Model Selection Criteria Based on Computationally Intensive Estimators of the Expected Optimism

Joseph E. Cavanaugh[1,2], Andrew A. Neath[3]
[1] Department of Biostatistics, The University of Iowa, Iowa City, IA, 52242, USA
[3] Department of Mathematics and Statistics, Southern Illinois University, Edwardsville, IL, 62026, USA
[2] Corresponding Author: joe-cavanaugh@uiowa.edu

Abstract

A model selection criterion based on a divergence or discrepancy measure is generally comprised of a goodness-of-fit term and a penalty term. The penalty term, which reflects model complexity, serves as an estimate of a quantity known as the expected optimism. Classical approaches to approximating the expected optimism often lead to simplistic penalizations. However, such approaches usually involve stringent assumptions that may fail to hold in practical applications. Modern computational statistical methods facilitate the development of improved estimators of the expected optimism. Selection criteria based on such penalty terms often provide more realistic measures of predictive efficacy than their classical counterparts, thereby resulting in superior model determinations. To survey this methodology, we outline the general framework for discrepancy-based model selection criteria, and review computationally intensive approaches for evaluating complexity penalizations.

Keywords: Bootstrap; cross validation; divergence measures; Monte Carlo simulation; variable selection

1. Introduction

A model selection criterion is often formulated by constructing an approximately unbiased estimator of an *expected discrepancy*, a measure that gauges the separation between the true model and a fitted candidate model. The expected discrepancy reflects how well, on average, the fitted candidate model predicts "new" data generated under the true model. A related measure, the *estimated discrepancy*, reflects how well the fitted candidate model predicts the data at hand.

In general, a model selection criterion consists of a goodness-of-fit term and a penalty term. The natural estimator of the expected discrepancy, the estimated discrepancy, corresponds to the goodness-of-fit term. However, the estimated discrepancy yields an overly optimistic assessment of how effectively the fitted model predicts new data. It therefore serves as a negatively biased estimator of the expected discrepancy. Correcting for this bias leads to the penalty term. Specifically, the penalty term provides an approximation to the expectation of the difference between the expected discrepancy and the estimated discrepancy, a measure known as the *expected optimism*.

Classical approaches to approximating the expected optimism often lead to simplistic penalty terms based on the dimension of the fitted candidate model and possibly the sample size. However, such approaches generally involve large-sample arguments, restrictive assumptions on the form of the candidate model, or both. The resulting penalty terms may fail to perform adequately in small-sample applications or in settings where the requisite assumptions do not hold.

Modern computational statistical methods, such as Monte Carlo simulation, bootstrapping, and cross validation, facilitate the development of flexible and accurate estimators of the expected optimism. Model selection criteria based on such penalty terms often provide more realistic measures of predictive efficacy than their classical counterparts, thereby resulting in superior model determinations.

In this note, we review the general paradigm for discrepancy-based model selection criteria, and discuss computationally intensive approaches to approximating the expected optimism.

2. Discrepancy-based selection criteria

Consider a collection of *n* response measurements $Y = \{y_1, \ldots, y_n\}$, where the y_i's may be scalars or vectors, often assumed to be independent. Let $\mathcal{M}_o$ denote the unknown "true" model; i.e., the model that presumably generates the sample Y.

Suppose that a parametric model is postulated for Y. Let θ denote the parameter vector for the model, and let $\mathcal{M}_\theta$ denote the candidate model. Let *k* denote the dimension of the candidate model: i.e., the number of functionally independent parameters in θ.

The quality of the candidate model $\mathcal{M}_\theta$ can be gauged by determining whether this model may be used to formulate accurate predictors of data generated under the true model $\mathcal{M}_o$. Consider a measure $\delta(Y,\theta)$ that assesses the effectiveness of model $\mathcal{M}_\theta$ in predicting the data Y. Suppose that $\delta(Y,\theta)$ is defined so that smaller values of $\delta(Y,\theta)$ are reflective of greater predictive efficacy. We will refer to $\delta(Y,\theta)$ as the *observed discrepancy*.

Once the observed discrepancy is defined, we may propose an estimator of θ based on minimizing this measure:

$$\hat{\theta} = \text{argmin}_\theta \, \delta(Y,\theta).$$

Such an estimator is called a *minimum discrepancy estimator* (MDE).

