Received: 2 March 2018 | Revised: 27 December 2018 | Accepted: 28 January 2019

DOI: 10.1002/wics.1460

WILEY WIREs COMPUTATIONAL STATISTICS

FOCUS ARTICLE

The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements

Joseph E. Cavanaugh[1] | Andrew A. Neath[2]

[1]Department of Biostatistics, University of Iowa, Iowa City, Iowa

[2]Department of Mathematics and Statistics, Southern Illinois University, Edwardsville, Illinois

Correspondence
Joseph E. Cavanaugh, Department of Biostatistics, University of Iowa, Iowa City, Iowa
Email: joe-cavanaugh@uiowa.edu

The Akaike information criterion (AIC) is one of the most ubiquitous tools in statistical modeling. The first model selection criterion to gain widespread acceptance, AIC was introduced in 1973 by Hirotugu Akaike as an extension to the maximum likelihood principle. Maximum likelihood is conventionally applied to estimate the parameters of a model once the structure and dimension of the model have been formulated. Akaike's seminal idea was to combine into a single procedure the process of estimation with structural and dimensional determination. This article reviews the conceptual and theoretical foundations for AIC, discusses its properties and its predictive interpretation, and provides a synopsis of important practical issues pertinent to its application. Comparisons and delineations are drawn between AIC and its primary competitor, the Bayesian information criterion (BIC). In addition, the article covers refinements of AIC for settings where the asymptotic conditions and model specification assumptions that underlie the justification of AIC may be violated.

This article is categorized under:
Software for Computational Statistics > Artificial Intelligence and Expert Systems
Statistical Models > Model Selection
Statistical and Graphical Methods of Data Analysis > Modeling Methods and Algorithms
Statistical and Graphical Methods of Data Analysis > Information Theoretic Methods

KEYWORDS
AIC, Kullback–Leibler information, model selection criterion

1 | INTRODUCTION

One of the most daunting challenges in statistical modeling is to select a suitable model from a candidate collection to characterize the underlying data. Obviously, a careful development and formulation of the candidate collection is vital, and the successful fulfillment of this goal requires an appropriate study design, a substantive understanding of the dynamics and scientific underpinnings of the associated phenomenon, access to data that can be employed to capture the salient features of the phenomenon, and an appreciation of the interrelationships among key variables.

Model selection criteria provide a useful tool in identifying a model of appropriate structure and dimension among a candidate collection. A selection criterion assesses whether a fitted model offers an optimal balance between goodness-of-fit and parsimony. A fitted model that provides such a balance should also be generalizable, in that it should effectively describe or predict new data arising from the same phenomenon. In principle, a selection criterion

will disqualify candidate models that are either too simplistic to adequately accommodate the data or unnecessarily complex.

The Akaike information criterion (AIC) was the first model selection criterion to gain widespread attention in the statistical community, and continues to be one of the most widely known and used model selection tools in statistical practice. The criterion was introduced by Hirotugu Akaike (1973) in his seminal paper "Information Theory and an Extension of the Maximum Likelihood Principle." The traditional maximum likelihood framework, as applied to statistical modeling, provides a cogent paradigm for estimating the unknown parameters of a model having a specified dimension and structure. Akaike extended this paradigm by considering a setting in which the model size and structure are also unknown, and must therefore be determined from the data. Thus, Akaike proposed and developed a framework wherein both model estimation and selection could be simultaneously achieved.

For a parametric candidate model of interest, the likelihood function reflects the conformity of the model to the observed data. As the size and complexity of the model are increased, the model becomes more capable of adapting to the nuances and subtleties of the data. As a consequence, choosing the fitted model that maximizes the empirical likelihood will invariably lead to the selection of the largest and most complex model in the candidate collection. Model selection based on the likelihood principle, therefore, requires an extension of the traditional likelihood framework.

2 | BACKGROUND

To present the development of AIC, consider the following model selection setting. Suppose we endeavor to find a suitable model to characterize a collection of outcome measurements y. We will assume that y has been generated according to an unknown density $g(y)$, which will be deemed the *true* or *generating model.*

A model developed and formulated to describe the data y will be referred to as a *candidate* or *approximating model.* We will require that any candidate model structurally corresponds to a parametric class of probability distributions. Specifically, we will assume that a certain candidate model is represented by a k-dimensional parametric class of density functions

$$\mathcal{F}(k) = \{f(y|\theta_k) \mid \theta_k \in \Theta(k)\}.$$

Here, the parameter space $\Theta(k)$ is comprised of k-dimensional vectors whose components are functionally independent.

We will let $L(\theta_k|y)$ denote the likelihood corresponding to the density $f(y|\theta_k)$; that is, $L(\theta_k|y) = f(y|\theta_k)$. Accordingly, $\hat{\theta}_k$ will denote a vector of estimates obtained by maximizing the likelihood $L(\theta_k|y)$ over the parameter space $\Theta(k)$.

Suppose we develop and formulate a collection of candidate models of various structures and dimensions k. These models could be based on different subsets of explanatory variables, different mean and variance/covariance structures, and even different distributional specifications for the outcome variable. Our objective will be to search among this collection for the fitted model that provides the "best" approximation to the generating model $g(y)$. Such a fitted model will ideally capture the salient features of $g(y)$, yet may omit subtle features of marginal importance that cannot be accurately estimated based on the data at hand.

