An Improved Akaike Information Criterion for State - Space Model Selection

Thomas Bengtsson
Department of Statistics, University of California – Berkeley
Joseph E. Cavanaugh
Department of Biostatistics, The University of Iowa

Abstract

Following the work of Hurvich, Shumway, and Tsai (1990), we propose an "improved" variant of the Akaike information criterion, AICi, for state-space model selection. The variant is based on Akaike's (1973) objective of estimating the Kullback-Leibler information (Kullback 1968) between the densities corresponding to the fitted model and the generating or true model.

The development of AICi proceeds by decomposing the expected information into two terms. The first term suggests that the empirical log likelihood can be used to form a biased estimator of the information; the second term provides the bias adjustment. Exact computation of the bias adjustment requires the values of the true model parameters, which are inaccessible in practical applications. Yet for fitted models in the candidate class that are correctly specified or overfit, the adjustment is asymptotically independent of the true parameters. Thus, in certain settings, the adjustment may be estimated via Monte Carlo simulations by using conveniently chosen simulation parameters as proxies for the true parameters.

We present simulation results to evaluate the performance of AICi both as an estimator of the Kullback-Leibler information and as a model selection criterion. Our results indicate that AICi estimates the information with less bias than traditional AIC. Furthermore, AICi serves as an effective tool for selecting a model of appropriate dimension.

Keywords: AIC, Kullback-Leibler information, Kullback's directed divergence, state-space model, time series analysis.

1 Introduction

In many time series applications, an investigator must choose an appropriate model to characterize the sample data. This determination should ideally be guided by scientific theory, but the researcher may also be well served by a data-driven selection method. To this end, Akaike (1973, 1974) introduced the Akaike information criterion, AIC, which discerns how "close" a fitted model is to the generating or true model. Akaike's work stimulated many other approaches to model selection, leading to the development of criteria such as SIC (Schwarz 1978), BIC (Akaike 1978), and HQ (Hannan and Quinn 1979).

Extending Akaike's original work, Sugiura (1978) proposed AICc, a corrected version of AIC justified in the context of linear regression with normal errors. The development of AICc was motivated by the need to adjust for AIC's propensity to favor high-dimensional models when the sample size is small relative to the maximum order of the models in the candidate class. Hurvich and Tsai (1989) show that AICc dramatically outperforms AIC in small-sample regression settings, and further extend AICc to include univariate Gaussian autoregressive models. Hurvich, Shumway, and Tsai (1990) generalize AICc to encompass univariate Gaussian autoregressive moving-average models, and Hurvich and Tsai (1993) handle the vector Gaussian autoregressive case. The demonstrated effectiveness of AICc in these settings motivates the need for a corrected or improved variant of AIC for state-space models.

Apart from Cavanaugh and Shumway (1997), who justify and investigate a bootstrap-based version of AIC, the development of model selection criteria expressly designed for state-space applications has received little attention in the literature. With respect to proposing a corrected variant of AIC in the state-space setting, one particular challenge is posed by the very generality of the model, which includes as special cases autoregressive, ARMA, and structural models. That is, the strategy commonly employed in the derivation of corrected versions of AIC – to consider a more limited modeling framework – clearly opposes the inclusive nature of the state-space class. Hence, it may prove difficult to conceive of a general state-space model formulation that is also sufficiently restrictive (in terms of the structure of the model and the distribution of the underlying data) to facilitate the justification of a corrected AIC. Nevertheless, the development of a corrected AIC for general state-space model selection would permit the comparison of different types of models, such as autoregressive integrated moving average (ARIMA) models and additive models involving stochastic trend and seasonality (cf. Harvey and Todd 1983, Kitagawa and Gersch 1984). Such a criterion would therefore be very appealing.

Motivated by the preceding notion, we propose a variant of AIC that achieves the same degree of effectiveness as AICc, but which can be used within the broad framework of the linear state-space model. This variant, called the "improved" Akaike information criterion or AICi, is based on an idea advanced by Hurvich, Shumway, and Tsai (1990) in the context of univariate Gaussian autoregressive models. The criterion involves the same goodness-of-fit term as AIC, yet features a penalty term that arises via a simulated bias correction. This developed bias adjustment can be justified and applied in a very general context, and the

resulting AIC variant fulfills our objective of obtaining an effective model selection tool for the state-space setting.

In what follows, we outline the development of AICi for state-space applications, and investigate the performance of the criterion in a simulation study.

2 The State-Space Model: Background and Notation

A linear state-space process $\mathbf{y}_t$ can be represented by two sets of equations:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}_t\mathbf{x}_t + \mathbf{v}_t, \text{ and} \\ \mathbf{x}_t &= \mathbf{\Phi}\mathbf{x}_{t-1} + \mathbf{w}_t, \text{ for } t = 1, 2, \ldots, T. \end{aligned} \tag{1}$$

In (1), the design matrix $\mathbf{A}_t$ relates the unobserved $q \times 1$ state vector $\mathbf{x}_t$ to the $p \times 1$ observed vector $\mathbf{y}_t$, while the transition structure $\mathbf{\Phi}$ relates $\mathbf{x}_t$ to its previous value $\mathbf{x}_{t-1}$ via an autoregression. The vectors $\mathbf{v}_t$ and $\mathbf{w}_t$ represent mutually and serially uncorrelated zero-mean error processes with covariance structures $\mathbf{R}$ and $\mathbf{Q}$, respectively. The model assumes a prior distribution for $\mathbf{x}_0$ with $E(\mathbf{x}_0) = \boldsymbol{\mu}$ and $cov(\mathbf{x}_0) = \mathbf{\Sigma}$. The state $\mathbf{x}_0$ is taken to be uncorrelated with $\mathbf{v}_t$ and $\mathbf{w}_t$ for all t. We assume normality of both error processes as well as for $\mathbf{x}_0$.

Let $\mathbf{\Theta} = \{\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{\Phi}, \mathbf{Q}, \mathbf{R}\}$ denote the set of parameters for (1), and let $\mathbf{Y}^t = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_t')'$ and $\mathbf{X}^t = (\mathbf{x}_0', \mathbf{x}_1', \ldots, \mathbf{x}_t')'$ represent vectors of observed data and unobserved states. The preceding notation and setup for the state-space model is identical to that of Shumway and Stoffer (2000, p306), and makes implementation of the EM algorithm for estimation of $\mathbf{\Theta}$ straightforward.