By replacing θ with $\hat{\theta}$ in $\delta(Y,\theta)$, we obtain a statistic $\delta(Y,\hat{\theta})$ known as the *estimated discrepancy*. The estimated discrepancy evaluates the predictive effectiveness of the fitted model $\mathcal{M}_{\hat{\theta}}$ based on the data used in its own construction. This statistic may be viewed as a goodness-of-fit measure for $\mathcal{M}_{\hat{\theta}}$. Comparing values of the statistic for various fitted models may facilitate the identification of models that are too simplistic to accommodate the data at hand. However, $\delta(Y,\hat{\theta})$ will always decrease as the complexity of the candidate model $\mathcal{M}_\theta$ is increased. Thus, choosing the fitted model in a candidate family that minimizes the estimated discrepancy will invariably result in selecting the most complex candidate model.

Obviously, the problem with the measure $\delta(Y,\hat{\theta})$ is that it leads to an overly optimistic assessment of predictive efficacy, one that is solely based on the conformity of the fitted model $\mathcal{M}_{\hat{\theta}}$ to the data used to fit the model. In principle, suppose that we could circumvent this problem by collecting a complete set of *n* new measurements on the response variable, say $Z = \{z_1, \ldots, z_n\}$, and assessing the predictive effectiveness of $\mathcal{M}_{\hat{\theta}}$ based on the data Z as opposed to the data Y. The measure $\delta(Z,\hat{\theta})$ could be used for this purpose. We could then view Y as a *fitting sample*, and Z as a *validation sample*. We will refer to $\delta(Z,\hat{\theta})$ as the *validatory discrepancy*.

The *expected discrepancy*, also known as the *expected divergence*, is defined as

$$\Delta(\mathcal{M}_o, \mathcal{M}_\theta) = \text{E}\left\{\delta(Z,\hat{\theta})\right\}, \tag{1}$$

where $\text{E}(\cdot)$ denotes the expectation under the true model $\mathcal{M}_o$. This measure reflects how well, on average, a fitted candidate model of the form $\mathcal{M}_\theta$ predicts new data generated under the true model $\mathcal{M}_o$. Since $\Delta(\mathcal{M}_o, \mathcal{M}_\theta)$ is based on averaging over the distributions of both Y and Z, the measure does not depend on data, but rather on constructs pertaining to both the true model $\mathcal{M}_o$ and the candidate model $\mathcal{M}_\theta$.

By comparing values of the expected discrepancy for various fitted models in a candidate family, one would be able to determine the optimal model structure. However, since the measure $\Delta(\mathcal{M}_o, \mathcal{M}_\theta)$ depends on the true model $\mathcal{M}_o$, it is inaccessible.

Model selection criteria are often formulated by constructing approximately unbiased estimators of the expected discrepancy. The definition (1) implies that $\delta(Z,\hat{\theta})$ could be used to estimate $\Delta(\mathcal{M}_o, \mathcal{M}_\theta)$ without bias, yet rarely is a validation sample Z actually available. A more pragmatic goal is to use the fit measure $\delta(Y,\hat{\theta})$ as a platform for estimating $\Delta(\mathcal{M}_o, \mathcal{M}_\theta)$, recognizing that $\delta(Y,\hat{\theta})$ will be inherently biased, and to formulate a bias adjustment for $\delta(Y,\hat{\theta})$.

To investigate this approach, consider writing $\Delta(\mathcal{M}_o, \mathcal{M}_\theta)$ as follows:

$$\begin{aligned}\Delta(\mathcal{M}_o, \mathcal{M}_\theta) &= \mathrm{E}\left\{\delta(Y, \hat{\theta})\right\} \\ &\quad + \left[\mathrm{E}\left\{\delta(Z, \hat{\theta}) - \delta(Y, \hat{\theta})\right\}\right]. \end{aligned} \tag{2}$$

The bracketed quantity (2) is often referred to as the *expected optimism* in judging the fit of a model using the same data as that which was used to construct the fit (Efron, 1983, 1986). The expected optimism is positive, implying that $\delta(Y, \hat{\theta})$ is negatively biased as an estimator of $\Delta(\mathcal{M}_o, \mathcal{M}_\theta)$. In order to correct for the negative bias, we must evaluate or approximate the bias adjustment represented by the expected optimism.

