Ann. Inst. Statist. Math.
Vol. 43, No. 2, 311-326 (1991)

ON THE ALMOST EVERYWHERE PROPERTIES OF THE KERNEL REGRESSION ESTIMATE*

MIROSŁAW PAWLAK

Department of Electrical and Computer Engineering, The University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada

(Received August 29, 1988; revised October 20, 1989)

Abstract. The regression $m(x) = E\{Y \mid X = x\}$ is estimated by the kernel regression estimate $\hat{m}(x)$ calculated from a sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent identically distributed random vectors from $R^d \times R$. The second order asymptotic expansions for $E\hat{m}(x)$ and $\operatorname{var}\hat{m}(x)$ are derived. The expansions hold for almost all (μ) $x \in R^d$, μ is the probability measure of X. No smoothing conditions on μ and m are imposed. As a result, the asymptotic distribution-free normality for a stochastic component of $\hat{m}(x)$ is established. Also some bandwidth-selection rule is suggested and bias adjustment is proposed.

Key words and phrases: Regression function, kernel estimate, asymptotic expansions, distribution-free properties, asymptotic normality, bandwidth-selection, bias adjustment.

1. Introduction

Let (X, Y), $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent identically distributed $R^d \times R$-valued random vectors, and let $m(x) = E\{Y \mid X = x\}$ be the regression function of Y on X with $E|Y| < \infty$. Let μ denote the probability measure of X.

We estimate $m(x)$ with the following kernel estimate

$$\hat{m}(x) = \sum_{i=1}^{n} Y_i K\left(\frac{x - X_i}{h}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right),$$

where K is a bounded nonnegative Borel kernel and $h = h(n) \in R^+$ is the smoothing parameter (bandwidth). In the above definition and in the paper, $0/0$ is treated as 0.

* This work was supported by NSERC Grant A8131.

Estimator $\hat{m}$ was motivated by the classical Rosenblatt-Parzen density function estimate and was introduced independently by Watson (1964) and Nadaraya (1964).

Stone (1977) found a class of non-parametric regression estimates which can be consistent for all distributions of X even those not possessing density. This result has been extended to the estimate $\hat{m}$ by Devroye and Wagner (1980) and Spiegelman and Sacks (1980). The pointwise distribution-free consistency of $\hat{m}$ was first studied by Dervoye (1981). He, assuming that $E|Y|^p < \infty$, $p \geq 1$, proved that $E|\hat{m}(x) - m(x)|^p$ converges to zero as n increases to infinity, for almost all (μ) $x \in R^d$. Weak and strong consistency at almost all (μ) $x \in R^d$ has been examined by Krzyżak and Pawlak (1984), Greblicki *et al.* (1984) and Zhao and Fang (1985). The distribution-free pointwise weak and strong rate of convergence has been investigated by Krzyżak and Pawlak (1987). For the distribution-free results concerning other kernel regression estimates, we refer to the paper of Greblicki and Pawlak (1987), see also (1985).

In this paper, contrary to the above authors, we do not examine another consistency problem, but rather we obtain asymptotic distribution-free expansions for $E\hat{m}(x)$ and $\operatorname{var}\hat{m}(x)$. The expansions are of the order $O((nh^d)^{-2})$. The worked out technique, however, allows us to consider the remainder terms of any order of smallness. Not one continuity assumption on m is made and the results are valid for all distributions of X.

As a result, the asymptotic formulas for $\operatorname{var}\hat{m}(x)$ and $E\hat{m}(x)$ are established, and the asymptotic distribution-free normality of $\hat{m}(x) - E\hat{m}(x)$ is derived. This, in turn, allows us to consider the bandwidth selection problem in the case of discontinuous regression functions and underlying distributions. Moreover, certain adjustment of bias of the estimate is proposed.

The asymptotic normality of $\hat{m}(x)$ and the asymptotic expressions for $\operatorname{var}\hat{m}(x)$ and $E\hat{m}(x)$ have been examined by a number of authors (see Rosenblatt (1969), Schuster (1972) and Collomb (1977)). They imposed very restrictive assumptions on the distribution of (X, Y) and on the smoothing sequence (see also Prakasa Rao (1983), Section 4.2).

For other properties (uniform consistency, robust estimates and estimation of a broad class of functionals of the conditional distribution function) of the kernel regression estimate with random design we refer to Mack and Silverman (1982), Härdle and Marron (1985) and Härdle and Tsybakow (1988). See also Collomb (1985) for further references.

2. Preliminaries

Throughout the paper, norms are either all $\mathcal{I}_\infty$ or all $\mathcal{I}_2$. By $S_{x,h}$ we denote an open sphere with a radius h centered at $x \in R^d$. Suppose that the following conditions are satisfied:

$$h(n) \to 0 \quad \text{as} \quad n \to \infty, \tag{2.1}$$

$$nh^d(n) \to \infty \quad \text{as} \quad n \to \infty, \tag{2.2}$$

$$c_1 I_{\{\|x\| \leq r\}} \leq K(x) \leq c_2 I_{\{\|x\| \leq r\}}, \tag{2.3}$$

where c_1, c_2 and r are positive numbers, and I is the indicator function.

For further considerations, we shall need

LEMMA 2.1. *Let $Z_1, \ldots, Z_n$ be independent identically distributed random variables. Then for any real $q > 1$*

$$E\{|S_n/n|^q\} \le c(q)\{(n^{-1}\operatorname{var} Z_1)^{q/2} + n^{-(q-1)}E|Z_1 - EZ_1|^q\},$$

where $S_n = \sum_{j=1}^n (Z_j - EZ_j)$ and $c(q)$ is a positive constant.

This inequality is due to Rosenthal (1970) (see also Burkholder (1973) for detail about the constant $c(q)$).

