Ann. Inst. Statist. Math.
Vol. 43, No. 1, 115–129 (1991)

ERROR INFERENCE FOR NONPARAMETRIC REGRESSION*

B. RUTHERFORD[1] AND S. YAKOWITZ[2]

[1] *The Reliability Department, Sandia Laboratories, Albuquerque, NM 87112, U.S.A.*
[2] *Systems and Industrial Engineering Department, University of Arizona,*
Tucson, AZ 85721, U.S.A.

(Received August 29, 1988; revised October 2, 1989)

Abstract. This study examines means for inferring the distribution of the error in nonparametric regression. The central objective is to develop confidence intervals for nonparametric regression. Our computational study would seem to affirm that our methods are potentially useful in cases of small sample size or heterogeneously distributed error. Theoretical developments offer sufficient conditions for asymptotic normality.

Key words and phrases: Confidence intervals, bootstrapping, asymptotic normality, error inference.

1. Introduction

Let $\{(X(i), Y(i)): i = 1, 2, \ldots\}$ denote a sequence of i.i.d. pairs, where the X's are in $\mathbb{R}^d$ and the Y's are real.

We presume a regression function

$$m(x) = E[Y \mid X = x]$$

exists but, aside from some smoothness properties, has unknown structure.

Currently, the primary emphasis of the nonparametric regression literature is on providing estimators $m_n(x)$ of $m(x)$, the subscript indicating that $m_n(x)$ is constructed from the first n terms of the data sequence. The thrust of the present study is in a different direction: a nonparametric regression construct $m_n(x)$ having been specified, our wish is to infer the distribution of $m_n(x) - m(x)$.

A central motivation for inference of the error law was to construct confidence intervals for $m(x)$. Another potential use is for hypothesis testing. Bhattacharya (1976) has given a definitive method for testing whether the regression function is constant. Our methodology, while not as focussed, could be adapted to test whether $m(x)$ is positive, for instance. Another potential use is for estimating

* This work was undertaken while Dr. Rutherford was with the University of Arizona. It was supported in part by NSF grant DPP 82-19439.

error rates and quantifying sub-optimality for nonparametric regression estimators serving as discrimination functions (as in Devroye and Wagner (1980), for instance).

An idea suggested by a reviewer is that error inference algorithms could provide an approach to local bandwidth selection. As we will see in Section 2, the kernel regression method depends on a real parameter "b" as well as the data. Recently, intense investigation has been dedicated to finding data-driven selections of this bandwidth parameter which are asymptotically optimal in a least-squares sense (e.g., Härdle and Marron (1985)). Müller and Stadtmüller (1987) and Vieu (1988) have found estimates of the bandwidth which are conditionally (on domain point x) optimal in various senses. Our developments could potentially pave the way for devising bandwidth estimators that asymptotically yield a minimum-length confidence interval. The idea would be to evaluate confidence intervals over a range of bandwidth values and then select the value b for which this is shortest. In this study, we have concentrated on the variability error; one could enhance our developments by accounting for bias error, as in Härdle and Bowman (1988), Müller and Stadtmüller (1987) and elsewhere, through approximation of derivatives of $m(\ )$. (Bias analysis is needed in the bandwidth selection application.)

Dikta (1988) has undertaken a parallel investigation for the nearest neighbor regression method. His analysis follows a different tack, being based on empirical distribution approximation; the experimental findings are in accord with ours.

In the following sections, we state our confidence band procedure and examine it through simulation experiments and asymptotic analysis. A ground rule is that like $m_n(x)$ itself, our error distribution estimator must be nonparametric and data-driven. This section closes with mention of related topics.

1.1 *Classical regression*

In classical Gaussian linear regression, the framework for error estimation is already in place. Working and Hotelling (1929) give an illuminating analysis of a linear regression problem. These notions extend to "trend surfaces"(e.g., Ripley (1981), Section 4.1) where $m(x)$ is assumed to be any regression function which is a linear combination of known functions.

Recent outgrowths of classical multivariable theory are even closer in spirit to the present study. Thus, in principle, kriging methods (e.g., Ripley (1981)) admit analysis of error distribution. One confronts two drawbacks here: (i) the normality assumption, and (ii) the need to infer a covariance function or variogram. Knafl *et al.* (1985) follow this avenue even further in providing error bands of a given level over the entire regression function domain.

