THE INSTITUTE OF STATISTICS

THE CONSOLIDATED UNIVERSITY
OF NORTH CAROLINA

A COMPARISON OF CROSS-VALIDATION
TECHNIQUES IN DENSITY ESTIMATION

(Comparison in Density Estimation)

by

J.S. Marron
University of North Carolina, Chapel Hill

Mimeo Series #1568

January 1985

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

A Comparison of Cross-Validation Techniques in Density Estimation[1]

(Comparison in Density Estimation)

by

J.S. Marron
University of North Carolina, Chapel Hill

AMS 1980 subject classifications: Primary 62G05, Secondary 62G20.

Key words and phrases: cross-validation, smoothing parameter selection, choice of nonparametric estimators, bandwidth selection.

[1]Research partially supported by ONR contract N00014-81-K-0373 and by Sonderforschungsbereich 123 of the Deutsche Forschungsgemeinschaft.

Summary

In the setting of nonparametric multivariate density estimation, theorems are established which allow a comparison of the Kullback-Leibler and the Least Squares cross-validation methods of smoothing parameter selection. The family of delta sequence estimators (including kernel, orthogonal series, histogram and histospline estimators) is considered. These theorems also show that either type of cross-validation can be used to compare different estimators (eg: kernel vs. orthogonal series).

1. Introduction

Consider the problem of trying to estimate a d-dimensional probability density function, $f(x)$, using a random sample, $X_1, \cdots, X_n$, from f. Most proposed estimators of f depend on a "smoothing parameter", say $\lambda \in \mathbb{R}^+$, whose selection is crucial to the performance of the estimator.

In this paper, for the large class of delta sequence estimators, theorems are obtained which allow comparison of two smoothing parameter selectors which are known to be asymptotically optimal. An important consequence of these results is that either smoothing parameter selector may be used for a data based comparison of two density estimators, for example, kernel vs. orthogonal series. Another attractive feature of these results is that they are set in a quite general framework, special cases of which provide simpler proofs of several recent asymptotic optimality results.

In sections 2 and 3 the family of delta sequence estimators and the smoothing parameter selectors are given. The theorems are stated in section 4, with some remarks in section 5. The rest of the paper consists of proofs.

2. Delta sequence estimators

A delta sequence density estimator, as studied by Watson and Leadbetter (1965), Földes and Révész (1974), and Walter and Blum (1979), is any estimator which can be written in the form

$$\hat{f}_\lambda(x) = n^{-1} \sum_{i=1}^{n} \delta_\lambda(x, X_i) ,$$

where the function $\delta_\lambda(x,y)$ is indexed by the smoothing parameter $\lambda \in \mathbb{R}^+$. Examples include

Kernel estimators: Given a kernel function $K(x)$, let

$$\delta_\lambda(x,X_i) = \lambda\, K(\lambda^{1/d}(x-X_i)) .$$

Histogram estimators: Let B denote a bounded subset of $\mathbb{R}^d$ and suppose that $A_1, \cdots, A_\lambda$ form a partition of B. For $k = 1, \cdots, \lambda$, let $1_k(x)$ denote the indicator of A_k and let μ denote Lebesgue measure. Define, for $\lambda \in \mathbb{Z}^+$,

$$\delta_\lambda(x,X_i) = \sum_{k=1}^{\lambda} \mu(A_k)^{-1}\, 1_k(x)\, 1_k(X_i) .$$

Orthogonal Series estimators: Given a nonnegative weight function w and a sequence of functions $\{\psi_k(x)\}$ which is orthonormal and complete with respect to the inner product

$$\int \psi_k(x)\, \psi_{k'}(x)\, w(x)\, dx ,$$

define, for $\lambda \in \mathbb{Z}^+$,

$$\delta_\lambda(x,X_i) = \sum_{k=1}^{\lambda} \psi_k(x)\, \psi_k(X_i)\, w(x) .$$

See Walter and Blum (1979) for a rather extensive list of other delta sequence density estimators.

