SMOOTHNESS INDEPENDENT OPTIMAL

ESTIMATION OF A NOT TOO SMOOTH DENSITY

by

James Stephen Marron
University of North Carolina at Chapel Hill

Key Words and Phrases: Kernel density estimation, pre-smoothed, cross-validation, optimal rates

AMS 1980 Subject Classifications: Primary 62G05, Secondary 62G20

ABSTRACT

In the problem of nonparametric density estimation, kernel estimators with cross-validated bandwidths are considered. An example is given to show that, even in the case where both density function and kernel have compact support, ordinary cross-validation is sub-optimal. A "pre-smoothed" modification of the cross-validated technique is proposed. By techniques similar to those of Hall (Biometrika 69 (1982) 383-390) it is shown that this density estimator achieves the well-known asymptotically optimal rate of convergence. This estimator does not make use of the precise amount of smoothness that is assumed on the density, but it is required that the density be not too smooth.

1. INTRODUCTION

Consider the problem of estimating a density function, f, using a sample $X_1,\ldots,X_n$ from f. The usual kernel estimator is defined as follows. Given a "kernel-function", K (with $\int K(x)dx = 1$), and a "bandwidth", $h \in \mathbb{R}$, let

$$\hat{f}_n(x,h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) .$$

There is a large literature concerning this estimator as may be seen from the survey by Wertz (1978). Most results make very precise assumptions about the amount of smoothness of f. Then h is chosen to depend on n in such a way that some error criterion (such as mean square error) is asymptotically minimized. It has been shown, see Farrell (1972) or Stone (1980) that this means of choosing the bandwidth gives an estimator which is asymptotically optimal, over the class of all estimators (not just kernel estimators), in the sense of rate of convergence.

Unfortunately, the particular choice of h(n) depends heavily on the precise amount of "smoothness" of f that is assumed. Thus, this means of choosing h is virtually useless to the practitioner because for an unknown f it is difficult to make accurate assumptions on the smoothness of f. For this reason there has been a considerable search for techniques which use the data to specify h.

A popular technique of this type is the cross-validated or pseudo-maximum likelihood method introduced by Habbema, Hermans, and van den Broeck (1974). To employ this method, first for j=1,...,n, form the "leave one out" kernel estimator,

$$\hat{f}_{nj}(x,h) = \frac{1}{(n-1)h} \sum_{\substack{i=1 \\ i \neq j}}^{n} K(\frac{x-X_i}{h}) .$$

Next, choose $\hat{h}$ to maximize the "estimated likelihood"

$$\hat{L}(h) = \prod_{j=1}^{n} \hat{f}_{nj}(X_j,h).$$

In a recent paper by Chow, Geman, and Wu (1982), an example is given which shows that if f is not compactly supported, then $\hat{h}$ diverges to ∞ (in probability) from which it follows that $\hat{f}_n(x,\hat{h})$ is not even a consistent estimator of f. So they assume that f is compactly supported and, with considerable effort, establish consistency.

Despite this promising result, the cross-validated estimator can still be very poorly behaved. In section 2 an example is given which shows that the cross-validated $\hat{h}$ can be strongly biased by observations X_i for which $f(X_i)$ is close to 0. A reasonable way to eliminate this effect is to find an interval, [a,b], on which f is known to be bounded above 0. The assumption of the existence of such an [a,b] seems easy to accept. Indeed, most practitioners should have no trouble finding reasonable candidates. Next, let

$$A = \{j: X_j \epsilon [a,b]\}, \tag{1.1}$$

redefine the estimated likelihood,

$$\hat{L}(h) = \prod_{j \epsilon A} \hat{f}_{nj}(X_j,h),$$

and take $\hat{h}$ to maximize $\hat{L}(h)$.

This estimator has been studied by Hall (1982a). His results show that, while the pathologies of section 2 cause no problems, the cross-validated estimator still behaves suboptimally with respect to the rate of convergence of mean square error. It is surprising to note that the dominant term in his calculations depends only on the behavior of f at the endpoints of [a,b]. This effect will be discussed heuristically in Remark 6.4.

In this paper (see section 3) a modification of cross-validation is proposed which eliminates this endpoint effect by "pre-smoothing" the set A. In section 5 it is seen that, under fairly mild conditions on f, this technique gives a density estimator which has excellent asymptotic properties.