In the development of AIC, the adequacy of approximation is determined by employing a well-known measure that can be used in assessing the similarity between the generating model $g(y)$ and a candidate model $f(y|\theta_k)$: the *Kullback–Leibler information* (Kullback, 1968; Kullback & Leibler, 1951), also known as the *Kullback–Leibler divergence.* The Kullback–Leibler information between $g(y)$ and $f(y|\theta_k)$ with respect to $g(y)$ is defined as

$$I(\theta_k) = \mathrm{E}\left\{\log \frac{g(y)}{f(y|\theta_k)}\right\},$$

where $\mathrm{E}(\cdot)$ denotes the expectation under $g(y)$. Using Jensen's inequality, one can establish that $I(\theta_k) \geq 0$ with equality if and only if $f(y|\theta_k)$ and $g(y)$ are the same density. $I(\theta_k)$ is not a formal metric; however, we may conceptualize the measure in an analogous manner to a distance: that is, as the separation between $f(y|\theta_k)$ and $g(y)$ grows, the magnitude of $I(\theta_k)$ will generally increase to reflect this disparity.

In the present context, $I(\theta_k)$ may be interpreted as providing an average gauge of model conformity. For a particular sample y generated under the true model $g(y)$, the measure $\log\{g(y)/f(y|\theta_k)\}$ assesses how well the candidate model $f(y|\theta_k)$ accommodates or conforms to y in comparison with the true model $g(y)$. Thus, $I(\theta_k)$ evaluates the mean of the measure $\log\{g(y)/f(y|\theta_k)\}$ taken over repeated realizations y generated under the true model $g(y)$.

Next, we define

$$d(\theta_k) = \mathrm{E}\{-2\log f(y|\theta_k)\}. \tag{1}$$

We can then assert

$$2I(\theta_k) = d(\theta_k) - \mathrm{E}\{-2\log g(y)\}.$$

Because the measure $\mathrm{E}\{-2\ \log\ g(y)\}$ does not depend on θ_k, any ranking of a set of candidate models based on values of $I(\theta_k)$ would be equivalent to a ranking based on values of $d(\theta_k)$. Thus, for the purpose of delineating among various candidate models, $d(\theta_k)$ serves as an appropriate surrogate for $I(\theta_k)$. We will refer to $d(\theta_k)$ as the *Kullback discrepancy.*

In principle, the separation between a fitted candidate model $f\left(y|\hat{\theta}_k\right)$ and the generating model $g(y)$ could be assessed by using the Kullback discrepancy evaluated at $\hat{\theta}_k$:

$$d\left(\hat{\theta}_k\right) = \mathrm{E}\{-2\log f(y|\theta_k)\}|_{\theta_k=\hat{\theta}_k}.$$

Because $d\left(\hat{\theta}_k\right)$ depends on the true distribution $g(\cdot)$, this measure is inaccessible. However, the work of Akaike (1973, 1974) leads to an estimator of $d\left(\hat{\theta}_k\right)$ that is asymptotically unbiased under conditions pertaining to the propriety of the candidate model $f(y|\theta_k)$.

Akaike's development of AIC is based on using the statistic $-2\log f\left(y|\hat{\theta}_k\right)$ as a platform for the estimation of $d\left(\hat{\theta}_k\right)$. Although $-2\log f\left(y|\hat{\theta}_k\right)$ serves as a negatively biased estimator, the bias adjustment

$$\mathrm{E}\left\{d\left(\hat{\theta}_k\right)\right\} - \mathrm{E}\left\{-2\log f\left(y|\hat{\theta}_k\right)\right\} \tag{2}$$

can often be asymptotically estimated by twice the dimension of θ_k.

Since k denotes the dimension of θ_k, under appropriate conditions, the expected value of

$$\mathrm{AIC} \ = -2\log f\left(y|\hat{\theta}_k\right) + 2k$$

will asymptotically approach the expected value of $d\left(\hat{\theta}_k\right)$, say

$$\Delta(k) = \mathrm{E}\left\{d\left(\hat{\theta}_k\right)\right\}. \tag{3}$$

More technically, we will establish that

$$\mathrm{E}\{\mathrm{AIC}\} + o(1) = \Delta(k). \tag{4}$$

Thus, in suitable settings, AIC provides an asymptotically unbiased estimator of $\Delta(k)$. $\Delta(k)$ is often referred to as the *expected Kullback discrepancy.*

In AIC, the statistic based on the empirical log-likelihood, $-2\log f\left(y|\hat{\theta}_k\right)$, is called the *goodness-of-fit term.* This statistic reflects the conformity of the fitted model $f\left(y|\hat{\theta}_k\right)$ to the data used in its own construction, y. The bias correction $2k$ is called the *penalty term.* Models which are too simplistic to adequately characterize the data at hand will often yield large goodness-of-fit terms yet small penalty terms. On the other hand, models that conform well to the data, yet do so at the expense of containing unnecessary or extraneous parameters, will often yield small goodness-of-fit terms yet large penalty terms. Models that provide an optimal compromise between fidelity to the data and parsimony will ideally correspond to small AIC values, with the sum of the two AIC terms reflecting this balance.

3 | DERIVATION

To theoretically justify AIC as an asymptotically unbiased estimator of $\Delta(k)$, we will focus on a specific parametric class $\mathcal{F}(k)$. For notational simplicity, in the subsequent theoretical development, we will suppress the dimension index k on the parameter vector θ_k and its estimator $\hat{\theta}_k$.