1.2 *Bootstrapping*

Efron and Tibshirani ((1986), Section 8) have mentioned bootstrapping in the context of nonparametric regression. They sketch their plan for bootstrapping in the nonparametric regression setting with extreme brevity, but, as we interpret it, the idea is to compute replications of the nonparametric estimator $m_n^*(x)$, this estimator being obtained from a random sample of size M from the set of observed pairs $\{(X(i), Y(i))\colon\ 1 \leq i \leq n\}$. Then confidence intervals could be assessed from these replicated values. Härdle ((1987), Chapter 4) follows out this idea and gives an example. Also, he has supplied quantile estimates, one based very directly on bootstrapping and another on use of extreme value theory. Härdle and Bowman (1988) is also closely related to the present study. Their application is to the fixed design-point regression problem, and the attack differs from ours. But this work represents a worthy investigation that presumably could be adapted to the present regression problem. Our aim was to condition the estimator on the design values $X(i)$; we think our scheme may have a slight advantage in this regard.

The plan to be described can be viewed in a bootstrapping context. Rutherford (1986) uses Edgeworth expansion arguments very close to those in the seminal bootstrapping study by Singh (1981) to analyze convergence rates of the distribution estimates.

2. Procedures for inferring the error distribution

The present study will concern the kernel nonparametric regression (NPR) function. A domain point x of interest, a sequence $\{b(n)\}$ of positive numbers, and a "kernel" pdf $k(\ )$ having been specified, the *kernel* NPR function for the data $\{(X(i),\, Y(i))\colon\ i = 1, \dots, n\}$ is defined by

$$m_n(x) = \sum Y(i)B(i;\ x,\, n). \tag{2.1}$$

Here and elsewhere, unless otherwise stated, the summation is from 1 *to* n and the weights $B(i;\ x,\, n)$ are

$$B(i;\ x,\, n) = k((x - X(i))/b(n)) \Big/ \left[\sum k((x - X(j))/b(n))\right]. \tag{2.2}$$

Usually, we will suppress the dependency on x and n by writing simply $B(i)$.

Our objective is to infer the distribution of $m_n(x)$, conditioned on the domain points $\boldsymbol{X}(n) = \{X(1), \dots, X(n)\}$. The plan we follow is to define the random variable

$$m_n^*(x) = \sum Y^*(i)B(i) \tag{2.3}$$

where the $B(i)$'s are those determined by (2.2) and the $Y^*(i)$'s are random variables determined by data-driven conditional density estimates $g_n(y \mid X(i))$ for the pdf $g(y \mid X(i))$ of the random variable $Y \mid X(i)$. In our computational studies, we used kernel methods to infer these requisite densities. Thus, specifically, we adopted approximations of the form

$$g_n(y \mid x') = q_n(x',\, y)/f_n(x'), \tag{2.4}$$

with

$$q_n(x',\, y) = 1/(na(n)c(n)^d) \sum k((x' - X(i))/c(n))k_A((y - Y(i))/a(n)) \tag{2.5a}$$

and

$$f_n(x') = 1/(nc(n)^d)\sum k((x' - X(i))/c(n)). \tag{2.5b}$$

Here $a(n)$ and $c(n)$ are positive numbers, $k(x)$ is a pdf on $\mathbb{R}^d$, and $k_A(y)$ is a pdf on $\mathbb{R}$.. Once the estimates $g_n(y \mid x')$ are at hand, in view of (2.3), our estimate of the pdf of $m_n(x)$ is given by

$$h_n(y) = (\pi_1 * \pi_2 * \cdots * \pi_n)(y). \tag{2.6}$$

Here "$*$" denotes convolution, and we define $\pi_j(y)$ by

$$\pi_j(y) = (1/B(j))g_n(y/B(j) \mid x(j)).$$

3. Computational studies

Two categories of confidence interval experiments were undertaken. The first category investigates behavior in the case of homoscedastic samples, a setting which is favorable to the alternative normal approximation. The second examines performance in the case that variance depends on the domain values.

3.1 *A common structure of the experiments*

The sample points $X(i)$ were selected uniformly from [0, 10].

The average confidence interval width and percent of coverage were obtained by averaging at domain points

$$t(i) = i/2, \qquad 2 \leq i \leq 18.$$

The unit intervals at the extremities of the domain were excluded because well-known end effects might have a complicating influence.

