Understanding information criteria for model selection

Tom Loredo

30 Aug 2010; minor edits 16 Jun 2014

There are many methods for selecting a "best" parametric model from among a set of competing models for data. The variety of methods reflects diversity of the goals for modeling (e.g., to identify a good explanatory hypothesis for the available data, or to predict future data based on available data), diversity in approaches to statistical inference (e.g., frequentist, Bayesian, complexity theory), and diversity in approximations invoked when formal methods require unavailable information or are computationally expensive. The most-used methods are approximate methods, dubbed *information criteria*. They are often described in terms of a tradeoff between goodness-of-fit and model complexity, but this is usually just a post hoc interpretation of the results of calculations based on other, more precise criteria. Lahiri (2001) collects articles reviewing the state of model selection from the perspective of statistics as of the turn of the century; in that collection, Rao and Wu (2001) survey over a dozen information criteria, situating them among more formal (and computationally demanding) methods.

In current applied modeling in the physical sciences and engineering, three information criteria are dominant. Two are decades-old classic criteria: the Akaike information criterion (AIC), and the Bayesian information criterion (BIC, also known as the Schwarz criterion); see Lahiri (2001) for references and Stoica and Selén (2004) for a review focusing on the AIC and BIC. The third is a relative newcomer whose foundations and performance are not yet fully understood: the deviance information criterion (DIC). To understand the complementary roles of these criteria, we start by defining a key ingredient in all of them, the *likelihood function* for the parameters of a model.

Let $\mathcal{M}_i$ denote a specific parametric model in a set $\{\mathcal{M}_i\}$ whose members are labeled by the integer i; it has parameters that we denote jointly by θ_i (a vector). A parametric model allows one to statistically predict values of data, D, via the *sampling distribution*, $p(D|\theta_i, \mathcal{M}_i)$: the probability that the data will have the values D if both the model and its parameter values are specified. This is a function of both the possible values of the data, and the possible values of the parameters. The likelihood function $\mathcal{L}_i(\theta_i)$ for the parameters of $\mathcal{M}_i$ is defined by $\mathcal{L}_i(\theta_i) \equiv p(D_{\rm obs}|\theta_i, \mathcal{M}_i)$, where $D_{\rm obs}$ denotes the actually observed values of the data. That is, the likelihood function is the sampling distribution, viewed as a function only of the parameters θ_i, with the data fixed at the observed values. When the data may be modeled as the sum of a predicted signal with additive Gaussian noise of known standard deviation, the likelihood function is closely related to the familiar χ^2 fitting statistic;

$$\mathcal{L}_i(\theta_i) \propto \exp\left[-\frac{1}{2}\chi^2(\theta_i)\right], \tag{1}$$

with χ^2 calculated using the observed data.

The AIC is a *predictive* criterion; it may be derived from both frequentist and Bayesian points of view. We outline a frequentist derivation here. Let $\mathcal{M}_*$ denote the true model describing how data would actually be distributed in repeated sampling. We denote the actual sampling distribution by $p(D|\mathcal{M}_*)$. It may be that $\mathcal{M}_*$ corresponds to one of the models in $\{\mathcal{M}_i\}$ with a particular choice of its parameters (i.e., $\mathcal{M}_* = (\mathcal{M}_i, \theta_i^*)$ for $i = i_*$, the label for the true model). Or it may be that $\mathcal{M}_*$ is not in the explored set. The AIC seeks the model in the set whose sampling distribution is closest to $p(D|\mathcal{M}_*)$, with the distance between distributions measured by an information-theoretic functional (map from functions to a real number), the Kullback-Leibler divergence (KLD), related to entropy. Denote the maximum-likelihood ("best-fit") parameters for model i by $\hat{\theta}_i(D_{\rm obs})$ (e.g., found by minimizing $\chi^2(\theta_i)$ using the observed data). AIC estimates the distance between $p(D|\hat{\theta}_i(D_{\rm obs}), \mathcal{M}_i)$ (the predictive distribution for future data using the best-fit parameters) and $p(D|\mathcal{M}_*)$ with a twofold asymptotic approximation (an approximation valid for large data set size, N). First, since $p(D|\mathcal{M}_*)$ is of course unknown, it is estimated from the data, roughly speaking via properties of the residuals from the best-fit model. Second, an asymptotic approximation is used in the formula for the KLD. The resulting estimated KLD is given by

$$\mathrm{AIC}(\mathcal{M}_i) = -2\ln\mathcal{L}_i(\hat{\theta}_i) + 2k_i, \tag{2}$$

where k_i is the number of parameters for $\mathcal{M}_i$. Larger distances correspond to worse predictions, so models with smaller AIC are preferred. But since the AIC estimates the KLD, and not a probabilistically calibrated

quantity (like a posterior probability or a significance level), there is no straightforward interpretation for how strongly one model may be preferred to another with a larger AIC.