LEMMA 2.2. *Let $(W_1, V_1), \ldots, (W_n, V_n)$ be pairs of independent identically distributed random variables. Let $S_n = \sum_{i=1}^n W_i$, $Q_n = \sum_{i=1}^n V_i$ and let $EW_1 = EV_1 = 0$.*

If $EW_1^2 < \infty$, $EV_1^3 < \infty$ then

(a) $E\{S_n Q_n^3\} = 3n(n-1)EV_1^2 E\{V_1 W_1\} + nE\{V_1^3 W_1\}$.

(b) $E\{S_n^2 Q_n^2\} = n(n-1)[EV_1^2 EW_1^2 + 2E^2\{V_1 W_1\}] + nE\{V_1^2 W_1^2\}$.

(c) $E\{S_n^2 Q_n^3\} = n(n-1)[3EV_1^2 E\{V_1 W_1^2\} + EV_1^3 EW_1^2 + 6E\{V_1 W_1\}E\{V_1^2 W_1\}] + nE\{V_1^3 W_1^2\}$.

PROOF. Let us consider the identity (c). Clearly,

$$\begin{aligned} E\{S_{n+1}^2 Q_{n+1}^3\} = {} & E\{S_n^2 Q_n^3\} + 3E\{S_n^2 Q_n\}EV_1^2 + ES_n^2 EV_1^3 \\ & + 6E\{S_n Q_n^2\}E\{V_1 W_1\} + 6E\{S_n Q_n\}E\{V_1^2 W_1\} + EQ_n^3 EW_1^2 \\ & + 3EQ_n^2 E\{W_1^2 V_1\} + E\{W_1^2 V_1^3\}. \end{aligned}$$

Since $E\{S_n^2 Q_n\} = nE\{V_1 W_1^2\}$, $ES_n^2 = nEW_1^2$, $EQ_n^2 = nEV_1^2$, $E\{S_n Q_n^2\} = nE\{W_1 V_1^2\}$, $E\{S_n Q_n\} = nE(V_1 W_1)$ and $EQ_n^3 = nEV_1^3$, it follows that

$$\begin{aligned} E\{S_{n+1}^2 Q_{n+1}^3\} = {} & E\{S_n^2 Q_n^3\} + 2n[3EV_1^2 E\{V_1 W_1^2\} \\ & + EV_1^3 EW_1^2 + 6E\{V_1 W_1\}E\{V_1^2 W_1\}] \\ & + E\{V_1^3 W_1^2\}. \end{aligned}$$

Noting that $ES_1^2 Q_1^3 = EW_1^2 V_1^3$ and then iterating the above recursive formula, one can easily find the postulated identity. Since the others' identities may be proved in the same way, the proof of Lemma 2.2 has been completed.

LEMMA 2.3. *Let g be a Borel measurable function and let $\int |g(x)|\mu(dx) < \infty$. If (2.3) holds then*

$$\frac{\int K\left(\frac{x-y}{h}\right) g(y)\mu(dy)}{\int K\left(\frac{x-y}{h}\right)\mu(dy)} \to g(x) \quad \text{as} \quad h \to 0$$

for almost all (μ) $x \in R^d$.

The proof of Lemma 2.3 for a more general class of kernels may be found in Greblicki *et al.* ((1984), Lemma 1).

We shall also need the following easily verified identity

$$u^{-1} = \sum_{i=0}^{p} (-1)^i \frac{(u-u_0)^i}{u_0^{i+1}} + (-1)^{p+1} \frac{(u-u_0)^{p+1}}{u u_0^{p+1}}, \tag{2.4}$$

where $p \geq 0$, u and $u_0 \neq 0$.

Furthermore, we shall use Corollary 10.50 in Wheeden and Zygmund (1977) which says that

$$\varphi_h(x) = \frac{\lambda(S_{x,h})}{\mu(S_{x,h})} \to \varphi(x) \quad \text{as} \quad h \to 0, \tag{2.5}$$

for almost all (μ) $x \in R^d$.

Here λ is the Lebesgue measure on R^d and $\varphi(x)$ is the Radon-Nikodym derivative of the μ-absolutely continuous part of λ. It is clear that $\varphi(x)$ is finite for almost all (μ) $x \in R^d$.

3. Main results

In the theorems presented in this section we give the asymptotic distribution-free expansions for $E\hat{m}(x)$ and $\operatorname{var} \hat{m}(x)$. Moreover, some consequences of the obtained expressions are established.

THEOREM 3.1. *Let* $E|Y|^{1+\epsilon} < \infty$, $\epsilon > 0$. *If* (2.1), (2.2) *and* (2.3) *hold then*

$$E\hat{m}(x) = m_h(x) + r_h(x) + O((nh^d)^{-2})$$

for almost all (μ) $x \in R^d$, *where*

$$m_h(x) = E\left\{m(X)K\left(\frac{x-X}{h}\right)\right\} \Big/ EK\left(\frac{x-X}{h}\right)$$

and

$$r_h(x) = E\left\{[m_h(x) - m(X)]K^2\left(\frac{x-X}{h}\right)\right\} \Big/ nE^2K\left(\frac{x-X}{h}\right).$$

THEOREM 3.2. *Let all the conditions of Theorem* 3.1 *be satisfied. If, in addition,* $E|Y|^{2+\epsilon} < \infty$, $\epsilon > 0$, *then*

$$\operatorname{var} \hat{m}(x) = v_h(x)/nEK\left(\frac{x-X}{h}\right) + O((nh^d)^{-2})$$

for almost all (μ) $x \in R^d$, *where*

$$v_h(x) = E\left\{\left[(Y - m_h(x))K\left(\frac{x-X}{h}\right)\right]^2\right\} \Big/ EK\left(\frac{x-X}{h}\right).$$

The proofs of Theorems 3.1 and 3.2 are deferred to the next section.