Of primary concern in state-space modeling is the recovery of the unobserved states in $\mathbf{X}^T$. Prediction of $\mathbf{x}_t$ using $\mathbf{Y}^{t-1}$ is referred to as one-step prediction, and prediction of $\mathbf{x}_t$ using $\mathbf{Y}^t$ is referred to as filtering. One-step prediction of $\mathbf{x}_t$ is accomplished by calculating the conditional mean $E(\mathbf{x}_t|\mathbf{Y}^{t-1})$, here denoted $\tilde{\mathbf{x}}_t(\mathbf{\Theta}, \mathbf{Y}^{t-1})$. Similarly, filter prediction of $\mathbf{x}_t$ is accomplished by calculating $E(\mathbf{x}_t|\mathbf{Y}^t)$, denoted $\tilde{\mathbf{x}}_t(\mathbf{\Theta}, \mathbf{Y}^t)$. Under the previously mentioned assumptions on $\mathbf{x}_0$, $\mathbf{v}_t$, and $\mathbf{w}_t$, the conditional means are the best predictors of $\mathbf{x}_t$; in the case of non-normal error processes, the conditional means are the best linear predictors of $\mathbf{x}_t$. For a given data and parameter structure, the one-step predictors $\tilde{\mathbf{x}}_t(\mathbf{\Theta}, \mathbf{Y}^{t-1})$ and the filters $\tilde{\mathbf{x}}_t(\mathbf{\Theta}, \mathbf{Y}^t)$ are generally obtained through the forward Kalman filter recursions (Kalman 1960). These recursions also generate the innovations $e_t(\mathbf{\Theta}, \mathbf{Y}^t) = \mathbf{y}_t - \mathbf{A}_t E(\mathbf{x}_t|\mathbf{Y}^{t-1})$ along with their covariance matrices $\mathbf{\Sigma}_t(\mathbf{\Theta}) = E\{e_t(\mathbf{\Theta}, \mathbf{Y}^t)e_t(\mathbf{\Theta}, \mathbf{Y}^t)'\}$.

The Kalman recursions assume that $\mathbf{\Theta}$ is known; realistically, however, estimates of the unknown parameters in $\mathbf{\Theta}$ will be needed. Estimates are usually obtained using maximum likelihood (ML), where the likelihood is formulated based on the innovations (see Schweppe 1965). In the simulations presented in section 5, we use the EM algorithm method of ML estimation, adapted to the state-space framework by Shumway and Stoffer (1982).

3 Kullback's Directed Divergence, AIC, AICc, and AICi

A well-known measure of separation between two densities is the Kullback-Leibler information, also known as the directed divergence (Kullback 1968). Here we use the directed divergence as a tool for choosing a fitted model that matches as closely as possible the one that presumably generated the data. Specifically, our approach follows Akaike's (1973) strategy of minimizing the separation between the densities corresponding to a fitted model and the true model. We now formalize the notion of selecting a fitted model from a candidate class. (In what follows, for notational simplicity, we write $\mathbf{Y}$ to mean $\mathbf{Y}^T$.)

Suppose the data $\mathbf{Y}$ is sampled according to an unknown parametric density $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$, where $\boldsymbol{\Theta}_0$ represents the generating (or true) parameter structure. Let $\Omega(k)$ denote a k-dimensional parameter space, and let $\mathcal{F}(k) = \{f(\mathbf{Y}|\boldsymbol{\Theta}_k)|\boldsymbol{\Theta}_k \in \Omega(k)\}$ denote a corresponding parametric family of densities. Further, let $\widehat{\boldsymbol{\Theta}}_k$ denote the parameter estimate obtained by maximizing the likelihood function $f(\mathbf{Y}|\boldsymbol{\Theta}_k)$ over $\Omega(k)$, and let $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$ represent the resulting empirical likelihood.

Our goal is to search among a class of families $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), \ldots, \mathcal{F}(k_L)\}$ for the fitted model $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k), k \in \{k_1, k_2, \ldots, k_L\}$, that best approximates $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$. We note that in many applications, some of the families in the class $\mathcal{F}$ may have the same dimension and yet be different. For ease of notation, we do not include an index to delineate between such families.

We refer to $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$ as the true or generating model, and to $f(\mathbf{Y}|\boldsymbol{\Theta}_k)$ as an approximating or candidate model (provided that $\boldsymbol{\Theta}_0 \neq \boldsymbol{\Theta}_k$). If $f(\mathbf{Y}|\boldsymbol{\Theta}_0) \in \mathcal{F}(k)$, and $\mathcal{F}(k)$ is such that no smaller family will contain $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$, we refer to $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$ as correctly specified. If $f(\mathbf{Y}|\boldsymbol{\Theta}_0) \in \mathcal{F}(k)$, yet $\mathcal{F}(k)$ is such that families smaller than $\mathcal{F}(k)$ also contain $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$, we say that $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$ is overfit. If $f(\mathbf{Y}|\boldsymbol{\Theta}_0) \notin \mathcal{F}(k)$, we say that $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$ is underfit.

To determine which of the fitted models $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_{k_1}), f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_{k_2}), \ldots, f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_{k_L})$ best resembles $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$, we need a measure that gauges the disparity between the true model $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$ and an approximating model $f(\mathbf{Y}|\boldsymbol{\Theta})$. To this end, we consider the Kullback-Leibler information between $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$ and $f(\mathbf{Y}|\boldsymbol{\Theta})$, given by $d(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}) = E[\log\{f(\mathbf{Y}|\boldsymbol{\Theta}_0)/f(\mathbf{Y}|\boldsymbol{\Theta})\}]$. Here and in our subsequent development, $E(\cdot)$ denotes the expectation under $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$.

Letting $\delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}) = E\{-2\log f(\mathbf{Y}|\boldsymbol{\Theta})\}$, we express $2d(\boldsymbol{\Theta}_0, \boldsymbol{\Theta})$ by the difference $\delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}) - \delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_0)$. Since $\delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}_0)$ does not depend on $\boldsymbol{\Theta}$, any ranking of a set of candidate models corresponding to $d(\boldsymbol{\Theta}_0, \boldsymbol{\Theta})$ would be identical to a ranking corresponding to $\delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta})$. Thus, for model selection purposes, we may use $\delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta})$ as a substitute for $d(\boldsymbol{\Theta}_0, \boldsymbol{\Theta})$.

Now, for a given maximum likelihood estimate $\widehat{\boldsymbol{\Theta}}_k$, the exact divergence between the true model and the fitted model is indicated by $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$. However, computing $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$ would require the true density $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$, and is for obvious reasons not possible. Addressing the lack of knowledge of $\boldsymbol{\Theta}_0$, Akaike (1973) noted that $-2\log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$ serves as a biased estimate of $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$, and that the bias adjustment

$$B_T(k, \boldsymbol{\Theta}_0) = E[\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k) - \{-2\log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)\}] \tag{2}$$

can often be approximated by $2k$. Specifically, if the following two assumptions are met, one can establish that $B_T(k, \boldsymbol{\Theta}_0)$ converges to $2k$ as T tends to infinity (e.g., see Cavanaugh 1997, p204).

(a) The fitted model is either correctly specified or overfit; i.e., $f(\mathbf{Y}|\boldsymbol{\Theta}_0) \in \mathcal{F}(k)$.

(b) A set of regularity conditions holds that will ensure the conventional asymptotic properties of the maximum likelihood estimator $\widehat{\boldsymbol{\Theta}}_k$.

Under assumptions (a) and (b), it follows that the expected value of

$$\text{AIC} = -2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k) + 2k$$

should asymptotically approach the expected value of $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$. Hence, for large samples, model selection based on AIC should lead to fitted models $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$ which are, in the sense of the average Kullback-Leibler information, closest to $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$.