The log-likelihood in the AIC arises from approximating the integral defining the KLD; the $2k_i$ term arises as a bias correction. The next order (in N) term in the approximation is known; empirical studies show using it improves performance when data sets are of modest size. The modified criterion is called the *corrected AIC* (AIC_c); it is given by

$$\mathrm{AIC}_c(\mathcal{M}_i) = -2\ln\mathcal{L}_i(\hat{\theta}_i) + 2k_i\frac{N}{N-k_i-1}. \tag{3}$$

It converges to AIC as N grows large with respect to k_i, so it may be used for both modest and large sample sizes.[1]

The BIC is an *explanatory* criterion; it seeks to identify the model in the explored set that offers the best explanation of the actually observed data, in the sense of the model with the largest Bayesian posterior probability. Its derivation makes no reference to probabilities of data sets other than that actually observed (thus there is no reference to a "true" model in the frequentist sense, i.e., as specifying a repeated-sampling distribution). The posterior probability for a model is proportional to its *marginal likelihood*, the average (*not* the maximum) of the likelihood function for the model's parameters. The averaging accounts for the effect of parameter uncertainty on the ability of the model to predict the observed data. The averaging is done using a prior probability distribution over the model's parameter space (describing before-data uncertainty in the parameters, most importantly, the search ranges for the model parameters). The BIC is an asymptotic approximation to the logarithm of the marginal likelihood. It assumes the best-fit parameters for a model lie within the parameter space (not on a boundary), which is a necessary condition for the likelihood function to asymptotically have a multivariate Gaussian shape. Then the asymptotic behavior of the marginal likelihood may be found by approximating the volume under the likelihood function by the product of the maximum likelihood and the volume of the Gaussian. Asymptotically, the width of the Gaussian scales like $1/\sqrt{N}$ in each parameter direction, so the volume scales like $(1/\sqrt{N})^{k_i}$. This gives the BIC, expressed as minus twice the logarithm of the marginal likelihood approximation;

$$\mathrm{BIC}(\mathcal{M}_i) = -2\ln\mathcal{L}_i(\hat{\theta}_i) + k_i\ln N. \tag{4}$$

A model with a smaller BIC than a competitor will have a larger posterior probability (if the models are considered equally probable a priori); differences in BIC values are thus approximately interpretable as -2 times the logarithm of the odds favoring one model over another. Although derived from the Bayesian point of view, investigators adopting the frequentist point of view and seeking an explanatory criterion are free to adopt the BIC, provided its frequentist performance is understood (more on this below).

Note that Schwarz derived the BIC to capture the leading-order asymptotic behavior of the Bayes factor, i.e., the leading-order terms that vary with sample size. He explicitly omitted an $O(1)$ term (i.e., constant with respect to N) associated with the prior. This term can be important for problems with finite sample sizes—i.e., for every real problem! Kass and Wasserman (1995) discuss this issue, showing how the BIC can be viewed as giving the Laplace approximation for the Bayes factor in the case of "unit information priors" that express the amount of information about the parameters contained in a single observation (this idea is challenging to generalize beyond simple models).

Recall that $-2\ln\mathcal{L}$ is just χ^2 in the common Gaussian noise setting. In this setting, both the AIC_c and the BIC have the form,

$$\mathrm{IC}(\mathcal{M}_i) = \chi^2(\hat{\theta}_i) + \kappa(k_i, N), \tag{5}$$

for different choices of the function $\kappa(k, N)$. For both criteria, $\kappa(k, N)$ grows with k for fixed N. For fixed k, it grows gently with N for the BIC, but is nearly constant for the AIC_c (decreasing slowly for modest N). The growth of $\kappa(k, N)$ with k suggests interpreting κ as a measure of model complexity, in the simplistic sense that models with more parameters are more complex than those with fewer. One may then interpret the criteria as grading models by their minimum χ^2, *penalized* by differing measures of model complexity.

