The Annals of Statistics
1984, Vol. 12, No. 4, 1285–1297

AN ASYMPTOTICALLY OPTIMAL WINDOW SELECTION RULE FOR KERNEL DENSITY ESTIMATES[1]

BY CHARLES J. STONE

University of California, Berkeley

Kernel estimates of an unknown multivariate density are investigated, with mild restrictions being placed on the kernel. A window selection rule is considered, which can be interpreted in terms of cross-validation. Under the mild assumption that the unknown density and its one-dimensional marginals are bounded, the rule is shown to be asymptotically optimal. This strengthens recent results of Peter Hall.

1. Introduction. Let $X_1, X_2, \cdots$ be independent $\mathbb{R}^d$-valued random variables having common unknown density p and consider the random sample $X_1, \cdots, X_n$ of size n. In this paper we will study the asymptotic behavior as the sample size tends to infinity of a certain window selection rule for kernel estimates of the unknown density based on the random sample.

The kernel estimates are of the form

$$p_{nh}(x) = (1/n) \textstyle\sum_1^n K_h(x - X_i),$$

where $K_h(x) = v_h^{-1}K(x/h)$. Here the "window" $h = (h_1, \cdots, h_d)$ belongs to $\mathbb{R}_+^d$, the collection of d-tuples of positive numbers; $v_h = h_1 \cdot \cdots \cdot h_d$ is the corresponding volume; $x/h = (x_1/h_1, \cdots, x_d/h_d)$ for $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$; and K is a function on $\mathbb{R}^d$ having integral one and satisfying some mild restrictions, which will be described in Section 2.

The integrated squared error loss $L_{nh} = \int (p_{nh} - p)^2$ of the estimate p_{nh} can be written as

$$L_{nh} = \int p_{nh}^2 - 2 \int p_{nh}p + \int p^2.$$

The goal of minimizing this loss is equivalent to that of minimizing

$$L_{nh} - \int p^2 = \int p_{nh}^2 - 2 \int p_{nh}p;$$

but this goal cannot be realized in practice, since $\int p_{nh}p$ is unknown. Observe, however, that

$$\int p_{nh}p = (1/n) \textstyle\sum_1^n \displaystyle\int K_h(x - X_i)p(x)\, dx$$

Received February 1984; revised June 1984.

[1] This research was supported in part by National Science Foundation Grant MCS83-01257.

AMS 1980 *subject classifications*. Primary 62G99; secondary 62H99.

Key words and phrases. Kernel density estimate, window selection rule, cross-validation, asymptotic optimality, Poissonization.

and hence that

$$E \int p_{nh} p = \int \int p(x)p(y)K_h(x-y)\,dx\,dy = EK_h(X-Y),$$

where X and Y are independent random variables each having density p. Consequently,

$$E \int p_{nh} p = E\left[\frac{1}{n(n-1)} \sum\sum_{i \neq j} K_h(X_i - X_j)\right],$$

where i and j are understood to range over $\{1, \cdots, n\}$. This leads to the unbiased estimate

$$\frac{1}{n(n-1)} \sum\sum_{i \neq j} K_h(X_i - X_j)$$

of $\int p_{nh} p$. A slight simplification leads to the estimate

$$\frac{1}{n^2} \sum\sum_{i \neq j} K_h(X_i - X_j)$$

of $\int p_{nh} p$; to the corresponding estimate

$$\begin{aligned} M_{nh} &= \int p_{nh}^2 - \frac{2}{n^2} \sum\sum_{i \neq j} K_h(X_i - X_j) \\ &= \frac{1}{n^2} \sum_1^n \sum_1^n K_h^{(2)}(X_i - X_j) - \frac{2}{n^2} \sum\sum_{i \neq j} K_h(X_i - X_j) \end{aligned}$$

of $L_{nh} - \int p^2$; and to the window selection rule, "choose the window h to minimize the criterion M_{nh}." This and other asymptotically equivalent criteria have been proposed and studied by Rudemo (1982), Bowman (1984), and Hall (1983a, 1983b). They point out that such criteria can also be thought of in terms of cross-validation. Specifically, let p_{nih} be the kernel estimate of p based on the random sample with the ith case removed:

$$p_{nih}(x) = 1/(n-1) \sum_{j \neq i} K_h(x - X_j).$$

Then

$$\frac{1}{n} \sum_1^n p_{nih}(X_i) = \frac{1}{n(n-1)} \sum\sum_{i \neq j} K_h(X_i - X_j)$$

is the cross-validation estimate of $\int p_{nh} p$.

