4, 2	31, 82	3, 8	1, 3	61, 5	6, 2	78, 86	9, 9	7, 3
BIC	4, 0	77, 90	10, 8	3, 0	6, 2	5, 0	81, 91	10, 8	4, 1
CAT	2, 0	20, 43	5, 16	8, 10	65, 31	3, 0	54, 58	19, 26	24, 16

For $n = 30$, the second entry in Table 2, AIC_C was strongly outperformed by the consistent methods SIC and BIC, but AIC_C was still the best of the efficient methods, by a wide margin.

In all cases, the value of the maximum cut-off had virtually no effect on the model chosen by AIC_C. For many of the other criteria, however, increasing the value of max tended to lead to increased overfitting of the model. To explore this further, Fig. 3(a), (b) plots the average criterion functions corresponding to the efficient and consistent methods, respectively, for $n = 23$. Except for AIC_C, the shapes corresponding to the efficient methods mirror the shape of AIC, and hence the criteria tend to favour large model orders, while the shape of AIC_C resembles that of DELTA. The consistent methods suffer from this overfitting problem as well, except for BIC.

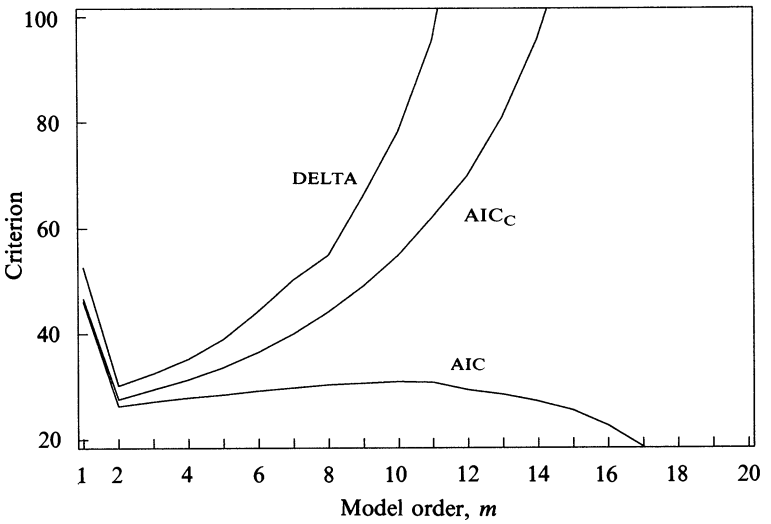

Fig. 2. Average criterion functions and Kullback-Leibler discrepancy in 100 realizations from an autoregressive model with $m_0 = 2$ and $n = 23$.

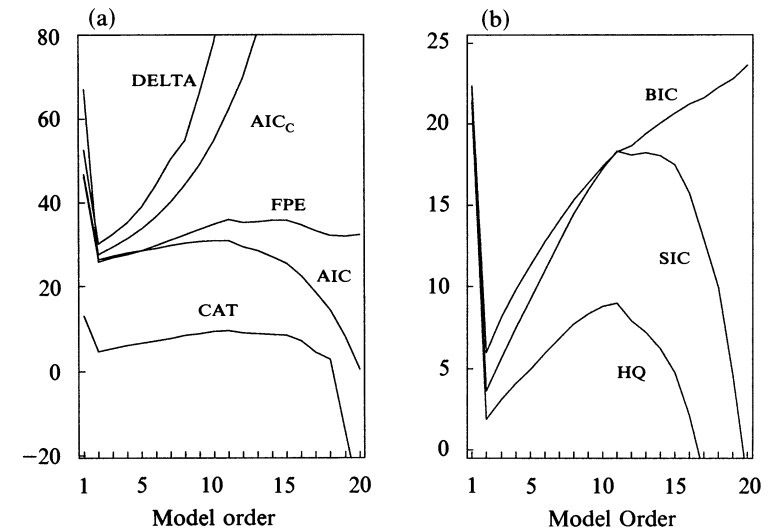

Fig. 3. Average of all criterion functions and Kullback-Leibler discrepancy in 100 realizations from an autoregressive model with $m_0 = 2$ and $n = 23$; (a) Efficient methods, (b) Consistent methods.

