The Annals of Statistics
1993, Vol. 21, No. 2, 681–693

NEAREST NEIGHBOR REGRESSION WITH HEAVY-TAILED ERRORS

By Hari Mukerjee

Wichita State University

There has been an increasing interest in modelling regression with heavy-tailed conditional error distributions, mostly in the parametric setting. Nonparametric regression procedures have been studied almost exclusively for the cases where the conditional variance of the regressed variable is finite in the region of interest. We initiate a study of the infinite variance case. Some results in strong uniform consistency of the nearest neighbor estimator with rates are proven. The technique used provides new results and insights when higher conditional moments exist. Some asymptotic distribution theory has also been obtained when the conditional errors are in the domain of attraction of a stable law.

1. Introduction. Since the work of Mandelbrot (1960, 1963, 1969), the use of stable distributions to model data with heavy-tailed distributions has become quite popular; the paper by DuMouchel (1983) contains a substantial bibliography. In the regression context the usual procedure is to employ a parametric model with (typically symmetric) stable error distributions. We study the nearest neighbor nonparametric regression procedure when the errors have heavy tails, prove strong uniform consistency of the estimator and its a.s. convergence rates, and initiate a study of its asymptotic distribution when the errors are in the domain of attraction of a stable law.

Suppose $(X_1, Y_1), \dots, (X_n, Y_n)$ are i.i.d. $R^d \times R$-valued random variables with finite conditional means $m(x) = E[Y_1 | X_1 = x]$. Let $\| \cdot \|$ and $| \cdot |$ denote the norms in R^d and R, respectively. For $x \in R^d$ and $n \geq 1$ let $\{D_{1n}(x), \dots, D_{nn}(x)\}$ be the order statistics of $\{X_1, \dots, X_n\}$, ordered by the distances $\{\|x - X_i\|\}$, with ties being broken by the chronological order. Let $\{W_{1n}(x), \dots, W_{nn}(x)\}$ denote the corresponding induced order statistics of $\{Y_1, \dots, Y_n\}$. The nearest neighbor (NN) regression estimator of $m(x)$ is given by

$$m_n(x) = \sum_{i=1}^{n} c_{in} W_{in}(x), \tag{1.1}$$

where $\{c_{in}\colon 1 \leq i \leq n,\ n \geq 1\}$ is a fixed double sequence of real numbers. The special case of k-nearest neighbor (k-NN) estimator is defined by the weight vectors

$$c_{in} = I(i \leq k)/k, \qquad n \geq 1, \tag{1.2}$$

Received May 1989; revised May 1992.

AMS 1991 *subject classifications*. Primary 60F15; secondary 62G05, 62G15.

Key words and phrases. Nearest neighbor regression, heavy-tailed distributions, strong consistency, convergence rates, stable distributions, confidence intervals.

where $k = k_n$ is a positive integer with $k \to \infty$ and $k/n \to 0$. The NN estimator and some variants have been studied extensively, mostly under the assumption that the conditional variance $\sigma^2(x) = \mathrm{Var}[Y_1|X_1 = x] < \infty$ for all x in the region of interest [see, e.g., Fix and Hodges (1951), Royall (1966), Devroye and Wagner (1977), Stone (1977), Devroye (1978), Mack (1981) and Cheng (1984), and the references therein]. Recently Mukerjee (1989) has shown that the k-NN estimator is strongly consistent with only a first moment-type assumption when $k_n \sim n^{\eta}$, $0 < \eta < 1$. In Section 2 we show that strong uniform consistency and a.s. rate of convergence results could be derived with only a $(1 + \delta)$-order conditional moment and that the order of moment assumed, the rate of convergence, and $\{k_n\}$ form an intimately connected triplet. The proof is based on truncation and an exponential probability bound for weighted sums of independent random variables with a triangular set of weights due to Hanson and Koopmans (1965) in conjunction with a result (Lemma 2.6) based on a combinatorial inequality due to Schläfli (1901). Our results greatly generalize those obtained by Cheng (1984) under second order and higher moment assumptions (see Remark 2.11). In Section 3 we use Brunk's (1970) independent observations regression model to extend the results to the case where $\{X_i\}$ may be a (marginally) dependent sequence and apply it to the case of a stationary $\{X_i\}$. In Section 4 we initiate a study of the asymptotic distribution of the k-NN estimator assuming stable errors and using the results of Logan, Mallows, Rice and Shepp (1973). In Section 5 we make some concluding remarks.

2. Consistency. Let S be the support of X_1. For any real-valued function T on R^d we denote $\sup_{x \in K}|T(x)|$ by $\|T\|_K$, $K \subset R^d$. Now consider the following assumptions:

$$(2.1)\qquad \begin{aligned} &\text{(i)}\ \max_{1\le i\le n} |c_{in}| \le 1/f(n) \quad \text{for some } f(n) \to \infty;\\ &\text{(ii)}\ \sup_n \sum_{i=1}^n |c_{in}| \le C < \infty \quad \text{for some } C \ge 1; \end{aligned}$$

and

$$(2.2)\qquad \begin{aligned} &\text{if}\quad H(t) = \sup_{x\in S} P(|Y_1| \ge t|X_1 = x), \quad \text{then}\\ &H(t) \to 0 \quad \text{as } t \to \infty \quad \text{and} \quad \int_0^\infty t^r|dH(t)| \equiv \sigma_r < \infty \quad \text{for some } r > 1. \end{aligned}$$

For a compact $B \subset R^d$, the a.s. convergence rate of $\|m_n - m\|_B$ for the k-NN estimator has been considered in the literature only in the cases corresponding to $r > 2 + d/2$ in (2.2). In this case the optimal convergence rate does not depend on r, as shown in Remarks 2.11. However, it does for

$1 < r \le 2 + d/2$. Let $\rho = \min\{r, 2\}$ and let θ_n be a positive sequence with $\theta_n \to \infty$.

