A Large-Sample Model Selection Criterion Based on Kullback's Symmetric Divergence

by

Joseph E. Cavanaugh * †

Department of Statistics, University of Missouri – Columbia

Abstract

The Akaike information criterion, AIC, is a widely known and extensively used tool for statistical model selection. AIC serves as an asymptotically unbiased estimator of a variant of Kullback's directed divergence between the true model and a fitted approximating model. The directed divergence is an asymmetric measure of separation between two statistical models, meaning that an alternate directed divergence may be obtained by reversing the roles of the two models in the definition of the measure. The sum of the two directed divergences is Kullback's symmetric divergence. Since the symmetric divergence combines the information in two related though distinct measures, it functions as a gauge of model disparity which is arguably more sensitive than either of its individual components. With this motivation, we propose a model selection criterion which serves as an asymptotically unbiased estimator of a variant of the symmetric divergence between the true model and a fitted approximating model. We examine the performance of the criterion relative to other well-known criteria in a simulation study.

Keywords: AIC, Akaike information criterion, I-divergence, J-divergence, Kullback-Leibler information, relative entropy.

*Correspondence: Joseph E. Cavanaugh, Department of Statistics, 222 Math Sciences Bldg., University of Missouri, Columbia, MO 65211.

†This research was supported by NSF grant DMS–9704436.

1. Introduction

One of the most important problems confronting an investigator in statistical modeling is the choice of an appropriate model to characterize the underlying data. This determination can often be facilitated by the use of a model selection criterion, which judges the propriety of a fitted model by assessing whether it offers an optimal balance between "goodness of fit" and parsimony.

The first model selection criterion to gain widespread acceptance was the Akaike (1973, 1974) information criterion, AIC. Many other criteria have been subsequently introduced and studied, including well-known measures by Mallows (1973), Parzen (1974), Schwarz (1978), Rissanen (1978), Akaike (1978), Hannan and Quinn (1979), and Hurvich and Tsai (1989).

AIC serves as an asymptotically unbiased estimator of a variant of Kullback's (1968, p. 5) directed divergence between the true model and a fitted approximating model. The directed divergence, also known as the Kullback-Leibler (1951) information, the I-divergence, or the relative entropy, assesses the dissimilarity between two statistical models. It is an asymmetric measure, meaning that an alternate directed divergence may be obtained by reversing the roles of the two models in the definition of the measure. The sum of the two directed divergences is Kullback's (1968, p. 6) symmetric divergence, also known as the J-divergence. Since the symmetric divergence combines the information in two related though distinct measures, it functions as a gauge of model disparity which is arguably more sensitive than either of its individual components. With this motivation, we propose a model selection criterion which serves as an asymptotically unbiased estimator of a variant of the symmetric divergence between the true model and a fitted approximating model. We examine the performance of this criterion relative to other well-known criteria in a simulation study where the objective is to determine the order of an autoregressive process.

In Section 2, we present a short discussion of relevant background material, including an overview of Kullback's directed divergence and AIC. In Section 3, we discuss Kullback's symmetric divergence and introduce a large-sample model selection criterion based on this measure. In Section 4, we present the results of our simulation study.

2. Kullback's Directed Divergence and AIC

We begin with a brief description of the model selection problem our methodology attempts to address. Suppose a collection of data Y has been generated according to an unknown parametric model or density $f(Y|\,\theta_o)$. We endeavor to find a fitted parametric model which provides a suitable approximation to $f(Y|\,\theta_o)$.

Let $\mathcal{F}(k) = \{f(Y|\,\theta_k) \mid \theta_k \in \Theta(k)\}$ denote a k-dimensional parametric family, i.e., a family in which the parameter space $\Theta(k)$ consists of k-dimensional vectors whose components are functionally independent. Let $\hat{\theta}_k$ denote a vector of estimates obtained by maximizing the likelihood function $f(Y|\,\theta_k)$ over $\Theta(k)$, and let $f(Y|\,\hat{\theta}_k)$ denote the corresponding fitted model.

Suppose our goal is to search among a collection of families $\{\mathcal{F}(k_1), \mathcal{F}(k_2), \ldots, \mathcal{F}(k_L)\}$ for the fitted model $f(Y|\,\hat{\theta}_k)$, $k \in \{k_1, k_2, \ldots, k_L\}$, which serves as the "best" approximation to $f(Y|\,\theta_o)$. For simplicity, we will assume $k_i = i$ for $i = 1, 2, \ldots, L$, so that the collection consists of families of dimensions 1 through L (cf. Akaike, 1973, p. 272). Since our objective amounts to choosing an appropriate dimension $k \in \{1, 2, \ldots, L\}$ for the fitted model $f(Y|\,\hat{\theta}_k)$, our model selection problem can therefore be viewed as a problem of dimension determination.

We refer to $f(Y|\,\theta_o)$ as the *true* or *generating* model. We refer to any $f(Y|\,\theta_k)$ other than $f(Y|\,\theta_o)$ as an *approximating* or *candidate* model. To determine which of the fitted models $\{f(Y|\,\hat{\theta}_1), f(Y|\,\hat{\theta}_2), \ldots, f(Y|\,\hat{\theta}_L)\}$ best resembles $f(Y|\,\theta_o)$, we require a measure which provides a suitable reflection of the disparity between the true model $f(Y|\,\theta_o)$ and an approximating model $f(Y|\,\theta_k)$. Kullback's directed divergence is one such measure.

For two arbitrary parametric densities $f(Y|\,\theta\,)$ and $f(Y|\,\theta^*)$, Kullback's *directed divergence* between $f(Y|\,\theta\,)$ and $f(Y|\,\theta^*)$ with respect to $f(Y|\,\theta\,)$ is defined as

$$I(\theta, \theta^*) = \mathrm{E}_\theta \left\{ \ln \frac{f(Y|\,\theta\,)}{f(Y|\,\theta^*)} \right\}, \tag{2.1}$$

where E_θ denotes the expectation under $f(Y|\,\theta\,)$. Thus, $I(\theta_o, \theta_k)$ defines the directed divergence between $f(Y|\,\theta_o)$ and $f(Y|\,\theta_k)$ with respect to $f(Y|\,\theta_o)$.