Remark 1. Let us note that the class of kernels satisfying condition (2.3) is practically confined to the uniform kernel, i.e. the kernel which equals 1 for $\|x\| \leq 1$ and 0 otherwise. This is due to the fact that the results of this paper rely on the distribution-free inequality in (4.3) which has been established by Devroye ((1981), Lemma 2.1). He has proved under condition (2.3) that $E|\hat{m}(x) - m(x)|^p \to 0$ as $n \to \infty$, $p \geq 1$ for almost all (μ) $x \in R^d$. It seems to be difficult to extend this result for a broader class of kernels. Nevertheless, assuming that μ is absolutely continuous and following the proof of Devroye's lemma one can extend the class of applicable kernels.

Specifically, if $c_1 H(\|x\|) \leq K(x) \leq c_2 H(\|x\|)$, where $H(t)$ is nonincreasing bounded function with $0 < H(0)$ and support $[0, a)$, $a < \infty$, then the following density-free version of (4.3) holds

$$E\left\{\sum_{i=1}^{n} |Y_i|^p K((x - X_i)/h) \Big/ \sum_{j=1}^{n} K((x - X_j)/h)\right\}$$
$$\leq 4 \int g(y) K((x - y)/h) f(y) dy \Big/ \int K((x - y)/h) f(y) dy$$
$$+ 8k^* \int_{S_{x,ah}} g(y) f(y) dy \Big/ \int K((x - y)/h) f(y) dy,$$

where $\sup K(x) = k^*$ and $f(x)$ is a density of X.

Furthermore, it is seen that if Y is a bounded random variable (the case occurring in the conditional distribution function estimation and discrimination problem) then the class of kernels can be as large as in Lemma 1 of Greblicki *et al.* (1984), i.e. including those without compact support.

Remark 2. Owing to (2.3) and (2.5) it is easy to prove that $h^d / EK((x-X)/h)$ is finite for almost all (μ) $x \in R^d$ and every $h > 0$. If, moreover, Y is bounded then Theorem 3.2 implies that $\operatorname{var} \hat{m}(x)$ converges to zero if $nh^d \to \infty$. On the other hand the bias $E\hat{m}(x) - m(x)$ tends to zero if both $nh^d \to \infty$ and $h \to 0$ are satisfied. More precisely, we have decomposed $E\hat{m}(x) - m(x)$ into two terms $m_h(x) - m(x)$ and $r_h(x) + O((nh^d)^{-2})$. The first term goes to zero if $h \to 0$, whereas the second one if $nh^d \to \infty$. Thus, the bias converges to zero if both (2.1) and (2.2) are satisfied, whereas the variance if only (2.2) holds. For comparison, the bias of the kernel density estimate tends to zero if $h \to 0$ and the variance if $nh^d \to \infty$. These observations suggest the following decomposition of the estimate

$$\hat{m}(x) = (\hat{m}(x) - m_h(x)) + m_h(x).$$

Now, the first term of the decomposition tends to zero (in probability for almost all (μ) $x \in R^d$) if $nh^d \to \infty$, while the second converges to $m(x)$ if only $h \to 0$.

Now, let us consider the expansions obtained in Theorems 3.1 and 3.2. First, let us note that because of Lemma 2.2, $m_h(x) \to m(x)$ as $h \to 0$ for almost all

(μ) $x \in R^d$. Next, we write

$$r_h(x) = \epsilon_h(x) EK^2\left(\frac{x-X}{h}\right) \Big/ nE^2K\left(\frac{x-X}{h}\right),$$

where

$$\epsilon_h(x) = E\left\{m(X)K\left(\frac{x-X}{h}\right)\right\} \Big/ EK\left(\frac{x-X}{h}\right) - E\left\{m(X)K^2\left(\frac{x-X}{h}\right)\right\} \Big/ EK^2\left(\frac{x-X}{h}\right).$$

Owing to Lemma 2.2, it follows that $\epsilon_h(x) \to 0$ as $h \to 0$ for almost all (μ) $x \in R^d$.

Further, by virtue of (2.3) and (2.5) we have

$$EK^2\left(\frac{x-X}{h}\right) \Big/ nE^2K\left(\frac{x-X}{h}\right) \leq (c_2/c_1cr^d)\varphi_{rh}(x)(nh^d)^{-1}, \tag{3.1}$$

where $c = \lambda(S_{0,1})$.

Thus, the second term in the bias expansion converges to zero at least as fast as $(nh^d)^{-1}$. This term, however, can vanish under some conditions. If, for example, K is the uniform kernel, then $r_h(x) \equiv 0$. These considerations yield

COROLLARY 3.1. *Under all the conditions of Theorem* 3.1

$$nh^d(E\hat{m}(x) - m_h(x)) \to 0 \quad \text{as} \quad n \to \infty \quad \text{for almost all } (\mu)\ x \in R^d.$$

Let us take the variance expansion into account. Let $\sigma^2(x)$ denote the conditional variance of Y, i.e.

$$\sigma^2(x) = E\{(Y - m(X))^2 \mid X = x\}.$$

The term $v_h(x)/nEK((x-X)/h)$ may be transformed to

$$\sigma_h^2(x) EK^2\left(\frac{x-X}{h}\right) \Big/ nE^2K\left(\frac{x-X}{h}\right), \tag{3.2}$$

where

$$\sigma_h^2(x) = m_h^2(x) + E\left\{(\sigma^2(X) + m^2(X))K^2\left(\frac{x-X}{h}\right)\right\} \Big/ EK^2\left(\frac{x-X}{h}\right) - 2m_h(x)E\left\{m(X)K^2\left(\frac{x-X}{h}\right)\right\} \Big/ EK^2\left(\frac{x-X}{h}\right).$$

It follows from Lemma 2.3 and $EY^2 < \infty$ that

$$\sigma_h^2(x) \to \sigma^2(x) \quad \text{as} \quad h \to 0 \tag{3.3}$$

for almost all (μ) $x \in R^d$.