THEOREM 2.1. *Assume that* (2.1) *and* (2.2) *hold.*

(a) *If* $f(n)/(n^{1/r}\log n) \to \infty$, *then* $\|m_n - Em_n\|_S \to 0$ *a.s.*
(b) $\|m_n - Em_n\|_S = O(1/\theta_n) + O(n^{-(r-1)/r})$ *a.s.*,

where

(i) $\theta_n = f(n)/(n^{1/r}\log n)$ *if* $f(n) = O(n^{\rho/r}\log n)$ *and*
(ii) $\theta_n = \sqrt{f(n)/(n^{(2-\rho)/r}\log n)}$ *if* $n^{\rho/r}\log n = o(f(n))$.

Note that r must be more than 2 and $\rho = 2$ in Theorem 2.1b(ii) if $f(n) = O(n)$, as is the case with the k-NN estimator.

For $n \ge 1$ and $1 \le i \le n$ let $Y_{in}^* = Y_i I(|Y_i| \le n^{1/r})$ and define $W_{in}^*(x)$ correspondingly for $x \in R^d$. Define m_n^* by

$$(2.3) \qquad m_n^*(x) = \sum_{i=1}^{n} c_{in} W_{in}^*(x), \qquad x \in R^d.$$

Our proof of Theorem 2.1 will be derived from the convergence properties of the three expressions on the rhs of the inequality

$$(2.4) \quad \|m_n - Em_n\|_S \le \|m_n - m_n^*\|_S + \|m_n^* - Em_n^*\|_S + \|Em_n^* - Em_n\|_S.$$

We first investigate these in the following lemmas.

LEMMA 2.2. *Under assumptions* (2.1)(i) *and* (2.2)

$$\|m_n - m_n^*\|_S = O(1/f(n)) \quad a.s.$$

PROOF. From (2.2) we have that $E|Y_1|^r < \infty$. Thus $P(|Y_n| > n^{1/r}$ i.o.$) = 0$ and $W_{in}(x) = W_{in}^*(x)$, $x \in S$, for almost all sample sequences if $n \ge n_0$ for some n_0 depending on the particular sequence. For such a sample sequence, using (2.1)(i),

$$\begin{aligned}\|m_n - m_n^*\|_S &= \sup_{x\in S}\left|\sum_{i=1}^{n} c_{in}W_{in}(x)I\big(|W_{in}(x)| > n^{1/r}\big)\right| \\ &\le \frac{1}{f(n)}\sum_{i=1}^{n}|Y_i|I\big(|Y_i| > n^{1/r}\big) \\ &\le \frac{1}{f(n)}\sum_{i=1}^{n_0}|Y_i|I\big(|Y_i| > n_0^{1/r}\big) = O\left(\frac{1}{f(n)}\right)\end{aligned}$$

from which the lemma follows. □

LEMMA 2.3. *Under assumptions* (2.1)(ii) *and* (2.2),

$$\|Em_n^* - Em_n\|_S = O(n^{-(r-1)/r}).$$

PROOF. Under the assumptions, the proof follows from

$$\begin{aligned}
&\|E(m_n^* - m_n|X_1,\dots,X_n)\|_S\\
&\quad\le \sup_{x\in S}\sum_{i=1}^{n}|c_{in}|E\big[|W_{in}(x)|I\big(|W_{in}(x)| > n^{1/r}|X_1,\dots,X_n\big]\\
&\quad\le \sum_{i=1}^{n}|c_{in}|n^{-(r-1)/r}\sup_{x\in S}E\big[|Y_1|^r|X_1 = x\big]\\
&\quad\le C\sigma_r n^{-(r-1)/r} = O\big(n^{-(r-1)/r}\big) \quad \text{a.s.} \qquad \square
\end{aligned}$$

The following lemma is well known for $\rho = 2$.

LEMMA 2.4. *If X is a random variable with $|X| \le M < \infty$, $EX = 0$, and $E|X|^\rho = \sigma_\rho < \infty$ for some $1 < \rho \le 2$, then*

$$|Ee^{tX} - 1| \le \sigma_\rho M^{2-\rho}t^2 \quad \textit{for } |t| \le 1/M.$$

PROOF. For $|t| \le 1/M$ we have

$$\begin{aligned}
|Ee^{tX} - 1| &= \left|\sum_{k=2}^{\infty} t^k EX^k/k!\right| \le \sum_{k=2}^{\infty}|t|^k E|X|^k/k! \le \sum_{k=2}^{\infty}|t|^k M^{k-\rho}\sigma_\rho/k!\\
&\le \big(\sigma_\rho M^{2-\rho}t^2/2\big)\sum_{k=2}^{\infty}|t|^{k-2}M^{k-2}/2^{k-2}\\
&= \big(\sigma_\rho M^{2-\rho}t^2/2\big)(1 - |t|M/2)^{-1} \le \sigma_\rho M^{2-\rho}t^2. \qquad \square
\end{aligned}$$

The following lemma is a simplified version of Theorem 1 in Hanson and Koopmans (1965).

