Statistica Sinica **7**(1997), 473-496

A BOOTSTRAP VARIANT OF AIC FOR STATE-SPACE MODEL SELECTION

Joseph E. Cavanaugh and Robert H. Shumway

University of Missouri and University of California, Davis

Abstract: Following in the recent work of Hurvich and Tsai (1989, 1991, 1993) and Hurvich, Shumway, and Tsai (1990), we propose a corrected variant of AIC developed for the purpose of small-sample state-space model selection. Our variant of AIC utilizes bootstrapping in the state-space framework (Stoffer and Wall (1991)) to provide an estimate of the expected Kullback-Leibler discrepancy between the model generating the data and a fitted approximating model. We present simulation results which demonstrate that in small-sample settings, our criterion estimates the expected discrepancy with less bias than traditional AIC and certain other competitors. As a result, our AIC variant serves as an effective tool for selecting a model of appropriate dimension. We present an asymptotic justification for our criterion in the Appendix.

Key words and phrases: Information theory, Kullback-Leibler information, time series.

1. Introduction

In time series modeling, an investigator is generally confronted with the problem of choosing an appropriate model from a class of candidate models. Many approaches to this problem have been proposed over the last twenty years, stimulated largely by the ground-breaking work of Akaike (1973, 1974). The Akaike information criterion, AIC, remains the most widely known and used tool for time series model selection, although many competitors and variants have gained acceptance since its introduction. Among these are FPE (Akaike (1969)), SIC (Schwarz (1978), Rissanen (1978)), BIC (Akaike (1978)), HQ (Hannan and Quinn (1979)), and more recently, AICc (Hurvich and Tsai (1989)).

AIC is both computationally and heuristically appealing, which partly explains its enduring popularity among practitioners. Yet the criterion suffers from one commonly observed drawback: it has a tendency to favor high dimensional models in a candidate class when the sample size is small relative to the larger model dimensions represented within the class. The development of "corrected" AIC, AICc, was motivated by the need to adjust for this weakness. First suggested by Sugiura (1978) and later investigated and generalized by Hurvich and

Tsai (1989, 1991, 1993) and Hurvich, Shumway, and Tsai (1990), AICc often dramatically outperforms AIC as a selection criterion in small-sample simulation studies. Yet the basic form of AICc is similar to that of AIC, meaning that the improvement in selection performance comes without an increase in computational cost.

Originally proposed for linear regression, AICc has been extended to univariate autoregressive modeling (Hurvich and Tsai (1989)), univariate autoregressive moving-average modeling (Hurvich, Shumway, and Tsai (1990)), and vector autoregressive modeling (Hurvich and Tsai (1993)). The demonstrated effectiveness of AICc as a selection criterion in these settings motivates the need for a corrected variant of AIC for state-space modeling. Yet the derivation of AICc is less general than that of AIC, involving distributional results which do not extend in an obvious manner to the state-space setting without the addition of certain restrictive assumptions. Thus, we propose a criterion which achieves the same degree of effectiveness as AICc, but which can be used within a broad state-space framework. This new AIC variant involves a bootstrap-based correction that can be justified and applied in a very general context, one which includes (but is not limited to) the state-space setting of interest. We call our criterion AICb.

The idea of using the bootstrap to improve the performance of a model selection rule has been suggested and investigated by Efron (1983, 1986), and is discussed by Efron and Tibshirani (1993), Chapter 17. Ishiguro and Sakamoto (1991) proposed an AIC variant called WIC based on Efron's methodology, and Ishiguro, Morita, and Ishiguro (1991) used this variant successfully in an aperture synthesis imaging problem. In a recent manuscript, Shibata (1997) proves the asymptotic equivalence of AICb and WIC under a general set of assumptions, and indicates the existence of other asymptotically equivalent bootstrap-based AIC variants. In small-sample settings, the type of variant which would perform optimally most likely depends on the nature of the modeling problem. For our state-space application of interest, our simulation results indicate that AICb outperforms WIC, although further investigation is needed before substantive conclusions can be drawn.

In Section 2, we briefly review the motivation behind AIC, and discuss why the criterion works poorly in small-sample applications. This leads to the introduction of AICb. In Section 3, we present an overview of the state-space model along with a brief discussion of Gaussian maximum likelihood parameter estimation in the state-space setting. Finally, in Sections 4 and 5, we compare the performance of AICb to that of other selection criteria in simulation studies based on small sample sizes. A theoretical asymptotic justification of AICb is presented in the Appendix.

2. Presentation of AICb

A well-known measure of separation between two models is given by the non-normalized Kullback-Leibler information, also known as the cross entropy or discrepancy. If θ_o represents the set of parameters for the "true" or generating model and θ represents the set of parameters for a candidate or approximating model, the discrepancy between the models is defined as

$$d_n(\theta, \theta_o) = E_o\{-2 \log L(\theta \mid Y_n)\},$$

where E_o denotes the expectation under the generating model, and $L(\theta \mid Y_n)$ represents the likelihood corresponding to the approximating model.

Now for a given set of estimates $\hat{\theta}_n$, we could determine the discrepancy between the fitted approximating model and the generating model if we could evaluate

$$d_n(\hat{\theta}_n, \theta_o) = E_o\{-2 \log L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}. \tag{2.1}$$

Yet evaluating (2.1) is not possible, since it requires knowledge of θ_o. Akaike (1973), however, noted that $-2 \log L(\hat{\theta}_n \mid Y_n)$ serves as a biased estimator of (2.1), and that the bias adjustment

$$E_o\{E_o\{-2 \log L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} - E_o\{-2 \log L(\hat{\theta}_n \mid Y_n)\} \tag{2.2}$$

can often be asymptotically estimated by twice the dimension of $\hat{\theta}_n$. Thus, if we let k represent the dimension of $\hat{\theta}_n$, then under appropriate conditions, the expected value of

$$\text{AIC} = -2 \log L(\hat{\theta}_n \mid Y_n) + 2k$$

should be asymptotically close to the expected value of (2.1), say $\Delta_n(k, \theta_o) = E_o\{d_n(\hat{\theta}_n, \theta_o)\}$. Alternatively, AIC should serve as an asymptotically unbiased estimator of the expected discrepancy $\Delta_n(k, \theta_o)$, where

$$\begin{aligned}\Delta_n(k, \theta_o) &= E_o\{E_o\{-2 \log L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} \\ &= E_o\{-2 \log L(\hat{\theta}_n \mid Y_n)\} + \\ &\quad [E_o\{E_o\{-2 \log L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} - E_o\{-2 \log L(\hat{\theta}_n \mid Y_n)\}]. \end{aligned} \tag{2.3}$$

Note that the "goodness of fit" term in AIC, $-2 \log L(\hat{\theta}_n \mid Y_n)$, estimates the first of the terms in (2.3), whereas the "penalty" term in AIC, $2k$, estimates the bias expression (2.2).

AIC provides us with an approximately unbiased estimator of $\Delta_n(k, \theta_o)$ in settings where n is large and k is comparatively small. In settings where n is small and k is comparatively large (e.g., $k \approx n/2$), $2k$ is often much smaller than the bias adjustment (2.2), making AIC substantially negatively biased as an estimator of $\Delta_n(k, \theta_o)$ (Hurvich and Tsai (1989)). If AIC severely underestimates $\Delta_n(k, \theta_o)$ for higher dimensional fitted models in the candidate set, the criterion may favor the higher dimensional models even when the expected discrepancy between these models and the generating model is rather large. Examples illustrating this phenomenon appear in Shumway (1988), page 169, and in Linhart and Zucchini (1986), pages 86-88, who comment (page 78) that "in some cases the criterion simply continues to decrease as the number of parameters in the approximating model is increased."

AICc was developed to yield an estimator of $\Delta_n(k, \theta_o)$ which is less biased in small-sample applications than traditional AIC (Hurvich and Tsai (1989)). Our criterion achieves the same goal through utilizing the bootstrap. Specifically, we propose a bootstrap-based estimator for the bias adjustment (2.2), which in small-sample settings should estimate (2.2) more accurately than $2k$.