In section 4, a means of measuring the smoothness of a function is given. This is formulated differently from the usual Lipschitz conditions on derivatives or tail conditions on Fourier transforms, but is seen to be much more natural for the density estimation problem. It also simplifies the formulation of the theorems of section 5.

Section 6 contains some remarks on the strengths and weaknesses of pre-smoothed cross-validation. Section 7 contains the proof of theorem 2. Section 8 contains an outline of the proof of theorem 1.

2. PATHOLOGIES IN ORDINARY CROSS-VALIDATION

To see why Chow, Geman, and Wu (1982) had to work so hard for their result, consider the following example. Suppose the density f has cumulative distribution function F so that, for some $\epsilon > 0$,

$$F(x) = e^{-\frac{1}{x}} \qquad \text{for } x\epsilon\ (0,\epsilon).$$

Such an f could easily be constructed to be infinitely differentiable. Let $X_{(1)}$ and $X_{(2)}$ denote the first two order statistics of $X_1,\ldots,X_n$. It can be shown by straight-forward computations that, for any $\alpha > 0$,

$$\lim_{n\to\infty} P[X_{(2)} - X_{(1)} \leq n^{-\alpha}] = 0 \ .$$

But for K comparctly supported, $\hat{L}(h) = 0$ unless $h \geq c(X_{(2)} - X_{(1)})$ for some constant c. Thus, the cross-validated $\hat{h}$ must converge to 0 slower than any algebraic rate.

By the familiar variance and bias decomposition (see (4.7) or Rosenblatt (1971)) the mean square error may be written:

$$E(\hat{f}_n(x,h)-f(x))^2 = O(\frac{1}{nh}) + O(h^{2p}),$$

where p represents the amount of smoothness that is assumed on f. Hence, it is seen that cross-validation behaves quite poorly in the mean square sense. Analogous, though not so dramatic examples can be constructed by taking, for k large,

$$F(x) = x^k \qquad x\epsilon(0,\epsilon) \ .$$

These examples indicate that, even when f is very smooth and compactly supported, cross-validated estimators, as proposed by Habbema, Hermans, and van den Broek (1974), can be drastically affected by observations where f is close to 0.

3. PRE-SMOOTHED CROSS-VALIDATION

One drawback to cross-validation as proposed by Habbema, Hermans, and van den Broek (1974), is that it can be computationally very expensive. Note that, for each h, computation of $\hat{L}(h)$ involves computing n different density estimators. To avoid this difficulty, Schuster and Gregory (1978) proposed the following device.

Assume the sample size is even and split the sample (randomly) into two equal subsets, which will be denoted here (after a change of indices) by $X_1,\ldots, X_n$ and $Y_1,\ldots,Y_n$. Then form two kernel estimators,

$$\hat{f}_n^X(x,h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h}) , \qquad \hat{f}_n^Y(x,h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-Y_i}{h}) .$$

Now take $\hat{h}_X$ and $\hat{h}_Y$ to maximize, respectively, the estimated likelihoods,

$$\hat{L}^X(h) = \prod_{j=1}^{n} \hat{f}_n^X(Y_j,h) , \qquad \hat{L}^Y(h) = \prod_{j=1}^{n} \hat{f}_n^Y(X_j,h).$$

As the final estimator, take

$$(3.1) \qquad \hat{f}_n(x) = \frac{\hat{f}_n^X(x,\hat{h}_X)+\hat{f}_n^Y(x,\hat{h}_Y)}{2} .$$

Intuitively speaking, it seems there should be some loss of efficiency compared to the "leave one out" type of cross-validation. However, in many cases, this consideration is outweighed by the computational tractability of this "subset" cross-validation. This type of cross-validation is used here because the proofs are technically simpler and thus the ideas involved are more readily apparent.

The optimality theorems of section 5 will be formulated in terms of $\hat{f}_n^X$ and $\hat{h}_X$. It is easily seen that similar theorems also apply to the $\hat{f}_n$ of (3.1). It is also apparent from the theorems that, at least asymptotically, "subset" cross-validation is competitive with "leave one out" cross-validation.

To simplify the notation, for the rest of this paper, redefine

$$\hat{f}_n(x,h) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-X_i}{h}).$$

Also redefine, for a randomly chosen (as in (1.1) for example) subset $A\subset\{1,\dots,n\}$,

$$\hat{L}(h) = \prod_{j\in A} \hat{f}_n(Y_j,h),$$

and let $\hat{h}$ maximize $\hat{L}(h)$. Now a means of choosing the set A will be given which will eliminate the endpoint effect observed by Hall (1982a).