LEMMA 2.5 (Hanson and Koopmans). *Suppose $X_1,\dots,X_n$ are independent mean-zero random variables. Assume that for every $\beta > 0$ there exists $T_\beta > 0$ such that the moment generating function $\phi_k(t) = Ee^{tX_k}$ exists for $1 \le k \le n$ and $|1 - \phi_k(t)| \le \beta|t|$ for $|t| \le T_\beta$ uniformly in k. Let $\{c_{in}\}$ be a double sequence of real numbers obeying* (2.1). *Then for every $\varepsilon > 0$ and $n \ge 1$*

$$P\left\{\left|\sum_{i=1}^{n} c_{in}X_i\right| \ge \varepsilon\right\} \le 2\rho_\varepsilon^{f(n)},$$

where $\rho_\varepsilon = e^{-\varepsilon T/4}$ with $T = \min\{1, T_{\varepsilon/(2C)}, T_{\varepsilon/(2C)}/C\}$.

Let ${}_NC_{\le d} = \sum_{i=0}^{d}{}_NC_m$, where ${}_NC_m = 0$ for $m > N$. Recalling the notation ${}_NC_m$ for the number of m-element subsets of an N-element set, ${}_NC_{\le d}$ is the number of subsets with at most d elements. The following lemma uses a combinatorial inequality due to Schläfli (1901) quoted in Dudley (1978).

LEMMA 2.6. *For every positive integer n the number of distinct values of $\{m_n(x)\colon x \in R^d\}$ is less than or equal to ${}_N C_{\le d}$ with probability 1, where $N = {}_n C_2$.*

PROOF. Fix a positive integer n. Let $\{x_1, \dots, x_n\}$ be a realization of $\{X_1, \dots, X_n\}$. For $1 \le i < j \le n$ let H_{ij} be the hyperplane bisecting the line segment formed by the pair (x_i, x_j). There are $N = {}_n C_2$ such hyperplanes. Schläfli (1901) showed that the maximum number of open regions formed by N hyperplanes in R^d is ${}_N C_{\le d}$, which is attainable if these hyperplanes are "in general position"; see Dudley (1978), page 921 and the references cited therein. If x and x' are in the same open region then, for every pair (i, j) with $i < j$, we have that x and x' are on the same side of H_{ij} so that $\|x - x_i\| < (>) \|x - x_j\|$ if and only if $\|x' - x_i\| < (>) \|x' - x_j\|$, and thus $D_{\alpha n}(x) = D_{\alpha n}(x')$ for $1 \le \alpha \le n$. Hence the number of values of $m_n(x)$, as x ranges over the open regions, is bounded above by ${}_N C_{\le d}$.

For any $i < j$ let O_{ij} denote the open half-space $\{y\colon \|y - x_i\| < \|y - x_j\|\}$ defined by H_{ij}. If $x \in \bigcap H_{i_m j_m} I(1 \le m \le p)$ for some $\{(i_m < j_m)\}$ and p, and $x \notin H_{ij}$ for any other (i, j), then x is in the boundary of one of the open regions contained in $\bigcap O_{i_m j_m} I(1 \le m \le p)$, which is nonnull wp 1 since no two hyperplanes are the same wp 1. For every y in this open region and for every $1 \le m \le p$, we have $\|y - x_{i_m}\| < \|y - x_{j_m}\|$. Since $i_m < j_m$, if $x_{i_m} = D_{\alpha n}(y) = D_{\beta n}(x)$ and $x_{j_m} = D_{\gamma n}(y) = D_{\delta n}(x)$, then $\alpha < \gamma$ and $\beta < \delta$, $1 \le m \le p$, by our definition of $\{D_{(\cdot)n}(\cdot)\}$ and the tie-breaking rule. Both x and y lie on the same side of H_{ij} for all other (i, j), and thus $m_n(x) = m_n(y)$. This along with the result above completes the proof of the lemma.

The restriction "wp 1" could be omitted in the lemma above, but the proof is messier, and we have no need for it.

