

An Optimality Result for a Cross Validation Estimate of Prediction Error

Andrew A. Neath, Southern Illinois University Edwardsville
Joseph E. Cavanaugh, The University of Iowa
Simon Davies, Centocor, Inc.

Key Words: Kullback divergence, Akaike information, minimum variance unbiased estimator, model selection criterion

1. Introduction

An important topic in linear regression theory is the statistical problem of determining which input variables are needed for estimating a response function. Such a decision is often facilitated by the use of a model selection criterion. For many situations, a candidate model's predictive ability is its most important attribute. In light of our interest in this property, we concentrate on model selection techniques based on cross validation, namely jackknifing. Our basic approach is to construct a measure which gauges the adequacy of an approximating model by assessing how effectively each case-deleted fitted model predicts the deleted case, as quantified by the Kullback-Leibler divergence.

A model selection criterion for linear regression based on cross validation, the predictive divergence criterion (PDC), is introduced. The goal of defining a parameter which reflects the predictive capabilities of a candidate model is attained through a derivation of the target value of PDC. We then develop an adjusted predictive divergence criterion (PDCa) which serves as the minimum variance unbiased estimator of this target.

2. Prediction Error

Consider a collection of data $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ where x_i is a $(p_L \times 1)$ vector of input variables and y_i is a real valued response. So, $y = (y_1, \dots, y_n)'$ is an $(n \times 1)$ response vector and $X_L = [x_1, \dots, x_n]'$ is an $(n \times p_L)$ design matrix of full column rank. The data \mathcal{D} will be treated as an iid sample from some multidimensional distribution F .

We take on the problem of determining which of the input variables are needed for creating the

“best” linear prediction function. Candidate models postulated for the data will be of the form

$$M : \quad y = X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2 I). \quad (2.1)$$

Here, X is an $(n \times p)$ design matrix with the column space property $C(X) \subseteq C(X_L)$. Then X_L represents the largest of the possible design matrices.

A fitted model corresponding to (2.1) is estimated from the data, most commonly using a least squares criterion or maximum likelihood. The predicted response at an input level x_o is expressed as $\eta_{\mathcal{D}}(x_o) = x_o' \hat{\beta}$. The goal of model selection for best prediction will be achieved by defining, for each candidate model M , a measure quantifying its predictive capability.

Let $Q[y, \eta]$ denote the error between a response y and its predicted value η . The prediction error for the forecast rule $\eta_{\mathcal{D}}$ is defined to be

$$err(\mathcal{D}, F) = E_F\{Q[y_o, \eta_{\mathcal{D}}(x_o)]\}, \quad (2.2)$$

where expectation is taken over a new realization (x_o, y_o) from the true distribution F . The data is treated as fixed since it is the fitted model whose predictive quality is to be judged.

Since F is unknown, it may at first glance seem appropriate to use the plug-in estimate $err(\mathcal{D}, \hat{F})$ where \hat{F} denotes the empirical distribution on the data \mathcal{D} . Call this the *apparent error rate* and write

$$err(\mathcal{D}, \hat{F}) = \frac{1}{n} \sum_{i=1}^n Q[y_i, \eta_{\mathcal{D}}(x_i)]. \quad (2.3)$$

However, the apparent error rate tends to be overly “optimistic” as a reflection of prediction error. The same data used to create the prediction function is also being used to judge its accuracy. For a detailed analysis of how this bias can be corrected, see Efron (1986).

Cross validation is a popular approach to circumventing the problem of having the training sample also serve as the test sample. Denote the data

set with the i^{th} observation (x_i, y_i) excluded as \mathcal{D}_{-i} with prediction function η_{-i} . Then the leave-one-out cross validation, or jackknife, estimate of prediction error is

$$CV = \frac{1}{n} \sum_{i=1}^n Q[y_i, \eta_{-i}(x_i)]. \quad (2.4)$$

The focus of our study here is cross validation and thus, our target measure is defined to be $E_F(CV)$ where expectation is again with respect to the true distribution F , now taken over the data \mathcal{D} . We examine a specific case in the next section by defining the error function via the Kullback-Leibler divergence, a well known measure of separation for model selection problems.

3. Kullback-Leibler Divergence

Let

$$\begin{aligned} Q_{KL}[y, M] &= -2 \log g_M(y) \\ &= \log \sigma^2 + \frac{(y - \eta)^2}{\sigma^2}, \end{aligned} \quad (3.1)$$

where g_M is the Gaussian probability density function determined by the candidate model M in (2.1). The error function Q_{KL} is defined in terms of the model variance σ^2 as well as the model predicted response η . The Kullback-Leibler divergence is the basis for the Akaike information criterion, AIC (Akaike, 1973, 1974), the first model selection criterion to gain widespread acceptance. A corrected version, AICc, has been proposed by Hurvich and Tsai (1989).