The approximation to the bias adjustment $B_T(k, \boldsymbol{\Theta}_0)$ by $2k$ is derived under fairly general assumptions and makes AIC applicable to many different statistical frameworks. However, because the approximation holds only for large samples, the utility of AIC in small-sample settings may be limited. For instance, with regression and autoregressive models, Hurvich and Tsai (1989) show that as k increases relative to the sample size T, AIC becomes increasingly negatively biased. The negative bias of AIC in small-sample applications often results in severe overfitting. One solution to this problem is to impose a strict cut-off for the maximum dimension to be considered in the model search. However, this approach is rather arbitrary and has no theoretical basis. An alternative strategy is to develop model selection criteria with better small-sample bias properties than AIC. As indicated previously, this is often achieved by considering only fitted models in a restricted candidate class.

In the context of normal linear regression where $f(\mathbf{Y}|\boldsymbol{\Theta}_0) \in \mathcal{F}(k)$, Sugiura (1978) shows that the bias adjustment (2) is exactly equal to $B_T(k, \boldsymbol{\Theta}_0) = \{2T(m+1)\}/(T-m-2)$, where m denotes the rank of the design matrix. Thus, for a particular fitted model, an exactly unbiased estimator of $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$ is obtained by evaluating

$$\text{AICc} = -2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k) + \frac{2T(m+1)}{T-m-2}.$$

AICc may be used in univariate Gaussian autoregressive applications to select the order m of the autoregression, but in such contexts the bias correction $\{2T(m+1)\}/(T-m-2)$ is not exact and AICc is only asymptotically unbiased for $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$. However, Hurvich and Tsai (1989) demonstrate that the criterion performs well in small-sample simulations. Further, they show that when the dimension of the largest model in the candidate class is large compared to the sample size, AICc does not suffer from the severe overfitting tendencies that often plague AIC.

To address the bias of AICc in autoregressive modeling applications, Hurvich, Shumway, and Tsai (1990) propose AICi as a refinement of AICc. AICi is based on the premise that the bias adjustment $B_T(k, \boldsymbol{\Theta}_0)$ only

loosely depends upon the true parameter $\boldsymbol{\Theta}_0$. Specifically, when assumptions (a) and (b) hold, $B_T(k, \boldsymbol{\Theta}_0)$ converges to $2k$, implying that the dependence of the bias adjustment on $\boldsymbol{\Theta}_0$ diminishes as the sample size tends to infinity. In such instances, it should be possible to accurately approximate the bias adjustment expressed in (2) via Monte Carlo simulation using an arbitrary but convenient choice of a parameter in place of $\boldsymbol{\Theta}_0$. To clearly distinguish the parameter used for simulations from the true (but unknown) parameter $\boldsymbol{\Theta}_0$, we denote the choice of the simulation parameter by $\boldsymbol{\Theta}_s$. Although the simulated approximation is only guaranteed to be near $B_T(k, \boldsymbol{\Theta}_0)$ for large T, in small to moderate sample-size applications, the approximation yields a more accurate estimate of $B_T(k, \boldsymbol{\Theta}_0)$ than $2k$. This claim is supported both by the simulation results that follow and by those reported in Hurvich, Shumway, and Tsai (1990).

The criterion AICi is then obtained by evaluating

$$\text{AICi} = -2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k) + \frac{1}{M} \sum_{j=1}^{M} [\delta(\boldsymbol{\Theta}_s, \widehat{\boldsymbol{\Theta}}_k(j)) - \{-2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k(j))\}],$$

where $\{\widehat{\boldsymbol{\Theta}}_k(1), \widehat{\boldsymbol{\Theta}}_k(2), \ldots, \widehat{\boldsymbol{\Theta}}_k(M)\}$ represent a set of estimates based on M samples generated under model (1) with $\boldsymbol{\Theta}_s$ as the parameter.

We now outline the development of AICi for the state-space framework.

4 Development of AICi for the State-Space Model

The derivation of AICi proceeds by decomposing the expected directed divergence $E\{\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)\}$ into two terms. The first term suggests that the empirical log likelihood can be used to form a biased estimator of the directed divergence, and the second term provides the bias adjustment. As indicated in the previous section, under assumptions (a) and (b), the bias adjustment is asymptotically independent of the true parameter $\boldsymbol{\Theta}_0$. It can therefore be estimated via Monte Carlo simulation by using an arbitrary simulation parameter $\boldsymbol{\Theta}_s$ as a proxy for $\boldsymbol{\Theta}_0$. Based on the true parameter $\boldsymbol{\Theta}_0$, we now develop explicit forms of the directed divergence and bias adjustment for models defined by (1). A convenient simulation parameter $\boldsymbol{\Theta}_s$ is then chosen to simplify the obtained divergence and bias expressions, and the simplified forms are used in simulations to estimate the bias adjustment $B_T(k, \boldsymbol{\Theta}_0)$.

Using the innovations form of the likelihood, we have

$$-2 \log f(\mathbf{Y}|\boldsymbol{\Theta}) = \sum_{t=1}^{T} \log |\boldsymbol{\Sigma}_t(\boldsymbol{\Theta})| + \sum_{t=1}^{T} e_t(\boldsymbol{\Theta}, \mathbf{Y}^t)' \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\Theta}) e_t(\boldsymbol{\Theta}, \mathbf{Y}^t). \tag{3}$$

(Here and throughout our development, we have ignored the constant involving 2π.) When evaluated at the MLE $\widehat{\boldsymbol{\Theta}}_k$, properties of the multivariate Gaussian distribution can be used to simplify the empirical log-likelihood (e.g., Johnson and Wichern 1998, p180-1). We have

$$-2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k) = \sum_{t=1}^{T} \log |\boldsymbol{\Sigma}_t(\widehat{\boldsymbol{\Theta}}_k)| + Tp, \tag{4}$$

where p denotes the dimension of the innovations.

From (3), it follows that the directed divergence between the true model $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$ and the candidate model $f(\mathbf{Y}|\boldsymbol{\Theta})$ is given by

$$\delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}) = \sum_{t=1}^{T} \log |\boldsymbol{\Sigma}_t(\boldsymbol{\Theta})| + \sum_{t=1}^{T} E\{e_t(\boldsymbol{\Theta}, \mathbf{Y}^t)' \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\Theta}) e_t(\boldsymbol{\Theta}, \mathbf{Y}^t)\}.$$

Using the mean and covariance properties of $e_t(\boldsymbol{\Theta}_0, \mathbf{Y}^t)$ under the model defined by $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$, the preceding can alternatively be expressed as

$$\begin{aligned} \delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta}) &= \sum_{t=1}^{T} \log |\boldsymbol{\Sigma}_t(\boldsymbol{\Theta})| + \sum_{t=1}^{T} tr\{\boldsymbol{\Sigma}_t(\boldsymbol{\Theta}_0) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\Theta})\} \\ &\quad + \sum_{t=1}^{T} E\Big[\{e_t(\boldsymbol{\Theta}, \mathbf{Y}^t) - e_t(\boldsymbol{\Theta}_0, \mathbf{Y}^t)\}' \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\Theta}) \{e_t(\boldsymbol{\Theta}, \mathbf{Y}^t) - e_t(\boldsymbol{\Theta}_0, \mathbf{Y}^t)\}\Big]. \end{aligned} \tag{5}$$

When evaluated at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_k$, $\delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta})$ yields the directed divergence between the true model and the fitted model; $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$ thereby reflects the separation between $f(\mathbf{Y}|\boldsymbol{\Theta}_0)$ and $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$.