As in Hall (1982a) it is assumed that f is bounded above 0 on an interval [a,b]. Those values of j for which Y_j is too near the endpoints of [a,b] are removed from A in a "smooth fashion" by the following device. Construct a function q(x) so that:

(3.2) q(x) is supported on [a,b],

(3.3) q(x) is infinitely differentiable on [a,b],

(3.4) $0 \le q(x) \le 1$,

(3.5) $q(x) > 0$ for some x.

For $j = 1,\dots,n$, given the value of Y_j, randomly put $j \in A$ with probability $q(Y_j)$. Assume this is done independently for each j and is also independent of $X_1,\dots,X_n$ and $Y_1,\dots,Y_{j-1}$, $Y_{j+1},\dots,Y_n$. Note that the cross-validation scheme of Hall (1982a) is of this type where q(x) is the indicator function of [a,b].

It will be convenient to define

$$p = P[j \in A] = \int q(y)f(y)dy$$

Conditional on the event $\{j \in A\}$, Y_j has density

$$\frac{q(y)f(y)}{p} .$$

4. A MEASURE OF SMOOTHNESS

Before giving the measure of smoothness, more restrictions will be placed on the kernel K. Let k denote a large, positive, even integer which will remain fixed throughout this paper. It is assumed that K is a real valued function for which

(4.1) $\int K(x)dx = 1$,

(4.2) $\int x^i K(x)dx = 0 \qquad$ for $i = 1,\ldots,k-1$,

(4.3) $\int x^k K(x)dx = Bk! > 0$,

(4.4) K is symmetric about 0,

(4.5) K is bounded and compactly supported,

(4.6) K is differentiable and K, K' have bounded variation.

The first two assumptions are crucial to the proper behavior of kernel estimators. The rest have been included for technical convenience.

Let $w(x)$ denote a nonnegative "weight function". A common means of measuring the error of the estimator $\hat{f}_n$ (in estimating f) is the (weighted) Mean Integrated Square Error (MISE),

$$E\int(\hat{f}_n(x)-f(x))^2 w(x)dx.$$

In the case of kernel estimation, this quantity is usually decomposed into variance and bias terms as follows (assuming f uniformly continuous)

$$\text{var } \hat{f}_n(y,h) = \frac{1}{n^2h^2}\sum_{i=1}^{n}\left(EK\left(\frac{X_i-y}{h}\right)^2 - \left[EK\left(\frac{X_i-y}{h}\right)\right]^2\right) =$$

$$= \frac{1}{nh^2}\int K\left(\frac{x-y}{h}\right)^2 f(x)dx - \frac{1}{nh^2}\left[\int K\left(\frac{x-y}{h}\right)f(x)dx\right]^2 =$$

$$= \frac{1}{nh}\int K(u)^2 f(y+hu)du - \frac{1}{n}\left[\int K(u)f(y+hu)du\right]^2 =$$

$$= \frac{1}{nh}f(y)\int K(u)^2 du + o\left(\frac{1}{nh}\right) .$$

Similarly,

$$E\hat{f}_n(y,h) = \int K(u)f(y+hu)du.$$

Thus,

$$E[\hat{f}_n(y,h)-f(y)]^2 = \frac{1}{nh}f(y)\int K(u)^2du + o\left(\frac{1}{nh}\right) + \left[\int K(u)f(y+hu)du-f(y)\right]^2 ,$$

from which it follows that,

$$(4.7)\qquad E\int(\hat{f}_n(y,h)-f(y))^2\ w(y)dy = \frac{1}{nh}(\int f(y)w(y)dy)(\int K(u)^2du) + o(\frac{1}{nh})$$

$$+ \int[\int K(u)f(y+hu)du - f(y)]^2\ w(y)dy.$$

For the bias term, adopt the notation

$$s_f(h) = \int[\int K(u)(f(x+hu)-f(x))du]^2\ w(x)dx.$$

Note that if f has a bounded p-th derivative (p k) and w is compactly supported, then by (4.1), (4.2), and Taylor's theorem, as h → 0,

$$s_f(h) = O_p(h^{2p}).$$

Thus the rate of convergence of $s_f(h)$ to 0 provides an "L^2-type" measure of how much of a Taylor's expansion f has (i.e., how smooth f is). This measure is very natural for density estimation because it correctly takes into account difficulties about more smoothness at one point than at another.

