Cross Validation Model Selection Criteria for Linear Regression Based on the Kullback-Leibler Discrepancy

Simon L. Davies
Pfizer Development Operations, Pfizer, Inc.

Andrew A. Neath
Department of Mathematics and Statistics, Southern Illinois University Edwardsville

Joseph E. Cavanaugh
Department of Biostatistics, The University of Iowa

Abstract

For many situations, the predictive ability of a candidate model is its most important attribute. In light of our interest in this property, we introduce a new cross validation model selection criterion, the predictive divergence criterion (PDC), together with a description of the target discrepancy upon which it is based. In the linear regression framework, we then develop an adjusted cross validation model selection criterion (PDCa) which serves as the minimum variance unbiased estimator of this target discrepancy. Furthermore, we show that this adjusted criterion is asymptotically a minimum variance unbiased estimator of the Kullback-Leibler discrepancy which serves as the basis for the Akaike information criteria AIC and AICc.

Keywords: Akaike information criterion, AIC, AICc, Kullback-Leibler information

1. Introduction

The first model selection criterion to gain widespread acceptance was the Akaike (1973) information criterion (AIC). AIC was developed as an estimator of the expected Kullback-Leibler discrepancy between the model generating the data and a fitted candidate model. AIC serves as an approximately unbiased estimator in instances where the sample size is large and the dimension of the candidate model is relatively small.

In other settings, AIC may be characterized by a large negative bias which limits its effectiveness as a model selection criterion. AICc is a corrected version of AIC originally developed by Sugiura (1978) for linear regression with normal errors. Hurvich and Tsai (1989) justify AICc for linear and non-linear regression, and exhibit its small-sample superiority over AIC in these settings. Davies (2002) and Davies, Neath and Cavanaugh (2005) show that AICc is the minimum variance unbiased estimator of its target discrepancy in a linear regression framework.

For many situations, the predictive ability of a candidate model 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. An early selection criterion based on the cross validatory evaluation of the mean square prediction error is PRESS (Allen, 1974). Cross validation provides a technique for developing an estimator of an expected discrepancy which need not be bias adjusted (Linhart and Zucchini, 1986). Our basic approach is to construct an empirical discrepancy which gauges the adequacy of an approximating model by assessing how effectively each case-deleted fitted model predicts the deleted case, as measured by the Kullback-Leibler divergence.

We introduce a new model selection criterion based on cross validation, the predictive divergence criterion (PDC), together with a description of the target discrepancy upon which it is based. In the linear regression framework, we then develop an adjusted cross validation model selection criterion (PDCa) which serves as the minimum variance unbiased estimator of this target discrepancy. Furthermore, we show that PDCa is asymptotically a minimum variance unbiased estimator of the expected Kullback-Leibler discrepancy which serves as the target measure for AIC and AICc.

2. The Akaike Information Criteria (AIC/AICc)

A general framework for discrepancy based selection criteria is used to motivate the predictive divergence criterion, PDC. We begin with a review of the development of AIC and AICc which will be conducive to our development of PDC.

Suppose we have an n-dimensional vector y of data

$$y = (y_1, ..., y_n)',$$

where we will assume the elements of y are independent. The likelihood for the generating or "true" model for the data vector y can be expressed as

$$L(\theta_o \; |y) = \prod_{i=1}^{n} g_i(y_i|\theta_o),$$

where $g_i(y_i|\theta_o)$ for $i = 1, ..., n$ are the individual probability density functions corresponding to the generating models for the elements of y. The likelihood function for the candidate or approximating model is expressed as

$$L(\theta \; |y) = \prod_{i=1}^{n} g_i(y_i|\theta).$$

Here, θ_o and θ represent vectors of functionally independent parameters for the generating model and the candidate model, respectively. Let Θ denote the parameter space for θ, and let k denote the dimension of θ.

Note that we are assuming the candidate model to be of the same parametric family as the generating model. This assumption is implied from a regularity condition needed to apply the theory used in the development of AIC and AICc. We will discuss this condition shortly.

A well-known measure of separation between the generating model and the candidate model is the Kullback-Leibler divergence (Kullback, 1968). For the i^{th} case y_i, such a discrepancy is defined as

$$d_i(\theta, \theta_o) = E_o\{-2 \ln g_i(y_i|\theta)\},$$

where E_o denotes the expectation under the generating model.

For the candidate model, let $\hat{\theta}$ denote the maximum likelihood estimator of θ. A useful measure of separation between the generating model and the fitted candidate model is the overall Kullback-Leibler discrepancy, which is given by

$$\begin{aligned} d_{AIC}(\hat{\theta}, \theta_o) &= \sum_{i=1}^{n} d_i(\hat{\theta}, \theta_o) \\ &= \sum_{i=1}^{n} E_o\{-2\ln g_i(y_i|\theta)\}|_{\theta=\hat{\theta}}. \end{aligned} \tag{2.1}$$

Thus, the expected overall Kullback-Leibler discrepancy is

$$\begin{aligned} \Delta_{AIC}(\theta_o, k) &= E_o\{d_{AIC}(\hat{\theta}, \theta_o)\} \\ &= \sum_{i=1}^{n} E_o\{E_o\{-2\ln g_i(y_i|\theta)\}|_{\theta=\hat{\theta}}\}. \end{aligned} \tag{2.2}$$

Model selection criteria based on $d_{AIC}(\hat{\theta}, \theta_o)$ are developed by finding a statistic that has an expectation which is equal to (2.2) (or at least approximately equal). Thus, the targeted measure is $\Delta_{AIC}(\theta_o, k)$. Note that the overall discrepancy (2.1) is not a statistic since its evaluation requires knowledge of θ_o. Therefore, $d_{AIC}(\hat{\theta}, \theta_o)$ cannot be used to estimate $\Delta_{AIC}(\theta_o, k)$. Yet the empirical log-likelihood measure

$$-2\ln L(\hat{\theta}\ |y) = \sum_{i=1}^{n}\{-2\ln g_i(y_i|\hat{\theta})\} \tag{2.3}$$

is a statistic and thus can be used to estimate $\Delta_{AIC}(\theta_o, k)$. However, (2.3) serves as a negatively biased estimator of $\Delta_{AIC}(\theta_o, k)$. If we write

$$\begin{aligned} \Delta_{AIC}(\theta_o, k) &= E_o\{d_{AIC}(\hat{\theta}, \theta_o)\} \\ &= E_o\{-2\ln L(\hat{\theta}\ |y)\} + E_o\{d_{AIC}(\hat{\theta}, \theta_o) - \{-2\ln L(\hat{\theta}\ |y)\}\}, \end{aligned}$$

then the bias adjustment

$$E_o\{d_{AIC}(\hat{\theta}, \theta_o) - \{-2\ln L(\hat{\theta}\ |y)\}\}$$

is referred to as the expected optimism (Efron, 1983, 1986) in judging the fit of a model using the same data as that which was used to construct the fit.