Intuitively, $I(\theta_o, \theta_k)$ can be interpreted in the following manner. For a particular sample

Y generated under the true model $f(Y|\theta_o)$, the measure $\ln\{f(Y|\theta_o)/f(Y|\theta_k)\}$ assesses how well the approximating model $f(Y|\theta_k)$ "fits" or conforms to Y in comparison to the true model $f(Y|\theta_o)$. Thus, $I(\theta_o, \theta_k)$ evaluates the average measure of fit $\ln\{f(Y|\theta_o)/f(Y|\theta_k)\}$ taken over many realizations Y generated under the true model $f(Y|\theta_o)$. It is well known that $I(\theta_o, \theta_k) \geq 0$ with equality if and only if $\theta_k = \theta_o$ (Kullback, 1968, pp. 14–15).

Now for $f(Y|\theta)$ and $f(Y|\theta^*)$, define

$$d(\theta, \theta^*) = \mathrm{E}_\theta\{-2\ln f(Y|\theta^*)\}. \tag{2.2}$$

From (2.1) and (2.2), note that we can write

$$2I(\theta_o, \theta_k) = d(\theta_o, \theta_k) - d(\theta_o, \theta_o).$$

Since $d(\theta_o, \theta_o)$ does not depend on θ_k, any ranking of a set of candidate models corresponding to values of $I(\theta_o, \theta_k)$ would be identical to a ranking corresponding to values of $d(\theta_o, \theta_k)$. Hence, for the purpose of discriminating among various candidate models, $d(\theta_o, \theta_k)$ serves as a valid substitute for $I(\theta_o, \theta_k)$.

The preceding discussion suggests that

$$d(\theta_o, \hat{\theta}_k) = \mathrm{E}_{\theta_o}\{-2\ln f(Y|\theta_k)\}|_{\theta_k=\hat{\theta}_k} \tag{2.3}$$

would provide a suitable measure of the separation between the generating model $f(Y|\theta_o)$ and a fitted candidate model $f(Y|\hat{\theta}_k)$. Yet evaluating $d(\theta_o, \hat{\theta}_k)$ is not possible, since doing so requires knowledge of θ_o. The work of Akaike (1973), however, suggests that $-2\ln f(Y|\hat{\theta}_k)$ serves as a biased estimator of $d(\theta_o, \hat{\theta}_k)$, and that the bias adjustment

$$\mathrm{E}_{\theta_o}\{d(\theta_o, \hat{\theta}_k)\} - \mathrm{E}_{\theta_o}\{-2\ln f(Y|\hat{\theta}_k)\} \tag{2.4}$$

can often be asymptotically estimated by twice the dimension of $\hat{\theta}_k$. (Here, $f(Y|\hat{\theta}_k)$ represents the empirical likelihood.)

Thus, since k denotes the dimension of $\hat{\theta}_k$, under appropriate conditions, the expected value of

$$\mathrm{AIC} = -2\ln f(Y|\hat{\theta}_k) + 2k$$

should asymptotically approach the expected value of $d(\theta_o, \hat{\theta}_k)$, say

$$\Delta(\theta_o, k) = \mathrm{E}_{\theta_o}\{d(\theta_o, \hat{\theta}_k)\}.$$

Specifically, one can establish that

$$\mathrm{E}_{\theta_o}\{\mathrm{AIC}\} + o(1) = \Delta(\theta_o, k). \tag{2.5}$$

The demonstration of (2.5) utilizes the assumption that $f(Y|\,\theta_o) \in \mathcal{F}(k)$ (see Linhart and Zucchini, 1986, pp. 243–245). This assumption is satisfied if the true parameter vector θ_o is an element of a subset of $\Theta(k)$ comprised of vectors in which k_o components are free to vary and the remaining $(k - k_o)$ components are constrained to equal zero, $0 < k_o \leq k$. Thus, the assumption is satisfied if the family $\mathcal{F}(k)$ consists of models which are either overparameterized (when $k > k_o$) or correctly specified (when $k = k_o$).

The requirement that $f(Y|\,\theta_o) \in \mathcal{F}(k)$ is frequently employed in the development of model selection criteria since it facilitates tractable derivations. Although the requirement may seem strong, a criterion developed under this assumption often achieves its intended objective even when the assumption is violated, e.g., when $\mathcal{F}(k)$ consists of models which are underparameterized (see Linhart and Zucchini, pp. 20–22).

Note that property (2.5) allows us to view AIC as an asymptotically unbiased estimator of $d(\theta_o, \hat{\theta}_k)$, in the sense that the expectations of the stochastic quantities AIC and $d(\theta_o, \hat{\theta}_k)$ are within $o(1)$ of one another (cf. Findley, 1985; Shibata, 1997). It also allows us to view AIC as an asymptotically unbiased estimator of $\Delta(\theta_o, k)$, in the sense that the expectation of the stochastic quantity AIC is within $o(1)$ of the nonstochastic quantity $\Delta(\theta_o, k)$ (cf. Hurvich and Tsai, 1989; Hurvich, Shumway, and Tsai, 1990). In what follows, we utilize both interpretations.

3. A Large-Sample Model Selection Criterion Based on Kullback's Symmetric Divergence

In the previous section, the directed divergence between $f(Y|\,\theta_o)$ and $f(Y|\,\theta_k)$ with respect to $f(Y|\,\theta_o)$ was defined via (2.1) as $I(\theta_o, \theta_k)$. Similarly, one can define the directed

divergence between $f(Y|\,\theta_o)$ and $f(Y|\,\theta_k)$ with respect to $f(Y|\,\theta_k)$ as $I(\theta_k, \theta_o)$. Kullback's *symmetric divergence* is then defined as

$$J(\theta_o, \theta_k) = I(\theta_o, \theta_k) + I(\theta_k, \theta_o). \tag{3.1}$$

Note that $J(\theta_o, \theta_k) = J(\theta_k, \theta_o)$, whereas $I(\theta_o, \theta_k) \neq I(\theta_k, \theta_o)$ unless $\theta_k = \theta_o$; thus $J(\theta_o, \theta_k)$ is symmetric in its arguments whereas $I(\theta_o, \theta_k)$ is not.