An Epanechnikov kernel function was selected for $k(\ )$ in (2.5a) and (2.5b). It has some appealing properties (e.g., Härdle (1987), Section 4.5), and is defined by

$$k(x) = \begin{cases} 15/16(1-x^2)^2, & \mid x \mid < 1, \\ 0, & \text{otherwise.} \end{cases}$$

The function $k_A(x, y)$ in (2.5a) was taken to be the product kernel $k(x)k(y)$, with $k(\ )$ as just stated. The bandwidth $b(n)$ was chosen to minimize the width of the 80% confidence interval.

The parameters $a(n)$ and $c(n)$ for the conditional density function were found using the Kullback-Leibler method (e.g., Marron (1987)) of cross-validation. For example, $c(n)$ in (2.5a) and (2.5b) is the value which maximizes

$$\sum \ln(f_{-i}(x(i);\ c(n))),$$

with $f_{-i}(x(i))$ being computed from (2.5b) but with the datum $x(i)$ omitted.

The convolution operation (2.6) was approximated by quadratures. The accuracy was confirmed by comparing different levels of discretization in the quadrature formulas, and also by comparison with simulations. In these computations, we noted that simulations were accurate enough to justify dispensing with convolution calculations to reduce processing time.

For purposes of comparison with our kernel method, we also computed intervals based on the asymptotic normal limiting law. Thus from Schuster (1972) it is known that

$$\sigma(n)(m_n(x) - E[m_n(x)]) \tag{3.1}$$

has the standard normal variable as its limit; here we have defined

$$\sigma(n)^2 = nb(n)\sigma(Y \mid x)^2/f(x) \int k^2(x)dx + o(nb(n)). \tag{3.2}$$

Here $\sigma(Y \mid x)^2$ is defined to be the variance of $Y \mid X = x$. In our computations, we estimated this by

$$\hat{\sigma}(Y \mid x)^2 = 1/n \cdot \left(\sum (Y(i) - m_n(x(i)))^2\right).$$

This estimate is sensible and consistent if the variance of $Y \mid x$ does not depend on x; that is, if the error were homoscedastic with respect to x.

In the tables to follow, "Asymp. Normal", is an abbreviation for "Asymptotic Normal". The associated entries were obtained by pretending that regardless of n, the distribution of $m_n(x) - m(x)$ is, in fact, the normal law with zero mean and standard deviation $\sigma(n)$ as determined by (3.2).

The regression function is

$$m(x) = ((x-2)^2 - 3)/20.$$

Three different distributions for the observations error were investigated, namely normal, exponential and uniform.

3.2 *Experiment* 1: *homoscedastic data*

The intention of this first study is to examine performance when the sample size is small; this is the realm for which our methodology is intended. Specifically, the sample size n is 20. Each entry in Table 1 results from 100 replications.

Table 1. Average confidence interval width and percent coverage sample size = 20.

Distribution of Y	Variance of $Y = 1.0$		Variance of $Y = 0.5$	
	Average width	Percent coverage	Average width	Percent coverage
GAUSSIAN				
Kernel Method	1.18	84	0.87	79
Asymp. Normal	1.15	79	0.82	77
UNIFORM				
Kernel Method	1.06	82	0.86	78
Asymp. Normal	1.16	80	0.82	72
EXPONENTIAL				
Kernel Method	0.99	82	0.82	75
Asymp. Normal	1.05	77	0.79	74

Figure 1 shows the results of a single replication in the uniform variate case. One sees here the true and nonparametric regression functions, and the loci of the endpoints of the two types of confidence intervals, as well as the 20 data points. Figure 2 is the conditional density function $g_{20}(y \mid x)$ at the domain point $x = 6.5$.

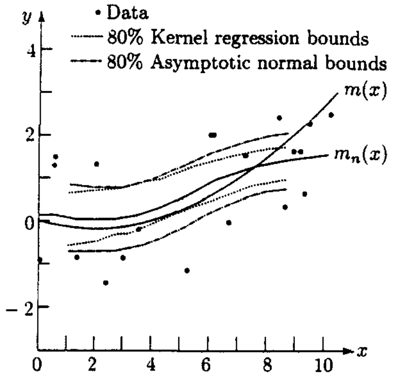

Fig. 1. A confidence interval estimation experiment.

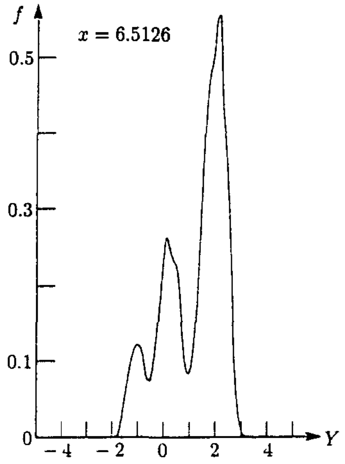

Fig. 2. Nonparametric conditional density estimator of Y.