A general approach for estimation of the expected optimism was developed by Akaike (1973, 1974) under the assumption that the candidate family includes the generating model (see Linhart and Zucchini, 1986, p. 245). Thus, we require that $L(\theta \mid y)$ subsumes $L(\theta_o \mid y)$, or that $\theta_o \in \Theta$. This regularity condition has two implications that warrant discussion. First, as we have seen, it is taken that the candidate parametric family is appropriately specified. The problem of model selection is then reduced to one of choosing the correct model dimension. Normal theory linear model selection provides a rich set of examples where model selection is focused on dimension selection. The second implication regarding this regularity condition is that results only apply to "overfit models," since the candidate likelihood must subsume the generating likelihood. An overall discrepancy, such as (2.1), can be decomposed as a discrepancy due to estimation and a discrepancy due to approximation. When a model is overfit, it contains all parameters needed to define the true model. Here, approximation error is zero and overall discrepancy is due only to estimation error. Underfit models suffering from large approximation error are easy to distinguish by an obvious lack of fit. In a practical sense, model selection criteria are judged by their ability to distinguish between models with little or no approximation error, precisely the condition for which Akaike-type criteria results will apply.

Yet cross validation procedures do not require this assumption and are applicable in a wider range of settings than the Akaike criteria.

Under the regularity condition, maximum likelihood asymptotic theory can be used to argue that in large-sample settings, the expected optimism can be estimated by $2k$, which is twice the dimension of $\hat{\theta}$. Therefore, the expected value of

$$\text{AIC} = -2 \ln L(\hat{\theta} \mid y) + 2k$$

should be asymptotically near the value of the expected overall discrepancy, (2.2). Specifically, one can establish that

$$E_o\{\text{AIC}\} + o(1) = \Delta_{AIC}(\theta_o, k). \tag{2.4}$$

We note that the result (2.4) extends beyond the linear regression setting to many other modeling frameworks.

AIC provides us with an approximately unbiased estimator of $\Delta_{AIC}(\theta_o, k)$ in settings where n is large and k is small. In other settings, $2k$ may be much smaller than the expected optimism, making AIC substantially negatively biased as an estimator of $\Delta_{AIC}(\theta_o, k)$. To correct for this negative bias, Hurvich and Tsai (1989) proposed "corrected" AIC for linear and nonlinear regression.

Suppose that the generating model for the data is given by

$$y = X\beta_o + \epsilon, \qquad \epsilon \sim N_n(0, \sigma_o^2\ I), \tag{2.5}$$

and that the candidate model postulated for the data is of the form

$$y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2\ I). \tag{2.6}$$

Here y is an $(n \times 1)$ observation vector, ϵ is an $(n \times 1)$ error vector, β_o and β are $(p \times 1)$ parameter vectors, and X is an $(n \times p)$ design matrix of full column rank. Assume β_o is such that for some $0 < p_o \leq p$, the last $(p - p_o)$ components of β_o are zero. Thus, model (2.5) is nested within model (2.6). The development for AICc requires the same condition as that which was imposed by Akaike. Let θ_o and θ respectively denote the $k = (p+1)$ dimensional vectors $(\beta_o^{'}, \sigma_o^2)^{'}$ and $(\beta^{'}, \sigma^2)^{'}$. Note that the nesting ensures that θ_o is an element of Θ, or that $L(\theta\ |y)$ subsumes $L(\theta_o\ |y)$.

Let $\hat{\beta}$ denote the least squares estimator of β, and let $\hat{\sigma}^2 = (y - X\hat{\beta})^{'}(y - X\hat{\beta})/n$. Hurvich and Tsai (1989, p. 300) define AICc as

$$\text{AICc} = n \ln \hat{\sigma}^2 + \frac{n(n+p)}{n-p-2}. \tag{2.7}$$

One can prove that in the linear regression setting,

$$E_o\{\text{AICc}\} = \Delta_{AIC}(\theta_o, k), \tag{2.8}$$

establishing that AICc is exactly unbiased for $\Delta_{AIC}(\theta_o, k)$. (The preceding holds up to $o(1)$ for other modeling frameworks in which AICc has been justified and developed.)

The derivation of AIC and verification of (2.4) in a general setting (Cavanaugh, 1997) along with the derivation of AICc and verification of (2.8) (Cavanaugh, 1997) clearly illustrates the way in which AICc improves upon the approximations leading to AIC. Furthermore, Davies, Neath and Cavanaugh (2005) show that AICc is the minimum variance unbiased estimator of the expected overall Kullback-Leibler discrepancy.

3. The Predictive Divergence Criterion (PDC)

In the cross validation setting, for the candidate model, let $\hat{\theta}_{-i}$ denote the maximum likelihood estimator of θ based on excluding the i^{th} case y_i from the vector y. For PDC, define the overall discrepancy to be

$$\begin{aligned} d_{PDC}(y, \theta_o) &= \sum_{i=1}^{n} d_i(\hat{\theta}_{-i}, \theta_o) \\ &= \sum_{i=1}^{n} E_o\{-2 \ln g_i(y_i|\theta)\}|_{\theta=\hat{\theta}_{-i}}. \end{aligned} \tag{3.1}$$

Thus, the expected overall PDC discrepancy is

$$\begin{aligned} \Delta_{PDC}(\theta_o, k) &= E_o\{d_{PDC}(y, \theta_o)\} \\ &= \sum_{i=1}^{n} E_o\{E_o\{-2 \ln g_i(y_i|\theta)\}|_{\theta=\hat{\theta}_{-i}}\}. \end{aligned} \tag{3.2}$$

Note that (3.1) and (3.2) are respectively analogous to (2.1) and (2.2).

