Biometrika (1984), **71**, 2, *pp.* 353–60
Printed in Great Britain

An alternative method of cross-validation for the smoothing of density estimates

BY ADRIAN W. BOWMAN
Statistical Laboratory, University of Manchester, Manchester, U.K.

SUMMARY

Cross-validation with Kullback–Leibler loss function has been applied to the choice of a smoothing parameter in the kernel method of density estimation. A framework for this problem is constructed and used to derive an alternative method of cross-validation, based on integrated squared error, recently also proposed by Rudemo (1982). Hall (1983) has established the consistency and asymptotic optimality of the new method. For small and moderate sized samples, the performances of the two methods of cross-validation are compared on simulated data and specific examples.

Some key words: Consistency; Integrated squared error; Kernel density estimation; Kullback–Leibler; Optimality; Sensitivity.

1. INTRODUCTION

The technique of cross-validation has proved to be useful in a number of statistical problems where satisfactory or straightforward solutions have not been found by direct application of more classical methods. A particular example is the control of the degree of smoothing when estimating nonparametrically an unknown probability density or regression function. A related problem occurs in the smoothing necessary to obtain a consistent estimator of the spectral density function of a time series. Cross-validation has achieved some success in these areas and a formal framework for its application to density estimation is described in §2. However, problems have recently come to light with this technique as currently implemented in density estimation. Rudemo (1982) takes an alternative approach and proposes that the smoothness of an estimate should be dictated by minimizing an estimate of a quadratic risk function, showing that a very similar criterion is achieved by invoking a heuristic cross-validatory argument. In §3 of the present paper, prepared independently of Rudemo (1982), the same criterion is derived by applying the framework of §2 with an appropriate loss function. Recent asymptotic results for the two available methods of cross-validation are reviewed and used in §4 to guide a simulation study where the optimality of the new method is monitored. Some examples are given in §5.

The estimation of an unknown density function has generated a considerable literature over the past few years. The kernel method, adopted here, has proved to be one of the most popular approaches and is well reviewed by Fryer (1977). Briefly, an estimate $f_n(x)$ of the true density function $f(x)$ is constructed by placing a kernel function $K(x;\, x_i, h)$ over each observation in the data set, $D = \{x_1, \ldots, x_n\}$, assumed to be a random sample from the distribution in question. The kernel function $K(x;\, y, h)$ is itself a density function with location parameter y and scale parameter h. The density estimate is then given by

$$f_n(x;\, D, h) = \frac{1}{n} \sum_{i=1}^{n} K(x;\, x_i, h).$$

It is generally accepted that the choice of kernel function K is less crucial than the value given to h. The latter is usually referred to as the smoothing parameter or window width. An optimal value, h^*, of this parameter may be defined to minimize a suitable loss function such as mean integrated squared error, $E[\int\{f(x)-f_n(x)\}^2\,dx]$. Scott & Factor (1981) give a concise description of this approach. The adoption of such a loss function may be regarded simply as a convenient means of constructing a well-defined problem and in practice there is likely to be a range of values of h throughout which the shape of the density estimate changes only slightly. This is demonstrated by Fryer (1976) in the rather special case of the normal distribution. Subjective choices of h have been suggested, as by Scott, Tapia & Thompson (1977, pp. 353–8), but attention has recently been given to methods which use the data to provide an appropriate value of the smoothing parameter.

2. Kullback–Leibler cross-validation

One of the first of the fully adaptive methods of choice of smoothing parameter was proposed by Habbema, Hermans & Van den Broek (1974) and by Duin (1976) who approached the problem via the idea of likelihood. If h is chosen to maximize

$$\prod_{i=1}^{n} f_n(x_i;\, D, h),$$

then it is easily seen that the unhelpful value of zero will be returned. Habbema *et al.* (1974) and Duin (1976) therefore chose to maximize

$$\prod_{i=1}^{n} f_{n-1}(x_i;\, D_i, h), \tag{1}$$

where D_i denotes the data set D with observation x_i omitted. This procedure, which leads to a reasonable degree of smoothing, was observed by Titterington (1980) to be a cross-validation in the sense of Stone (1974a). Bowman (1980) demonstrated this in detail for the case of discrete data, showing that maximizing (1) is equivalent to minimizing

$$\frac{1}{n} \sum_{i=1}^{n} I\{\delta_{x_i}, f_{n-1}(.;\, D_i, h)\}, \tag{2}$$

where I is the discrete Kullback–Leibler loss function and δ_{x_i} is the degenerate distribution with probability concentrated at x_i. However, a direct transfer of this formulation to the continuous case is inadequate since the use of Kullback–Leibler loss

