Criteria for Linear Model Selection
Based on Kullback's Symmetric Divergence

by

Joseph E. Cavanaugh

Department of Biostatistics, The University of Iowa

Summary

Model selection criteria frequently arise from constructing estimators of discrepancy measures used to assess the disparity between the 'true' model and a fitted approximating model. The Akaike (1973) information criterion and its variants result from utilizing Kullback's (1968) directed divergence as the targeted discrepancy. 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 (1968) symmetric divergence.

In the framework of linear models, a comparison of the two directed divergences indicates an important distinction between the measures. When used to evaluate fitted approximating models which are improperly specified, the directed divergence which serves as the basis for AIC is more sensitive towards detecting overfitted models, whereas its counterpart is more sensitive towards detecting underfitted models. Since the symmetric divergence combines the information in both measures, it functions as a gauge of model disparity which is arguably more balanced than either of its individual components. With this motivation, we propose a new class of criteria for linear model selection based on targeting the symmetric divergence. Our criteria may be regarded as analogues of AIC and two of its variants: 'corrected' AIC or AICc (Sugiura, 1978; Hurvich & Tsai, 1989), and 'modified' AIC or MAIC (Fujikoshi & Satoh, 1997). We examine the selection tendencies of the new criteria in a simulation study. Our results indicate that the new criteria perform favorably against their AIC analogues.

Key words: AIC, Akaike information criterion, I–divergence, J–divergence, Kullback–Leibler information, regression, relative entropy.

Acknowledgments. The author is indebted to the editor, the associate editor, and two referees whose thoughtful suggestions helped to improve the original version of this manuscript. This research was supported by the National Science Foundation, grant DMS–9704436.

1. Introduction

An important component of any linear modeling problem consists of determining an appropriate size and form for the design matrix. Improper specification may substantially impact both estimators of the model parameters and predictors of the response variable: underspecification may lead to results which are severely biased, whereas overspecification may lead to results with unnecessarily high variability. Model selection criteria, such as the Akaike (1973) information criterion or Mallows' (1973) conceptual predictive criterion, provide a powerful and useful tool for choosing a suitable design matrix. A selection criterion scores every fitted model in a candidate class in accordance with how effectively the model conforms to the data based on its size. Ideally, undesirable scores are assigned not only to models which omit essential variables, but also to models which adequately accommodate the data yet involve extraneous or irrelevant variables.

The first model selection criterion to gain wide–spread acceptance was the Akaike (1973) information criterion, AIC. AIC is applicable in a broad array of modeling frameworks, since its justification only requires conventional large–sample properties of maximum likelihood estimators. The criterion of Mallows' (1973), Cp, is used primarily for variable selection in linear regression, and is arguably the most popular criterion for this purpose. Since the introduction of AIC and Cp, many other criteria have been proposed and studied, including well–known measures by Parzen (1974), Schwarz (1978), Rissanen (1978), Akaike (1978), Hannan & Quinn (1979), Hurvich & Tsai (1989), Bozdogan (1990), and Wei (1992).

AIC serves as an estimator of a variant of Kullback's (1968 p.5) directed divergence between the 'true' model (i.e., the model which presumably gave rise to the data) and a fitted approximating model. The corrected Akaike information criterion, AICc, is an adjusted version of AIC originally proposed for linear regression with normal errors (Sugiura, 1978; Hurvich & Tsai, 1989). For fitted models in the candidate class which are correctly specified or overfitted, AIC is asymptotically unbiased and AICc is exactly unbiased as an estimator of its target measure. Recently, Fujikoshi & Satoh (1997) introduced a modification of AIC, MAIC, for multivariate linear regression with normal errors. Assuming that the true model is a member of the largest family of models in the candidate class, MAIC serves as an approximately unbiased estimator of its target measure for each of the fitted models in the class, including those which are underfitted.

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.

In the framework of linear models, a comparison of the two directed divergences indicates an important distinction between the measures. When used to evaluate fitted approximating models which are improperly specified, the directed divergence which serves as the basis for AIC is more sensitive towards detecting overfitted models, whereas its counterpart is more sensitive towards detecting underfitted models. Since the symmetric divergence reflects the sensitivities of both directed divergences, it may therefore serve as a more balanced discrepancy measure than either of its individual components.

Motivated by the preceding principle, we propose a new class of model selection criteria which target the symmetric divergence in the same manner that AIC, AICc, and MAIC target the directed divergence. We examine the selection tendencies of these new criteria in a simulation study. Our results indicate that the new criteria perform favorably against their AIC analogues.

In Section 2, we introduce and discuss Kullback's directed and symmetric divergence. We characterize and contrast these measures, and present a simulation example to illustrate the efficacy of the symmetric divergence as a discrepancy measure for linear model selection. In Section 3, we introduce the new class of selection criteria. A simulation study which illustrates the performance of the criteria is presented in Section 4. Proofs are given in the Appendix.

2. Kullback's directed and symmetric divergence

We begin with a brief outline of our model selection problem. Suppose a collection of data $\boldsymbol{y}$ has been generated according to an unknown parametric density $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$, one which corresponds to the normal linear model

$$\boldsymbol{y} = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \overset{\mathrm{d}}{=} \mathrm{N}_n(0, \sigma_0^2\, \boldsymbol{I}). \tag{1}$$

Here, we assume that the design matrix $\boldsymbol{X}_0$ is $n \times p_0$ of rank p_0, and the parameter vector $\boldsymbol{\beta}_0$ is $p_0 \times 1$. Note that $\boldsymbol{\theta}_0 = (\sigma_0^2, \boldsymbol{\beta}_0)$. We endeavor to find a fitted linear model which provides a suitable approximation to (1).

Let $\mathcal{F}(p) = \{f(\boldsymbol{y}|\,\boldsymbol{\theta}_p) \mid \boldsymbol{\theta}_p \in \Theta(p)\}$ denote a $(p+1)$–dimensional parametric family of densities, where each density $f(\boldsymbol{y}|\,\boldsymbol{\theta}_p)$ corresponds to a normal linear model

$$\boldsymbol{y} = \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \overset{\mathrm{d}}{=} \mathrm{N}_n(0, \sigma_p^2\, \boldsymbol{I}). \tag{2}$$

Here, we assume that $\boldsymbol{X}_p$ is $n \times p$ of rank p, and $\boldsymbol{\beta}_p$ is $p \times 1$. Note that $\boldsymbol{\theta}_p = (\sigma_p^2, \boldsymbol{\beta}_p)$. Let $\hat{\boldsymbol{\theta}}_p = (\hat{\sigma}_p^2, \hat{\boldsymbol{\beta}}_p)$ denote a vector of estimates obtained by maximizing the likelihood function $f(\boldsymbol{y}|\,\boldsymbol{\theta}_p)$ over $\Theta(p)$. Let $f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)$ represent the resulting empirical likelihood.

We remark that the assumption of full–rank design matrices in (1) and (2) is employed merely for expositional convenience, and is not fundamental to the subsequent development. The criteria in Section 3 may be used in settings where this assumption does not hold. For the criterion definitions in such applications, p should represent the rank of the design matrix in (2) (as opposed to the number of columns comprising the matrix).

Suppose our goal is to search among a class of families $\{\mathcal{F}(p_1), \mathcal{F}(p_2), \ldots, \mathcal{F}(p_L)\}$ for the fitted model $f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)$, $p \in \{p_1, p_2, \ldots, p_L\}$, which serves as the 'best' approximation to $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$. We note that in many applications, some of the families in this class may have the same dimension and yet be different: i.e., the models in some families may have design matrices with the same size and yet different column spaces. For ease of notation, we do not include an index to delineate between such families.

We refer to $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$ and (1) as the *true* or *generating* model. We refer to $f(\boldsymbol{y}|\,\boldsymbol{\theta}_p)$ and (2) as an *approximating* or *candidate* model.