Such a bias correction, say $\widehat{\mathrm{eo}}$, is then added to the estimated discrepancy $\delta(Y, \hat{\theta})$ to produce an approximately unbiased estimator of the expected discrepancy $\Delta(\mathcal{M}_o, \mathcal{M}_\theta)$:

$$\delta(Y, \hat{\theta}) + \widehat{\mathrm{eo}}. \tag{3}$$

The statistic $\delta(Y, \hat{\theta}) + \widehat{\mathrm{eo}}$ may then be used as a model selection criterion: among a candidate collection of fitted models, the fitted model corresponding to the smallest value of $\delta(Y, \hat{\theta}) + \widehat{\mathrm{eo}}$ should be favored.

Simple approximations to $\widehat{\mathrm{eo}}$ can often be found by employing the following assumptions:

(A.1) The sample size *n* is large relative to the dimension of the candidate model *k*.

(A.2) The true model $\mathcal{M}_o$ is subsumed by the candidate model $\mathcal{M}_\theta$, so that the fitted model $\mathcal{M}_{\hat{\theta}}$ is either correctly specified or overfit.

Mathematically, assumption (A.1) often translates to the use of asymptotic results for the minimum discrepancy estimator $\hat{\theta}$ that hold in the limit as *n* approaches infinity, provided that the candidate model dimension *k* is assumed fixed. For certain discrepancies, under assumptions (A.1) and (A.2), the expected optimism can be approximated by a simple function of *k*. (See, for example, Theorem 2.4 of Mattheou, Lee, and Karagrigoriou, 2009.) In fact, in some instances, the expected optimism asymptotically reduces to a multiple of *k*, as shown in section 2.4 of Linhart and Zucchini (1986). Such a simplification leads to the penalty terms of the Akaike (1973, 1974) information criterion (AIC) and Mallow's (1973) C_p. See also Linhart and Zucchini (1985) and Cavanaugh (1999).

Variable selection methods have been extensively studied in the framework of Gaussian linear regression. In such a framework, under only assumption (A.2), exact expressions for the expected optimism that only depend on *k* and *n* can be derived for certain discrepancies. Examples may be found in Sugiura (1978) and Cavanaugh (2004).

Assumptions (A.1) and (A.2) lead to model selection criteria based on simplistic penalty terms that are easily computed. However, in settings where (A.1) or (A.2) are violated, such a penalty term may provide a poor approximation to the expected optimism, leading to a selection criterion that serves as an inferior estimator of the expected discrepancy. As a result, the selection criterion might choose a model in a candidate collection which is quite different from the model that would be favored by the expected discrepancy. Simulation studies that illustrate this phenomenon appear in Hurvich and Tsai (1989), Hurvich, Shumway, and Tsai (1990), Cavanaugh and Shumway (1997), Cavanaugh (2004), and Kim and Cavanaugh (2005).

As previously mentioned, modern computational statistical methods, such as Monte Carlo simulation, bootstrapping, and cross validation, facilitate the development of flexible and accurate estimators of the expected optimism. Specifically, estimators based on Monte Carlo simulation may be developed by relaxing the large-sample assumption (A.1), and estimators based on bootstrapping and cross-validation may be developed by relaxing both (A.1) and the model specification assumption (A.2). Model selection criteria featuring such penalty terms often provide more realistic measures of predictive efficacy than their classical counterparts. In the next section, we discuss these approaches.

3. Computationally intensive estimators of the expected optimism

3.1 Monte-Carlo simulation

Consider a discrepancy where under assumptions (A.1) and (A.2), the expected optimism can be approximated by a simple function of k and possibly n. Such a result implies that the expected optimism does not depend on the form of the true model $\mathcal{M}_o$ for correctly specified or overspecified candidate models, at least when the sample size is large. If the sample size is small or moderate, it may therefore follow that the expected optimism may only depend loosely on the form of the true model. Based on this notion, Hurvich, Shumway, and Tsai (1990) proposed an "improved" Akaike information criterion, AICi, where the estimator of the expected optimism is based on Monte Carlo simulation. (See also Kim and Cavanaugh, 2005, and Bengtsson and Cavanaugh, 2006.) Although the development of AICi arises from the use of Kullback's directed divergence as the targeted discrepancy, the idea behind the criterion may be extended to any other discrepancy where the preceding result applies.

Consider a collection of candidate models and a sample size n. The computation of the Monte Carlo estimate of the expected optimism proceeds as follows.