Suppose that $\{\hat{\theta}_n^*(i);\ i = 1, \ldots, N\}$ represents a set of N bootstrap replicates of $\hat{\theta}_n$. Let E_* denote the expectation with respect to the bootstrap distribution of $\hat{\theta}_n^*$. In the Appendix, we show that under suitable conditions, the difference between

$$2\ [E_*\{-2 \log L(\hat{\theta}_n^* \mid Y_n)\} - \{-2 \log L(\hat{\theta}_n \mid Y_n)\}]$$

and the bias expression (2.2) converges almost surely to zero (as $n \to \infty$). This follows from the observation that (2.2) can be decomposed into the sum of

$$E_o\{E_o\{-2 \log L(\theta \mid Y_n)\}|_{\theta=\hat{\theta}_n}\} - E_o\{-2 \log L(\theta_o \mid Y_n)\} \tag{2.4}$$

and

$$E_o\{-2 \log L(\theta_o \mid Y_n)\} - E_o\{-2 \log L(\hat{\theta}_n \mid Y_n)\}, \tag{2.5}$$

and that the difference between

$$E_*\{-2 \log L(\hat{\theta}_n^* \mid Y_n)\} - \{-2 \log L(\hat{\theta}_n \mid Y_n)\} \tag{2.6}$$

and either (2.4) or (2.5) tends almost surely to zero (as $n \to \infty$).

Now by the strong law of large numbers, as $N \to \infty$, $N^{-1} \sum_{i=1}^{N} -2 \log L(\hat{\theta}_n^*(i) \mid Y_n)$ converges almost surely to $E_*\{-2 \log L(\hat{\theta}_n^* \mid Y_n)\}$. Thus, for $N \to \infty$,

$$\Big\{\frac{1}{N} \sum_{i=1}^{N} -2 \log L(\hat{\theta}_n^*(i) \mid Y_n)\Big\} - \{-2 \log L(\hat{\theta}_n \mid Y_n)\}$$

is almost surely the same as (2.6). This leads us to the following large-sample estimator of $\Delta_n(k, \theta_o)$:

$$\text{AICb} = -2 \log L(\hat{\theta}_n \mid Y_n)$$

$$+2\Big[\{\frac{1}{N}\sum_{i=1}^{N}-2\log L(\hat{\theta}_n^*(i)\ |Y_n)\}-\{-2\log L(\hat{\theta}_n\ |Y_n)\}\Big]$$
$$=-2\log L(\hat{\theta}_n\ |Y_n)+2\left\{\frac{1}{N}\sum_{i=1}^{N}-2\log\frac{L(\hat{\theta}_n^*(i)\ |Y_n)}{L(\hat{\theta}_n\ |Y_n)}\right\}.$$

Note that AICb is composed of the same "goodness of fit" term as AIC, and an inherently positive "penalty" term which is asymptotically equivalent to the bias term in (2.3) (as both n, $N \to \infty$).

We should note that the asymptotic justifications of AIC and AICc both involve the assumption that the "true" parameter vector θ_o corresponds to a model in the candidate class. (See Hurvich and Tsai (1989).) This is admittedly a strong assumption, and one which is also required in our asymptotic justification of AICb. The behavior of AIC and AICc when this condition is not met has been investigated by Hurvich and Tsai (1991). In future work, we hope to explore the same issue with regard to AICb.

Our asymptotic defense of AICb demonstrates that our criterion fulfills the same large-sample objective as AIC, in that it provides an approximately unbiased estimator of $\Delta_n(k,\theta_o)$. Yet the computational burden required to evaluate AICb is justifiable only if it can be shown that AICb is superior to AIC in settings where the sample size is small enough to cast doubt on asymptotic arguments. Thus, in Sections 4 and 5, we describe and present a collection of simulation results to examine the small-sample behavior of AICb and AIC, as well as that of certain other criteria of interest.

3. The State-Space Model and Gaussian ML Estimation

The state-space model has the form

$$y_t = Ax_t + v_t, \tag{3.1}$$
$$x_t = \Phi x_{t-1} + w_t, \tag{3.2}$$

for $t = 1,\ldots,n$ time periods, where y_t is an observed vector process, x_t is an unobserved vector state process, A is a known design matrix, Φ is an unknown transition matrix, and v_t and w_t are zero-mean vector noise processes. Equations (3.1) and (3.2) are respectively called the observation equation and the state equation. We let R denote the covariance matrix of the observation noise process v_t, and let Q denote the covariance matrix of the state noise process w_t. We also let μ and Σ respectively denote the mean and covariance matrix of the initial state x_o.

It is routinely assumed that

- x_o, the w_t, and the v_t are mutually independent, (3.3) and often additionally assumed that

- x_o, the w_t, and the v_t are multivariate normal. (3.4)

To represent the unknown parameters, we let θ denote a kx1 vector that uniquely determines the model coefficients and correlation structure: i.e., $\mu \equiv \mu(\theta)$, $\Sigma \equiv \Sigma(\theta)$, $\Phi \equiv \Phi(\theta)$, $Q \equiv Q(\theta)$, $R \equiv R(\theta)$. We let Y_t denote the observed data up until time t (i.e., $Y_t = [y_1, \ldots, y_t]$).

The likelihood $L(\theta \,|Y_n)$ is generally written in its *innovations form* (Schweppe (1965)). The innovation at time t is defined as

$$e_t(\theta) = y_t - Ax_t^{t-1}(\theta) \quad \text{where} \quad x_t^{t-1}(\theta) = E(x_t \,|Y_{t-1}).$$

We let $\Sigma_t(\theta)$ denote the covariance matrix of $e_t(\theta)$. (Note that $E(e_t(\theta)) = 0$.) The well-known Kalman filter equations (Kalman (1960), Kalman and Bucy (1961)) provide us with a recursive algorithm for evaluating successive values of $e_t(\theta)$ and $\Sigma_t(\theta)$, as well as the state estimators $x_t^{t-1}(\theta)$ and $x_t^t(\theta) = E(x_t \,|Y_t)$ and their respective covariance matrices $P_t^{t-1}(\theta)$ and $P_t^t(\theta)$. The starting values $x_o^o(\theta) = \mu$ and $P_o^o(\theta) = \Sigma$ are used to initialize the filter.

Under the assumptions (3.3) and (3.4), the innovations are mutually independent and multivariate normal. Thus, for the log of the likelihood $L(\theta \,|Y_n)$, we can write

$$\log L(\theta \,|Y_n) \propto -\frac{1}{2}\sum_{t=1}^{n} \log|\Sigma_t(\theta)| - \frac{1}{2}\sum_{t=1}^{n} e_t(\theta)'\Sigma_t^{-1}(\theta)e_t(\theta). \tag{3.5}$$

Since (3.5) is generally a highly non-linear function of the parameters, the maximum likelihood estimates are usually found by using an iterative optimization algorithm. Maximum likelihood estimation can also be carried out via the EM algorithm. Details are provided in Shumway and Stoffer (1982).

Henceforth, we assume $\hat{\theta}_n$ denotes the set of Gaussian maximum likelihood (GML) estimates for the kx1 vector θ.

If the normality assumption (3.4) is not imposed, $L(\theta \,|Y_n)$ does not represent the joint density of the innovations. In this case, $\hat{\theta}_n$ is viewed as the set of estimates which minimizes the loss function $-\log L(\theta \,|Y_n)$. Although our asymptotic justification of AICb assumes that the parameter estimates are obtained through Gaussian maximum likelihood (or some asymptotically equivalent method), our development does not require (3.4). Thus, we expect AICb to be fairly robust to violations of this assumption.

A nonparametric bootstrap procedure for the state-space model is presented as a four-step algorithm by Stoffer and Wall (1991). In the first step, the estimated innovations $e_t(\hat{\theta}_n)$ are evaluated and standardized. In the second step, the standardized innovations are resampled, and in the third step, the resampled innovations are used to construct a bootstrap sample of y_t's, say $Y_n^*(i)$. This construction is accomplished through utilizing analogues of equations (3.1) and (3.2), where the terms v_t and w_t are replaced by functions of the innovations. In the fourth step, the bootstrap sample $Y_n^*(i)$ is used to compute a bootstrap GML vector $\hat{\theta}_n^*(i)$. Repeating steps two through four N times results in a sample of bootstrap GML vectors $\{\hat{\theta}_n^*(i);\ i = 1, \ldots, N\}$. The sampling distribution of $\hat{\theta}_n$ is estimated by the relative frequency distribution of the $\hat{\theta}_n^*(i)$.