If $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(p)$, and $\mathcal{F}(p)$ is such that no smaller family will contain $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$, we refer to $f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)$ as *correctly specified*. If $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(p)$, yet $\mathcal{F}(p)$ is such that families smaller than $\mathcal{F}(p)$ also contain $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$, we refer to $f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)$ as *overfitted*. If $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \notin \mathcal{F}(p)$, we refer to $f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)$ as *underfitted*.

To determine which of the fitted models $\{f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_1}), f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_2}), \ldots, f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_L})\}$ best resembles $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$, we require a measure which provides a suitable reflection of the disparity between the true model $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$ and an approximating model $f(\boldsymbol{y}|\,\boldsymbol{\theta}_p)$. Kullback's directed and symmetric divergence both fulfill this objective.

For two arbitrary parametric densities $f(\boldsymbol{y}|\,\boldsymbol{\theta})$ and $f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)$, Kullback's *directed divergence* between $f(\boldsymbol{y}|\,\boldsymbol{\theta})$ and $f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)$ with respect to $f(\boldsymbol{y}|\,\boldsymbol{\theta})$ is defined as

$$I(\boldsymbol{\theta}, \boldsymbol{\theta}_*) = \mathrm{E}_{\boldsymbol{\theta}} \left(\ln \frac{f(\boldsymbol{y}|\,\boldsymbol{\theta})}{f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)} \right), \tag{3}$$

and Kullback's *symmetric divergence* between $f(\boldsymbol{y}|\,\boldsymbol{\theta})$ and $f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)$ is defined as

$$J(\boldsymbol{\theta}, \boldsymbol{\theta}_*) = \mathrm{E}_{\boldsymbol{\theta}} \left(\ln \frac{f(\boldsymbol{y}|\,\boldsymbol{\theta})}{f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)} \right) + \mathrm{E}_{\boldsymbol{\theta}_*} \left(\ln \frac{f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)}{f(\boldsymbol{y}|\,\boldsymbol{\theta})} \right). \tag{4}$$

Here, $\mathrm{E}_{\boldsymbol{\theta}}$ denotes the expectation under $f(\boldsymbol{y}|\,\boldsymbol{\theta})$. Note that $J(\boldsymbol{\theta}, \boldsymbol{\theta}_*)$ is symmetric in its arguments whereas $I(\boldsymbol{\theta}, \boldsymbol{\theta}_*)$ is not: thus, an alternate directed divergence, $I(\boldsymbol{\theta}_*, \boldsymbol{\theta})$, may be obtained by switching the roles of $f(\boldsymbol{y}|\,\boldsymbol{\theta})$ and $f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)$ in (3). The sum of the two directed divergences yields the symmetric divergence: $J(\boldsymbol{\theta}, \boldsymbol{\theta}_*) = I(\boldsymbol{\theta}, \boldsymbol{\theta}_*) + I(\boldsymbol{\theta}_*, \boldsymbol{\theta})$.

In the model selection setting, each of $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$, $I(\boldsymbol{\theta}_p, \boldsymbol{\theta}_0)$, and $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ reflects the separation between the true model $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$ and the approximating model $f(\boldsymbol{y}|\,\boldsymbol{\theta}_p)$. It is well known that $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) \geq 0$ with equality if and only if $\boldsymbol{\theta}_p = \boldsymbol{\theta}_0$ (Kullback, 1968 pp.14–15); the same property then follows for $I(\boldsymbol{\theta}_p, \boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$.

In the model selection problem outlined, we search for a preferred model among a collection of fitted candidate models $\{f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_1}), f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_2}), \ldots, f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_L})\}$. For the purpose of assessing the proximity between a certain fitted candidate model $f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)$ and the true model $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$, we consider the measures $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p) = J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)|_{\boldsymbol{\theta}_p = \hat{\boldsymbol{\theta}}_p}$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p) = I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)|_{\boldsymbol{\theta}_p = \hat{\boldsymbol{\theta}}_p}$. In practical settings, neither $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ or $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ can be directly employed, since both measures depend upon the true parameter vector $\boldsymbol{\theta}_0$. However, each measure can be estimated. As the next section will indicate, AIC, AICc, and MAIC all target a variant of $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. Our intention is to propose estimators of a variant of $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ which are analogues of AIC, AICc, and MAIC. Yet before proceeding in this direction, we discuss an issue of obvious pertinence to the present objective: which of $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ or $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ provides a more suitable measure of disparity between $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0)$ and $f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)$?

To address this question, we must consider the relationship between the measures $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$. Corresponding to the true model (1) and the candidate model (2), we can easily show

$$\begin{aligned} I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) &= \frac{n}{2}\left(\ln\left(\frac{\sigma_p^2}{\sigma_0^2}\right) + \frac{\sigma_0^2}{\sigma_p^2}\right) + \frac{1}{2}\left((\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\boldsymbol{\beta}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\boldsymbol{\beta}_p)/\sigma_p^2\right) - \frac{n}{2}, \quad (5) \\ I(\boldsymbol{\theta}_p, \boldsymbol{\theta}_0) &= \frac{n}{2}\left(\ln\left(\frac{\sigma_0^2}{\sigma_p^2}\right) + \frac{\sigma_p^2}{\sigma_0^2}\right) + \frac{1}{2}\left((\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\boldsymbol{\beta}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\boldsymbol{\beta}_p)/\sigma_0^2\right) - \frac{n}{2}. \quad (6) \end{aligned}$$

For the interpretations of these measures which follow, a useful fact is that the function $f(x) = \ln(1/x) + x$ is positive and increasing in x for $x > 1$.

Consider $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$ as functions of $\hat{\boldsymbol{\theta}}_p = (\hat{\sigma}_p^2, \hat{\boldsymbol{\beta}}_p)$. Note that the form of $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ suggests that $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ will tend to be large when $(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)$ is large and $\hat{\sigma}_p^2$ is small. For models which are correctly specified or overfitted,

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}_0}\left((\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)\right) &= p\sigma_0^2, \\ \mathrm{E}_{\boldsymbol{\theta}_0}\left(\hat{\sigma}_p^2\right) &= \left(1 - \frac{p}{n}\right)\sigma_0^2. \end{aligned}$$

Thus, the form of $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ implies that $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ will tend to be sensitive towards overfitting, assuming large values when p is large and $\hat{\sigma}_p^2$ is substantially deflated.

On the other hand, the form of $I(\boldsymbol{\theta}_p, \boldsymbol{\theta}_0)$ suggests that $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$ will tend to be large when $(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)$ and $\hat{\sigma}_p^2$ are both large. For models which are underfitted,

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}_0}\left((\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)\right) &= p\sigma_0^2 + (\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0), \\ \mathrm{E}_{\boldsymbol{\theta}_0}\left(\hat{\sigma}_p^2\right) &= \left(1 - \frac{p}{n}\right)\sigma_0^2 + (\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)/n, \end{aligned}$$

where $\boldsymbol{H}_p$ is the projection matrix onto the column space of $\boldsymbol{X}_p$. The size of the quadratic form $(\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)$ is dictated by the extent to which the design matrix of the candidate model is underspecified. Thus, the form of $I(\boldsymbol{\theta}_p, \boldsymbol{\theta}_0)$ implies that $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$ will tend to be sensitive towards underfitting, assuming large values when $(\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)$ is large and $\hat{\sigma}_p^2$ is substantially inflated.

In evaluating the adequacy of a fitted candidate model, the preceding indicates that $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ may better reflect the error due to *estimation variability*, whereas $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$ may better reflect the error due to *estimation bias*. This suggests that there might be an advantage to combining $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$: that the composite measure may provide a more balanced gauge of model disparity than either of its individual components. Thus, in settings where the collection $\{f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_1}), f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_2}), \ldots, f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_{p_L})\}$ consists of both underfitted and overfitted models, $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ may serve to better indicate which models are improperly specified than $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. As a consequence, an estimator of $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ may be preferable to an estimator of $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ as a model selection criterion, provided that the former estimator is accurate enough to sufficiently reflect the sensitivity of $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$.

The following example serves to illustrate the effectiveness of $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ as discrepancy measures for linear model selection.