1. Identify the smallest model in the candidate collection. For this candidate model structure, choose a convenient, fixed value for the parameter vector θ. Let $\mathcal{M}_f$ denote the model based on this fixed parameter vector.

2. Use the model $\mathcal{M}_f$ to generate multiple fitting samples $Y(1),\ldots,Y(R)$ and multiple validation samples $Z(1),\ldots,Z(R)$.

3. For a given candidate model structure $\mathcal{M}_\theta$, obtain MDE replicates $\hat{\theta}(1),\ldots,\hat{\theta}(R)$ using the fitting samples $Y(1),\ldots,Y(R)$.

4. Compute the estimate of the expected optimism as follows:

$$\widehat{\text{eo}} = \frac{1}{R}\sum_{i=1}^{R}\left\{\delta(Z(i),\hat{\theta}(i)) - \delta(Y(i),\hat{\theta}(i))\right\}.$$

5. Repeat steps (2) through (4) for each candidate model structure under consideration, thereby obtaining a penalization for every model in the candidate collection.

The resulting penalizations may be tabulated and used for any set of fitted models from the candidate family based on the sample size n. Model selection criteria may then be constructed by augmenting the goodness-of-fit statistics with these penalizations, as indicated in (3).

In smaller sample settings, simulation results demonstrate that model selection criteria based on the preceding approach outperform their classical counterparts; see, for instance, Hurvich, Shumway, and Tsai (1990), Kim and Cavanaugh (2005), and Bengtsson and Cavanaugh (2006). In such settings, simplistic penalty terms derived under (A.1) and (A.2) grossly underestimate the expected optimism for the larger models in the candidate collection. As a result, selection criteria with simplistic penalizations favor larger models, even when the expected discrepancy indicates that such models have poor predictive capabilities. In contrast, criteria with Monte Carlo penalizations provide improved estimators of the expected discrepancy, leading to more parsimonious model selections that better reflect the values of the target measure.

The Monte Carlo approach relaxes the large-sample assumption (A.1) often employed in the development of simplistic penalty terms. However, the justification of this approach depends on the model specification assumption (A.2). Bootstrapping and cross validation have been used to develop estimators of the expected optimism that relax both (A.1) and (A.2). Next, we outline these approaches.

3.2 Bootstrapping

The idea of using the bootstrap to improve the performance of a model selection rule was introduced by Efron (1983, 1986). To present the basic approach, let $\{Y^*(i) \mid i = 1,\ldots,B\}$ represent a collection of B bootstrap samples, and let $\{\hat{\theta}^*(i) \mid i = 1,\ldots,B\}$ represent a collection of B bootstrap replicates of $\hat{\theta}$ corresponding to the B bootstrap samples.

The bootstrap estimator of the expected optimism is based on the familiar "plug–in" principle. The estimator is given by

$$\widehat{\text{eo}} = \frac{1}{B}\sum_{i=1}^{B}\left\{\delta(Y,\hat{\theta}^*(i)) - \delta(Y^*(i),\hat{\theta}^*(i))\right\}.$$

In comparing the preceding to (2), note that Y plays the role of the validation sample and a bootstrap sample Y^* plays the role of the fitting sample. The discrepancy $\delta(Y^*(i),\hat{\theta}^*(i))$ evaluates the predictive effectiveness of the $\hat{\theta}^*(i)$ fitted model based on the data $Y^*(i)$ used to fit the model. In the discrepancy $\delta(Y,\hat{\theta}^*(i))$, the predictive effectiveness of the $\hat{\theta}^*(i)$ model is assessed using the original data Y.

Standard bootstrapping procedures require that the data to be resampled consists of independent, identically distributed (iid) replicates. For modeling problems where the y_i's are assumed independent and a set of potential covariates X_i is associated with each outcome y_i, a multivariate distribution is often assumed for the pairs (y_i,X_i). The variates (y_i,X_i) can then be viewed as iid replicates arising from the multivariate distribution, and can be randomly drawn with replacement to construct the bootstrap samples. This approach is referred to as *nonparametric* bootstrapping.

If a model is fit to the data and the residuals are obtained, the bootstrap samples are sometimes constructed by resampling the residuals, and assembling the bootstrap samples based on using the resampled residuals in conjunction with the fitted model. If the residuals are resampled by taking random draws with replacement from the residuals for the original model fit, the procedure is called *semi-parametric*. If the residuals are resampled by generating random samples from an assumed parametric distribution, the procedure is called *parametric*.