It should be noted that the symmetric divergence was first introduced by Jeffreys (1946; 1983, p. 179), who was primarily interested in its use in providing a prior probability density for parameters. Kullback (1968), however, appears to have been principally responsible for introducing, investigating, and popularizing the widespread statistical applications of both the symmetric and the directed divergence.

Each of $I(\theta_o, \theta_k)$, $I(\theta_k, \theta_o)$, and $J(\theta_o, \theta_k)$ reflects the separation between the true and the approximating model. Intuitively, $J(\theta_o, \theta_k)$ can be interpreted in the following manner. For a particular sample Y generated under the true model $f(Y|\,\theta_o)$, the measure $\ln\{f(Y|\,\theta_o)/f(Y|\,\theta_k)\}$ assesses how well the approximating model $f(Y|\,\theta_k)$ "fits" Y in comparison to the true model $f(Y|\,\theta_o)$. Likewise, for a particular sample Z generated under the approximating model $f(Z|\,\theta_k)$, the measure $\ln\{f(Z|\,\theta_k)/f(Z|\,\theta_o)\}$ assesses how well the true model $f(Z|\,\theta_o)$ "fits" Z in comparison to the approximating model $f(Z|\,\theta_k)$. Thus, $J(\theta_o, \theta_k)$ evaluates the average combined measure of fit

$$[\ln\{f(Y|\,\theta_o)/f(Y|\,\theta_k)\}] + [\ln\{f(Z|\,\theta_k)/f(Z|\,\theta_o)\}]\,,$$

where the average is taken over many realizations of Y and Z generated, respectively, under $f(Y|\,\theta_o)$ and $f(Z|\,\theta_k)$. It is easily seen that $J(\theta_o, \theta_k) \geq 0$ with equality if and only if $\theta_k = \theta_o$.

Using (2.1), (2.2), and (3.1), we can write

$$2J(\theta_o, \theta_k) = \{d(\theta_o, \theta_k) - d(\theta_o, \theta_o)\} + \{d(\theta_k, \theta_o) - d(\theta_k, \theta_k)\}.$$

Since $d(\theta_o, \theta_o)$ does not depend on θ_k, for the purpose of discriminating among various candidate models, one could propose

$$K(\theta_o, \theta_k) = d(\theta_o, \theta_k) + \{d(\theta_k, \theta_o) - d(\theta_k, \theta_k)\} \tag{3.2}$$

as a substitute for the measure $J(\theta_o, \theta_k)$. Note that the relationship between $K(\theta_o, \theta_k)$ and $J(\theta_o, \theta_k)$ is analogous to the relationship between $d(\theta_o, \theta_k)$ and $I(\theta_o, \theta_k)$. Measures such as $K(\theta_o, \theta_k)$, $J(\theta_o, \theta_k)$, $d(\theta_o, \theta_k)$, and $I(\theta_o, \theta_k)$ are often called *discrepancies* (see Linhart and Zucchini, 1986, pp. 11–12).

Now evaluating (3.2) at $\theta_k = \hat{\theta}_k$ would lead to an appealing measure of separation between the generating model and the fitted candidate model, namely

$$K(\theta_o, \hat{\theta}_k) = d(\theta_o, \hat{\theta}_k) + \{d(\hat{\theta}_k, \theta_o) - d(\hat{\theta}_k, \hat{\theta}_k)\}. \tag{3.3}$$

(For clarity, we emphasize that

$$d(\hat{\theta}_k, \theta_o) = \mathrm{E}_{\theta_k}\{-2 \ln f(Y \,|\, \theta_o)\}|_{\theta_k = \hat{\theta}_k} \quad \text{and} \quad d(\hat{\theta}_k, \hat{\theta}_k) = \mathrm{E}_{\theta_k}\{-2 \ln f(Y \,|\, \theta_k)\}|_{\theta_k = \hat{\theta}_k};$$

$d(\theta_o, \hat{\theta}_k)$ is exhibited in (2.3).) Although $K(\theta_o, \hat{\theta}_k)$ is inaccessible, one might speculate that it is possible to construct an asymptotically unbiased estimator of $K(\theta_o, \hat{\theta}_k)$: i.e., an estimator with an expected value that asymptotically approaches the expected value of $K(\theta_o, \hat{\theta}_k)$, say

$$\Omega(\theta_o, k) = \mathrm{E}_{\theta_o}\{K(\theta_o, \hat{\theta}_k)\}. \tag{3.4}$$

The impetus for this notion comes from recognizing that $K(\theta_o, \hat{\theta}_k)$ and $\Omega(\theta_o, k)$ serve as the respective analogues of $d(\theta_o, \hat{\theta}_k)$ and $\Delta(\theta_o, k)$, and from recalling that (2.5) justifies AIC as an asymptotically unbiased estimator of $d(\theta_o, \hat{\theta}_k)$.

The following proposition derives a statistic with an expectation which is within $o(1)$ of $\Omega(\theta_o, k)$. To establish the result, we assume the usual regularity conditions required to ensure the consistency and asymptotic normality of the maximum likelihood vector $\hat{\theta}_k$. We also utilize the assumption that $f(Y \,|\, \theta_o) \in \mathcal{F}(k)$, i.e., that $\theta_o \in \Theta(k)$. Thus, we can take θ_o to be a k-dimensional vector with k_o non-zero components ($0 < k_o \leq k$).

