



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

M. Stone

Journal of the Royal Statistical Society. Series B (Methodological), Volume 39, Issue 1 (1977), 44-47.

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

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 α indexes the model, choose α to maximize

$$L(\alpha, \hat{\theta}_\alpha) - p_\alpha, \quad (1.1)$$

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

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_α 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, σ^2 , Mallows' C_p (Gorman and Toman, 1966) is given by

$$C_p = (\text{RSS}_\alpha / \sigma^2) - (n - 2p_\alpha), \quad (1.2)$$

where RSS_α is the residual sum of squares for model α 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

$$S = \{(x_i, y_i), i = 1, \dots, n\}$$

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

$$\{f(y|x, \alpha, S), \alpha \in \mathcal{A}\}, \quad (2.1)$$

whose members are indexed by the choice parameter α . 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 α 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 α 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

$$f(y|x, \alpha, S) \equiv f_\alpha(y|x, \hat{\theta}_\alpha(S)), \tag{2.2}$$

where

$$\{f_\alpha(y|x, \theta_\alpha), \theta_\alpha \in \Theta_\alpha\} \tag{2.3}$$

are the densities for a conventional parametric model α 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 $f^{(i)}(y)$, $i = 1, \dots, n$, were presented as predicting densities for y_i , $i = 1, \dots, n$, respectively. As a measure of their success, take the log-density assessment

$$A = \sum_i \log f^{(i)}(y_i). \tag{3.1}$$

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 $f^{(i)}(y) = f(y|x_i, \alpha)$, $i = 1, \dots, n$, would have the assessment

$$A(\alpha) = \sum_i \log f(y_i|x_i, \alpha), \tag{3.2}$$

whence we see that choice of α to maximize $A(\alpha)$ would be equivalent to maximum likelihood "estimation" of α 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 α 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

$$A(\alpha) = \sum_i \log f(y_i|x_i, \alpha, S_{-i}). \tag{3.3}$$

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 α as fixed and omit it from the notation. Writing l for $\log f$, with f given by (2.2) and (2.3), A in (3.3) equals $\sum_i l(y_i|x_i, \hat{\theta}(S_{-i}))$. With $L(\theta) = \sum_j l(y_j|x_j, \theta)$, we have that $\hat{\theta}(S)$ [$\hat{\theta}$ for short] maximizes $L(\theta)$ and $\hat{\theta}(S_{-i})$ [$\hat{\theta}_{-i}$ for short] maximizes $L(\theta) - l(y_i|x_i, \theta)$. We suppose that $\theta = (\theta_1 \dots \theta_p)^T \in \Theta$ an open region of R^p and that f is twice-differentiable with respect to θ . Write

$$l' = \left(\frac{\partial l}{\partial \theta_1} \dots \frac{\partial l}{\partial \theta_p} \right)^T, \quad l'' = \left(\frac{\partial^2 l}{\partial \theta_i \partial \theta_j} \right)$$

with similar notation for L . We suppose that $\hat{\theta}$ and $\hat{\theta}_{-i}$ are unique solutions of $L'(\theta) = \mathbf{0}$ and $L'(\theta) - l'(y_i|x_i, \theta) = \mathbf{0}$ respectively. Then by Taylor's theorem

$$A = L(\hat{\theta}) + \sum_i (\hat{\theta}_{-i} - \hat{\theta})^T l'\{y_i|x_i, \hat{\theta} + a_i(\hat{\theta}_{-i} - \hat{\theta})\}, \tag{4.1}$$

$$L'(\hat{\theta}_{-i}) = L''\{\hat{\theta} + b_i(\hat{\theta}_{-i} - \hat{\theta})\}(\hat{\theta}_{-i} - \hat{\theta}) \tag{4.2}$$

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

$$L'(\hat{\theta}_{-i}) = l'(y_i|x_i, \hat{\theta}_{-i}). \tag{4.3}$$

From (4.1), (4.2) and (4.3), supposing L'' in (4.2) is invertible,

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

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 $n \rightarrow \infty$ where θ_0 is the supposed unique value of θ maximizing $E\{l(y|x, \theta)\}$;
- (ii) $\hat{\theta}_{-i} \xrightarrow{P} \theta_0$ as $n \rightarrow \infty$ for $i = 1, 2, \dots$;
- (iii) $n^{-1} L''(\hat{\theta} + b_i(\hat{\theta}_{-i} - \hat{\theta})) \xrightarrow{P} E\{l''(y|x, \theta_0)\} = L_2$, say;
- (iv) $n^{-1} \sum_i l'\{y_i|x_i, \hat{\theta} + a_i(\hat{\theta}_{-i} - \hat{\theta})\} l'(y_i|x_i, \hat{\theta}_{-i})^T \xrightarrow{P} E\{l'(y|x, \theta_0) l'(y|x, \theta_0)^T\} = L_1$, say.

So we have, heuristically, established that A is asymptotically

$$L(\hat{\theta}) + \text{trace}(L_2^{-1} L_1). \tag{4.5}$$

Since θ_0 maximizes $E\{l(y|x, \theta)\}$, it follows that $E\{l''(y|x, \theta_0)\}$ is negative-definite. Hence the correction term in (4.5), written in the form $E\{l'(y|x, \theta_0)^T L_2^{-1} l'(y|x, \theta_0)\}$ 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

$$\begin{aligned} E\{l(y|x, \theta_0)\} &= E\left\{ \int f(y|x, \theta^*) \log f(y|x, \theta_0) dy \right\} \\ &\leq E\left\{ \int f(y|x, \theta^*) \log f(y|x, \theta^*) dy \right\} = E\{l(y|x, \theta^*)\} \end{aligned}$$

and θ_0 is the supposed unique maximizer of $E\{l(y|x, \theta)\}$. Further, differentiating the identity $\int f(y|x, \theta) l'(y|x, \theta) dy = \mathbf{0}$ with respect to θ , setting $\theta = \theta_0$ and taking expectations with respect to x , we find $L_1 = -L_2$ (the well-known identity). Hence the correction term in (4.5) is $\text{trace}(-I_{p \times p}) = -p$ and asymptotically

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

which is identical to (1.1) once the missing α 's are restored.

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

If we consider two models α_1, α_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 χ^2 with $d = p_{\alpha_2} - p_{\alpha_1}$ degrees of freedom. Hence, by

(4.6), $A(\alpha_2) - A(\alpha_1)$ is asymptotically $\frac{1}{2}\chi_d^2 - d$. 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

- AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory* (B. N. Petrov and F. Czaki, eds), pp. 267-281. Budapest: Akademiai Kiadó.
- GORMAN, J. W. and TOMAN, R. J. (1966). Selection of variables for fitting equations to data. *Technometrics*, **8**, 27-51.
- STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with Discussion). *J. R. Statist. Soc. B*, **36**, 111-147.
-