JSPI-02-43

Model Selection Criteria Based on Kullback Information Measures for Nonlinear Regression

by

Hyun–Joo Kim

Department of Mathematics and Computer Science, Truman State University

Joseph E. Cavanaugh

Department of Biostatistics, University of Iowa

Abstract

In statistical modeling, selecting an optimal model from a class of candidates is a critical issue. During the past three decades, a number of model selection criteria have been proposed based on estimating Kullback's (1968, p. 5) directed divergence between the model generating the data and a fitted candidate model. The Akaike (1973, 1974) information criterion, AIC, was the first of these. AIC is justified in a very general framework, and as a result, offers a crude estimator of the directed divergence: one which exhibits a potentially high degree of negative bias in small-sample applications (Hurvich and Tsai, 1989). The "corrected" Akaike information criterion (Hurvich and Tsai, 1989), AICc, adjusts for this bias, and consequently often outperforms AIC as a selection criterion. However, AICc is less broadly applicable than AIC since its justification depends upon the structure of the candidate model. AIC_I (Hurvich, Shumway, and Tsai, 1990) is an "improved" version of AIC featuring a simulated bias correction.

Recently, model selection criteria have been proposed based on estimating Kullback's (1968, p. 6) symmetric divergence between the generating model and a fitted candidate model (Cavanaugh, 1999, 2004). KIC, KICc, and KIC_I are criteria devised to target the symmetric divergence in the same manner that AIC, AICc, and AIC_I target the directed divergence.

AICc has been justified for the nonlinear regression framework by Hurvich and Tsai (1989). In this paper, we justify KICc for this framework, and propose versions of AIC_I and KIC_I suitable for nonlinear regression applications. We evaluate the selection performance of AIC, AICc, AIC_I, KIC, KICc, and KIC_I in a simulation study. Our results generally indicate that the "improved" criteria outperform the "corrected" criteria, which in turn outperform the non-adjusted criteria. Moreover, the KIC family performs favorably against the AIC family.

Key Words: AIC, Akaike information criterion, I-divergence, J-divergence, Kullback-Leibler information, nonlinear regression.

1. Introduction

In statistical modeling, one of the main objectives is to select a suitable model from a candidate class to characterize the underlying data. Model selection criteria provide a useful tool in this regard. A selection criterion assesses whether a fitted model offers an optimal balance between goodness-of-fit and parsimony. Ideally, a criterion will identify candidate models which are either too simplistic to accommodate the data or unnecessarily complex.

The first model selection criterion to gain widespread acceptance was the Akaike (1973, 1974) information criterion, AIC. AIC is applicable in a broad array of modeling frameworks, since its large-sample justification only requires conventional asymptotic properties of maximum likelihood estimators. However, in settings where the sample size is small, AIC tends to favor inappropriately high dimensional candidate models (Hurvich and Tsai, 1989); this limits its effectiveness as a model selection criterion.

AIC serves as an estimator of Kullback's (1968, p. 5) directed divergence between the generating or "true" model (i.e., the model which presumably gave rise to the data) and a fitted candidate model. The "corrected" AIC, AICc, is an adjusted version of AIC originally proposed for linear regression with normal errors (Sugiura, 1978; Hurvich and Tsai, 1989). For fitted models in the candidate class which are correctly specified or overfit, AIC is asymptotically unbiased and AICc is exactly unbiased as an estimator of its target measure.

In small-sample applications, AICc often dramatically outperforms AIC as a selection criterion. Since the basic form of AICc is similar to that of AIC, the improvement in selection performance comes without an increase in computational cost. However, AICc is less broadly applicable than AIC since its justification relies upon the structure of the candidate model.

Another adjusted variant of AIC is AIC_I, an "improved" version of AIC proposed by Hurvich, Shumway, and Tsai (1990) for Gaussian autoregressive model selection. The derivation of AIC_I proceeds by decomposing the expected directed divergence into two terms. The first term suggests that the empirical log likelihood can be used to form a biased estimator of the directed divergence; 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 which are correctly specified or overfit, the adjustment is asymptotically independent of the true parameters. Thus, for large-sample applications, the adjustment may be approximated via Monte Carlo simulation using arbitrary values for the parameters.

The directed divergence, also known as the Kullback-Leibler (1951) information or the I-divergence, accesses the dissimilarity between two statistical models. It is an asymmetric measure, meaning that an alternative 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. When used to evaluate fitted candidate models, the directed divergence which serves as the basis for AIC is arguably less sensitive than the symmetric divergence towards detecting improperly specified models. This premise has been used to justify the development of a new family of selection criteria (Cavanaugh, 1999, 2004). KIC, KICc, and KIC_I are criteria constructed to target the symmetric divergence in the same manner that AIC, AICc, and AIC_I target the directed divergence. KIC has been justified under the same general conditions as AIC (Cavanaugh, 1999); however, KICc has only been justified for linear regression with normal errors (Cavanaugh, 2004). As with AIC_I, KIC_I must be formulated based upon the structure of the candidate modeling framework.

Hurvich and Tsai (1989) established that AICc serves as an approximately unbiased estimator of Kullback's directed divergence for nonlinear regression candidate models with normal errors. In this paper, we justify KICc in the same framework. We also propose versions of AIC_I and KIC_I suitable for nonlinear regression applications. We evaluate the selection performance of AIC, AICc, AIC_I, KIC, KICc, and KIC_I in a simulation study. Our results generally indicate that the "improved" criteria outperform the "corrected" criteria, which in turn outperform the non-adjusted criteria. Moreover, the KIC family performs favorably against the AIC family.

In Section 2, we propose and discuss the criteria. Our simulation study is presented and summarized in Section 3. The formal justification of KICc for nonlinear regression appears in the Appendix.

2. Selection Criteria Based on Kullback Information Measures

The nonlinear regression model is frequently used in many areas of the physical, chemical, engineering, and biological sciences. The traditional regression model assumes that the mean structure is linear in the model coefficients: i.e., $E[y] = X'\delta$, where y is the response variable, X is a vector of regressor variables, and δ is an unknown parameter vector. However, one often expects a nonlinear relationship between $E[y]$ and X, perhaps because of the theory which supports the underlying phenomenon. Many nonlinear models fall into categories that are designed for certain situations: thus, there are various families of nonlinear models corresponding to specific functional forms of the mean response.

Assume a collection of data Y has been generated according to an unknown parametric density $f(Y|\theta_o)$, one which corresponds to the normal regression model

$$Y = h_o(\delta_o, \mathbf{X}_o) + \epsilon, \qquad \epsilon \sim N(0, \sigma_o^2 I). \tag{2.1}$$

Suppose that the candidate model postulated for the data is of the form

$$Y = h(\delta, \mathbf{X}) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I). \tag{2.2}$$

Here, Y is an $n \times 1$ response vector, δ_o and δ are $p_o \times 1$ and $p \times 1$ parameter vectors, and $\mathbf{X}_o$ and $\mathbf{X}$ are $n \times s_o$ and $n \times s$ design matrices with rows X_{oi} and X_i. The mean vectors are assumed to have the layouts

$$h_o(\delta_o, \mathbf{X}_o) = (g_o(\delta_o, X_{o1}), \cdots, g_o(\delta_o, X_{on}))' \qquad \text{and} \qquad h(\delta, \mathbf{X}) = (g(\delta, X_1), \cdots, g(\delta, X_n)).'$$

To establish asymptotic inferential results, the mean response function g is generally required to be twice continuously differentiable in δ. For the sake of brevity, we will subsequently write $h(\delta, \mathbf{X})$ as $h(\delta)$ and $h_o(\delta_o, \mathbf{X}_o)$ as $h_o(\delta_o)$.