Establishing the result (4) requires the strong assumption that the parametric class $\mathcal{F}(k)$ encompasses the true density $g(y)$. Under this assumption, we may define a "true" parameter vector θ_o that has the same dimension as θ, and express $g(y)$ using the parametric form $f(y|\theta_o)$. Here, θ_o would be an interior point of the parameter space $\Theta(k)$.

The requirement that $f(y|\theta_o) \in \mathcal{F}(k)$, which facilitates mathematical tractability, implies that the candidate model is either correctly specified or overspecified. From a practical standpoint, such a requirement seems inherently problematic, since the

candidate collection would obviously include models that are underspecified. The practical ramifications of this condition are later discussed.

To justify the asymptotic result (4), consider representing $\Delta(k)$ as follows:

$$\begin{aligned}
&\Delta(k)\\
&= \mathrm{E}\{d(\hat{\theta})\}\\
&= \mathrm{E}\{-2\log f(y|\hat{\theta})\}\\
&+ \left[\mathrm{E}\{-2\log f(y|\theta_o)\} - \mathrm{E}\{-2\log f(y|\hat{\theta})\}\right] \qquad (5)\\
&+ \left[\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\theta_o)\}\right]. \qquad (6)
\end{aligned}$$

The following lemma establishes that (5) and (6) are both within $o(1)$ of k. The proof largely follows the development in Cavanaugh (1997).

Henceforth, we require a set of regularity conditions that will ensure the consistency and asymptotic normality of the maximum likelihood vector $\hat{\theta}$.

Lemma

$$\mathrm{E}\{-2\log f(y|\theta_o)\} - \mathrm{E}\{-2\log f(y|\hat{\theta})\} = k + o(1), \qquad (7)$$

$$\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\theta_o)\} = k + o(1). \qquad (8)$$

Proof To begin, we define

$$\mathrm{I}(\theta) = \mathrm{E}\left[-\frac{\partial^2 \log f(y|\theta)}{\partial\theta\partial\theta'}\right] \text{ and } \mathcal{J}(\theta, y) = \left[-\frac{\partial^2 \log f(y|\theta)}{\partial\theta\partial\theta'}\right]. \qquad (9)$$

Thus, $\mathrm{I}(\theta)$ denotes the *expected Fisher information matrix* and $\mathcal{J}(\theta, y)$ denotes the *observed Fisher information matrix.*

The justification of the result hinges on two second-order Taylor series expansions: the first which expands $-2\log f(y|\theta_o)$ about the point $\hat{\theta}$, and the second which expands $d(\hat{\theta})$ about θ_o.

First, consider taking a second-order expansion of $-2\log f(y|\theta_o)$ about $\hat{\theta}$, and evaluating the expectation of the result. Because $-2\log f(y|\theta)$ is minimized at $\theta = \hat{\theta}$, the gradient of this function is zero when evaluated at $\theta = \hat{\theta}$. Thus, the first-order term in the expansion disappears, resulting in

$$\mathrm{E}\{-2\log f(y|\theta_o)\} = \mathrm{E}\{-2\log f(y|\hat{\theta})\} + \mathrm{E}\left\{(\hat{\theta}-\theta_o)'\{\mathcal{J}(\hat{\theta},y)\}(\hat{\theta}-\theta_o)\right\} + o(1).$$

Therefore, we obtain the approximation

$$\mathrm{E}\{-2\log f(y|\theta_o)\} - \mathrm{E}\{-2\log f(y|\hat{\theta})\} = \mathrm{E}\left\{(\hat{\theta}-\theta_o)'\{\mathcal{J}(\hat{\theta},y)\}(\hat{\theta}-\theta_o)\right\} + o(1). \qquad (10)$$

Next, consider taking a second-order expansion of $d(\hat{\theta})$ about θ_o, again evaluating the expectation of the result. Because $d(\theta)$ is minimized at $\theta = \theta_o$, the gradient of this function is zero when evaluated at $\theta = \theta_o$. Thus, the first-order term in the expansion again disappears, resulting in

$$\mathrm{E}\{d(\hat{\theta})\} = \mathrm{E}\{-2\log f(y|\theta_o)\} + \mathrm{E}\left\{(\hat{\theta}-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}-\theta_o)\right\} + o(1).$$

We thereby obtain the approximation

$$\mathrm{E}\{d(\hat{\theta})\} - \mathrm{E}\{-2\log f(y|\theta_o)\} = \mathrm{E}\left\{(\hat{\theta}-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}-\theta_o)\right\} + o(1). \qquad (11)$$

By assumption, the true parameter vector θ_o is an interior point of the parameter space $\Theta(k)$. Therefore, each of the quadratic forms

$$\left(\hat{\theta}-\theta_o\right)'\left\{\mathcal{J}\left(\hat{\theta},y\right)\right\}\left(\hat{\theta}-\theta_o\right) \text{ and } \left(\hat{\theta}-\theta_o\right)'\left\{\mathrm{I}(\theta_o)\right\}\left(\hat{\theta}-\theta_o\right)$$

converge to centrally distributed chi-square random variables with k degrees of freedom. The expectations of both quadratic forms are thereby within $o(1)$ of k. This result along with the approximations (10) and (11) establish the results of the lemma, (7) and (8).