5. OPTIMALITY THEOREMS

It is assumed that all quantities are defined as above. In particular, q and K are assumed to satisfy (3.2)-(3.5) and (4.1)-(4.6) respectively. In addition, the following is assumed about the underlying density, f,

(5.1) f is bounded,

(5.2) f is bounded above 0 on [a,b],

(5.3) f is differentiable and f, f' are of bounded variation,

(5.4) $\liminf_{h \to 0} h^{-k} s_f(h) = \infty$.

Now the main result of this paper can be stated.

Theorem 1: Choosing h to maximize $\hat{L}(h)$ is the same as choosing h to minimize

$$E\int[\hat{f}_n(x,h)-f(x)]^2 \frac{q(x)}{f(x)} dx + o_p(E\int[\hat{f}_n(x,h)-f(x)]^2 \frac{q(x)}{f(x)} dx).$$

Hence, if MISE with weight function $w(x) = q(x)/f(x)$ is accepted as an error criterion, then the pre-smoothed cross-validated $\hat{h}$ yields a density estimator which not only attains the optimal rate of convergence in the exponent sense, but the coefficient is also asymptotically optimal in the sense of minimizing MISE for this particular kernel estimator. Even if the admittedly artificial choice of $w(x) = q(x)/f(x)$ is not accepted the estimator will still have optimal exponent of convergence for a wide class of weight functions. It should be noted that this estimator makes no use of prior knowledge of the precise amount of smoothness of f.

Assumptions (5.3) and (5.4) seem quite strong. To see how much can be said in their absence, the following is presented as the first step in the proof of theorem 1.

Theorem 2: Under assumptions (5.1) and (5.2), choosing h to maximize $\hat{L}(h)$ is the same as choosing h to minimize,

$$\sum_{j\epsilon A} \frac{[\hat{f}_n(Y_j,h)-f(Y_j)]^2}{f(Y_j)^2} + o_p\Big(\sum_{j\epsilon A} \frac{[f_n(Y_j,h)-f(Y_j)]^2}{f(Y_j)^2}\Big) - nh^k \frac{2B}{p} \int q^{(k)}(x)f(x)dx +$$

$$+ o(nh^k) + 0_p(h^{-1/2}) + 0_p(n^{1/2}s_f(h)^{1/2}) \tag{5.5}$$

Note that if the first term dominates the other terms, then the pre-smoothed cross-validated choice of $\hat{h}$ is intuitively very attractive. Assumptions (5.3) and (5.4) are sufficient conditions for this dominance.

6. REMARKS

Remark 6.1: The most well-taken objection against pre-smoothed cross-validated density estimation is the very restrictive assumption (5.4). This says, somewhat paradoxically, that f cannot be too smooth. However, an inspection of the proof of theorem 1 shows that the first term in (5.5) may not be dominant unless some such assumption is made.

Of course, the class of f's that satisfy the assumption can be made larger by taking k larger and larger, but this device requires using a kernel function, K, which seems more and more artificial. Also assumption (5.4) may be difficult for the practitioner to accept. Still this does not preclude pre-smoothed cross-validation from having small sample properties that are superior to other types of cross-validation.

Remark 6.2: The small sample properties of the estimator presented in this paper could be improved by using the "leave one out" method of cross-validation. In view of the results of Hall (1982a,b) no trouble should be encountered in proving theorems 1 and 2 for a pre-smoothed version of "leave one out" cross-validation.

Remark 6.3: From theorem 2, it can be seen that if assumption (5.4) is violated, then the third term of (5.5) will be dominant. Thus the pre-smoothed cross-validated h could be "far too small" or "far too large" depending on the sign of the third term.

This difficulty can be eased somewhat by installing the following artificial "safety net." Find a density g, which is supported on [a,b], such that

$$\liminf_{h \to 0} \ h^{-k} \ s_g(h) = \infty \ .$$

By suitably adding observations from g to the data, it may be assumed that, for some $\alpha > 0$, the data consists of a sample from $(1-\alpha)f+\alpha g$. Heuristically, it seems reasonable that

$$\liminf_{h \to 0} h^{-k} s_{\alpha f+(1-\alpha)g}(h) = \infty \quad .$$

Thus this mixture density may be estimated optimally and a reasonable estimate of f may be obtained by subtraction.