Define the parameter vectors θ_o and θ_k as $\theta_o = (\delta_o', \sigma_o^2)'$ and $\theta_k = (\delta', \sigma^2)'$. The subscript k on θ_k will refer to the dimension of the vector; i.e., $k = p + 1$.

Let $f(Y|\theta_k)$ denote the likelihood under the model (2.2). We will use $\hat{\delta}$ and $\hat{\sigma}^2$ to represent the maximum likelihood estimators (MLEs) of δ and σ^2, which are generally obtained using the Gauss-Newton method or some other iterative procedure. We will let $\hat{\theta}_k$ denote the

MLE for θ_k; i.e., $\hat{\theta}_k = (\hat{\delta}', \hat{\sigma}^2)'$. Accordingly, $f(Y|\hat{\theta}_k)$ will represent the empirical likelihood corresponding to $f(Y|\theta_k)$.

Let $\mathcal{F}(k) = \{f(Y|\theta_k) \mid \theta_k \in \Theta(k)\}$ denote the k-dimensional parametric family of densities corresponding to candidate models (2.2) of a particular size. 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 $f(Y|\theta_o)$. We note that in many applications, some of the families in the candidate collection may have the same dimension and yet be different; e.g., for some families of linear regression models, the design matrices may have the same rank and yet different column spaces. For ease of notation, we do not include an index to delineate between such families.

If $f(Y|\theta_o) \in \mathcal{F}(k)$, and $\mathcal{F}(k)$ is such that no smaller family will contain $f(Y|\theta_o)$, we refer to $f(Y|\hat{\theta}_k)$ as *correctly specified*. If $f(Y|\theta_o) \in \mathcal{F}(k)$, yet $\mathcal{F}(k)$ is such that families smaller than $\mathcal{F}(k)$ also contain $f(Y|\theta_o)$, we refer to $f(Y|\hat{\theta}_k)$ as *overfit*. If $f(Y|\theta_o) \notin \mathcal{F}(k)$, we refer to $f(Y|\hat{\theta}_k)$ as *underfit*.

To determine which of the fitted models $\{f(Y|\hat{\theta}_{k_1}), f(Y|\hat{\theta}_{k_2}), \ldots, f(Y|\hat{\theta}_{k_L})\}$ 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 a candidate model $f(Y|\theta_k)$. Kullback's directed and symmetric divergence both fulfill this objective.

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

$$I(\theta, \theta_*) = \mathrm{E}_\theta \left[\ln \left\{ \frac{f(Y|\theta)}{f(Y|\theta_*)} \right\} \right], \tag{2.3}$$

and Kullback's *symmetric divergence* between $f(Y|\theta)$ and $f(Y|\theta_*)$ is defined as

$$J(\theta, \theta_*) = \mathrm{E}_\theta \left[\ln \left\{ \frac{f(Y|\theta)}{f(Y|\theta_*)} \right\} \right] + \mathrm{E}_{\theta_*} \left[\ln \left\{ \frac{f(Y|\theta_*)}{f(Y|\theta)} \right\} \right]. \tag{2.4}$$

Here, E_θ denotes the expectation under $f(Y|\theta)$. Note that $J(\theta, \theta_*)$ is symmetric in its arguments whereas $I(\theta, \theta_*)$ is not. Thus, an alternate directed divergence, $I(\theta_*, \theta)$, may be obtained by switching the roles of $f(Y|\theta)$ and $f(Y|\theta_*)$ in (2.3). The sum of the two directed divergences yields the symmetric divergence: $J(\theta, \theta_*) = I(\theta, \theta_*) + I(\theta_*, \theta)$.

For the purpose of assessing the proximity between a certain fitted candidate model $f(Y|\hat{\theta}_k)$ and the true model $f(Y|\theta_o)$, we consider the measures

$$I(\theta_o, \hat{\theta}_k) = I(\theta_o, \theta_k)|_{\theta_k=\hat{\theta}_k} \qquad \text{and} \qquad J(\theta_o, \hat{\theta}_k) = J(\theta_o, \theta_k)|_{\theta_k=\hat{\theta}_k}.$$

Of these two, Cavanaugh (1999, 2004) conjectures that $J(\theta_o, \hat{\theta}_k)$ may be preferred, since it combines $I(\theta_o, \hat{\theta}_k)$ with its counterpart $I(\hat{\theta}_k, \theta_o)$, a measure which serves a related yet distinct function. To gauge the disparity between $f(Y|\hat{\theta}_k)$ and $f(Y|\theta_o)$, $I(\theta_o, \hat{\theta}_k)$ assesses how well samples generated under the true model $f(Y|\theta_o)$ conform to the fitted candidate model $f(Y|\hat{\theta}_k)$, whereas $I(\hat{\theta}_k, \theta_o)$ assesses how well samples generated under the fitted candidate model $f(Y|\hat{\theta}_k)$ conform to the true model $f(Y|\theta_o)$. As a result of these contrasting roles, $I(\theta_o, \hat{\theta}_k)$ tends to be more sensitive towards reflecting overfit models, whereas $I(\hat{\theta}_k, \theta_o)$ tends to be more sensitive towards reflecting underfit models. Accordingly, $J(\theta_o, \hat{\theta}_k)$ may be more adept at detecting misspecification than either of its components.

In what follows, we will show how the AIC family of model selection criteria arises through estimating a variant of $I(\theta_o, \hat{\theta}_k)$. We will then show how an alternate family of selection criteria, the KIC family, arises through estimating a variant of $J(\theta_o, \hat{\theta}_k)$.

For two arbitrary parametric densities $f(Y|\theta)$ and $f(Y|\theta_*)$, let

$$d(\theta, \theta_*) = \mathrm{E}_\theta[-2 \ln f(Y|\theta_*)]. \tag{2.5}$$

From (2.3) and (2.5), note that we can write

$$2I(\theta_o, \theta_k) = d(\theta_o, \theta_k) - d(\theta_o, \theta_o). \tag{2.6}$$

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 at hand, $d(\theta_o, \theta_k)$ serves as a valid substitute for $I(\theta_o, \theta_k)$.

Now for a given set of MLEs $\hat{\theta}_k$,

$$d(\theta_o, \hat{\theta}_k) = d(\theta_o, \theta_k)|_{\theta_k=\hat{\theta}_k}$$

would provide a meaningful measure of separation between the true model and a fitted candidate model. Evaluating $d(\theta_o, \hat{\theta}_k)$ is not possible since doing so requires knowledge of

θ_o. However, the work of Akaike (1973, 1974) suggests that $-2\ln f(Y|\,\hat{\theta}_k)$ serves as a biased estimator of $d(\theta_o, \hat{\theta}_k)$, and that in many applications (including those beyond the scope of nonlinear regression models), the bias adjustment

$$\mathrm{B}_1(k, \theta_o) = \mathrm{E}_{\theta_o}[d(\theta_o, \hat{\theta}_k)] - \mathrm{E}_{\theta_o}[-2\ln f(Y|\,\hat{\theta}_k)] \tag{2.7}$$

can be asymptotically estimated by twice the dimension of $\hat{\theta}_k$. Specifically, if we assume that $\hat{\theta}_k$ satisfies the conventional large-sample properties of MLEs, and that $f(Y|\,\hat{\theta}_k)$ is either correctly specified or overfit ($f(Y|\,\theta_o) \in \mathcal{F}(k)$), it can be shown that

$$\mathrm{B}_1(k, \theta_o) \simeq 2k. \tag{2.8}$$

(See, for instance, Cavanaugh, 1997, p. 204.) With this motivation, we define the criterion

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

As the sample size increases, the difference between the expected value of AIC and the expected value of $d(\theta_o, \hat{\theta}_k)$ should tend to zero. Accordingly, if we define

$$\begin{aligned} \Delta(k, \theta_o) &= \mathrm{E}_{\theta_o}[d(\theta_o, \hat{\theta}_k)] \\ &= \mathrm{E}_{\theta_o}[-2\ln f(Y|\,\hat{\theta}_k)] + \mathrm{B}_1(k, \theta_o), \end{aligned} \tag{2.9}$$

we may regard AIC as an asymptotically unbiased estimator of $\Delta(k, \theta_o)$.