LEMMA 2.7. *Let $\sigma_\rho = \sup_{x \in S} E[|Y_1|^\rho | X_1 = x]$, $A = 2^{\rho+3} C^2 \sigma_\rho$, $N = {}_n C_2$, and $M_n = 2n^{1/r}$. Under assumptions* (2.1) *and* (2.2), *for every positive integer n and for every $\varepsilon > 0$,*

$$P\{\|m_n^* - Em_n^*\|_S \ge \varepsilon\} \le 2\,{}_N C_{\le d} \exp\left[-f(n)\varepsilon \min\{\varepsilon/(AM_n^{2-\rho}), 1/(4M_n C)\}\right].$$

PROOF. Fix a positive integer n and $\varepsilon > 0$. Let $Z_{in}(x) = W_{in}^*(x) - E[W_{in}^*(x)|X_1, \dots, X_n]$. Note that $|Z_{in}(x)| \le 2n^{1/r}$, $E[Z_{in}(x)|X_1, \dots, X_n] = 0$ a.s. and

$$\begin{aligned} &E\left[|Z_{in}(x)|^r | X_1, \dots, X_n\right] \\ &\quad \le 2^{r-1}\left\{E\left[|W_{in}^*(x)|^r | X_1, \dots, X_n\right] + \left(E\left[|W_{in}^*| | X_1, \dots, X_n\right]\right)^r\right\} \\ &\quad \le 2^r E\left[|W_{in}^*|^r | X_1, \dots, X_n\right] \le 2^r \sup_{x \in S} E\left[|Y_1|^r | X_1 = x\right] \le 2^r \sigma_r < \infty \quad \text{a.s.}, \end{aligned}$$

by standard inequalities and (2.2). Also note that this series of inequalities hold with r replaced by ρ. Applying Lemma 2.4 to $Z_{in}(x)$ we have

$$|E[\exp(tZ_{in}(x))|X_1, \dots, X_n] - 1| \le 2^\rho \sigma_\rho M_n^{2-\rho} t^2 \quad \text{a.s.,}$$
$$|t| \le 1/M_n, 1 \le i \le n, x \in S.$$

For any $\beta > 0$ let $T_\beta = \min\{\beta/(2^\rho \sigma_\rho M_n^{2-\rho}), 1/M_n\}$. Then $2^\rho \sigma_\rho M_n^{2-\rho} t^2 \le \beta|t|$ for $|t| \le T_\beta$. Since $C \ge 1$ and $1/M_n < 1$, we have that $T \equiv \min\{1, T_{\varepsilon/(2C)}, T_{\varepsilon/(2C)}/C\} = T_{\varepsilon/(2C)}/C$. Applying Lemma 2.5 to $\{Z_{in}(x)\}$, for every $x \in S$,

$$P\left\{\left|\sum_{i=1}^n c_{in} Z_{in}(x)\right| \ge \varepsilon | X_1, \dots, X_n\right\} \le 2\rho_\varepsilon^{f(n)} \quad \text{a.s.,}$$

where

$$\rho_\varepsilon = \exp[-\varepsilon T/4] = \exp[-\varepsilon T_{\varepsilon/2C}/(4C)]$$
$$= \exp\left[-\varepsilon \min\left\{\varepsilon/\left(2^{\rho+3} C^2 \sigma_\rho M_n^{2-\rho}\right), 1/(4M_n C)\right\}\right].$$

The proof of the lemma follows by noting that Lemma 2.6 implies

$$P\{\|m_n^* - E[m_n^*|X_1, \dots, X_n]\|_S \ge \varepsilon | X_1, \dots, X_n\}$$
$$\le {}_N C_{\le d} \sup_{x \in S} P\{|m_n^*(x) - E[m_n^*(x)|X_1, \dots, X_n]| \ge \varepsilon | X_1, \dots, X_n\}$$
$$\le {}_N C_{\le d} \sup_{x \in S} P\left\{\left|\sum_{i=1}^n c_{in} Z_{in}(x)\right| \ge \varepsilon | X_1, \dots, X_n\right\} \quad \text{a.s.} \qquad \square$$

Proof of Theorem 2.1. Let $\varepsilon > 0$ be arbitrary in Lemma 2.7. Since $M_n = 2n^{1/r}$ and $\rho > 1$ we have $1/M_n = o(M_n^{2-\rho})$. Thus the rhs of the probability inequality in the lemma becomes

$$2\,{}_N C_{\le d} \exp[-\varepsilon f(n)/(4M_n C)] = 2\,{}_N C_{\le d} \exp[-\beta_n \log n]$$

for some $\beta_n \to \infty$, and hence is summable if $f(n)/(n^{1/r} \log n) \to \infty$. Thus $\|m_n^* - Em_n^*\|_S \to 0$ a.s. under this assumption. The proof Theorem 2.1(a) is completed by using (2.4), Lemma 2.2 and Lemma 2.3.

Using (2.4) and Lemmas 2.2 and 2.3, and the fact that $\theta_n = o(f(n))$ under both sets of conditions in Theorem 2.1(b), it is sufficient to show that $\|m_n^* - Em_n^*\|_S = O(1/\theta_n)$ a.s. to prove Theorem 2.1(b). Let $\varepsilon = K/\theta_n$ in Lemma 2.7 for some positive constant K, not depending on n, and let $e_n = [f(n)/\theta_n] \min\{1/(\theta_n n^{(2-\rho)/r}), 1/n^{1/r}\}$. The rhs of the probability inequality in Lemma 2.7 is summable if $e_n = \log n$ and K is sufficiently large. When $\theta_n n^{(2-\rho)/r} = O(n^{1/r})$, that is, $\theta_n n^{1/r} = O(n^{\rho/r})$, and $e_n = \log n$, we have $\log n = e_n = f(n)/(\theta_n n^{1/r})$ so that $\theta_n = f(n)/(n^{1/r} \log n)$ and $f(n) = \theta_n n^{1/r} \log n = O(n^{\rho/r} \log n)$, proving Theorem 2.1(b)(i). When $n^{\rho/r} = o(\theta_n n^{1/r})$ and $e_n = \log n$, we have $\log n = e_n = f(n)/(\theta_n^2 n^{(2-\rho)/r})$ so that $\theta_n^2 = f(n)/(n^{(2-\rho)/r} \log n)$ and $n^{\rho/r} \log n = f(n) n^{2\rho/r}/(\theta_n^2 n^{2/r}) = o(f(n))$, proving Theorem 2.1(b)(ii). $\square$