$$I(p, q) = \int p(x) \log\{p(x)/q(x)\}\, dx,$$

with δ_{x_i} now a Dirac delta function, makes (2) infinite. A simple solution is not to assess the performance of the density estimate directly, as in (2), but to compare its performance with that of the true density. The relevant loss function is then

$$H(p, q, r) = I(p, q) - I(p, r),$$

which, for Kullback–Leibler loss, reduces to

$$\int p(x)\log\{r(x)/q(x)\}\,dx.$$

Cross-validation now requires the minimization of

$$\frac{1}{n}\sum_{i=1}^{n} H\{\delta_{x_i}, f_{n-1}(.\,;\, D_i, h), f(\,.\,)\}, \tag{3}$$

which, in the Kullback–Leibler case, is

$$\frac{1}{n}\sum_{i=1}^{n} \log\{f(x_i)/f_{n-1}(x_i;\, D_i, h)\}. \tag{4}$$

A log transformation of (1) shows that (1) and (4) are equivalent criteria for choice of h. This method of choice will be referred to as Kullback–Leibler cross-validation.

The theoretical properties of this procedure have only recently begun to come to light. Aitchison & Aitken (1976) developed a kernel estimator for use with categorical data and Bowman (1980) established consistency when Kullback–Leibler cross-validation is applied. However, Schuster & Gregory (1981) showed that with continuous data the technique will produce inconsistent estimators if the true probability density function has a sufficiently long tail. An intuitive explanation of this is that nonnegligible weight must be given to each observation by the density estimate constructed from the remaining $(n-1)$ data points. If the underlying distribution has a long tail or if outliers are present, this technique will tend to produce smoothing parameters which are too large, as demonstrated by the sensitivity curve of Scott & Factor (1981). A similar curve is shown in Fig. 1. The graph monitors the value of h chosen by Kullback–Leibler cross-validation on a data set consisting of 24 approximate normal order statistics

$$\left\{\Phi^{-1}\left(\frac{i-\frac{1}{2}}{24}\right);\ i = 1, \ldots, 24\right\}$$

and a 25th, x, against which h is plotted. As x moves away from the rest of the data the value of h is forced to increase.

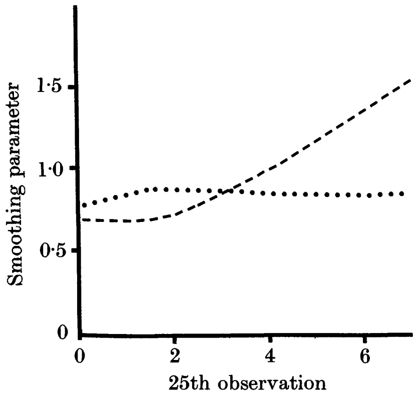

Fig. 1. Sensitivity curves, based on 24 approximate normal order statistics, for Kullback–Leibler, shown by dashed line, and squared error, dotted line, cross-validation.

In discussion of density estimates it is common to use mean integrated squared error as a measure of performance and, with respect to this loss function, Chow, Geman & Wu (1983) established the consistency of estimators based on Kullback–Leibler cross-validation when the underlying density has bounded support. Hall (1982) shows that for large samples the performance is suboptimal with respect to mean integrated squared error but it should be remembered that the method is based on an entirely different loss function, namely Kullback–Leibler.

3. Integrated squared error cross-validation

In a cross-validatory framework, it is natural to consider the use of loss functions other than Kullback–Leibler. Stone (1974b) considered absolute and squared error loss functions on multinomial data and a form of cross-validation with squared error loss wȧs applied to nonparametric regression by Wahba & Wold (1975) and later authors.

In the present problem, only a few of the many possibilities for alternative loss functions are sensible because of the nature of the Dirac delta function. In work which has recently come to the author's attention, Rudemo (1982), and R. Gratton in a Ph.D. thesis at the University of Newcastle, proposed to choose the smoothing parameter by minimizing an estimate of a quadratic loss function and provided cross-validatory arguments to support the estimates used. The following derivation produces the same criterion by applying the framework of § 2. It appeared in its earliest form in a Ph.D. thesis by the author at the University of Glasgow.

