An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion

M. Stone


Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281977%2939%3A1%3C44%3AAAEOCO%3E2.0.CO%3B2-A

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Journal of the Royal Statistical Society. Series B (Methodological) is published by Royal Statistical Society. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/rss.html.

Journal of the Royal Statistical Society. Series B (Methodological)
©1977 Royal Statistical Society

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR

http://www.jstor.org/
Tue May 20 10:55:54 2003
An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike’s Criterion

By M. Stone

University College London

[Received July 1976. Revised November 1976]

SUMMARY

A logarithmic assessment of the performance of a predicting density is found to lead to asymptotic equivalence of choice of model by cross-validation and Akaike’s criterion, when maximum likelihood estimation is used within each model.

Keywords: predicting density; model choice; Akaike’s information criterion; cross-validation

1. INTRODUCTION

Akaike (1973) proposed a criterion for model choice equivalent to the following: If \( \alpha \) indexes the model, choose \( \alpha \) to maximize

\[
L(\alpha, \hat{\theta}_{\alpha}) - p_{\alpha}
\]  

(1.1)

where \( L(\alpha, \theta_{\alpha}) \) is the log-likelihood function, \( \hat{\theta}_{\alpha} \) is the maximum likelihood estimate of the parameter \( \theta_{\alpha} \) in the model \( \alpha \) and \( p_{\alpha} \) is the dimensionality of \( \theta_{\alpha} \).

Akaike’s derivation of (1.1) was for hierarchical models but, as he finally remarked, this restriction is unnecessary. Looking at (1.1), we see \( p_{\alpha} \) as a correction term without which we would be maximizing \( L(\alpha, \hat{\theta}_{\alpha}) \); models with parameters of high dimensionality are given a severe handicap by this correction term.

For normal multiple linear regression models with known variance, \( \sigma^2 \), Mallows’ \( C_p \) (Gorman and Toman, 1966) is given by

\[
C_p = \frac{\text{RSS}_{\alpha}}{\sigma^2} - (n - 2p_{\alpha})
\]  

(1.2)

where \( \text{RSS}_{\alpha} \) is the residual sum of squares for model \( \alpha \) and \( n \) is the sample size. From (1.2) we see that maximizing (1.1) is equivalent to minimizing \( C_p \).

Akaike’s criterion stemmed from a recognition that unreserved maximization of likelihood provides an unsatisfactory method of choice between models that differ appreciably in their parametric dimensionality. Since the method of cross-validatory choice (Stone, 1974) is also concerned with the latter problem, it is perhaps unsurprising that a relationship can be established between the two approaches.

2. THE CHOICE PROBLEM

Adopting the notation of Stone (1974), we suppose we have a data-base

\[ S = \{(x_i, y_i), i = 1, \ldots, n\} \]

for \( n \) items and that our problem is the choice of predicting density for \( y \) given \( x \) from a prescribed class of formal predicting densities

\[
\{f(y|x, \alpha, S) \mid \alpha \in \mathcal{A}\},
\]  

(2.1)

whose members are indexed by the choice parameter \( \alpha \). All densities for \( y \) are with respect to a common fixed measure with generic element \( dy \). The operational interpretation of (2.1) is that the choice of \( \alpha \) specifies a predicting density of \( y \) for each \( x \), whose form depends in a prescribed way on \( S \). The notation is not intended to carry any other probabilistic interpretation.
It is useful to distinguish two complementary cases of (2.1):

**Case 1.** $f(y | x, \alpha, S) = f(y | x, \alpha)$ independent of $S$;

**Case 2.** $f(y | x, \alpha, S)$ properly dependent on $S$.

In Case 1, (2.1) becomes formally equivalent to a statistical model with $\alpha$ as conventional parameter. In Case 2, our attention will be focused on a general example which we will call Example A after Akaike (1973). Its prescription is

$$f(y | x, \alpha, S) \equiv f_{\alpha}(y | x, \hat{\theta}_{\alpha}(S)),$$

where

$$\{f_{\alpha}(y | x, \theta_{\alpha}), \theta_{\alpha} \in \Theta_{\alpha}\}$$

are the densities for a conventional parametric model $\alpha$ and $\hat{\theta}_{\alpha}(S)$ is the supposed unique maximum likelihood estimator maximizing $L(\alpha, \theta_{\alpha}) = \sum_{i} \log f_{\alpha}(y_{i} | x_{i}, \theta_{\alpha})$.