3. Smoothing parameter selectors.

The two methods of choosing the smoothing parameter λ that are discussed in this paper are the Kullback-Leibler and the Least Squares methods of cross-validation. Both make use of the leave-one-out estimators:

$$\hat{f}_{\lambda,j-}(x) = (n-1)^{-1} \sum_{i \neq j} \delta_\lambda(x,X_i) , \quad j=1, \cdots, n .$$

The Kullback-Leibler (also known as psuedo-likelihood) method first appeared in Habbema, Hermans and van den Broeck (1974) and was refined in Marron (1984). This involves choosing λ to maximize

$$KL(\lambda) = \prod_{j=1}^{n} [\hat{f}^{+}_{\lambda,j-}(X_j)^{u(X_j)} e^{-\hat{p}(\lambda)}] ,$$

where $\hat{f}^{+}_{\lambda,j-}(x)$ is the positive part of $\hat{f}_{\lambda,j-}(x)$

$$\hat{f}^{+}_{\lambda,j-}(x) = \hat{f}_{\lambda,j-}(X) \vee 0 ,$$

where $u(x)$ is a nonnegative weight function which is supported on a set where f is bounded above 0 (for example the indicator of such a set), and where

$$\hat{p}(\lambda) = \int \hat{f}_\lambda(x)\, u(x)\, dx .$$

The Least Squares method was introduced by Rudemo (1982) and Bowman (1984). This involves choosing λ to minimize

$$LS(\lambda) = \int \hat{f}_\lambda(x)^2 w(x)\, dx - 2n^{-1} \sum_{j=1}^{n} \hat{f}_{\lambda,j-}(x)\, w(X_j) ,$$

where $w(x)$ is a nonnegative weight function.

4. Theorems

In this section it will be demonstrated that choosing λ by the methods in the last section is, in a strong sense, asymptotically equivalent to minimizing the following distances:

Average Square Error:

$$d_A(\hat{f}_\lambda, f) = n^{-1} \sum_{j=1}^{n} [\hat{f}_\lambda(X_j) - f(X_j)]^2 f(X_j)^{-1} w(X_j) ,$$

Integrated Square Error:

$$d_I(\hat{f}_\lambda, f) = \int [\hat{f}_\lambda(x) - f(x)]^2 \, w(x) \, dx ,$$

Mean Integrated Square Error:

$$d_M(\hat{f}_\lambda, f) = E \int [\hat{f}_\lambda(x) - f(x)]^2 \, w(x) \, dx ,$$

where $w(x)$ is a nonnegative weight function.

Note that d_M admits the variance-bias square decomposition

$$(4.1) \qquad d_M(\hat{f}_\lambda, f) = \int n^{-1} \operatorname{var}[\delta_\lambda(x, X_i)] \, w(x) \, dx + \int B(x)^2 \, w(x) \, dx ,$$

where $B(x)$ denotes the pointwise bias,

$$(4.2) \qquad B(x) = \int \delta_\lambda(x,y) \, f(y) \, dy - f(x) .$$

Marron and Härdle (1984) have shown that, for large n, under reasonable assumptions, these distances are essentially the same in the sense that

$$(4.3) \qquad \limsup_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left| \frac{d_A(\hat{f}_\lambda, f) - d_M(\hat{f}_\lambda, f)}{d_M(\hat{f}_\lambda, f)} \right| = 0 \quad \text{a.s.},$$

$$(4.4) \qquad \limsup_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left| \frac{d_I(\hat{f}_\lambda, f) - d_M(\hat{f}_\lambda, f)}{d_M(\hat{f}_\lambda, f)} \right| = 0 \quad \text{a.s.},$$

where Λ_n is a finite set whose cardinality grows algebraically fast.

The approximations (4.3) and (4.4) are vital to the theorems of this paper. Other assumptions include the existence of constants $C, C', \delta > 0$ so that

$$(4.5) \qquad \#(\Lambda_n) \le n^C ,$$

$$(4.6) \qquad C^{-1} n^\delta \le \lambda \le C \, n^{1-\delta} , \qquad \lambda \in \Lambda_n ,$$

(4.7) $w(x) \leq C \ , \quad x \in \mathbb{R} \ ,$

(4.8) $f(x) \leq C \ , \quad x \in S \ ,$

where S denotes the support of w,

(4.9) $B(x) \leq C\, n^{-\delta} \ , \quad \lambda \in \Lambda_n, \ x \in S$

(4.10) $$\lim_{n\to\infty} \sup_{\lambda \in \Lambda_n} \left| \frac{\int \mathrm{var}[\sigma_\lambda(x, X_i)]\, w(x)dx}{C'\,\lambda} - 1 \right| = 0 \ .$$

Another useful assumption is that for k = 2, 2, ⋯ there is a constant C_k so that for m = 2, ⋯, k,

(4.11) $$\int \cdots \int \delta_\lambda(x_{i_1}, x_{j_1}) \cdots \delta_\lambda(x_{i_k}, x_{j_k}) dx_1 \cdots dx_m \leq C_k\, \lambda^{k-m/2} \ ,$$

where $i_1, j_1, \cdots, i_k, j_k = 1, \cdots, m$ subject to $i_1 \neq j_1, \cdots, i_k \neq j_k$, and to each of 1, ⋯, m appearing at least twice in the list $i_1, j_1, \cdots, i_k, j_k$. In the case of kernel estimation (4.11) is a consequence of integration by substitution. Marron and Härdle (1984) show how such a condition is satisfied for the histogram and orthogonal series estimators.