Since the squared error loss function is well behaved, an obvious candidate for the density estimation problem is integrated squared error

$$I(p, q) = \int \{p(x) - q(x)\}^2\, dx.$$

Criterion (3) then reduces to minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \int f_{n-1}^2(x;\, D_i, h)\, dx - \frac{2}{n}\sum_{i=1}^{n} f_{n-1}(x_i;\, D_i, h) + \frac{2}{n}\sum_{i=1}^{n} f(x_i) - \int f^2(x)\, dx. \qquad (5)$$

This method of choice will be referred to as squared error cross-validation. The last two terms are independent of h and may therefore be ignored in the minimization. The apparent necessity of carrying out integration to evaluate this criterion function is removed by choosing a suitable kernel, such as the normal, for which the integration may be carried out analytically. If the normal density function is denoted by

$$N(x, h^2) = (2\pi)^{-\frac{1}{2}} h^{-1} \exp(-\tfrac{1}{2}x^2/h^2),$$

then that part of expression (5) dependent on h reduces to

$$\frac{1}{(n-1)} N(0, 2h^2) + \frac{(n-2)}{n(n-1)^2} \textstyle\sum_{i \neq j} N(x_i - x_j, 2h^2) - \frac{2}{n(n-1)} \sum_{i \neq j} N(x_i - x_j, h^2).$$

This yields a simple function to be minimized. The criterion examined in detail by Rudemo (1982) is slightly different from (5) but this difference is unimportant in practice.

The aim of Rudemo (1982) was to minimize an estimated loss function and it is clear that cross-validation achieves this since, for fixed h, the expected values of (4) and (5) are

$I(f, f_{n-1})$, where I is Kullback–Leibler loss and integrated squared error respectively. This shows the cross-validatory methods to be similar to those of Wahba (1981) and Davis (1981), who choose the smoothing parameter to minimize an estimate of the mean integrated squared error between the true density and an estimate based on orthogonal series.

One means of comparing the properties of these two methods of cross-validation is to produce sensitivity curves, as displayed in Fig. 1, where it is immediately clear that the squared error loss function is not unduly sensitive to outlying observations. This augurs well for consistency and, indeed, Hall (1983) has proved the powerful result that, subject to certain conditions, smoothing parameters chosen by squared error cross-validation produce density estimators which are not only consistent but are also asymptotically optimal in the sense that mean integrated squared error is asymptotically minimized. For a wide range of distributions, then, squared error cross-validation will, for large samples, produce density estimators which perform as well as those based on the theoretical optimal smoothing parameters. The main condition imposed in the proof of this result refers to the tails of the distribution and requires that $E(X^2 |\log|X||^{2+\varepsilon})$ is finite for some $\varepsilon > 0$. This is only slightly stronger than requiring the existence of a finite variance.

4. A simulation study

The small sample performances of the two procedures may be easily investigated by simulating data from a few well-chosen distributions. This may also indicate how quickly the asymptotic results begin to apply. Rudemo (1982) examines the performance of his criterion for mixtures of log normal distributions but the results of a more extensive study are given in Table 1. Further, optimality is monitored by giving the results in terms of ratios with respect to estimated optimal values of mean integrated squared error. These estimated values are simply the average integrated squared errors incurred when the asymptotically optimal smoothing parameters are employed. Evaluation in terms of Kullback–Leibler loss is discussed below. The chosen distributions are the standard normal, a bimodal mixture of normals, a Student's t with 5 degrees of freedom and the standard Cauchy. The first two test for sensitivity to changes of shape in the main body of the distribution and allow direct comparison with the results of Scott & Factor (1981) who showed that Kullback–Leibler cross-validation compares very

Table 1. *Integrated squared errors of density estimates from simulated data as ratios to estimated optimal values. Averages over simulations, with standard errors in brackets*

		$n = 25$	$n = 50$	$n = 100$
		$n_{\text{sim}} = 100$	$n_{\text{sim}} = 50$	$n_{\text{sim}} = 25$
Standard	KL	1·52 (0·29)	1·32 (0·14)	1·21 (0·07)
normal	ISE	1·96 (0·40)	1·73 (0·23)	1·42 (0·09)
Normal mixture	KL	1·25 (0·17)	1·33 (0·15)	1·14 (0·06)
$\frac{1}{2}N(-1{\cdot}5, 1)+\frac{1}{2}N(1{\cdot}5, 1)$	ISE	1·26 (0·17)	1·31 (0·12)	1·14 (0·06)
Student's $t_{(5)}$	KL	1·71 (0·29)	1·38 (0·16)	1·91 (0·12)
	ISE	1·89 (0·39)	1·47 (0·13)	1·49 (0·12)
Standard	KL	3·79 (0·45)	6·59 (0·47)	14·15 (0·67)
Cauchy	ISE	1·30 (0·20)	1·49 (0·16)	1·28 (0·07)

KL and ISE refer to Kullback–Leibler and squared error cross-validation respectively; n, sample size; n_{sim}, simulation size.

favourably with two other data-based smoothing procedures. The second two distributions serve to indicate the relative performances with respect to long-tailed distributions. Note that Kullback–Leibler cross-validation is known to be inconsistent for both of these distributions while squared error cross-validation is known to be consistent for Student's $t_{(5)}$. Hall's (1983) theory does not cover the Cauchy distribution. Normal kernels were used in all cases.