3. LOG-DENSITY ASSESSMENT

Suppose $f^{(i)}(y)$, $i = 1, \ldots, n$, were presented as predicting densities for $y_{i}$, $i = 1, \ldots, n$, respectively. As a measure of their success, take the log-density assessment

$$A = \sum_{i} \log f^{(i)}(y_{i}).$$

Observe that $A$ is the logarithm of $\prod_{i} f^{(i)}(y_{i})$ which may be termed the predicting probability density evaluated at the observations.

For Case 1, use of $f^{(i)}(y) = f(y | x_{i}, \alpha)$, $i = 1, \ldots, n$, would have the assessment

$$A(\alpha) = \sum_{i} \log f(y_{i} | x_{i}, \alpha),$$

whence we see that choice of $\alpha$ to maximize $A(\alpha)$ would be equivalent to maximum likelihood “estimation” of $\alpha$ for the “log-likelihood” given by the right-hand side of (3.2). Thus Case 1 introduces no innovations.

For Case 2, it would be unrealistic to assess the choice of $\alpha$ with $f^{(i)}(y) = f(y | x_{i}, \alpha, S)$ because $S$ itself contains $y_{i}$. It is more realistic to use the cross-validatory

$$f^{(i)}(y) = f(y | x_{i}, \alpha, S_{-i})$$

where $S_{-i} = S - (x_{i}, y_{i})$. This gives us

$$A(\alpha) = \sum_{i} \log f(y_{i} | x_{i}, \alpha, S_{-i}).$$

We will show in the next section that for Example A, $A(\alpha)$, given by (3.3), is asymptotically equivalent, under weak conditions, to Akaike’s criterion (1.1), which, as we have seen, “corrects” maximum likelihood as a method of choice of model.

4. ASYMPTOTIC EQUIVALENCE

For simplicity, we treat $\alpha$ as fixed and omit it from the notation. Writing $I$ for $\log f$, with $f$ given by (2.2) and (2.3), $A$ in (3.3) equals $\sum_{i} I(y_{i} | x_{i}, \hat{\theta}(S_{-i}))$. With $L(\theta) = \sum_{i} I(y_{i} | x_{i}, \theta)$, we have that $\hat{\theta}(S)$ [\hat{\theta} for short] maximizes $L(\theta)$ and $\hat{\theta}(S_{-i})$ [\hat{\theta}_{-i} for short] maximizes $L(\theta) - I(y_{i} | x_{i}, \theta)$. We suppose that $\theta = (\theta_{1} \ldots \theta_{p})^{T} \in \Theta$ an open region of $R^{p}$ and that $f$ is twice-differentiable with respect to $\theta$. Write

$$I' = \left( \frac{\partial I}{\partial \theta_{1}} \ldots \frac{\partial I}{\partial \theta_{p}} \right)^{T}, \quad I'' = \left( \frac{\partial^{2} I}{\partial \theta_{i} \partial \theta_{j}} \right).$$
with similar notation for $L$. We suppose that $\theta$ and $\theta_{-i}$ are unique solutions of $L'(\theta) = 0$ and $L'(\theta_{-i}) = L'(y_i | x_i, \theta) = 0$ respectively. Then by Taylor's theorem

$$A = L(\theta) + \sum_i (\theta_{-i} - \theta)^T L'(y_i | x_i, \theta_{-i} + a_i(\theta_{-i} - \theta)), \quad (4.1)$$

$$L'(\theta_{-i}) = L^*(\theta + b_i(\theta_{-i} - \theta))(\theta_{-i} - \theta) \quad (4.2)$$

with $|a_i| \leq 1$, $|b_i| \leq 1$, $i = 1, \ldots, n$. Also

$$L'(\theta_{-i}) = L'(y_i | x_i, \theta_{-i}). \quad (4.3)$$

From (4.1), (4.2) and (4.3), supposing $L^*$ in (4.2) is invertible,

$$A = L(\theta) + \sum_i L'(y_i | x_i, \theta_{-i})^T [L^*(\theta + b_i(\theta_{-i} - \theta))]^{-1} L'(y_i | x_i, \theta + a_i(\theta_{-i} - \theta)). \quad (4.4)$$