Example: Contrasting $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ in the regression framework

Envision a setting where a sample of size $n = 26$ is generated from a true model

$$y = 1 + x_1 + x_2 + x_3 + \epsilon, \quad \text{where } \epsilon \overset{\mathrm{d}}{=} \mathrm{N}(0, 36). \tag{7}$$

We consider using both $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ to determine which of the following fitted candidate models best describes the data:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\epsilon}, \tag{8}$$

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\epsilon}, \tag{9}$$

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\epsilon}. \tag{10}$$

As previously mentioned, $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ depend on $\boldsymbol{\theta}_0$, and therefore are 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.

Refer to (5) and (6). Viewing σ_0^2 and $\boldsymbol{X}_0\boldsymbol{\beta}_0$ as fixed, note that both $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ and $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) = I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) + I(\boldsymbol{\theta}_p, \boldsymbol{\theta}_0)$ may be regarded as functions of σ_p^2 and the quadratic form $\mathrm{Q}_p = (\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\boldsymbol{\beta}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\boldsymbol{\beta}_p)$. In the present setting, Figure 1 features three–dimensional plots of $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ and $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ as functions of σ_p^2 and Q_p. The minimum value of each function occurs at the point $(\sigma_p^2, \mathrm{Q}_p) = (\sigma_0^2, 0) = (36, 0)$. Moving in any direction from this point, the increase in $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ exceeds that of $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$. The difference in the curvature of the divergences is particularly evident in the back upper–left corners of the plots.

To assess the behavior of $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ as selection rules, we generate 10 000 samples from the true model (7). For every sample, we compute the fitted models (8), (9), (10). We then evaluate $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ for all three models, and record the fitted model corresponding to the minimum value of each measure. The regressors for all models are generated from a uniform (0,10) distribution. The results are featured below, along with the average values of $\hat{\sigma}_p^2$, $\hat{\mathrm{Q}}_p$, $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ for each candidate model.

Fitted					Selections	Selections
Model	$\hat{\mathrm{E}}_{\boldsymbol{\theta}_0}(\hat{\sigma}_p^2)$	$\hat{\mathrm{E}}_{\boldsymbol{\theta}_0}(\hat{\mathrm{Q}}_p)$	$\hat{\mathrm{E}}_{\boldsymbol{\theta}_0}(I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p))$	$\hat{\mathrm{E}}_{\boldsymbol{\theta}_0}(J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p))$	for $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$	for $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$
(8)	48.7	472.4	6.03	13.78	854	398
(9)	30.5	143.4	3.68	6.43	8408	9026
(10)	35.9	354.4	6.15	11.69	738	576

Note that $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ chooses the correctly specified model, (9), more frequently than $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. The reason for this can be explained by inspecting the plots of $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ and $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ in Figure 1, particularly in the neighborhoods of the points $(\hat{\mathrm{E}}_{\boldsymbol{\theta}_0}(\hat{\sigma}_p^2), \hat{\mathrm{E}}_{\boldsymbol{\theta}_0}(\hat{\mathrm{Q}}_p))$ corresponding to each of the candidate models. For the correctly specified model, (9), the point $(\hat{\mathrm{E}}_{\boldsymbol{\theta}_0}(\hat{\sigma}_p^2), \hat{\mathrm{E}}_{\boldsymbol{\theta}_0}(\hat{\mathrm{Q}}_p)) = (30.5, 143.4)$ lies to the back right of the point at which the discrepancy surfaces attain their minimum, $(36, 0)$. Moving from the point $(30.5, 143.4)$ towards either of the points $(48.7, 472.4)$ (for model (8)) or $(35.9, 354.4)$ (for model (10)), the curvature of $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ is more pronounced than that of $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$, especially in the direction of $(48.7, 472.4)$. Thus, for a particular sample, it is more likely for $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ than for $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ to be minimized at the coordinate $(\hat{\sigma}_p^2, \hat{\mathrm{Q}}_p)$ corresponding to the correctly specified model.

As mentioned previously, $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ combines the information in $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$, two measures which are related and yet distinct. Over the 10 000 samples generated for this example, the correlations between $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$ for the fitted candidate models (8), (9), and (10) are 0.483, 0.909, and 0.716, respectively. This reinforces the notion that $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$ are not redundant, and advances the premise that $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ improves upon $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ by incorporating the additional information in $I(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)$.

3. Model selection criteria based on Kullback's symmetric divergence

In this section, we propose model selection criteria based on the symmetric divergence. We begin by discussing how AIC, AICc, and MAIC are justified as estimators of a variant of $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. We then discuss how analogous criteria can be derived as estimators of a variant of $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. Proofs are given in the Appendix.

For two arbitrary parametric densities $f(\boldsymbol{y}|\,\boldsymbol{\theta})$ and $f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)$, define

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}_*) = \mathrm{E}_{\boldsymbol{\theta}}\left(-2\ln f(\boldsymbol{y}|\,\boldsymbol{\theta}_*)\right). \tag{11}$$

From (3) and (11), note that we can write

$$2I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0).$$

Since $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ does not depend on $\boldsymbol{\theta}_p$, any ranking of a set of candidate models corresponding to values of $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ would be identical to a ranking corresponding to values of $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$. Hence, for the purpose of discriminating among various candidate models, $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ serves as a valid substitute for $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$.

Similarly, using (4) and (11), we can write

$$2J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) = \left(d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)\right) + \left(d(\boldsymbol{\theta}_p, \boldsymbol{\theta}_0) - d(\boldsymbol{\theta}_p, \boldsymbol{\theta}_p)\right).$$

Again, for the purpose of discriminating among various candidate models, we can use

$$K(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p) + \left(d(\boldsymbol{\theta}_p, \boldsymbol{\theta}_0) - d(\boldsymbol{\theta}_p, \boldsymbol{\theta}_p)\right) \tag{12}$$

as a substitute for the measure $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$. Measures such as $K(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$, $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$, $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$, and $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ are often called *discrepancies*. (See Linhart & Zucchini, 1986 pp.11–12.)

Now consider the measure

$$d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p) = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)|_{\boldsymbol{\theta}_p = \hat{\boldsymbol{\theta}}_p},$$

which is equivalent to $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ for the intended purpose. The work of Akaike (1973, 1974) suggests that $-2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p)$ serves as a biased estimator of $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$, and that in many applications (including those beyond the scope of linear models), the bias adjustment

$$\mathrm{B}_1(p, \boldsymbol{\theta}_0) = \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)\right) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(-2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p)\right) \tag{13}$$

can be asymptotically estimated by twice the dimension of $\hat{\boldsymbol{\theta}}_p$. Thus, since $(p+1)$ denotes the dimension of $\hat{\boldsymbol{\theta}}_p$, under appropriate conditions, the expected value of

$$\mathrm{AIC} = -2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p) + 2(p+1) \tag{14}$$

should asymptotically approach the expected value of $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. Specifically, if we assume that $\hat{\boldsymbol{\theta}}_p$ satisfies the conventional properties of maximum likelihood estimators, and that $f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p)$ is either correctly specified or overfitted, we can establish

$$\mathrm{B}_1(p, \boldsymbol{\theta}_0) - 2(p+1) = o(1). \tag{15}$$

(See, for instance, Cavanaugh, 1997 p.204.) Since (15) is a general result, AIC is applicable in a broad array of settings.

The work of Sugiura (1978) and Hurvich & Tsai (1989) indicates that in the framework of normal linear regression models, the bias adjustment (13) can be evaluated exactly for correctly specified and overfitted models. For the true model (1) and the candidate model (2), it can be shown that when $f(\boldsymbol{y}|\, \boldsymbol{\theta}_0) \in \mathcal{F}(p)$,

$$\mathrm{B}_1(p, \boldsymbol{\theta}_0) = \frac{2n(p+1)}{(n-p-2)}. \tag{16}$$

(See Cavanaugh, 1997 pp.204–205.) Thus, an exactly unbiased estimator of $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ is given by

$$\mathrm{AICc} = -2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p) + \frac{2n(p+1)}{(n-p-2)}. \tag{17}$$

Note from (14) and (17) that AIC $-$ AICc $= O(n^{-1})$.