A model selection criterion based on $d_{PDC}(y, \theta_o)$ is constructed by finding a statistic that has an expectation which is equal to (3.2) (or at least approximately equal). Thus, the targeted measure here is $\Delta_{PDC}(\theta_o, k)$. Again, the overall discrepancy (3.1) is not a statistic and cannot be used to estimate $\Delta_{PDC}(\theta_o, k)$. Yet the case-deleted empirical log-likelihood measure

$$\sum_{i=1}^{n} -2 \ln g_i(y_i|\hat{\theta}_{-i})$$

is a statistic and can therefore be used to estimate $\Delta_{PDC}(\theta_o, k)$ Moreover, this statistic is exactly unbiased for $\Delta_{PDC}(\theta_o, k)$, since

$$\begin{aligned} \Delta_{PDC}(\theta_o, k) &= E_o\{d_{PDC}(y, \theta_o)\} \\ &= \sum_{i=1}^{n} E_o\{E_o\{-2 \ln g_i(y_i|\theta)\}|_{\theta=\hat{\theta}_{-i}}\} \\ &= E_o\left\{\sum_{i=1}^{n} -2 \ln g_i(y_i|\hat{\theta}_{-i})\right\}. \end{aligned}$$

Note that in the preceding, for the second expression on the right-hand side, the inner expectation is taken with respect to the distribution of y_i, and the outer expectation is taken with respect to the sampling distribution of $\hat{\theta}_{-i}$.

Thus, we define the predictive divergence criterion

$$\mathrm{PDC} = \sum_{i=1}^{n} -2 \ln g_i(y_i|\hat{\theta}_{-i}).$$

PDC serves as an exactly unbiased estimator of $\Delta_{PDC}(\theta_o, k)$ regardless of the sample size, the relationship between $L(\theta_o \;|y)$ and $L(\theta \;|y)$ or the distribution of the underlying data.

For the linear regression setting in Section 2, PDC is defined as

$$\mathrm{PDC} = \sum_{i=1}^{n} \left\{ \ln \hat{\sigma}^2_{-i} + \frac{(y_i - \hat{y}_{i,-i})^2}{\hat{\sigma}^2_{-i}} \right\},$$

where $\hat{y}_{i,-i}$ is the fitted value for y_i with the i^{th} observation deleted from the data, and $\hat{\sigma}^2_{-i}$ is the maximum likelihood estimate of the variance with the i^{th} observation deleted from the data.

4. The Minimum Variance Unbiased Estimator of the Expected Overall PDC Discrepancy

We now establish the expected value of PDC in the linear regression setting. This will provide us with an expression for $\Delta_{PDC}(\theta_o, k)$.

For this derivation and subsequent justifications, we return to the regularity condition used in the development of the Akaike criteria. Thus, we assume henceforth that $L(\theta \;|y)$ subsumes $L(\theta_o \;|y)$. Imposing this condition limits the generality of the results, yet ensures mathematical tractability and is methodologically defensible for the reasons stated in Section 2.

Consider

$$\Delta_{PDC}(\theta_o, k) \;\; = \;\; \sum_{i=1}^{n} \left[E_o\{\ln \hat{\sigma}^2_{-i}\} + E_o \left\{ \frac{(y_i - \hat{y}_{i,-i})^2}{\hat{\sigma}^2_{-i}} \right\} \right]. \qquad (4.1)$$

To evaluate (4.1), we make use of the fact that $\{((n-1)\hat{\sigma}^2_{-i})/\sigma^2_o\}$ has a chi-square distribution with $(n - p - 1)$ degrees of freedom. From this, we have

$$E_o \left\{ \ln \frac{(n-1)\hat{\sigma}^2_{-i}}{\sigma^2_o} \right\} = \ln 2 + \psi \left(\frac{n-p-1}{2} \right) \qquad (4.2)$$

and

$$E_o \left\{ \frac{\sigma^2_o}{(n-1)\hat{\sigma}^2_{-i}} \right\} = \frac{1}{n-p-3}, \qquad (4.3)$$

where ψ is the *digamma* or *psi* function. (An overview of this function can be found in the Appendix.) Also, to evaluate $E_o\{(y_i - \hat{y}_{i,-i})^2\}$, we can make use of the fact that the i^{th} deleted residual is defined by $e_{i,-i} = y_i - \hat{y}_{i,-i} = e_i/(1 - h_{ii})$, where e_i is the residual for case i based on the full data y and h_{ii} is the i^{th} diagonal element of the hat matrix, $H = X(X'X)^{-1}X'$. Using the expectation and variance of $e_{i,-i}$, we have

$$E_o\{(y_i - \hat{y}_{i,-i})^2\} \;=\; \frac{\sigma_o^2}{1 - h_{ii}}. \tag{4.4}$$

Utilizing results from (4.2) and (4.3), together with (4.4) and the independence of $\hat{\sigma}^2_{-i}$ and $(y_i - \hat{y}_{i,-i})^2$, we obtain the following expression for (4.1):

$$\Delta_{PDC}(\theta_o, k) \;=\; n\ln\sigma_o^2 + n\ln\frac{2}{n-1} + n\psi\left(\frac{n-p-1}{2}\right) + \frac{n-1}{n-p-3}\sum_{i=1}^{n}\frac{1}{1-h_{ii}}. \tag{4.5}$$

Davies, Neath and Cavanaugh (2005) show that a simplified expression for the expected overall Kullback-Leibler discrepancy is given by

$$\Delta_{AIC}(\theta_o, k) = n\ln\sigma_o^2 + n\ln\frac{2}{n} + n\psi\left(\frac{n-p}{2}\right) + \frac{n(n+p)}{n-p-2}. \tag{4.6}$$

The similarity of expressions (4.5) and (4.6) suggests that for large samples

$$\Delta_{PDC}(\theta_o, k) \approx \Delta_{AIC}(\theta_o, k).$$

This is established formally in the next section.