Additional assumptions needed only for the KL cross-validation function include the existence of a constant C so that

(4.12) $\delta_\lambda(x,x) \leq C\,\lambda \ , \quad x \in \mathbb{R} \ ,$

(4.13) $u(x) = w(x)\, f(x) \ ,$

(4.14) $f(x) \geq C^{-1} \ , \quad x \in S \ ,$

(4.15) $$\sup_{j,\lambda,x} \left| \hat{f}_{\lambda,j-}(x) - f(x) \right| \to 0 \quad \text{a.s.},$$

where $\sup_{j,\lambda,x}$ denotes supremum over $j=1, \cdots, n$, $\lambda \in \Lambda_n$, $x \in S$.

Before stating the theorems, it is convenient to define

$$R = n^{-1} \sum_{j=1}^{n} f(X_j)\, w(X_j) - E[f(X_j)\, w(X_j)] ,$$

$$(4.16) \qquad S = 2n^{-1} \sum_{j=1}^{n} [u(X_j)\,(1 - \log f(X_j)) - R] ,$$

$$T = - \int f(x)^2\, w(x)dx - 2R .$$

It is important to note that R, S and T are independent of λ. For this reason, the fact that both maximizing $KL(\lambda)$ and minimizing $LS(\lambda)$ are asymptotically equivalent to minimizing the distances d_A, d_I and d_M is demonstrated by (4.3), (4.4) and

Theorem 1: *Under the assumptions* (4.3), (4.5)-(4.15),

$$-2n^{-1} \log KL(\lambda) = d_A(\hat{f}_\lambda, f) + S + o(d_M(\hat{f}_\lambda, f)) ,$$

in the sense that

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left| \frac{2n^{-1} \log KL(\lambda) + d_A(\hat{f}_\lambda, f) + S}{d_M(\hat{f}_\lambda, f)} \right| = 0 \quad \text{a.s.}$$

Theorem 2: *Under the assumptions* (4.5)-(4.11)

$$LS(\lambda) = d_I(\hat{f}_\lambda, f) + T + o(d_I(\hat{f}_\lambda, f)),$$

in the sense that

$$\lim_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left| \frac{LS(\lambda) - d_I(\hat{f}_\lambda, f) - T}{d_M(\hat{f}_\lambda, f)} \right| = 0 \quad \text{a.s.}$$

Theorems 1 and 2 are stated in this nonstandard form because this provides the best comparison between KL and LS. Easy consequences of Theorems 1 and 2, respectively, are

Corollary 1: Under the assumptions (4.3)-(4.15), if $\hat{\lambda}$ is the maximizer of $KL(\lambda)$ over Λ_n, then

$$\lim_{n\to\infty} \frac{d(\hat{f}_{\hat{\lambda}},f)}{\inf_{\lambda\in\Lambda_n} d(\hat{f}_\lambda,f)} = 1 \quad \text{a.s.},$$

where d is any of d_A, d_I, or d_M.

Corollary 2: Under the assumption (4.3)-(4.11), if $\hat{\lambda}$ is the minimizer of $LS(\lambda)$ over Λ_n, then

$$\lim_{n\to\infty} \frac{d(\hat{f}_{\hat{\lambda}},f)}{\inf_{\lambda\in\Lambda_n} d(\hat{f}_\lambda,f)} = 1 \quad \text{a.s.},$$

where d is any of d_A, d_I, or d_M.

5. Remarks

5.1 The main point of this paper is comparison of the KL and LS cross-validation functions. Theorem 1 shows that $KL(\lambda)$ is based on the distance $d_A(\hat{f}_\lambda,f)$, while Theorem 2 shows that $LS(\lambda)$ is based on the somewhat more compelling distance $d_I(\hat{f}_\lambda,f)$. A more significant advantage of $LS(\lambda)$ is that the term $o(d_M(\hat{f}_\lambda,f))$ in Theorem 2 represents error from one source, while in Theorem 1 it represents error from three sources, one of which is the same as that of Theorem 2. A final disadvantage of KL is the stronger assumptions required, especially the uniform convergence assumption (4.15) and the fact that (4.14) requires $\mathcal{S}$ to be compact.