Proposition. *Let*

$$\mathrm{KIC} = -2 \ln f(Y \,|\, \hat{\theta}_k) + 3k.$$

Then under the aforementioned conditions,

$$\mathrm{E}_{\theta_o}\{\mathrm{KIC}\} + o(1) = \Omega(\theta_o, k). \tag{3.5}$$

Proof. Referring to (3.3) and (3.4), note that we can write $\Omega(\theta_o, k)$ as

$$\begin{aligned}\Omega(\theta_o, k) &= \mathrm{E}_{\theta_o}\{-2\ln f(Y\,|\,\hat{\theta}_k)\} \\ &\quad + \left[d(\theta_o, \theta_o) - \mathrm{E}_{\theta_o}\{-2\ln f(Y\,|\,\hat{\theta}_k)\}\right] \qquad (3.6) \\ &\quad + \left[\mathrm{E}_{\theta_o}\{d(\theta_o, \hat{\theta}_k)\} - d(\theta_o, \theta_o)\right] \qquad (3.7) \\ &\quad + \mathrm{E}_{\theta_o}\{d(\hat{\theta}_k, \theta_o) - d(\hat{\theta}_k, \hat{\theta}_k)\}. \qquad (3.8)\end{aligned}$$

Clearly, the result will be established if we can verify that each of (3.6), (3.7), (3.8) is within $o(1)$ of k.

Define

$$\mathrm{I}(\theta_k) = \mathrm{E}_{\theta_k}\left\{-\frac{\partial^2 \ln f(Y\,|\,\theta_k)}{\partial\theta_k \partial\theta_k'}\right\} \quad \text{and} \quad \mathcal{I}(\theta_k, Y) = \left\{-\frac{\partial^2 \ln f(Y\,|\,\theta_k)}{\partial\theta_k \partial\theta_k'}\right\}.$$

Thus, $\mathrm{I}(\theta_o)$, $\mathrm{I}(\hat{\theta}_k)$, and $\mathcal{I}(\hat{\theta}_k, Y)$ respectively denote the true, the expected, and the observed Fisher information matrix (cf. Efron and Hinkley, 1978).

First, consider taking a second-order expansion of $-2\ln f(Y\,|\,\theta_o)$ about $\hat{\theta}_k$. Since the log-likelihood $\ln f(Y\,|\,\theta_k)$ is maximized at $\theta_k = \hat{\theta}_k$, one can establish

$$-2\ln f(Y\,|\,\theta_o) = -2\ln f(Y\,|\,\hat{\theta}_k) + (\hat{\theta}_k - \theta_o)'\mathcal{I}(\hat{\theta}_k, Y)(\hat{\theta}_k - \theta_o) + r_1(\theta_o, \hat{\theta}_k), \qquad (3.9)$$

where $r_1(\theta_o, \hat{\theta}_k)$ is $o_p(1)$ and $\mathrm{E}_{\theta_o}\{r_1(\theta_o, \hat{\theta}_k)\}$ is $o(1)$. Taking the expectation of both sides of (3.9) with respect to θ_o yields

$$d(\theta_o, \theta_o) - \mathrm{E}_{\theta_o}\{-2\ln f(Y\,|\,\hat{\theta}_k)\} = \mathrm{E}_{\theta_o}\{(\hat{\theta}_k - \theta_o)'\mathcal{I}(\hat{\theta}_k, Y)(\hat{\theta}_k - \theta_o)\} + o(1). \qquad (3.10)$$

Next, consider taking a second-order expansion in the second argument of $d(\theta_o, \hat{\theta}_k)$ about θ_o, and a second-order expansion in the second argument of $d(\hat{\theta}_k, \theta_o)$ about $\hat{\theta}_k$. Since the discrepancies $d(\theta_o, \theta_k)$ and $d(\hat{\theta}_k, \theta_k)$ are respectively minimized at $\theta_k = \theta_o$ and $\theta_k = \hat{\theta}_k$, one can establish

$$\begin{aligned}d(\theta_o, \hat{\theta}_k) &= d(\theta_o, \theta_o) + (\hat{\theta}_k - \theta_o)'\mathrm{I}(\theta_o)(\hat{\theta}_k - \theta_o) + r_2(\theta_o, \hat{\theta}_k), \qquad (3.11) \\ d(\hat{\theta}_k, \theta_o) &= d(\hat{\theta}_k, \hat{\theta}_k) + (\hat{\theta}_k - \theta_o)'\mathrm{I}(\hat{\theta}_k)(\hat{\theta}_k - \theta_o) + r_3(\theta_o, \hat{\theta}_k), \qquad (3.12)\end{aligned}$$

where $r_2(\theta_o, \hat{\theta}_k)$ and $r_3(\theta_o, \hat{\theta}_k)$ are both $o_p(1)$ and $\mathrm{E}_{\theta_o}\{r_2(\theta_o, \hat{\theta}_k)\}$ and $\mathrm{E}_{\theta_o}\{r_3(\theta_o, \hat{\theta}_k)\}$ are both $o(1)$. Taking the expectation of both sides of (3.11) and (3.12) with respect to θ_o yields

$$\mathrm{E}_{\theta_o}\{d(\theta_o, \hat{\theta}_k)\} - d(\theta_o, \theta_o) = \mathrm{E}_{\theta_o}\{(\hat{\theta}_k - \theta_o)' \mathrm{I}(\theta_o)(\hat{\theta}_k - \theta_o)\} + o(1), \tag{3.13}$$

$$\mathrm{E}_{\theta_o}\{d(\hat{\theta}_k, \theta_o) - d(\hat{\theta}_k, \hat{\theta}_k)\} = \mathrm{E}_{\theta_o}\{(\hat{\theta}_k - \theta_o)' \mathrm{I}(\hat{\theta}_k)(\hat{\theta}_k - \theta_o)\} + o(1). \tag{3.14}$$

Now the quadratic forms

$$(\hat{\theta}_k - \theta_o)' \mathcal{I}(\hat{\theta}_k, Y)(\hat{\theta}_k - \theta_o), \quad (\hat{\theta}_k - \theta_o)' \mathrm{I}(\theta_o)(\hat{\theta}_k - \theta_o), \quad (\hat{\theta}_k - \theta_o)' \mathrm{I}(\hat{\theta}_k)(\hat{\theta}_k - \theta_o)$$

all converge to centrally distributed chi-square random variables with k degrees of freedom. Thus, the expectations (under θ_o) of each of these quadratic forms is within $o(1)$ of k. This fact along with (3.10), (3.13), and (3.14) verifies that (3.6), (3.7), and (3.8) are each within $o(1)$ of k, thereby establishing (3.5). □

In the model selection problem outlined in Section 2, we search for a preferred model among a collection of fitted candidate models $\{f(Y|\hat{\theta}_1), f(Y|\hat{\theta}_2), \ldots, f(Y|\hat{\theta}_L)\}$. For the purpose of assessing the proximity between a certain fitted candidate model $f(Y|\hat{\theta}_k)$ and the true model $f(Y|\theta_o)$, either of the measures $J(\theta_o, \hat{\theta}_k)$ or $I(\theta_o, \hat{\theta}_k)$ could be entertained. Ideally, the fitted model corresponding to the minimum value of KIC will have a small symmetric divergence $J(\theta_o, \hat{\theta}_k)$, whereas the fitted model corresponding to the minimum value of AIC will have a small directed divergence $I(\theta_o, \hat{\theta}_k)$. Naturally, the question arises as to which of $J(\theta_o, \theta_k)$ or $I(\theta_o, \theta_k)$ is a better disparity measure for the application at hand.