The expected overall discrepancy for PDC can be written

$$\begin{aligned}\Delta_{PDC}(\theta_o, k) \;=\;& \Delta_{AIC}(\theta_o, k) + n\ln\frac{n}{n-1} + n\left\{\psi\left(\frac{n-p-1}{2}\right) - \psi\left(\frac{n-p}{2}\right)\right\}\\ &-\frac{n(n+p)}{n-p-2} + \frac{n-1}{n-p-3}\sum_{i=1}^{n}\frac{1}{1-h_{ii}}.\end{aligned} \tag{4.7}$$

So $\Delta_{PDC}(\theta_o, k) = \Delta_{AIC}(\theta_o, k) + c$, where c represents the remainder of the terms in (4.7). Since AICc is the minimum variance unbiased estimator of $\Delta_{AIC}(\theta_o, k)$, the minimum variance unbiased estimator of $\Delta_{PDC}(\theta_o, k)$ is AICc $+\, c$. Equivalently, using (2.7), we define the adjusted predictive divergence criterion PDCa* as

$$\begin{aligned}\text{PDCa}^* \;=\;& n\ln\hat{\sigma}^2 + n\ln\frac{n}{n-1} + n\left\{\psi\left(\frac{n-p-1}{2}\right) - \psi\left(\frac{n-p}{2}\right)\right\}\\ &+\frac{n-1}{n-p-3}\sum_{i=1}^{n}\frac{1}{1-h_{ii}}.\end{aligned} \tag{4.8}$$

PDCa* serves as the minimum variance unbiased estimator of $\Delta_{PDC}(\theta_o, k)$. For practical purposes, we can exclude

$$n \ln \frac{n}{n-1} + n \left\{ \psi \left(\frac{n-p-1}{2} \right) - \psi \left(\frac{n-p}{2} \right) \right\} \tag{4.9}$$

from formula (4.8), and define the criterion as

$$\text{PDCa} = n \ln \hat{\sigma}^2 + \frac{n-1}{n-p-3} \sum_{i=1}^{n} \frac{1}{1-h_{ii}},$$

where $n \ln \hat{\sigma}^2$ is the goodness-of-fit term and $\frac{n-1}{n-p-3} \sum_{i=1}^{n} \frac{1}{1-h_{ii}}$ is the penalty term. The exclusion of (4.9) from the PDCa* formula in (4.8) is valid since we show in the next section that (4.9) is asymptotically equal to zero. Moreover, one can establish that convergence to zero is at a rate of $O(1/n)$. Our simulation results in Section 6 will show that PDCa* and PDCa have similar behavioral properties, even in small-sample applications.

5. Asymptotic Equivalence

From (4.5) and (4.6) we have

$$\begin{aligned} \Delta_{PDC}(\theta_o, k) - \Delta_{AIC}(\theta_o, k) &= n \ln \frac{n}{n-1} + n \left\{ \psi \left(\frac{n-p-1}{2} \right) - \psi \left(\frac{n-p}{2} \right) \right\} \\ &\quad + \left\{ \frac{n-1}{n-p-3} \sum_{i=1}^{n} \frac{1}{1-h_{ii}} - \frac{n(n+p)}{n-p-2} \right\}. \end{aligned} \tag{5.1}$$

We will show that the difference (5.1) approaches zero asymptotically (as $n \to \infty$ and p is held constant).

Making use of L'Hospital's rule for the first term in (5.1), one can establish that

$$n \ln \frac{n}{n-1} = 1 + o(1). \tag{5.2}$$

For the second term in (5.1), we can make use of the following approximation for the digamma function (see (A.1) in the Appendix):

$$\psi(\nu) = \ln \nu - \frac{1}{2\nu} + O\left(\frac{1}{\nu^2}\right),$$

for $\nu > 1$. Using the preceding expression, we have

$$\begin{aligned} \psi\left(\nu - \frac{1}{2}\right) - \psi(\nu) &= \ln\left(\nu - \frac{1}{2}\right) - \frac{1}{2(\nu - \frac{1}{2})} - \ln \nu + \frac{1}{2\nu} + O\left(\frac{1}{\nu^2}\right) \\ &= \ln\left(1 - \frac{1}{2\nu}\right) + O\left(\frac{1}{\nu^2}\right). \end{aligned} \tag{5.3}$$

Based on (5.3), we can write

$$\begin{aligned} n\left\{\psi\left(\frac{n-p-1}{2}\right)-\psi\left(\frac{n-p}{2}\right)\right\} &= n\ln\left(1-\frac{1}{n-p}\right)+O\left(\frac{n}{(n-p)^2}\right) \\ &= -1+o(1). \end{aligned} \tag{5.4}$$

For the final term in (5.1), we need to examine the asymptotic behavior of $\sum_{i=1}^{n} \frac{1}{1-h_{ii}}$. Recall the property that

$$\sum_{i=1}^{n} h_{ii} = p, \tag{5.5}$$

and thus the h_{ii}'s tend to decrease as n increases. Let $h_{max} = \max_i h_{ii}$. Consider a Taylor series expansion in h_{ii} of $f(h_{ii}) = \frac{1}{1-h_{ii}}$ about the point $h_{ii} = 0$. We have

$$\frac{1}{1-h_{ii}} = 1 + h_{ii} + \frac{h_{ii}^2}{(1-h_{ii}^*)^3}, \tag{5.6}$$

where $0 < h_{ii}^* < h_{ii}$. Summing each of the terms in (5.6) from 1 to n and using (5.5) leads to the inequality

$$(n+p) < \sum_{i=1}^{n} \frac{1}{1-h_{ii}} < (n+p) + \frac{1}{(1-h_{max})^3} \ \ (\sum_{i=1}^{n} h_{ii}^2). \tag{5.7}$$

Focusing on the right-hand side of (5.7), and again using (5.5), we can write

$$\frac{1}{(1-h_{max})^3} \ \ (\sum_{i=1}^{n} h_{ii}^2) \ \ \leq \ \ \frac{h_{max}}{(1-h_{max})^3} \ \ p. \tag{5.8}$$

We will now employ Huber's condition (Huber, 1981, p. 164), which states

$$h_{max} = o(1).$$

This condition along with (5.7) and (5.8) implies

$$\sum_{i=1}^{n} \frac{1}{1-h_{ii}} = (n+p) + o(1). \tag{5.9}$$

By (5.9), we have the following asymptotic simplification for the final term in (5.1):

$$\begin{aligned} \frac{n-1}{n-p-3}\sum_{i=1}^{n}\frac{1}{1-h_{ii}} - \frac{n(n+p)}{n-p-2} &= \frac{(n+p)(p+2)}{(n-p-3)(n-p-2)} + o(1) \\ &= o(1). \end{aligned} \tag{5.10}$$