[1] A further generalization of the AIC is worth noting. In assessing predictions, the derivation of the AIC considers future data sets similar to the observed one, in the sense of having the same sample size and similar design (e.g., choice of sample spacing). If one targets predictions of different data sets (e.g., significantly smaller or larger), the factor multiplying k_i may be different, leading to what some have called the generalized information criterion (GIC; see Stoica and Selén 2004).

But it is important to keep in mind that this is an a posteriori interpretation of calculations that make no explicit reference to complexity or a priori model penalties. The criteria address different goals, and both the maximum likelihood terms, and the "complexity" terms, appear in each criterion for completely different reasons.

The more recent DIC is an intrinsically Bayesian criterion. It is like the BIC in that it comes from a Bayesian formulation in which averaging the likelihood is important; but it accounts for parameter uncertainty more accurately than does the BIC. It is like the AIC in that it is a predictive rather than an explanatory criterion; it makes explicit reference to probabilities for not-yet-observed data. Roughly speaking, the DIC seeks to identify the model that we expect will make the best predictions of future data (like AIC), but instead of focusing on the best-fit version of each model, it accounts for parameter uncertainty by numerically averaging over a model's parameter space using Monte Carlo methods. It arose in the context of Bayesian models analyzed using Markov chain Monte Carlo (MCMC) posterior sampling methods. In this setting, one may not even bother calculating the maximum likelihood parameter values for each model. One instead generates a large sample of parameter values, drawn from a posterior distribution; the sample accurately encodes the finite sample size parameter uncertainties, with no recourse to asymptotic approximations. The DIC uses log-likelihood ("deviance") values for all the parameter values in the posterior sample, not just a single maximum log-likelihood value. Also, it has a more sophisticated notion of the "number of parameters" of a model that tries to account for the effective degrees of freedom learnable from the data (which may be different from the number of formal parameters in large, complex models where some parameters may be poorly constrained, or where sets of parameters may have strongly correlated estimates). Though computationally more complicated than the AIC or BIC, it uses quantities readily available from MCMC runs, so it has become popular among investigators using the Bayesian approach. Its foundations and relationship to other criteria are still matters of debate and research (see Plummer 2008, and the discussion in Spiegelhalter et al. 2002), but empirically it appears to perform very well as a predictive criterion. We henceforth focus on the simpler and better-understood AIC and BIC, but for investigators using MCMC methods, the DIC deserves consideration.

The AIC is a predictive criterion and the BIC an explanatory one. In principle, for a well-posed problem, they should not be competitors; they address different goals, and presumably only one will address the goal at hand. In practice, the goals of a multi-model analysis may not be so clear cut. We may be seeking to explain the observed data (i.e., to interpret the chosen model as a realistic (if approximate) description of the data generation process); but we may also intend to use the model for predictions (e.g., to plan future observations). We may be considering rival, purely phenomenological models of little explanatory interest (e.g., polynomials of various order); we may use such models for a component of the data that is not of ultimate scientific interest (e.g., a background or detector properties), but it is needed in the context of explanatory modeling of other components of the data; the ultimate goal is thus explanatory. Given the "blurry" goals of practical model selection, it is worth comparing the performance of rival criteria in various settings.

Suppose we have a nested set of models (so a smaller model is a special case of larger models, with extra parameters set to default values; e.g., low-order polynomials are higher-order polynomials with zero coefficients for the higher-order terms). Suppose further that we are in a repeated-sampling setting, with the actual model generating sets of data among the considered models (with a fixed set of parameters). Then it is known that the AIC will tend to overfit (choose too big a model) with finite sample sizes. As sample size grows, the probability of underfitting (choosing too small a model) vanishes, but the probability of overfit stays finite. (Some investigators argue that this overfitting tendency is due to reliance on the best-fit model in the AIC, i.e., that it is a consequence of incomplete accounting for parameter uncertainty; see Kass and Raftery 1995 for discussion and references.) This warns that if the explanatory significance of the chosen model is important, the AIC may be a poor choice.