Table 2 was constructed in a similar manner to Table 1, the difference being that here n is 40. In all cases, the confidence interval was constructed so as to have containment probability of 80%. The superior performance for the small sample case of the kernel method is clear in Table 1. In Table 2, the improvement is marginal, reflecting that with more samples, the asymptotic approximation becomes more justifiable.

Table 2. Average confidence interval width and percent coverage sample size = 40.

Distribution of Y	Variance of $Y = 1.0$		Variance of $Y = 0.5$	
	Average width	Percent coverage	Average width	Percent coverage
GAUSSIAN				
Kernel Method	0.82	78	0.65	80
Asymp. Normal	0.88	77	0.59	73
UNIFORM				
Kernel Method	0.78	79	0.65	81
Asymp. Normal	1.16	80	0.64	80
EXPONENTIAL				
Kernel Method	0.79	86	0.63	86
Asymp. Normal	0.87	84	0.63	83

3.3 *Experiment* 3: *heteroscedastic data*

Here n is 20 and the observations in each run were normal. They were chosen so that

$$\operatorname{Var}(Y \mid X = x) = \begin{cases} 0.5, & x \leq 5 \\ 1.00, & x > 5. \end{cases}$$

As before, each tabulated entry was on the basis of 100 replications.

Table 3. Average confidence interval width and percent coverage sample size = 20.

	Average width	Percent coverage
Kernel Method	0.95	79
Asymp. Normal	1.01	76

The final result we report is a repetition of the above experiment with the modification that now n is 200 observations. The table is based on 20 replications.

Table 4. Average confidence interval width and percent coverage sample size = 200.

	Average width	Percent coverage
Kernel Method	0.49	89
Asymp. Normal	0.36	65

3.4 *Commentary on the experiments*

Our interpretation of these studies and others we have undertaken is that the kernel method is superior to the simplistic normal approximation rule in the small sample case, and at least as good for larger samples. Thus, ignoring the extra programming and computational effort involved, the technique is attractive. The overall improvement in terms of shorter confidence intervals and percent coverage in Tables 1 through 3 is about 5%. The dramatic improvement in Table 4 stems from the characteristic that the asymptotic normal rules do not account for changing variance. Had we incorporated such a feature, it would have lost accuracy in the homoscedastic case. Bear in mind that throughout all these runs, we used the same kernel method code.

4. Asymptotic convergence of the error distribution estimate

4.1 *Introduction and basic assumptions*

In this section we seek to establish that under certain conditions the distribution function of $(m_n^*(x) - m(x))$, with $m_n^*(x)$ as determined by (2.6), is asymptotically normal. Perhaps more importantly, we show that it converges to the actual distribution of $(m_n(x) - m(x))$ and therefore is suitable for inference of the regression error. The orders of the errors of these approximations are provided.

We will adopt the following assumptions:

About the Data Sequence:

D.1 For some positive constants c and C, the relations

$$\mathrm{Var}(Y \mid X = x') > c \quad \text{and} \quad E[|Y|^3 \mid X = x'] < C \tag{4.1}$$

hold for all x' in some neighborhood of x.

D.2 The marginal variable X has a continuous pdf $f(\ )$ and x is in its support.

D.3 In a neighborhood of x, the regression function has a bounded second derivative (or Hessian, if $d > 1$), $m''(\ )$.

About the Regression Kernels:

K.1 The support of $k(\ )$ is the d-dimensional open unit ball, and $k_A(\ )$ has support in the open interval $(-1, 1)$. There are positive numbers $C1$ and $C2$ such that for any z in the respective supports,

$$C1 \leq k(z) \leq C2. \tag{4.2}$$

K.2 For $K(z) = k(z)$ or $k_A(z)$

$$\int zK(z)dz = 0. \tag{4.3}$$

K.3 $K(z)$ is twice continuously differentiable on its support set.

About the Bandwidth Parameters:

B.1 The parameters $b(n)$ and $c(n)$ are both $o(n^{-1/(4+d)})$, and $a(n)$ is $o(n^{-1/5})$.