Moreover, observing that

$$EK^2\left(\frac{x-X}{h}\right)\Big/ nE^2K\left(\frac{x-X}{h}\right) \geq (c_1^2/c_2^2cr^d)\varphi_{rh}(x)(nh^d)^{-1}$$

and using (2.5) and (3.1) we have

COROLLARY 3.2. *Suppose that all the conditions of Theorem* 3.2 *hold. Then*

$$\limsup_{n\to\infty} nh^d \operatorname{var}\hat{m}(x) \leq (c_2/c_1cr^d)\varphi(x)\sigma^2(x)$$

and

$$\liminf_{n\to\infty} nh^d \operatorname{var}\hat{m}(x) \geq (c_1^2/c_2^2cr^d)\varphi(x)\sigma^2(x)$$

for almost all (μ) $x \in R^d$.

The expression in (3.2) may be further decomposed noting that

$$EK^2\left(\frac{x-X}{h}\right)\Big/ nE^2K\left(\frac{x-X}{h}\right) = \varphi_{\alpha h}(x)\rho_h(x)(nh^d)^{-1},$$

where

$$\rho_h(x) = EK_0\left(\frac{x-X}{h}\right)EK^2\left(\frac{x-X}{h}\right)\Big/ E^2K\left(\frac{x-X}{h}\right),$$

$$K_0(x) = I_{S_{0,\alpha}}(x), \qquad \alpha = (\lambda(S_{0,1}))^{-1/d}.$$

Let us denote

$$\lim_{h\to 0}\rho_h(x) = \rho \tag{3.4}$$

for almost all (μ) $x \in R^d$.

Thus, (2.5), (3.2) and (3.3) follow.

COROLLARY 3.3. *Under all the conditions of Theorem* 3.2 *and* (3.4)

$$nh^d \operatorname{var}\hat{m}(x) \to \varphi(x)\sigma^2(x)\rho \quad as \quad n\to\infty$$

for almost all (μ) $x \in R^d$.

The function $\rho_h(x)$ plays the role of a similarity measure between the uniform kernel K_0 and the kernels satisfying (2.3). If $K = K_0$ then clearly $\rho_h(x) \equiv 1$. If, moreover, $\alpha \leq r$ then due to (3.1) $\rho_h(x) \leq c_2/c_1$. It is not simple, however, to determine $\lim_{h\to 0}\rho_h(x)$ for general μ. For μ being absolutely continuous or atomic, or a mixture of both of them, that limit may be found easily, which can be seen from the discussion below.

The next corollaries give us the important result of the previous considerations, that is, the asymptotic distribution-free normality of $\hat{m}(x) - E\hat{m}(x)$ and $\hat{m}(x) - m_h(x)$.

COROLLARY 3.4. *Let all the conditions of Theorem* 3.2 *be satisfied. Let* (3.4) *hold. If* $\varphi(x)$, $\sigma^2(x)$, $\rho \neq 0$ *then*

$$(nh^d)^{1/2}(\hat{m}(x) - E\hat{m}(x)) \to N(0, \varphi(x)\sigma^2(x)\rho)$$

in distribution as $n \to \infty$ *for almost all* (μ) $x \in R^d$.

The proof of Corollary 3.4 is postponed to the next section.

Combining the above result with Corollary 3.1, we have

COROLLARY 3.5. *Under all the assumptions of Corollary* 3.4

$$(nh^d)^{1/2}(\hat{m}(x) - m_h(x)) \to N(0, \varphi(x)\sigma^2(x)\rho)$$

in distribution as $n \to \infty$ *for almost all* (μ) $x \in R^d$.

Let us make a series of assumptions regarding the measure μ. At first, let μ have a density f, i.e. let μ be absolutely continuous with respect to λ. Theorem 9.13 in Wheeden and Zygmund (1977) says that

$$\lim_{h\to 0} h^{-d} \int K\left(\frac{x-y}{h}\right) f(y)\lambda(dy) = f(x) \int K(y)\lambda(dy),$$

for almost all (λ) $x \in R^d$.

This, together with the fact that $f(x) > 0$ for almost all (μ) $x \in R^d$ gives us

$$\rho_h(x) \to \int K^2(y)\lambda(dy) \Big/ \left(\int K(y)\lambda(dy)\right)^2$$

and

$$\varphi_{\alpha h}(x) \to 1/f(x) \qquad \text{as} \qquad h \to 0$$

for almost all (μ) $x \in R^d$.

From this and (3.4) we have

COROLLARY 3.6. *Let* μ *have a density* f. *Suppose that all the conditions of Theorem* 3.2 *hold. Then*

$$nh^d \operatorname{var} \hat{m}(x) \to \sigma_0^2(x) \qquad \text{as} \qquad n \to \infty$$

for almost all (μ) $x \in R^d$, *where*

$$\sigma_0^2(x) = (\sigma^2(x)/f(x)) \int K^2(y)\lambda(dy) \Big/ \left(\int K(y)\lambda(dy)\right)^2. \tag{3.5}$$

Let, in turn, $\mu = \mu_c + \mu_a$, where μ_c is absolutely continuous with respect to λ and μ_a denotes the atomic part of μ. It is not difficult to verify, assuming that $K(0)$ exists, that

$$\int K\left(\frac{x-y}{h}\right)\mu(dy) \to \mu_a(\{x\})K(0) \qquad \text{as} \qquad h \to 0.$$

Employing this observation one easily concludes that $\rho_h(x) \to 1$ as $h \to 0$ for almost all (μ) $x \in R^d$. Moreover, the remainder term in the variance expansion is $O(n^{-2})$. This and (3.3) yield

COROLLARY 3.7. *Let μ have both absolutely continuous and atomic part. Let $K(0)$ exist. If the conditions of Theorem 3.2 are satisfied then*

$$nh^d \operatorname{var} \hat{m}(x) \to 0$$

and

$$n \operatorname{var} \hat{m}(x) \to \sigma^2(x)/\mu_a(\{x\}) \qquad \text{as} \qquad n \to \infty$$

for almost all (μ) $x \in R^d$.