In the same setting, the BIC tends to underfit with finite sample size, giving larger prediction errors (e.g., mean-square residuals) than models chosen by the AIC. This arises because of the larger penalty (for $N > 7$) in the BIC for a model of a given size. This warns against use of the BIC when finite-sample prediction is paramount (it is an explanatory criterion after all). As sample size grows, the BIC will choose the correct model with probability converging to unity, arguing in its favor if the actual identity of a correct repeated-sampling model (and not merely the quality of predictive distributions) is scientifically important.

Now suppose the data-generating model is not in the explored set. Then the notions of underfit and

overfit no longer apply in a simple way, though it may often be true that one of the considered models includes predictive distributions closer to the true one than any competing models. If our models are merely phenomenological tools for prediction, the framework leading to the AIC is natural, and it should be favored over the BIC. In some fundamental sense, it may seem that the BIC (or any explanatory criterion) is irrelevant in this scenario. However, even the best explanatory models are seldom believed to be exactly true,[2] so even in an explanatory setting a BIC user may be interested in how the BIC behaves when the "true model" (in a repeated-sampling sense) differs from the explored models. Also, from a frequentist point of view, one may simply regard the BIC as another ad hoc model selection statistic and investigate its properties. Statisticians have shown that, when the parameter posterior for a model has an asymptotic limit (which we might hope to be true when a model closely approximates reality), Bayesian posterior probabilities concentrate on the model whose predictive distribution is closest to the actual sampling distribution, with distance measured by the KLD.[3] Since the BIC asymptotically approximates the marginal likelihood (twice its negative logarithm), it presumably inherits this behavior (when the BIC approximation is valid). This identifies the sense in which we should think of explanatory models as approximations of true models when thinking about Bayesian model selection in settings where repeated sampling (and thus prediction) may be relevant. Interestingly, minimizing the KLD is just what the AIC explicitly aims to do; presumably, by targeting it directly (as a predictive criterion), it does a better job picking predictively close models with modest-sized samples than does the BIC, though a less approximate predictive Bayesian criterion, such as the DIC, may do better still. I do not know of studies specifically exploring this.

As a final remark on performance, one should note that most theoretical studies of performance are asymptotic. It is legitimate to ask what the relevance of such studies is for real-world analyses where data sets are always comfortably finite. Some of the strange asymptotic behaviors of methods arise because with infinite data, even small differences between sampling distributions become large compared to asymptotically infinitesimal measurement uncertainties. But with finite sample size, an approximate model may be empirically indistiguishable from the "true" model, and peculiar asymptotic behavior may never become evident for achievable sample sizes. Indeed, restricting consideration to realistically finite data, it may be that there is no objective, operational meaning to the notion of a unique true sampling distribution. Finite sample properties are notoriously difficult to study theoretically, but simulation studies can shed light on these issues, and hopefully ongoing theoretical work will further illuminate them.

Berk, R. (1966) "Limiting Behavior of Posterior Distributions when the Model is Incorrect," *Ann. Math. Stat.* 37, 51–58 [online]

Ghosal, S., Lember, J., and van der Vaart, A. (2008) "Nonparametric Bayesian model selection and averaging," *Elec. J. Stat.* 2, 63–89 [online]

Kass, R., and Wasserman, L. (1995) "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion," *JASA* 90, 928–934 [JSTOR]

Lahiri, P. (ed.) (2001) *Model Selection*, Institute of Mathematical Statistics Lecture Notes–Monograph Series, Volume 38, (Beachwood, OH) [online]

Plummer, M. (2008) "Penalized loss functions for Bayesian model comparison," *Biostatistics* **9**, 523–539 [online]

Rao, C. R. and Wu, Y. (2001) "On model selection," in *Model Selection*, ed. P. Lahiri, Institute of Mathematical Statistics Lecture Notes–Monograph Series, Volume 38, (Beachwood, OH) 1–57 [online]

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002) "Bayesian measures of model complexity and fit (with discussion)" *J. Roy. Stat. Soc.* B64, 583–639 [online]

Stoica, P., and Selén, Y. (2004) "Model-order selection: A review of information criterion rules," *IEEE Sig.*

[2]Even if we are confident in the physical model for the signal, the statistical model for the noise in the data will often be approximate.

[3]In general, a unique asymptotic limit may not exist when no model is true; see Berk 1966 for the general case, and Ghosal et al. 2008 for examples where a limit exists.

Proc. Mag. 21, 36–47 [online]