Using the results from (5.2), (5.4), and (5.10), for the difference (5.1), we see that asymptotically

$$\Delta_{PDC}(\theta_o, k) - \Delta_{AIC}(\theta_o, k) \quad = \quad o(1).$$

The preceding establishes that for large samples, we have $\Delta_{PDC}(\theta_o, k) \approx \Delta_{AIC}(\theta_o, k)$, or equivalently $E_o\{\text{PDCa}\} \approx E_o\{\text{AICc}\}$. Consequently, one can see that PDCa, the minimum variance unbiased estimator of its target discrepancy $\Delta_{PDC}(\theta_o, k)$, is an asymptotically minimum variance unbiased estimator of $\Delta_{AIC}(\theta_o, k)$, the target discrepancy for AIC and AICc.

6. Simulations

Consider a setting where a sample of size n is generated from a true linear regression model, having a design matrix of rank p_o. Suppose our objective is to search among a candidate collection of nested families for the fitted model which serves as the best approximation to the truth. The strength of the approximation is reflected via the expected overall discrepancy, either Δ_{AIC} or Δ_{PDC}.

Assume that our candidate models are of the form (2.6), corresponding to design matrices of ranks $p = 2, 3, ..., P$, and that the design matrix of rank p_o $(2 < p_o < P)$ is correctly specified. Hence, fitted models for which $2 \leq p < p_o$ are underfit, and those for which $p_o < p \leq P$ are overfit. In all design matrices, we assume that the initial column is a vector consisting of all ones. We will refer to p as the *order* of the model, and to p_o as the *true order.*

We examine the behavior of AIC, AICc, PDC, and PDCa and their target discrepancies by simulating a setting where the criteria are used to select p. In each of the simulation sets, samples of data are generated from a true model of the form $y = 1 + x_1 + \ldots + x_{p_o-1} + \epsilon$. The true model errors, ϵ, are generated from a distribution with median 0 and standard deviation 2. The input variables $x_1, \ldots, x_{P-1}$ are generated independently from a uniform $(0, 10)$ distribution. The sets feature a simulation size of 5000 samples.

Set 1 features a sample size of $n = 15$, a candidate class size of $P = 6$, a true order of $p_o = 4$, and normally distributed errors. Results are displayed in Table 1(a,b,c) and Figure 1(a). PDCa and AICc choose the correct order on over 90% of the samples. Cross validation procedures such as PDC tend to be affected by small samples. PDC does select the correct model on over 80% of the samples, although it favors underfit models at a higher rate than AICc. The criterion based strictly on an asymptotic justification, AIC, performs poorly with only about 60% correct selections.

The simulation results for criterion PDCa* are presented to verify the claim that the exclusion of the terms in (4.9) has no practical consequences. Even with a sample size of only $n = 15$, the behavioral properties of PDC and PDCa* are similar.

Set 2 and Set 3 feature a slightly larger sample size with $n = 25$, a candidate class of size $P = 11$, and a true order of $p_o = 5$. The results are featured in Tables 2(a,b,c) and 3(a,b,c) and in Figure 1(b,c). In Set 2, we again consider normally distributed errors. PDCa obtains the most correct order selections (95%) among the criteria of interest. The probability of selecting underfit models has become negligible for the predictive divergence criteria. PDC (91%) outperforms AICc (89%) in this example. AIC and AICc tend to favor overspecified models relative to PDC and PDCa, with the overfitting tendencies of AIC much more extreme than those of AICc. Note from Figure 1(b) how Δ_{PDC} provides a greater distinction than Δ_{AIC} between the correctly specified model and overfit models.

The performance of the criteria when the true model is not in the candidate family can be investigated by simulating a nonnormal error distribution. For Set 3, we generate the errors from an exponential distribution with standard deviation 2, shifted to a median of 0. A definite right-skewness to the errors is now present. The candidate models will continue to be based on the assumption of normally distributed errors.

It has been mentioned that a merit of PDC as a selection criterion is that the true model need not be represented among the candidate models to achieve unbiasedness. Table 3(a,b,c) and Figure 1(c) show the results of Set 3. As indicated in Table 3(c), PDCa now demonstrates a marked bias, underestimating the true prediction error Δ_{PDC} as reflected by the averages for PDC. Although the values of Δ_{AIC} and the averages of AIC and AICc are not featured, these results also indicate a substantial negative bias.

Curiously, PDCa still outperforms PDC in terms of correct selections: although PDCa is negatively biased for Δ_{PDC}, it estimates the target discrepancy with less variability. The use of PDCa to measure prediction error is inappropriate here due to model misspecification.

Set 4 and Set 5 demonstrate the effect of increasing sample size on the selection criteria. These sets feature a candidate class of size $P = 13$, a true order of $p_o = 6$, and sample sizes of $n = 50$ (Set 4) and $n = 100$ (Set 5). We again consider normally distributed errors. The results are displayed in Tables 4(a,b,c) and 5(a,b,c) and in Figure 2(a,b). First, note from Figure 2 how $\Delta_{PDC} - \Delta_{AIC}$ converges to zero, as established in Section 5.

The behavior of a model selection criterion is governed by its underlying discrepancy. If the focus is on predictive ability (PDC, PDCa), or accurate fitted values (AIC, AICc), then the probability of correct order selection may not converge to 1 as the sample size increases (Shao, 1993). Note, however, that the relative difference between the discrepancy for the correctly specified model and the discrepancy for an overfit model becomes smaller as the sample size becomes larger. An overfit model is incorrect only due to estimation error, which tapers off as n increases. Predictive ability among overfit models is then nearly identical. The need and ability to select the true order as opposed to a higher order diminishes for large samples, as evidenced by the discrepancy Δ_{PDC}.