Recently, Fujikoshi & Satoh (1997) proposed a modification of AIC, MAIC, based on an approximate evaluation of the bias adjustment (13) for underfitted as well as correctly specified and overfitted models. They develop their variant for normal multivariate linear regression models under the assumption that the true model is a member of the largest family in the candidate class. Suppose that $\mathcal{F}(P)$ denotes this family, and that $\hat{\boldsymbol{\theta}}_P = (\hat{\sigma}_P^2, \hat{\boldsymbol{\beta}}_P)$ denotes the corresponding MLE. For a particular family $\mathcal{F}(p)$ in the candidate class (possibly $\mathcal{F}(P)$) and its associated MLE $\hat{\boldsymbol{\theta}}_p = (\hat{\sigma}_p^2, \hat{\boldsymbol{\beta}}_p)$, define

$$\lambda(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P) = \frac{(n-p)\hat{\sigma}_P^2}{(n-P)\hat{\sigma}_p^2}.$$

Fujikoshi & Satoh (1997 pp.709–711) suggest estimating (13) by

$$b_1(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P) = \frac{2n(p+1)}{(n-p-2)} + 2p\left(\lambda(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P) - 1\right) - 2\left(\lambda(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P) - 1\right)^2. \tag{18}$$

MAIC is then defined as

$$\text{MAIC} = -2\ln f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p) + b_1(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P).$$

Assuming $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(P)$ and

$$(\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0) = O(n), \tag{19}$$

it can be shown that for correctly specified or overfitted models,

$$\text{B}_1(p, \boldsymbol{\theta}_0) - \text{E}_{\boldsymbol{\theta}_0}\left(b_1(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)\right) = O(n^{-2}),$$

and that for underfitted models,

$$\text{B}_1(p, \boldsymbol{\theta}_0) - \text{E}_{\boldsymbol{\theta}_0}\left(b_1(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)\right) = O(n^{-1}).$$

(See Fujikoshi & Satoh, 1997 pp.710–711.)

Now recalling (12), consider estimating the measure

$$K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p) = K(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)|_{\boldsymbol{\theta}_p = \hat{\boldsymbol{\theta}}_p}.$$

For the intended purpose, note that $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ is equivalent to $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ in the same way that $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ is equivalent to $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. If the empirical log–likelihood $-2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p)$ is considered as a platform for an estimator of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$, the challenge is then to correct for the bias. The bias adjustment may be expressed as

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}_0}\left(K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)\right) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(-2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p)\right) &= \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)\right) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(-2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p)\right) \quad (20) \\ &\quad + \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)\right) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_p)\right). \quad (21) \end{aligned}$$

Note that the difference on the right–hand side of (20) is the same as $\mathrm{B}_1(p, \boldsymbol{\theta}_0)$, the bias adjustment for $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ expressed in (13). For the difference (21), let

$$\mathrm{B}_2(p, \boldsymbol{\theta}_0) = \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)\right) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_p)\right). \quad (22)$$

We can then write

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left(K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)\right) = \mathrm{E}_{\boldsymbol{\theta}_0}\left(-2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p)\right) + \mathrm{B}_1(p, \boldsymbol{\theta}_0) + \mathrm{B}_2(p, \boldsymbol{\theta}_0). \quad (23)$$

The penalty terms of AIC, AICc, and MAIC provide us with estimators of $\mathrm{B}_1(p, \boldsymbol{\theta}_0)$; our goal is to seek similar estimators of $\mathrm{B}_2(p, \boldsymbol{\theta}_0)$. This will lead us to a set of criteria which are analogous to AIC, AICc, and MAIC, targeting $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ in the same way that the Akaike–type criteria target $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$.

First, we propose an analogue of AIC based on estimating $\mathrm{B}_2(p, \boldsymbol{\theta}_0)$ in the same manner that the penalty term of AIC estimates $\mathrm{B}_1(p, \boldsymbol{\theta}_0)$. If we assume that $\hat{\boldsymbol{\theta}}_p$ satisfies the conventional properties of maximum likelihood estimators, and that $f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p)$ is either correctly specified or overfitted, we can establish

$$\mathrm{B}_2(p, \boldsymbol{\theta}_0) - (p + 1) = o(1). \quad (24)$$

(See Cavanaugh, 1999 pp.337–338.) Based on (15) and (24), we define the criterion

$$\mathrm{KIC} = -2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p) + 3(p + 1) \quad (25)$$

via (23). This criterion was previously introduced by Cavanaugh (1999), who illustrates its performance in a simulation study where the objective is to choose the order of an autoregression.

Next, we propose an analogue of AICc based on estimating $\mathrm{B}_2(p, \boldsymbol{\theta}_0)$ in the same manner that the penalty term of AICc estimates $\mathrm{B}_1(p, \boldsymbol{\theta}_0)$. In the setting of normal linear models, the bias adjustment (22) can be evaluated exactly for correctly specified and overfitted models. For the true model (1) and the candidate model (2), it can be shown that when $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(p)$,

$$\mathrm{B}_2(p, \boldsymbol{\theta}_0) = n \ln \left(\frac{n}{2}\right) - n\psi \left(\frac{n-p}{2}\right), \tag{26}$$

where ψ denotes the digamma function. (See Lemma 1 in the Appendix.) Bernardo (1976) presents a simple algorithm for computing precise values of ψ. However, an accurate substitute for (26) is suggested by the approximation

$$n \ln \left(\frac{n}{2}\right) - n\psi \left(\frac{n-p}{2}\right) = n \ln \left(\frac{n}{n-p}\right) + \frac{n}{n-p} + O(n^{-2}). \tag{27}$$

(See Kotz & Johnson, 1982 p.373.) Based on (16), (26), and (27), we define the criterion

$$\mathrm{KICc} = -2 \ln f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p) + n \ln \left(\frac{n}{n-p}\right) + \frac{n\left((n-p)(2p+3)-2\right)}{(n-p-2)(n-p)} \tag{28}$$

via (23). Assuming that $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(p)$, note that the bias incurred in estimating $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ with KICc is $O(n^{-2})$. Also, using (25) and (28), we can show that KIC $-$ KICc $= O(n^{-1})$.

We remark that KICc is behaviorally similar to a variant of AIC suggested by McQuarrie, Shumway & Tsai (1997). This variant arises by adjusting the goodness–of–fit term of AICc, $-2 \ln f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)$, so that its expected value is equal to $d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0)$ for correctly specified and overfitted models.

Finally, we introduce an analogue of MAIC based on an estimator of $\mathrm{B}_2(p, \boldsymbol{\theta}_0)$ which has validity regardless of whether $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(p)$. As with the justification of MAIC, we assume a normal linear modeling framework where $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(P)$, in which $\mathcal{F}(P)$ denotes the largest family in the candidate class. We also assume (19).

Define

$$\delta(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P) = \frac{(n-P-2)\hat{\sigma}_p^2}{\hat{\sigma}_P^2} + p - (n-2) \tag{29}$$

and

$$b_2(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P) = -\left(-2 \ln f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)\right) + 2\delta(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P). \tag{30}$$

In the Appendix (Lemma 2), we establish that $\delta(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)$ serves as an unbiased estimator of the quadratic form $(\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)/\sigma_0^2$. (Recall that the magnitude of this quadratic form reflects the extent to which the candidate model is underfitted.) We then argue that for correctly specified, overfitted, and underfitted models,

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left(b_2(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)\right) = \mathrm{B}_2(p, \boldsymbol{\theta}_0) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0). \tag{31}$$

The preceding implies $b_2(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)$ is exactly unbiased for $\mathrm{B}_2(p, \boldsymbol{\theta}_0)$ to within a constant amount.