Now assume that

$$\liminf_n \sum_{i=1}^{n} |c_{in}| > 0. \tag{2.5}$$

Then $\max_{1\le i\le n}|c_{in}| \ge c/n$ for all n sufficiently large and some $c > 0$, which implies that $f(n) = O(n)$. Since $\theta_n = O(n^{(\rho-1)/r}) = O(n^{(r-1)/r})$ for all $r > 1$ under the conditions (i) and $\theta_n = o(\sqrt{f(n)}) = o(\sqrt{n}) = o(n^{(r-1)/r})$ under the conditions (ii) where $r > 2$, the rate result may be stated as

$$\theta_n \| m_n - Em_n \|_S = O(1) \quad \text{a.s.} \tag{2.6}$$

under the additional assumption (2.5).

We now consider the uniform convergence of m_n to m on B, a compact subset of S. Consider the following assumptions:

(2.7) there exists a positive sequence $\{d(n)\}$ such that $d(n) \to \infty$, $d(n)/n \to 0$, $|\Sigma_{i=1}^{n} c_{in} - 1| \to 0$ and $\Sigma_{i=d(n)+1}^{n} |c_{in}| \to 0$

and

(2.8) m is bounded on S and is uniformly continuous on the intersection of S and an open neighborhood of B.

THEOREM 2.8. *Under the assumptions* (2.1), (2.2), (2.7) *and* (2.8),

$$\| m_n - m \|_B \to 0 \quad a.s.\ iff(n)/(n^{1/r}\log n) \to \infty.$$

PROOF. Using Theorem 2.1 it is sufficient to show that $\| Em_n - m \|_B \to 0$. Now

$$|Em_n(x) - m(x)| \le E\left| \sum_{i=1}^{d(n)} c_{in} m(D_{in}(x)) - m(x) \right| + \sum_{i=d(n)+1}^{n} |c_{in}| E\big[|m(D_{in}(x))| + |m(x)|\big].$$

For $x \in R^d$ let $R_n(x) = \| x - D_{d(n),n}(x) \|$. Then $\| R_n \|_B \to 0$ a.s. from a result in Devroye (1978). Using this, (2.7) and (2.8),

$$\begin{aligned} &\left| \sum_{i=1}^{d(n)} c_{in} m(D_{in}(x)) - m(x) \right| \\ &\quad = \left| \sum_{i=1}^{d(n)} c_{in} m(x)(1 + o_u(1)) - m(x) \right| \\ &\quad \le \left| \left(\sum_{i=1}^{d(n)} c_{in} - 1 \right) m(x) \right| + \sum_{i=1}^{d(n)} |c_{in}| |m(x)| o_u(1) = o_u(1) \quad \text{a.s.,} \end{aligned}$$

where $o_u(1) \to 0$ as $n \to \infty$ uniformly in $x \in B$. Using (2.7) and the boundedness of m by (2.8),

$$\sum_{i=d(n)+1}^{n} |c_{in}|\left[|m(D_{in}(x))| + |m(x)|\right] \to 0 \quad \text{a.s. uniformly in } x \in B.$$

Thus $\|Em_n - m\|_B \to 0$ which completes the proof of the theorem. □

To obtain an a.s. convergence rate for $\|m_n - m\|_B$ we consider only the k-NN case and note that (1.2) implies (2.1) and (2.5) with $f(n) = k$ and $C = 1$. Now consider the following assumptions:

(2.9) the marginal distribution of X_1 has a bounded density g w.r.t. the Lebesgue measure on R^d

and

(2.10) m and g are continuously differentiable up to second order in the intersection of S and an open neighborhood of B.

The following result follows from Theorem 1 in Mack (1981).

Lemma 2.9 (Mack). *Suppose that* (1.2), (2.9) *and* (2.10) *hold and that* $k = o(n)$, $\log n = o(k)$, $\inf_{x \in B} g(x) > 0$, $P(\|x - X_1\| > s) = O(s^{-t})$ *for some* $t > 0$ *as* $s \to \infty$. *Then*

$$\|Em_n - m\|_B = O\left((k/n)^{2/d}\right) + O(1/k).$$

Theorem 2.10. *Assume that the conditions of Theorem* 2.1(b) *and Lemma* 2.10 *hold for the k-NN estimator. Then*

$$\|m_n - m\|_B = O(1/\theta_n) + O\left((k/n)^{2/d}\right) \quad a.s.$$

with θ_n *as in Theorem* 2.1.