Of course, for smooth f such an estimator will be suboptimal, but this is, asymptotically, better than leaving the third term of (5.5) unbounded. It is noted that this device of adding noise to the data set is very unappealing to the practitioner.

Remark 6.4: Another use of theorem 2 is that some heuristics can be given regarding the endpoint effect observed by Hall (1982a). Suppose for $\epsilon > 0$ that $q(x) = 1$ for $x \in (a+\epsilon, b-\epsilon)$. Then $q^{(k)}(x)$ is supported on $(a,a+\epsilon)$ and $(b-\epsilon,b)$. Also, for ϵ small, $q^{(k)}(x)$ is both very positive and very negative on each of the two intervals of support. Thus as ϵ tends to 0, the behavior of the integral,

$$\int f(x)q^{(k)}(x)dx,$$

is determined by $f'(a)$ and $f'(b)$. Hence, these derivatives appear in the third term of (5.5). To see how this connects with the results of Hall (1982a), recall he treats the case $\epsilon = 0$.

Remark 6.5: All the results presented here seemingly should be reasonably straightforward to extend to a multivariate density as well as to estimation of derivatives of f. In view of the results of Marron (1982) the multivariate version can be applied to give optimal solutions to the classification problem. It should be noted that the weight function used in this paper is almost exactly the one that arises naturally in Marron (1982).

Remark 6.6: There are several interesting problems which follow from the results of this paper, in addition to the major difficulty mentioned in Remark (6.1). For example, conceivably K could be chosen in some fashion analagous to the optimal kernel results of Epanechnikov (1969) and Sacks and Ylvisaker (1981). Analogously, perhaps the interval [a,b] and the pre-smoothing function q(x) can be chosen in some optimal fashion. Means of adapting this technique to other error criteria, such as sup norm can also be investigated.

7. PROOF OF THEOREM 1

This proof is based on techniques developed in Hall (1982a). Define the "likelihood with respect to A" by

$$L = \prod_{j\in A} f(Y_j).$$

Choosing h to maximize $\hat{L}(h)$ is the same as maximizing

$$\log\left(\frac{\hat{L}(h)}{L}\right) = \sum_{j\in A} \log\left(\frac{\hat{f}_n(Y_j,h)}{f(Y_j)}\right) .$$

Let

$$\Delta_n = \sup_{x\in[a,b]} \left|\frac{\hat{f}_n(x,h)-f(x)}{f(x)}\right| .$$

Suppose h = h(n) is chosen so that

$$h \to 0,\ (nh)^{-1} \log n \to 0.$$

Then, by theorem A of Silverman (1978)

$$\Delta_n \to 0 \text{ in probability.}$$

For $j \in A$, let

$$\Delta_{nj} = \frac{\hat{f}_n(Y_j,h)-f(Y_j)}{f(Y_j)} .$$

Note that

$$\log(\frac{\hat{L}(h)}{L}) = \sum_{j\in A} \log(1+\Delta_{nj}) = \sum_{j\in A} \Delta_{nj} \quad \frac{1}{2}(1+o_p(1)) \sum_{j\in A} \Delta_{nj}^2 .$$

Thus it is desired to choose h to maximize

$$(7.1) \qquad \sum_{j\in A} \Delta_{nj} - \frac{1}{2} \sum_{j\in A} \Delta_{nj}^2 + o_p(\sum_{j\in A} \Delta_{nj}^2) \quad .$$

Bounds will be found for the first term using the projections which Hall (1982a) has attributed to Hǎjek (1968). Define

$$\ell(x,y) = \frac{1}{nhf(y)} K(\frac{x-y}{h}) = \frac{1}{nhf(y)} K(\frac{y-x}{h}) \ ,$$

$$g_1(x) = E[\ell(x,Y_j)|j\in A]$$

$$g_2(y) = E\ \ell(X_i,y),$$

$$G \quad = E[\ell(X_i,Y_j)|j\in A].$$

By the change of variable u = (y-x)/h, (3.3), (4.1), (4.2), and (4.3),

$$\begin{aligned} (7.2) \qquad g_1(x) &= \int[\frac{1}{nhf(y)} K(\frac{y-x}{h})] \frac{q(y)f(y)}{p}\, dy = \\ &= \frac{1}{np} \int K(u)q(x+hu)du = \\ &= \frac{1}{np} \int K(u)[q(x)+huq'(x)+\ldots+\frac{(hu)^{k+1}}{(k+1)!} q^{(k+1)}(\xi)]du = \\ &= \frac{1}{np} [q(x)+ Bh^k q^{(k)}(x) + o(h^k)]. \end{aligned}$$