In the present context, nonparametric bootstrapping is arguably the most natural procedure, since this method does not require the use of a fitted model. However, in settings where the y_i are dependent, such as time series applications, a semi-parametric or nonparametric approach might be necessary, since the residuals can often be assumed iid whereas the original data cannot.

3.3 Cross validation

Stone (1977) proposed an analogue of AIC based on cross validation and established the asymptotic equivalence of AIC and this cross validatory analogue. Further work on this criterion appears in Davies, Neath, and Cavanaugh (2005), Cavanaugh, Davies, and Neath (2008), and Konishi and Kitagawa (2008). A comprehensive survey of cross-validatory methods for model selection is presented by Arlot and Celisse (2010).

To outline the cross validatory approach to estimation of the expected optimism, assume that $y_1,\ldots,y_n$ are independent. In such a setting, the observed discrepancy $\delta(Y,\theta)$ can often be decomposed into the sum of n individual contributions $\delta_1(y_1,\theta),\ldots,\delta_n(y_n,\theta)$, where each contribution $\delta_i(y_i,\theta)$ corresponds to a specific observation y_i:

$$\delta(Y,\theta) = \sum_{i=1}^{n}\delta_i(y_i,\theta).$$

For example, if $\delta(Y,\theta)$ represents the negative log-likelihood for θ based on the data Y (i.e., the negative of the log of the joint density of Y), then $\delta_i(y_i,\theta)$ would represent the negative log-likelihood contribution corresponding to y_i (i.e., the negative of the log of the marginal density of y_i). If $\delta(Y,\theta)$ represents the sum of the squared deviations between the observations and their mean values under the candidate model $\mathcal{M}_\theta$, then $\delta_i(y_i,\theta)$ would represent the specific squared deviation for the i^{th} observation y_i.

Let $Y[i]$ denote the data set Y with the i^{th} case y_i excluded. Let $\hat{\theta}[i]$ denote an estimator of θ based on $Y[i]$. Often, $Y[i]$ is referred to as the *case-deleted data set*, and the fitted model based on $\hat{\theta}[i]$ as the *case-deleted fitted model*.

Recall the general definition of the expected optimism:

$$\mathrm{E}\left\{\delta(Z,\hat{\theta}) - \delta(Y,\hat{\theta})\right\}.$$

Under suitable conditions, the cross-validatory statistic

$$\sum_{i=1}^{n} \delta_i(y_i, \hat{\theta}[i])$$

serves as an asymptotically unbiased estimator of $\mathrm{E}\left\{\delta(Z,\hat{\theta})\right\}$. Thus, a cross-validatory estimator of the expected optimism is given by

$$\widehat{\mathrm{eo}} = \left(\sum_{i=1}^{n} \delta_i(y_i, \hat{\theta}[i])\right) - \left(\sum_{i=1}^{n} \delta_i(y_i, \hat{\theta})\right). \tag{4}$$

The preceding result is established in Chapter 10 of Konishi and Kitagawa (2008). (See also Cavanaugh, Davies, and Neath (2008).) In (4), note that the estimated discrepancy

$$\delta(Y,\hat{\theta}) = \sum_{i=1}^{n} \delta_i(y_i, \hat{\theta})$$

serves as an estimator of its own expected value.

The model selection criterion based on the penalty term (4) is given by

$$\begin{aligned} \delta(Y,\hat{\theta}) + \widehat{\mathrm{eo}} &= \sum_{i=1}^{n} \delta_i(y_i, \hat{\theta}) \\ &\quad + \left[\left(\sum_{i=1}^{n} \delta_i(y_i, \hat{\theta}[i])\right) - \left(\sum_{i=1}^{n} \delta_i(y_i, \hat{\theta})\right)\right] \\ &= \sum_{i=1}^{n} \delta_i(y_i, \hat{\theta}[i]). \end{aligned} \tag{5}$$

In general, the evaluation of (5) requires n case-deleted model fits. However, in the context of Gaussian linear models, certain discrepancies will lead to contributions $\delta_i(y_i, \hat{\theta}[i])$ that may be computed based only on the original model fit.

Key References

Arlot S. and Celisse A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4: 40-79.

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

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

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

Konishi S. and Kitagawa G. (2008). *Information Criteria and Statistical Modeling*, Springer.

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

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.