Proof. The conditions imply the rate result in Theorem 2.1(b) in the form of (2.6). The proof is completed by using Lemma 2.9 and the fact that $\theta_n = o(k)$. □

Remarks 2.11. (i) Theorem 2.1 clearly displays the interrelationships of the doublet $(r, f(n))$ in part (a) and of the triplet $(r, f(n), \theta_n)$ in part (b), and they hold for all $r > 1$. In the k-NN case condition (b)(ii) cannot be satisfied if $r \le 2$, and for $r > 2$ the result can be stated as $\theta_n = \sqrt{k/\log n}$ if $n^{2/r} \log n = o(k)$. For $r > 2$ both conditions may be used depending on the choice of k.

(ii) Under the conditions of Theorem 2.10 the best order of convergence of $\|m_n - m\|_B$ is obtained when θ_n is of the same order of magnitude as $(n/k)^{2/d}$. For $r \le 2 + d/2$, using (b)(i) of Theorem 2.1, this gives a convergence rate of $(\log n/n^{(r-1)/r})^{2/(2+d)}$ corresponding to $k^{2+d} \sim n^{(2r+d)/r}(\log n)^d$; note that $r \le 2 + d/2$ if and only if $(2r + d)/[r(2 + d)] \le 2/r$ so that condition (b)(ii) cannot be satisfied for this k. If we want to use (b)(ii) we need k, and hence $(k/n)^{2/d}$, of a larger order of magnitude, and thus the rate given above is optimal. For $r > 2 + d/2$, using (b)(ii), the best rate is $(\log n/n)^{2/(4+d)}$ corresponding to $k^{4+d} \sim n^4(\log n)^d$ [note that $r > 2 + d/2$ if and only if $4/(4+d) > 2/r$ and thus (b)(ii) applies]. If we use a (smaller) $k = O(n^{p/r}\log n) = O(n^{2/r}\log n)$ to apply (b)(i), then $\theta_n = O(n^{1/r}) = o(n^{2/(4+d)})$ when $r > 2 + d/2$, and thus the rate given above is optimal. It is interesting to note that this optimal convergence rate, which is independent of r for $r > 2 + d/2$, differs from that for $r = 2 + d/2$ by the factor $(\log n)^{(2+d)/(4+d)}$. Thus the "power of n" factor in the convergence rate is a continuous function of r for all $r > 1$.

(iii) The best known results for the k-NN case are the two results of Cheng's (1984) that correspond to two isolated cases of our theorem:

(a) $r = 2$ and $k/(\sqrt{n}\log n) \to \infty$ for uniform consistency without rates.

(b) $r = d + 2$ and $k = [Dn^{2/(2+d)}]$ for some $D > 0$ for a uniform convergence rate of $\|m_n - m\|_B$ equal to $\beta_n \log n/n^{1/(2+d)}$ for some arbitrary $\beta_n \to \infty$.

Part (a) follows from our Theorem 2.8. In part (b) Cheng gets $\theta_n = n^{1/(2+d)}/\log n$, the same that we get from our Theorem 2.1 using (b)(i); the extra term β_n comes from higher bias (which dominates) because Cheng does not assume differentiability of the density of X_1. Using the conditions of Lemma 2.9 the rate will be $1/\theta_n$ with Cheng's r and k, but, even with $r = 2 + d/2$, the optimal rate would be $(\log n)^{2/(2+d)}n^{-2/(4+d)}$ as shown above.

(iv) For kernel regression with $d = 1$, $r > 5/2$, $b_n = n^{-\alpha}$ where $1/5 < \alpha < 1 - 2/r$, Mack and Silverman (1982) obtained the rate $[\log n/(nb_n)]^{1/2}$, where b_n is the bandwidth. Since $1 - \alpha < 2/r$, we have $n^{2/r}\log n = o(nb_n)$, and the rate agrees with the optimal rate given above if we identify nb_n with k.

3. An independent observations regression model. We note that Theorem 2.1 does not depend on d or the marginal distribution of X_1; they come into the picture only in the convergence properties of Em_n to m. The results in Section 2 can be immensely generalized using the independent observations regression model (IORM) due to Brunk (1970) and applying it to regression models where $\{X_i\}$ may be fixed or a (marginally) dependent sequence. In this model F_x is a d.f. for each $x \in R^d$ with mean $m(x)$, $\{X_n\}$ is an R^d-valued stochastic process, $\{Y_n\}$ is a real-valued stochastic process, and, conditional on $\{X_n\} = \{x_n\}$, $\{Y_n\}$ is an independent sequence with Y_n distributed as F_{x_n}. We show an application to a case where $\{X_n\}_{n=-\infty}^{\infty}$ is (marginally) a one-dimensional strictly stationary sequence, while we observe only $\{X_1, X_2, \dots\}$. Let S be the support of X_1, G its d.f., and let G_n denote the e.d.f. generated by $\{X_1, \dots, X_n\}$. Suppose

$$\text{for each real } t, \quad E[I(X_n \le t)| \cdots X_{-1}, X_0] \to G(t) \quad \text{uniformly in almost all sample sequences.} \tag{3.1}$$