In the interpretation of $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$, $\widehat{\boldsymbol{\Theta}}_k$ is viewed as fixed. However, since $\widehat{\boldsymbol{\Theta}}_k$ is based on the data, we consider averaging $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$ over different values of $\widehat{\boldsymbol{\Theta}}_k$ arising from its sampling distribution. The resulting measure, the expected directed divergence $E\{\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)\}$, reflects the *average* separation between the true model and those fitted models having the same structure as $f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$.

The expected directed divergence $E\{\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)\}$ involves a double expectation. First, as indicated by (5), we average over the distribution of $\mathbf{Y}$ to obtain $\delta(\boldsymbol{\Theta}_0, \boldsymbol{\Theta})$. Second, we average $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$ over the sampling distribution of $\widehat{\boldsymbol{\Theta}}_k$ to obtain $E\{\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)\}$. For conceptual clarity, the double expectation may be represented as a single expectation by introducing a data vector $\mathbf{Y}_*$ that is independent of $\mathbf{Y}$ yet shares the same distribution as $\mathbf{Y}$. With $\mathbf{Y}_*^t$ defined accordingly, by (5), $E\{\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)\}$ may be expressed as

$$\begin{aligned} E\{\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)\} &= \sum_{t=1}^{T} E\{\log |\boldsymbol{\Sigma}_t(\widehat{\boldsymbol{\Theta}}_k)|\} + \sum_{t=1}^{T} E[tr\{\boldsymbol{\Sigma}_t(\boldsymbol{\Theta}_0) \boldsymbol{\Sigma}_t^{-1}(\widehat{\boldsymbol{\Theta}}_k)\}] \\ &\quad + \sum_{t=1}^{T} E\Big[\{e_t(\widehat{\boldsymbol{\Theta}}_k, \mathbf{Y}_*^t) - e_t(\boldsymbol{\Theta}_0, \mathbf{Y}_*^t)\}' \boldsymbol{\Sigma}_t^{-1}(\widehat{\boldsymbol{\Theta}}_k) \{e_t(\widehat{\boldsymbol{\Theta}}_k, \mathbf{Y}_*^t) - e_t(\boldsymbol{\Theta}_0, \mathbf{Y}_*^t)\}\Big]. \end{aligned} \tag{6}$$

Now consider using $-2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$ as an estimator of the directed divergence $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$. For the expectation of $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$, we can write

$$\begin{aligned} E\{\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)\} &= E\{-2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)\} + E[\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k) - \{-2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)\}] \\ &= E\{-2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)\} + B_T(k, \boldsymbol{\Theta}_0). \end{aligned}$$

This decomposition suggests that $-2 \log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k)$ serves as a biased estimator of the directed divergence,

with bias adjustment provided by $B_T(k, \boldsymbol{\Theta}_0)$. Using (4) and (6), we have

$$\begin{aligned} B_T(k, \boldsymbol{\Theta}_0) &= -Tp + \sum_{t=1}^{T} E[tr\{\boldsymbol{\Sigma}_t(\boldsymbol{\Theta}_0)\boldsymbol{\Sigma}_t^{-1}(\widehat{\boldsymbol{\Theta}}_k)\}] \\ &\quad + \sum_{t=1}^{T} E\Big[\{e_t(\widehat{\boldsymbol{\Theta}}_k, \mathbf{Y}_*^t) - e_t(\boldsymbol{\Theta}_0, \mathbf{Y}_*^t)\}'\boldsymbol{\Sigma}_t^{-1}(\widehat{\boldsymbol{\Theta}}_k)\{e_t(\widehat{\boldsymbol{\Theta}}_k, \mathbf{Y}_*^t) - e_t(\boldsymbol{\Theta}_0, \mathbf{Y}_*^t)\}\Big]. \end{aligned} \tag{7}$$

As previously discussed, the bias adjustment is asymptotically independent of the unknown parameter $\boldsymbol{\Theta}_0$. $B_T(k, \boldsymbol{\Theta}_0)$ may therefore be approximated via Monte Carlo simulation using a conveniently chosen simulation parameter $\boldsymbol{\Theta}_s$. Following Hurvich, Shumway, and Tsai (1990), we aim to specify a simulation parameter which ensures that the process $\mathbf{y}_t$ is Gaussian white noise: i.e., $\mathbf{y}_t \overset{iid}{\sim} N(0, \mathbf{I})$, $t = 1, 2, \ldots, T$.

With $\mathbf{y}_t \overset{iid}{\sim} N(0, \mathbf{I})$, it follows that

$$e_t(\boldsymbol{\Theta}_s, \mathbf{Y}^t) = \mathbf{y}_t - \mathbf{A}_t\tilde{\mathbf{x}}_t(\boldsymbol{\Theta}_s, \mathbf{Y}^{t-1}) = \mathbf{y}_t - E(\mathbf{y}_t|\mathbf{Y}^{t-1}) = \mathbf{y}_t,$$

and

$$\boldsymbol{\Sigma}_t(\boldsymbol{\Theta}_s) = E\{e_t(\boldsymbol{\Theta}_s, \mathbf{Y}^t)e_t(\boldsymbol{\Theta}_s, \mathbf{Y}^t)'\} = E(\mathbf{y}_t\mathbf{y}_t') = \mathbf{I},$$

where the expectations are taken with respect to $f(\mathbf{Y}|\boldsymbol{\Theta}_s)$. Substituting these expressions into (7) yields

$$\begin{aligned} B_T(k, \boldsymbol{\Theta}_s) &= -Tp + \sum_{t=1}^{T} E[tr\{\boldsymbol{\Sigma}_t^{-1}(\widehat{\boldsymbol{\Theta}}_k)\}] \\ &\quad + \sum_{t=1}^{T} E[\{\mathbf{A}_t\tilde{\mathbf{x}}_t(\widehat{\boldsymbol{\Theta}}_k, \mathbf{Y}_*^{t-1})\}'\boldsymbol{\Sigma}_t^{-1}(\widehat{\boldsymbol{\Theta}}_k)\{\mathbf{A}_t\tilde{\mathbf{x}}_t(\widehat{\boldsymbol{\Theta}}_k, \mathbf{Y}_*^{t-1})\}]. \end{aligned} \tag{8}$$

