*J. R. Statist. Soc.* B (1990)
**52**, *No.* 1, *pp.* 223-232

# Bandwidth Choice for Density Derivatives

By WOLFGANG HÄRDLE, J. S. MARRON† and M. P. WAND

*Universität Bonn, FRG* *University of North Carolina at Chapel Hill, USA* *Australian National University, Canberra, Australia*

[Received February 1988. Final revision June 1989]

SUMMARY

An adaptation of least squares cross-validation is proposed for bandwidth choice in the kernel estimation of the derivatives of a probability density. The practicality of the method is demonstrated by an example and a simulation study. Theoretical justification is provided by an asymptotic optimality result.

*Keywords*: CROSS-VALIDATION DENSITY DERIVATIVES; KERNEL ESTIMATORS

## 1. INTRODUCTION

The kernel approach provides an attractive method for estimation of both probability densities and their derivatives. Such estimators have been successfully used in the exploration and presentation of data; see, for example, Silverman (1986). Density derivatives are of particular interest for the evaluation of modes and inflection points. They are also of theoretical importance, as they occur both directly and indirectly in asymptotic expansions of error criteria for density estimation. In addition, density derivatives are of practical importance for estimating scores in certain additive models; see Härdle and Stoker (1990). Another application is to the empirical verification of uniqueness of equilibria of market demand, where the estimation of derivatives of densities enters through so-called income effects; see Hildenbrand and Hildenbrand (1986).

As with any type of smoothing method, the performance of kernel estimators is heavily dependent on the choice of smoothing parameter. If the effective amount of local averaging is too small, the resulting curve estimate is subject to too much sample variability, which appears in the form of a curve which is too wiggly. In contrast, too much local averaging results in the introduction of an unacceptably large bias, in the sense that features of the true curve will be smoothed away.

A practical approach to the problem of smoothing parameter selection for density estimation is provided by least squares cross-validation, which was proposed by Rudemo (1982) and Bowman (1984). Strong theoretical justification has been provided by several asymptotic optimality results which demonstrate that the selected smoothing parameter is, in the limit, effectively the same as the squared error optimal choice; see Hall (1983), Stone (1984) and Burman (1985).

In this paper the cross-validation idea is extended to the estimation of density derivatives. The extension is motivated and made precise in Section 2. The practical

†*Address for correspondence*: Department of Statistics, University of North Carolina at Chapel Hill, CB 3260, Phillips Hall, Chapel Hill, NC 27599-3260, USA.

0035-9246/90/52223 $2.00

effectiveness of this method is demonstrated through an example and a simulation study in Section 3.

Section 4 provides theoretical underpinning for the cross-validation method. In particular, two types of asymptotic optimality results are established. It is shown that the cross-validated smoothing parameter is asymptotically the same as the squared error optimal choice, and also that the squared error performance is effectively the same. Section 5 contains the proofs of the theoretical results.

## 2. CROSS-VALIDATION FOR DENSITY DERIVATIVES

We consider here the estimation of the $k$th derivative $f^{(k)}(x)$ of a probability density $f(x)$ from a random sample $X_1, \ldots, X_n$. A kernel estimator of $f^{(k)}(x)$, motivated by taking the $k$th derivative of the kernel estimate of $f$, is given by

$$\hat{f}_h^{(k)}(x) = n^{-1} \sum_{i=1}^{n} h^{-k-1} K^{(k)}\{(x-X_i)/h\},$$

where $h$ is called the bandwidth or smoothing parameter and $K$ is the kernel function which is assumed to be a symmetric probability density. Gasser *et al.* (1985) have developed an interesting asymptotic theory for the optimal choice of $K$ which shows that the best choice of the function $K^{(k)}$ is not necessarily the $k$th derivative of the optimal kernel for estimating $f$. However, the present form is used here because we prefer its intuitive content and are concerned about numerical instabilities (see Section 3 for more details on this, as well as a strong reason for not using the normal kernel, especially for large $k$). As already noted, the choice of the amount of smoothing, quantified here by $h$, is crucial to the performance of $\hat{f}_h^{(k)}(x)$.