Then Bhattacharya (1972) showed that for every $\varepsilon > 0$ there exists $0 < k(\varepsilon) < \infty$ such that

$$P(\|G_n - G\|_S > \varepsilon) \le \exp(-n/k(\varepsilon)). \tag{3.2}$$

THEOREM 3.1. *Assume the IORM with the distribution of* $\{X_n\}$ *as described above. Then under assumptions* (2.1), (2.2), (2.7), (2.8) *and* (3.1)

$$\|m_n - m\|_B \to 0 \quad a.s. \text{ if } f(n)/n^{1/r}\log n \to \infty.$$

PROOF. By Theorem 2.1 we have $\|m_n - Em_n\|_B \to 0$ a.s. under the assumptions. In proving $\|Em_n - m\|_B \to 0$ in Theorem 2.9 the (marginal) independence of $\{X_n\}$ was used only to prove that $\|b_n\|_B = \sup_{x\in S}\|x - D_{d(n),n}(x)\| \to 0$ a.s.; Devroye (1978) proved this using the compactness of B and Hoeffding's inequality. The same proof goes through in our case using the inequality (3.2) instead. □

4. Asymptotic distribution. Besides the point estimator $m_n(x)$ of $m(x)$, we need to know something about the distribution of (properly normalized) $m_n(x) - m(x)$ in order to do statistical inference about $m(x)$. We consider the k-NN case only and assume that the conditional distribution of $Y_1 - m(X_1)$ given $X_1 = x$ is in the domain of attraction of a stable law with exponent α with $1 < \alpha < 2$ and that it is the same for all $x \in S$ (i.e., the conditional errors are i.i.d.). Let $\{Z_i\}$ be an i.i.d. sequence with the same distribution, independent of $\{X_i\}$. Let h be the density of the stable distribution satisfying

$$x^{\alpha+1}h(x) \to p \quad \text{and} \quad x^{\alpha+1}h(-x) \to q. \tag{4.1}$$

Logan, Mallows, Rice and Shepp (1973) have shown that the Student statistic

$$T_n = \sqrt{n}\sum_{i=1}^n Z_i \Big/ \sqrt{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2} \to_{\mathscr{L}} T, \tag{4.2}$$

where T has a two-parameter density $f_{\alpha,p/q}$ that they express in a computable form. They also point out that the self-normalized sum

$$S_n = \frac{1}{n^{1/\alpha}}\sum_{i=1}^n Z_i \Big/ \sqrt{\frac{1}{n^{2/\alpha}}\sum_{i=1}^n Z_i^2} \equiv U_n/V_n \to_{\mathscr{L}} T \quad \text{also}, \tag{4.3}$$

and they derive the limiting distribution from the joint limiting distribution of (U_n, V_n) after noting that it is the same whether the Z_i's are only in the domain of attraction of the stable law in question or they actually have the distribution of the stable law. Using this result we can prove the following:

THEOREM 4.1. *Fix* $x \in S$. *Consider the k-NN estimator under the distributional assumptions described above. If the conditions of Lemma* 2.10 *hold and* $k^{1-d/2\alpha+d/2} = o(n)$ *then*

$$T_n(x) = k[m_n(x) - m(x)] \Big/ \sqrt{\sum_{i=1}^k [W_{in}(x) - m_n(x)]^2} \to_{\mathscr{L}} T. \tag{4.4}$$

PROOF. The numerator of $T_n(x)$ may be written as

$$\sum_{i=1}^{k} \{[W_{in}(x) - m(D_{in}(x))] + [m(D_{in}(x)) - m(x)]\}$$

$$=_d \sum_{i=1}^{k} \{Z_i + [m(D_{in}(x)) - m(x)]\}.$$

From assumption (2.10) we have $m(D_{in}(x)) - m(x) = m'(x)(D_{in}(x) - x)(1 + o(1))$. Under the assumptions of Lemma (2.10) Mack (1981) shows that, with $R_n(x)$ = the distance of x from its kth NN among the X_i's and c = the volume of the unit ball in R^n,

$$R_n(x) = O_p\big((k/n)^{2/d}\big) \quad \text{and} \quad \frac{c}{k}\sum_{i=1}^{k} \frac{x - D_{in}(x)}{2R_n(x)} \to 1 \quad \text{a.s.},$$

so that

$$\text{(4.5)} \quad \frac{1}{k^{1/\alpha}} \sum_{i=1}^{k} \{m(D_{in}(x)) - m(x)\} = o_p(1) \quad \text{if } k^{1-1/\alpha+2/d} = o(n^{2/d}).$$

Let $\overline{m}_n(x) = (1/k)\Sigma_{i=1}^{k} m(D_{in}(x))$ and $\overline{Z} = (1/k)\Sigma_{i=1}^{k} Z_i$. Then we could write

$$\sum_{i=1}^{k} [W_{in}(x) - m_n(x)]^2 =_d \sum_{i=1}^{k} \{(Z_i - \overline{Z}) + [m(D_{in}(x)) - \overline{m}_n(x)]\}^2.$$

Since $R_n(x) = o_p(1)$, assumption (2.10) implies that

$$\text{(4.6)} \qquad m(D_{in}(x)) - \overline{m}_n(x) = o_p(1).$$

The proof of the theorem is completed using (4.3) and the estimates (4.5) and (4.6) along with our assumption on k. □

It may be noted that if $k = [n^\delta]$, then for $d = 1$ and in the limit as $\alpha \to 2$, the above condition on k implies that $\delta < 4/5$; the usual condition for asymptotic mean-zero normality under the assumption of finite conditional variance.