In the light of these results and Corollary 3.5 we have

COROLLARY 3.8. *Let the conditions of Corollaries 3.5 and 3.6 hold. Then*

$$(nh^d)^{1/2}(\hat{m}(x) - m_h(x)) \to N(0, \sigma_0^2(x))$$

in distribution as $n \to \infty$ for almost all (μ) $x \in R^d$, where $\sigma_0^2(x)$ is defined in (3.5).

COROLLARY 3.9. *Let the conditions of Corollaries 3.5 and 3.7 hold. Then*

$$n^{1/2}(\hat{m}(x) - m_h(x)) \to N(0, \sigma^2(x)/\mu_a(\{x\}))$$

in distribution as $n \to \infty$ for almost all (μ) $x \in R^d$.

Up to now we have examined the asymptotic normality of the stochastic component $\hat{m}(x) - m_h(x)$. To show that $\hat{m}(x) - m(x)$ has the same limit distribution one needs to consider the deterministic part $m_h(x) - m(x)$. This requires some smoothness conditions on $m(x)$.

Let, e.g. $m(x)$ be Lipschitz of order α, $0 < \alpha \leq 1$, in the neighborhood of x. Then, $|m_h(x) - m(x)| \leq ch^\alpha$ (Krzyżak and Pawlak (1987), Lemma 2).

Therefore we have

COROLLARY 3.10. *Let all the conditions of Corollary 3.5 hold. If m is Lipschitz of order α, $0 < \alpha \leq 1$, in the neighborhood of x and if $nh^{d+2\alpha} \to 0$, then*

$$(nh^d)^{1/2}(\hat{m}(x) - m(x)) \to N(0, \varphi(x)\sigma^2(x)\rho)$$

in distribution as $n \to \infty$.

The above considerations allow us to treat the case of optimal h, i.e. when $nh^{d+2\alpha} \to c$, $c > 0$. To do this, let us make specific assumptions concerning $m(x)$ and distribution of X. Let, e.g. $m(x)$ be continuous at $x \in R$ and let it have left and right first derivatives in a neighborhood of x. Assume also that a density of X

and its derivatives are continuous at x. Then letting $h(n) = v(x)n^{-1/3}$, $v(x) > 0$ and after a simple algebra we have

$$n^{1/3}(\hat{m}(x) - m(x)) \to N(v(x)\gamma(x), \sigma_0^2(x)/v(x)), \tag{3.6}$$

in distribution as $n \to \infty$, where

$$\gamma(x) = -\left[m^{(1)}(x+0)\int_{-\infty}^{0} yK(y)dy + m^{(1)}(x-0)\int_{0}^{\infty} yK(y)dy\right] \Big/ \int K(y)dy$$

and $\sigma_0^2(x)$ is given by expression (3.5).

The above result can be employed for the problem of selection of h. To do this one has to choose $v(x)$ (the factor $n^{-1/3}$ is optimal since it assures an optimal asymptotic rate) which minimizes the L_q local error, $q \geq 1$.

Owing to (3.6) such optimal $v(x)$ is determined by minimizing

$$\int |v(x)\gamma(x) + z(\sigma_0(x)/v^{1/2}(x))|^q \phi(z)dz,$$

where ϕ denotes the standard normal density function.

Clearly, that $v(x)$ depends on $\sigma^2(x)$, $f(x)$, $m^{(1)}(x+0)$ and $m^{(1)}(x-0)$. Those values can be easily estimated from the available data. Such "plug-in" scheme for $q = 2$ has been examined by Tsybakov (1987). He requires, however, much smoother conditions for $m(x)$ and $f(x)$ (see also Mack and Müller (1987)). In Hall (1984) a similar approach has been studied with respect to a global error. Hall and Wand (1988*a*, 1988*b*) investigate such rule in the context of nonparametric density estimation. For an alternative approach of choice of the bandwidth based on a cross-validation method we refer to Härdle and Marron (1985).

The above discussion can be easily extended for other types of singularities in $m(x)$ and $f(x)$, e.g. discontinuity of $m(x)$ and $f(x)$, discontinuity of first derivatives, etc. van Eeden (1985) examines that problem in the context of density estimation.

The established higher-order expansions can be employed to design more efficient estimates. That is, we are able to make adjustment of the estimate to improve its quality.

According to Theorem 3.1 $E\hat{m}(x) = m_h(x) + r_h(x) + O((nh^d)^{-2})$, where $m_h(x) \to m(x)$ and $r_h(x) \to 0$. The order of $r_h(x)$ is $(nh^d)^{-1}$.

Set

$$\bar{m}(x) = t_h(x)\hat{m}(x),$$

where $t_h(x) = m_h(x)/(m_h(x) + r_h(x))$.

Owing to Theorem 3.1 one can easily get

$$E\bar{m}(x) = m_h(x) + O((nh^d)^{-2}).$$

That is, $\bar{m}(x)$ is biased $O((nh^d)^{-2})$ compared to $O((nh^d)^{-1})$ for the original estimate. Moreover, since $t_h(x) \to 1$ as $h \to 0$ then the asymptotic behaviour of $\operatorname{var}\bar{m}(x)$ and $\operatorname{var}\hat{m}(x)$ are equivalent.