An alternative, asymptotically equivalent, cross-validation criterion (estimate of $L_{nh} - \int p^2$) considered by these authors is

$$\frac{1}{n} \sum_1^n \int p_{nih}^2 - \frac{2}{n} \sum_1^n p_{nih}(X_i).$$

Hall showed that choosing $h \in H_n$ to minimize this cross-validation criterion is asymptotically optimal under certain conditions on K, H_n and p. In particular, K is assumed to be nonnegative. (If p is sufficiently smooth, then faster rates of convergence of the integrated squared error loss to zero can be obtained when the nonnegativity restriction on K is dropped; see Müller and Gasser, 1979.) The unknown density p is assumed to have a uniformly continuous square-integrable second derivative and to have finite second moment. Moreover,

$$H_n = \{(h_1, \cdots, h_1): \varepsilon \le n^{1/(4+d)} h_1 \le \lambda\},$$

where $0 < \varepsilon < \lambda < \infty$. On the other hand, two of the restrictions imposed on K in Section 2 of this paper, compact support and Hölder continuity, are not required in Hall's results. (No serious attempt has been made here to eliminate or weaken these restrictions on K, for it is numerically more efficient to compute M_{nh} when K is a suitably chosen function with compact support; also minimizing M_{nh} by a numerical search technique is more attractive when K is at least mildly smooth. When $d = 1$ these two considerations suggest using the triangular kernel K defined by $K(x) = 1 - |x|$ for $|x| \le 1$ and $K(x) = 0$ elsewhere; with this choice of K, after a preliminary sort of $X_1, \cdots, X_n$, the determination of M_{nh} for any given value of h requires only $O(n)$ computations.)

The purpose of this paper is to show that choosing $h \in \mathbb{R}_+^d$ to minimize M_{nh} is asymptotically optimal under a surprisingly mild assumption on p, namely that p and its one-dimensional marginals are bounded. In this level of generality, there are no known theoretical results on the asymptotic behavior of the optimal window h or the optimal rate of convergence to zero of the integrated squared error of estimation.

The main result is described in Section 2 and proven in Section 3. The formulation of the result and the method of proof were influenced to some extent by several recent theoretical investigations of the Final Prediction Error (FPE) and other closely related model selection criteria in the regression context: Shibata (1981), Breiman and Freedman (1983), Rice (1983) and Chen (1983). The relatively long proof of Lemma 3 in Section 3 is given in Section 4. It uses "Poissonization", which has been employed by Rosenblatt (1975), Krieger and Pickands (1981) and Nadaraya (1983) in related contexts; interestingly, it also uses multiple stochastic integration with respect to a Poisson process. In a footnote to problem 5 of XII, 6 Feller (1980) gives credit to Domb (1952) for the use of Poissonization to obtain elegant derivations of various formulas in combinatorial probability.

A result similar to Theorem 1 was obtained for histogram density estimates in Stone (1984). The method of proof was also similar, except that the Poissonization argument used to prove the analog of Lemma 3 was much simpler.

Under various restrictions, Krieger and Pickands (1981) and Sacks and Ylvisaker (1981) obtained asymptotically optimal selection rules for kernel estimates of the density at a fixed point. In the later paper the entire kernel, not just the window, was optimized.

2. Statement of the main result. As mentioned above, the kernel K is required to have integral one. In addition, it is required to be symmetric about the origin, to have compact support, and to be Hölder continuous; that is, such that for some positive constants β and c,

$$|K(y) - K(x)| \le c|y - x|^{\beta} \quad \text{for} \quad x, y \in \mathbb{R}^d$$

(here $|x| = (x_1^2 + \cdots + x_d^2)^{1/2}$ for $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$). The function K is not required to be nonnegative. Let $K^{(2)}$ denote the convolution of K with itself, so that $K^{(2)}(x) = \int K(x - y)K(y)\,dy$. Then $K^{(2)}$ satisfies the same assumptions as K; in addition, $K^{(2)}(0) = \int K^2(y)\,dy > 0$. The kernel K is further restricted by requiring that $K^{(2)}(0) < 2K(0)$ (which necessarily holds if K is nonnegative and $K(0) = \max_x K(x)$).

Let h, v_h, x/h and K_h be defined as in Section 1 and note that $0 < v_h \le |h|^d$. Also define $K_h^{(2)}$ by $K_h^{(2)}(x) = v_h^{-1}K^{(2)}(x/h)$. Then K_h and $K_h^{(2)}$ each have integral one and $K_h^{(2)}$ is the convolution of K_h with itself. Let p_{nh} and L_{nh} be defined as in Section 1, and observe that $\int p_{nh}^2$ and L_{nh} are both continuous on $\mathbb{R}_+^d$.