The penalty term of our criterion is then obtained by simulating the two sums in (8). Specifically, suppose $\mathbf{Y}(1), \mathbf{Y}(2), \ldots, \mathbf{Y}(M), \mathbf{Y}_*(1), \mathbf{Y}_*(2), \ldots, \mathbf{Y}_*(M)$, represent $2M$ vectors of data generated with $\mathbf{y}_t \overset{iid}{\sim} N(0, \mathbf{I})$. Then, with $\widehat{\boldsymbol{\Theta}}_k(1), \widehat{\boldsymbol{\Theta}}_k(2), \ldots, \widehat{\boldsymbol{\Theta}}_k(M)$, representing the ML estimates corresponding to $\mathbf{Y}(1), \mathbf{Y}(2), \ldots, \mathbf{Y}(M)$, the quantity

$$\begin{aligned} \widehat{B}_T(k, \boldsymbol{\Theta}_s) &= -Tp + \frac{1}{M}\sum_{j=1}^{M}\Bigg(\sum_{t=1}^{T}[tr\{\boldsymbol{\Sigma}_t^{-1}(\widehat{\boldsymbol{\Theta}}_k(j))\}] \\ &\quad + \sum_{t=1}^{T}[\{\mathbf{A}_t\tilde{\mathbf{x}}_t(\widehat{\boldsymbol{\Theta}}_k(j), \mathbf{Y}_*^{t-1}(j))\}'\boldsymbol{\Sigma}_t^{-1}(\widehat{\boldsymbol{\Theta}}_k(j))\{\mathbf{A}_t\tilde{\mathbf{x}}_t(\widehat{\boldsymbol{\Theta}}_k(j), \mathbf{Y}_*^{t-1}(j))\}]\Bigg) \end{aligned} \tag{9}$$

provides a Monte Carlo approximation to (8).

AICi is now obtained by

$$\text{AICi} = -2\log f(\mathbf{Y}|\widehat{\boldsymbol{\Theta}}_k) + \widehat{B}_T(k, \boldsymbol{\Theta}_s). \tag{10}$$

The penalty term of AICi, $\widehat{B}_T(k, \boldsymbol{\Theta}_s)$, serves as a Monte Carlo approximation to (8), which should be asymptotically equivalent to the bias adjustment (7) under the assumptions given in section 2.

We emphasize that the developments leading to (9) can be adapted to accommodate a more general framework of the state-space model, allowing for covariates and correlated error processes (e.g., de Jong 1989).

5 Simulations

In what follows, we present simulation results to evaluate the performance of AICi both as an estimator of the Kullback-Leibler information and as a model selection criterion. We consider two settings. In the first, the candidate class consists of models where the state process is a univariate autoregression. In the second, the class is comprised of models where the state process is an additive combination of a univariate autoregression and a seasonal component. In both settings, the goal is to choose the order of the autoregression.

For each candidate class of interest, the penalty term of AICi is computed via simulation for various model orders, and the resulting values are tabulated and later used in the evaluation of (10).

Next we outline the computation of the penalty term, then present and summarize our simulation results.

5.1 Penalty Term Computation

As emphasized by Hurvich, Shumway, and Tsai (1990), once values of the penalty term of AICi are tabulated for a given candidate class, AICi may be conveniently evaluated for any application involving that class. We now describe the computation of $\widehat{B}_T(k, \boldsymbol{\Theta}_s)$ for each candidate class employed in the simulations.

In our first candidate class of models, the state process is a univariate Gaussian autoregression z_t of order p. The observed process y_t consists of the state-process z_t corrupted by additive observation noise v_t. The resulting state-space process can be characterized by two sets of equations:

$$\begin{aligned} y_t &= z_t + v_t, \\ z_t &= \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + \ldots + \varphi_p z_{t-p} + \epsilon_t. \end{aligned} \tag{11}$$

Here, $v_t \overset{iid}{\sim} N(0, \sigma_R^2)$ and $\epsilon_t \overset{iid}{\sim} N(0, \sigma_Q^2)$. We shall denote such a state-space process by ARN(p).

Model (11) is put in the general state-space form of (1) by writing the observation equation as

$$y_t = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} z_t \\ z_{t-1} \\ \vdots \\ z_{t-p+1} \end{pmatrix} + v_t, \tag{12}$$

and the state-equation as

$$\begin{pmatrix} z_t \\ z_{t-1} \\ \vdots \\ z_{t-p+1} \end{pmatrix} = \begin{pmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_{p-1} & \varphi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} z_{t-1} \\ z_{t-2} \\ \vdots \\ z_{t-p} \end{pmatrix} + \begin{pmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{13}$$

Here, the transition structure $\boldsymbol{\Phi}$ is as defined in (13), $\mathbf{Q}$ is a $p \times p$ matrix with all zero entries except for the entry in the upper left-hand corner, which is σ_Q^2, and $\mathbf{R} = \sigma_R^2$.

Our first simulation sets are based on a sample size of $T = 18$ and consider a candidate class of ARN(p) models of orders 1 through 10. Thus, we compute ten values of $\widehat{B}_T(k, \boldsymbol{\Theta}_s)$ corresponding to $T = 18$ and

$p = 1, 2, \ldots, 10$ ($k = p + 2$). For a given model order, we evaluate $\widehat{B}_T(k, \boldsymbol{\Theta}_s)$ via (9) based on M=1000 replications. The estimates $\widehat{\boldsymbol{\Theta}}_k(j)$ are obtained using 100 iterations of the EM algorithm.

The simulated values of the penalty term $\widehat{B}_T(k, \boldsymbol{\Theta}_s)$ are given in Table 1. In our first collection of simulation sets, we make use of these values in evaluating AICi via (10).

Order, p	1	2	3	4	5	6	7	8	9	10
$\widehat{B}_T(k, \boldsymbol{\Theta}_s)$	5.7	8.6	11.4	15.6	19.5	29.4	48.9	65.8	111.2	217.8

Table 1: AICi penalty term for ARN(p) models, $T = 18$

In our second candidate class of models, the state process is an additive combination of (i) a univariate Gaussian autoregression z_t of order p, and (ii) a seasonal Gaussian process s_t of known periodicity d. The observed process y_t consists of the state process $z_t + s_t$ corrupted by additive observation noise v_t. The resulting state-space process can be characterized by three sets of equations:

$$\begin{aligned} y_t &= z_t + s_t + v_t, \\ z_t &= \varphi_1 z_{t-1} + \varphi_2 z_{t-2} + \ldots + \varphi_p z_{t-p} + \epsilon_t, \\ s_t &= -s_{t-1} - s_{t-2} - \ldots - s_{t-d+1} + \eta_t. \end{aligned} \tag{14}$$

Here, $v_t \overset{iid}{\sim} N(0, \sigma_R^2)$, $\epsilon_t \overset{iid}{\sim} N(0, \sigma_Q^2)$, and $\eta_t \overset{iid}{\sim} N(0, \sigma_\eta^2)$, We shall denote such a state-space process by ARSN(p, d).