Now the bias properties of the penalty term for AICc under assumption (19) are as follows (Fujikoshi & Satoh, 1997 p.711). For correctly specified and overfitted models

$$\frac{2n(p+1)}{(n-p-2)} - \mathrm{B}_1(p, \boldsymbol{\theta}_0) = 0, \tag{32}$$

and for underfitted models,

$$\frac{2n(p+1)}{(n-p-2)} - \mathrm{B}_1(p, \boldsymbol{\theta}_0) = O(1). \tag{33}$$

By combining (17) and (30), we arrive at the criterion

$$\mathrm{MKIC} = 2\delta(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P) + \frac{2n(p+1)}{(n-p-2)}.$$

Assuming that $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(P)$ and that (19) holds, note by (23), (31), (32), and (33) that for correctly specified and overfitted models,

$$\mathrm{E}_{\boldsymbol{\theta}_0}(\mathrm{MKIC}) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)\right) + d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = 0,$$

and for underfitted models,

$$\mathrm{E}_{\boldsymbol{\theta}_0}(\mathrm{MKIC}) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)\right) + d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = O(1).$$

We note that the KIC family of criteria may be negatively biased for underfitted models, MKIC to a lesser extent than KICc or KIC. However, when targeting $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$, our investigations suggest that precise estimation of the discrepancy is much less essential for underfitted models than for overfitted models. (In fact, it is for this reason we have chosen to use (16) as opposed to (18) as an estimator of $\mathrm{B}_1(p, \boldsymbol{\theta}_0)$ in constructing MKIC.) Relative to the minimum value of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$, values of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ for models which are underspecified tend to be quite large, whereas values of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ for models which are slightly overspecified tend to be comparable. Thus, an estimator of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ which is crude for underfitted models yet precise for overfitted models can adequately reflect the selection tendencies of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$.

Having introduced KIC, KICc, and MKIC as analogues of the Akaike–type criteria AIC, AICc, and MAIC, we now examine the selection tendencies of the criteria in a simulation study.

4. Simulations

Consider a setting where a sample of size n is generated from a true regression model of the form (1) corresponding to a design matrix of rank p_0. Assume our objective is to search among a candidate class of families for the fitted model which serves as the best approximation to (1).

Suppose our candidate models are of the form (2). In all design matrices, we require that the initial column is a vector consisting of all ones. Assuming we have $(P-1)$ regressor variables of interest, we consider candidate models corresponding to design matrices of ranks $p = 2, 3, \ldots, P$, where the smallest models ($p = 2$) contain only one regressor, and the largest model ($p = P$) contains all $(P-1)$ regressors. We require that one of the candidate models in our class is correctly specified, i.e., contains the same regressors as the true model.

We examine the behavior of AIC, AICc, MAIC, KIC, KICc, and MKIC in simulation sets where the criteria are used to select a fitted model from the candidate class. In each set, 1000 samples are generated from the true model. For every sample, the candidate models are fit to the data, the criteria are evaluated, and the fitted model favored by each criterion is recorded. Over the 1000 samples, the proportion of correct model selections is determined for each of the criteria. Thus, in each simulation set, we can compare the criterion success rates in choosing the correctly specified fitted model.

Our simulation study consists of two parts. In the first part, we assume that the candidate models are nested. Thus, with $(P-1)$ regressor variables of interest, we entertain $(P-1)$ candidate models based on a sequence of design matrices of ranks $p = 2, 3, \ldots, P$. Each successive design matrix contains all of the regressors in its predecessors. The candidate model having the design matrix of rank p_0 is correctly specified $(2 \leq p_0 \leq P)$. Moreover, fitted models for which $2 \leq p < p_0$ are underfitted, and those for which $p_0 < p \leq P$ are overfitted. We refer to p as the *order* of the model, and to p_0 as the *true order*.

In the second part, we assume that the candidate models correspond to all possible subsets of the regressor variables. With $(P-1)$ regressor variables of interest, a total of $(2^{(P-1)} - 1)$ candidate models may be constructed. (Recall that intercept–only models are excluded.) We again entertain candidate models based on design matrices of ranks $p = 2, 3, \ldots, P$; however, for each rank p, we must consider $\binom{P-1}{p-1}$ different models representing various combinations of $(p-1)$ regressors. One of the candidate models having a design matrix of rank p_0 is correctly specified $(2 \leq p_0 \leq P)$. Fitted models corresponding to design matrices that do not contain all of the regressors in the true model are underfitted. Fitted models corresponding to design matrices for which $p > p_0$ that contain all of the regressors in the true model are overfitted.

For practical applications, the all possible regressions (APR) framework is more realistic than the nested models (NM) framework. However, the latter setting is often used in simulation studies for model selection criteria so that large candidate models may be considered without making the number of models in the candidate class excessively high. (See, for instance, McQuarrie & Tsai, 1998.) When P is large and n is relatively small, criteria with penalty terms justified asymptotically are often outperformed by variants with penalty terms refined for finite samples. Thus, simulations based on nested models may allow us to better assess the efficacy of the 'corrected' and 'modified' criteria.

In addition to the form of the candidate class (dictated by P), the performance of a model selection criterion is also affected by two other factors: the size of the sample n, and the extent to which the true regression surface is obscured in the sample by random error. To quantify the latter factor, we define a *signal–to–noise ratio* (SNR) for the true model as a ratio of two variances: the variance of the linear form in the regressor variables relative to the variance of the error component. Of course, in traditional regression applications, the linear form in the regressors is regarded as deterministic and thereby has a variance of zero. However, our SNR definition is sensible in the context of our simulation sets since the regressors are randomly generated. Moreover, our definition is amenable to a familiar interpretation: if a correctly specified model is fit to data generated under a true model with a signal–to–noise ratio of SNR, the coefficient of determination for the fit will be approximately SNR/(1+SNR).

We present a total of five collections of simulation sets, three in the NM setting and two in the APR setting. A collection is characterized by three factors: the structure of the true model (dictated by $\boldsymbol{\beta}_0$ and σ_0^2), the form of the candidate class (dictated by P), and the signal–to–noise ratio (SNR). In every collection, the sample size n is varied over a sequence of values deemed appropriate based on the SNR. A simulation set involving 1000 replications is compiled for each sample size represented in the sequence.

The results of each collection are summarized graphically: for every criterion, the proportion of correct model selections is plotted against the sample size n. The figures also feature the proportion of correct model selections obtained by $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ (or equivalently, by $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $d(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$). As previously mentioned, $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ depend on $\boldsymbol{\theta}_0$, and therefore are 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.

Nested models

In the nested models (NM) framework, we consider three collections of simulation sets corresponding to signal–to–noise ratios of 1.0 : 1.0, 1.0 : 1.5, and 1.0 : 2.0. In all sets, $P = 13$. The true model is parameterized by $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1, 1, 2)$, so that $p_0 = 7$. Thus, in the candidate class, the smallest 5 models are underfitted, and the largest 6 models are overfitted. The regressors are generated from a N$(0, 10)$ distribution.

In the first collection, the true model variance is set at $\sigma_0^2 = 90$, making the signal–to–noise ratio 1.0 : 1.0. The collection features 12 sets, with sample sizes of $n = 22, 26, 30, \ldots, 66$. The results of the 12 sets are graphically summarized in Figure 2(a).

In the sets corresponding to smaller sample sizes ($n = 22, 26, 30$), the form of the true regression surface cannot be easily delineated. The criteria have difficulty identifying the correctly specified model, often choosing models that are underfitted. In these sets, MKIC obtains the highest proportion of correct order selections, followed closely by AICc. As the sample size is increased, the propensity to underfit is attenuated, and the performance of the criteria improves. The rates for MKIC and AICc become virtually identical. MAIC overtakes MKIC/AICc, yet is eventually outdistanced by KICc. For the larger sample sizes ($n \geq 54$), all of the criteria except AIC achieve correct selection rates above 80%.