Stoffer and Wall (1991) establish that the asymptotic behavior of the bootstrap GML estimator $\hat{\theta}_n^*$ is the same as that of the GML estimator $\hat{\theta}_n$. Their justification relies upon an asymptotic theory proposed by Ljung and Caines (1979) for a general class of estimators. We utilize results from both Stoffer and Wall (1991) and Ljung and Caines (1979) in providing a formal asymptotic justification for AICb in the Appendix.

4. Description of Simulations

Two different types of time series models are used in our simulation sets: the univariate autoregressive model, and the univariate autoregressive model with observation noise. The univariate autoregressive model of order p can be written as

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + \epsilon_t; \quad \epsilon_t \ \sim \ i.i.d.\ (0, \sigma_Q^2).$$

We denote this model as AR(p). The univariate autoregressive model of order p with observation noise can be written as

$$y_t = z_t + v_t; \quad v_t \ \sim \ i.i.d.\ (0, \sigma_R^2);$$

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + \epsilon_t; \quad \epsilon_t \ \sim \ i.i.d.\ (0, \sigma_Q^2).$$

We denote this model as ARN(p).

The ARN(p) model is expressed in state-space form by writing the observation equation (3.1) as

$$y_t = (1, 0, \ldots, 0) \begin{pmatrix} z_t \\ z_{t-1} \\ \vdots \\ z_{t-p+1} \end{pmatrix} + v_t,$$

and the state equation (3.2) as

$$\begin{pmatrix} z_t \\ z_{t-1} \\ \vdots \\ z_{t-p+1} \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} z_{t-1} \\ z_{t-2} \\ \vdots \\ z_{t-p} \end{pmatrix} + \begin{pmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Here, the covariance matrix Q of the state noise vector is a pxp matrix with all zero entries except for the entry in the upper left-hand corner, which is σ_Q^2. The observation noise is scalar, and has variance $R = \sigma_R^2$.

The AR(p) model is expressed in state-space form in the same manner, except that the noise process v_t does not appear in the observation equation.

For each of the models considered in our simulations, the eigenvalues of Φ are all within the unit circle. This ensures that the state process z_t is weakly stationary.

The parameter vectors for the AR(p) and ARN(p) models are, respectively, the $(p+1)$ x 1 and $(p+2)$ x 1 vectors

$$\theta = (\phi_1, \ldots, \phi_p, \sigma_Q^2)' \quad \text{and} \quad \theta = (\phi_1, \ldots, \phi_p, \sigma_R^2, \sigma_Q^2)'.$$

The parameter estimates $\hat{\theta}_n$ are obtained using the EM algorithm (Shumway and Stoffer (1982)). The true parameter values are used to initialize the algorithm. For the initial state vector x_o, the mean vector μ is fixed at zero, and the covariance matrix Σ is found by solving the equation $\Sigma = \Phi\Sigma\Phi' + Q$. (See Harvey (1989), pages 120 and 121.) In fitting the models, the mean of the observed process y_t is subtracted from each y_t, $t = 1, \ldots, n$.

In addition to AICb and AIC, the other criteria considered in our simulations are FPE (Akaike (1969)), SIC (Schwarz (1978), Rissanen (1978)), BIC (Akaike (1978)), HQ (Hannan and Quinn (1979)), AICc (Hurvich and Tsai (1989)), and WIC (Ishiguro, Morita, and Ishiguro (1991)). (The justifications of some of these criteria do not extend in an obvious manner to the state-space setting. Thus, their definitions and usages for ARN(p) model selection are somewhat ad hoc.)

The complete set of criteria is listed below. In the definitions involving k, k is $(p+1)$ when applied to the AR(p) model and $(p+2)$ when applied to the ARN(p) model. The estimate of the steady-state innovations variance is denoted by $\hat{\sigma}_n^2$: i.e., $\hat{\sigma}_n^2 = \Sigma_t(\hat{\theta}_n)$ where t is "large". In the definition of WIC, $Y_n^*(i)$ represents the bootstrap sample corresponding to the bootstrap GML vector $\hat{\theta}_n^*(i)$.

$$\text{AICb} = -2\log L(\hat{\theta}_n \,|Y_n) + 2\left\{\frac{1}{N}\sum_{i=1}^{N} -2\log\frac{L(\hat{\theta}_n^*(i)\ |Y_n)}{L(\hat{\theta}_n\ |Y_n)}\right\} \tag{4.1}$$

$$\text{WIC} = -2\log L(\hat{\theta}_n \,|Y_n) + \left\{\frac{1}{N}\sum_{i=1}^{N} -2\log\frac{L(\hat{\theta}_n^*(i)\ |Y_n)}{L(\hat{\theta}_n^*(i)\ |Y_n^*(i))}\right\} \tag{4.2}$$

$$\text{AIC} = -2\log L(\hat{\theta}_n \,|Y_n) + 2k \tag{4.3}$$

$$\text{AICc} = \left(n\log\hat{\sigma}_n^2 + n\right) + \frac{2n(p+1)}{n-p-2} \tag{4.4}$$

$$\text{FPE} = n \left(\frac{n+k}{n-k}\right) \hat{\sigma}_n^2 \tag{4.5}$$

$$\text{HQ} = n \log \hat{\sigma}_n^2 + 2k \log\log n \tag{4.6}$$

$$\text{BIC} = (n-p) \log \left(\frac{n\hat{\sigma}_n^2}{n-p}\right) + p \log \left\{ \frac{(\sum_{t=1}^{n} y_t^2) - n\hat{\sigma}_n^2}{p} \right\} \tag{4.7}$$

$$\text{SIC} = -2 \log L(\hat{\theta}_n \; | Y_n) + k \log n \tag{4.8}$$

In each simulation set, 100 realizations of size n are generated from a known model of order p_o. For each of the realizations, candidate models of orders 1 through P are fit to the data ($p_o < P$), the criteria (4.1) through (4.8) are evaluated, and the fitted candidate model selected by each criterion is determined. In the computation of AICb and WIC, $N = 250$ bootstrap replications $\hat{\theta}_n^*(i)$ are used. The distribution of selections by each criterion is recorded for the 100 realizations and presented in tabular form. (On an occasional realization, a criterion is minimized for two different model orders. If the minima agree out to two decimal places, the case is treated as a tie, and both selections are recorded.)

For each of the AIC-type criteria (AICb, WIC, AIC, and AICc), the average criterion value over the 100 realizations is computed for each of the candidate model orders 1 through P. The value of $\Delta_n(k, \theta_o)$ is simulated for each of these orders. The averages for AICb, WIC, AIC, and AICc are then compared to the simulated values of $\Delta_n(k, \theta_o)$ by plotting the criterion averages and the simulated $\Delta_n(k, \theta_o)$ against several of the initial orders. Using this approach, we can judge the relative effectiveness of AICb, WIC, AIC, and AICc as unbiased estimators of $\Delta_n(k, \theta_o)$.

5. Presentation of Simulation Results

The first simulation set involves a generating model and sample size originally considered in a simulation study presented by Hurvich and Tsai (1989) to assess the effectiveness of AICc in autoregressive model selection. The model is the AR(2) model

$$z_t = 0.99 z_{t-1} - 0.80 z_{t-2} + \epsilon_t; \quad \epsilon_t \;\sim\; i.i.d.\; N(0,1). \tag{5.1}$$

The sample size is $n = 23$. The candidate class consists of AR(p) models where $1 \leq p \leq 12$.

Although AICc performs well here, Table 1 indicates that AICb results in considerably more correct order selections than any other criterion. Moreover, AICb does not incorrectly select any high dimensional models, whereas most of the other criteria exhibit a propensity to overfit.

Table 1. Model selections for (5.1) (maximum order: 12)

Order	AICb	WIC	AIC	AICc	FPE	HQ	BIC	SIC
1	1	1	1	1	0	1	1	2
2	90	72	45	69	25	23	60	77
3	8	18	11	7	6	5	6	7
4	1	1	0	1	0	0	0	1
5	0	2	4	1	1	1	0	0
6	0	1	1	1	1	2	1	0
7	0	1	4	1	4	3	1	0
8	0	0	2	1	0	0	2	1
9 to 12	0	4	33	18	63	65	29	12

Figure 1 illustrates that over the first eight model orders, the average AICb curve more closely follows the simulated $\Delta_n(k, \theta_o)$ curve than either the average AIC or AICc curve. The WIC curve follows $\Delta_n(k, \theta_o)$ effectively, but WIC results in fewer correct order selections than AICb.