Note that the seasonal component can easily be written in the form of the state equation in (1):

$$\begin{pmatrix} s_t \\ s_{t-1} \\ \vdots \\ s_{t-d+2} \end{pmatrix} = \begin{pmatrix} -1 & \cdots & \cdots & \cdots & -1 \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} s_{t-1} \\ s_{t-2} \\ \vdots \\ s_{t-d+1} \end{pmatrix} + \begin{pmatrix} \eta_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{15}$$

Since the state process consists of structures that correspond to specific attributes of the underlying phenomenon, a model such as (14) is often referred to as *structural*. With (14), the structures are z_t, a process for which the dynamics are governed by the autoregressive parameters $\varphi_1, \varphi_2, \ldots, \varphi_p$; and s_t, a non-stationary process that cycles through d "seasons" each "yearly" period. Harvey (1989) details econometric models with $d = 4$, for quarterly measurements, and $d = 12$, for monthly measurements.

To put (14) is state-space form, let $\mathbf{x}_t = (z_t, \ldots, z_{t-p}, s_t, \ldots, s_{t-d+2})$,$'$ $\mathbf{A}_1$ be the observation operator used in (12), and $\mathbf{A}_2$ be a $1 \times (d-1)$ vector defined as $(1, 0, \ldots, 0)$. The observation equation can then be written as

$$y_t = (\mathbf{A}_1 \ \mathbf{A}_2)\, \mathbf{x}_t + v_t. \tag{16}$$

Here, we let $\mathbf{R} = \sigma_R^2$. For the state equation, let the $(p + d - 1) \times 1$ vector $\boldsymbol{\gamma}_t$ be defined as $\boldsymbol{\gamma}_t = (\epsilon_t, 0, \ldots, 0, \eta_t, 0, \ldots, 0)'$, and let $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ be the transition structures used in (13) and (15), respectively.

The state-equation can then be written as

$$\mathbf{x}_t = \begin{pmatrix} \mathbf{\Phi}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{\Phi}_2 \end{pmatrix} \mathbf{x}_{t-1} + \boldsymbol{\gamma}_t. \tag{17}$$

For the objects in (1), $\mathbf{\Phi}$ is as given in (17), with $\mathbf{Q}$ defined as follows. Let $\mathbf{Q}_1$ be a $p \times p$ matrix with all zero entries except for the entry in the upper left-hand corner, which is σ_Q^2. Similarly, let $\mathbf{Q}_2$ be a $(d-1) \times (d-1)$ matrix with all zero entries except for the entry in the upper left-hand corner, which is σ_η^2. Then, $\mathbf{Q}$ can be defined as the $(p + d - 1) \times (p + d - 1)$ matrix

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{Q}_2 \end{pmatrix}. \tag{18}$$

Our second simulation sets are based on a sample size of $T = 18$ and consider a candidate class of ARSN$(p, 4)$ models of orders 1 through 10. Again, we compute ten values of $\widehat{B}_T(k, \mathbf{\Theta}_s)$ corresponding to $T = 18$ and $p = 1, 2, \ldots, 10$ ($k = p + 2$). For a given model order, $\widehat{B}_T(k, \mathbf{\Theta}_s)$ is evaluated via (9) based on M=1000 replications. The estimates $\widehat{\mathbf{\Theta}}_k(j)$ are obtained using 100 iterations of the EM algorithm.

The simulated values of the penalty term $\widehat{B}_T(k, \mathbf{\Theta}_s)$ are given in Table 2. In our second collection of simulation sets, we make use of these values in evaluating AICi via (10).

Order, p	1	2	3	4	5	6	7	8	9	10
$\widehat{B}_T(k, \mathbf{\Theta}_s)$	4.8	7.6	11.0	21.8	36.61	59.8	186.8	4128.1	5493.4	13789.9

Table 2: AICi penalty term for ARSN$(p, 4)$ models, $T = 18$

5.2 Simulation Settings

We compile six simulation sets, two based on ARN(p) models and four based on ARSN$(p, 4)$ models. The generating models are listed in Table 3. For the noise variances, we set $\sigma_R^2 = 0.1$, $\sigma_Q^2 = 1.0$, and $\sigma_\eta^2 = 0.1$. These variances lead to signal-to-noise ratios that are common in practice. In each simulation set, 1000 realizations of size $T = 18$ are generated, and for every realization, candidate models based on autoregressions of order 1 through 10 are fit to the data.

Set	State-Space Model
I.	$y_t = z_t + v_t,\ z_t = .99z_{t-1} - .80z_{t-2} + \epsilon_t.$
II.	$y_t = z_t + v_t,\ z_t = -.90z_{t-3} + \epsilon_t.$
III.	$y_t = z_t + s_t + v_t,\ z_t = .99z_{t-1} - .80z_{t-2} + \epsilon_t.$
IV.	$y_t = z_t + s_t + v_t,\ z_t = -.90z_{t-3} + \epsilon_t.$
V.	$y_t = z_t + s_t + v_t,\ z_t = -.80z_{t-1} + \epsilon_t.$
VI.	$y_t = z_t + s_t + v_t,\ z_t = 1.40z_{t-1} - .49z_{t-2} + \epsilon_t.$

Table 3: Models for simulation sets

In addition to AICi, the other criteria considered in the simulations are FPE (Akaike 1969), SIC, BIC, HQ, AIC, and AICc. The complete set of criteria is listed below. In the definitions, $k = p + 2$, and $\widehat{\sigma}^2$ equals

the estimate of the steady-state innovations variance, i.e., $\widehat{\sigma}^2 = \mathbf{\Sigma}_t(\widehat{\mathbf{\Theta}}_k)$ for "large" t.

$$\begin{aligned}
\text{AIC} &= -2\log f(Y|\widehat{\mathbf{\Theta}}_k) + 2k, \\
\text{AICc} &= -2\log f(Y|\widehat{\mathbf{\Theta}}_k) + \frac{2T(p+1)}{T-p-2}, \\
\text{FPE} &= T\frac{(T+p)}{(T-p)}\widehat{\sigma}^2, \\
\text{HQ} &= -2\log f(Y|\widehat{\mathbf{\Theta}}_k) + 2k\log(\log T), \\
\text{BIC} &= (T-p)\log\left(\frac{T\widehat{\sigma}^2}{T-p}\right) + p\log\left\{\frac{\left(\sum_{t=1}^{T} y_t^2\right) - T\widehat{\sigma}^2}{p}\right\},
\end{aligned}$$

and

$$\text{SIC} = -2\log f(Y|\widehat{\mathbf{\Theta}}_k) + k\log T.$$

(The version of AICc used here is based on the definition provided in Hurvich, Shumway, and Tsai 1990.)

Since the development of AICc, FPE, HQ, and BIC do not extend in any obvious manner to the ARN(p) or ARSN(p, d) frameworks, it should be emphasized that their application in the present setting is somewhat ad-hoc. However, the performance of these criteria (as well as AIC and SIC) will provide a useful baseline with which to evaluate the effectiveness of AICi.

For each of the AIC-type criteria (AICi, AICc, and AIC), the average criterion value over the 1000 realizations is computed for each of the candidate model orders 1 through 10. Also, since the true parameters are known (see Table 3 and the associated variances), the value of $E\{\delta(\mathbf{\Theta}_0, \widehat{\mathbf{\Theta}}_k)\}$ can be simulated for each model order. This allows us to plot the criterion averages and simulated values of $E\{\delta(\mathbf{\Theta}_0, \widehat{\mathbf{\Theta}}_k)\}$ against the model orders, and enables us to judge the relative effectiveness of the criteria as unbiased estimators of $\delta(\mathbf{\Theta}_0, \widehat{\mathbf{\Theta}}_k)$.