A substantive comparison of $J(\theta_o, \theta_k)$ and $I(\theta_o, \theta_k)$ may be made by considering the manner in which $J(\theta_o, \hat{\theta}_k)$ and $I(\theta_o, \hat{\theta}_k)$ each gauge the separation between the true model and a particular fitted candidate model. The measure $I(\theta_o, \hat{\theta}_k)$ evaluates how well the fitted candidate model conforms on average to "new" samples generated under the true model. Thus, $I(\theta_o, \hat{\theta}_k)$ assesses the extent of the divergence between $f(Y|\theta_o)$ and $f(Y|\hat{\theta}_k)$, using $f(Y|\theta_o)$ as the benchmark for comparison. On the other hand, the measure $I(\hat{\theta}_k, \theta_o)$ evaluates how well the true model conforms on average to "new" samples generated under the fitted candidate model. Thus, $I(\hat{\theta}_k, \theta_o)$ assesses the extent of the divergence between $f(Y|\theta_o)$ and $f(Y|\hat{\theta}_k)$, using $f(Y|\hat{\theta}_k)$ as the benchmark for comparison. Although $I(\theta_o, \hat{\theta}_k)$ and $I(\hat{\theta}_k, \theta_o)$ are clearly

associated, the two measures are not redundant, since each judges the dissimilarity between $f(Y|\,\theta_o)$ and $f(Y|\,\hat{\theta}_k)$ in a different manner. In many instances where there is a meaningful discrepancy between $f(Y|\,\theta_o)$ and $f(Y|\,\hat{\theta}_k)$, one of the measures will be more pronounced than the other. As a result, the value of $J(\theta_o, \hat{\theta}_k)$ may better reflect an important disparity between $f(Y|\,\theta_o)$ and $f(Y|\,\hat{\theta}_k)$ than the value of $I(\theta_o, \hat{\theta}_k)$.

For the collection of fitted candidate models $\{f(Y|\,\hat{\theta}_1), f(Y|\,\hat{\theta}_2), \ldots, f(Y|\,\hat{\theta}_L)\}$, the estimates $\{\hat{\theta}_1, \hat{\theta}_2 \ldots, \hat{\theta}_L\}$ lie in the union of the parameter spaces for the corresponding candidate families, namely $\Theta(L)$. Assuming $\theta_o \in \Theta(L)$, the measures $J(\theta_o, \theta_k)$, $I(\theta_o, \theta_k)$, and $I(\theta_k, \theta_o)$ each tend to grow as θ_k moves from θ_o across $\Theta(L)$. Proceeding along certain directions from θ_o, the growth in $I(\theta_o, \theta_k)$ tends to exceed that of $I(\theta_k, \theta_o)$; along certain other directions, the opposite is true. Yet since $J(\theta_o, \theta_k) = I(\theta_o, \theta_k) + I(\theta_k, \theta_o)$, $J(\theta_o, \theta_k)$ exhibits the increasing tendencies in both $I(\theta_o, \theta_k)$ and $I(\theta_k, \theta_o)$. Thus, $J(\theta_o, \theta_k)$ functions as a more sensitive measure of model disparity than either of its individual components. It follows that $J(\theta_o, \hat{\theta}_k)$ may serve to better discriminate between a particular fitted candidate model and the true model than $I(\theta_o, \hat{\theta}_k)$. As a consequence, an estimator of $J(\theta_o, \hat{\theta}_k)$ (or $K(\theta_o, \hat{\theta}_k)$) may be preferable to an estimator of $I(\theta_o, \hat{\theta}_k)$ (or $d(\theta_o, \hat{\theta}_k)$) as a model selection criterion, provided that the former estimator is accurate enough to sufficiently reflect the sensitivity of $J(\theta_o, \hat{\theta}_k)$ (or $K(\theta_o, \hat{\theta}_k)$).

Although the preceding arguments are informal and warrant further investigation, they advance a perspective which promotes the use of $J(\theta_o, \theta_k)$ over $I(\theta_o, \theta_k)$ as a basis for the construction of model selection criteria. This perspective is further explored in the next section, where we examine the behavior of KIC, AIC, and other well-known criteria in a simulation study where the objective is to determine the order of an autoregressive process.

We close this section by discussing important considerations regarding the asymptotic justifications of KIC and AIC provided by (3.5) and (2.5). As previously outlined, in practice, one often considers a collection of fitted models representing various candidate families. The fitted model corresponding to the minimum value of the selection criterion is favored. When $f(Y|\,\theta_o) \in \mathcal{F}(k)$, (3.5) and (2.5) imply that KIC and AIC will respectively serve as approximately unbiased estimators of $K(\theta_o, \hat{\theta}_k)$ and $d(\theta_o, \hat{\theta}_k)$, provided that the sample size

is large. However, when $f(Y|\,\theta_o) \notin \mathcal{F}(k)$ or when the sample size is small, the penalty terms of $3k$ (for KIC) and $2k$ (for AIC) may be far less than the bias adjustments they are designed to approximate. As a result, KIC and AIC may tend to grossly underestimate their respective target measures.