When n is large and k is comparatively small, the degree of bias incurred in estimating $\Delta(k, \theta_o)$ with AIC is negligible. However, when n is small and k is relatively large (e.g., $k \simeq n/2$), $2k$ is often much smaller than $\mathrm{B}_1(k, \theta_o)$, making AIC substantially negatively biased as an estimator of $\Delta(k, \theta_o)$. If AIC severely underestimates $\Delta(k, \theta_o)$ for high dimensional fitted models in the candidate class, the criterion may favor these models even though they may correspond to large values of $d(\theta_o, \hat{\theta}_k)$ (Hurvich and Tsai, 1989).

AICc and AIC_I were proposed to serve as estimators of $\Delta(k, \theta_o)$ which are less biased in small-sample applications than traditional AIC (Hurvich and Tsai, 1989; Hurvich, Shumway, and Tsai, 1990). However, since the justification of AICc and the computation of AIC_I are contingent upon the structure of the candidate modeling framework, these criteria are less generally applicable than AIC.

AICc was originally proposed by Sugiura (1978) in the setting of linear regression models with normal errors. In this framework, the bias adjustment (2.7) can be evaluated exactly for correctly specified and overfit models. Where p represents the rank of the design matrix for the candidate model, it can be shown that when $f(Y|\theta_o) \in \mathcal{F}(k)$,

$$\mathrm{B}_1(k, \theta_o) = \frac{2n(p+1)}{(n-p-2)}. \tag{2.10}$$

(See Cavanaugh, 1997, pp. 204–205.) Thus, an exactly unbiased estimator of $\Delta(k, \theta_o)$ is given by

$$\mathrm{AICc} = -2 \ln f(Y|\hat{\theta}_k) + \frac{2n(p+1)}{(n-p-2)}.$$

Although relation (2.10) does not hold precisely in the normal nonlinear regression framework, the arguments and results of Hurvich and Tsai (1989) suggest that $\mathrm{B}_1(k, \theta_o)$ is well approximated by $\{2n(p+1)\}/(n-p-2)$ even for relatively small n.

AIC_I was originally proposed for Gaussian autoregressive models by Hurvich, Shumway, and Tsai (1990). In this framework, the relation (2.10) only holds approximately. However, when $f(Y|\theta_o) \in \mathcal{F}(k)$, the bias adjustment $\mathrm{B}_1(k, \theta_o)$ is asymptotically independent of the true model parameters θ_o. Thus, for large n, $\mathrm{B}_1(k, \theta_o)$ may be approximated via Monte Carlo simulation after setting θ_o equal to a conveniently chosen vector.

To propose an AIC_I for the normal nonlinear regression framework, we utilize the following results:

$$\begin{aligned} -2 \ln f(Y|\theta_k) &= n \ln \sigma^2 + \frac{\{Y - h(\delta)\}'\{Y - h(\delta)\}}{\sigma^2}, \\ -2 \ln f(Y|\hat{\theta}_k) &= n(\ln \hat{\sigma}^2 + 1), \qquad (2.11) \\ \mathrm{E}_{\theta_o}[d(\theta_o, \hat{\theta}_k)] &= \mathrm{E}_{\theta_o}\left[n \ln \hat{\sigma}^2 + \frac{n\sigma_o^2}{\hat{\sigma}^2} + \frac{\{h_o(\delta_o) - h(\hat{\delta})\}'\{h_o(\delta_o) - h(\hat{\delta})\}}{\hat{\sigma}^2} \right]. \qquad (2.12) \end{aligned}$$

(In the preceding relations and throughout our development, we have neglected the additive constant $n \ln 2\pi$.) Note that by using (2.11) and (2.12) in conjunction with (2.7) and (2.9), we may write

$$\begin{aligned}\Delta(k,\theta_o) &= \mathrm{E}_{\theta_o}[d(\theta_o,\hat{\theta}_k)] \\ &= \mathrm{E}_{\theta_o}[-2\ln f(Y|\,\hat{\theta}_k)] \\ &\quad +\mathrm{E}_{\theta_o}\left[\frac{n\sigma_o^2}{\hat{\sigma}^2}+\frac{\{h_o(\delta_o)-h(\hat{\delta})\}'\{h_o(\delta_o)-h(\hat{\delta})\}}{\hat{\sigma}^2}-n\right]. \end{aligned} \tag{2.13}$$

The form of AIC_I is suggested by (2.13). To evaluate AIC_I, we set the parameters σ_o^2 and δ_o at conveniently chosen values, generate R samples according to model (2.1), solve for the R sets of corresponding MLEs $\{(\hat{\sigma}^2(1),\hat{\delta}(1)),\ldots,(\hat{\sigma}^2(R),\hat{\delta}(R))\}$ under model (2.2), and compute the criterion via

$$\mathrm{AIC}_I = -2\ln f(Y|\,\hat{\theta}_k) + \frac{1}{R}\sum_{j=1}^{R}\left[\frac{n\sigma_o^2}{\hat{\sigma}^2(j)}+\frac{\{h_o(\delta_o)-h(\hat{\delta}(j))\}'\{h_o(\delta_o)-h(\hat{\delta}(j))\}}{\hat{\sigma}^2(j)}-n\right].$$

In frameworks where it is not possible to evaluate $\mathrm{B}_1(k,\theta_o)$ exactly, AIC_I may estimate $\Delta(k,\theta_o)$ with less bias than AICc, and may outperform AICc as a selection criterion (Hurvich, Shumway, and Tsai, 1990).

Next, we propose selection criteria devised to target $J(\theta_o,\hat{\theta}_k)$ in the same manner that AIC, AICc, and AIC_I target $I(\theta_o,\hat{\theta}_k)$.

Similar to (2.6), using (2.4) and (2.5), we can write

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

Discarding the constant $d(\theta_o,\theta_o)$ from the preceding yields

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

For the purpose of discriminating among various candidate models, $K(\theta_o,\theta_k)$ is equivalent to $J(\theta_o,\theta_k)$. 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 consider estimating

$$K(\theta_o,\hat{\theta}_k) = K(\theta_o,\theta_k)|_{\theta_k=\hat{\theta}_k}.$$

If $-2\ln f(Y|\hat{\theta}_k)$ is regarded as a platform for an estimator of this measure, the challenge is then to correct for the bias. The bias adjustment may be expressed as

$$\begin{aligned} \mathrm{E}_{\theta_o}[K(\theta_o,\hat{\theta}_k)] - \mathrm{E}_{\theta_o}[-2\ln f(Y|\hat{\theta}_k)] &= \mathrm{E}_{\theta_o}[d(\theta_o,\hat{\theta}_k)] - \mathrm{E}_{\theta_o}[-2\ln f(Y|\hat{\theta}_k)] \quad (2.14) \\ &\quad + \mathrm{E}_{\theta_o}[d(\hat{\theta}_k,\theta_o)] - \mathrm{E}_{\theta_o}[d(\hat{\theta}_k,\hat{\theta}_k)]. \quad (2.15) \end{aligned}$$

Note that the difference on the right-hand side of (2.14) is the same as $\mathrm{B}_1(k,\theta_o)$, the bias adjustment for $d(\theta_o,\hat{\theta}_k)$ expressed in (2.7). For the difference (2.15), define

$$\mathrm{B}_2(k,\theta_o) = \mathrm{E}_{\theta_o}[d(\hat{\theta}_k,\theta_o)] - \mathrm{E}_{\theta_o}[d(\hat{\theta}_k,\hat{\theta}_k)]. \quad (2.16)$$

The penalty terms of AIC, AICc, and AIC_I provide us with estimators of $\mathrm{B}_1(k,\theta_o)$; our goal is to seek similar estimators of $\mathrm{B}_2(k,\theta_o)$. This will lead us to a set of criteria which are analogous to AIC, AICc, and AIC_I, targeting $K(\theta_o,\hat{\theta}_k)$ in the same way that the AIC-type criteria target $d(\theta_o,\hat{\theta}_k)$.