Reference to the Experiments:

We comment that the popular Epanechnikov kernel used in our experiments of Section 3 does not satisfy the condition K.1, since it is not bounded away from 0 on the interval $(-1, 1)$. We thought that for small, predetermined sample sizes, this was not a serious contradiction. For example, one could flatten out the slope of the kernel over very small neighborhoods of the endpoints and know that with very high probability, this new kernel will return the same values as the original kernel, because it is very improbable that within 20 observations, $x - X(i)$ will fall into these minuscule regions where the functions differ. The property K.1 is influential only if the sample size is large or unbounded.

It is known (e.g., Marron (1987)) that cross-validation bandwidth selectors yield sequences that are $O(n^{-1/(d+4)})$, in contradiction to B.1, which requires $o(n^{-1/(d+4)})$. Our plan calls for using cross validation and then making the parameter slightly smaller, to assure that bias is not the limitation. Another more complicated avenue for confronting this distinction could be to extend the algorithm to include an estimate of the bias term.

4.2 *Convergence analysis*

We remind the reader of the pdf construct (2.4) and (2.5) of the random variable $m_n^*(x)$. That is,

$$(4.4) \qquad g_n^*(y \mid x) = \frac{1/(nc(n)^d a(n)) \sum k((x - X(i))/c(n)) k_A((y - Y(i))/a(n))}{1/(nc(n)^d) \sum k((x - X(i))/c(n))}.$$

We will let σ_n^* denote the standard deviation of $m_n^*(x)$, and $H_n(\ )$ the cdf of $(m_n^*(x) - m_n(x))$. In the same vein, σ_n and $F_n(x)$ are, respectively, the standard deviation of $m_n(x)$ and cdf of $(m_n(x) - m(x))$.

The purpose of the present section is to prove the following statement:

THEOREM 4.1. *Under the hypotheses* D.1, D.2, D.3, K.1, K.2, K.3 *and* B.1,

$$\sup_y |H_n(\sigma_n^* y) - F_n(\sigma_n y)| = O_p(\tau(n) q^*(n)),$$

where

$$q^*(n) = (w(n) + n^{1/2} w(n)^{2+d/2} + (nw(n)^d)^{(-1/2)}),$$
$$w(n) = \max\{a(n), b(n), c(n)\},$$

and $\tau(n)$ is any sequence increasing without bound.

Remark. The point of the theorem is that $F_n(y)$ is a useful object; it immediately gives confidence intervals for regression estimation error, and could serve for hypothesis testing and other decision problems involving $m_n(x)$. The function $H_n(y)$ can be calculated entirely on the basis of observations. The theorem tells us that these distribution functions converge, and it gives us rates.

The proof is divided into three lemmas.

LEMMA 4.1. *Suppose $b(n) \geq n^{-a}$ for some a in the open unit interval. Define $\bar{n}$ =Number of i, $1 \leq i \leq n$, such that $|X(i) - x| < b(n)$. Then as $n \to \infty$,*

$$\bar{n}/(nb(n)^d) \to Vf(x), \quad \textit{in probability}, \tag{4.5}$$

V being the volume of the d-dimensional unit ball.

PROOF. The probability that a point falls into the ball B of radius $b(n)$ centered at x is,

$$p_n = \int_B f(z)dz = Vb(n)^d f(x) + o(b(n)^d). \tag{4.6}$$

Thus, $\bar{n}$ is a binomial variable with

$$E[\bar{n}] = nVb(n)^d f(x) + o(nb(n)^d), \tag{4.7}$$

and the variance of the binomial variable with parameter (p_n, n) is bounded by $np_n = O(nb(n)^d)$. Now the lemma follows by Chebyshev's inequality.

The results of this section are founded on a standard form of the Berry-Esseen Theorem (e.g., Bhattacharya and Rao ((1976), p. 104)) which states that if $\{Z(i): 1 \leq i \leq n\}$ are independent random variables, each with zero mean and finite absolute third moment, then for $Q_n(z)$ the df of the sample average

$$(1/n)(Z(1) + \cdots + Z(n)),$$

we have

$$\sup_z |Q_n(\sigma_n z) - \Phi(z)| \leq C(n)n^{-1/2}, \tag{4.8}$$

where

$$\sigma_n^2 = n^{-2} \sum EZ(i)^2. \tag{4.9}$$

Here $\Phi(z)$ is the standard normal df and

$$C(n) = 2.75 \left((1/n) \sum E[|Z(i)|^3]\right) \Big/ (\sqrt{n}\sigma_n)^3. \tag{4.10}$$

For our developments, it is useful to note in passing that, with respect to row-wise independent arrays $\{Z(i;\, n)\}$, if the absolute third moments $E[|Z(i,\, n)|^3]$ are uniformly bounded for all i and n, and if the variances of $Z(i;\, n)$ are uniformly bounded away from 0, then $C(n)$ in (4.8) may be replaced by a number C not depending on n.