Next we present the results of our simulation sets.

5.3 Simulation Results

The order selections from our six simulation sets are presented in Tables 4 through 9. For a given criterion, each cell entry shows the number of times (out of 1000) a certain model order is chosen. For simulation sets I and III, Figures 1 and 2 feature plots of the the criterion averages for AICi, AICc, AIC, and the simulated values of $E\{\delta(\mathbf{\Theta}_0, \widehat{\mathbf{\Theta}}_k)\}$.

In sets I and II, the candidate class consists of models where the state process is a univariate autoregression. As shown in Table 4, AICi and AICc obtain the most correct order selections in set I. However, AICc chooses models that are excessively overparameterized more often than AICi, and our criterion slightly outperforms AICc in this regard. The remaining criteria perform relatively poorly, often choosing models that are grossly overspecified.

As an estimator of $\delta(\mathbf{\Theta}_0, \widehat{\mathbf{\Theta}}_k)$, Figure 1 illustrates that AICi exhibits less negative bias than either AICc or AIC. The strong propensity of AIC to select overspecified models can explained by this plot: as the model

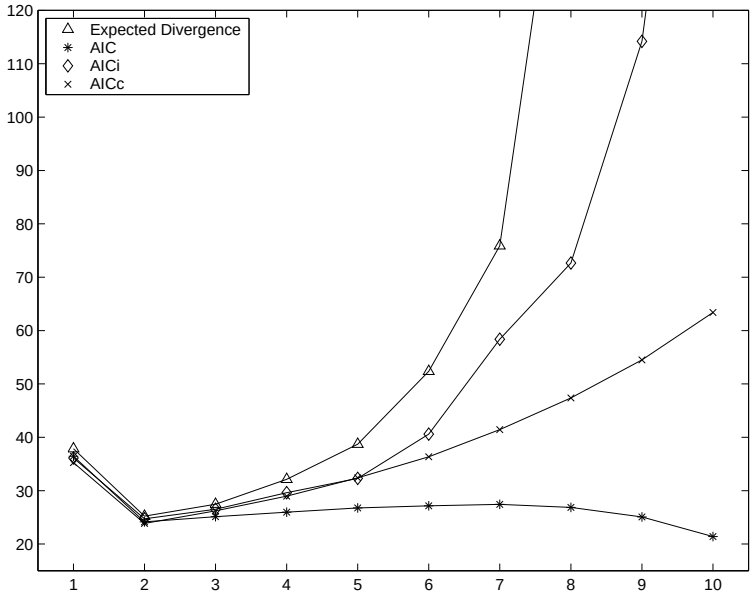

Figure 1: Expected divergence and average criterion values for simulation set I.

order is increased beyond the true model order (of $p = 2$), the negative bias of AIC becomes more extreme. For the largest model orders in the candidate class, the bias of AIC is so pronounced that the average value of AIC is less than the average value for the true model order. As a consequence, AIC often favors such models, even though they are grossly overparameterized and correspond to extremely high values of the Kullback-Leibler information.

Order	SIC	BIC	HQ	FPE	AIC	AICc	AICi
1	40	16	21	10	19	59	58
2	517	403	372	170	341	817	819
3	39	30	42	24	39	60	79
4	30	23	37	25	37	20	23
5	6	9	14	22	19	9	14
6	21	28	24	27	25	11	7
7	20	25	27	26	32	6	0
8	40	62	48	64	51	6	0
9	77	120	90	127	94	9	0
10	210	284	325	505	343	3	0

Table 4: Results for simulation set I. Generating model: ARN(2).

In set II, as illustrated in Table 5, AICi outperforms all other criteria in terms of obtaining the most correctly-specified model selections and the least overspecified model selections, but the performance of AICc is competitive. The remaining criteria share a pronounced tendency to favor grossly overfit models.

In sets III and VI, the candidate class is comprised of models where the state process is an additive combination of a univariate autoregression and a structural seasonal component. Sets III and VI are analogous

Order	SIC	BIC	HQ	FPE	AIC	AICc	AICi
1	58	7	26	6	22	107	95
2	38	32	23	8	18	61	55
3	404	258	286	123	266	699	769
4	55	46	49	31	50	60	48
5	23	21	19	23	19	13	25
6	21	34	24	20	26	20	8
7	24	34	27	30	28	12	0
8	54	86	71	75	77	9	0
9	123	184	166	223	173	17	0
10	200	298	309	461	321	2	0

Table 5: Results for simulation set II. Generating model: ARN(3).

to sets I and II, respectively, since they are based on the same autoregressions. Thus, sets I through IV allow us to investigate the impact of additional structural complexity on the selection performance of the criteria.

The results from set III are featured in Table 6. In comparing sets I and III on the basis of correct order selections, the performance of all criteria markedly deteriorates except for that of AICi. In particular, the inclusion of the seasonal process severely degrades the performance of AICc, now exhibiting a greater tendency to choose both underspecified and overspecified models. For the other criteria, the addition of the seasonal component generally results in more underfit selections (i.e., $p = 1$).

Figure 2 illustrates that AICi provides an estimator of $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$ with less negative bias than either AICc or AIC. Relative to Figure 1 (which pertains to set I), the negative bias of AICc and AIC is more pronounced. As can be seen, AIC dramatically underestimates $\delta(\boldsymbol{\Theta}_0, \widehat{\boldsymbol{\Theta}}_k)$ for all model orders beyond $p = 4$.

Order	SIC	BIC	HQ	FPE	AIC	AICc	AICi
1	72	84	35	8	31	121	129
2	343	254	210	44	185	660	815
3	32	30	24	10	23	42	50
4	12	8	18	11	16	24	5
5	25	30	14	15	17	24	1
6	39	57	45	29	42	40	0
7	49	52	52	46	49	16	0
8	87	118	113	110	119	28	0
9	150	162	196	247	206	33	0
10	191	205	293	480	312	12	0

Table 6: Results for simulation set III. Generating model: ARSN(2,4).

The results from set IV are featured in Table 7. In comparing the results for sets II (see Table 5) and IV based on correct order selections, the performance of all criteria deteriorates, including that of AICi. However, the deterioration is the least pronounced for AICi, which outperforms all other criteria in terms of obtaining the most correctly specified model selections and the least overspecified model selections. Again, inclusion of the seasonal component causes AICc to choose more underspecified and overspecified models.