A window selection rule h_n is a $\mathbb{R}_+^d$-valued function of $X_1, \cdots, X_n$. Clearly

$$L_{nh_n}/\min_h L_{nh} \ge 1.$$

The indicated minimum is actually taken on at some $h \in \mathbb{R}_+^d$. For it is easily seen that

$$\liminf_{h \to \partial \mathbb{R}_+^d}\left(L_{nh} - \int p^2\right) \ge 0;$$

also if the coordinates of h are all large, then

$$\int p_{nh}^2 \sim v_h^{-1}K^{(2)}(0) \quad \text{and} \quad \int p_{nh}p \sim v_h^{-1}K(0),$$

so

$$L_{nh} - \int p^2 \sim v_h^{-1}(K^{(2)}(0) - 2K(0)) < 0.$$

(Here we have used the restriction that $K^{(2)}(0) < 2K(0)$.) The window selection rule h_n is said to be asymptotically optimal provided that

$$\lim_n (L_{nh_n}/\min_h L_{nh}) = 1 \text{ with probability one.}$$

Consider the window selection rule $\hat{h}_n$ defined to be a value of $h \in \mathbb{R}_+^d$ that minimizes the criterion M_{nh} introduced in Section 1. (It follows as in the previous paragraph that the minimum of M_{nh} is taken on at some $h \in \mathbb{R}_+^d$.) The one-dimensional marginals of p are defined to be the densities of the coordinates of X, where X has density p. The main result of this paper can now be stated simply as follows.

THEOREM 1. *If p and its one-dimensional marginals are bounded, then $\hat{h}_n$ is asymptotically optimal.*

Suppose p satisfies the assumptions of Theorem 1. Then, in the notation of Section 3, $\|p_h - p\| \to 0$ as $h \to 0$. Thus it follows from Theorem 1 together with

Lemma 1 and Lemma 4 of Section 3 that $\hat{h}_n$ and $L_{n\hat{h}_n}$ both converge to zero with probability one as $n \to \infty$. For contrasting results when the Fourier transform of p vanishes outside a compact set C and the Fourier transform of K is the indicator function of C, see Ibragimov and Khasminskii (1982).

Burman (1984) has concurrently used arguments of Shibata (1980, 1981) to obtain a more general asymptotic optimality result for density estimation (with "in probability" instead of "with probability one" in the definition of asymptotic optimality). When specialized to kernel density estimation, the window h is selected from a finite set $H_n = \{h_1, \cdots, h_{N_n}\}$ subject to certain restrictions on N_n and the deterministic sequence $h_1, h_2, \cdots$; p is assumed to be bounded; and K is required to have finite 8th moment, but K is not required to be symmetric or continuous or to have compact support.

For related work in which integrated squared error loss is replaced by other measures of loss see Chow, Geman and Wu (1983); Devroye and Györfi (1983); Stone (1983); Birgé (1983); Marron (1984); and Bowman, Hall and Titterington (1984). For a recent review of a wide variety of smoothing techniques in statistics see Titterington (1984).

3. Proof of Theorem 1. Throughout this section and the next one, it is assumed that p is bounded. Let p_h denote the convolution of K_h and p, so that

$$p_h(x) = \int K_h(x-y)p(y)\,dy = Ep_{nh}(x).$$

Set $\|p_h - p\| = (\int (p_h - p)^2)^{1/2}$ and let $s \wedge t$ denote the minimum of $s, t \in \mathbb{R}$.

LEMMA 1. *There are positive constants b and c such that*

$$\|p_h - p\|^2 \geq c(|h|^{bd} \wedge 1) \geq c(v_h^b \wedge 1) \quad \textit{for} \quad h \in \mathbb{R}_+^d.$$

PROOF. Let ϕ and ρ denote the Fourier transforms of K and p respectively. Then ϕ is bounded and continuous; it is real-valued since K is symmetric; it vanishes at infinity by the Riemann-Lebesgue lemma; it equals one at the origin and is not identically one on any neighborhood of the origin. The Fourier transform ϕ_h of K_h is given by $\phi_h(t) = \phi(ht)$, where $ht = (h_1t_1, \cdots, h_dt_d)$; and the Fourier transform of p_h is $\phi_h\rho$. According to Parseval's identity and the boundness of the density p, $\int |\rho|^2 = (2\pi)^d \int p^2 < \infty$ and

$$(2\pi)^d \|p_h - p\|^2 = \int |\phi_h\rho - \rho|^2 = \int (1-\phi_h)^2 |\rho|^2.$$