The essential idea of least squares cross-validation, for the estimation of $f$ (the special case $k = 0$ here), is to use the bandwidth which minimizes the function

$$\mathrm{CV}(h) = \int \hat{f}_h(x)^2 \, \mathrm{d}x - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{h,i}(X_i),$$

where $\hat{f}_{h,i}$ denotes the leave-one-out kernel estimator (defined for general $k$ later). This method of bandwidth selection can be motivated by observing that the function $\mathrm{CV}(h)$ provides a reasonable, and indeed unbiased, estimate of the first two terms in the expansion of the integrated square error,

$$\begin{aligned} d_{\mathrm{I}}(\hat{f}_h, f) &= \int \{\hat{f}_h(x) - f(x)\}^2 \, \mathrm{d}x \\ &= \int \hat{f}_h^2 - 2\int \hat{f}_h f + \int f^2. \end{aligned}$$

So, since the third term is independent of $h$, the minimizer of $\mathrm{CV}(h)$ may be expected to be reasonably close to the minimizer of $d_{\mathrm{I}}$.

This idea can be extended to the estimation of derivatives of the density by observing that

$$\begin{aligned} d_{\mathrm{I}}(\hat{f}_h^{(k)}, f^{(k)}) &= \int \{\hat{f}_h^{(k)}(x) - f^{(k)}(x)\}^2 \, \mathrm{d}x \\ &= \int \hat{f}_h^{(k)2} - 2\int \hat{f}_h^{(k)} f^{(k)} + \int f^{(k)2}. \end{aligned}$$

As before, the last term is independent of $h$, so it has no effect on the location of the minimizer, and the first term is available to the experimenter. By integration by parts, the second term may be estimated by the second term of the cross-validation function,

$$\mathrm{CV}_k(h) = \int \hat{f}_h^{(k)}(x)^2 \, \mathrm{d}x - 2n^{-1}(-1)^k \sum_{i=1}^{n} \hat{f}_{h,i}^{(2k)}(X_i),$$

where

$$\hat{f}_{h,i}^{(2k)}(x) = (n-1)^{-1} \sum_{j \neq i} h^{-2k-1} K^{(2k)}\{(x-X_j)/h\}.$$

Hence the bandwidth $\hat{h}_k$, which minimizes $\mathrm{CV}_k$, should be close to $h_k^*$, the minimizer of $d_\mathrm{I}(\hat{f}_h^{(k)}, f^{(k)})$.

$\mathrm{CV}_k(h)$ can be simplified to give the computationally more straightforward version

$$\mathrm{CV}_k(h) = (-1)^k n^{-1} h^{-2k-1} \biggl[ n^{-1} \sum_i \sum_j (K*K)^{(2k)}\{(X_i - X_j)/h\} - 2(n-1)^{-1} \sum_{i \neq j}\sum K^{(2k)}\{(X_i - X_j)/h\} \biggr],$$

where here and throughout an asterisk denotes convolution. This can either be used directly, or easily adapted to give an efficient fast Fourier transform approximation, as described in Section 3.5 of Silverman (1986). For some kernels, the fact that $(K*K)^{(2k)} = K^{(k)}*K^{(k)}$ can also be useful.

## 3. EXAMPLE AND SIMULATIONS

We tested the bandwidth selection method described in Section 2 on several data sets for estimation of $f^{(1)}(x)$, the first derivative of $f(x)$. We had the best success when the kernel was standard normal. Piecewise polynomial kernels, with asymptotic optimality properties of the type described in Gasser *et al.* (1985), sometimes gave a numerically unstable derivative cross-validation function, especially for the smaller data sets. This seems to be caused by the fact that $\mathrm{CV}_k$ makes use of the $2k$th derivative of $K$, which for $k = 1$ is discontinuous for some popular kernels. One approach to this problem would be to use piecewise polynomials that have an optimality property under smoothness constraints, as developed in Müller (1984), although we have not tried this.