LEMMA 4.2. *Let $F_n(z)$ denote the df of $m_n(x) - m(x)$. Then we have*

$$\sup_y |F_n(\sigma_n y) - \Phi(y)| = O_p(\tau(n)q(n)), \tag{4.11}$$

where

$$q(n) = (b(n) + n^{1/2}b(n)^{2+d/2} + (nb(n)^d)^{(-1/2)}), \tag{4.12}$$

σ_n being the standard deviation of $m_n(x)$, and $\tau(n)$ is any sequence converging to infinity.

Remark. In the classical case that $b(n) = n^{-a}$ for some constant a in the unit interval and $d = 1$, one can check that the normal approximation becomes accurate whenever a is in the interval (1/5, 1). The 1/5 threshold makes sense, for otherwise the bias term dominates. Schuster (1972) postulates $a > 1/5$ for his proof of asymptotic normality.

PROOF. Let $S(n)$ be the indices i ($i \leq n$) such that $|X(i) - x| \leq b(n)$. We let $\bar{n}$ be the number of elements in $S(n)$. In the notation surrounding (4.8), the variables $\bar{n}B(i)Y(i)$, $i \in S(n)$, will play the role of the $Z(i)$'s, and $\bar{n}$ will replace n, in the Berry-Esseen theorem.

From postulate K.1, it is readily confirmed that for all i in $S(n)$, we have

$$C1/C2 \leq \bar{n}B(i) \leq C2/C1. \tag{4.13}$$

With the $B(i)$'s thus constrained, postulate D.1 then implies that

$$\left\{(1/\bar{n})\sum E[|\bar{n}B(i)Y(i)|^3]\right\}\Big/(\sqrt{\bar{n}}\sigma_n)^3$$

is uniformly bounded. Now the Berry-Esseen theorem, applied to the $Y(i)$'s, conditioned by the $X(i)$'s, implies that for

$$\bar{m} = \sum B(i)m(X(i)) \tag{4.14}$$

we have

$$\sup_y |P[(m_n(x) - \bar{m}) < \sigma_n y] - \Phi(y)| = O(\bar{n}^{-1/2}).$$

Toward analyzing the asymptotic bias, we have that

$$\begin{aligned} F_n(\sigma_n y) &= P[(m_n(x) - m(x)) \leq \sigma_n y] \\ &= P[(m_n(x) - \bar{m}) \leq \sigma_n(y + (m(x) - \bar{m})/\sigma_n)] \\ &= \Phi(y) + [\Phi(y + (m(x) - \bar{m})/\sigma_n) - \Phi(y)] + O(\bar{n}^{-1/2}) \\ &= \Phi(y) + \phi(y)((m(x) - \bar{m})/\sigma_n) \\ &\quad + o((m(x) - \bar{m})/\sigma_n) + O(\bar{n}^{-1/2}), \end{aligned} \tag{4.15}$$

where $\phi(\ )$ is the standard normal pdf.

Toward achieving a probability bound for $\bar{m} - m(x)$, write

$$\bar{m} - m(x) = \frac{(1/nb(n)^d)\sum k((x - X(j))/b(n))(m(X(j)) - m(x))}{(1/nb(n)^d)\sum k((x - X(i))/b(n))}. \tag{4.16}$$

One recognizes the denominator to be the kernel density estimator $f_n(x)$ of the marginal $f(x)$. Regarding equation (4.16), we introduce the notation

$$\bar{m} - m(x) = \mathrm{Num}(n)/f_n(x).$$

In the following, w is some positive number less than $f(x)$, and Γ is some positive number satisfying

$$\Gamma^2 < V,$$

with

$$V = \frac{1}{f(x)}\mathrm{Var}(Y \mid x)\int k^2(x')dx'.$$

In terms of these constants and the notation of (4.16), we define the events E_1, E_2 and E_3 as follows:

$$\begin{aligned} E_1(n) &= \text{``}f_n(x) > f(x) - w\text{''}, \\ E_2(n) &= \text{``}|\mathrm{Num}(n)| < \tau(n)q(n)(nb(n))^{1/2}\text{''}, \\ E_3(n) &= \text{``}\sigma_n > \Gamma/\sqrt{b(n)}n\text{''}. \end{aligned}$$

When all three of these events hold simultaneously, then, as we invite the reader to confirm, (4.15) implies that

$$\sup_y |F_n(\sigma_n y) - \Phi(y)| < \tau(n)q(n). \tag{4.17}$$

Thus Lemma 4.2 will be confirmed if only we can show that the probabilities of the events $E_j(n)$ converge to 1, as $n \to \infty$, $j = 1, 2, 3$.