4 | PROPERTIES

For candidate models that are correctly specified or overspecified, the previous lemma justifies AIC as an asymptotically unbiased estimator of the expected Kullback discrepancy $\Delta(k)$. From a practical standpoint, AIC estimates $\Delta(k)$ with negligible bias in settings where the sample size n is large and the model dimension k is relatively small. However, in settings where n is small and k is relatively large (e.g., $k \approx n/2$), $2k$ is often much smaller than the bias adjustment, meaning that AIC is substantially negatively biased as an estimator of $\Delta(k)$.

In small to moderate sample size applications where the candidate collection includes models of high dimension, AIC may severely underestimate $\Delta(k)$ for the larger fitted models. As a consequence, the criterion may favor the larger models even when the expected discrepancy between these models and the generating model is high relative to simpler, more parsimonious models. This phenomenon is illustrated in Linhart and Zucchini (1986, pp. 86–88), who caution (p. 78) that "in some cases the criterion simply continues to decrease as the number of parameters in the approximating model is increased."

The complexity penalization for a model selection criterion is governed by the dimension k, yet the behavior of any criterion depends crucially on the form in which the dimension enters the penalty term. Based on its penalization, AIC is *asymptotically efficient* in the sense of Shibata (1980, 1981); however, it is not *consistent.* To define the latter, suppose that the generating model is of finite dimension, and that the dimension and structure of this model are represented among the candidate models under consideration. A consistent criterion will asymptotically select the fitted candidate model having the correct size and structure with probability one. On the other hand, suppose that the generating model is of an infinite dimension, and therefore lies outside of the collection of candidate models under consideration. An asymptotically efficient criterion will asymptotically select the fitted candidate model that is predictively optimal, in that the model-based predictors minimize the mean squared error of prediction.

From a theoretical perspective, asymptotic efficiency is arguably the strongest optimality property of AIC. However, the property is somewhat surprising. Establishing the asymptotic unbiasedness of AIC as an estimator of the expected Kullback discrepancy requires that the candidate model of interest subsumes the true model. Yet the asymptotic efficiency of AIC ensures that in large-sample settings, even when none of the candidate models are of the correct dimension and structure, the criterion will choose the fitted model that most effectively predicts new data.

5 | APPLICATION

AIC is applicable in a broad array of modeling frameworks, since its justification only requires conventional large-sample properties of maximum likelihood estimators. The successful application of AIC is not contingent on each of the models in the candidate collection closely approximating the generating model $g(y)$, although as previously mentioned, the derivation may imply otherwise. This attribute of AIC is further discussed in what follows.

The framework of hypothesis testing is often used for model comparison. Likelihood ratio test (LRT) procedures are ubiquitous for this purpose. In most testing settings, including that for the LRT, the null model is nested within a larger alternative model. The latter model is assumed to be true, and the test is conducted to determine whether the simpler, more parsimonious null model can also be deemed suitable. AIC, however, can be used to compare non-nested models. As emphasized by Burnham and Anderson (2002, p. 88) "A substantial advantage in using information-theoretic criteria is that they are valid for non-nested models. Of course, traditional likelihood ratio tests are defined only for nested models, and this represents another substantial limitation in the use of hypothesis testing in model selection."

AIC can also be used to compare models based on different probability distributions for the outcome variable: for example, normal versus gamma, Poisson versus negative binomial. However, if the models in the candidate collection are based on different distributions, then all of the terms in each empirical likelihood must be retained when the values of AIC are evaluated, including constants that are not data dependent. (If the models in the candidate collection are each based on the same distribution, such terms may be discarded in AIC computations.) This property of AIC is particularly useful in applications where an appropriate distribution must be determined for the outcome, in addition to the model size and structure. For this reason, AIC is ideally suited to generalized linear modeling applications.

AIC cannot be used to compare models based on different transformations of the outcome variable: for example, log versus square root. Thus, the criterion cannot be used to select an optimal transformation.

In routine model selection applications, the optimal fitted model is identified by the minimum value of AIC. However, the criterion values are important; models with similar values should receive the same "ranking" in assessing criterion preferences. A common rule of thumb is to treat any fitted models that yield an AIC value within two units of the minimum AIC value as viable candidates. Burnham and Anderson (2002, p. 70) feature guidelines for practitioners for the assessment and interpretation of AIC differences (Table 1).

The Bayesian information criterion (BIC) was introduced by Schwarz (1978) as a competitor to AIC. Schwarz derived BIC to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. In large-sample settings, the fitted model favored by BIC ideally corresponds to the candidate model which is a posteriori most probable; that is, the model which is rendered most plausible by the data at hand.

AIC and BIC share the same goodness-of-fit term, but the penalty terms differ based on the manner in which the dimension k is incorporated: BIC employs a complexity penalization of $k \log n$ as opposed to $2k$. Consequently, BIC tends to choose fitted models that are more parsimonious than those favored by AIC. The differences in selected models may be especially pronounced in large-sample settings.

From a practical perspective, AIC and BIC might be distinguished as follows. AIC could be advocated when the primary goal of the modeling application is *predictive*; that is, to build a model that will effectively predict new outcomes. BIC could be advocated when the primary goal of the modeling application is *descriptive*; that is, to build a model that will feature the most meaningful factors influencing the outcome, based on an assessment of relative importance. As the sample size grows, predictive accuracy improves as subtle effects are admitted to the model. AIC will increasingly favor the inclusion of such effects; BIC will not.