For the normal distribution and normal mixture there is little difference between the two methods, although Kullback–Leibler cross-validation is consistently superior to the squared error version by a small margin in the standard normal case. This is intriguing since the latter method should be favoured by the use of the integrated squared error loss function to measure performance. It is for the long-tailed distributions, however, that major differences become apparent. In the case of the Cauchy, the superiority of the squared error version is clear for sample sizes as small as 25 whereas for the $t_{(5)}$ distribution greater sample sizes are necessary before a marked difference can be detected.

Integrated squared error is a convenient measure of performance and is appropriate for a general comparison of the two smoothing methods since it treats equally seriously discrepancies between the true and estimated density functions at different points in the sample space. If Kullback–Leibler cross-validation is favoured by measuring performance in terms of Kullback–Leibler loss, there is again little difference between the methods for the normal based distributions and superiority now emerges over squared error cross-validation for the long-tailed distributions. However, for the latter distributions the Kullback–Leibler errors are very large and the consistency results described above should be borne in mind.

5. EXAMPLES AND DISCUSSION

The simulation study clearly indicates that cross-validatory choice with integrated squared error loss function provides a very useful alternative to the Kullback–Leibler version, achieving much more satisfactory results for long-tailed distributions, and with little or no deterioration for distributions with normal tails. To highlight this, both methods were applied to two data sets published in the literature: the chrondite meteor data of Ahrens (1965) and a set of data simulated from the standard Cauchy distribution by Schuster & Gregory (1981). Rudemo (1982) gives an example of smoothing data from a point process.

The chrondite meteor data has only 22 observations, a sample size which must be regarded as very small for the purposes of estimating the density function. However, this data set has been used on several occasions to compare techniques of density estimation and Scott, in the discussion of Good & Gaskins (1980), illustrates the performance of Kullback–Leibler cross-validation among other methods. With normal kernels this produces a smoothing parameter of 0·95. Squared error cross-validation produces a smoothing parameter of 0·71 and there is little qualitative difference in the estimates corresponding to these two values, as Fig. 2(a) shows.

The simulation of 100 observations from a Cauchy distribution by Schuster & Gregory (1981) was carried out to illustrate the problem which arises when Kullback–Leibler cross-validation is applied to data from a long-tailed distribution. For this set of data the smoothing parameter appropriate to normal kernels is 1·91, which produces a density estimate which is far too flat. Schuster & Gregory (1981) propose to overcome this by

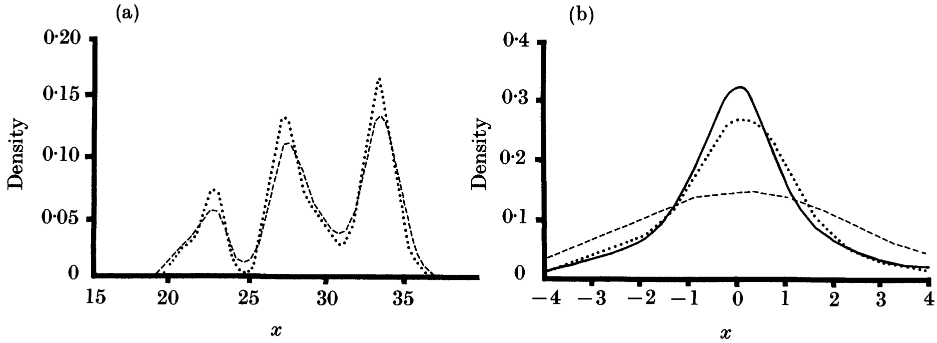

Fig. 2. Density estimates for: (a) the chrondite meteor data; (b) the simulated standard Cauchy data. Dashed line, Kullback–Leibler cross-validation; dotted line, squared error cross-validation; solid line, true density; x denotes the measurement scale of observed data in each case.

employing variable kernels, but application of squared error cross-validation yields a smoothing parameter of 0·58 and an estimate which is close to the true density, as illustrated in Fig. 2b.