Now ρ is continuous and $\rho(0) = 1$, so there is a nonempty bounded open ball C centered at the origin of $\mathbb{R}^d$ such that $|\rho|^2 \geq ½$ on C. Also $\|p_h - p\|^2$ is bounded away from zero for h outside any neighborhood of the origin. Suppose the desired conclusion is false. It then follows easily from the power series for the cosine function and a compactness argument that there is a unit vector $u \in \mathbb{R}^d$ such that

$$\int_C dt\left(\int (ut \cdot x)^k K(x)\,dx\right)^2 = 0$$

for every positive even integer k. By continuity, for each such k,

$$\int (ut \cdot x)^k K(x)\, dx = 0 \quad \text{for all} \quad t \in C.$$

Choose $j \in \{1, \cdots, d\}$ such that $u_j \neq 0$. By proper choice of t it follows that

$$\int x_j^k K(x)\, dx = 0$$

for every even integer k. By the symmetry of K, this equality holds for every positive integer k. But this is clearly impossible, since K has integral one and compact support. (Suppose, say, that $j = 1$ and define K_1 by

$$K_1(x_1) = \int \cdots \int K(x_1, \cdots, x_d)\, dx_2 \cdots dx_d.$$

Then K_1 has integral one and compact support and $\int_{-\infty}^{\infty} x_1^k K_1(x_1)\, dx_1 = 0$ for every positive integer k. Consequently the Fourier transform of K_1 is identically equal to one, which contradicts the conclusion of the Riemann-Lebesgue lemma.)

Set

$$J_{nh} = \| p_h - p \|^2 + 1/nv_h,$$

$$J_{nhr} = v_h^r \wedge 1 + 1/nv_h \quad \text{for} \quad r > 0,$$

$$G_{nh} = n^{-1} \textstyle\sum_1^n p_h(X_i) - Ep_n(X),$$

and

$$G_n = n^{-1} \textstyle\sum_1^n p(X_i) - Ep(X).$$

A modified form of Theorem 1 will first be proven, in which h ranges over a finite subset H_n of $\mathbb{R}_+^d$, the number of whose elements increases at most algebraically fast in n; the original form of the theorem then follows (see the end of this section).

Condition 1. $\#(H_n) \leq An^a$ for $n \geq 1$, where A and a are positive constants.

Lemma 2. *If Condition* 1 *holds, then*

$$\lim_n \max_{h \in H_n} J_{nh}^{-1} | G_{nh} - G_n | = 0 \textit{ with probability one}$$

and

$$\lim_n \max_{h \in H_n} J_{nh}^{-1} \left| \int (p_{nh} - p_h)(p_h - p) \right| = 0 \textit{ with probability one.}$$

Proof. Set

$$Z_{ih} = p_h(X_i) - p(X_i) - (Ep_h(X) - Ep(X)).$$

Then Z_{ih}, $i \geq 1$, are independent and identically distributed random variables each having mean zero. Since p is bounded, there is a positive constant c independent of h such that $|Z_{ih}| \le c$ and $\operatorname{Var}(Z_{ih}) \le cu_h^2$, where $u_h = \|p_h - p\|$. Observe that $G_{nh} - G_n = \bar{Z}_{nh} = (Z_{1h} + \cdots + Z_{nh})/n$. By Bernstein's inequality (see Hoeffding, 1963)

$$\Pr(|\bar{Z}_{nh}| \ge t) \le 2\exp[-\tau\lambda/2(1 + \lambda/3)],$$

where $0 \le \lambda \le t/u_h^2$ and $\tau = nt/c$. Choose $\varepsilon > 0$. Suppose that $u_h \ge n^{\varepsilon-1/2}$ Set $t = n^{\varepsilon-1/2}u_h$ and $\lambda = n^{\varepsilon-1/2}/u_h \le 1$. Then $\lambda\tau = n^{2\varepsilon}/c$. Suppose instead that $u_h < n^{\varepsilon-1/2}$. Set $t = n^{2\varepsilon-1}$ and $\lambda = 1$. Again, $\lambda\tau = n^{2\varepsilon}/c$. Thus in either case it follows from Bernstein's inequality that

$$\Pr(|\bar{Z}_{nh}| \ge t) \le 2\exp(-n^{2\varepsilon}/3c).$$

Hence by Condition 1.

$$\lim_n \Pr(|\bar{Z}_{nh}| \ge n^{\varepsilon-1/2}u_h + n^{2\varepsilon-1} \quad \text{for some} \quad h \in H_n) = 0.$$