5.2 Despite the poor showing of $KL(\lambda)$ in the above respects, it should be noted that KL and LS are not really comparable because for LS, w must not depend of f, while for KL, w=u/f, with u independent of f. This is an advantage of KL because for the important applications of density estimation to discrimination and to minimum Hellinger distance estimation, the latter form is more natural.

5.3 The second important point of the results of this paper is that theoretical backing is given to the idea of Rudemo (1982) to use either KL or LS for data based comparison of density estimators. For example, to choose between a kernel and an orthogonal series estimator the one estimator with the smaller minimum KL(λ) (or LS(λ)) should have smaller $d_A(\hat{f}_\lambda, f)$ (or $d_I(\hat{f}_\lambda, f)$ respectively). Hans-Georg Müller has pointed out that caution must be used in interpreting these results with respect to the problem of kernel selection because the asymptotics of this paper do not properly describe what is happening in that setting. It is conjectured that Rudemo's idea will also work there.

5.4 As noted in the introduction, the general framework of the results of this paper contain all or part of the results of a number of recent papers as special cases. These include the results of Burman (1984), Hall (1983a,b), Marron (1984), and Stone (1984a,b). In most cases the techniques of the present paper provide a substantical simplification of the proofs in the earlier papers. Also, the unified approach makes it seem easy to provide theoretical backing to some interesting heuristics of Bowman, Hall, and Titterington (1984).

5.5 To save space, some of the assumptions of the theorems of this paper have been made more restrictive than necessary. For example, (4.5) can be weakened to Λ_n an interval by a straightforward continuity argument (see Härdle and Marron (1984) for details). The condition (4.6) can also be substantically weakened (compare Burman (1984); Stone (1984a), and Stone (1984b)). Another straight forward extension is to the case of λ vector or matrix valued as discussed by Deheuvels (1977).

6. Proof of Theorems 1 and 2

It is convenient to define, for j=1, ⋯, n

$$\Delta_j = \left[\frac{\hat{f}_{\lambda,j-}(X_j) - f(X_j)}{f(X_j)}\right] 1_S(X_j) , \quad \Delta_j^+ = \left[\frac{f^+_{\lambda,j-}(X_j) - f(X_j)}{f(X_j)}\right] 1_S(X_j) .$$

Note that by (4.14) and (4.15),

$$\sup_{j,\lambda} |\Delta_j^+| \leq \sup_{j,\lambda} |\Delta_j| \to 0 \quad \text{a.s.},$$

where the suprema are taken over j=1, ⋯, n, $\lambda \in \Lambda_n$. For n=1, 2, ⋯ define the event

$$U_n = \{\Delta_j^+ = \Delta_j \text{ for each } \lambda \in \Lambda_n \text{ and each } j=1, \cdots, n\}.$$

Note that

$$\lim_{n\to\infty} P[U_n] = 1 .$$

From the above, it follows that (on the event U_n)

$$-2n^{-1} \log KL(\lambda) - S = -2n^{-1} \sum_{j=1}^{n} [u(X_j)\,(1 + \log(1 + \Delta_j)) - \hat{p}(\lambda) - R] =$$

$$= -2n^{-1} \sum_{j=1}^{n} [u(X_j)\,(1 + \Delta_j) - \hat{p}(\lambda) - R] +$$

$$+ d_A'(\hat{f}_\lambda, f) - 2n^{-1} \sum_{j=1}^{n} r_j u(X_j) ,$$

where d'_A is the leave-one-out version of d_A given by

$$d'_A(\hat{f}_\lambda,f) = n^{-1} \sum_{j=1}^{n} [\hat{f}_{\lambda,j-}(X_j) - f(X_j)]^2 f(X_j)^{-1} w(X_j) ,$$

and where r_j denotes the remainder term of the log Taylor expansion. Theorem 1 follows easily from this, (4.3), (4.15) and the two lemmas:

Lemma 1: Under the assumptions of Theorem 1,

$$\limsup_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left| \frac{d_A(\hat{f}_\lambda,f) - d'_A(\hat{f}_\lambda,f)}{d_M(\hat{f}_\lambda,f)} \right| = 0 \quad \text{a.s.}$$

Lemma 2: Under the assumptions (4.5)-(4.11)

$$\limsup_{n\to\infty} \sup_{\lambda\in\Lambda_n} \left| \frac{n^{-1} \sum_{j=1}^{n} \hat{f}_{\lambda,j-}(X_j)\, w(X_j) - \int \hat{f}_\lambda(x)\, f(x)\, w(x)dx - R}{d_M(\hat{f}_\lambda,f)} \right| = 0 \quad \text{a.s.}$$

The proof of Lemma 1 follows in a straightforward manner from

$$\hat{f}_{\lambda,j-}(x) - \hat{f}_\lambda(x) = (n-1)^{-1} \hat{f}_\lambda(x) - (n-1)^{-1} \delta_\lambda(x,x)$$

and the assumption (4.12). The proof of Lemma 2 is in section 7.