Comparing the 'modified' criteria, we note that MKIC outperforms MAIC for smaller sample sizes ($n = 22, 26, 30$), where the propensity to underfit is greatest. However, MAIC eventually overtakes MKIC. If the sample sizes were further increased (beyond $n = 66$), the correct selection rates for the two criteria would begin to converge. With the 'corrected' criteria, AICc initially outperforms KICc ($n \leq 42$), yet is subsequently surpassed by KICc ($n \geq 46$). Because KICc shares the same goodness–of–fit term as AICc yet features a more stringent penalty term, if the sample sizes were further increased (beyond $n = 66$), the correct selection rates for KICc would continue to exceed those of AICc. With the original criteria, KIC outperforms AIC over all sets. As with KICc/AICc, this tendency would persist for larger sample sizes.

As a selection rule, $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ consistently obtains a proportion of correct order selections that is at least as high as that obtained by $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. Differences in the correct selection rates are particularly evident for the smaller sample sizes.

In the second collection, the true model variance is set at $\sigma_0^2 = 135$, making the signal–to–noise ratio 1.0 : 1.5. The collection features 12 sets, with sample sizes of $n = 26, 32, 38, \ldots, 92$. (The sample sizes are increased more rapidly here compared to the first collection due to the smaller value of SNR.) The results of the 12 sets are featured in Figure 2(b).

The pattern exhibited in the results for the second collection mirrors the pattern exhibited in the first. In the sets corresponding to smaller sample sizes ($n = 26, 32, 38$) where underfitting is problematic, MKIC obtains the highest proportion of correct order selections, followed closely by AICc. As the sample size is increased, MAIC surpasses MKIC/AICc, yet is eventually outperformed by KICc. The correct selection rates for MKIC and AICc become indistinguishable. For the larger sample sizes ($n \geq 56$), all of the criteria except AIC achieve correct selection rates exceeding 75%.

Between MKIC and MAIC, the former is superior for smaller–sample sets and the latter for larger–sample sets. KICc is initially outperformed by AICc, yet KICc subsequently surpasses AICc. KIC again dominates AIC over all sets. Also, $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ consistently obtains correct selection rates at least as high as $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$.

In the third and final collection, the true model variance is set at $\sigma_0^2 = 180$, resulting in a signal–to–noise ratio of 1.0 : 2.0. The collection features 13 sets, with sample sizes of $n = 32, 46, 60, \ldots, 200$. (Again, the sample size increments are larger here than in the previous collection to account for the smaller value of SNR.) The results of the 13 sets are illustrated in Figure 2(c).

In the sets corresponding to larger sample sizes ($n \geq 88$), all of the criteria except AIC achieve correct selection rates exceeding 75%. KICc outperforms the remaining criteria, followed by KIC. The criteria MAIC, MKIC, and AICc exhibit similar correct selection rates. AIC initially trails MAIC/MKIC/AICc, yet begins to obtain similar rates for the latter sets. The initial sets featuring the smaller sample sizes ($n \leq 74$) indicate tendencies similar to those found in the first two collections. Again, over all sets, the proportion of correct order selections for $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ is at least as high as that for $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$.

All possible regressions

In the all possible regressions (APR) framework, we consider two collections of simulation sets. In all sets, $P = 6$. Therefore, the candidate class consists of 31 models based on 5 potential regressors: 1 model containing all 5 regressors, 5 with 4 regressors, 10 with 3 regressors, 10 with 2 regressors, and 5 with only 1 regressor. The regressors are again generated from a N$(0, 10)$ distribution.

In the first collection, the true model is parameterized by $\boldsymbol{\beta}_0 = (1, 2)$ ($p_0 = 2$) with $\sigma_0^2 = 40$. Thus, the signal–to–noise ratio is 1.0 : 1.0. In the second, $\boldsymbol{\beta}_0 = (1, 1, 1)$ ($p_0 = 3$) and $\sigma_0^2 = 10$. Thus, the signal–to–noise ratio is 1.0 : 0.5. Both collections feature 11 sets, with sample sizes of $n = 20, 28, 36, \ldots, 100$. The results of the first collection are featured in Figure 3(a); the results of the second are displayed in Figure 3(b).

The correct selection tendencies in both APR collections are quite similar. Moreover, these tendencies tend to mirror those exhibited over the larger sample sizes ($n \geq 88$) for the third collection in the nested models (NM) setting. (Although the sample sizes employed here vary over a range of smaller values than those used in the third NM collection, the models in the candidate class are based on design matrices of much lower rank. Thus, the sample sizes are effectively larger due to the reduced dimensions of the candidate models.) The correct selection rates for the APR collections, however, do not achieve levels as high as those in the NM collections. In part, this is due to the larger number of incorrectly specified models represented in the APR candidate classes.

Referring to Figures 3(a) and 3(b), we note that KICc clearly outperforms the remaining criteria. The correct selection rates of KIC are also relatively high, and begin to approach those of KICc as the sample size is increased. The criteria MAIC, MKIC, and AICc exhibit similar correct selection rates. MAIC marginally outperforms MKIC, which in turn, marginally outperforms AICc. The difference in rates for MAIC, MKIC, and AICc diminish as the sample size is raised. AIC obtains the lowest correct selection rates.

Comparing the 'modified' criteria, we note that MKIC is consistently outperformed by MAIC (albeit marginally). However, with the 'corrected' criteria, KICc always obtains higher correct selection rates than AICc. With the original criteria, KIC markedly outperforms AIC over all sets.

As selection rules, $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ and $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ are both very effective in delineating the correctly specified model, exhibiting correct selection rates that quickly approach 1.0 as the sample size is increased. Again, however, in sets where the rules choose improperly specified models, the correct selection rates for $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ always exceed those for $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$.

5. Conclusions and discussion

Model selection criteria often result from constructing approximately unbiased estimators of suitable discrepancy functions, evaluated to measure the disparity between the true or generating model and a fitted approximating model. Kullback's directed divergence is one of many possible measures which can be used as a discrepancy function. The arguments and simulation results in Sections 2 and 4, however, indicate that Kullback's symmetric divergence may improve upon the directed divergence since it better reflects the risks of both overfitting and underfitting.

Of course, even if $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ is deemed a more appropriate discrepancy than $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ for model selection applications, criteria which target $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ (or a variant thereof) improve upon those which target $I(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ only if the criteria provide sufficiently accurate estimators of $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$. The simulation results in Section 4 indicate that the proposed class of criteria based on $J(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ perform effectively. Specifically, MKIC shows promise as a smaller–sample selection criterion (where underfitting may be likely), whereas KICc and KIC show promise as larger–sample selection criteria (where underfitting is not as problematic yet overfitting remains probable). However, it is clear that many different estimators of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ are possible, both in the setting of linear models as well as in other modeling frameworks. Criteria based on more sophisticated estimators of $K(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_p)$ have the potential to effectively guard against both underfitting and overfitting over a wide array of different applications. Developing such criteria provides the impetus for future work.