4. Discussion

A common pattern in many of the criterion functions studied here is the eventual decline with increasing m, leading to overfitting of the model; see, for example, Fig. 1-3. Here, we show that the expectation of AIC has this pattern, thereby obtaining a partial theoretical explanation for the overfitting problem. In the autoregressive case, if the approximating family includes the operating model, and P_m is the operating white noise variance, then $n\hat{P}_m/P_m$ is approximately distributed as χ^2_{n-m}, and

$$E(\log \hat{P}_m) \doteq \log 2 + \psi(\tfrac{1}{2}n - \tfrac{1}{2}m) + \log (P_m/n),$$

where ψ is the digamma function (Johnson & Kotz, 1970, p. 198, eqn (67)). Thus,

$$E(\text{AIC}) \doteq n\{\log 2 + \psi(\tfrac{1}{2}n - \tfrac{1}{2}m) + \log (P_m/n) + 1\} + 2(m+1). \tag{5}$$

As a function of m, the right-hand side of (5) has the same concave shape as found in the AIC plots of Fig. 1-3. For the linear regression case, (5) is exact, if $\hat{P}_m$, P_m are replaced by $\hat{\sigma}^2$, σ_0^2, respectively.

We have shown that AIC yields a biased estimate of the Kullback-Leibler information, and that this bias tends to cause overfitting of the model, in the cases of regression and autoregressive time series. We have also demonstrated that a bias-correction in AIC is able to overcome the above deficiencies.

Additional time series models in which AIC, SIC and HQ have been applied include nonstationary autoregressions (Tsay, 1984) and mixed autoregressive moving averages, ARMA (Hannan, 1980). Here, we explore the potential applicability of AIC_C for these models, based on theoretical and simulation results.

For stationary autoregressive models, Shibata (1976) obtained the asymptotic distribution of the order selected by AIC. Hannan (1980) and Tsay (1984) generalized Shibata's result to nonstationary autoregressive and ARMA models, respectively. Since the difference between AIC, equation (3), and AIC_C, equation (4), is a nonstochastic term of order $1/n$, Theorem 1 of Shibata (1976, p. 119), Theorem 2 of Hannan (1980, p. 1073) and Theorem 1 of Tsay (1984, p. 1427) can be extended directly to AIC_C.

Next, we present simulation results on the behaviour of AIC_C for nonstationary autoregressive and ARMA model selection. All models were estimated by conditional maximum likelihood. One hundred realizations of the nonstationary third-order autoregression

$$(1-B^2)(1+0{\cdot}95B)x_t = \varepsilon_t$$

were generated, with sample size $n=15$, and ε_t independent identically distributed standard normal. Here, B is the backshift operator, $Bx_t = x_{t-1}$. Table 3 lists the frequency of the order selected by the criteria AIC_C, AIC, HQ and SIC. Of these four criteria, AIC_C performs best. For the case of ARMA model selection, 100 realizations of the first-order moving average model $x_t = \varepsilon_t + 0{\cdot}95\varepsilon_{t-1}$ were generated, with sample size $n=15$, and ε_t independent identically distributed standard normal. Table 4 gives the frequency of the

Table 3. *Frequency of order selected by various criteria in* 100 *realizations of nonstationary third-order autoregressive model with* $n=15$

	Selected model order					
Criterion	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
AIC_C	8	11	45	7	11	18
AIC	3	2	10	11	10	64
HQ	3	2	10	11	10	64
SIC	5	4	19	12	9	51

Table 4. *Frequency of model selected by various criteria in* 100 *realizations of first-order moving average with* $n=15$. *Candidate models are pure autoregressive,* AR, *pure moving average,* MA, *and mixed autoregressive-moving average,* ARMA