Obviously, the value $t_h(x)$ is unknown and it can be estimated as follows

$$\hat{t}(x) = \hat{m}(x)/(\hat{m}(x) + \hat{r}(x)),$$

where

$$\hat{r}(x) = (\hat{m}(x) - \hat{m}'(x))\hat{\beta}(x)$$

and

$$\hat{\beta}(x) = \sum_{i=1}^{n} K^2((x - X_i)/h) \Bigg/ \left(\sum_{i=1}^{n} K((x - X_i)/h)\right)^2 .$$

Here $\hat{m}'(x)$ stands for the kernel estimate with $K(x)$ replaced by $K^2(x)$.

One can conjecture that such defined adjustment $\bar{m}(x)$ (with $t_h(x)$ replaced by $\hat{t}(x)$) improves the original estimate quality. Some finite sample experiments would be desirable here, it is, however, beyond the scope of this paper.

The asymptotic expansions for $\operatorname{var}\hat{m}(x)$ and $E\hat{m}(x)$ have been established by Collomb (1977). By requiring the existence of f, assuming the continuity of $\sigma^2(x)$ and $f(x)$ and imposing $EY^2 < \infty$, $nh^d/n^w \to \infty$, $w > 0$, $\int K(y)dy = 1$, he reports that

$$\operatorname{var}\hat{m}(x) = (\sigma^2(x)/f(x)) \int K^2(y)dy(nh^d)^{-1} + o((nh^d)^{-1}).$$

This first order formula fails if, e.g. $\sigma^2(x) \equiv 0$ (which is a case in the absence of noise). Then, simple algebra resulting from Theorem 3.2 yields

$$\operatorname{var}\hat{m}(x) = hn^{-1}\|\operatorname{grad} m(x)\|^2 \int \|y\|^2 K(y)dy/f(x) + O((nh^d)^{-2}),$$

where appropriate smoothness conditions for m and f have been assumed.

Schuster (1972) showed the asymptotic normality of $\hat{m}(x)$, under conditions much more restrictive than in Corollary 3.8 (see also Rosenblatt (1969)). The asymptotic normality of the nearest neighbor type regressions estimates have been proved by Mack (1981) and Stute (1984). In the latter reference, the asymptotic normality (at almost all (μ) $x \in R$) of the nearest neighbor version of the kernel estimate is derived. The required assumptions are: a continuity of the distribution function of X, $nh^3 \to \infty$, $nh^5 \to 0$ as $n \to \infty$ and finiteness of EY^2.

4. Proofs

Proof of Theorem 3.1. Let us denote

$$a_n = \sum_{i=1}^{n} Y_i K_i/nEK_1, \qquad b_n = \sum_{i=1}^{n} K_i/nEK_1, \qquad \text{where} \qquad K_i = K\left(\frac{x - X_i}{h}\right).$$

Clearly, $\hat{m}(x) = a_n/b_n$. Using (2.4) with $p = 3$ and taking $u = b_n$, $u_0 = Eb_n$ we have

$$\hat{m}(x) = a_n - a_n(b_n - Eb_n) + a_n(b_n - Eb_n)^2 - a_n(b_n - Eb_n)^3 + a_n(b_n - Eb_n)^4/b_n. \tag{4.1}$$

The next step in our proof is to evaluate the expected values of all terms in (4.1). We first bound the last summand. By Hölder's and Jensen's inequalities, the expected value of the term does not exceed

$$E^{1/p}\left\{\sum_{i=1}^{n}|Y_i|^p K_i \Big/ \sum_{j=1}^{n} K_j\right\} E^{1/q}\{|b_n - Eb_n|^{4q}\}, \tag{4.2}$$

where $p > 1$ and $p^{-1} + q^{-1} = 1$.

Taking

$$|K_i| \le c_2,$$
$$E|K_i/EK_1|^q \le (c_2/EK_1)^{q-1}, \qquad q > 1,$$

and Lemma 2.1 into account we get

$$E|b_n - Eb_n|^{4q} \le c(4q)\{(c_2/nEK_1)^{2q} + (2c_2/nEK_1)^{4q-1} + (2/n)^{4q-1}\}.$$

By virtue of (2.3) and (2.5) the right side of the above inequality is not greater than

$$c(4q)\{[(c_2/c_1cr^d)\varphi_{rh}(x)(nh^d)^{-1}]^{2q} + [(2c_2/c_1cr^d)\varphi_{rh}(x)(nh^d)^{-1}]^{4q-1} + (2/n)^{4q-1}\},$$

where c is the constant defined in (3.1).

Thus,

$$E^{1/q}|b_n - Eb_n|^{4q} = O((nh^d)^{-2}) \qquad \text{a.e. } (\mu).$$

Devroye ((1981), Lemma 2.1) proved that

$$E\left\{\sum_{i=1}^{n}|Y_i|^p K_i \Big/ \sum_{j=1}^{n} K_j\right\} \le 7(c_2/c_1)\int_{S_{x,rh}} g(y)\mu(dy)/\mu(S_{x,rh}), \tag{4.3}$$

where $g(x) = E\{|Y|^p \mid X = x\}$.

This, together with the fact that $\int_{S_{x,h}} g(y)\mu(dy)/\mu(S_{x,h}) \to g(x)$ as $h \to 0$ a.e. (μ), see Wheeden and Zygmund ((1977), p. 189) follow that the expected value of the last term in (4.1) is $O((nh^d)^{-2})$ a.e. (μ).