Thus,

$$(7.3) \qquad G = Eg_1(X_i) = \frac{1}{np}[p+ Bh^k \int q^{(k)}(x)f(x)dx] + o(n^{-1}h^k).$$

In a similar fashion

$$(7.4)\qquad E[(g_2(Y_j)-\tfrac{1}{n})^2 \big| j\epsilon A] = \int[\frac{1}{nf(y)}\,(\int K(v)f(y+hv)dv - f(y))]^2\, \frac{q(y)f(y)}{p}\, dy =$$

$$= \frac{1}{n^2 p}\; s_f(h).$$

Hence, by the Schwartz inequality,

$$(7.5)\qquad (G - \tfrac{1}{n})^2 = (E[g_2(Y_j) - \tfrac{1}{n}\big| j\epsilon A])^2 \;\le\; \frac{1}{n^2 p}\, s_f(h).$$

Let #(A) denote the cardinality of the set A. Note that,

$$(7.6)\qquad \sum_{j\epsilon A}\Delta_{nj} = \sum_{j\epsilon A}\;\sum_{i=1}^{n}\ell(X_i,Y_j) - \#(A) =$$

$$= \sum_{j\epsilon A}\;\sum_{i=1}^{n}[\ell(X_i,Y_j)-g_1(X_i)-g_2(Y_j)+G] +$$

$$+ \#(A)\sum_{i=1}^{n}[g_1(X_i)-G] + n\sum_{j\epsilon A}[g_2(Y_j)-G] + \#(A)(nG-1)$$

Each term will be bounded in turn.

To bound the first term in (7.6), let

$$r(x,y) = \ell(x,y) - g_1(x) - g_2(y) + G.$$

Routine computations show that

$$E[(\sum_{j\epsilon A}\;\sum_{i=1} r(X_i,Y_j))^2\big| A] = \sum_{j\epsilon A}\;\sum_{i=1}\; E[r(X_i,Y_j)^2\big| j\epsilon A]\ ,$$

and

$$E[r(X_i,Y_j)^2\big| j\epsilon A] = E[\ell(X_i,Y_j)^2 - G^2\big| j\epsilon A] - E[g_1(X_i)^2 - G^2] -$$

$$-E[g_2(Y_j)^2 - G^2\big| j\epsilon A] \le$$

$$\le E[\ell(X_i,Y_j)^2\big| j\epsilon A] =$$

$$= \iint[\frac{1}{nhf(y)}\;K(\frac{x-y}{h})]^2(\frac{q(y)f(y)}{p})f(x)dxdy =$$

$$= \frac{1}{n^2hp}\;\iint K(u)^2\;\frac{q(y)}{f(y)}\;f(u+hy)dudy =$$

$$= O(n^{-2}h^{-1})\ .$$

Hence,

$$E(\sum_{j\epsilon A}\sum_{i=1}^{n} r(X_i,Y_j))^2 = E(\#(A)\cdot n\cdot E[r(X_i,Y_j)^2 \big| j\epsilon A]) = O(h^{-1}).$$

Thus by the Markov inequality

$$(7.7) \quad \sum_{j\epsilon A}\sum_{i=1}^{n}[\ell(X_i,Y_j)- g_1(X_i)- g_2(Y_j) + G] = O_p(h^{-1/2}).$$

For the second term on the right of (7.6), using (7.2), (7.3) and the central limit theorem

$$\sum_{i=1}^{n}(g_1(X_i)- G) = \sum_{i=1}^{n}(\frac{1}{np}[q(X_i) + Bh^k q^{(k)}(X_i) + o(h^k)] -$$

$$- \frac{1}{np}[Eq(X_i) + Bh^k \; E\; q^{(k)}(X_i) + o(h^k)] =$$

$$= \frac{1}{np}\sum_i(q(X_i)- Eq(X_i)) + O_p(n^{-1/2}h^k) + o(h^k).$$

Thus,

$$(7.8) \quad \#(A)\sum_{i=1}^{n}(g(X_i) - G) = M_n + O_p(n^{1/2}h^k),$$

where M_n does not depend on h.