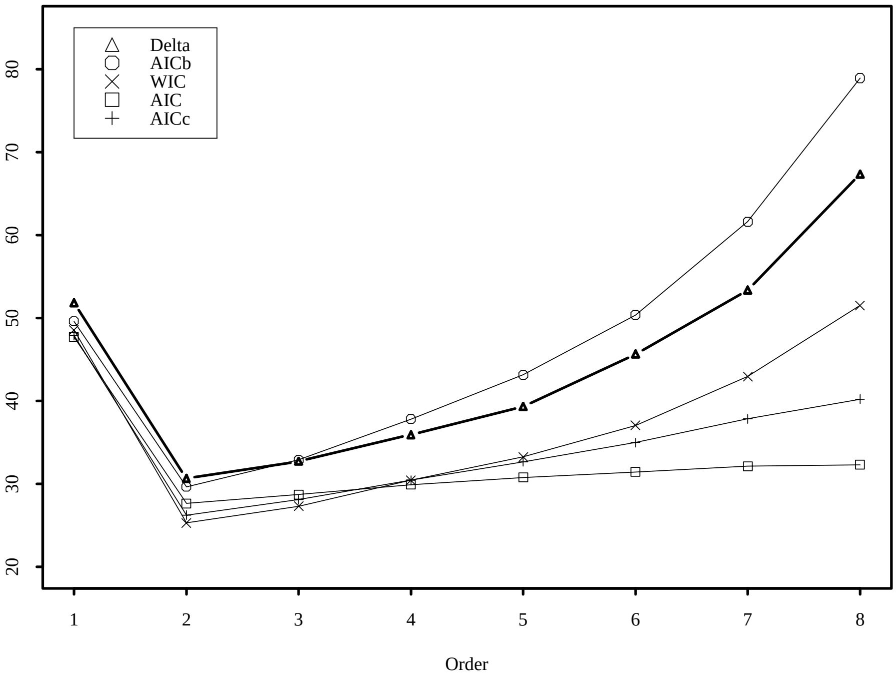

Figure 1. Criterion averages and simulated $\Delta_n(k, \theta_o)$ for (5.1) (orders 1 through 8)

The second simulation set uses as the generating model the ARN(1) model

$$y_t = z_t + v_t; \quad v_t \sim i.i.d.\ N(0, 0.2); \tag{5.2}$$

$$z_t = 0.60 z_{t-1} + \epsilon_t; \quad \epsilon_t \sim i.i.d.\ N(0,1).$$

The sample size is $n = 15$. The candidate class consists of ARN(p) models where $1 \leq p \leq 8$.

As shown in Table 2, AICb obtains the most correct order selections, followed by AICc. Of the remaining criteria, only SIC and WIC perform acceptably, although WIC exhibits a tendency to favor higher dimensional models. AIC, FPE, HQ, and BIC all exhibit this overfitting tendency to an even greater degree than WIC.

Table 2. Model selections for (5.2) (maximum order: 8)

Order	AICb	WIC	AIC	AICc	FPE	HQ	BIC	SIC
1	79	57	45	74	27	18	13	64
2	3	3	5	8	6	4	5	8
3	4	6	3	3	4	2	4	2
4	2	4	6	6	9	6	8	4
5	2	4	5	4	6	5	12	5
6	3	5	6	3	12	15	17	5
7	3	7	13	2	17	17	19	7
8	4	14	17	0	19	33	24	5

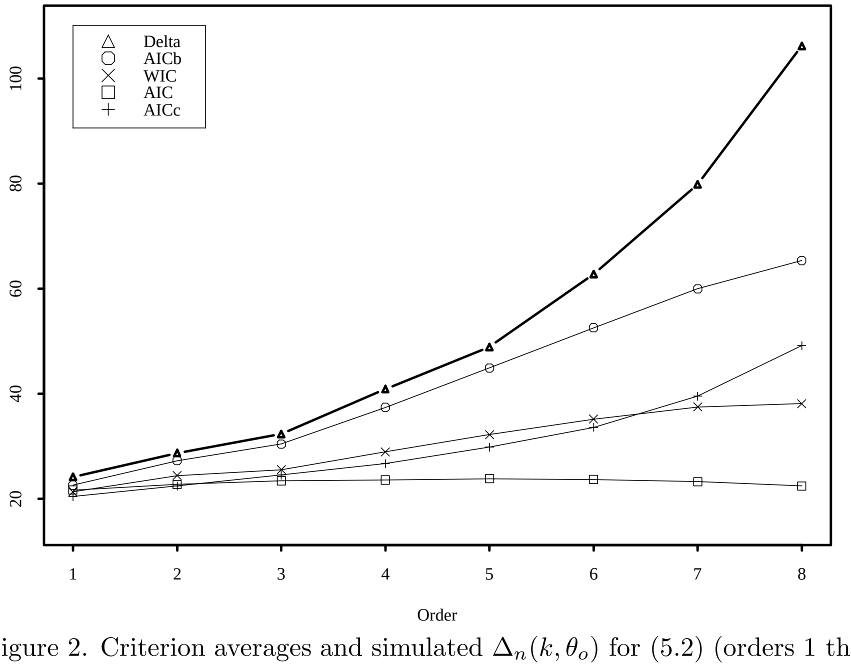

Figure 2. Criterion averages and simulated $\Delta_n(k, \theta_o)$ for (5.2) (orders 1 through 8)

Figure 2 demonstrates that the average AICb and AICc curves reflect the general shape of the simulated $\Delta_n(k, \theta_o)$ curve, although the AICc curve better represents the increasing slope in $\Delta_n(k, \theta_o)$ past model order 6. The AIC curve remains relatively constant over all model orders. The WIC curve initially follows the AICc curve, but appears flat past model order 6.

The third simulation set is based on the generating AR(2) model

$$z_t = 1.40 z_{t-1} - 0.49 z_{t-2} + \epsilon_t; \quad \epsilon_t \sim \textit{i.i.d. } N(0,1). \tag{5.3}$$

The sample size is $n = 50$. The candidate class consists of AR(p) models where $1 \leq p \leq 14$.

Table 3 indicates that SIC and BIC obtain the most correct order selections. The AIC-type criteria all perform comparably due to the relatively larger sample size used in this set: AICb and AICc obtain the most correct selections, followed by WIC and AIC.

Table 3. Model selections for (5.3) (maximum order: 14)

Order	AICb	WIC	AIC	AICc	FPE	HQ	BIC	SIC
1	1	1	3	4	2	5	11	11
2	78	73	73	78	63	79	88	86
3	13	12	9	7	9	5	1	2
4	3	4	8	6	8	6	0	1
5	2	6	0	0	0	0	0	0
6	2	3	3	1	2	1	0	0
7	0	0	1	0	1	0	0	0
8	1	0	2	2	2	1	0	0
9 to 14	0	2	2	2	13	3	0	0

Figure 3 illustrates that over the first eight model orders, the average AICb curve tracks the simulated $\Delta_n(k, \theta_o)$ curve quite closely. The WIC and AICc curves also effectively reflect the general shape of the $\Delta_n(k, \theta_o)$ curve.

The fourth and final simulation set is based on a generating ARN(2) model where the observation noise and state noise have scaled t distributions with five degrees of freedom. This simulation set is included so that the sensitivity of the criteria to the normality assumption (3.4) may be assessed. Since the asymptotic justification of AICb does not require such an assumption, the performance of AICb should not be greatly impaired by using heavy-tailed distributions for the model errors.

The generating model is a modification of the AR(2) model used in the first simulation set:

$$\begin{aligned} y_t &= z_t + v_t; \quad v_t \sim \textit{i.i.d. } t_5(0.15); \\ z_t &= 0.99 z_{t-1} - 0.80 z_{t-2} + \epsilon_t; \quad \epsilon_t \sim \textit{i.i.d. } t_5(1). \end{aligned} \tag{5.4}$$

Here, $t_{d.f.}(\sigma)$ represents a t distribution with $d.f.$ degrees of freedom, scaled to have a standard deviation of σ. The sample size is $n = 23$. The candidate class consists of ARN(p) models where $1 \leq p \leq 12$.