The normal kernel is attractive also for larger data sets, as the function $\mathrm{CV}_k(h)$ may be efficiently calculated by a fast Fourier transform algorithm as mentioned at the end of Section 2. To see how much loss in efficiency could be expected from using the normal kernel, we calculated an analog of Table 3.1 of Silverman (1986). The analog of Silverman's $C(K)$ (although see Marron and Nolan (1989) for a more convincing derivation) for estimating the $k$th derivative is

$$C_k(K) = \{\textstyle\int (K^{(k)})^2\}^{4/(5+2k)} (\textstyle\int x^{2+k} K^{(k)})^{(4k+2)/(5+2k)}.$$

Table 1 shows the efficiency (in the sense of Silverman) of the normal kernel, with respect to some of the optimal kernels of Müller (1984) (indexed by the amount of 'smoothness' $\mu$ in Müller's notation), defined as

TABLE 1
*Efficiencies of the normal kernel, with respect to Müller's optimal kernels*

| $k$ | $\mu$ | $eff(K_{k,\mu}, \phi^{(k)})$ |
|---|---|---|
| 0 | 2 | $\frac{10}{7}\left(\frac{\pi}{7}\right)^{1/2} \approx 0.9570$ |
| 0 | 3 | $\frac{700}{1287}\pi^{1/2} \approx 0.9640$ |
| 1 | 2 | $\frac{140}{297}\pi^{1/2} \approx 0.8355$ |
| 1 | 3 | $\frac{2520}{1573}\left(\frac{\pi}{11}\right)^{1/2} \approx 0.8562$ |
| 2 | 2 | $\frac{22680}{17303}\left(\frac{\pi}{11}\right)^{1/2} \approx 0.7005$ |
| 2 | 3 | $\frac{55440}{37349}\left(\frac{\pi}{13}\right)^{1/2} \approx 0.7297$ |

$$\mathrm{eff}(K_{k,\mu}, \phi^{(k)}) = \{C_k(K_{k,\mu})/C_k(\phi^{(k)})\}^{(5+2k)/4}.$$

In view of the well-known fact that there is very little loss in efficiency when $k = 0$, we were rather surprised to see a fairly substantial loss in efficiency for the other cases. The loss is not too serious for $k = 1$, but is worse with increasing $k$. It appears that $k$ need not be too large before this loss in efficiency will outweigh the numerical and intuitive advantages of the Gaussian kernel.

Derivative cross-validation usually, but not always, gave a larger bandwidth than the ordinary cross-validation. This was expected from the asymptotic rate of convergence results, which say that a larger bandwidth is required to estimate higher derivatives. In particular, in the simplest setting of $K$ non-negative and $f$ sufficiently smooth, reasonable bandwidths are of the order $n^{-1/(2k+5)}$, which increases in $k$; see, for example, Stone (1980).

An interesting case was an application to a data set on food expenditures in 1973 from the *Family Expenditure Survey, Annual Base Tapes (1968–1983)* (Department of Employment, 1984). The data utilized in this paper were made available by the Economic and Social Research Council's data archive at the University of Essex. Because of the large size of this data set, we worked with a condensed version, where each observation consists of the average of groups of 50 order statistics.

The estimation of the derivative of the probability density is useful for several reasons in this context. One is that it is a major component of the Engel curve, which is vital for empirical verification of the law of demand. Another is that it figures heavily in the estimation of elasticities. See Hildenbrand and Hildenbrand (1988) for definition, motivation and analysis of these quantities together with the economic conclusions which have been drawn. One more related application is that it represents the most difficult to estimate component of the average derivative functional, which, together with the Engel curve, is also important for empirical verification of the law of demand; see Härdle and Stoker (1990).