First, we propose an analogue of AIC based on estimating $\mathrm{B}_2(k,\theta_o)$ in the same manner that the penalty term of AIC estimates $\mathrm{B}_1(k,\theta_o)$. If we assume that $f(Y|\hat{\theta}_k)$ is either correctly specified or overfit ($f(Y|\theta_o) \in \mathcal{F}(k)$), it can be shown that for large n,

$$\mathrm{B}_2(k,\theta_o) \simeq k. \quad (2.17)$$

(See Cavanaugh, 1999, pp. 337–338.) As with (2.8), the preceding applies to any modeling framework in which $\hat{\theta}_k$ satisfies the conventional properties of MLEs. Motivated by the large-sample approximations (2.8) and (2.17), we define the criterion

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

As the sample size increases, the difference between the expected value of KIC and the expected value of $K(\theta_o,\hat{\theta}_k)$ should tend to zero. Accordingly, if we define

$$\begin{aligned} \Omega(k,\theta_o) &= \mathrm{E}_{\theta_o}[K(\theta_o,\hat{\theta}_k)] \\ &= \mathrm{E}_{\theta_o}[-2\ln f(Y|\hat{\theta}_k)] + \mathrm{B}_1(k,\theta_o) + \mathrm{B}_2(k,\theta_o), \quad (2.18) \end{aligned}$$

we may regard KIC as an asymptotically unbiased estimator of $\Omega(k,\theta_o)$.

Cavanaugh (2004) proposed an analogue of AICc for normal linear regression models based on estimating $\mathrm{B}_2(k, \theta_o)$ in the same manner that the penalty term of AICc estimates $\mathrm{B}_1(k, \theta_o)$. In the normal linear regression framework, the bias adjustment (2.16) can be evaluated exactly for correctly specified and overfit models. When $f(Y \mid \theta_o) \in \mathcal{F}(k)$, it can be shown that

$$\mathrm{B}_2(k, \theta_o) = n \ln\left(\frac{n}{2}\right) - n\psi\left(\frac{n-p}{2}\right), \tag{2.19}$$

where $\psi(\cdot)$ denotes the *psi* or *digamma* function. Although $\psi(\cdot)$ does not have a closed form representation, an accurate substitute for (2.19) is suggested by the large-sample approximation

$$\left\{n \ln\left(\frac{n}{2}\right) - n\psi\left(\frac{n-p}{2}\right)\right\} \simeq n \ln\left(\frac{n}{n-p}\right) + \frac{n}{n-p}. \tag{2.20}$$

(See Kotz and Johnson, 1982, p. 373.) Based on (2.18), (2.10), (2.19), and (2.20), we define the criterion

$$\mathrm{KICc} = -2 \ln f(Y \mid \hat{\theta}_k) + n \ln\left(\frac{n}{n-p}\right) + \frac{n\{(n-p)(2p+3)-2\}}{(n-p-2)(n-p)}.$$

For the normal nonlinear regression framework, the justification of KICc as an approximately unbiased estimator of $K(\theta_o, \hat{\theta}_k)$ is provided in the Appendix.

Finally, we introduce KIC_I as an analogue of AIC_I based on augmenting AIC_I with a simulated approximation to $\mathrm{B}_2(k, \theta_o)$. We utilize the following results:

$$\mathrm{E}_{\theta_o}[d(\hat{\theta}_k, \hat{\theta}_k)] = \mathrm{E}_{\theta_o}[n(\ln \hat{\sigma}^2 + 1)], \tag{2.21}$$

$$\mathrm{E}_{\theta_o}[d(\hat{\theta}_k, \theta_o)] = \mathrm{E}_{\theta_o}\left[n \ln \sigma_o^2 + \frac{n\hat{\sigma}^2}{\sigma_o^2} + \frac{\{h(\hat{\delta}) - h_o(\delta_o)\}'\{h(\hat{\delta}) - h_o(\delta_o)\}}{\sigma_o^2}\right]. \tag{2.22}$$

Note that by using (2.11), (2.12), (2.21), and (2.22) in conjunction with (2.7), (2.16), and (2.18), we may write

$$\begin{aligned}\Omega(k, \theta_o) &= \mathrm{E}_{\theta_o}[K(\theta_o, \hat{\theta}_k)] \\ &= \mathrm{E}_{\theta_o}[-2 \ln f(Y \mid \hat{\theta}_k)] \\ &\quad + \mathrm{E}_{\theta_o}\left[n \ln\left(\frac{\sigma_o^2}{\hat{\sigma}^2}\right) + \frac{n\sigma_o^2}{\hat{\sigma}^2} + \frac{n\hat{\sigma}^2}{\sigma_o^2}\right. \\ &\qquad \left. + \left(\frac{1}{\hat{\sigma}^2} + \frac{1}{\sigma_o^2}\right)\{h_o(\delta_o) - h(\hat{\delta})\}'\{h_o(\delta_o) - h(\hat{\delta})\} - 2n\right]. \end{aligned} \tag{2.23}$$

The form of KIC_I is suggested by (2.23). To evaluate KIC_I, we set the parameters σ_o^2 and δ_o at conveniently chosen values, generate R samples according to model (2.1), solve for the R sets of corresponding MLEs $\{(\hat{\sigma}^2(1), \hat{\delta}(1)), \ldots, (\hat{\sigma}^2(R), \hat{\delta}(R))\}$ under model (2.2), and compute the criterion via

$$\begin{aligned} \text{KIC}_I &= -2\ln f(Y|\hat{\theta}_k) \\ &+\frac{1}{R}\sum_{j=1}^{R}\Bigg[n\ln\left\{\frac{\sigma_o^2}{\hat{\sigma}^2(j)}\right\} + \frac{n\sigma_o^2}{\hat{\sigma}^2(j)} + \frac{n\hat{\sigma}^2(j)}{\sigma_o^2} \\ &\qquad + \left\{\frac{1}{\hat{\sigma}^2(j)} + \frac{1}{\sigma_o^2}\right\}\{h_o(\delta_o) - h(\hat{\delta}(j))\}'\{h_o(\delta_o) - h(\hat{\delta}(j))\} - 2n\Bigg]. \end{aligned}$$

Having now presented AIC, AICc, AIC_I, KIC, KICc, and KIC_I as selection criteria for nonlinear regression applications, we evaluate the selection performance of these criteria in a simulation study.

3. Simulations

Simulation Sets Based on Nested Candidate Models

Consider a setting where a sample of size n is generated from an exponential regression model of the form (2.1) with

$$h_o(\delta_o, \mathbf{X}_o) = (\alpha_o \exp(X'_{o1}\beta_o), \cdots, \alpha_o \exp(X'_{on}\beta_o))'; \tag{3.1}$$

i.e., $g_o(\delta_o, X_{oi}) = \alpha_o \exp(X'_{oi}\beta_o)$ and $\delta_o = (\alpha_o, \beta'_o)'$. Here, $\mathbf{X}_o$ is an $n \times s_o$ covariate matrix of rank s_o with rows X_{oi}, α_o is a scale parameter, and β_o is an $s_o \times 1$ regression parameter vector. Suppose our objective is to search among a candidate collection of nested families for the fitted model which serves as the best approximation to (2.1).