Model	AIC_C	AIC	HQ	SIC	Model	AIC_C	AIC	HQ	SIC
AR(1) − AR(4)	5	2	2	1	ARMA(1, 1)	10	4	4	5
AR(5)-AR(6)	13	38	38	31	ARMA(1, 2)	12	7	7	9
MA(1)	20	1	1	5	ARMA(2, 1)	10	1	1	4
MA(2) − MA(4)	10	3	3	5	ARMA(2, 2)	20	24	24	26
MA(5) − MA(10)	0	20	20	14					

model selected by AIC_C, AIC, HQ and SIC. Here, the candidates included a variety of pure autoregressive, pure moving average and mixed ARMA models. Of the four criteria, AIC_C selected the correct model most frequently, in 20 cases. Further, the models selected by AIC_C, although often incorrect, were typically much more parsimonious than those selected by the other criteria.

ACKNOWLEDGMENT

The authors are grateful to the referee for suggesting the additional applications of AIC_C discussed in § 4.

APPENDIX

Derivation of AIC_C *for autoregressive models*

Here, we suppose that univariate time series data $x_0, \ldots, x_{n-1}$ are available. The operating model is that the data form a piece of a realization of a Gaussian zero-mean weakly stationary stochastic process with autocovariance $c_r = E(x_t x_{t-r})$ and spectral density

$$f(\omega) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} c_r \exp(ir\omega).$$

Suppose that $g(\omega)$ is an even nonnegative integrable function on $[-\pi, \pi]$. An approximation due to Whittle (1953, p. 133) is that the corresponding log likelihood $l(g)$ is such that

$$-2l(g) \doteq n \log(2\pi) + \frac{n}{2\pi} \int_{-\pi}^{\pi} \{\log g(\omega) + I(\omega)/g(\omega)\}\, d\omega,$$

where

$$I(\omega) = \frac{1}{2\pi n} \left| \sum_{t=0}^{n-1} x_t \exp(-i\omega t) \right|^2$$

is the periodogram. Since $I(\omega)$ is an asymptotically unbiased estimator of $f(\omega)$, we have

$$E\{-2l(g)\} \doteq d(f, g) = n \log(2\pi) + \frac{n}{2\pi} \int_{-\pi}^{\pi} \{\log g(\omega) + f(\omega)/g(\omega)\}\, d\omega.$$

Thus, the discrepancy function $d(f, g)$ serves as an approximation to the Kullback-Leibler information.

The approximating model is the order-m autoregressive model with parameters $\hat{a} = (1, \hat{a}_1, \ldots, \hat{a}_m)^T$ and white noise variance $\hat{P}_m$, fitted to the data by maximum likelihood or some other asymptotically equivalent method, e.g. least-squares or the Burg method. The resulting approximating spectral density is

$$\hat{f}(\omega) = \frac{\hat{P}_m/(2\pi)}{|\sum \hat{a}_k \exp(i\omega k)|^2},$$

where $\hat{a}_0 = 1$, and the sum is over $k = 0, \ldots, m$.

We now assume that the approximating family includes the operating model. Then the process is an AR(m) process which is potentially degenerate to a lower-order autoregression. Let $a = (1, a_1, \ldots, a_m)^T$, P_m be the solutions to the population Yule-Walker equations $R_m a = (P_m, 0, \ldots, 0)^T$, where

$$R_m = \begin{bmatrix} c_0 & c_1 & \ldots & c_m \\ c_1 & c_0 & \ldots & c_{m-1} \\ \vdots & \vdots & \ldots & \vdots \\ c_m & c_{m-1} & \ldots & c_0 \end{bmatrix}.$$