Imprecise bias correction appears to be particularly problematic in settings where the candidate collection consists of families which are excessively overparameterized (i.e., $f(Y|\,\theta_o) \in \mathcal{F}(k)$ and $k >> k_o$) and the sample size is small (i.e., n is small or k is large relative to n). (For example, see Hurvich and Tsai, 1989; Hurvich, Shumway, and Tsai, 1990; Cavanaugh, 1997; Cavanaugh and Shumway, 1997.) In such settings, the larger fitted models are often characterized by relatively large discrepancies $K(\theta_o, \hat{\theta}_k)$ and $d(\theta_o, \hat{\theta}_k)$, yet relatively small values of KIC and AIC. Thus, the minimum criterion values often correspond to grossly overparameterized fitted models, models with values of $K(\theta_o, \hat{\theta}_k)$ and $d(\theta_o, \hat{\theta}_k)$ which are far from optimal. In Section 5, we discuss approaches which have been used to refine the penalty term of AIC for small-sample applications, and suggest utilizing these approaches to make analogous refinements to the penalty term of KIC.

4. Simulations

A univariate autoregressive (AR) process of order p can be represented as

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \epsilon_t, \quad \epsilon_t \sim iid\ N(0, \sigma^2).$$

Given a set of observations $Y = \{y_1, y_2, \ldots, y_n\}$ from such a process, suppose our objective is to determine an appropriate order p for the autoregression, where $1 \leq p \leq P$. Thus, to model the data Y, we consider a collection of P candidate families, where each family consists of AR models of a particular order p. Note that $\theta_k = (\sigma^2, \phi_1, \phi_2, \ldots, \phi_p)'$ and $k = p + 1$.

We examine the behavior of KIC, AIC, and certain other selection criteria by simulating a setting where the criteria are used to select p. In each of six simulation sets, 1000 samples are generated from a specified AR model of order p_o ($1 \leq p_o \leq P$). For every sample, AR candidate models of orders 1 through P are fit to the data, the criteria are evaluated, and the fitted AR model favored by each criterion is recorded. Over the 1000 samples, the order selections are tabulated, summarized, and reported.

The other criteria considered in our simulation sets are AICc (Hurvich and Tsai, 1989), FPE (Akaike, 1969), HQ (Hannan and Quinn, 1979), BIC (Akaike, 1978), and SIC (Schwarz, 1978). These criteria are defined as follows:

$$\begin{aligned}
\text{AICc} &= \left(n \ln \hat{\sigma}^2 + n\right) + \frac{2n(p+1)}{n-p-2}, \\
\text{FPE} &= n\left(\frac{n+p}{n-p}\right)\hat{\sigma}^2, \\
\text{HQ} &= n \ln \hat{\sigma}^2 + 2p \ln \ln n, \\
\text{BIC} &= (n-p)\, \ln\left(\frac{n\hat{\sigma}^2}{n-p}\right) + p \ln\left\{\frac{(\sum_{t=1}^{n} y_t^2) - n\hat{\sigma}^2}{p}\right\}, \\
\text{SIC} &= -2 \ln f(Y \,|\, \hat{\theta}_k) + k \ln n.
\end{aligned}$$

Here, $\hat{\sigma}^2$ denotes the maximum likelihood estimator of σ^2.

We remark that for autoregressive modeling, SIC, HQ, and BIC are *consistent* whereas AIC, AICc, and FPE are *asymptotically efficient* in the sense of Shibata (1980). Suppose that the generating model is of a finite order and that this order is represented in the collection of candidate families under consideration. A consistent criterion will asymptotically select the fitted candidate model having the correct order with probability one. On the other hand, suppose that the generating model is of an infinite order and therefore lies outside of the collection of candidate families under consideration. An asymptotically efficient criterion will asymptotically select the fitted candidate model which minimizes the one-step mean squared error of prediction. For autoregressive model selection, we note that KIC is asymptotically efficient within a broad class of generating models; however, AIC, AICc, and FPE are asymptotically efficient within an even larger class. (For details, see Bhansali, 1993, p. 55.)

The generating models featured in our simulation sets are as follows:

$$\begin{aligned}
&(1) \qquad y_t = 0.99 y_{t-1} - 0.80 y_{t-2} + \epsilon_t, \\
&(2) \qquad y_t = 0.10 y_{t-1} + 0.60 y_{t-2} + \epsilon_t, \\
&(3) \qquad y_t = 0.70 y_{t-1} - 0.50 y_{t-2} + 0.60 y_{t-3} + \epsilon_t.
\end{aligned}$$

Here, ϵ_t represents a Gaussian white noise process with mean 0 and variance 1. For each generating model, two simulation sets are considered: one in which the sample size is $n = 40$

and the other in which $n = 60$. For all six sets, the maximum model order used for the candidate collection is $P = 8$.

The results of the six sets are summarized in Table 1. Note that in all six sets, KIC obtains substantially more correct order selections than any of the asymptotically efficient criteria. KIC also consistently outperforms HQ in terms of correct order selections, and outperforms BIC in sets 3 through 6. KIC is generally outperformed by SIC; however, KIC does not exhibit as strong a tendency as SIC to choose underparameterized models. Moreover, it should be noted that this type of simulation study tends to favor consistent criteria such as SIC, since in each set, the generating model is of a finite order, and this order is represented in the collection of candidate families under consideration.

Figure 1 provides some insight as to why KIC tends to outperform AIC as a selection criterion. Consider the second set of simulations, based on generating model (2) with a sample size of $n = 60$. In Figure 1, simulated values of $\Omega(\theta_o, p)$ and $\Delta(\theta_o, p)$ are plotted versus the order p for $p = 1, 2, \ldots, 8$. These values are obtained by averaging $K(\theta_o, \hat{\theta}_k)$ and $d(\theta_o, \hat{\theta}_k)$, respectively, over the 1000 replications. The average values for each of KIC and AIC are also plotted versus p.