Assume our candidate models are of the form (2.2) with an exponential response function:

$$h(\delta, \mathbf{X}) = (\alpha \exp(X'_1\beta), \cdots, \alpha \exp(X'_n\beta))'; \tag{3.2}$$

i.e., $g(\delta, X_i) = \alpha \exp(X'_i\beta)$ and $\delta = (\alpha, \beta')'$. Here, $\mathbf{X}$ is an $n \times s$ covariate matrix of rank s with rows X_i, α is a scale parameter, and β is an $s \times 1$ regression parameter vector. Let $\hat{\alpha}$ and $\hat{\beta}$ denote the MLEs of α and β, and let $\hat{\delta} = (\hat{\alpha}, \hat{\beta}')'$ denote the MLE of δ.

In fitting candidate models to the data, we will consider nested design matrices $\mathbf{X}$ of ranks $s = 1, 2, \ldots, S$. We will assume that the design matrix of rank s_o $(1 < s_o < S)$ is correctly specified. Hence, fitted models for which $1 \leq s < s_o$ are underfit, and those for which $s_o < s \leq S$ are overfit. We will refer to s as the *order* of the model and to s_o as the *true order*. The dimension of the model is given by $k = s + 2$, and the size of the parameter vector δ is $p = s + 1$.

We examine the behavior of AIC, AICc, AIC_I, KIC, KICc, and KIC_I as order selection criteria. For the simulated bias adjustments of AIC_I and KIC_I, recall that the true model parameters may be set to convenient values. The values chosen for the parameters are $\alpha_o = 1$, $\beta_o = (0, 0, \ldots, 0)'$, and $\sigma_o^2 = 1$. With these specifications, (2.1) and (3.1) imply that the response vector Y consists of independent, identically distributed standard normal variates. Additionally, AIC_I and KIC_I may be defined as follows:

$$\begin{aligned}
\text{AIC}_I &= n(\ln \hat{\sigma}^2 + 1) \\
&\quad + \frac{1}{R}\sum_{j=1}^{R}\left[\frac{n}{\hat{\sigma}^2(j)} + \frac{\{h(\hat{\delta}(j))'h(\hat{\delta}(j))\}}{\hat{\sigma}^2(j)} - n\right], \\
\text{KIC}_I &= n(\ln \hat{\sigma}^2 + 1) \\
&\quad + \frac{1}{R}\sum_{j=1}^{R}\Bigg[-n\ln \hat{\sigma}^2(j) + \frac{n}{\hat{\sigma}^2(j)} + n\hat{\sigma}^2(j) \\
&\qquad + \left\{\frac{1}{\hat{\sigma}^2(j)} + 1\right\}\left\{h(\hat{\delta}(j))'h(\hat{\delta}(j))\right\} - 2n\Bigg].
\end{aligned}$$

(Two R routines are available from the authors which will provide the simulated bias adjustments for AIC_I and KIC_I for nonlinear regression applications.)

In the initial six simulation sets, 1000 samples are generated from a true model where $\alpha_o = 1$ and $\beta_o = (1, 1, \ldots, 1)'$. In the first three of these sets, $s_o = 3$ and $\sigma_o^2 = 1$. Thus, in scalar form, the true model can be written as

$$y_i = \exp(x_{1i} + x_{2i} + x_{3i}) + e_i, \qquad e_i \sim iid\ N(0, 1). \tag{3.3}$$

In the next three sets, $s_o = 5$ and $\sigma_o^2 = 4$, meaning that the scalar representation of the true model is

$$y_i = \exp(x_{1i} + x_{2i} + x_{3i} + x_{4i} + x_{5i}) + e_i, \qquad e_i \sim iid\ N(0, 4). \tag{3.4}$$

The regressors for all models are produced using a Uniform$(-1, 1)$ distribution. For each of the true models, three different sample sizes n are employed: 50, 75, and 100. For the sets with $n = 50$ and $n = 75$, candidate models of orders 1 through 7 are considered; for the sets with $n = 100$, orders 1 through 10 are entertained. The simulated bias adjustments for AIC_I and KIC_I are based on $R = 200$ replications. For every sample in a set, the fitted model favored by each criterion is recorded. Over the 1000 samples, the order selections are tabulated and summarized.

The order selection results for the three sets corresponding to model (3.3) are featured in Table 1; those corresponding to model (3.4) appear in Table 2. Note that each J-divergence criterion obtains more correct selections than its I-divergence counterpart. Also, within each family of criteria, the "improved" criterion outperforms the "corrected" criterion, which in turn outperforms the non-adjusted criterion. In each of the sets, KIC_I obtains the most correct selections, followed by KICc. KIC ranks third in every set except the fourth.

Figures 1 and 2 help to explain the order selection behaviors of the criteria. These figures are based on the results of simulation set 1 in Table 1. In Figure 1, the criterion averages for the AIC family and the simulated expected discrepancy $\Delta(k, \theta_o)$ are plotted versus the model order; in Figure 2, the criterion averages for the KIC family and simulated expected discrepancy $\Omega(k, \theta_o)$ are plotted versus the model order. The following conclusions can be drawn.

- Past the true model order, the curves for $\Omega(k, \theta_o)$ and the KIC family (Figure 2) exhibit more extreme slopes than the corresponding curves for $\Delta(k, \theta_o)$ and the AIC family (Figure 1). As a result, the J-divergence criteria are less likely than their I-divergence counterparts to choose overfit models.

- AIC and KIC tend to underestimate $\Delta(k, \theta_o)$ and $\Omega(k, \theta_o)$ for overfit models. As a result, these criteria often select models of an inappropriately high order.

- AICc and KICc also tend to underestimate $\Delta(k, \theta_o)$ and $\Omega(k, \theta_o)$ for overfit models, although the degree of underestimation is much less than that exhibited by AIC and KIC.

- The AIC_I and KIC_I curves track the $\Delta(k, \theta_o)$ and $\Omega(k, \theta_o)$ curves very closely. Thus, the "improved" criteria tend to estimate the expected discrepancies with the least amount of bias.

In addition to evaluating the criteria on the basis of order selections, it is also of interest to investigate whether the criteria choose the fitted model which most accurately predicts the response. We define the mean squared error of prediction (MSEP) as

$$\text{MSEP} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \text{E}_{\theta_o}[y_i])^2$$

(cf. Myers, 1990, p. 180). For every sample in a set, MSEP is computed for each of the fitted candidate models. Over the 1000 samples, the number of times that each criterion chooses the fitted model corresponding to the smallest MSEP is recorded. The average MSEP for the models selected by each criterion is also recorded.

The MSEP results for the six sets are featured in Tables 1 and 2. Within each family of criteria, the "improved" criterion produces the most minimum MSEP selections, followed respectively by the "corrected" criterion and the non-adjusted criterion. Also, each J-divergence criterion obtains more minimum MSEP selections than its I-divergence counterpart.

The results for the average MSEP of the selected models are less straightforward to characterize. Sets 1, 2, 3, and 6 exhibit the previous pattern. Within each family of criteria, the "improved" criterion yields the smallest average MSEP, followed respectively by the "corrected" criterion and the non-adjusted criterion. Also, each J-divergence criterion produces a smaller average MSEP than its I-divergence counterpart. This pattern does not hold in sets 4 and 5, however. In these sets, the KIC family does not perform as favorably relative to the AIC family.