In the present context of studying a cross validation statistic given by (2.4), we introduce the predictive divergence criterion

$$PDC = \frac{1}{n} \sum_{i=1}^n \left[\log \sigma_{-i}^2 + \frac{(y_i - \eta_{-i}(x_i))^2}{\sigma_{-i}^2} \right], \quad (3.2)$$

where σ_{-i}^2 is the case-deleted maximum likelihood estimate of the model variance and $\eta_{-i}(x_i)$ is the case-deleted predicted value for y_i . The targeted measure is $E_F(PDC)$.

We note that cross validation in general, and PDC in particular, provides a computation of prediction error which does not require any specification of the underlying true distribution F . However, in order to obtain analytical results on the target of interest, we will consider the same conditions as those imposed for the development of the Akaike statistics. This leads to a method for improving PDC through a more accurate estimate of its target $E_F(PDC)$. Our mission in the next section will

be to derive the minimum variance unbiased estimator for $E_F(PDC)$.

4. Minimum Variance Unbiased Estimation

Now suppose that the true distribution F is such that the response vector y is related to a design matrix X_T of input variables according to a linear model

$$y = X_T \beta_T + \epsilon_T, \quad \epsilon_T \sim N_n(0, \sigma_T^2 I). \quad (4.1)$$

Suppose further that $C(X_T) \subseteq C(X)$. That is, we are assuming the candidate model (2.1) subsumes the true model (4.1). Imposing this condition reduces the generality of the results, yet ensures mathematical tractability and is methodologically defensible. Model selection criteria are best judged by their capacity to distinguish between models with little or no approximation error, precisely the condition for which Akaike-type criteria results, and those of this section, will apply.

We now provide an expression for the target parameter $E_F(PDC)$. See Davies, Neath, Cavanaugh (2004 a,b) for details. Write

$$\begin{aligned} E_F(PDC) &= \log \sigma_T^2 + \log \frac{2}{n-1} \\ &+ \psi \left(\frac{n-p-1}{2} \right) \\ &+ \frac{n-1}{n-p-3} \left[\frac{1}{n} \sum_{i=1}^n E_F \left(\frac{1}{1-H_{ii}} \right) \right], \end{aligned} \quad (4.2)$$

where ψ is the digamma or psi function and H_{ii} is the i^{th} diagonal element of the hat matrix $H = X(X'X)^{-1}X'$. Hence, the task is to find the MVUE of (4.2) under candidate model (2.1) where it is only known that the candidate model is not underspecified.

We will use a traditional approach to the problem of determining minimum variance unbiasedness. If one can find an unbiased estimator which is a function of a complete sufficient statistic, an appeal to the Lehmann-Scheffé Theorem establishes the desired property. Details on the following are also available in the papers of Davies, Neath, Cavanaugh (2004 a,b).

Proposition 4.1: $T(X, y) = (X'X, y'y, X'y)$ is a complete sufficient statistic.

Proposition 4.2: $\hat{\sigma}^2 = \frac{1}{n} \left| y - X\hat{\beta} \right|^2$ is a function of $T(X, y)$.

Proposition 4.3:
 $E_F \log \hat{\sigma}^2 = \log \sigma_T^2 + \log \frac{2}{n} + \psi\left(\frac{n-p}{2}\right)$.

Theorem 4.1: In the linear regression setting under the described regularity conditions, the alternate predictive divergence criterion

$$\begin{aligned} PDCa &= \log \hat{\sigma}^2 + \log \frac{n}{n-1} & (4.3) \\ &+ \left[\psi\left(\frac{n-p-1}{2}\right) - \psi\left(\frac{n-p}{2}\right) \right] \\ &+ \frac{n-1}{n-p-3} \left[\frac{1}{n} \sum_{i=1}^n \frac{1}{1-h_{ii}} \right] \end{aligned}$$

is the minimum variance unbiased estimator for $E_F(PDC)$.

5. Concluding Remarks

Cross validation is an effective tool for assessing prediction error based on a training sample alone. A natural question may arise as to what precisely a cross validation statistic is estimating. In other words, a determination of the target parameter for cross validation is desired.

We have considered the specific case where prediction error is measured by the Kullback-Leibler divergence. We introduce the corresponding cross validation statistic as the predictive divergence criterion (PDC). Under reasonable regularity conditions, $E_F(PDC)$, the target for PDC, is derived. Under these same conditions, an optimality result is presented where an alternate predictive divergence criterion (PDCa) has the same target $E_F(PDC)$, but achieves minimum variance among the class of unbiased estimators. So, $E_F(PDCa) = E_F(PDC)$, with $Var_F(PDCa) < Var_F(PDC)$. An improvement on cross validation, for estimating a target created by cross validation, is achieved.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (Edited by B.N. Petrov and F. Csáki), 267-281. Akadémia Kiadó, Budapest.
- Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, AC-19, 716-723.
- Davies, S. Neath, A. Cavanaugh, J. (2004). On the minimum variance unbiasedness property of AICc and MCp. Technical Report, Department of Biostatistics, The University of Iowa.
- Davies, S. Neath, A. Cavanaugh, J. (2004). Cross Validation Model Selection Criteria Based on the Kullback-Leibler Discrepancy. Technical Report, Department of Biostatistics, The University of Iowa.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 78, 316-331.
- Hurvich, C. Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.