An interesting feature of the mathematical structure here is that Lemma 2 contains the hardest part of the proof of Theorem 2 as well. To see this write

$$LS(\lambda) = d_I(\hat{f}_\lambda,f) - 2n^{-1} \sum_{j=1}^{n} \hat{f}_{\lambda,j-}(X_j)\, w(X_j) +$$

$$+ 2n^{-1} \int \hat{f}_\lambda(x)\, f(x)\, w(x)dx - \int f(x)^2\, w(x)dx .$$

Theorem 2 follows easily from this, (4.16) and Lemma 2.

7. Proof of Lemma 2

The conclusion of Lemma 2 may be written as

$$\sup_{\lambda\in\Lambda_n} n^{-1}(n-1)^{-1} \left| \sum_{i\neq j} U_{i,j} \right| d_M(\hat{f}_\lambda, f)^{-1} \to 0 \quad \text{a.s.}$$

where

$$U_{i,j} = \delta_\lambda(X_j, X_i)\, w(X_j) - \int \delta_\lambda(x, X_i)\, f(x)\, w(x)dx - f(X_j)\, w(X_j) +$$
$$+ \int f(x)^2\, w(x)dx \ .$$

For $j=1, \cdots, n$, let

$$W_j = E[U_{i,j} \mid X_j] \ ,$$

and for $i\neq j$ define

$$V_{i,j} = U_{i,j} - W_j \ .$$

Observe that

$$\text{(7.1)} \qquad \begin{aligned} &E[V_{i,j} \mid X_i] = E[V_{i,j} \mid X_j] = 0 \ , \\ &E[W_j] = 0 \ . \end{aligned}$$

To finish the proof of Lemma 2 it is enough to show that

$$\text{(7.2)} \qquad \sup_{\lambda\in\Lambda_n} n^{-2} \left| \sum_{i\neq j} V_{i,j} \right| d_M(\hat{f}_\lambda, f)^{-1} \to 0 \quad \text{a.s.}$$

and that

$$\text{(7.3)} \qquad \sup_{\lambda\in\Lambda_n} n^{-1} \left| \sum_{j=1}^{n} W_j \right| d_M(\hat{f}_\lambda, f)^{-1} \to 0 \quad \text{a.s.}$$

To verify (7.3), note that by the Borel-Cantelli Lemma, it is enough to show that for $\varepsilon > 0$,

$$\text{(7.4)} \qquad \sum_{n=1}^{\infty} \#(\Lambda_n) \sup_{\lambda\in\Lambda_n} P\left[\left|n^{-1} \sum_{j=1}^{n} W_j\right| > \varepsilon\, d_M(\hat{f}_\lambda, f)\right] < \infty \ .$$

For this, using the notation (4.2), write

$$W_j = B(X_j)\ w(X_j) - \int B(x)\ f(x)\ w(x)dx\ .$$

From the assumptions (4.7), (4.8) and (4.9) it follows that

$$|W_j| \le C\ n^{-\delta}\ ,$$

$$\sigma^2 = \text{var } W_j \le C^2 \int B(x)^2\ w(x)dx\ .$$

Now Bernstein's Inequality (see (2.13) of Hoeffding (1963)) with (in Hoeffding's notation)

$$\lambda = bt/\sigma^2,\ \tau = nt/b,\ b = C\ n^{-\delta},\ t = \varepsilon \cdot d_M(\hat{f}_\lambda, f)$$

gives

$$P[|n^{-1} \sum_{j=1}^{n} W_j| > \varepsilon\ d_M(\hat{f}_\lambda,f)] \le \exp(-nt^2/2(\sigma^2 + bt/3)) \le$$

$$\le \exp(-n\varepsilon^2 d_M(\hat{f}_\lambda,f)/2C^2) \le \exp(-n^\delta \varepsilon^2/2C^2)\ ,$$

for n sufficiently large. (7.4) is a consequence of this.