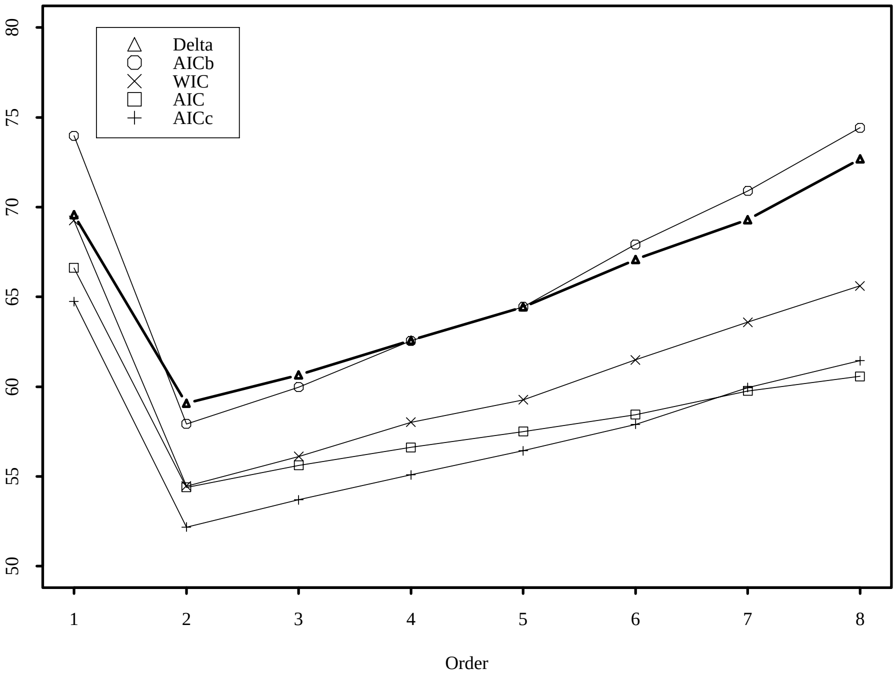

Figure 3. Criterion averages and simulated $\Delta_n(k, \theta_o)$ for (5.3) (orders 1 through 8)

Table 4 indicates that AICb obtains the most correct order selections, followed by SIC. As in the first simulation set, AICb does not exhibit the overfitting tendency of many of the other criteria.

Table 4. Model selections for (5.4) (maximum order: 12)

Order	AICb	WIC	AIC	AICc	FPE	HQ	BIC	SIC
1	4	1	0	0	0	0	1	1
2	84	62	48	73	25	21	74	80
3	8	20	2	6	1	0	2	6
4	3	6	4	5	8	6	2	1
5	1	5	1	3	1	1	1	2
6	0	1	0	0	1	1	0	0
7	0	1	1	1	0	0	1	0
8	0	0	3	2	5	5	1	1
9 to 12	0	4	41	10	60	66	18	9

Figure 4 again illustrates that over the first eight model orders, the average AICb curve closely tracks the simulated $\Delta_n(k,\theta_o)$ curve. The WIC curve also reflects the shape of $\Delta_n(k,\theta_o)$. The AICc curve follows $\Delta_n(k,\theta_o)$ to a lesser degree, whereas the AIC curve appears flat past the true model order.

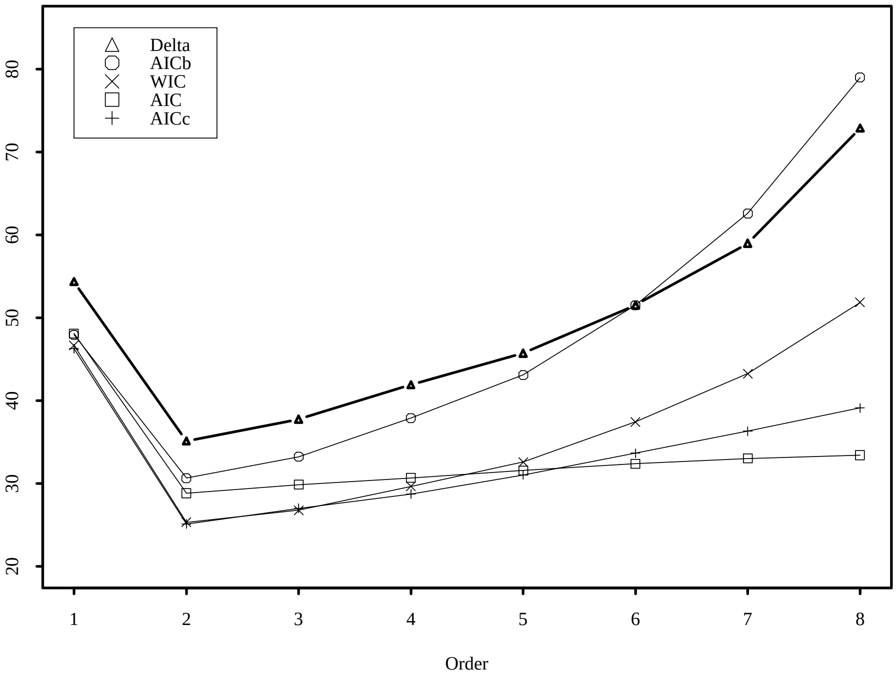

Figure 4. Criterion averages and simulated $\Delta_n(k,\theta_o)$ for (5.4) (orders 1 through 8)

We close this section with a brief discussion of two computational issues. These issues are relevant not only in evaluating the results of the preceding simulations, but also in assessing how AICb and its competitors may perform in practice.

First, one may question how the behavior of the criteria is affected by the choice of the maximum order P for the class of candidate models. In the simulations sets, the criteria which perform poorly tend to choose an excessive number of high dimensional models. How would these criteria behave if lower maximum orders were employed?

To address this question, the criterion selections for each of the four simulation sets are recompiled using smaller maximum orders than those originally considered. The number of correct order selections corresponding to the original and the new maximum orders are reported in Table 5.

Table 5. Number of correct order selections based on various maximum orders for the candidate classes

Set	Max. Order	AICb	WIC	AIC	AICc	FPE	HQ	BIC	SIC
(5.1)	4	90	78	81	86	78	78	90	89
	8	90	74	69	83	51	54	83	88
	12	90	72	45	69	25	23	60	77
(5.2)	4	87	74	68	80	51	50	41	79
	6	82	64	59	76	39	34	27	70
	8	79	57	45	74	27	18	13	64
(5.3)	4	82	81	79	82	77	83	88	86
	8	78	74	73	78	69	80	88	86
	14	78	73	73	78	63	79	88	86
(5.4)	4	85	69	79	84	72	77	87	89
	8	84	66	69	81	53	59	86	88
	12	84	62	48	73	25	21	74	80

Table 5 indicates that as the maximum order is decreased, those criteria which are prone to overfitting become more competitive, and as a result, the disparities in the correct selection rates of the criteria become less extreme. Although AICb still tends to outperform other AIC-type criteria for smaller maximum orders, its advantage in such settings is less pronounced. This is perhaps expected in light of Figures 1 through 4, which show that the average AICb, WIC, AIC, and AICc curves are only substantially discrepant for larger model orders.

Second, one may question how the behavior of AICb and WIC is affected by the choice of N, the number of bootstrap replications used in the evaluation of these criteria. As N increases, the averages which comprise the "penalty" terms of AICb and WIC stabilize. Choosing a value of N which is too small may result in inaccurate estimation of the bias expression (2.2), yet choosing a value of N which is too large will waste computational time. How is the behavior of AICb and WIC affected by the selection of N?

To gain insight into this question, each of the four simulation sets are re-run using N of 50, 100, 150, 200, and 250 (and a maximum candidate model order of $P = 8$). The number of correct order selections for AICb and WIC are reported in Table 6.

Table 6 indicates that while a value of 50 for N appears insufficient, values of 100 and higher seem to yield acceptable results. Although the number of correct selections for both AICb and WIC tends to increase with increasing N, the differences among the results for N of 100 to 250 are of debatable importance. Thus, our choice of $N = 250$ in the original simulation sets may seem somewhat conservative. (This value of N was chosen since in these sets and in others not reported, smaller values seemed to marginally diminish selection performance while larger values did not seem to appreciably improve the results.)