Let us take the other terms in (4.1) into consideration. Clearly, $Ea_n = m_h(x)$ and by Lemma 2.3

$$\begin{aligned} E\{a_n(b_n - Eb_n)\} &= E\{Y_1K_1^2\}/nE^2K_1 - E\{Y_1K_1\}/nEK_1 \\ &= E\{Y_1K_1^2\}/nE^2K - m_h(x)/n. \end{aligned}$$

Next, employing (2.3), (2.5) and Lemma 2.3 we get

$$\begin{aligned} E\{a_n(b_n - Eb_n)^2\} &= E\{(a_n - Ea_n)(b_n - Eb_n)^2\} + m_h(x)\operatorname{var} b_n \\ &= n^{-2}[E\{Y_1K_1^3\}/E^3K_1 - 2E\{Y_1K_1^2\}/E^2K_1 \\ &\qquad - m_h(x)EK_1^2/E^2K_1 + 2m_h(x)] \\ &\quad + m_h(x)EK_1^2/nE^2K_1 - m_h(x)/n \\ &= m_h(x)EK_1^2/nE^2K_1 - m_h(x)/n + O((nh^d)^{-2}) \qquad \text{a.e. } (\mu). \end{aligned}$$

Furthermore,

$$(4.4)\quad E\{a_n(b_n - Eb_n)^3\} = m_h(x)E\{(b_n - Eb_n)^3\} + E\{(a_n - Ea_n)(b_n - Eb_n)^3\}.$$

The first term in the above expression is equal to

$$n^{-2}E\{((K_1 - EK_1)/EK_1)^3\} = O((nh^d)^{-2}) \qquad \text{a.e. } (\mu).$$

Employing Lemma 2.2(a) with $W_i = (Y_iK_i - EY_iK_i)/nEK_1$ and $V_i = (K_i - EK_i)/nEK_1$, one can bound the second term in (4.4) by

$$\begin{aligned}&3n^{-2}EK_1^2\left[\frac{E\{|Y_1|K_1^2\}}{E^2K_1} + |m_h(x)|\right]\Big/ E^2K_1\\ &\quad + n^{-3}\left[\frac{E|Y_1|K_1^4}{E^4K_1} + 3\frac{E|Y_1|K_1^3}{E^3K_1} + 3\frac{E|Y_1|K_1}{E^2K_1}\right.\\ &\qquad\left. + |m_h(x)|\frac{EK_1^3}{E^3K_1} + 3|m_h(x)|\frac{EK_1^2}{E^2K_1} + 3|m_h(x)|\right].\end{aligned}$$

From (2.3) and (2.5), it follows that the above expression is $O((nh^d)^{-2})$ a.e. (μ). The proof of Theorem 3.1 has been completed.

PROOF OF THEOREM 3.2. Making use of the notation of the proof of Theorem 3.1, squaring (4.1) and taking the expected value, we get

$$\begin{aligned}(4.5)\quad E\hat{m}^2(x) = {}& Ea_n^2 - 2E\{a_n^2(b_n - Eb_n)\} + 3E\{a_n^2(b_n - Eb_n)^2\}\\ &- 4E\{a_n^2(b_n - Eb_n)^3\} + 3E\{a_n^2(b_n - Eb_n)^4\}\\ &- 2E\{a_n^2(b_n - Eb_n)^5\} + E\{a_n^2(b_n - Eb_n)^6\} + Et_n^2 + 2E\{a_nt_n\}\\ &- 2E\{a_n(b_n - Eb_n)t_n\} + 2E\{a_n(b_n - Eb_n)^2t_n\}\\ &- 2E\{a_n(b_n - Eb_n)^3t_n\},\end{aligned}$$

where

$$t_n = a_n(b_n - Eb_n)^4/b_n.$$

Proceeding as in the proof of Theorem 3.1, we have

$$Ea_n^2 = \operatorname{var}\{Y_1K_1\}/nE^2K_1 + m_h^2(x)$$

and

$$\begin{aligned}E\{a_n^2(b_n - Eb_n)\} = {}& 2m_h(x)n^{-1}[E\{Y_1K_1^2\}/E^2K_1 - m_h(x)]\\ &+ O((nh^d)^{-2}) \qquad \text{a.e. } (\mu).\end{aligned}$$

With the help of Lemma 2.2(b) the third term in (4.5) is equal to

$$3m_h^2(x)EK_1^2/nE^2K_1 - 3m_h^2(x)/n + O((nh^d)^{-2}) \qquad \text{a.e. } (\mu).$$

Similarly, due to Lemma 2.2(c) and a little algebra one can verify that the fourth term in (4.5) is $O((nh^d)^{-3})$ a.e. (μ). In turn, using Hölder's inequality, Lemma

2.1 and the previous arguments we can easily find that the terms fifth, sixth and seventh in (4.5) are $O((nh^d)^{-3})$ a.e. (μ).

Because of Lemma 2.1 and arguing as in the proof of Theorem 3.1, we get

$$Et_n^2 = O((nh^d)^{-4}) \qquad \text{a.e. } (\mu).$$

From this and (4.2) we have

$$E\{a_n t_n\} \le (\operatorname{var} a_n)^{1/2} E^{1/2} t_n^2 + |m_h(x)| E|t_n| = O((nh^d)^{-2}) \qquad \text{a.e. } (\mu).$$

The other terms in (4.4) may be evaluated identically and they are of the order $O((nh^d)^{-2})$ a.e. (μ).

Since

$$E^2 \hat{m}(x) = m_h^2(x) + 2m_h(x) E\{(m_h(x) - m(X))K_1^2\}/nEK_1^2 + O((nh^d)^{-2}) \qquad \text{a.e. } (\mu)$$

the proof of Theorem 3.2 has been completed.