The predictive/descriptive modeling delineation of AIC and BIC may be viewed as a practical manifestation of the large-sample optimality properties mentioned in the previous section. AIC is asymptotically efficient yet not consistent, whereas BIC is consistent yet not asymptotically efficient. Despite this delineation, as discussed in the next section, both AIC and BIC can be interpreted in a predictive context.

6 | PREDICTIVE INTERPRETATION

George Box is famously attributed with the quote "All models are wrong, some are useful." This quote reflects the statistical philosophy that in most applications, models can only be expected to approximate reality, and cannot be expected to precisely capture or mirror reality. For this reason, some statisticians posit that the best platform for assessing the propriety of a fitted model is generalizability: the ability of a fitted model to accurately describe or predict new data.

The justification of AIC can be reframed to emphasize the predictive objective. Suppose that future data z is to be generated from the true model $g(\cdot)$, independent of the observed data y. Consider the problem of employing y to fit a model for the prediction of z. Thus, in the present context, y may be viewed as a fitting or training sample and z as a validation sample.

The model $f\left(z|\hat{\theta}_k\right)$, where $\hat{\theta}_k$ is computed from y, serves as a predictive density for z. Rather than merely providing point predictions, or interval predictions, the predictive density allows for a general accounting of prediction uncertainty.

The relative magnitude of the predictive density $f\left(z|\hat{\theta}_k\right)$ provides a measure of plausibility across the sample space of possible outcomes. Predictive accuracy can then be assessed by $f\left(z|\hat{\theta}_k\right)$; the larger the value of the predictive density, the more accurate the prediction of the new data z. Conversely, we can define a measure of *prediction error* through the quantity $-2\log f\left(z|\hat{\theta}_k\right)$. Under this framework, the mean prediction error is equivalent to the expected Kullback discrepancy $\Delta(k)$. That is,

$$\Delta(k) = \mathrm{E}\left\{-2\log f\left(z|\hat{\theta}_k\right)\right\}. \tag{12}$$

TABLE 1 Strength of evidence provided in support of a fitted candidate model based on the difference between the model's Akaike information criterion (AIC) value and the minimum AIC value

$AIC_i - AIC_{min}$	Level of empirical support for model i
0–2	Substantial
4–7	Considerably less
>10	Essentially none

The introduction of the validation data z removes the need to define $\Delta(k)$ through the iterated expectation implied by (1) and (3) in that $\hat{\theta}_k$ depends on the original data y, independent of z. In the definition of $\Delta(k)$ provided by (12), the expectation averages over the joint distribution of both the fitting sample y and the validation sample z.

The development leading to (3) suggests that the fitted candidate model that best predicts new data z generated under $g(\cdot)$ is the model for which $\Delta(k)$ is minimized. Not surprisingly, this fitted model is also closest to the truth $g(\cdot)$ in the sense of Kullback–Leibler information, as implied by the development leading to (12).

However, additional insights can be garnered by viewing the model selection problem from the perspective of prediction. The statistic $-2\log f\left(y|\hat{\theta}\right)$ is observable, yet underestimates the expected prediction error $\Delta(k)$, since the same data used to fit the model that yields the predictions is also being used to judge predictive accuracy. Access to new data is typically needed to provide an unbiased estimator of the prediction error. Yet we have established that AIC provides an asymptotically unbiased estimator of $\Delta(k)$. The derivation of AIC leads to the introduction of a penalty term necessary to counteract the "double use" of the data y, allowing for an estimate of the prediction error that does not require the availability of new data.

The predictive perspective on model selection may be further illuminated by reflecting on how the AIC predictive objective compares to that of BIC. In the Bayesian framework, a model M_k is specified by both a likelihood function $f(y|\theta_k)$ and a prior distribution $p(\theta_k)$ over the parameter space $\Theta(k)$. Then,

$$f(y \mid M_k) = \int f(y|\theta_k)\, p(\theta_k)\, d\theta_k$$

serves as the *prior predictive density* under the Bayes model M_k. The goal for BIC is to select the model which is a posteriori most probable, given the observed data y. The posterior probability on model M_k arises as

$$\pi(M_k \mid y) = \frac{\pi_k\, f(y \mid M_k)}{\sum_{l=1}^{L} \pi_l\, f(y \mid M_l)},$$

where $\pi_1, \ldots, \pi_L$ are the prior probabilities on the candidate models $M_1, \ldots, M_L$..

BIC is derived under a prior with equal model probabilities $\pi_1 = \ldots = \pi_L$. The posterior probability $\pi(M_k \mid y)$ is thereby maximized when $f(y \mid M_k)$ is maximized over the candidate collection, for the observed y. Therefore, model selection under BIC can also be motivated from a predictive viewpoint. Each candidate model puts forth a predictive density. The selected model is the one for which the observed data y is most accurately predicted. Because the predictive density is specified a priori, no difficulties arise from using the same data to fit and validate the prediction model.

The differing perspectives on the role of prediction in model selection illustrate how the impetus for AIC satisfies frequentist objectives while the motivation for BIC fulfills Bayesian objectives. With AIC, predictive optimality is sought over repeated (future) observations; with BIC, predictive optimality is based on the data at hand. Because we conceptualize model selection under AIC as an attempt to identify the model that will most accurately reflect future data arising from the same phenomenon as the observed data, AIC-based model selection is arguably better aligned with the goals of replicability in scientific research problems.