Table 6. Number of correct order selections for AICb and WIC based on various numbers of bootstrap replications (maximum orders: 8)

	AICb					WIC				
	Bootstrap Replications					Bootstrap Replications				
Set	50	100	150	200	250	50	100	150	200	250
(5.1)	79	84	87	87	90	67	73	74	76	74
(5.2)	64	75	75	74	79	46	51	52	50	57
(5.3)	66	73	79	77	78	56	64	66	68	74
(5.4)	73	76	84	84	84	59	69	69	70	66

Of course, an appropriate choice for N depends on several factors: most importantly, the sample size n, the dimension of the candidate model which minimizes the expected discrepancy $\Delta_n(k, \theta_o)$, and the dimension of the largest model in the candidate class. In practice, we recommend monitoring the values of AICb (or WIC) for increasing values of N, until the criterion values are stable enough to clearly discern the minimum. If for even large N, the minimum tends to oscillate among the criterion values corresponding to two or more fitted candidate models, this may serve as an indication that the expected discrepancies for these models are not significantly different. In such an instance, it would be reasonable to select the model having the smallest dimension among these final few candidates.

6. Conclusion

For large-sample applications, AICb is designed to serve the same purpose as traditional AIC, in that it provides an asymptotically unbiased estimate of the expected discrepancy $\Delta_n(k, \theta_o)$ between the generating model and a fitted approximating model. However, for small-sample applications, our simulation results illustrate that in the state-space setting of interest, AICb seems to outperform traditional AIC in three important ways:

- AICb provides an estimate of the expected discrepancy $\Delta_n(k, \theta_o)$ which is considerably less biased than AIC.
- When data is generated from a known finite dimensional model, AICb has a higher success rate in identifying the correct model dimension than AIC.
- AICb does not exhibit the same tendency to overfit that AIC exhibits.

AICb also appears to outperform WIC and AICc in the same three ways, although to a somewhat lesser degree.

AICb is developed in the context of a general model formulation and a nonrestrictive set of conditions. Our justification and application of the criterion focuses on the state-space framework, yet the criterion could certainly be used in other model selection settings. (See Shibata (1997).) And although AICb is more computationally expensive to evaluate than either AIC or AICc, it has a simplistic form, and would be convenient to compute as part of an overall bootstrap-based analysis.

Acknowledgement

The authors would like to express their appreciation to an associate editor and two anonymous referees for providing thoughtful and insightful comments which helped to improve the original version of this manuscript.

Appendix

Here, we present a formal justification of AICb as a large-sample estimator of $\Delta_n(k, \theta_o)$.

We require the following fundamental assumptions:

- The parameter space Θ is a compact subset of k-dimensional Euclidean space.
- Derivatives of the log-likelihood up to order three exist with respect to θ, and are continuous and suitably bounded over Θ.
- θ_o is an interior point of Θ.
- For all $\theta \in \Theta$, the eigenvalues of $\Phi(\theta)$ are within the unit circle, and $AQ(\theta)A' + R(\theta)$ is positive definite.

Our arguments rely on an asymptotic theory developed by Ljung and Caines (1979) for a general class of estimators. This theory can be used to justify the strong consistency and asymptotic normality of the state-space GML estimator $\hat{\theta}_n$, even in the absence of the normality assumption (3.4). (See Caines (1988), page 499.) The results of Ljung and Caines (1979) were utilized by Stoffer and Wall (1991) to provide an asymptotic justification of a nonparametric state-space bootstrap procedure based on GML estimation. Neither the development in Stoffer and Wall (1991) nor the development which follows requires the normality assumption (3.4).

We begin by briefly outlining the theory of Ljung and Caines (1979). (For further details, see Ljung and Caines (1979), Theorem 1 and its corollary.)

Let

$$V_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} \{\log |\Sigma_t(\theta)| + e_t(\theta)' \Sigma_t^{-1}(\theta) e_t(\theta)\}.$$

Let $V_n^{(1)}(\theta)$ denote the kx1 vector of first partials of $V_n(\theta)$ with respect to θ, and let $V_n^{(2)}(\theta)$ denote the kxk matrix of second partials of $V_n(\theta)$ with respect to θ.

Let

$$W_n(\theta) = E_o\{V_n(\theta)\}, \quad W_n^{(1)}(\theta) = E_o\{V_n^{(1)}(\theta)\}, \quad W_n^{(2)}(\theta) = E_o\{V_n^{(2)}(\theta)\}, \text{ and}$$

$$U_n(\theta) = n\ E_o\{(V_n^{(1)}(\theta))(V_n^{(1)}(\theta))^{'}\}.$$

Let $\bar{\theta}_n$ represent the unique global minimum of $W_n(\theta)$ (assumed to exist). We assume $W_n(\theta) \to W(\theta)$ uniformly in θ as $n \to \infty$, $W(\theta)$ has a unique global minimum at θ_o, $\sqrt{n}\ W_n^{(1)}(\theta_o) \to 0$ as $n \to \infty$, and $W^{(2)}(\theta_o)$ is invertible. We assume $U(\theta_o) = \lim_{n\to\infty} U_n(\bar{\theta}_n)$ exists, and $U(\theta_o)$ is invertible.

Let $P_n(\theta) = (W_n^{(2)}(\theta))^{-1}\ U_n(\theta)(W_n^{(2)}(\theta))^{-1}$. Theorem 1 of Ljung and Caines (1979) provides us with the following results:

$$(\hat{\theta}_n - \bar{\theta}_n) \to 0 \quad \text{almost surely as } n \to \infty, \tag{A.1}$$

$$\sqrt{n}\ P_n(\bar{\theta}_n)^{-1/2}\ (\hat{\theta}_n - \bar{\theta}_n) \to N_k(0, I) \ \text{ as } \ n \to \infty. \tag{A.2}$$

We use much of the preceding development in what follows.

We begin our justification by obtaining a useful expansion for $\Delta_n(k, \theta_o)$.

Lemma 1.

$$\begin{aligned}\Delta_n(k,\theta_o) &= E_o\{E_o\{-2\log L(\theta\ |Y_n)\}|_{\theta=\hat{\theta}_n}\}\\ &= E_o\{-2\log L(\hat{\theta}_n\ |Y_n)\} + \frac{n}{2}\ E_o\{(\hat{\theta}_n - \bar{\theta}_n)^{'}\ W_n^{(2)}(\eta_n)\ (\hat{\theta}_n - \bar{\theta}_n)\}\\ &\quad + \frac{n}{2}\ E_o\{(\hat{\theta}_n - \bar{\theta}_n)^{'}\ V_n^{(2)}(\beta_n)\ (\hat{\theta}_n - \bar{\theta}_n)\}. \end{aligned}\tag{A.3}$$

Here, η_n and β_n are random vectors which lie between $\hat{\theta}_n$ and $\bar{\theta}_n$.

Proof. First, we expand $E_o\{-2\log L(\theta\ |Y_n)\}|_{\theta=\hat{\theta}_n}$ about $\bar{\theta}_n$ to obtain

$$E_o\{-2\log L(\theta\ |Y_n)\}|_{\theta=\hat{\theta}_n} = E_o\{-2\log L(\bar{\theta}_n\ |Y_n)\} + \frac{n}{2}(\hat{\theta}_n - \bar{\theta}_n)' W_n^{(2)}(\eta_n)(\hat{\theta}_n - \bar{\theta}_n). \tag{A.4}$$

Here, η_n is a random vector which lies between $\hat{\theta}_n$ and $\bar{\theta}_n$.

Next, we expand $-2\log L(\bar{\theta}_n\ |Y_n)$ about $\hat{\theta}_n$, and take expectations of both sides of the resulting expression to obtain

$$E_o\{-2\log L(\bar{\theta}_n\ |Y_n)\} = E_o\{-2\log L(\hat{\theta}_n\ |Y_n)\} + \frac{n}{2}\ E_o\{(\hat{\theta}_n - \bar{\theta}_n)^{'} V_n^{(2)}(\beta_n)(\hat{\theta}_n - \bar{\theta}_n)\}. \tag{A.5}$$

Here, β_n is a random vector which lies between $\hat{\theta}_n$ and $\bar{\theta}_n$.

The lemma is established by taking expectations with respect to both sides of (A.4), and substituting (A.5) for $E_o\{-2\log L(\bar{\theta}_n\ |Y_n)\}$ in the result.

We next derive a result which leads to a bootstrap estimator of both

$$\frac{n}{2}\ E_o\{(\hat{\theta}_n - \bar{\theta}_n)'\ W_n^{(2)}(\eta_n)\ (\hat{\theta}_n - \bar{\theta}_n)\} \tag{A.6}$$

and

$$\frac{n}{2}\ E_o\{(\hat{\theta}_n - \bar{\theta}_n)'\ V_n^{(2)}(\beta_n)\ (\hat{\theta}_n - \bar{\theta}_n)\}. \tag{A.7}$$

Note by comparing (A.3) to (2.3) in Section 2 that the sum of (A.6) and (A.7) is equivalent to the bias expression (2.2).