For the sets based on model (3.4), the configuration is more conducive to underfitting than for the sets based on model (3.3). Although the propensity of the criteria to choose an underfit model is attenuated as the sample size is increased, the sample sizes in sets 4 and 5 are small enough to result in under-specified selections. These selections often correspond to large values of MSEP, which tend to inflate the average MSEP.

Simulation Sets Where the True Model is Vague

Consider the same simulation setting as that described in the previous subsection. In the four sets to follow, however, we define each true model so there is no unambiguous optimal order for the class of fitted candidate models.

In each set, 1000 samples of size 50 are generated from a true model where $\alpha_o = 1$ and $\sigma_o^2 = 1$. The β_o vectors considered are $\beta_o = (1, 1, 1, 0.05, 0.01)'$, $\beta_o = (1, 0.9, 0.7, 0.1, 0.05)'$, $\beta_o = (1, 0.5, 0.2, 0.1, 0.05)'$, and $\beta_o = (1, 0.7, 0.45, 0.25, 0.1, 0.05, 0.01)'$. Thus, in scalar form, the true models can be written as follows:

$$y_i = \exp(x_{1i} + x_{2i} + x_{3i} + 0.05\, x_{4i} + 0.01\, x_{5i}) + e_i, \tag{3.5}$$

$$y_i = \exp(x_{1i} + 0.9\, x_{2i} + 0.7\, x_{3i} + 0.1\, x_{4i} + 0.05\, x_{5i}) + e_i, \tag{3.6}$$

$$y_i = \exp(x_{1i} + 0.5\, x_{2i} + 0.2\, x_{3i} + 0.1\, x_{4i} + 0.05\, x_{5i}) + e_i, \tag{3.7}$$

$$y_i = \exp(x_{1i} + 0.7\, x_{2i} + 0.45\, x_{3i} + 0.25\, x_{4i} + 0.1\, x_{5i} + 0.05\, x_{6i} + 0.01\, x_{7i}) + e_i, \tag{3.8}$$

with $e_i \sim iid\ N(0, 1)$.

Candidate models of orders 1 to 7 are fit to the data. However, a clearly-defined optimal order does not exist for the fitted models in the candidate class. Consider, for instance, the generating model (3.5). Although the order of this model is 5, it is questionable whether the inclusion of the fourth or fifth regressor justifies the cost of estimating the corresponding parameter. Thus, whether the optimal model order is 3, 4, or 5 is uncertain.

We refer to models (3.5) through (3.8) as *vague*. The parameter values are configured so that each consecutive model is more vague than its predecessor. For these models, it is pointless to investigate how often the criteria choose the fitted model with the same order as the true model. However, it is still meaningful to explore whether the criteria tend to choose a fitted model which accurate predicts the response.

As with the simulation results compiled for the previous subsection, for every sample in a set, MSEP is computed for each of the fitted candidate models. Over the 1000 samples, the number of times that each criterion chooses the fitted model corresponding to the smallest MSEP is recorded. The average MSEP for the models selected by each criterion is also recorded. The results for the four sets are featured in Table 3.

For the first two sets (based on models (3.5) and (3.6)), the results for the minimum MSEP selections are consistent with those reported in Tables 1 and 2. Moreover, the results for the average MSEP are congruous with those reported in Table 1 and in set 6 of Table 2. However, as the generating model becomes increasingly vague, the AIC family of criteria begins to marginally outperform the KIC family. In the last two sets (based on models (3.7) and (3.8)), all of the criteria exhibit difficulty in identifying the fitted model corresponding to the minimum MSEP. However, the average MSEPs, which are similar across all criteria, are comparable to the average MSEPs reported in the first two sets, as well as in set 1 of Table 1 (where the same values of n and σ_o^2 are employed). Thus, although no criterion consistently chooses the fitted model corresponding to the minimum MSEP, no criterion persistently selects a fitted model with a large MSEP.

Conclusions

An extensive collection of simulation sets not featured here reflect selection patterns similar to those in the sets reported. In most settings, the "improved" criteria outperform the "corrected" criteria, which in turn outperform the non-adjusted criteria. Moreover, the KIC family performs favorably against the AIC family.

Our results support two conclusions regarding model selection criteria based on Kullback information measures. First, the performance of a criterion appears to be largely dictated by how well its penalty term approximates the corresponding bias adjustment. Second, for the purpose of delineating between correctly specified and misspecified models, the symmetric divergence may serve as a more sensitive discrepancy measure than the directed divergence.

Acknowledgements

We wish to extend our sincere appreciation to the associate editor and to two referees for carefully reading the original version of this manuscript, and for preparing helpful and constructive critiques which served to greatly improve the exposition and content. We also extend our thanks to the editor, Professor Subir Ghosh.

Table 1: Order Selections, Minimum MSEP Selections, and Average MSEP for Selections Generating Model (3.3)

Set	S / s_o / n	Selections	Criterion					
			AIC	AICc	AIC_I	KIC	KICc	KIC_I
	7	Underfit	1	1	2	2	2	3
1	3	Correctly Specified	676	811	814	848	908	911
	50	Overfit	323	188	184	150	90	86
		Minimum MSEP	616	741	744	778	838	840
		Average MSEP	0.0605	0.0526	0.0517	0.0508	0.0476	0.0469
	7	Underfit	0	0	0	0	0	0
2	3	Correctly Specified	724	788	838	860	904	938
	75	Overfit	276	212	162	140	96	62
		Minimum MSEP	676	739	789	810	853	887
		Average MSEP	0.0385	0.0362	0.0345	0.0338	0.0319	0.0306
	10	Underfit	0	0	0	0	0	0
3	3	Correctly Specified	694	768	792	857	889	912
	100	Overfit	306	232	208	143	111	88
		Minimum MSEP	650	724	747	811	843	866
		Average MSEP	0.0308	0.0278	0.0271	0.0249	0.0237	0.0232

Table 2: Order Selections, Minimum MSEP Selections, and Average MSEP for Selections Generating Model (3.4)

Set	S s_o n	Selections	Criterion					
			AIC	AICc	AIC_I	KIC	KICc	KIC_I
4	7	Underfit	13	24	26	27	51	62
	5	Correctly Specified	696	808	824	818	868	873
	50	Overfit	291	168	150	155	81	65
		Minimum MSEP	577	683	699	693	745	752
		Average MSEP	0.3558	0.3420	0.3393	0.3515	0.3788	0.3809
5	7	Underfit	1	2	2	2	3	7
	5	Correctly Specified	736	806	824	857	910	917
	75	Overfit	263	192	174	141	87	76
		Minimum MSEP	650	717	733	763	815	822
		Average MSEP	0.1977	0.1919	0.1896	0.1873	0.1817	0.1878
6	10	Underfit	0	0	0	0	0	0
	5	Correctly Specified	695	786	796	855	893	897
	100	Overfit	305	214	204	145	107	103
		Minimum MSEP	623	711	721	779	815	819
		Average MSEP	0.1643	0.1518	0.1487	0.1419	0.1357	0.1348

Table 3: Minimum MSEP Selections and Average MSEP for Selections

Generating Model	Selections	Criterion					
		AIC	AICc	AIC_I	KIC	KICc	KIC_I
(3.5)	Minimum MSEP	381	484	487	523	586	587
	Average MSEP	0.0616	0.0545	0.0537	0.0526	0.0481	0.0476
(3.6)	Minimum MSEP	232	278	278	311	353	353
	Average MSEP	0.0647	0.0596	0.0588	0.0573	0.0542	0.0540
(3.7)	Minimum MSEP	147	183	168	168	166	169
	Average MSEP	0.0664	0.0606	0.0586	0.0605	0.0604	0.0578
(3.8)	Minimum MSEP	181	200	208	182	166	176
	Average MSEP	0.0746	0.0726	0.0732	0.0752	0.0775	0.0798