Thus to verify the first conclusion of Lemma 2 it is enough to show that for some $\varepsilon > 0$

$$\lim_n \max_{u>0} \frac{n^{\varepsilon-1/2}u + n^{2\varepsilon-1}}{u^2 + 1/nu^{2/b}} = 0,$$

where the positive number b is defined as in Lemma 1. For $0 < \varepsilon < \tfrac{1}{2}(1 + b)$, this result is easily shown by considering separately: $0 < u \le n^{\varepsilon-1/2}$, $n^{\varepsilon-1/2} < u < n^{-b/2(1+b)}$, and $u > n^{-b/2(1+b)}$. The second conclusion of the Lemma follows from the same argument applied to

$$Z_{ih} = \int (K_h(x - X_i) - p_h(x))(p_h(x) - p(x))\,dx.$$

Let P_n denote the empirical distribution of $X_1, \cdots, X_n$ defined by

$$P_n(B) = n^{-1}\#\{i\colon 1 \le i \le n \quad \text{and} \quad X_i \in B\} \quad \text{for} \quad B \subseteq \mathbb{R}^d.$$

The proof of the next result is postponed to Section 4.

LEMMA 3. *If Condition* 1 *holds, then for all* $r > 0$

$$\lim_n \max_{h\in H_n} J_{nhr}^{-1} \left| \int\int_{x\neq y} K_h(x - y)(P_n(dx) - P(dx))(P_n(dy) - P(dy)) \right| = 0 \quad \textit{with probability one.}$$

LEMMA 4. *If Condition* 1 *holds, then for all* $r > 0$

$$\lim_n \max_{h\in H_n} J_{nhr}^{-1} \left| \int (p_{nh} - p_h)^2 - K^{(2)}(0)/nv_h \right| = 0 \quad \textit{with probability one.}$$

PROOF. Observe that

$$\int (p_{nh} - p_h)^2 = \int \left(\int K_h(z - x)(P_n(dx) - P(dx)) \right)^2 dz$$

$$= \int \int K_h^{(2)}(x - y)(P_n(dx) - P(dx))(P_n(dy) - P(dy))$$

$$= \int \int_{x \neq y} K_h^{(2)}(x - y)(P_n(dx) - P(dx))(P_n(dy) - P(dy))$$

$$+ K^{(2)}(0)/nv_h,$$

so the desired result follows from Lemma 3 (applied to $K_h^{(2)}$ instead of K_h).

Suppose now that h is constrained to lie in H_n, that $\hat{h}_n$ minimizes M_{nh} over H_n, and that Condition 1 holds. To verify that $\hat{h}_n$ is asymptotically optimal, it suffices to show that with probability one

$$\lim_n \max_{h,h' \in H_n} \frac{| L_{nh'} - L_{nh} - (M_{nh'} - M_{nh}) |}{L_{nh} + L_{nh'}} = 0.$$

For this it is enough to show that

(1) $$\lim\inf_n \min_{h \in H_n} (L_{nh}/J_{nh}) > 0 \text{ with probability one}$$

and

(2) $$\lim_n \max_{h,h' \in H_n} \frac{| L_{nh'} - L_{nh} - (M_{nh'} - M_{nh}) |}{J_{nh} + J_{nh'}} = 0 \text{ with probability one.}$$

Since

$$L_{nh} = \int (p_{nh} - p)^2$$

$$= \int (p_{nh} - p_h)^2 + \| p_h - p |^2 + 2 \int (p_{nh} - p_h)(p_h - p),$$

(1) follows from Lemmas 1, 2 and 4. Observe next (see Section 1) that

$$L_{nh} - M_{nh} - 2G_n - \int p^2$$

$$= 2(G_{nh} - G_n) + 2 \int \int_{x \neq y} K_h(x - y)(P_n(dx) - P(dx))(P_n(dy) - P(dy)).$$

Thus (2) follows from Lemmas 1, 2 and 3.

Since K is Hölder continuous, the original form of Theorem 1 can be derived from the modified form based on Condition 1; small, moderate and large values of the coordinates of h must be handled separately, the details being left to the reader. (Recall the assumption that the one-dimensional marginals of p are bounded. Accordingly, for given n, if one of the coordinates of h is very small, then $K_h^{(2)}(X_i - X_j) = K_h(X_i - X_j) = 0$ for $1 \le i < j \le n$ except on an event having very small probability.)

4. Proof of Lemma 3. The proof is based on "Poissonization." Given a positive number λ, let $N(dx)$ be a Poisson process on $\mathbb{R}^d$ with $EN(B) = \lambda P(B)$. By definition, $N(B)$ has a Poisson distribution; and if $B_1, \cdots, B_k$ are disjoint, then $N(B_1), \cdots, N(B_k)$ are independent. Set $M(dx) = N(dx) - \lambda P(dx)$. Also, given a positive integer ℓ, let P^ℓ denote the probability measure on $\mathbb{R}^{\ell d}$ defined by $P^\ell(dx_1 \cdots dx_\ell) = P(dx_1) \cdots P(dx_\ell)$.