The convergence of the probabilities of the events $E_1(n)$ and $E_3(n)$ to 1 follows from standard consistency results in the literature (e.g., Prakasa-Rao (1983)), and thus our effort concentrates on analysis of $E_2(n)$. We will show that for some number K and all n

$$E[(\mathrm{Num}(n))^2] < K(b(n)^2/nb(n)^d + b(n)^4). \tag{4.18}$$

Toward verifying (4.18), define

$$c(u) = k((u-x)/b(n))(m(u) - m(x)).$$

Then we have

$$\begin{aligned} E[c(X)^2] &= \int k^2((x-u)/b(n))(m(u) - m(x))^2 f(u)du \\ &= (m'(x)b(n))^2 f(x)b(n)^d \int v^2k^2(v)dv[1 + O(b(n))] \\ &= O(b(n)^{2+d}). \end{aligned} \tag{4.19}$$

Cross terms $E[c(X(i))c(X(j))] = E[c(X(i))]E[c(X(j))]$ are accounted for as follows:

$$\begin{aligned} E[c(X)] &= \int k((u-x)/b(n))(m(u)-m(x))f(u)du \\ &= b(n)^d \int k(v)m'(x)(b(n)v + (1/2)m''(x)b(n)^2v^2 \\ &\quad + o(b(n)^2))f(x+b(n)v)dv, \end{aligned}$$

where we have substituted the variable $v = (u-x)/b(n)$. Use property K.2 of unbiasedness of $k(\ )$ to get

$$E^2[c(x)] = O(b(n)^{4+2d}).$$

In summary,

$$\begin{aligned} (4.20)\quad & E[\mathrm{Num}(n)^2] \\ &= 1/(n^2b(n)^{2d})\sum_i \left\{ E[c(X(i))^2] + \sum_j E[c(X(i))c(X(j))] \right\} \\ &\le 1/(n^2b(n)^{2d})\{nE[c(X)^2] + n(n-1)E^2[c(X)]\}. \end{aligned}$$

This is tantamount to the statement that

$$(4.21)\qquad E[\mathrm{Num}(n)^2] = O(b(n)^{2-d}/n + b(n)^4).$$

The above and a use of the "basic inequality" of Loève ((1955), p. 157), for example, gives

$$P[\mathrm{Num}(n)/\sqrt{nb(n)} < \tau(n)q(n)] \to 1$$

as $n \to \infty$. This is a restatement that $P[E_2(n)] \to 1$.

This accounts for all but the last term in (4.12). Get the last term by replacing $\bar{n}$ in (4.15) by $b(n)^dn$, according to Lemma 4.1.

Recall the density construct for $Y \mid (X = x)$ stated in (4.4), and its use in construction of $m_n^*(x)$.

LEMMA 4.3. *Let $H_n(y)$ denote the df of $(m_n^*(x) - m_n(x))$ and σ_n^* its standard deviation. Then*

$$\sup_y |H_n(\sigma_n^* y) - \Phi(y)| = O_p(q^*(n)\tau(n)).$$

where

$$q^*(n) = w(n) + (nb(n)^d)^{-1/2} + (nb(n)^d)^{1/2}w(n)^2,$$

with $w(n) = \max\{a(n), b(n), c(n)\}$ and $\tau(n)$ any sequence growing without bound.

PROOF. The moment conditions for the Berry-Esseen theorem are verified as follows: since $k_A(y)$ has support restricted to $|y| < 1$ (Hypothesis K.1), $k_A((y-$

$Y(i))/a(n))/a(n)$ assigns its entire mass to the interval $[Y(i) - a(n),\ Y(i) + a(n)]$. Whence we have

$$
\begin{aligned}
(4.22)\quad & E[|Y^*(i)|^3] \\
&= \frac{1/a(n) \sum k((X(i) - X(j))/c(n)) \int k_A((y - Y(j))/a(n)) |y|^3 dy}{\sum k((X(i) - X(j))/c(n))} \\
&\leq \frac{\sum k((X(i) - X(j))/c(n))(|Y(i)| + a(n))^3}{\sum k((X(i) - X(j))/c(n))}.
\end{aligned}
$$