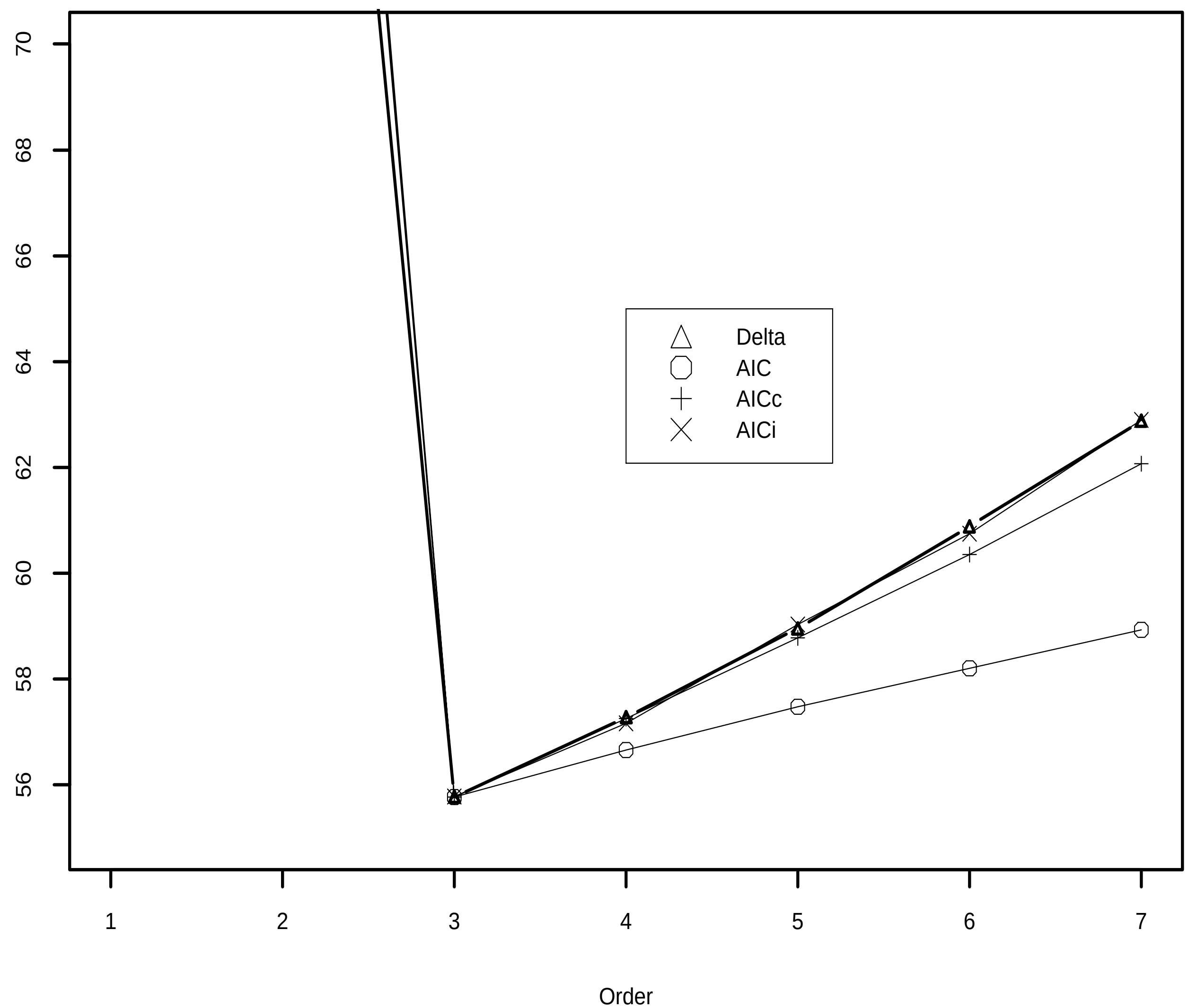

Figure 1: Criterion Averages and Simulated $\Delta(k, \theta_o)$ (Table 1, Set 1)

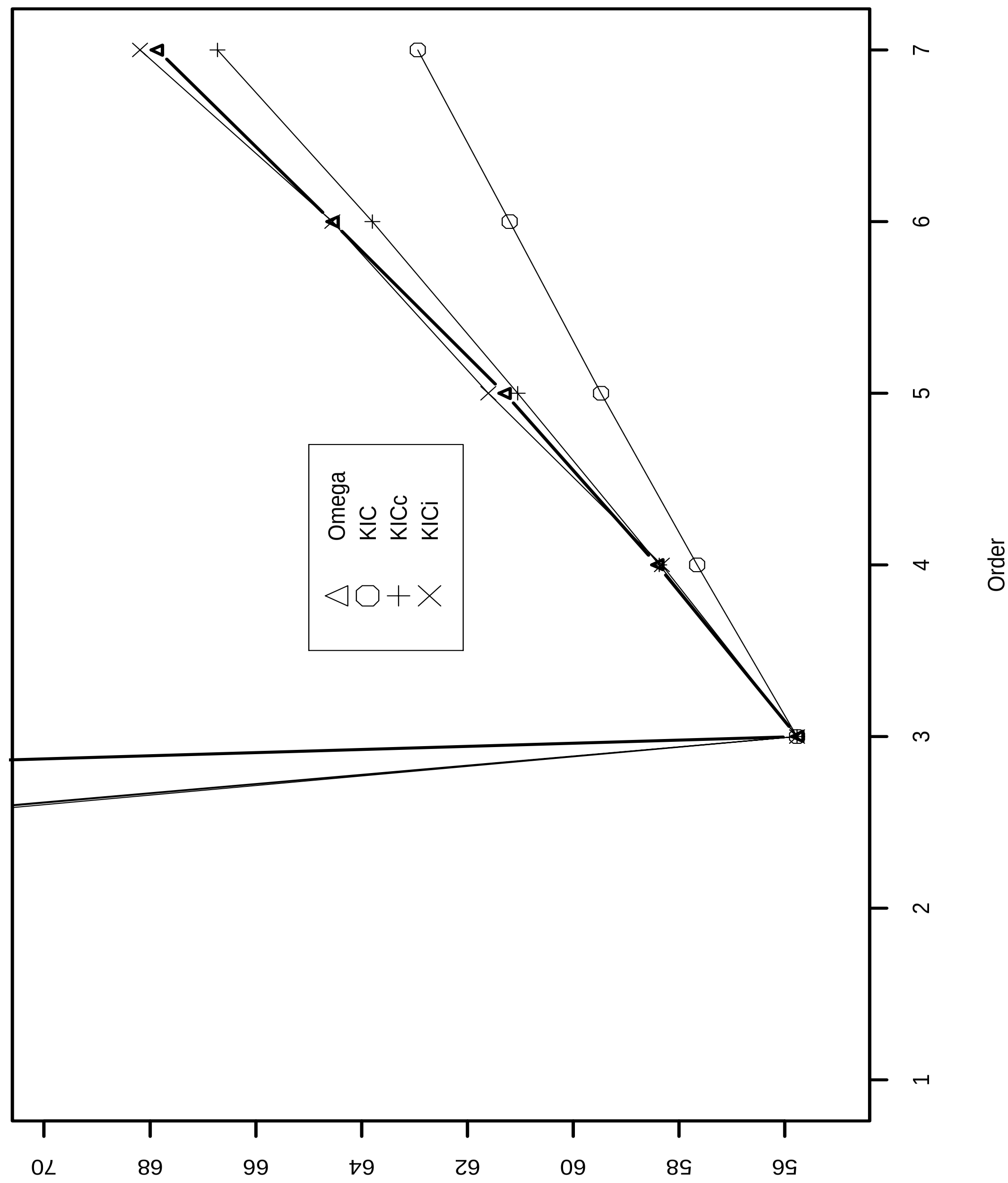

Figure 2: Criterion Averages and Simulated $\Omega(k, \theta_o)$ (Table 1, Set 1)

Appendix: Justification of KICc for Nonlinear Regression.