Lemma 2.

$$\frac{n}{2}\ E_*\{(\hat{\theta}_n^* - \hat{\theta}_n)'\ V_n^{(2)}(\gamma_n)\ (\hat{\theta}_n^* - \hat{\theta}_n)\} = E_*\{-2\log L(\hat{\theta}_n^*\ |Y_n)\} - \{-2\log L(\hat{\theta}_n\ |Y_n)\}.$$

Here, γ_n is a random vector which lies between $\hat{\theta}_n^*$ and $\hat{\theta}_n$.

Proof. Consider expanding $-2\log L(\hat{\theta}_n^*\ |Y_n)$ about $\hat{\theta}_n$ to obtain

$$-2\log L(\hat{\theta}_n^*\ |Y_n) = -2\log L(\hat{\theta}_n\ |Y_n) + \frac{n}{2}\ (\hat{\theta}_n^* - \hat{\theta}_n)'\ V_n^{(2)}(\gamma_n)\ (\hat{\theta}_n^* - \hat{\theta}_n).$$

Here, γ_n is a random vector which lies between $\hat{\theta}_n^*$ and $\hat{\theta}_n$.

Taking expectations of both sides of this expression with respect to the bootstrap distribution of $\hat{\theta}_n^*$, we have

$$E_*\{-2\log L(\hat{\theta}_n^*\ |Y_n)\} = \{-2\log L(\hat{\theta}_n\ |Y_n)\} + \frac{n}{2}\ E_*\{(\hat{\theta}_n^* - \hat{\theta}_n)'\ V_n^{(2)}(\gamma_n)\ (\hat{\theta}_n^* - \hat{\theta}_n)\}.$$

Thus, the result is established.

At the end of the Appendix, we state and prove a final lemma (Lemma 3) that will show as $n \to \infty$, the difference between $(n/2)E_*\{(\hat{\theta}_n^* - \hat{\theta}_n)' V_n^{(2)}(\gamma_n)\ (\hat{\theta}_n^* - \hat{\theta}_n)\}$ and either (A.6) or (A.7) converges almost surely to zero. By the strong law of large numbers, as $N \to \infty$, $N^{-1}\sum_{i=1}^{N} -2\log L(\hat{\theta}_n^*(i)\ |Y_n)$ converges almost surely to $E_*\{-2\log L(\hat{\theta}_n^*\ |Y_n)\}$. Thus, with Lemma 2, we will be able to conclude that for $n, N \to \infty$,

$$\Big\{\frac{1}{N}\sum_{i=1}^{N} -2\log L(\hat{\theta}_n^*(i)\ |Y_n)\Big\} - \{-2\log L(\hat{\theta}_n\ |Y_n)\} \tag{A.8}$$

is almost surely the same as either (A.6) or (A.7). This will justify AICb as a large-sample estimator of $\Delta_n(k, \theta_o)$, in that it will show the "penalty" term of AICb (twice (A.8)) is asymptotically equal to the sum of (A.6) and (A.7), or equivalently, to the bias expression (2.2).

We first introduce some additional notation and results.

In the bootstrap setting, to parallel the definitions for

$$V_n(\theta),\ V_n^{(1)}(\theta),\ V_n^{(2)}(\theta),\ W_n(\theta),\ W_n^{(1)}(\theta),\ W_n^{(2)}(\theta),\ U_n(\theta),\ \hat{\theta}_n,\ \text{and}\ \bar{\theta}_n,$$

let

$$V_n^*(\theta) = \frac{1}{n}\sum_{t=1}^{n}\{\log|\Sigma_t(\theta)| + e_t^*(\theta)'\Sigma_t^{-1}(\theta)e_t^*(\theta)\}$$

where the innovations $e_t^*(\theta)$ correspond to a bootstrap sample, let $V_n^{*(1)}(\theta)$ denote the kx1 vector of first partials of $V_n^*(\theta)$ with respect to θ, and let $V_n^{*(2)}(\theta)$ denote the kxk matrix of second partials of $V_n^*(\theta)$ with respect to θ. Let

$$W_n^*(\theta)=E_*\{V_n^*(\theta)\},\ W_n^{*(1)}(\theta)=E_*\{V_n^{*(1)}(\theta)\},\ W_n^{*(2)}(\theta)=E_*\{V_n^{*(2)}(\theta)\},\ \text{and}$$
$$U_n^*(\theta)=n\ E_*\{(V_n^{*(1)}(\theta))(V_n^{*(1)}(\theta))'\}.$$

Also, let $\hat{\theta}_n^* = \text{argmin}\ _{\theta\in\Theta}\ V_n^*(\theta)$, and $\bar{\theta}_n^* = \text{argmin}\ _{\theta\in\Theta}\ W_n^*(\theta)$.

We make use of the following important result from Lemma 3 of Ljung and Caines (1979):

$$V_n^{(2)}(\theta) - W_n^{(2)}(\theta) \to 0 \quad \text{almost surely as } n\to\infty,\ \ \text{uniformly in } \theta. \tag{A.9}$$

Now from Lemma 1 of Stoffer and Wall (1991),

$$W_n^*(\theta) = V_n(\theta) \text{ for all } \theta\in\Theta, \tag{A.10}$$

meaning

$$\bar{\theta}_n^* = \hat{\theta}_n. \tag{A.11}$$

Also, from Lemma 2 of Stoffer and Wall (1991),

$$U_n^*(\hat{\theta}_n) - U_n(\bar{\theta}_n) \to 0 \text{ almost surely as } n\to\infty. \tag{A.12}$$

Let $P_n^*(\theta) = (W_n^{*(2)}(\theta))^{-1}U_n^*(\theta)(W_n^{*(2)}(\theta))^{-1}$. Stoffer and Wall (1991) appeal to Theorem 1 of Ljung and Caines (1979) to establish the following analogue of (A.2) for the bootstrap GML estimator $\hat{\theta}_n^*$:

$$\sqrt{n}P_n(\hat{\theta}_n)^{-1/2}(\hat{\theta}_n^* - \hat{\theta}_n) \to N_k(0,I) \text{ as } n\to\infty. \tag{A.13}$$

One can also appeal to Theorem 1 and Lemma 3 of Ljung and Caines (1979) to establish the bootstrap analogues of (A.1) and (A.9):

$$(\hat{\theta}_n^* - \hat{\theta}_n) \to 0 \text{ almost surely as } n\to\infty, \tag{A.14}$$

$$V_n^{*(2)}(\theta) - W_n^{*(2)}(\theta) \to 0 \text{ almost surely as } n\to\infty,\ \text{uniformly in } \theta. \tag{A.15}$$

We now present the statement and proof of our final lemma.

Lemma 3.

(a) $nE_*\{(\hat{\theta}_n^* - \hat{\theta}_n)'V_n^{(2)}(\gamma_n)(\hat{\theta}_n^* - \hat{\theta}_n)\} - nE_o\{(\hat{\theta}_n - \bar{\theta}_n)'V_n^{(2)}(\beta_n)(\hat{\theta}_n - \bar{\theta}_n)\} \to 0$ *almost surely as* $n\to\infty$.

(b) $nE_*\{(\hat{\theta}_n^* - \hat{\theta}_n)' V_n^{(2)}(\gamma_n)(\hat{\theta}_n^* - \hat{\theta}_n)\} - nE_o\{(\hat{\theta}_n - \bar{\theta}_n)' W_n^{(2)}(\eta_n)(\hat{\theta}_n - \bar{\theta}_n)\} \to 0$ *almost surely as* $n \to \infty$.

Here, γ_n *is a random vector which lies between* $\hat{\theta}_n^*$ *and* $\hat{\theta}_n$*, and* β_n *and* η_n *are random vectors which lie between* $\hat{\theta}_n$ *and* $\bar{\theta}_n$.

Proof. First, we consider the expression $nE_*\{(\hat{\theta}_n^* - \hat{\theta}_n)' V_n^{(2)}(\gamma_n)(\hat{\theta}_n^* - \hat{\theta}_n)\}$. We show that as $n \to \infty$, this quantity differs from

$$\mathrm{tr}\{(W_n^{(2)}(\bar{\theta}_n))^{-1} U_n(\bar{\theta}_n)\} \tag{A.16}$$

by an amount tending to zero almost surely.