Referring to Tables 1(c), 2(c), 4(c), 5(c), we see $E_o\{\text{PDC}\}$ and $E_o\{\text{PDCa}\}$ are essentially equal. PDCa is more accurate as an estimator of Δ_{PDC}. PDC and other cross validation methods have an advantage in that a correct specification of the form of the true model is not required. However, when this form can be characterized and corresponding assumptions can be imposed, we show through the development of PDCa how one can improve upon cross validation.

Acknowledgments

The authors would like to express their appreciation to the referees for providing thoughtful and insightful comments which helped to improve the original version of this manuscript.

Appendix: The Digamma Function

The *digamma* or *psi* function $\psi(\nu)$ can be defined in terms of the gamma function $\Gamma(\nu)$ $(\nu > 0)$:

$$\psi(\nu) = \frac{d}{d\nu} \ln \Gamma(\nu) = \frac{\Gamma'(\nu)}{\Gamma(\nu)}.$$

This function can be expressed as

$$\psi(\nu) = -C - \sum_{j=0}^{\infty} \left(\frac{1}{j+\nu} - \frac{1}{j+1} \right), \qquad \nu > 0,$$

where $C = 0.577215664901\ldots$ is Euler's constant.

It can be shown (Gradshteyn and Ryzhik, 1965, p. 546) that

$$\int_0^\infty z^{\nu-1} e^{-\mu z} (\ln z)\ dz = \frac{1}{\mu^\nu} \Gamma(\nu) \left[\psi(\nu) - (\ln \mu) \right], \qquad \mu > 0, \nu > 0.$$

Thus, if χ^2 has central chi-square distribution with d degrees of freedom, then

$$\begin{aligned} E\left\{\ln \chi^2\right\} &= \int_0^\infty (\ln z) \left\{ \frac{1}{2^{d/2}\Gamma(d/2)} \right\} z^{d/2-1} e^{-z/2}\ dz \\ &= \ln 2 + \psi\left(\frac{d}{2}\right). \end{aligned}$$

For integer $\nu \geq 1$, values of $\psi(\nu)$ can be found via

$$\psi(\nu) = -C + \sum_{k=1}^{\nu-1} \frac{1}{k}.$$

For real $\nu \geq 1$, values of $\psi(\nu)$ can be approximated to any degree of accuracy via the expansion

$$\begin{aligned} \psi(\nu) &= \ln \nu - \frac{1}{2\nu} - \sum_{k=1}^{\infty} \frac{B_{2k}}{2k\nu^{2k}} \\ &= \ln \nu - \frac{1}{2\nu} - \frac{1}{12\nu^2} + \frac{1}{120\nu^4} - \frac{1}{252\nu^6} + \ldots, \end{aligned}$$

where the B_{2k} are *Bernoulli numbers.* This expansion leads to the useful approximation

$$\psi(\nu) = \ln \nu - \frac{1}{2\nu} + O\left(\frac{1}{\nu^2}\right). \tag{A.1}$$

For further discussion of the digamma function, see Abramowitz and Stegun (1972, pp. 258-259).

Table 1(a). Order selections for AIC, AICc, PDC, PDCa, PDCa*: Set 1.

p	AIC	AICc	PDC	PDCa	PDCa*
2	0	17	239	87	79
3	3	46	433	163	152
4	3070	4618	4092	4702	4716
5	898	265	202	47	52
6	1029	54	34	1	1

Table 1(b). Percentages of underfit, correctly fit, and overfit models selected by AIC, AICc, PDC, PDCa, PDCa*: Set 1.

Type of Fitted Model	AIC	AICc	PDC	PDCa	PDCa*
Underfit	00.06%	1.26%	13.44%	5.00%	4.62%
Correctly Fit	61.40%	92.36%	81.84%	94.04%	94.32%
Overfit	38.54%	6.38%	4.72%	0.96%	1.06%

Table 1(c). Averages and standard deviations (in parentheses) for PDC and PDCa: Set 1.

p	2	3	4	5	6
PDC	66.5	62.7	52.5	61.6	75.8
	(6.0)	(7.2)	(11.6)	(17.0)	(24.0)
PDCa	66.7	62.9	51.4	60.0	73.5
	(5.5)	(5.7)	(6.7)	(7.1)	(7.7)

Table 2(a). Order selections for AIC, AICc, PDC, PDCa, PDCa*: Set 2.

p	AIC	AICc	PDC	PDCa	PDCa*
2	0	0	1	0	0
3	0	0	2	0	0
4	0	1	12	4	3
5	2637	4433	4570	4726	4714
6	540	367	308	224	229
7	367	115	76	38	44
8	333	57	25	7	9
9	297	18	5	1	1
10	338	8	1	0	0
11	488	1	0	0	0

Table 2(b). Percentages of underfit, correctly fit, and overfit models selected by AIC, AICc, PDC, PDCa, PDCa*: Set 2.

Type of Fitted Model	AIC	AICc	PDC	PDCa	PDCa*
Underfit	0.00%	0.02%	0.30%	0.08%	0.06%
Correctly Fit	52.74%	88.66%	91.40%	94.52%	94.28%
Overfit	47.26%	11.32%	8.30%	5.40%	5.66%

Table 2(c). Averages and standard deviations (in parentheses) for PDC and PDCa: Set 2.

p	2	3	4	5	6	7	8	9	10	11
PDC	113.5	107.2	96.9	72.6	76.9	82.2	88.8	97.3	108.1	122.2
	(6.9)	(7.1)	(7.0)	(9.0)	(9.8)	(10.7)	(12.0)	(13.6)	(16.1)	(20.0)
PDCa	113.6	107.4	97.3	72.2	76.3	81.4	87.9	96.1	106.6	120.2
	(6.9)	(7.0)	(6.8)	(8.2)	(8.4)	(8.7)	(8.9)	(9.2)	(9.5)	(9.9)

Table 3(a). Order selections for AIC, AICc, PDC, PDCa, PDCa*: Set 3.

p	AIC	AICc	PDC	PDCa	PDCa*
2	0	0	43	0	0
3	1	3	96	4	4
4	0	7	508	12	12
5	2505	4403	3936	4736	4725
6	578	379	309	210	218
7	444	144	79	31	34
8	307	42	23	7	7
9	313	18	6	0	0
10	359	2	0	0	0
11	493	2	0	0	0

Table 3(b). Percentages of underfit, correctly fit, and overfit models selected by AIC, AICc, PDC, PDCa, PDCa*: Set 3.