7 | REFINEMENTS

A number of AIC variants have been proposed and developed since the introduction of the original criterion. In general, these variants have been designed to achieve either or both of two objectives: (a) to relax the assumptions or expand the setting under which the criterion can be applied and (b) to improve the small-sample performance of the criterion.

In the Gaussian linear regression framework, Sugiura (1978) established that the bias adjustment (2) can be exactly evaluated for correctly specified or overspecified models. The resulting criterion, with a refined penalty term, is known as "corrected" AIC, or AICc. Hurvich and Tsai (1989) extended AICc to the frameworks of Gaussian nonlinear regression models and time series autoregressive models. Subsequent work has extended AICc to other modeling frameworks, such as Gaussian multivariate linear regression models (Bedrick & Tsai, 1994), autoregressive moving average models (Hurvich, Shumway, & Tsai, 1990), vector autoregressive models (Hurvich & Tsai, 1993), and certain generalized linear models and linear mixed models (Azari, Li, & Tsai, 2006; Hurvich & Tsai, 1995).

As noted by Cavanaugh (1997), "AIC is justified in a very general framework, and as a result, offers a crude estimator of the expected discrepancy: one which exhibits a potentially high degree of negative bias in small-sample applications. AICc corrects for this bias, but is less broadly applicable than AIC since its justification depends upon the form of the candidate model."

The penalty terms of AIC and AICc are asymptotically equivalent. However, in small to moderate sample size applications, AICc generally estimates the expected discrepancy $\Delta(k)$ with substantially less bias than AIC, and thereby guards against the propensity of AIC to favor models of an inappropriately high dimension. For this reason, the routine use of AICc may be advisable for those modeling frameworks in which the corrected criterion has been developed and justified.

The Takeuchi (1976) information criterion (TIC), was derived by obtaining a general, large-sample approximation to each of (5) and (6) that does not rely on the assumption that the true density $g(y)$ is a member of the parametric class $\mathcal{F}(k)$. The resulting approximation is given by the trace of the product of two matrices: an information matrix based on the score vector, and the inverse of the expected Fisher information matrix. Specifically, let

$$\mathrm{J}(\theta) = \mathrm{E}\left[\left\{\frac{\partial \log f(y|\ \theta)}{\partial \theta}\right\}\left\{\frac{\partial \log f(y|\ \theta)}{\partial \theta}\right\}'\right].$$

The penalty term of TIC is based on an estimator of

$$2\ \text{trace}\ \left\{\mathrm{J}(\hat{\theta})\left[\mathrm{I}(\hat{\theta})\right]^{-1}\right\},$$

where $\mathrm{I}(\theta)$ is as defined in (9). Under the assumption that $g(y) \in \mathcal{F}(k)$, the information matrices $\mathrm{J}(\theta_o)$ and $\mathrm{I}(\theta_o)$ are equivalent. Thus, in large-sample settings, the trace is close to k, and the penalty term of TIC essentially reduces to that of AIC.

An investigative comparison of the penalty terms of AIC and TIC helps to illuminate why AIC may be used to delineate candidate models that lack requisite structure from those that are sufficient to adequately approximate the salient features of the generating model $g(y)$. For the latter fitted candidate models, the penalty terms of TIC and AIC should be similar. For the former fitted candidate models, the penalty term of AIC could be quite biased relative to that of TIC; yet in such settings, the accuracy of the data-based TIC penalization could be poor due to high variability. However, the goodness-of-fit terms for underspecified models will tend to be much larger than those for adequately specified models, and the magnitude of the differences in these terms should be much greater than the magnitude of the differences in the bias adjustments. For this reason, the more sophisticated, data-based complexity penalization of TIC may not always yield a discernable practical advantage over the more simplistic penalization of AIC. For further discussion, see Kitagawa (1987, pp. 1060–1062).

Bozdogon (1987) proposed a variant of AIC that corrects for its lack of consistency. The variant, called CAIC, has a penalty term defined as $(k \log n + k)$. The initial component of this term, $k \log n$, is the same as the penalty term of BIC, and dominates the remaining component k in large-sample settings. For this reason, in such settings, the selection behaviors of CAIC more closely mirror that of BIC than AIC. The consistency of CAIC follows from that of BIC; however, CAIC is not asymptotically efficient.

Pan (2001) introduced a variant of AIC for applications in the framework of generalized linear models fitted using generalized estimating equations. The criterion is called QIC, since the goodness-of-fit term is based on the empirical quasi-likelihood.

Konishi and Kitagawa (1996) extended the setting in which AIC has been developed to a general framework where (a) the method used to fit the candidate model is not necessarily maximum likelihood, and (b) the true density $g(y)$ is not necessarily a member of the parametric class $\mathcal{F}(k)$. Their resulting criterion is called the generalized information criterion (GIC). The penalty term of GIC reduces to that of TIC when the fitting method is maximum likelihood.

AIC variants based on computationally intensive methods have also been proposed, including cross-validation (Davies, Neath, & Cavanaugh, 2005; Stone, 1977), bootstrapping (Cavanaugh & Shumway, 1997; Ishiguro, Sakamoto, & Kitagawa, 1997; Neath, Cavanaugh, & Reidle, 2012; Shibata, 1997), and Monte Carlo simulation (Bengtsson & Cavanaugh, 2006; Hurvich, Shumway, & Tsai, 1990). The justifications for these AIC variants are strongly tied to the predictive impetus for AIC presented in the previous section, in that the developments attempt to provide a surrogate for the new data z and evaluate the efficacy of the fitted model in predicting this data.