In what follows, we will require that $f(Y|\theta_o) \in \mathcal{F}(k)$; i.e., that the fitted model is correctly specified or overfit. Under this assumption, the true model and the candidate model can be written using the same response function, the same design matrix, and parameter vectors of a common dimension:

$$Y = h(\delta_o, \mathbf{X}) + \epsilon, \qquad \epsilon \sim N(0, \sigma_o^2 I), \tag{A.1}$$

$$Y = h(\delta, \mathbf{X}) + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I). \tag{A.2}$$

Here, the common design matrix $\mathbf{X}$ is $n \times p$. The parameter vector for the true model (A.1), δ_o, and for the candidate model (A.2), δ, are both of dimension $p \times 1$.

The i^{th} row of the design matrix will be denoted by X_i. The mean vectors are assumed to have the layouts

$$h(\delta_o, \mathbf{X}) = (g(\delta_o, X_1), \cdots, g(\delta_o, X_n))' \qquad \text{and} \qquad h(\delta, \mathbf{X}) = (g(\delta, X_1), \cdots, g(\delta, X_n)),'$$

where the common mean response function g is twice continuously differentiable in δ. As before, for brevity, we will write $h(\delta, \mathbf{X})$ as $h(\delta)$ and $h(\delta_o, \mathbf{X})$ as $h(\delta_o)$.

We will assume that the final $(p - p_o)$ components of δ_o are zero $(p_o \leq p)$. This is permissible since we are requiring that the candidate model is either correctly or over specified.

As before, we will let $\theta_o = (\delta_o', \sigma_o^2)'$ and $\theta_k = (\delta', \sigma^2)'$, and use $f(Y|\theta_o)$ and $f(Y|\theta_k)$ to denote the densities corresponding to models (A.1) and (A.2).

Now recall (2.18) from Section 2:

$$\Omega(k, \theta_o) = \mathrm{E}_{\theta_o}[-2 \ln f(Y|\hat{\theta}_k)] + \mathrm{B}_1(k, \theta_o) + \mathrm{B}_2(k, \theta_o). \tag{A.3}$$

The first of the three terms on the right-hand side of (A.3) suggests that $-2 \ln f(Y|\hat{\theta}_k)$ serves as a biased estimator for $\Omega(k, \theta_o)$. As shown in Hurvich and Tsai (1989, pp. 299–300), for large n, $\mathrm{B}_1(k, \theta_o)$ is approximately equal to the penalty term of AICc; i.e.,

$$\mathrm{B}_1(k, \theta_o) \simeq \frac{2n(p+1)}{(n-p-2)}. \tag{A.4}$$

By (2.16), (2.21), and (2.22), the bias adjustment $\mathrm{B}_2(k, \theta_o)$ can be written as

$$\mathrm{B}_2(k, \theta_o) = \mathrm{E}_{\theta_o}\left[n \ln\left(\frac{\sigma_o^2}{\hat{\sigma}^2}\right) + \frac{n\hat{\sigma}^2}{\sigma_o^2} + \frac{\{h(\hat{\delta}) - h(\delta_o)\}'\{h(\hat{\delta}) - h(\delta_o)\}}{\sigma_o^2} - n \right]. \qquad \text{(A.5)}$$

To simplify (A.5), we will use the following large-sample results for MLEs in the normal nonlinear regression framework (Gallant, 1987, p. 17). The linear expansion of $h(\hat{\delta})$ at $\delta = \delta_o$ is given by $h(\hat{\delta}) \simeq h(\delta_o) + \mathbf{V}(\hat{\delta} - \delta_o)$, where $\mathbf{V} = \partial h(\delta)/\partial\delta$ evaluated at $\delta = \delta_o$. Under the true model, the difference $(\hat{\delta} - \delta_o)$ is approximately multivariate normal, $\mathrm{N}(0,\, \sigma_o^2(\mathbf{V}'\mathbf{V})^{-1})$, the ratio $(n\hat{\sigma}^2/\sigma_o^2)$ is approximately distributed as a χ^2 random variable with $(n-p)$ degrees of freedom, and $\hat{\delta}$ and $\hat{\sigma}^2$ are approximately independent.

Using the preceding, we have

$$\mathrm{E}_{\theta_o}\left[\frac{n\hat{\sigma}^2}{\sigma_o^2}\right] \simeq n - p, \qquad \text{(A.6)}$$

and

$$\mathrm{E}_{\theta_o}\left[\frac{\{h(\hat{\delta}) - h(\delta_o)\}'\{h(\hat{\delta}) - h(\delta_o)\}}{\sigma_o^2}\right] \simeq \mathrm{E}_{\theta_o}\left[\frac{(\hat{\delta} - \delta_o)'\mathbf{V}'\mathbf{V}(\hat{\delta} - \delta_o)}{\sigma_o^2}\right] \simeq p. \qquad \text{(A.7)}$$

Using (A.6) and (A.7) in conjunction with (A.5), we may argue

$$\mathrm{B}_2(k, \theta_o) \simeq \mathrm{E}_{\theta_o}\left[n \ln\left(\frac{\sigma_o^2}{\hat{\sigma}^2}\right)\right]. \qquad \text{(A.8)}$$

Now recall that the expectation of the log of a random variable with a central χ^2 distribution having *df* degrees of freedom is $\ln 2 + \psi(df/2)$, where $\psi(\cdot)$ denotes the *psi* or *digamma* function. (See, for instance, McQuarrie and Tsai, 1998, p. 67.) The term $\mathrm{E}_{\theta_o}\left[\ln(n\hat{\sigma}^2/\sigma_o^2)\right]$ has the form of the expectation of the log of central χ^2 random variable with $(n-p)$ degrees of freedom. This fact along with (A.3), (A.4), and (A.8) yields

$$\Omega(k, \theta_o) \simeq \mathrm{E}_{\theta_o}[f(Y|\,\hat{\theta}_k)] + \frac{2n(p+1)}{(n-p-2)} + n \ln\left(\frac{n}{2}\right) - n\psi\left(\frac{n-p}{2}\right). \qquad \text{(A.9)}$$

An accurate substitute for $\{n\ln(n/2) - n\psi((n-p)/2)\}$ is provided by the large-sample approximation (2.20). Employing this approximation in (A.9) justifies

$$\mathrm{KICc} = -2\ln f(Y|\,\hat{\theta}_k) + n\ln\left(\frac{n}{n-p}\right) + \frac{n\{(n-p)(2p+3) - 2\}}{(n-p-2)(n-p)}$$

as an approximately unbiased estimator of $\Omega(k, \theta_o)$.

References

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

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

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

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters* **42**, 333–343.

Cavanaugh, J. E. (2004). Criteria for linear model selection based on Kullback's symmetric divergence. To appear in *Australian and New Zealand Journal of Statistics.*

Gallant, A. R. (1987). *Nonlinear Statistical Models.* Wiley: New York, New York.

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

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

Kotz, S. and Johnson, N. L., editors (1982). *Encyclopedia of Statistical Sciences, Volume 2.* Wiley: New York, New York.

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

McQuarrie, A. D. R. and Tsai, C.–L. (1998). *Regression and Time Series Model Selection.* World Scientific: River Edge, New Jersey.

Myers, R. H. (1990). *Classical and Modern Regression with Applications (Second Edition)* Duxbury: Pacific Grove, California.

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