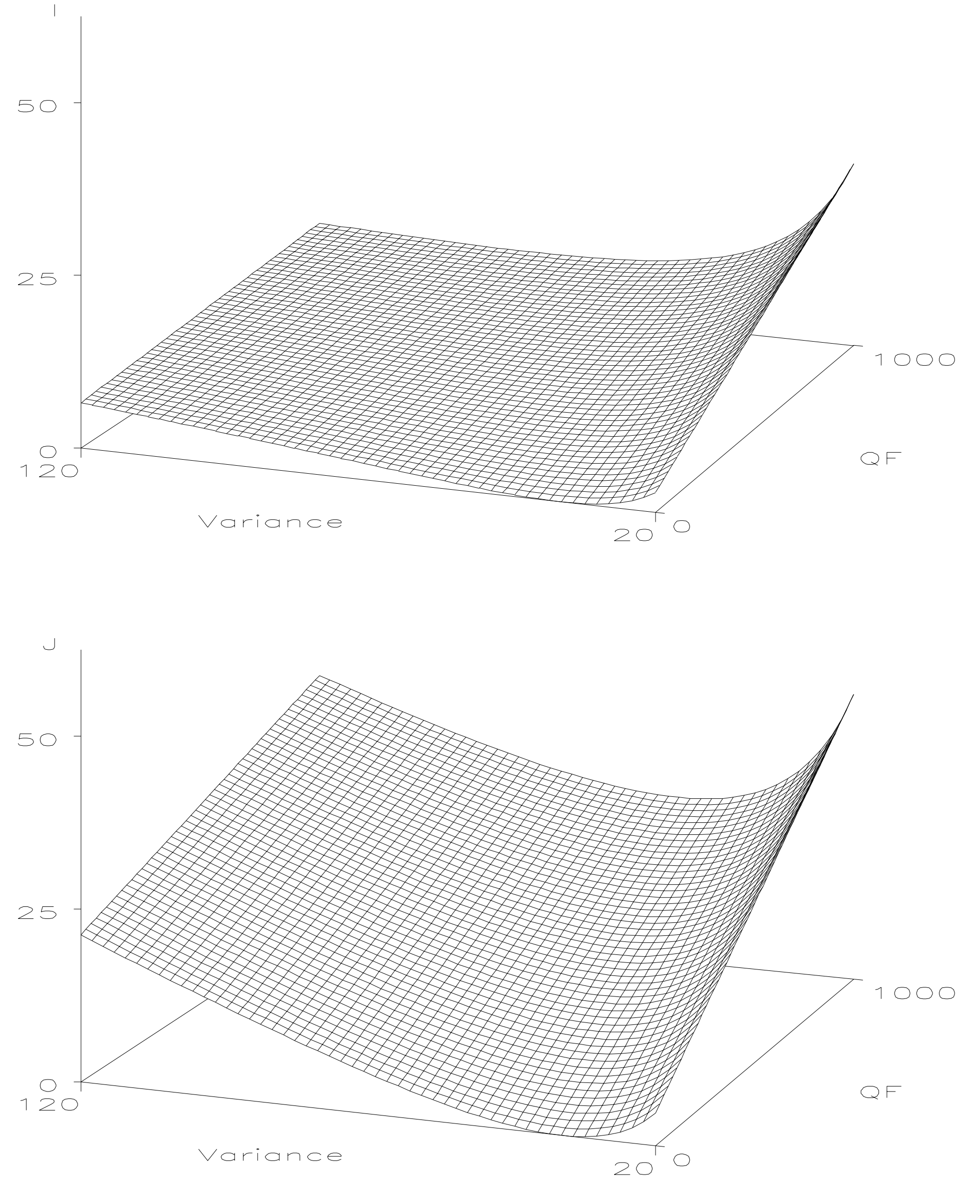

Figure 1: Graphs of $I(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$ and $J(\boldsymbol{\theta}_0, \boldsymbol{\theta}_p)$.

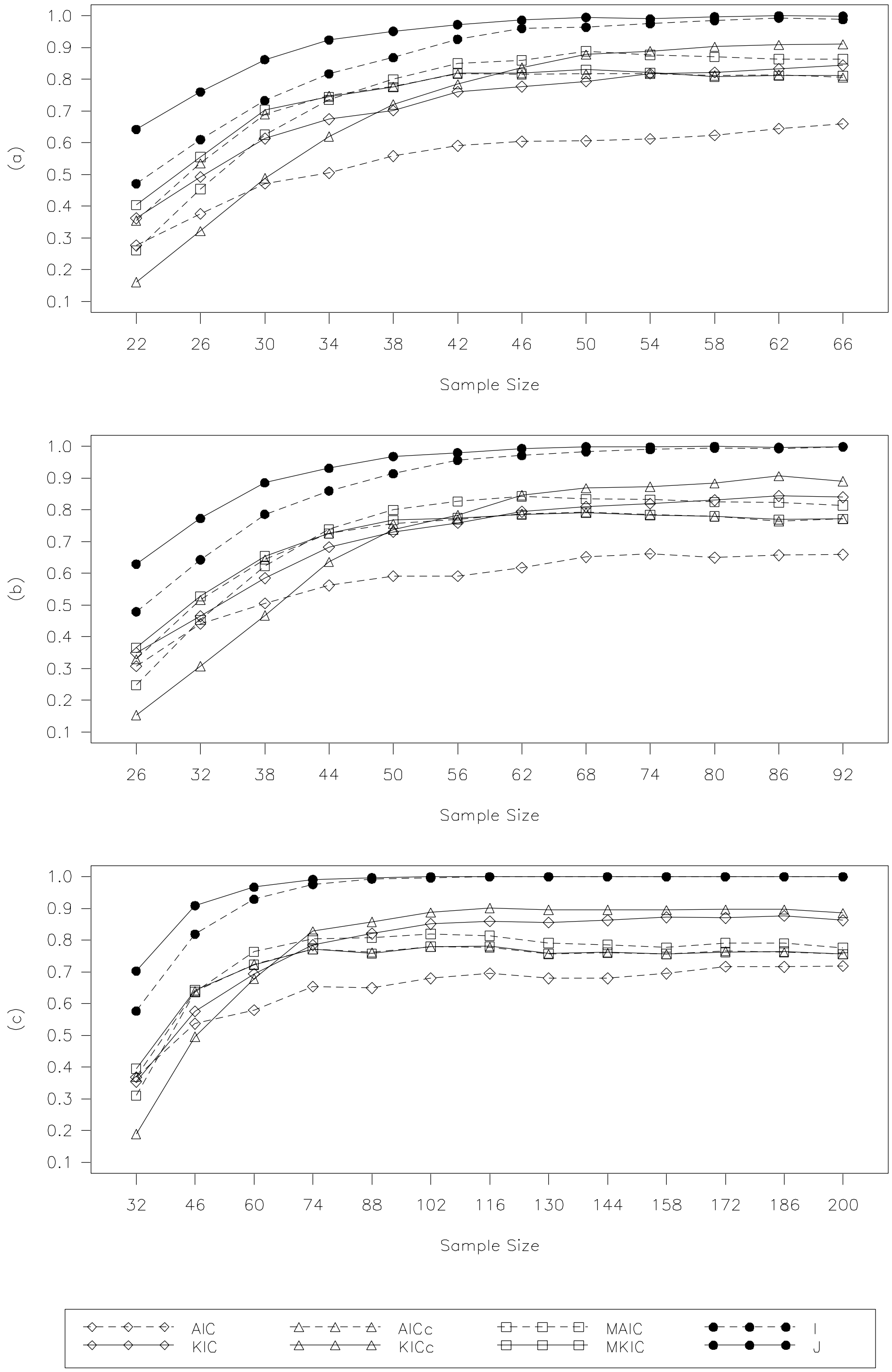

Figure 2: Proportion of correct order selections. (a) first NM collection (SNR = 1.0 : 1.0); (b) second NM collection (SNR = 1.0 : 1.5); (c) third NM collection (SNR = 1.0 : 2.0).