With the cross-validatory variants, if an n-fold approach is employed, the cases in the fitting sample y are sequentially deleted, and each case-deleted fitted model is used to predict the deleted case. With the bootstrap variants, fitted models based on bootstrap samples drawn from y are repeatedly used to predict the original sample y, which serves as a validation sample.

As with TIC, cross-validatory and bootstrap AIC variants may be justified without employing the assumption that the true density $g(y)$ is a member of the parametric class $\mathcal{F}(k)$ (Cavanaugh, Davies, & Neath, 2008; Shibata, 1997). Despite the computational expense required for their evaluation, these variants tend to outperform both TIC and traditional AIC in small to moderate sample size applications.

With the Monte Carlo simulation variants, an approximate bias adjustment (2) is simulated by noting

$$\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2\log f(y|\hat{\theta}_k)\} = \mathrm{E}\{-2\log f(z|\ \hat{\theta}_k)\} - \mathrm{E}\{-2\log f(y|\ \hat{\theta}_k)\}.$$

Both fitting samples y and validation samples z are repeatedly generated under a predetermined, simplistic baseline model, and an approximate adjustment is evaluated by computing replicated differences

$$\left\{-2\log f\left(z|\hat{\theta}_k\right)\right\}-\left\{-2\log f\left(y|\hat{\theta}_k\right)\right\}.$$

An approximate bias adjustment is evaluated by averaging the replicated differences.

As with AIC and AICc, the justifications of Monte Carlo variants of AIC require that the true density $g(y)$ is a member of the parametric class $\mathcal{F}(k)$. Monte Carlo AIC variants often behave similarly to AICc, yet may be conveniently applied in modeling frameworks where the development of an AICc penalization is problematic (Bengtsson & Cavanaugh, 2006).

8 | ADDITIONAL CRITERIA/COMPLEXITY PENALIZATION

Although AIC is arguably the most pervasively known and utilized model selection criterion, since its introduction, a vast array of selection criteria have been developed and investigated. Some of the most familiar alternatives to AIC were initially proposed for variable determination within the context of traditional linear regression.

A popular yet controversial practice in linear regression is to select the fitted model that corresponds to the minimum mean squared error (MSE). This is equivalent to choosing the fitted model corresponding to the maximum adjusted coefficient of determination, R^2_{adj}. If p denotes the rank of the design matrix for a candidate model, then MSE is defined as the sum of squared residuals (SSE), divided by the associated degrees of freedom, $(n-p)$. The adjusted coefficient of determination can then be defined in terms of MSE as

$$R^2_{\text{adj}}=1-\frac{\text{MSE}}{\text{SST}/(n-1)},$$

where SST denotes the total sum of squares.

Although the addition of variables to a candidate model can only decrease SSE, admitting variables that only marginally diminish SSE can increase MSE. However, choosing the candidate model that corresponds to the minimum MSE offers no discernable protection from overfitting. This behavior is evident based on the expected value of the statistic. If σ^2_0 denotes the error variance of the generating model, then $\text{E}\{\text{MSE}\}=\sigma^2_0$ for correctly specified and overspecified models. On the other hand, $\text{E}\{\text{MSE}\}>\sigma^2_0$ for underspecified models, with the size of $\text{E}\{\text{MSE}\}$ corresponding to the overall importance of the omitted variables. Thus, underfitted models should produce inflated values of MSE, but overfitted models should yield roughly the same values of MSE as the fitted generating model, regardless of the degree of overspecification.

One of the most popular criteria for variable selection in linear regression is the conceptual predictive statistic, C_p, proposed by Mallows (1973). Mallows developed C_p as an approximately unbiased estimator of an expected discrepancy based on a sum of squared prediction errors, scaled by the variance σ^2_0. The statistic is defined as

$$\text{C}_p=\frac{\text{SSE}}{\text{MSE}_*}-n+2p,$$

where MSE_* denotes the mean squared error for the largest fitted model in the candidate collection. (In the justification of C_p, it is assumed that the largest model is correctly specified or overspecified, meaning that $\text{E}\{\text{MSE}_*\}=\sigma^2_0$.)

One of the reasons for the enduring popularity of C_p is its ease of interpretability: if the fitted model is correctly specified or overspecified, C_p will tend to be close to p; yet if the fitted model is underspecified, then C_p will exceed p, with the excess governed by the aggregate importance of the variable exclusions. Thus, unlike AIC or BIC, where only relative comparisons of the criterion values are meaningful, values of C_p can be interpreted in isolation.

An analogue of C_p based on n-fold cross validation is the prediction sum of squares statistic (PRESS) (Allen, 1974). PRESS is defined as the sum of squared case-deleted residuals (i.e., the sum of the n squared residuals resulting from using each case-deleted fitted model to predict the deleted case). PRESS estimates the same expected discrepancy as C_p, only without the scale factor σ^2_0.