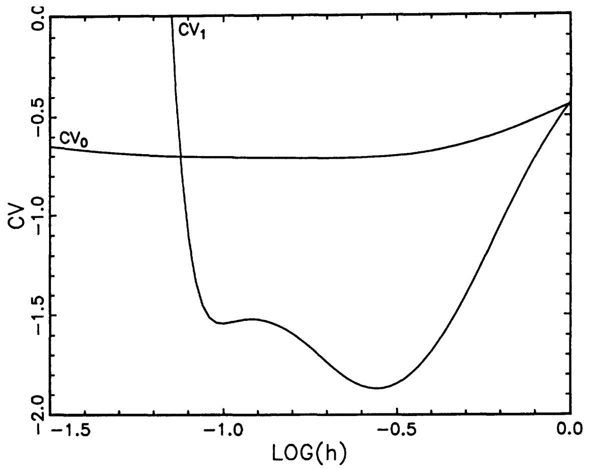

Fig. 1. Cross-validation functions $CV_1$ and $CV_0$ for food expenditure data, $K$ standard normal

Fig. 1 shows a superimposition of the cross-validation functions $CV_0(h)$ and $CV_1(h)$. A feature that was typical of all the examples that we have considered is that the minimum is much better defined for the derivative. This appears related to the fact that bandwidths chosen by cross-validation have better stability properties in settings where curve estimation is more difficult, such as higher dimensional estimation. See Section 4 of Marron (1986) for a discussion of this seeming paradox.

An interesting feature of this data set, that we did not observe for any other, is that if we extend the range of $h$s, for which minimization is performed, to include some very small values then the function $CV_1(h)$ has its global minimum at an unreasonably small value. This is not a practical problem for this data set, because the bandwidths in the extended range represent amounts of smoothing which give a far too wiggly curve to be seriously considered. However, it is worth noting because similar phenomena have been observed for ordinary density cross-validation; see Rudemo (1982) and Scott and Terrell (1987).

The bandwidth selection rule was also applied to some simulated data. 15 samples of size 750 of data having the extreme value density

$$f(x) = e^x e^{-e^x}$$

were generated to assess the performance of $\hat{h}_1$ in the estimation of

$$f^{(1)}(x) = e^x e^{-e^x}(1 - e^x).$$

The selected bandwidths are listed in Table 2. With only 15 data sets, the results are far from conclusive, but they do give some insight.

Most of them are in reasonably close agreement with the bandwidth which minimizes the mean integrated square error, which was roughly 0.34 in this case. To give an idea about the performance of the resulting curve estimates, we chose the sample which gave the median value of $d_I(\hat{f}_{\hat{h}_1}^{(1)}, f^{(1)})$ among our 15 replications. Fig. 2 shows the resulting curve estimate $\hat{f}_{\hat{h}_1}^{(1)}$ as a broken line and the true underlying curve $f^{(1)}$ as the full curve.

TABLE 2
*Values of $\hat{h}_I$ for 15 samples from the extreme value density with $n = 750$ using the standard normal kernel*

| | | | | |
|---|---|---|---|---|
| 0.16 | 0.44 | 0.20 | 0.45 | 0.33 |
| 0.44 | 0.30 | 0.46 | 0.43 | 0.28 |
| 0.18 | 0.34 | 0.34 | 0.47 | 0.16 |

## 4. THEORETICAL RESULTS

For ordinary density estimation, i.e. for $k = 0$, the effective asymptotic performance of the cross-validated bandwidth has been established by the optimality results of Hall (1983), Stone (1984) and Burman (1985). In this section, it is seen how these results may be extended to general $k$.

Assume that $K$ is a compactly supported probability density with $2k$ bounded derivatives and that $f$ has $2k+2$ continuous bounded derivatives. The assumption of compact support of $K$ does not include the Gaussian kernel used in Section 3. The results proven here can be extended to this case by a straightforward truncation argument. This is not explicitly done because the increased technical complexity of the proof only detracts from the main points.

The bandwidths under consideration are assumed to come, for each $n$, from a set $H_n$ so that $\sup_{h \in H_n} h \leqslant n^{-\delta}$, $\inf_{h \in H_n} h \geqslant n^{(-1+\delta)/(k+1)}$ and $\operatorname{card}(H_n) \leqslant n^{\rho}$ for some constants $\delta > 0$ and $\rho > 0$.