Nonparametric discrimination is one of the chief areas of application of density estimation and here multivariate data are often involved. Both methods of cross-validation extend without conceptual difficulty to this case. Habbema, Hermans & Remme (1978) describe the application of Kullback–Leibler cross-validation to very general types of multivariate data. There is also current interest in the use of variable kernel estimators and cross-validation again covers this generalization without difficulty. However, simulations carried out by the author suggest that Kullback–Leibler cross-validation may not be an entirely satisfactory means of choosing the order of nearest neighbour distances which these estimators require.

Another modification to which both methods of cross-validation are appropriate is the large sample version advocated by Wahba & Wold (1975) and discussed by Gregory & Schuster (1979) in the context of kernel density estimation.

The original form of this work was carried out as part of a Ph.D. thesis at the University of Glasgow and owes much to the supervision of Prof. D. M. Titterington and the financial support of the Science and Engineering Research Council. I am also indebted to Dr P. Hall for his correspondence and helpful comments on an earlier form of this paper, and to the constructive remarks of a referee.

References

Ahrens, L. A. (1965). Observations on the Fe-Si-Mg relationship in chrondites. *Cosmochem. Geochim. Acta* **29**, 801–6.

Aitchison, J. & Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413–20.

Bowman, A. W. (1980). A note on consistency of the kernel method for the analysis of categorical data. *Biometrika* **67**, 682–4.

Chow, Y.-S., Geman, S. & Wu, L.-D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* **11**, 25–38.

Davis, K. B. (1981). Estimation of the scaling parameter for a kernel-type density estimate. *J. Am. Statist. Assoc.* **76**, 632–6.

Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *I.E.E.E. Trans. Comput. C* **25**, 1175–9.

Fryer, M. J. (1976). Some errors associated with the non-parametric estimation of density functions. *J. Inst. Math. Applic.* **18**, 371–80.

FRYER, M. J. (1977). A review of some non-parametric methods of density estimation. *J. Inst. Math. Applic.* **20**, 335–54.

GOOD, I. J. & GASKINS, R. A. (1980). Density estimation and bump-hunting by the penalised likelihood method exemplified by scattering and meteorite data (with discussion). *J. Am. Statist. Assoc.* **75**, 42–56.

GREGORY, G. G. & SCHUSTER, E. F. (1979). Contributions to non-parametric maximum likelihood methods of density estimation. In *12th Annual Symposium on the Interface of Computer Science and Statistics*, Ed. J. F. Gentleman, pp. 427–31. University of Waterloo, Ontario.

HABBEMA, J. D. F., HERMANS, J. & REMME, J. (1978). Variable kernel density estimation in discriminant analysis. In *Compstat* 1978. Ed. L. C. A. Corsten and J. Hermans, pp. 178–85. Vienna: Physica Verlag.

HABBEMA, J. D. F., HERMANS, J. & VAN DEN BROEK, K. (1974). A stepwise discriminant analysis program using density estimation. In *Compstat* 1974, Ed. G. Bruckmann, pp. 101–10. Vienna: Physica Verlag.

HALL, P. (1982). Cross-validation in density estimation. *Biometrika* **69**, 383–90.

HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11**, 1156–74.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65–78.

SCHUSTER, E. F. & GREGORY, G. G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In *13th Annual Symposium on the Interface of Computer Science and Statistics*, Ed. W. F. Eddy, pp. 295–8. New York: Springer-Verlag.

SCOTT, D. W. (1980). Discussion of paper by I. J. Good and R. A. Gaskins. *J. Am. Statist. Assoc.* **75**, 61–2.

SCOTT, D. W. & FACTOR, L. E. (1981). Monte Carlo study of three data-based nonparametric probability density estimators. *J. Am. Statist. Assoc.* **76**, 9–15.

SCOTT, D. W., TAPIA, R. A. & THOMPSON, J. R. (1977). Kernel density estimation revisited. *Nonlinear Anal., Theory, Methods Applic.* **1**, 339–72.

STONE, M. (1974a). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. R. Statist. Soc.* B **36**, 111–147.

STONE, M. (1974b). Cross-validation and multinomial prediction. *Biometrika* **61**, 509–15.

TITTERINGTON, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics* **22**, 259–68.

WAHBA, G. (1981). Data-based optimal smoothing of orthogonal series density estimates. *Ann. Statist.* **9**, 146–56.

WAHBA, G. & WOLD, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation. *Comm. Statist.* **4**, 1–18.

[*Received February* 1983. *Revised October* 1983]