Let k and ℓ denote positive integers with $\ell \le k$. Let $\Gamma_{k\ell}^0$ denote the collection of all k-tuples $i_1, \cdots, i_k$ of integers in $\{1, \cdots, \ell\}$ such that:

(a) each $i \in \{1, \cdots, \ell\}$ appears one or more times among $i_1, \cdots, i_k$;

(b) if $i, i' \in \{1, \cdots, \ell\}$ and $i < i'$, then i appears before i' among $i_1, \cdots, i_k$.

Given $x \in (x_1, \cdots, x_k) \in \mathbb{R}^k$ and $\gamma = (i_1, \cdots, i_k) \in \cup_1^k \Gamma_{k\ell}^0$, set $x_\gamma = (x_{i_1}, \cdots, x_{i_k})$. Let $\Gamma_{k\ell}$ denote the subcollection of all $\gamma = (i_1, \cdots, i_k) \in \Gamma_{k\ell}^0$ such that each $i \in \{1, \cdots, \ell\}$ appears two or more times among $i_1, \cdots, i_k$. Observe that $\Gamma_{k\ell}$ is empty for $\ell > [k/2]$, where $[c]$ is the greatest integer no greater than c. By definition, $\Gamma_{k1} = \{(1, \cdots, 1)\}$ for $k \ge 2$; while Γ_{42} consists of the three 4-tuples $(1, 1, 2, 2)$, $(1, 2, 1, 2)$ and $(1, 2, 2, 1)$.

LEMMA 5. *Let g be a (Borel) function on $\mathbb{R}^k$ such that*

$$\textstyle\sum_{\ell=1}^k \sum_{\gamma \in \Gamma_{k\ell}^0} |g(x_\gamma)| P^\ell(dx) < \infty.$$

Then

$$E \int \cdots \int g(x_1, \cdots, x_k) M(dx_1) \cdots M(dx_k) = \textstyle\sum_{\ell=1}^{[k/2]} \lambda^\ell \sum_{\gamma \in \Gamma_{k\ell}} \displaystyle\int g(x_\gamma) P^\ell(dx).$$

PROOF. It suffices to prove the result for functions g of the product form $g(x_1, \cdots, x_k) = \prod_1^k \Psi_j(x_j)$, where Ψ_j, $1 \le j \le k$, are bounded; the general result follows by the usual L^1 approximation argument. For functions of the indicated product form the desired result follows in a straightforward manner from the formula

$$E \exp\left(\textstyle\sum_1^k t_i \displaystyle\int \Psi_i dM\right) = e^\phi,$$

where

$$\phi = \lambda \int (e^{\sum t_i \Psi_i} - 1 - \textstyle\sum t_i \Psi_i)\, dP.$$

Observe that

$$E \textstyle\prod_1^k \displaystyle\int \Psi_i \, dM = \frac{\partial^k e^\phi}{\partial t_1 \cdots \partial t_k} \Big|_0;$$

here $|_0$ means that $t_1 = \cdots = t_k = 0$. Note that $\phi|_0 = 0$ and $\partial\phi/\partial t_j|_0 = 0$. Thus it follows, for example, that

$$E\int \Psi_1\,dM \int \Psi_2\,dM = \frac{\partial^2 e^{\phi}}{\partial t_1 \partial t_2}\,|_0 = \left(\frac{\partial^2 \phi}{\partial t_1 \partial t_2} + \frac{\partial \phi}{\partial t_1}\frac{\partial \phi}{\partial t_2}\right)e^{\phi}|_0$$

$$= \frac{\partial^2 \phi}{\partial t_1 \partial t_2}\,|_0 = \lambda \int \Psi_1 \Psi_2\,dP = \lambda \int g(x_1, x_1)\,dP$$

where $g(x_1, x_2) = \Psi_1(x_1)\Psi_2(x_2)$.

For results related to Lemma 5 see Ogura (1972) and Krausz (1975).

LEMMA. 6. *For each positive integer k there is a positive constant c_k such that*

$$E\left[\left(\int\int_{x\neq y} K_h(x-y)M(dx)M(dy)\right)^{2k}\right] \le c_k v_h^{-2k} \textstyle\sum_{\ell=2}^{2k} \lambda^{\ell} v_h^{[(\ell+1)/2]}$$

for $\lambda > 0$ and $h \in \mathbb{R}_+^d$

PROOF. It follows from Lemma 5 that the indicated expectation is a finite linear combination of terms of the form

$$\lambda^{\ell} \int \cdots \int \textstyle\prod_m K_h^{\nu_m}(x_{i_m} - x_{j_m})P(dx_1) \cdots P(dx_{\ell}),$$

where $1 \le i_m < j_m \le \ell$ and $\nu_m > 0$ for all m, $2 \le \ell \le 2k$, $\sum_m \nu_m = 2k$, and each $i \in \{1, \cdots, \ell\}$ appears at least once in the sequence $i_1, j_1, i_2, j_2, \cdots$. It follows easily from the boundedness of p and the definition of K_h that terms of this form are bounded in absolute value by a constant multiple of $\lambda^{\ell} v_h^{-2k} v_h^{[(\ell+1)/2]}$. The desired result now follows immediately.