The cross-validated bandwidth $\hat{h}_k$ is asymptotically the same as the optimal bandwidth $h_k^*$ (both chosen as minimizers over the set $H_n$) in the following sense.

*Theorem 1.* Under the above assumptions, as $n \to \infty$,

$$\hat{h}_k / h_k^* \to 1, \qquad \text{almost surely.}$$

The fact that this result means that the cross-validated bandwidth is useful for estimation is demonstrated by the following theorem.

*Theorem 2.* Under the above assumptions, as $n \to \infty$,

$$d_{\mathrm{I}}(\hat{f}^{(k)}_{\hat{h}_k}, f^{(k)}) / d_{\mathrm{I}}(\hat{f}^{(k)}_{h_k^*}, f^{(k)}) \to 1, \qquad \text{almost surely.}$$

*Remark 1.* Since $d_{\mathrm{I}}$ is random, i.e. changes for different data sets, one may prefer as an error criterion its expected value, the mean integrated square error,

$$d_{\mathrm{M}}(\hat{f}^{(k)}_h, f^{(k)}) = E d_{\mathrm{I}}(\hat{f}^{(k)}_h, f^{(k)}).$$

The proof of theorems 1 and 2 may be adapted in a straightforward fashion to give the $d_{\mathrm{M}}$ analogues of those results:

$$\hat{h}_k / h_k^{\cdot} \to 1, \qquad \text{almost surely,}$$

$$d_{\mathrm{M}}(\hat{f}^{(k)}_{\hat{h}_k}, f^{(k)}) / d_{\mathrm{M}}(\hat{f}^{(k)}_{h_k^{\cdot}}, f^{(k)}) \to 1, \qquad \text{almost surely,}$$

where $h_k^{\cdot}$ denotes the minimizer of $d_{\mathrm{M}}(\hat{f}^{(k)}_h, f^{(k)})$. However, no statement is made here about $E d_{\mathrm{I}}(\hat{f}^{(k)}_{\hat{h}_k}, f^{(k)})$.

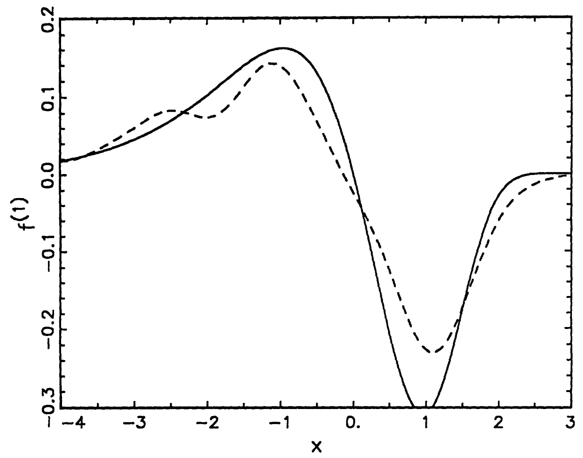

Fig. 2. Simulation target curve (———) and curve estimate (---) for the data set giving median performance, from the extreme value distribution, using the standard normal kernel

*Remark 2.* The assumption that $f$ has $2k+2$ derivatives is substantially stronger than that made, for $k=0$, by most of the researchers cited at the beginning of this section. With more technical effort (and messy notation) than seems justified to us, this assumption can be weakened somewhat, but observe that the assumption of at least $2k$ derivatives appears to be essential for the establishment of asymptotic optimality for this method of cross-validation.

## 5. PROOFS

Theorems 1 and 2 are a consequence of the following two lemmas.