The final m entries of $\sqrt{n}(\hat{a}-a)$ are asymptotically normal $N(0, P_m R_{m-1}^{-1})$, and $n\hat{P}_m/P_m$ is asymptotically distributed as χ^2_{n-m}, independently of $\hat{a}$ (Brockwell & Davis, 1987, p. 254). Assuming for the sake of mathematical tractability that these asymptotic results are exact for finite n, and using Kolmogorov's formula (Brockwell & Davis, 1987, p. 184) as well as basic properties of the Yule-Walker equations, we have

$$\begin{aligned}
E\{d(f,\hat{f})/n\} &= \log(2\pi) + \frac{1}{2\pi} E\int_{-\pi}^{\pi} \{\log \hat{f}(\omega) + f(\omega)/\hat{f}(\omega)\}\, d\omega \\
&= E(\log \hat{P}_m) + E\int_{-\pi}^{\pi} \frac{f(\omega)}{\hat{P}_m} \left| \sum_{k=0}^{m} \hat{a}_k \exp(i\omega k) \right|^2 d\omega \\
&= E(\log \hat{P}_m) + E(\hat{a}^{\mathrm{T}} R_m \hat{a}/\hat{P}_m) \\
&= E(\log \hat{P}_m) + E\{P_m + (\hat{a}-a)^{\mathrm{T}} R_m (\hat{a}-a)/\hat{P}_m\} \\
&= E(\log \hat{P}_m) + E\left\{\frac{P_m(1+n^{-1}\chi^2_m)}{P_m(n^{-1}\chi^2_{n-m})}\right\} \\
&= E(\log \hat{P}_m) + nE(1/\chi^2_{n-m}) + \frac{m}{n-m} E\{F(m, n-m)\} \\
&= E(\log \hat{P}_m) + \frac{n}{n-m-2} + \frac{m}{n-m}\,\frac{n-m}{n-m-2} \\
&= E(\log \hat{P}_m) + \frac{n+m}{n-m-2}.
\end{aligned}$$

Thus, we obtain the approximately unbiased estimator of $E\{d(f,\hat{f})\}$ as

$$\mathrm{AIC_C} = n \log \hat{P}_m + n\frac{1+m/n}{1-(m+2)/n}.$$

It follows (Shibata, 1980, p. 160) that $\mathrm{AIC_C}$ is asymptotically efficient. Note that $\mathrm{AIC_C}$ obtained here is equivalent to the formula derived in § 2 for the regression case.

REFERENCES

AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22**, 203-17

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.

AKAIKE, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Ann. Inst. Statist. Math.* A **30**, 9-14.

ALLEN, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics* **16**, 125-7.

BHANSALI, R. J. (1986). Asymptotically efficient selection of the order by the criterion autoregressive transfer function. *Ann. Statist.* **14**, 315-25.

BHANSALI, R. J. & DOWNHAM, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* **64**, 547-51.

BROCKWELL, P. J. & DAVIS, R. A. (1987). *Time Series: Theory and Methods.* New York: Springer Verlag.

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.

GALLANT, A. R. (1986). *Nonlinear Statistical Models.* New York: Wiley.

HANNAN, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8**, 1071-81.

HANNAN, E. J. & QUINN, B. G. (1979). The determination of the order of autoregression. *J. R. Statist. Soc.* B **41**, 190-5.

JOHNSON, N. L. & KOTZ, S. (1970). *Continuous Univariate Distributions*-1. New York: Wiley.
LINHART, H. & ZUCCHINI, W. (1986). *Model Selection.* New York: Wiley.
MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **12**, 591–612.
PARZEN, E. (1977). Multiple time series modeling: determining the order of approximating autoregressive schemes. In *Multivariate Analysis IV*, Ed. P. Krishnaiah, pp. 283–95. Amsterdam: North-Holland.
PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series.* New York: Academic Press.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–4.
SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117–26.
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.
SHIBATA, R. (1986). Consistency of model selection and parameter estimation. *J. Appl. Prob.* **23A**, (Essays in time series and allied processes), Ed. J. Gani and M. B. Priestley, pp. 127–41.
SUGIURA, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist.* A**7**, 13–26.
TSAY, R. S. (1984). Order selection in nonstationary autoregressive models. *Ann. Statist.* **12**, 1425–33.
WHITTLE, P. (1953). The analysis of multiple stationary time series. *J.R. Statist. Soc.* B **15**, 125–39.

[*Received May* 1988. *Revised July* 1988]