PROOF OF COROLLARY 3.4. It follows from (4.1) that

$$\hat{m}(x) - E\hat{m}(x) = a_n - Ea_n - m_h(x)(b_n - Eb_n) + \xi_n,$$

where ξ_n can be easily given in an explicit form. Applying Chebyshev's inequality and the results from the proofs of Theorem 3.1 and Theorem 3.2 we have

$$(nh^d)^{1/2} \xi_n \to 0 \qquad \text{in probability as} \qquad n \to \infty \qquad \text{a.e. } (\mu).$$

Let us note that

$$a_n - Ea_n - m_h(x)(b_n - Eb_n) = n^{-1} \sum_{j=1}^{n} \eta_{j,n} = T_n, \qquad \text{say,}$$

where

$$\eta_{j,n} = (Y_i - m_{h(n)}(x)) K\left(\frac{x - X_i}{h(n)}\right) \Big/ EK\left(\frac{x - X}{h(n)}\right).$$

Owing to Theorem 3.2 and Corollary 3.3 we get

$$nh^d \operatorname{var} T_n \to \varphi(x)\sigma^2(x)\rho \qquad \text{as} \qquad \to \infty \qquad \text{a.e. } (\mu).$$

Furthermore, for $p > 2$ and with the help of Lemma 2.3, (2.3) and (2.5) we get

$$\begin{aligned}
\sum_{j=1}^{n} E|n^{-1}\eta_{j,n}|^p &= n^{-(p-1)} E|\eta_{1,n}|^p \\
&\le (2c_2)^{(p-1)} [E\{|Y_1|^p K_1\}/EK_1 + |m_h(x)|^p]/(nEK_1)^{p-1} \\
&= O((nh^d)^{p-1}) \qquad \text{a.e. } (\mu).
\end{aligned}$$

This enables us to verify the Liapunov's condition. That is,

$$\frac{\sum_{j=1}^{n} E|n^{-1}\eta_{j,n}|^p}{(\operatorname{var} T_n)^{p/2}} = O((nh^d)^{1-p/2}) \qquad \text{a.e. } (\mu),$$

where $p > 2$. The proof of Corollary 3.4 has been completed.

Acknowledgements

The author thanks all referees for careful reading of the manuscript.

References

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

Collomb, G. (1977). Quelques proprietes de la methode du noyau pair l'estimation non parameterique de la regression en un point fixe', *C. R. Acad. Sci. Paris Ser. A*, **285**, 289–292.

Collomb, G. (1985). Nonparametric regression: an up-to-date bibliography, *Statistics*, **16**, 300–324.

Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates, *Ann. Statist.*, **9**, 1310–1319.

Devroye, L. and Wagner, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation, *Ann. Statist.*, **8**, 231–239.

Greblicki, W. and Pawlak, M. (1985). Fourier and Hermite series estimate of regression functions, *Ann. Inst. Statist. Math.*, **37**, 443–454.

Greblicki, W. and Pawlak, M. (1987). Necessary and sufficient consistency conditions for a recursive kernel regression estimate, *J. Multivariate Anal.*, **23**, 67–76.

Greblicki, W., Krzyżak, A. and Pawlak, M. (1984). Distribution-free pointwise consistency of kernel regression estimate, *Ann. Statist.*, **12**, 1570–1575.

Hall, P. (1984). Asymptotic properties of integrated square error and cross—validation for kernel estimation of a regression function, *Z. Wahrsch. Verw. Gebiete*, **67**, 175–196.

Hall, P. and Wand, M. P. (1988*a*). On the minimization of the absolute distance in kernel density estimation, *Statist. Probab. Lett.*, **6**, 311–314.

Hall, P. and Wand, M. P. (1988*b*). Minimization L_1 distance in nonparametric density estimation, *J. Multivariate Anal.*, **26**, 59–88.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation, *Ann. Statist.*, **13**, 1465–1481.

Härdle, W. and Tsybakow, A. B. (1988). Robust nonparametric regression with simultaneous scale curve estimation, *Ann. Statist.*, **16**, 120–135.

Härdle, W., Janssen, P. and Serfling, R. (1988). Strong uniform consistency rates for estimators of conditional functionals, *Ann. Statist.*, **16**, 1428–1449.

Krzyżak, A. and Pawlak, M. (1984). Distribution-free consistency of a nonparametric regression estimate and classification, *IEEE Trans. Inform. Theory*, **30**, 78–81.

Krzyżak, A. and Pawlak, M. (1987). The pointwise rate of convergence of the kernel regression estimate, *J. Statist. Plann. Inference*, **16**, 159–166.

Mack, Y. P. (1981). Local properties of *k-NN* regression estimates, *SIAM J. Algebraic Discrete Methods*, **2**, 311–323.

Mack, Y. P. and Müller, H. G. (1987). Adaptive nonparametric estimation of a multivariate regression function, *J. Multivariate Anal.*, **23**, 169–182.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong consistency of kernel regression estimate, *Z. Wahrsch. Verw. Gebiete*, **61**, 405–415.

Nadaraya, E. A. (1964). On estimating regression, *Theory Probab. Appl.*, **9**, 141–142.

Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*, Academic Press, New York.

Rosenblatt, M. (1969). Conditional probability density and regression estimates, *Multivariate Analysis II* (ed. P. R. Krishnaiah), 25–31, Academic Press, New York.

Rosenthal, H. P. (1970). On the subspace of $L^p (p > 2)$ spanned by the sequences of independent random variables, *Israel J. Math.*, **8**, 273–303.

Schuster, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of distinct points, *Ann. Math. Statist.*, **43**, 84–88.

Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression, *Ann. Statist.*, **8**, 240–246.

Stone, C. J. (1977). Consistent nonparametric regression, *Ann. Statist.*, **5**, 595–645.
Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates, *Ann. Statist.*, **12**, 917–926.
Tsybakov, A. B. (1987). On the choice of the bandwidth in kernel nonparametric regression, *Theory Probab. Appl.*, **32**, 142–148.
van Eeden, C. (1985). Mean integrated squared error of kernel estimators when the density and its derivatives are not necessarily continuous, *Ann. Inst. Statist. Math.*, **37**, 461–472.
Watson, G. S. (1964). Smooth regression analysis, *Sankhyā*, **26**, 359–372.
Wheeden, R. L. and Zygmund, A. (1977). *Measure and Integral*, Dekker, New York.
Zhao, L. C. and Fang, Z. (1985). Strong convergence of kernel estimates of nonparametric regression functions, *Chinese Ann. Math. Ser. B*, **6**, 147–155.