*Lemma 1.*

$$\sup_{h \in H_n} |\{d_I(\hat{f}_h^{(k)}, f^{(k)}) - A(h)\}/A(h)| \to 0, \qquad \text{almost surely,}$$

where

$$A(h) = \int (K^{(k)})^2 n^{-1} h^{-(2k+1)} + (\int u^2 K/2)^2 \int (f^{(k+2)})^2 h^4.$$

*Lemma 2.*

$$\sup_{h, h' \in H_n} |B(h, h')| \to 0, \qquad \text{almost surely,}$$

where

$$B(h, h') = [\mathrm{CV}(h) - d_I(\hat{f}_h^{(k)}, f^{(k)}) - \{\mathrm{CV}(h') - d_I(\hat{f}_{h'}^{(k)}, f^{(k)})\}]/\{A(h) + A(h')\}.$$

Calculations of the type leading to equation (3.20) of Silverman (1986), for example, show that $A(h)$ asymptotically approximates $d_M(\hat{f}_h^{(k)}, f^{(k)})$ in the sense that

$$\sup_{h \in H_n} |\{d_M(\hat{f}_h^{(k)}, f^{(k)}) - A(h)\}/A(h)| \to 0.$$

This can be used to verify the claims made in remark 1.

The proof of lemma 1 follows very closely the proof of theorem 1 of Marron and

Härdle (1986). To see how to adapt the current set-up to the notation of that paper, define

$$g(x) := h^k f^{(k)}(x),$$

$$\hat{g}(x) := h^k \hat{f}_h^{(k)}(x),$$

$$\lambda := h^{-1},$$

$$\delta_\lambda(x, y) := h^{-1} K^{(k)}(x-y)/h,$$

$$w(x)\ \mathrm{d}F(x) := \mathrm{d}x.$$

The results of Marron and Härdle (1986) cannot be directly applied here because the target function $g(x)$ in that paper is not allowed to depend on $h$. However, an inspection of the proof in that paper shows that the result still holds, even in the current slightly more general context. This completes the proof of lemma 1.

Lemma 2 is a consequence of the following lemma.

*Lemma 3.*

$$\sup_{h\in H_n} \left| \left\{ n^{-1} \sum_{i=1}^{n} \hat{f}_{h,i}^{(2k)}(X_i) - \int \hat{f}^{(2k)} f - R \right\} \Big/ A(h) \right| \to 0, \qquad \text{almost surely,}$$

where

$$R = n^{-1} \sum_{i=1}^{n} f^{(2k)}(X_i) - \int f^{(2k)} f.$$

To prove lemma 3, define

$$U_{i,j} := h^{-2k-1} K^{(2k)}\{(X_i - X_j)/h\} - h^{-2k-1} \int K^{(2k)}\{(x - X_j)/h\}\, f(x)\ \mathrm{d}x$$
$$- f^{(2k)}(X_i) + \int f^{(2k)}(x)\, f(x)\ \mathrm{d}x.$$

$$V_i := E(U_{i,j} \,|\, X_i),$$

$$W_{i,j} := U_{i,j} - V_i.$$

To finish the proof it is sufficient to show that

$$\sup_{h\in H_n} \left| n^{-1} \sum_{i=1}^{n} V_i \Big/ A(h) \right| \to 0, \qquad \text{almost surely,} \tag{1}$$

and that

$$\sup_{h\in H_n} \left| n^{-2} \sum_{i\neq j}\sum W_{i,j} \Big/ A(h) \right| \to 0, \qquad \text{almost surely.} \tag{2}$$

To verify expression (1), by the Borel-Cantelli lemma it is sufficient to show that for $\epsilon > 0$

$$\sum_{n=1}^{\infty} \mathrm{card}(H_n)\ \sup_{h\in H_n} P\left\{ \left| n^{-1} \sum_{i=1}^{n} V_i \right| > \epsilon A(h) \right\} < \infty.$$

Hence by the Chebyshev inequality, it is sufficient to show that there is a constant $\gamma > 0$, so that for $m = 1, 2, \ldots$ there are constants $C_m$ such that

$$\sup_{h\in H_n} E\left\{n^{-1} \sum_{i=1}^{n} V_i \Big/ A(h)\right\}^{2m} \leqslant C_m n^{-\gamma m}. \tag{3}$$