Type of Fitted Model	AIC	AICc	PDC	PDCa	PDCa*
Underfit	0.02%	0.20%	12.94%	0.32%	0.32%
Correctly Fit	50.10%	88.06%	78.72%	94.72%	94.50%
Overfit	49.88%	11.74%	8.34%	4.96%	5.18%

Table 3(c). Averages and standard deviations (in parentheses) for PDC and PDCa: Set 3.

p	2	3	4	5	6	7	8	9	10	11
PDC	113.7	107.5	97.5	79.3	84.1	90.0	97.3	106.7	118.4	133.6
	(7.5)	(7.9)	(9.4)	(27.8)	(29.3)	(31.1)	(33.1)	(36.3)	(39.1)	(43.3)
PDCa	113.7	107.6	97.3	70.0	74.1	79.1	85.6	93.8	104.2	117.9
	(7.2)	(7.4)	(7.8)	(13.9)	(14.0)	(14.1)	(14.2)	(14.4)	(14.6)	(14.9)

Table 4(a). Order selections for AIC, AICc, PDC, PDCa, PDCa*: Set 4.

p	AIC	AICc	PDC	PDCa	PDCa*
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	3016	4073	4271	4348	4333
7	636	489	443	400	410
8	388	222	186	157	159
9	241	106	57	55	55
10	197	60	25	26	27
11	179	22	11	10	12
12	155	17	5	4	4
13	188	11	2	0	0

Table 4(b). Percentages of underfit, correctly fit, and overfit models selected by AIC, AICc, PDC, PDCa, PDCa*: Set 4.

Type of Fitted Model	AIC	AICc	PDC	PDCa	PDCa*
Underfit	0.00%	0.00%	0.00%	0.00%	0.00%
Correctly Fit	60.32%	81.46%	85.42%	86.96%	86.66%
Overfit	39.68%	18.54%	14.58%	13.04%	13.34%

Table 4(c). Averages and standard deviations (in parentheses) for PDC and PDCa: Set 4.

p	2	3	4	5	6	7	8	9	10	11	12	13
PDC	234.7	223.6	208.2	184.0	129.8	132.0	134.4	137.2	140.2	143.6	147.3	151.4
	(9.7)	(9.7)	(9.6)	(9.2)	(11.0)	(11.2)	(11.4)	(11.5)	(11.7)	(12.0)	(12.3)	(12.6)
PDCa	234.9	223.8	208.6	184.5	129.6	131.8	134.2	136.9	139.8	143.1	146.8	150.9
	(9.7)	(9.7)	(9.6)	(9.2)	(11.0)	(11.1)	(11.2)	(11.3)	(11.4)	(11.6)	(11.8)	(11.9)

Table 5(a). Order selections for AIC, AICc, PDC, PDCa, PDCa*: Set 5.

p	AIC	AICc	PDC	PDCa	PDCa*
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	3436	3899	4032	4069	4058
7	576	523	491	482	486
8	305	237	221	199	202
9	208	143	125	124	125
10	163	83	60	62	63
11	119	56	39	33	34
12	92	34	24	15	16
13	101	25	8	16	16

Table 5(b). Percentages of underfit, correctly fit, and overfit models selected by AIC, AICc, PDC, PDCa, PDCa*: Set 5.

Type of Fitted Model	AIC	AICc	PDC	PDCa	PDCa*
Underfit	0.00%	0.00%	0.00%	0.00%	0.00%
Correctly Fit	68.72%	77.98%	80.64%	81.38%	81.16%
Overfit	31.28%	22.02%	19.36%	18.62%	18.84%

Table 5(c). Averages and standard deviations (in parentheses) for PDC and PDCa: Set 5.

p	2	3	4	5	6	7	8	9	10	11	12	13
PDC	464.9	440.7	408.2	357.8	247.2	248.7	250.3	251.9	253.6	255.5	257.4	259.4
	(13.6)	(13.3)	(13.1)	(12.6)	(14.9)	(15.0)	(15.0)	(15.2)	(15.2)	(15.3)	(15.4)	(15.5)
PDCa	465.1	441.0	408.5	359.3	247.1	248.6	250.2	251.8	253.5	255.3	257.2	259.2
	(13.6)	(13.3)	(13.1)	(12.6)	(14.8)	(14.9)	(15.0)	(15.1)	(15.1)	(15.2)	(15.3)	(15.4)