5. Concluding remarks.

We have derived the strong convergence properties of the NN estimators with infinite conditional variances and have shown that the rate of a.s. convergence, the choice of $\{k_n\}$, and the order of moment of the conditional errors form an interrelated triplet. This greatly generalizes existing results even for the finite variance case. However, the asymptotic distribution theory has been derived ,only under the assumption of i.i.d. conditional errors. Generalization to nonidentically distributed errors (but still in the domain of attraction of the same stable law) will be useful. Further generalization involving errors attracted to stable laws with a range of values of the exponent may be possible using Tucker's (1968) results on convolutions of such variables.

Our asymptotic distribution depends on two parameters that are typically unknown. One has to guess their values (or a range of values) to construct approximate confidence intervals. For some data in economics the variables of interest are positive so that the ratio p/q may be taken to be infinity.

An alternative to the use of the distribution theory above is to use the results of Benjamini's (1983) that provide some justifications to the folklore that confidence intervals based on the assumption that T_n in (4.2) has a t-distribution with $n - 1$ degrees of freedom are conservative when the Z_i's have heavy tails. In particular, he shows, analytically and by simulations, that for scale mixtures of the standard normal, a family that includes the symmetric stables, the confidence intervals are indeed conservative if the critical value t_c is more than 1.8 for all reasonable sample sizes. To use this in our problem we must make the additional assumption of symmetry, but we can broaden the family of distributions. It is not clear what to do with this procedure if the distributions are asymmetric. Moreover, the problem of the nonidentically distributed case needs to be worked on as in the procedure suggested above.

Acknowledgment. The author is grateful to a referee whose critical comments vastly improved the presentation and interpretation of the main results.

REFERENCES

BENJAMINI, Y. (1983). Is the t test really conservative when the parent distribution is long-tailed? *J. Amer. Statist. Assoc.* **78** 645–654.

BHATTACHARYA, P. K. (1972). Probabilities of large deviations of sums of random variables. *Sankhyā Ser. A* **34** 9–16.

BRUNK, H. D. (1970). Estimation of isotonic regression. In *Nonparametric Techniques in Statistical Inference* (M. L. Puri, ed.) 177–195. Cambridge Univ. Press.

CHENG, P. E. (1984). Strong consistency of nearest neighbor regression function estimators. *J. Multivariate Anal.* **15** 63–72.

DEVROYE, L. P. (1978). The uniform convergence of nearest neighbor regression function estimates and their application in optimization. *IEEE Trans. Inform. Theory* **24** 142–151.

DEVROYE, L. P. and WAGNER, T. J. (1977). The strong uniform consistency of nearest neighbor density estimates. *Ann. Statist.* **5** 536–540.

DUDLEY, R. M. (1978). Central limit theorem for empirical measures. *Ann. Probab.* **6** 899–929.

DUMOUCHEL, W. H. (1983). Estimating the stable index α in order to measure tail thickness: A critique. *Ann. Statist.* **11** 1019–1031.

FIX, E. and HODGES, J. L. (1951). Discrimination analysis, Nonparametric discrimination: Consistency properties. Technical report 4, Project 21-49-004, USAF School of Aviation Medicine, Random Field, Texas.

HANSON, D. L. and KOOPMANS, L. H. (1965). On the convergence rate of the law of large numbers for linear combinations of independent random variables. *Ann. Math. Statist.* **36** 559–564.

LOGAN, B., MALLOWS, C., RICE, S. and SHEPP, L. (1973). Limit distributions of self-normalized sums. *Ann. Probab.* **1** 788–809.

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

MACK, Y. P. and SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61** 405–415.

MANDELBROT, B. (1960). The Pareto–Lévy law and the distribution of income. *Internat. Econom. Rev.* **1** 79–106.

MANDELBROT, B. (1963). The variation of certain speculative prices. *Journal of Business of the University of Chicago* **26** 394–419.

MANDELBROT, B. (1969). Long-run linearity, locally Gaussian processes, *H*-spectra, and infinite variances. *Internat. Econom. Rev.* **10** 82–111.

MUKERJEE, H. (1989). A strong law of large numbers for nonparametric regression. *J. Multivariate Anal.* **30** 17–26.

ROYALL, R. M. (1966). A class of nonparametric estimates of a smooth regression function. Ph.D. dissertation, Stanford Univ.

SCHLÄFLI, L. (1901, posth.). Theorie der vielfachen Kontinuität. In *Gesammelte Mathematische Abhandlungen I*. Teubner, Leipzig. (Basel, Birkhäuser, 1950).

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

TUCKER, H. G. (1968). Convolutions of distributions attracted to stable laws. *Ann. Math. Statist.* **39** 1381–1390.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
1845 FAIRMOUNT
WICHITA STATE UNIVERSITY
WICHITA, KANSAS 57208-1595