To establish inequality (3), observe that $\{\Sigma_{i=1}^{n} V_i\}$ is a martingale with respect to the sequence of sigma fields generated by $\{X_1, \ldots, X_n\}$. An application of equation (21.5) of Burkholder (1973) (which is essentially Rosenthal's inequality), with $\Phi(x) = x^{2m}$, to the finitely (from $1, \ldots, n$) indexed martingale, gives

$$E\left(\sum_{i=1}^{n} V_i\right)^{2m} \leqslant C(n^m h^{4m} + n),$$

for some constant $C$. Inequality (3) follows from this and the definition of $A(h)$. This completes the proof of inequality (3), and hence also that of expression (1).

To verify expression (2), by a development similar to that leading to inequality (3), it is sufficient to show that

$$\sup_{h\in H_n} E\left\{n^{-2} \sum_{i>j}\sum W_{i,j} \Big/ A(h)\right\}^{2m} \leqslant C_m n^{-\gamma m}. \tag{4}$$

Since

$$E(W_{i,j} | X_i) = E(W_{i,j} | X_j) = 0,$$

$\{\Sigma\Sigma_{i>j} W_{i,j}\}$ is a martingale with respect to the same sequence of sigma fields as before. Applying the same inequality to this finitely indexed martingale gives

$$E\left(\sum_{i>j}\sum W_{i,j}\right)^{2m} \leqslant C(n^{2m}h^{-(2k+1)m} + n^{m+1}h^{-(2k+1)2m}),$$

for another constant $C$. A consequence of this is inequality (4). This completes the proof of expression (2) and hence that of lemmas 2 and 3.

## ACKNOWLEDGEMENTS

The research of the first two authors was supported by the Deutsche Forschungsgemeinschaft, SFB 303; that of the second author was also partially supported by the National Science Foundation.

## REFERENCES

Bowman, A. (1984) An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**, 353–360.

Burkholder, D. L. (1973) Distribution function inequalities for martingales. *Ann. Probab.*, **1**, 19–42.

Burman, P. (1985) A data dependent approach to density estimation. *Z. Wahrsch. Ver. Geb.*, **69**, 609–628.

Department of Employment (1984) *Family Expenditure Survey, Annual Base Tapes (1968–1983)*. London: Her Majesty's Stationery Office.

Gasser, T., Müller, H.-G. and Mammitzsch, V. (1985) Kernels for nonparametric curve estimation. *J. R. Statist. Soc.* B, **47**, 238–252.

Hall, P. (1983) Large sample optimality of least square cross-validation in density estimation. *Ann. Statist.*, **11**, 1156–1174.

Härdle, W. and Stoker, T. (1990) Investigating smooth multiple regression by the method of average derivatives. *J. Am. Statist. Ass.*, to be published.

Hildenbrand, K. and Hildenbrand, W. (1986) On the mean income effect: a data analysis of the U.K. family expenditure survey. In *Contributions to Mathematical Economics, in Honor of Gerard Debreu* (eds W. Hildenbrand and A. Mas-Colell), pp. 247–268. Amsterdam: North-Holland.

Marron, J. S. (1986) Will the art of smoothing ever become a science? *Contemp. Math.*, **9**, 169–178.

Marron, J. S. and Härdle, W. (1986) Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multiv. Anal.*, **20**, 91–113.

Marron, J. S. and Nolan, D. (1989) Canonical kernels for density estimation. *Statist. Probab. Lett.*, **7**, 195–199.

Müller, H. G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *Ann. Statist.*, **12**, 766–774.

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

Scott, D. W. and Terrell, G. R. (1987) Biased and unbiased cross-validation in density estimation. *J. Am. Statist. Ass.*, **82**, 1131–1146.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Stone, C. J. (1980) Optimal convergence rates for nonparametric estimators. *Ann. Statist.*, **8**, 1348–1360.

——— (1984) An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, **12**, 1285–1297.