Unlike AIC, BIC, and the variants of AIC mentioned in the previous section, C_p, PRESS, and $\text{MSE}/R^2_{\text{adj}}$ are not developed in a likelihood-based paradigm. However, in the traditional regression framework, assuming normal errors, the goodness-of-fit term of likelihood-based criteria has a simplistic form that facilitates illuminating comparisons:

$$-2\log f\left(y|\hat{\theta}\right)=n\log\left(\text{SSE}/n\right)+n(\log 2\pi+1).$$

TABLE 2 Large-sample complexity penalization of popular selection criteria in the Gaussian linear regression framework

Criterion	Large–sample complexity penalization
MSE, R^2_{adj}	k
AIC, AICc, TIC, C_p, PRESS	$2k$
CAIC	$k \log n + k$
BIC	$k \log n$

For MSE, R^2_{adj}, TIC, C_p, and PRESS, the characterization is based on operationally equivalent criteria of the form $-2\log f\left(y|\hat{\theta}_k\right) + a_n\, k$. The results assume the existence of models in the candidate collection that are adequately specified.

Note that the only statistic involved in the goodness-of-fit term is SSE, which also appears in the definitions of C_p, MSE, and R^2_{adj}. In fact, in large-sample traditional regression settings where the candidate collection consists of at least some models that are adequately specified (i.e., correctly specified or overspecified), operationally equivalent criteria of the form

$$-2\log f\left(y|\hat{\theta}\right) + a_n k \tag{13}$$

can be derived for most selection criteria, regardless of whether they are inherently likelihood-based. Here, a_n represents a sequence that may depend on the sample size n. These operationally equivalent criteria will exhibit the same selections asymptotically as their counterparts.

Based on the representation (13), Table 2 features the large-sample complexity penalizations of the criteria MSE/R^2_{adj}, C_p, PRESS, AICc, and TIC, as well as the complexity penalizations of AIC, BIC, and CAIC.

The criteria in Table 2 that feature a complexity penalization of $2k$ (large-sample or otherwise) are asymptotically efficient. In practice, these criteria should be favored in predictive applications, where the primary objective is to select a fitted model that will most accurately approximate new data arising from the same phenomenon as the observed data. The criteria that feature a complexity penalization involving $k \log n$ are consistent. These criteria might be preferred in applications where the principal goal is to select a fitted model that will include the most substantive and salient factors that govern the dynamics of the outcome. The criteria that feature a large-sample complexity penalization of k should only be used if overfitting is not a concern, and rather, the overall focus pertains to avoiding models that are too simplistic to adequately characterize the underlying phenomenon.

9 | CONCLUSION

The Akaike information criterion is a widely used tool in model selection, which can be effectively applied in many large-sample settings where a collection of candidate models is fit using maximum likelihood. The objective of the criterion is to identify the fitted candidate model that is closest to the generating model in the sense of Kullback–Leibler information. AIC is particularly well suited to predictive modeling applications. Variants of AIC have been developed to expand the applicability of the criterion and to improve its small-sample performance.

CONFLICT OF INTEREST

The authors declare no conflicts of interest for this article.

RELATED WIREs ARTICLES

The Bayesian information criterion: Background, derivation, and applications
Variable selection in linear models

FURTHER READING

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, England: University Press.
Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. New York, NY: Springer.
Lahiri, P. (Ed.). (2001). *Model selection. Institute of mathematical statistics lecture notes—Monograph series* (Vol. 18). Beachwood, OH: Institute of Mathematical Statistics.

REFERENCES

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

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

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, *16*, 125–127.

Azari, R., Li, L., & Tsai, C. L. (2006). Longitudinal data model selection. *Computational Statistics and Data Analysis*, *50*, 3053–3066.

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

Bengtsson, T., & Cavanaugh, J. E. (2006). An improved Akaike information criterion for state-space model selection. *Computational Statistics and Data Analysis*, *50*, 2635–2654.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multi-model inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

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

Cavanaugh, J. E., Davies, S. L., & Neath, A. A. (2008). Discrepancy-based model selection criteria using cross validation. In F. Vonta, M. Nikulin, N. Limnios, & C. Huber (Eds.), *Statistical models and methods for biomedical and technical systems* (pp. 473–485). Boston, MA: Birkhaüser.

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

Davies, S. L., Neath, A. A., & Cavanaugh, J. E. (2005). Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy. *Statistical Methodology*, *2*, 249–266.

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

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

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

Hurvich, C. M., & Tsai, C. L. (1995). Model selection for extended quasi–likelihood models in small samples. *Biometrics*, *51*, 1077–1084.

Ishiguro, M., Sakamoto, Y., & Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics*, *49*, 411–434.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series: Rejoinder. *Journal of the American Statistical Association*, *82*, 1060–1063.

Konishi, S., & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, *83*, 875–890.

Kullback, S. (1968). *Information theory and statistics*. New York, NY: 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, NY: Wiley.

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

Neath, A. A., Cavanaugh, J. E., & Reidle, B. (2012). A bootstrap method for assessing uncertainty in Kullback–Leibler discrepancy model selection problems. In I. Vonta & A. Karagrigoriou (Eds.), *Mathematics in engineering, science and aerospace; special issue on "Theory and applications of divergence and information measures"* (Vol. 3, pp. 381–391). Cambridge, UK: Cambridge Scientific Publishers.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*, 120–125.

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

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

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, *68*, 45–54.

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

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B*, *39*, 44–47.

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

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Sciences*, *153*, 12–18 in Japanese.

How to cite this article: Cavanaugh JE, Neath AA. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Comput Stat*. 2019;e1460. https://doi.org/10.1002/wics.1460