Note that the shapes of the $\Omega(\theta_o, p)$ and $\Delta(\theta_o, p)$ curves suggest that $K(\theta_o, \hat{\theta}_k)$ tends to be more effective than $d(\theta_o, \hat{\theta}_k)$ in delineating between fitted models of the correct order and fitted models which are either too small or too large. This illustrates the perspective advanced in Section 3: i.e., as the dissimilarity between a fitted model and the true model becomes more pronounced, $J(\theta_o, \hat{\theta}_k)$ tends to grow to a greater extent than $I(\theta_o, \hat{\theta}_k)$, exhibiting the increasing tendencies in both $I(\theta_o, \hat{\theta}_k)$ and $I(\hat{\theta}_k, \theta_o)$. Thus, a model selection criterion which estimates $J(\theta_o, \hat{\theta}_k)$ (or $K(\theta_o, \hat{\theta}_k)$) may be preferable to one which estimates $I(\theta_o, \hat{\theta}_k)$ (or $d(\theta_o, \hat{\theta}_k)$), provided that the former criterion is accurate enough to adequately reflect the sensitivity of $J(\theta_o, \hat{\theta}_k)$ (or $K(\theta_o, \hat{\theta}_k)$). To this end, note that for $p \geq p_o$, the average KIC and AIC curves each track their respective target curves comparably. This indicates that the criteria each achieve the property of asymptotic unbiasedness to roughly the same degree. However, since KIC targets a more sensitive discrepancy measure than AIC, KIC has a higher success rate in identifying the correct model order.

The results reported in Table 2 provide further support to recommend $J(\theta_o, \theta_k)$ over $I(\theta_o, \theta_k)$ as a basis for the construction of model selection criteria. Here, we consider the use of $J(\theta_o, \hat{\theta}_k)$ and $I(\theta_o, \hat{\theta}_k)$ (or equivalently, $K(\theta_o, \hat{\theta}_k)$ and $d(\theta_o, \hat{\theta}_k)$) as selection criteria in our six simulation sets. The table features the number of correct order selections obtained by each divergence measure in each set.

As previously mentioned, $J(\theta_o, \hat{\theta}_k)$ and $I(\theta_o, \hat{\theta}_k)$ depend on θ_o, and are therefore not accessible in practical applications. Nonetheless, inspecting the performance of these measures as selection rules is instructive, since it may help to indicate whether there is an advantage in targeting one of the measures over the other. Note that in each of the sets, $J(\theta_o, \hat{\theta}_k)$ obtains considerably more correct order selections than $I(\theta_o, \hat{\theta}_k)$.

5. Further Directions

The results in Section 4 suggest that KIC should function as an effective model selection criterion in large-sample applications. The results also suggest that $J(\theta_o, \theta_k)$ may provide a foundation for the development of model selection criteria which is preferable to that provided by $I(\theta_o, \theta_k)$. This motivates the need to further explore the properties of $J(\theta_o, \theta_k)$ as a measure of model disparity, as well as the need to develop small-sample estimators of $K(\theta_o, \hat{\theta}_k)$.

As emphasized at the end of Section 3, in settings where the sample size is small and the candidate collection consists of families which are excessively overparameterized, AIC may exhibit a tendency to choose models which are "overfit." In such instances, the penalty term of AIC provides an insufficient degree of bias correction for the larger fitted models; as a result, AIC tends to grossly underestimate $d(\theta_o, \hat{\theta}_k)$.

Recent work has led to successful small-sample refinements of the penalty term of AIC. One type of refinement is based on assuming a particular modeling framework for the candidate family $\mathcal{F}(k)$, and using the characteristics of that framework to derive either an exact expression or a more precise approximation for the bias adjustment (2.4). This approach was first suggested for linear regression by Sugiura (1978), and later extended and advanced by Hurvich and Tsai (1989, 1993), Hurvich, Shumway, and Tsai (1990), and Bedrick and

Tsai (1994) for nonlinear regression, multivariate regression, autoregressive and autoregressive moving-average modeling, and vector autoregressive modeling. An alternative type of refinement is based on using the bootstrap to approximate the adjustment (2.4): see, for instance, Efron (1983, 1986), Cavanaugh and Shumway (1997), and Shibata (1997).

In small-sample applications where excessively overparameterized families are entertained, KIC tends to underestimate $K(\theta_o, \hat{\theta}_k)$ in the same manner that AIC tends to underestimate $d(\theta_o, \hat{\theta}_k)$. In future work, we hope to use the aforementioned approaches which have led to small-sample refinements of the penalty term of AIC to develop such refinements for the penalty term of KIC.

Acknowledgements

The author wishes to extend his appreciation to the referee for detailed and helpful reviews. The author also extends his thanks to Professors Andrew Neath, Lawrence Ries, and Alex Karagrigoriou for useful insights and suggestions.

Table 1. Criterion Selections.

Set	Model	n	Order	Criterion						
				KIC	AIC	AICc	FPE	HQ	BIC	SIC
1	(1)	40	$< p_o$	1	1	1	0	1	1	3
			$= p_o$	874	707	770	637	762	922	918
			$> p_o$	125	292	229	363	237	77	79
2	(1)	60	$< p_o$	0	0	0	0	0	0	0
			$= p_o$	881	706	740	667	825	951	953
			$> p_o$	119	294	260	333	175	49	47
3	(2)	40	$< p_o$	80	55	59	48	64	43	114
			$= p_o$	808	679	717	603	709	684	817
			$> p_o$	112	266	224	349	227	273	69
4	(2)	60	$< p_o$	16	7	9	6	11	8	26
			$= p_o$	864	698	735	649	816	790	914
			$> p_o$	120	295	256	345	173	202	60
5	(3)	40	$< p_o$	122	57	76	48	80	79	176
			$= p_o$	737	630	676	543	647	664	740
			$> p_o$	141	313	248	409	273	257	84
6	(3)	60	$< p_o$	20	5	6	4	16	21	51
			$= p_o$	831	686	718	620	775	795	881
			$> p_o$	149	309	276	376	209	184	68

Table 2. Correct Order Selections for $J(\theta_o, \hat{\theta}_k)$ and $I(\theta_o, \hat{\theta}_k)$.