Set $N = N(\mathbb{R}^d)$.

LEMMA 7. *For each positive integer k there is a positive constant c_k such that*

$$E\left[\left(\int\int_{x\neq y} K_h(x-y)(N(dx) - NP(dx))(N(dy) - NP(dy))\right)^{2k}\right]$$

$$\le c_k(\lambda + \lambda^{2k} + v_h^{-2k} \textstyle\sum_{\ell=2}^{2k} \lambda^{\ell} v_h^{[(\ell+1)/2]}) \quad \textit{for} \quad \lambda > 0 \quad \textit{and} \quad h \in \mathbb{R}_+^d.$$

PROOF. Observe first that

$$\int\int_{x\neq y} K_h(x-y)(N(dx) - NP(dx))(N(dy) - NP(dy))$$

$$= \int\int_{x\neq y} K_h(x-y)M(dx)M(dy) - 2(N-\lambda)\int p_h\,dM + (N-\lambda)^2 \int p_h p.$$

Now $|\int p_h p|$ is bounded in h and $E(N-\lambda)^{4k}$ is bounded above by a constant multiple of $\lambda + \lambda^{2k}$. Also $p_h(x)$ is bounded in h and x and

$$E \exp\left(t \int p_h \, dM\right) = \exp\left(\lambda \int (e^{tp_h} - 1 - tp_h)p\right),$$

so each cumulant of $\int p_h \, dM$ is a multiple of λ that is bounded in h. Since this random variable has mean zero, its $4k$th moment is bounded above by a constant multiple of $\lambda + \lambda^{2k}$. The desired result now follows from Lemma 6.

LEMMA 8. *For each positive integer k there is a positive constant c_k such that*

$$E\left[\left(\int\int_{x \neq y} K_h(x-y)(P_n(dx) - P(dx))(P_n(dy) - P(dy))\right)^{2k}\right]$$
$$\leq c_k n^{-4k}\left(n^{2k} + v_h^{-2k} \textstyle\sum_{\ell=2}^{2k} n^{\ell} v_h^{[(\ell+1)/2]}\right) \qquad \text{for} \quad n \geq 1 \quad \text{and} \quad h \in \mathbb{R}_+^d.$$

PROOF. Set $N_n(dx) = nP_n(dx)$ and

$$Z = \int\int_{x \neq y} K_h(x-y)(N_n(dx) - nP(dx))(N_n(dy) - nP(dy)).$$

Let μ_n denote the $2k$th moment of Z and set $\mu_0 = 0$. Let $R(\lambda)$ denote the $2k$th moment of the random variable obtained through replacing n in the definition of Z by a Poisson random number N having mean λ, N being independent of X_i, $i \geq 1$. Then

$$R(\lambda) = \textstyle\sum_n \Pr(N = n)\mu_n = \sum_n (\lambda^n/n!)e^{-\lambda}\mu_n$$

determines a polynomial of degree $2k$ in λ with $R(0) = 0$, and by Lemma 7 there is a positive constant c_k' such that

$$0 \leq \sum_{j=1}^{2k} \frac{R^{(j)}(0)}{j!} \lambda^j = R(\lambda) \leq c_k'\left(\lambda + \lambda^{2k} + v_h^{-2k} \textstyle\sum_{\ell=1}^{2k} \lambda^{\ell} v_h^{[(\ell+1)/2]}\right)$$

for $\lambda > 0$ and $h \in \mathbb{R}_+^d$. By a straightforward argument, there is a positive constant c_k'' such that

$$\sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} \lambda^j \leq c_k''\left(\lambda + \lambda^{2k} + v_h^{-2k} \textstyle\sum_{\ell=1}^{2k} \lambda^{\ell} v_h^{[(\ell+1)/2]}\right)$$

for $\lambda > 0$ and $h \in \mathbb{R}_+^d$. (For suppose otherwise and note that for each fixed $c > 0$, if

$$\frac{|R^{(j)}(0)|}{j!} \lambda^j \gg c_k'\left(\lambda + \lambda^{2k} + v_h^{-2k} \textstyle\sum_{\ell=1}^{2k} \lambda^{\ell} v_h^{[(\ell+1)/2]}\right)$$