Using (3.5) of Ljung and Caines (1979) along with (A.11), we can write

$$\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)' = \sqrt{n} V_n^{*(1)}(\hat{\theta}_n)' (V_n^{*(2)}(\alpha_n))^{-1}, \tag{A.17}$$

where α_n is a random vector between $\hat{\theta}_n^*$ and $\hat{\theta}_n$. Now by (A.15), (A.10), (A.9), and the consistency results (A.14) and (A.1),

$$V_n^{*(2)}(\alpha_n) = W_n^{(2)}(\bar{\theta}_n) + o_{a.s.}(1). \tag{A.18}$$

Also, by (A.9), and the consistency results (A.14) and (A.1),

$$V_n^{(2)}(\gamma_n) = W_n^{(2)}(\bar{\theta}_n) + o_{a.s.}(1). \tag{A.19}$$

Using representation (A.17) along with (A.18) and (A.19), we can establish that

$$n(\hat{\theta}_n^* - \hat{\theta}_n)' V_n^{(2)}(\gamma_n)(\hat{\theta}_n^* - \hat{\theta}_n) = n(V_n^{*(1)}(\hat{\theta}_n))' (W_n^{(2)}(\bar{\theta}_n))^{-1} (V_n^{*(1)}(\hat{\theta}_n)) + o_{a.s.}(1).$$

Applying the bootstrap expectation operator to both sides of the preceding expression and utilizing (A.12), we obtain

$$\begin{aligned} &nE_*\{(\hat{\theta}_n^* - \hat{\theta}_n)' V_n^{(2)}(\gamma_n)(\hat{\theta}_n^* - \hat{\theta}_n)\} \\ &= nE_*\{(V_n^{*(1)}(\hat{\theta}_n))' (W_n^{(2)}(\bar{\theta}_n))^{-1} (V_n^{*(1)}(\hat{\theta}_n))\} + o_{a.s.}(1) \\ &= \mathrm{tr}\{(W_n^{(2)}(\bar{\theta}_n))^{-1} [nE_*\{(V_n^{*(1)}(\hat{\theta}_n))(V_n^{*(1)}(\hat{\theta}_n))'\}]\} + o_{a.s.}(1) \\ &= \mathrm{tr}\{(W_n^{(2)}(\bar{\theta}_n))^{-1} U_n^*(\hat{\theta}_n)\} + o_{a.s.}(1) \\ &= \mathrm{tr}\{(W_n^{(2)}(\bar{\theta}_n))^{-1} U_n(\bar{\theta}_n)\} + o_{a.s.}(1). \end{aligned} \tag{A.20}$$

Next, consider the quadratic expressions

$$n(\hat{\theta}_n - \bar{\theta}_n)' V_n^{(2)}(\beta_n)(\hat{\theta}_n - \bar{\theta}_n) \text{ and } n(\hat{\theta}_n - \bar{\theta}_n)' W_n^{(2)}(\eta_n)(\hat{\theta}_n - \bar{\theta}_n).$$

We show that as $n \to \infty$, the difference between these quadratics tends almost surely to zero. We then show that as $n \to \infty$, the difference between the expectation of either quadratic and (A.16) tends to zero. This combined with (A.20) will establish the lemma.

Using (3.5) of Ljung and Caines (1979), we can write

$$\sqrt{n}(\hat{\theta}_n - \bar{\theta}_n)' = \sqrt{n} V_n^{(1)}(\bar{\theta}_n)' (V_n^{(2)}(\delta_n))^{-1}, \tag{A.21}$$

where δ_n is a random vector between $\hat{\theta}_n$ and $\bar{\theta}_n$. Now by (A.9) and the consistency result (A.1), we have

$$V_n^{(2)}(\delta_n) = W_n^{(2)}(\bar{\theta}_n) + o_{a.s.}(1). \tag{A.22}$$

Also by (A.9) and (A.1),

$$V_n^{(2)}(\beta_n) = W_n^{(2)}(\bar{\theta}_n) + o_{a.s.}(1) \text{ and } W_n^{(2)}(\eta_n) = W_n^{(2)}(\bar{\theta}_n) + o_{a.s.}(1). \tag{A.23}$$

Using representation (A.21) along with (A.22) and (A.23), we can argue that as $n \to \infty$,

$$n(\hat{\theta}_n - \bar{\theta}_n)' V_n^{(2)}(\beta_n)(\hat{\theta}_n - \bar{\theta}_n) \text{ and } n(\hat{\theta}_n - \bar{\theta}_n)' W_n^{(2)}(\eta_n)(\hat{\theta}_n - \bar{\theta}_n)$$

each differ from

$$n(V_n^{(1)}(\bar{\theta}_n))' (W_n^{(2)}(\bar{\theta}_n))^{-1} (V_n^{(1)}(\bar{\theta}_n))$$

by an amount which tends almost surely to zero.

Thus, as $n \to \infty$,

$$nE_o\{(\hat{\theta}_n - \bar{\theta}_n)' V_n^{(2)}(\beta_n)(\hat{\theta}_n - \bar{\theta}_n)\} \text{ and } nE_o\{(\hat{\theta}_n - \bar{\theta}_n)' W_n^{(2)}(\eta_n)(\hat{\theta}_n - \bar{\theta}_n)\}$$

each differ from

$$\begin{aligned} & nE_o\{(V_n^{(1)}(\bar{\theta}_n))' (W_n^{(2)}(\bar{\theta}_n))^{-1} (V_n^{(1)}(\bar{\theta}_n))\} \\ &= \mathrm{tr}\{(W_n^{(2)}(\bar{\theta}_n))^{-1} [nE_o\{(V_n^{(1)}(\bar{\theta}_n))(V_n^{(1)}(\bar{\theta}_n))'\}]\} \\ &= \mathrm{tr}\{(W_n^{(2)}(\bar{\theta}_n))^{-1} U_n(\bar{\theta}_n)\} \end{aligned}$$

by an amount which converges to zero.

This completes the proof of the lemma.

References

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

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19**, 716-723.

Akaike, H. (1978). Time series analysis and control through parametric models. In *Applied Time Series Analysis* (Edited by D. F. Findley), 1-23. Academic Press, New York.

Caines, Peter E. (1988). *Linear Stochastic Systems.* Wiley, New York.

Cavanaugh, J. E. (1993). Small-sample model selection in the general state-space setting. Ph.D. dissertation, University of California, Davis, Division of Statistics. (Printed by UMI Dissertation Services, Ann Arbor, MI.)

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78**, 316-331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

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.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, New York.

Hurvich, C. M., Shumway, R. H. and Tsai, C. L. (1990). Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. *Biometrika* **77**, 709-719.

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

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

Hurvich, C. M. and Tsai, C. L. (1993). A corrected Akaike information criterion for vector autoregressive model selection. *J. Time Ser. Anal.* **14**, 271-279.

Ishiguro, M. and Sakamoto, Y. (1991). WIC: An estimation-free information criterion. Research memorandum, Institute of Statistical Mathematics, Tokyo.

Ishiguro, M., Morita, K. I. and Ishiguro, M. (1991). Application of an estimator-free information criterion (WIC) to aperture synthesis imaging. In *Radio Interferometry: Theory, Techniques, and Applications* (Edited by T. J. Cornwell and R. A. Perley), International Astronomical Union Coll. 131, Conference Series, 19, 243-248. Astronomical Society of the Pacific, San Francisco.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Engineering, Transactions ASME* **82**, 35-45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction theory. *J. Basic Engineering, Transactions ASME* **83**, 95-108.

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

Ljung, L. and Caines, P. E. (1979). Asymptotic normality of prediction error estimators for approximate system models. *Stochastics* **3**, 29-46.

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

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

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Trans. Inform. Theory* **11**, 61-70.

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statist. Sinica* **7**, 375-394.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.* **3**, 253-264.

Shumway, R. H. (1988). *Applied Statistical Time Series Analysis.* Prentice-Hall, New Jersey.

Stoffer, D. S. and Wall, K. D. (1991). Bootstrapping state-space models: Gaussian maximum likelihood estimation and the Kalman filter. *J. Amer. Statis. Assoc.* **86**, 1024-1033.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Statist.* **A7**, 13-26.

Department of Statistics, University of Missouri, Columbia, MO 65211, U.S.A.

Division of Statistics, University of California, Davis, CA 95616, U.S.A.

(Received November 1994; accepted September 1996)