Set	Model	n	Divergence	
			$J(\theta_o, \hat{\theta}_k)$	$I(\theta_o, \hat{\theta}_k)$
1	(1)	40	906	865
2	(1)	60	918	882
3	(2)	40	846	806
4	(2)	60	888	835
5	(3)	40	805	762
6	(3)	60	834	785

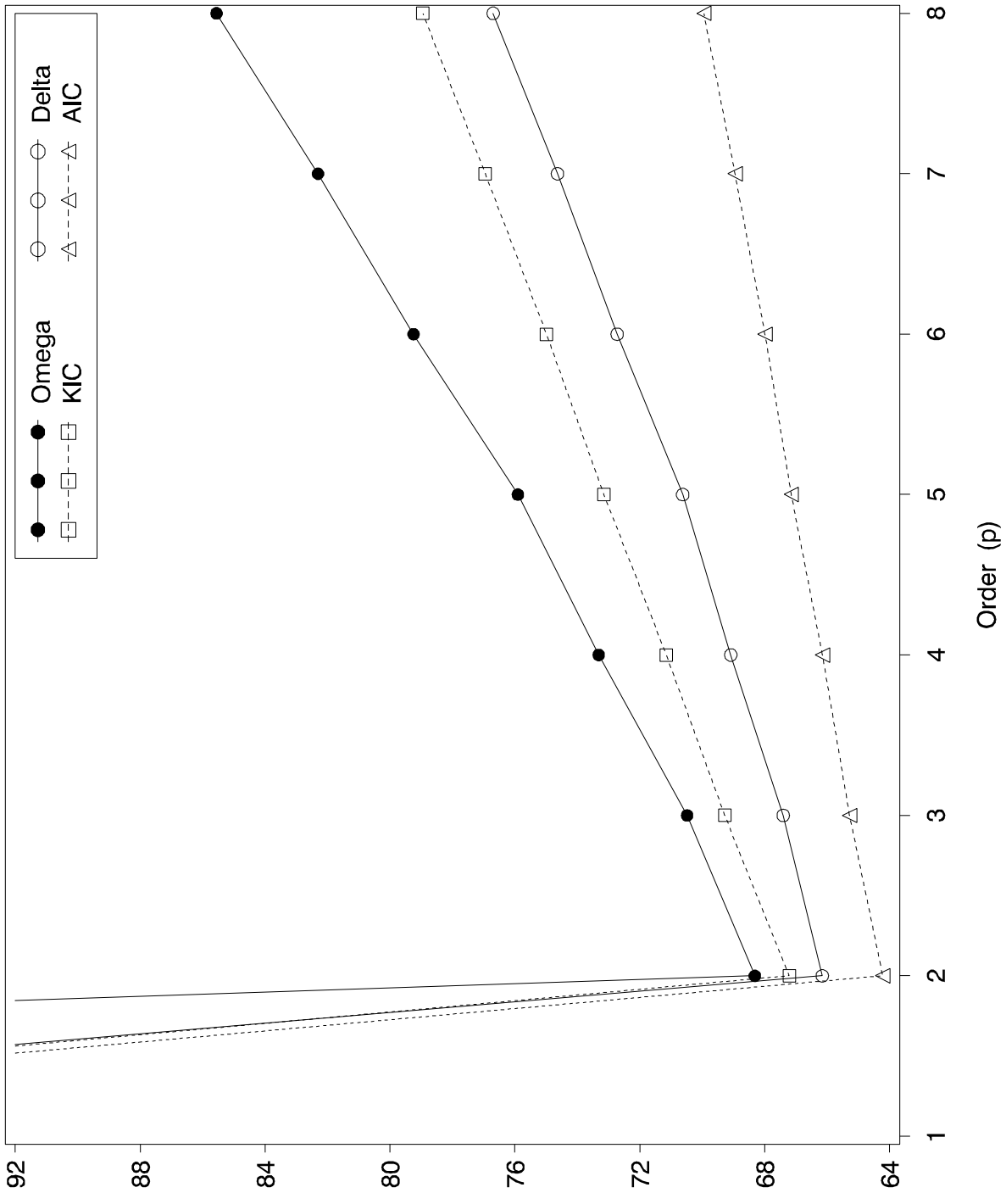

Figure 1. $\Omega(\theta_o, p)$, $\Delta(\theta_o, p)$, and Average Values of KIC, AIC. (Simulation Set 2.)

References

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

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov and F. Csáki, eds., *2nd International Symposium on Information Theory* (Akadémia Kiadó, Budapest) pp. 267–281.

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

Akaike, H. (1978), Time series analysis and control through parametric models, in: D. F. Findley, ed., *Applied Time Series Analysis* (Academic Press, New York) pp. 1–23.

Bedrick, E. J. and Tsai, C. L. (1994), Model selection for multivariate regression in small samples, *Biometrics* **50**, 226–231.

Bhansali, R. J. (1993), Order selection for linear time series models: A review, in: T. S. Rao, ed., *Developments in Time Series Analysis* (Chapman and Hall, London) pp. 50–66.

Cavanaugh, J. E. (1997), Unifying the derivations of the Akaike and corrected Akaike information criteria, *Statistics & Probability Letters* **33**, 201–208.

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

Efron, B. (1983), Estimating the error rate of a prediction rule: Improvement on cross-validation, *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. (1986), How biased is the apparent error rate of a prediction rule?, *Journal of the American Statistical Association* **81**, 461–470.

Efron, B. and Hinkley, D. V. (1978), Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, *Biometrika* **65**, 457–481.

Findley, D. F. (1985), On the unbiasedness property of AIC for exact or approximating linear stochastic time series models, *Journal of Time Series Analysis* **6**, 229–252.

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

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

Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society A* **186**, 453–461.

Jeffreys, H. (1983), *Theory of Probability* (Clarendon Press, Oxford, 3rd ed.).

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

Kullback, S. and Leibler, R. A. (1951), On information and sufficiency, *Annals of Mathematical Statistics* **22**, 76–86.

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

Mallows, C. L. (1973), Some comments on C_p, *Technometrics* **15**, 661–675.

Parzen, E. (1974), Some recent advances in time series modeling, *IEEE Transactions on Automatic Control* **AC-19**, 389–409.

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

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

Shibata, R. (1980), Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Annals of Statistics* **80**, 147–164.

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

Sugiura, N. (1978), Further analysis of the data by Akaike's information criterion and the finite corrections, *Communications in Statistics* **A7**, 13–26.