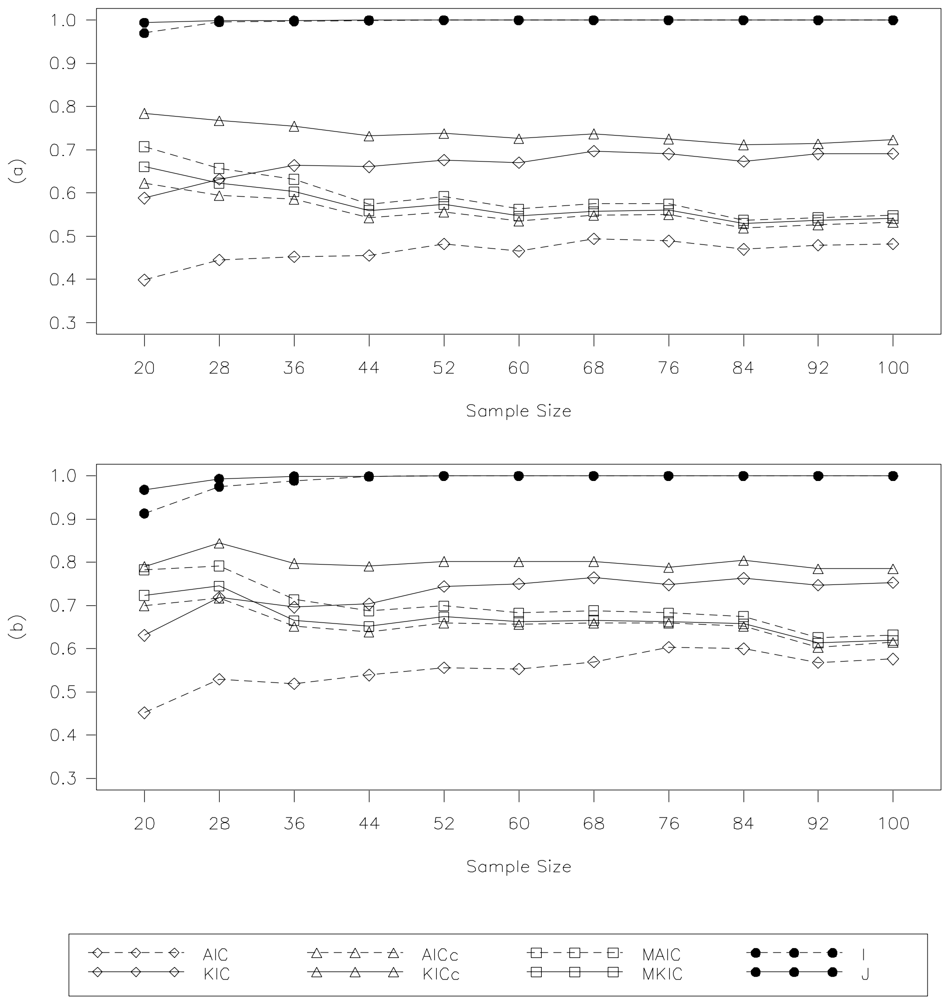

Figure 3: Proportion of correct order selections. (a) first APR collection (SNR = 1.0 : 1.0); (b) second APR collection (SNR = 1.0 : 0.5).

Appendix: Derivation of results from Section 3

Here, we outline the justifications for the estimators of

$$\mathrm{B}_2(p, \boldsymbol{\theta}_0) = \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)\right) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_p)\right) \tag{34}$$

which lead to the penalty terms of KICc and MKIC.

We consider a setting where data is generated from a true model of the form (1), and a candidate model of the form (2) is fit to the data. The following results will be used.

For any candidate model (2) which is fit to the data, we have

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left(\frac{n\hat{\sigma}_p^2}{\sigma_0^2}\right) = (n-p) + (\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)/\sigma_0^2, \tag{35}$$

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left((\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)/\sigma_0^2\right) = p + (\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)/\sigma_0^2. \tag{36}$$

Neglecting the log–likelihood constant $n \ln 2\pi$, we also have

$$\begin{aligned} d(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0) &= n \ln \sigma_0^2 + \frac{n\hat{\sigma}_p^2}{\sigma_0^2} + (\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)^\top(\boldsymbol{X}_0\boldsymbol{\beta}_0 - \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_p)/\sigma_0^2, & (37)\\ d(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_p) &= n \ln \hat{\sigma}_p^2 + n = -2 \ln f(\boldsymbol{y}|\, \hat{\boldsymbol{\theta}}_p), & (38)\\ d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) &= n \ln \sigma_0^2 + n. & (39) \end{aligned}$$

We begin with the justification of (26), which supports the estimator of (34) utilized by KICc.

Lemma 1. *Assume $f(\boldsymbol{y}|\, \boldsymbol{\theta}_0) \in \mathcal{F}(p)$. Then*

$$B_2(p, \boldsymbol{\theta}_0) = n \ln\left(\frac{n}{2}\right) - n\psi\left(\frac{n-p}{2}\right),$$

where ψ denotes the digamma function.

Proof. Note that when $f(\boldsymbol{y}|\, \boldsymbol{\theta}_0) \in \mathcal{F}(p)$, $(\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I} - \boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0) = 0$. Thus, by (35), (36), (37), and (38), we have

$$\begin{aligned} \mathrm{B}_2(p, \boldsymbol{\theta}_0) &= \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\hat{\boldsymbol{\theta}}_p, \boldsymbol{\theta}_0)\right) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(d(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_p)\right)\\ &= \mathrm{E}_{\boldsymbol{\theta}_0}\left(n \ln\left(\frac{\sigma_0^2}{\hat{\sigma}_p^2}\right)\right). \end{aligned}$$

Now the expectation of the log of a random variable with a central χ^2 distribution having *df* degrees of freedom is $\ln 2 + \psi(df/2)$. (See, for instance, McQuarrie & Tsai, 1998 p.67.) The result then follows from the fact that $(n\hat{\sigma}_p^2)/\sigma_0^2$ has a central χ^2 distribution with $(n-p)$ degrees of freedom.

We next present the justification of (31), which supports the estimator of (34) utilized by MKIC.

Let $\boldsymbol{X}_P$ denote the design matrix corresponding to the largest family in the candidate class, $\mathcal{F}(P)$, and let $\boldsymbol{H}_P$ denote the projection matrix onto the column space of $\boldsymbol{X}_P$. Recall the definitions of $\delta(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)$ and $b_2(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)$ as provided by (29) and (30).

Lemma 2. *Assume* $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(P)$. *Then*

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}_0}\left(\delta(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)\right) &= (\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I}-\boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0), & (40)\\ \mathrm{E}_{\boldsymbol{\theta}_0}\left(b_2(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\theta}}_P)\right) &= \mathrm{B}_2(p, \boldsymbol{\theta}_0) - d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0). & (41) \end{aligned}$$

Proof. Under the assumption $f(\boldsymbol{y}|\,\boldsymbol{\theta}_0) \in \mathcal{F}(P)$, the random variables

$$y^\top(\boldsymbol{H}_P - \boldsymbol{H}_p)y/\sigma_0^2 \qquad \text{and} \qquad y^\top(\boldsymbol{I}-\boldsymbol{H}_P)y/\sigma_0^2$$

are independent, the latter having a central χ^2 distribution with $(n-P)$ degrees of freedom, the former having a non–central χ^2 distribution with $(P-p)$ degrees of freedom and non–centrality parameter $(\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I}-\boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)/2\sigma_0^2$. Also, recall that the expectation of the reciprocal of a random variable with a central χ^2 distribution having *df* degrees of freedom is $1/(df-2)$. Using these facts, we can establish

$$\mathrm{E}_{\boldsymbol{\theta}_0}\left(\frac{n\hat{\sigma}_p^2}{\hat{\sigma}_P^2}\right) = n + \frac{n}{n-P-2}\left((P-p) + (\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I}-\boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)/\sigma_0^2\right).$$

Result (40) thereby follows.

Now using (34), (35), (36), (37), (38), and (39), we can easily verify that

$$\mathrm{B}_2(p, \boldsymbol{\theta}_0) = d(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) - \mathrm{E}_{\boldsymbol{\theta}_0}\left(-2\ln f(\boldsymbol{y}|\,\hat{\boldsymbol{\theta}}_p)\right) + 2(\boldsymbol{X}_0\boldsymbol{\beta}_0)^\top(\boldsymbol{I}-\boldsymbol{H}_p)(\boldsymbol{X}_0\boldsymbol{\beta}_0)/\sigma_0^2. \qquad (42)$$

Result (41) then follows via (42) and (40).

References

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

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 *Applied Time Series Analysis*, ed D. F. Findley, pp.1–23. New York: Academic Press.

Bernardo, J. M. (1976). Psi (digamma) function. *Applied Statistics* **25**, 315–317.

Bozdogan, H. (1990). On the information–based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics A* **19**, 221–278.

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* **44**, 333–344.

Fujikoshi, Y. & Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika* **84**, 707–716.

Hannan, E. J. & 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. & Tsai, C.–L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

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

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

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

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

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

McQuarrie, A., Shumway, R. & Tsai, C.–L. (1997). The model selection criterion AICu. *Statistics & Probability Letters* **34**, 285–292.

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

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.

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

Wei, C. Z. (1990). On predictive least squares principles. *Annals of Statistics* **20**, 1–42.