Next suppose that $S$ is a random sample from some joint distribution $P$ of $(x, y)$. Let $E$ denote expectation with respect to $P$. With this supposition we can expect:

(i) $\hat{\theta} \xrightarrow{P} \theta_0$ as $n \to \infty$ where $\theta_0$ is the supposed unique value of $\theta$ maximizing $E[l(y | x, \theta)];$
(ii) $\hat{\theta}_{-i} \xrightarrow{P} \theta_0$ as $n \to \infty$ for $i = 1, 2, \ldots$;
(iii) $n^{-1} L^*(\theta + b_i(\theta_{-i} - \theta)) \xrightarrow{P} E[l'(y | x, \theta_0)] = L_{2i},$ say;
(iv) $n^{-1} \sum_i L'(y_i | x_i, \theta_{-i})^T \xrightarrow{P} E[l'(y | x, \theta_0) l'(y | x, \theta_0)^T] = L_1,$ say.

So we have, heuristically, established that $A$ is asymptotically

$$L(\hat{\theta}) + \text{trace}(L_{2i}^{-1} L_{1i}). \quad (4.5)$$

Since $\theta_0$ maximizes $E[l(y | x, \theta)],$ it follows that $E[l'(y | x, \theta)]$ is negative-definite. Hence the correction term in (4.5), written in the form $E[l'(y | x, \theta_0)^T L_{2i}^{-1} l'(y | x, \theta_0)]$ is seen to be negative. However, little more can be said about it without further assumptions of a statistical character. The key assumption that gives us our asymptotic equivalence with Akaike's criterion is: The conditional distribution of $y$ given $x$ in the distribution $P$ is $f(y | x, \theta^*)$ for some unique $\theta^* \in \Theta$, that is, the conventional model $\{f(y | x, \theta) : \theta \in \Theta\}$ is true. In fact, this assumption implies $\theta^* = \theta_0.$ For

$$E[l(y | x, \theta_0)] = E \left\{ \int f(y | x, \theta^*) \log f(y | x, \theta_0) \, dy \right\}$$

$$\leq E \left\{ \int f(y | x, \theta^*) \log f(y | x, \theta^*) \, dy \right\} = E[l(y | x, \theta^*)]$$

and $\theta_0$ is the supposed unique maximizer of $E[l(y | x, \theta)].$ Further, differentiating the identity $\int f(y | x, \theta) l'(y | x, \theta) \, dy = 0$ with respect to $\theta$, setting $\theta = \theta_0$ and taking expectations with respect to $x$, we find $L_1 = -L_{2i}$ (the well-known identity). Hence the correction term in (4.5) is trace $(-I_{p \times p}) = -p$ and asymptotically

$$A = L(\hat{\theta}) - p \quad (4.6)$$

which is identical to (1.1) once the missing $\alpha$'s are restored.

While the key assumption italicized above gives us the general equivalence, weaker assumptions will suffice for particular choices of $\{f_a(y | x, \theta_a), \theta_a \in \Theta_a\}$.

If we consider two models $\alpha_1, \alpha_2$ of type (2.3) with

$$\Theta_{\alpha_1} \subset \Theta_{\alpha_2}$$

and suppose that both are true, then it is well known that, under regularity conditions, $2(L(\alpha_2, \hat{\theta}_{\alpha_2}) - L(\alpha_1, \hat{\theta}_{\alpha_1}))$ is asymptotically $\chi^2$ with $d = p_{\alpha_2} - p_{\alpha_1}$ degrees of freedom. Hence, by
(4.6), $A(\alpha_2) - A(\alpha_1)$ is asymptotically $1/4 X_\alpha^2 - d$. This shows how the simpler model will be favoured by the choice criterion $A(\alpha)$.

ACKNOWLEDGEMENTS

I am indebted to Dr H. Tong for impressing on me the inevitability of a strong connection between cross-validation and the Akaike criterion; also to A. Mabbett for stimulating discussions and to referees for helpful suggestions.

REFERENCES