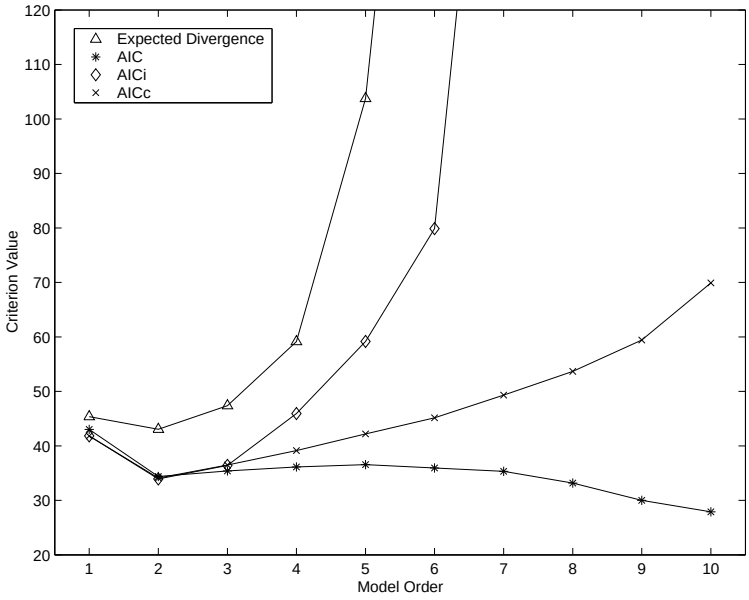

Figure 2: Expected divergence and average criterion values for simulation set III.

Moreover, as with the comparison between sets I and III, the addition of the seasonal component generally results in more underfit selections; specifically, more selections of the minimal order $p = 1$.

Order	SIC	BIC	HQ	FPE	AIC	AICc	AICi
1	69	75	37	4	32	158	202
2	20	14	14	1	12	43	76
3	173	122	97	13	83	449	699
4	64	74	51	16	50	105	23
5	33	38	27	12	27	34	0
6	58	58	48	31	49	46	0
7	64	82	69	58	63	26	0
8	125	157	147	154	155	61	0
9	160	187	207	235	212	59	0
10	234	193	303	476	317	19	0

Table 7: Results for simulation set IV. Generating model: ARSN(3,4).

The results from set V are featured in Table 8. In set V, the true order of the autoregression is $p = 1$. Thus, the criteria can only be evaluated based on their overfitting properties, and AICi clearly performs best in this regard. As with the previous simulation sets, AIC, FPE, and HQ exhibit the most extreme tendencies to favor overspecified models.

The results from the last set, set VI, are featured in Table 9. Here, all criteria appear to have difficulty in distinguishing between the correct model order $p = 2$ and the minimal model order $p = 1$. AICi again obtains the most correct order selections, yet chooses the order 1 model more frequently than the order 2 model. AICc, BIC, and SIC also choose the order 1 model more often than the order 2 model. Again, AICi

Order	SIC	BIC	HQ	FPE	AIC	AICc	AICi
1	426	430	238	51	214	727	852
2	50	29	44	12	45	84	102
3	24	19	16	2	15	32	39
4	25	19	23	8	23	35	7
5	19	11	22	10	22	13	0
6	21	26	34	27	36	15	0
7	33	42	45	50	46	10	0
8	83	118	122	142	119	46	0
9	123	137	176	241	185	35	0
10	196	169	280	457	295	3	0

Table 8: Results for simulation set V. Generating model: ARSN(1,4).

does not favor overfit models. All other criteria, with the possible exception of AICc, exhibit inordinate overfitting tendencies.

Order	SIC	BIC	HQ	FPE	AIC	AICc	AICi
1	264	278	129	26	108	477	520
2	185	166	139	47	129	342	424
3	22	26	25	9	25	38	49
4	16	13	10	6	10	23	7
5	28	30	29	20	29	29	0
6	44	50	42	29	39	34	0
7	55	58	56	50	58	15	0
8	72	89	110	117	115	22	0
9	133	127	190	253	203	17	0
10	181	163	270	443	284	3	0

Table 9: Results for simulation set VI. Generating model: ARSN(2,4).

6 Conclusion

Based on the work of Hurvich, Shumway, and Tsai (1990), we develop a model selection criterion for the linear state-space model. The criterion, AICi, is straightforward to evaluate once its penalty term has been tabulated via Monte Carlo simulations.

The results of our simulations show that AICi performs effectively as a model selection criterion in small-sample settings. Our results also indicate that AICi estimates the Kullback-Leibler information with less bias than traditional AIC or corrected AIC, that AICi does not exhibit a propensity to favor grossly overparameterized models, and that AICi generally outperforms its competitors in terms of correct order selections. Importantly, if the candidate class consists of models with structural complexity, the order selection properties of AICi appear to be far superior to those of its competitors.

Our development suggests that an AICi may be devised for additional settings in which the forecast

density may be conveniently expressed and evaluated. Such settings could include observation-driven models for serial counts (cf. Brockwell and Davis 1991, p291-303) as well as other practical time series frameworks.

Acknowledgements

The authors wish to extend their appreciation to two referees whose valuable suggestions helped to significantly improve the original version of the manuscript. They also thank Professor Peter Bickel for useful technical comments.

The work of the first author was mainly completed while a postdoctoral fellow at the Geophysical Statistics Project at NCAR, Boulder, CO, and was supported by the National Science Foundation under grants DMS 9815344 and DMS 9312686.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institue of Statistical Mathematics 21*, 243–247.

Akaike, H. (1973). *2nd International Symposium on Information Theory*, Chapter Information theory and an extension of maximum likelihood principle, pp. 267–281. Akademia Kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC-19*, 716–723.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Mathematical Statistics A*(30), 9–14.

Brockwell, P. and R. Davis (1991). *Time Series: Theory and Methods* (2 ed.). Springer-Verlag.

Cavanaugh, J. (1997). Unifying the derivations of the Akaike and corrected Akaike information criteria. *Statistics and Probability Letters 31*, 201–208.

Cavanaugh, J. and R. Shumway (1997). A bootstrap-corrected variant of AIC for state-space model selection. *Statistica Sinica 7*, 473–496.

de Jong, P. (1989). Smoothing and interpolation with the state-space model. *Journal of the American Statistical Association 84*, 1085–1088.

Hannan, E. and B. Quinn (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*(41), 190–195.

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

Harvey, A. and P. Todd (1983). Forecasting economic time series with structural and Box-Jenkins models: A case study with comments. *Journal of Business and Economic Statistics 1*, 299–315.

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

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

Hurvich, C. and C. Tsai (1993). A corrected Akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis 14*, 271–279.

Johnson, R. and D. Wichern (1998). *Applied Multivariate Statistical Analysis* (4 ed.). Prentice-Hall.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Transactions ASME Journal of Basic Engineering 9*, 35–45.

Kitagawa, G. and W. Gersch (1984). A smoothness priors modeling of times series with trend and seasonality. *Journal of the American Statistical Association 79*, 378–389.

Kullback, S. (1968). *Information Theory and Statistics*. Dover.

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

Schweppe, F. (1965). Evaluation of likelihood functions for Gaussian signals. *IEEE Transactions on Information Theory 11*, 61–70.

Shumway, R. and D. Stoffer (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis 3*, 253–264.

Shumway, R. and D. Stoffer (2000). *Time Series Analysis and Its Applications*. Springer-Verlag.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communication in Statistics* $A(7)$, 13–26.