For the third term on the right of (7.6),

$$E[\sum_{j\epsilon A}(g_2(Y_j)-G)]^2 = E(\#(A)\; E[(g_2(Y_j)-G)^2|j\epsilon A]) =$$

$$= np\; E[(g_2(Y_j)- \frac{1}{n})^2 + 2(g_2(Y_j)-\frac{1}{n})(\frac{1}{n} - G)+(\frac{1}{n} - G)^2\big|j\epsilon A].$$

Hence, by (7.4) and (7.5),

$$E[\sum_{j\epsilon A}(g_2(Y_j)-G)]^2 = O(n^{-1}s_f(h))$$

from which it follows that

$$(7.9) \quad n\sum_{j\epsilon A}(g_2(Y_j)- G) = O_p(n^{1/2}s_f(h)^{1/2}).$$

Now the terms of (7.6) have been bounded by (7.7), (7.8), (7.9) and (7.3). The result is

$$\sum_{j\epsilon A} \Delta_{nj} = O_p(h^{-1/2}) + M_n + O_p(n^{1/2}h^k) + O_p(n^{1/2} s f(h)^{1/2}) + + nh^k \; \frac{B}{p} \int q^{(k)}(x) f(x) dx + o(nh^k).$$

The claim of theorem 2 follows directly from this and (7.1).

8. OUTLINE OF PROOF OF THEOREM 1

To see the idea behind this proof, note that

$$E\left[\frac{[\hat{f}_n(Y_j,h)-f(Y_j)]^2}{f(Y_j)^2} \,\middle|\, j\epsilon A\right] = E\int[\hat{f}_n(y,h)-f(y)]^2 \frac{q(y)}{pf(y)}\, dy.$$

Thus, speaking heuristically, if all the seemingly lower order terms in the statement of theorem 2 are indeed of lower order, then theorem 1 will follow from a law of larger numbers.

The main work in making the above argument rigorous is contained in the following lemma.

Lemma:

$$\frac{1}{\#(A)} \sum_{j\epsilon A} \frac{[\hat{f}_n(Y_j,h)-f(Y_j)]^2}{f(Y_j)^2} = \frac{1}{nh} \left(\int q(y)dy\right)\left(\int K(u)^2 du\right) + s_f(h) + o_p\left(\frac{1}{nh}\right) + o_p(s_f(h)).$$

The proof of this is not given here because it is too similar to the proof of theorem 1 of Hall (1982b). Note that assumptions (4.5) and (4.6) are somewhat stronger than those used by Hall.

Using the lemma, (4.7), and the fact that #(A) has a Binomial (n,p) distribution, the proof of theorem 1 will be complete when it is shown that the last terms of (5.5) are dominated by $\frac{1}{h} + ns_f(h)$. However, this follows easily from assumption (5.4).

ACKNOWLEDGEMENT

The author is indebted to Peter Hall for his development of the techniques used in proving the results in this paper.

10. REFERENCES

CHOW, Y.S., GEMAN, S., and WU, L.D. (1982). Consistent cross-validated density estimation. (To appear).

EPANECHNIKOV, V. (1969). Nonparametric estimates of a multivariate probability density. *Theor. Prob. Appl.* **14**, 153-158.

FARRELL, R.H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43**, 170-180.

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.

HǍJEK, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives. *Ann. Math. Statist.* **39**, 325-346.

HALL, P. (1982a). Cross-validation in density estimation. *Biometrika* **69**, 383-390.

HALL, P. (1982b). Limit theorems for stochastic measures of the accuracy of nonparametric density estimators. *Stoch. Processes Appln.* **13**, 11-25.

MARRON, J.S. (1982). Optimal rates of convergence in nonparametric classification (Ph.D. dissertation, UCLA).

ROSENBLATT, M. (1971). Curve Estimates. *Ann. Math. Statist.* **42**, 1815-1842.

SACKS, J. and YLVISAKER, D. (1981). Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.* **9**, 334-346.

SCHUSTER, E.F. and GREGORY, G.G. (1978). Choosing the shape factor(s) when estimating a density. *Inst. Math. Statist. Bull.* **7**, 292.

SILVERMAN, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.* **6**, 177-184.

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

WERTZ, W. (1978). *Statistical Density Estimation: A Survey*. Angewandte Statistique und Okonometrie **13**, Vandenhoeck and Ruprecht.