(where $a \gg b > 0$ means that a/b is "very large"), then

$$\frac{|R^{(j)}(0)|}{j!} (c\lambda)^j \gg \sum_{j=1}^{2k} \frac{R^{(j)}(0)}{j!} (c\lambda)^j \geq 0;$$

by normalization and a compactness argument, there would then be a nonzero polynomial in c of degree $2k$ which equals zero at more than $2k$ distinct points.) Consequently,

$$\mu_n = \sum_{j=1}^{2k} \frac{n! R^{(j)}(0)}{(n-j)! j!} \le \sum_{j=1}^{2k} \frac{|R^{(j)}(0)|}{j!} n^j$$
$$\le c_k''(n + n^{2k} + v_h^{-2k} \sum_{\ell=1}^{2k} n^{\ell} v_h^{[(\ell+1)/2]}),$$

which yields the desired result.

Lemma 3 follows from Lemma 8 and a Chebychev type inequality involving the $2k$th moment by considering four cases separately: $v_h \ge 1$, $n^{-1/(r+1)} \le v_h < 1$, $n^{-2} \le v_h < n^{-1/(r+1)}$, and $0 < v_h < n^{-2}$.

Acknowledgment. I wish to thank the Associate Editor and two referees for a prompt, conscientious and helpful review of this paper.

REFERENCES

BIRGÉ, L. (1983). On estimating a density using Hellinger distance and some other strange facts. Technical Report MSRI 045-83, Mathematical Sciences Research Institute, Berkeley.

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** 341–351.

BREIMAN, L. and FREEDMAN, D. A. (1983). How many variables should be entered in a regression equation? *J. Amer. Statist. Assoc.* **78** 131–136.

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

CHEN, K.-W. (1983). Asymptotically optimal selection of a piecewise polynomial estimator of a regression function. Ph.D. Dissertation, Dept. of Statist., Univ. of California, Berkeley.

CHOW, Y.-S., GEMAN, S. and WU, L. D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* **11** 25–38.

DEVROYE, L. and GYÖRFI, L. (1983). Nonparametric density estimation: The L_1 view. Manuscript.

DOMB, C. (1952). On the use of a random parameter in combinatorial problems. *Proc. Phys. Soc.* Sect A **65** 305–309.

FELLER, W. (1980). *An Introduction to Probability Theory and its Applications*, Vol. I, 3rd ed. Wiley, New York.

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

HALL, P. (1983). Asymptotic theory of minimum integrated square error for multivariate density estimation. *Proc. Sixth Internat. Symp. Multivariate Anal.* Pittsburgh, 25–29 July 1983, to appear.

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

IBRAGIMOV, I. A. and KHASMINSKI, R. Z. (1982). Estimation of distribution density belonging to a class of entire functions. *Theory Probab. Appl.* **27** 551–562.

KRAUSZ, H. K. (1975). Identification of nonlinear systems using random impulse train inputs. *Biol. Cybernetics* **19** 217–230.

KRIEGER, A. M. and PICKANDS, J. III (1981). Weak convergence and efficient density estimation at a point. *Ann. Statist.* **9** 1066–1078.

MARRON, J. S. (1984). An asymptotically efficient solution of the bandwidth problem of kernel density estimation. Manuscript.

MÜLLER, H.-G. and GASSER, T. (1979). Optimal convergence properties of kernel estimates of derivatives of a density function. In: *Smoothing Techniques for Curve Estimation* 144–154. T. Gasser and M. Rosenblatt, eds. Springer-Verlag, Berlin.

NADARAYA, E. A. (1983). A limit distribution of the square error deviation of nonparametric estimators of the regression function. *Z. Wahrsch. verw. Gebiete* **64** 37–48.

OGURA, H. (1972). Orthogonal functionals of the Poisson Process. *IEEE Trans. Information Theory* **IT-18** 473–481.

RICE, J. (1983). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12** 1225–1240.

ROSENBLATT, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3** 1–14.

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

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

SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68** 45–54.

STONE, C. J. (1983). Optimal uniform rate of convergence for nonparametric estimators of a density function or its derivatives. *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday*, 393–406, M. H. Rizvi, J. S. Rustagi, and D. Siegmund (eds.). Academic, New York.

STONE, C. J. (1984). An asymptotically optimal histogram selection rule. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* **II.** Wadsworth, Belmont, California. To appear.

TITTERINGTON, D. M. (1985). Common structure of smoothing techniques in statistics. *Internat. Statist. Rev.* **53**, to appear.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720