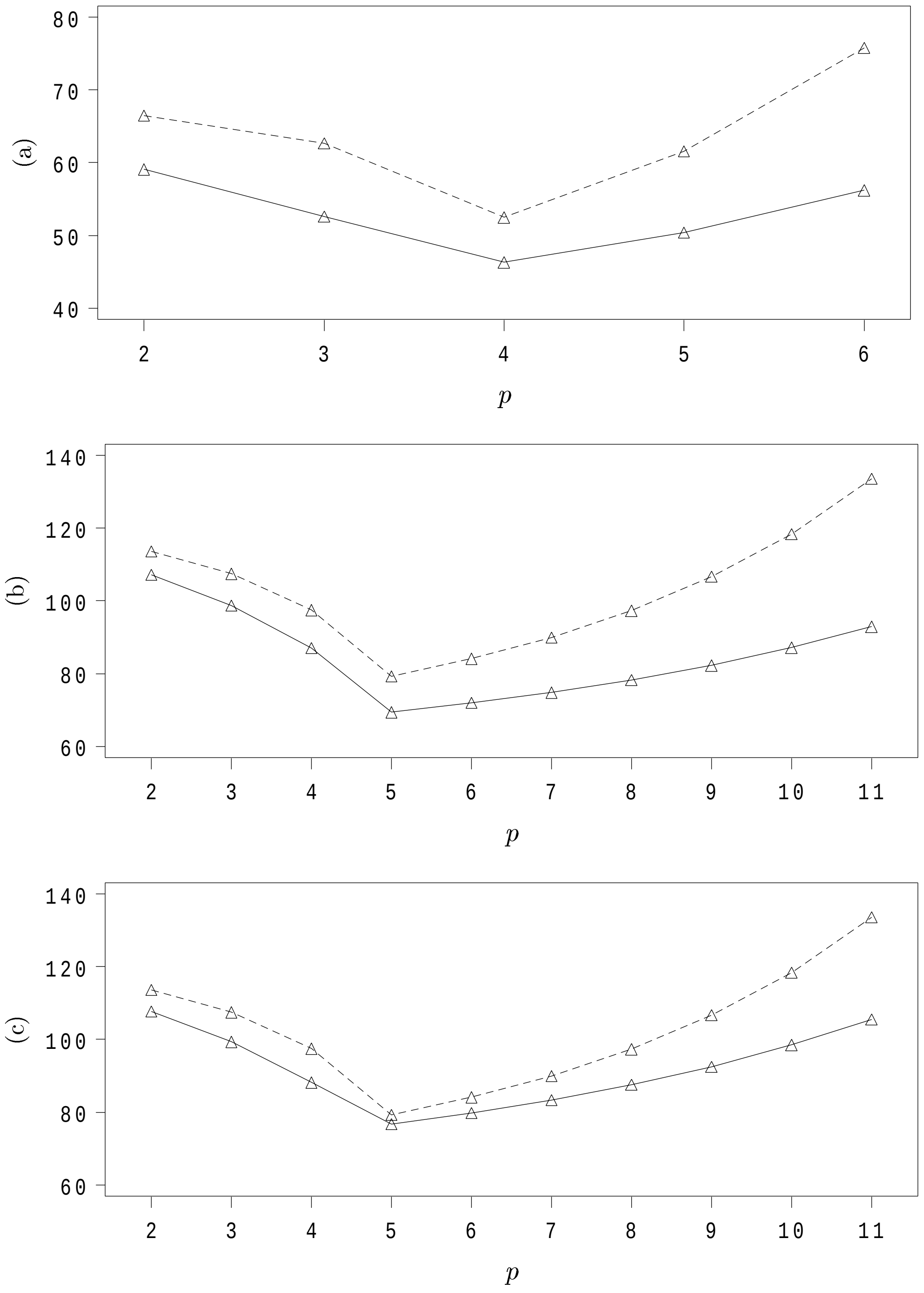

Figure 1. Δ_{AIC} (solid line) and Δ_{PDC} (dotted line): (a) Set 1; (b) Set 2; (c) Set 3.

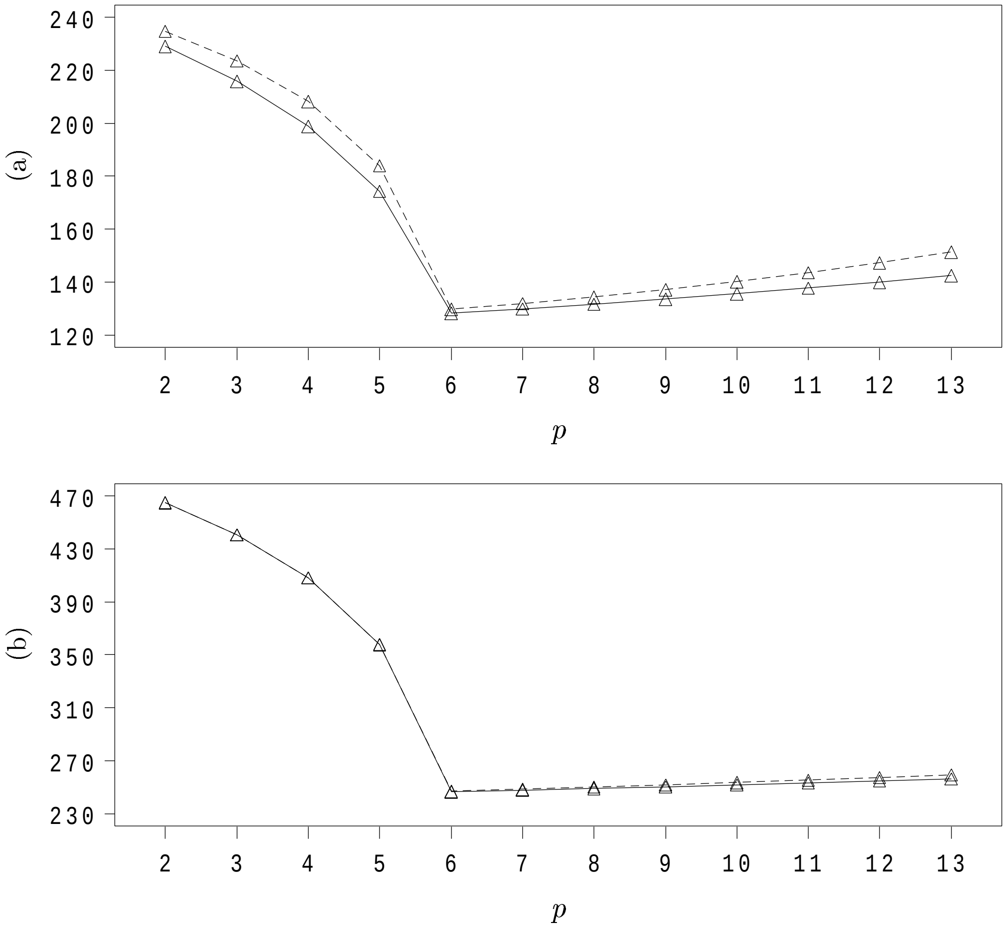

Figure 2. Δ_{AIC} (solid line) and Δ_{PDC} (dotted line): (a) Set 4; (b) Set 5.

References

Abramowitz, M. and Stegun, I.A., editors (1972). Psi (Digamma) Function. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, 9th printing (Dover, New York).

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

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

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

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

Davies, S. L. (2002). Discrepancy-based model selection criteria using cross-validation. Doctoral dissertation, Department of Statistics, University of Missouri - Columbia.

Davies, S. L., Neath A. A. and Cavanaugh, J. E. (2005). On the minimum variance unbiasedness property of AICc and MCp. Technical report, Department of Biostatistics, The University of Iowa.

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.

Gradshteyn, I. S. and Ryzhik, I. M. (1965). *Tables of Integrals, Series, and Products* (Academic Press, New York).

Huber, P. J. (1981). *Robust Statistics* (Wiley, New York).

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

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

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

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**, 486-495.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics* **A7**, 13-26.