In (4.22), the summations are for j ranging from 1 to n. The reader will recognize that the final term is a kernel regression estimator of $E[(|\,Y| + a(n))^3 | X = X(i)]$. Lemma 2 of Devroye and Wagner (1980) implies that under our hypotheses D.1 and K.1 of the uniform boundedness of the conditional third moments of Y and of $k(z)$, that $C(n)$ (in (4.8)) is bounded in n, and so from the Berry-Esseen result,

$$P[(m_n^* - m^*) \leq \sigma_n^* y] - \Phi(y) = O(\bar{n}^{-1/2}),$$

where

$$(4.23)\qquad m^* = \sum B(i) E[Y^* \mid X = X(i)]$$

is the expectation of $m_n^*(x)$. Analogously to (4.15) one readily confirms that

$$\sup_y |H_n(\sigma_n^* y) - \Phi(y)| = O_p((m^* - m_n(x))/\sigma_n^* + \bar{n}^{-1/2}).$$

Toward analyzing the bias term $m^* - m_n(x)$, we find it most direct to appeal to the analysis of $m(x) - m$ as in the proof of Lemma 4.2, and then to analyze $m - m^*$. In fact, from inspection of (4.23),

$$m_n(x) - m^* = \sum B(i)[m(X(i)) - m_n^*(X(i))].$$

Now the bias analysis for Lemma 4.2 applies to the terms in brackets, and the $B(i)$'s, since they sum to 1, do not affect the convergence rate. Thus the argument there can be repeated to get that

$$m^* - m_n(x) = O_p(w(n)^{2-d}/n + w(n)^4).$$

Proof of Theorem 4.1. Lemma 4.1 gives the probabilistic rate of the uniform distance of $F_n(\sigma_n y)$ and $\Phi(y)$ while Lemma 4.2 gives the rate between $\Phi(y)$ and $H_n(\sigma_n^* y)$. Adding these relations and application of the triangle inequality yields Theorem 4.1.

Remark. We have considered the possibility of improving the result through approximating $m(x) - m$ by $\sum((d/dx)B_i(x))(x - X(i))$, with $B_i(x) = B_i$ as defined in (2.2). It does not improve the asymptotic rates.

References

Bhattacharya, P. K. (1976). An invariance principle in regression analysis, *Ann. Statist.*, **4**, 621–624.

Bhattacharya, R. and Rao, C. R. (1976). *Normal Approximations and Asymptotic Expansions*, Wiley, New York.

Devroye, L. and Wagner, T. J. (1980). Distribution-free consistancy results in nonparametric discrimination and regression function estimation, *Ann. Statist.*, **8**, 231–259.

Dikta, G. (1988). Bootstrap approximations of nearest neighbor regression functions, Preprint, University of Glessen.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statist. Sci.*, **1**, 54–74.

Härdle, W. (1987). *Applied Nonparametric Regression*, Preprint of a textbook.

Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bounds, *J. Amer. Statist. Assoc.*, **83**, 102–110.

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

Knafl, G., Sacks, J. and Ylvisaker, D. (1985). Confidence bands for regression functions, *J. Amer. Statist. Assoc.*, **80**, 683–691.

Loève, M. (1955). *Probability Theory*, 3rd ed., Nostrand, Princeton, New Jersey.

Marron, J. S. (1987). A comparison of cross-validation techniques in density estimation, *Ann. Statist.*, **15**, 152–162.

Müller, H. and Stadtmüller, U. (1987). Variable bandwidth kernel estimators of bandwidth, *Ann. Statist.*, **15**, 163–181.

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

Ripley, B. D. (1981). *Spatial Statistics*, Wiley, New York.

Rutherford, B. (1986). Bootstrap and related methods for approximate confidence bounds in nonparametric regression, Ph. D. Dissertation, Systems and Industrial Engineering Department, University of Arizona.

Schuster, E. F. (1972). Joint asymptotic distribution of the estimated regression function at a finite number of discrete points, *Ann. Math. Statist.*, **7**, 139–149.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap, *Ann. Statist.*, **9**, 1187–1195.

Vieu, P. (1988). Nonparametric regression: Optimal local bandwidth choice, *J. Roy. Statist. Soc. Ser. B* (to appear).

Working, H. and Hotelling, H. (1929). Applications of the theory of error to the interpretation of trends, *J. Amer. Statist. Assoc.*, **24**, 73–85.