To verify (7.2), as in the proof of (7.3) above, together with the Chebyshev Inequality, it is enough to show that there is a constant $\gamma > 0$, so that for k = 1, 2, $\cdots$ there are constants C_k so that

$$\sup_{\lambda \in \Lambda_n} E[n^{-2} \sum_{i \ne j} V_{ij}\ d_M(\hat{f}_\lambda,f)^{-1}]^{2k} \le C_k\ n^{-\gamma k}\ .$$

But by the cumulant expansion of the 2k-th centered moment (see, for example, Kendall and Stuart (1963)), this may be obtained from

$$(7.5) \qquad |n^{-2k}\ d_M(\hat{f}_\lambda,\hat{f})^{-k}\ \Sigma \text{cum}_k(V_{i_1,j_1}, \cdots, V_{i_k,j_k})| \le C_k\ n^{-\gamma k}$$

where cum_k is the k-th order cumulant and Σ denotes summation over $i_1, j_1, \cdots, i_k, j_k = 1, \cdots, n$ subject to $i_1 \ne j_1, \cdots, i_k \ne j_k$.

To check (7.5), note that by (7.1) and the moment expansion of cum_k, most of the terms in the summation will be 0. In particular cum_k can be nonzero only when each of $i_1, j_1, \cdots, i_k, j_k$ is the same as one of the others. For each such term, let m denote the number of unique elements of $\{1, \cdots, n\}$ appearing among $i_1, j_1, \cdots, i_k, j_k$. By assumption (4.11) there is a constant C_k so that

$$|\mathrm{cum}_k(V_{i_1,j_1}, \cdots, V_{i_k,j_k})| \leq C_k \lambda^{k-m/2} .$$

But there is also a constant C_k so that for $m = 2, \cdots, k$, the number of nonzero terms in the summation of (7.5) with exactly m distinct indices is bounded by

$$C_k n^m .$$

Hence, by (4.1) and (4.10) there is a constant C_k so that the left side of (7.5) is bounded by

$$C_k n^{-2k}(n^{-1}\lambda)^{-k} \sum_{m=2}^{k} n^m \lambda^{k-m/2} = C_k \sum_{m=2}^{k} n^{-k+m} \lambda^{-m/2} .$$

A consequence of this is (7.5). This completes the proof of Lemma 2.

8. References

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

BOWMAN, A.W., HALL, P. and TITTERINGTON, D.M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. Biometrika **71**, 341-352.

BURMAN, P. (1984). A data dependent approach to density estimation. Manuscript.

DEHEUVELS, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. Rev. Statist. Appl. **25**, 5-42.

FÜLDES, A. and RÉVÉSZ, P. (1974). A general method for density estimation. Studia Sci. Math. Hungar. **9**, 81-82.

HABBEMA, J.D.F., HERMANS, J. and VAN DEN BROEK, K. (1974). A stepwise discrimination analysis program using density estimation. Compstat 1974: Proceedings in computational statistics (G. Bruckman, ed.) 101-110. Vienna: Physica Verlag.

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

HALL, P. (1983b). Asymptotic theory of minimum integrated square error for multivariate density estimation. Proc. Sixth Internat. Symp. Multivariate Anal. Pittsburgh, 25-29.

HÄRDLE, W. and MARRON, J.S. (1984). Optimal bandwidth selection in nonparametric regression estimation. N.C. Inst. of Statist. Mimeo Series #1546.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. **58**, 13-30.

KENDALL, M.G. and STUART, A. (1963). The Advanced Theory of Statistics, Vol. 1: Distribution Theory, Butler and Tanner Ltd.

MARRON, J.S. (1984). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. N.C. Inst. of Statist. Mimeo Series #1545.

MARRON, J.S. and HÄRDLE, W. (1984). Random approximations to some measures of accuracy in nonparametric curve estimation. N.C. Inst. of Statist. Mimeo Series #1566.

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

STONE, C.J. (1984a). An asymptotically optimal histogram selection rule. Proceedings of the Berkeley Symposium in Honor of Jerzy Neyman and Jack Kiefer **II**.

STONE, C.J. (1984b). An asymptotically optimal window selection rule for kernel density estimates. Ann. Statist. **12**, 1285-1297.

WALTER, G. and BLUM, J. (1979). Probability density estimation using delta sequences. Ann. Statist. **7**, 328-340.

WATSON, G.S. and LEADBETTER, M.R. (1965). Hazard Analysis II. Sankhyā **A26**, 101-116.