Statistical Science
2004, Vol. 19, No. 4, 663–675
DOI 10.1214/088342304000000756

Kernel Smoothers: An Overview of Curve Estimators for the First Graduate Course in Nonparametric Statistics

William R. Schucany

Abstract. An introduction to nonparametric regression is accomplished with selected real data sets, statistical graphics and simulations from known functions. It is pedagogically effective for many to have some initial intuition about what the techniques are and why they work. Visual displays of small examples along with the plots of several types of smoothers are a good beginning. Some students benefit from a brief historical development of the topic, provided that they are familiar with other methodology, such as linear regression. Ultimately, one must engage the formulas for some of the linear curve estimators. These mathematical expressions for local smoothers are more easily understood after the student has seen a graph and a description of what the procedure is actually doing. In this article there are several such figures. These are mostly scatterplots of a single response against one predictor. Kernel smoothers have series expansions for bias and variance. The leading terms of those expansions yield approximate expressions for asymptotic mean squared error. In turn these provide one criterion for selection of the bandwidth. This choice of a smoothing parameter is done a rich variety of ways in practice. The final sections cover alternative approaches and extensions. The survey is supplemented with citations to some excellent books and articles. These provide the student with an entry into the literature, which is rapidly developing in traditional print media as well as on line.

Key words and phrases: Local polynomial regression, AIC, variable bandwidths, cross validation, windows.

1. INTRODUCTION

Nonparametric curve estimators are valuable tools in statistical practice. There is a rich variety of such curves and surfaces. A very basic curve estimator is one for a continuous density function. Histograms are widely used rough estimates of probability density functions (p.d.f.). These blocky displays have a venerable history. They also have some deficiencies relative to estimators that have the same continuity as an assumed model p.d.f. $f(x)$; see Sheather (2005).

William R. Schucany is Professor, Department of Statistical Science, Southern Methodist University, Dallas, Texas 75275-0332, USA (e-mail: schucany@smu.edu).

The dynamic graphics that are available on line (*http://www.stat.sc.edu/rsrch/gasp/*) provide a nice introduction to the issue of bin size for histograms. The Java script by Webster West demonstrates the effect of user-controlled continuous variation of bin widths for the Old Faithful data (see Figure 1). The 107 times between eruptions of the geyser in Yellowstone Park are evidently bimodal. The student can visualize a smoothly changing array of bin sizes from large enough to hide the two modes to small enough to produce spikes at each of the data points. These graphics for the distributions of univariate observations have been extended to higher dimensions. These topics are covered well by Scott (1992). Next we change to curves for patterns of association.

The same tensions exist for a curve that models the association between a response Y and a predictor X.

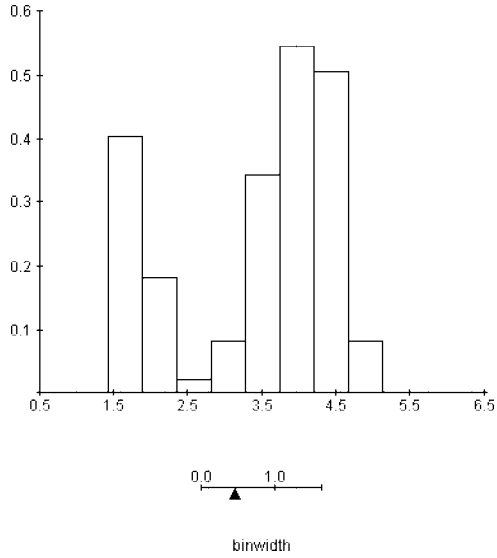

FIG. 1. *Histogram of eruption time intervals.*

One may visualize the relationship in scatterplots of the y_i against the x_i. When there is no sound reason to force a simple straight line, then we may let the data speak for themselves. Figure 2 displays some real data in which a nonlinear association is obvious. The scatterplot is Figure 1.1 in Wand and Jones (1995) of data from Ullah (1985). Chu and Marron (1991) used this same example in their introduction to kernel regression. The validity of a feature such as the dip in the forties is a challenge to curve estimation methodology. (As always, one must be wary of selection bias in any cross-sectional study.)

There are several alternatives for producing a smooth curve to model some characteristic of the distribution of Y given each value of x. Typically we are interested in the *regression* function

$$m(x) = E[Y|X = x].$$

In addition we may be willing to impose some smoothness constraints on this unknown $m(\cdot)$ and an additive error model for the n pairs,

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

The conventional approach is to treat the ε_i as independent and identically distributed. The assumption of constant variance can be relaxed to allow a variance function $\sigma(x_i)$. See Ruppert, Wand and Carroll (2003) for more on heteroscedastic models and extensions to mixed models.

An accessible introduction to the basic ideas of nonparametric regression can be found in Altman (1992). The introductory sections of books that I recommend to beginning students are Chapter 1 of Eubank (1999); Sections 2.1–2.3 of Hart (1997) and Sections 5.1–5.6 of Wand and Jones (1995). The book by Fan and Gijbels (1996) is a thorough treatment of kernel methods for local polynomials from one perspective. Loader (1999) provided another perspective on several basic

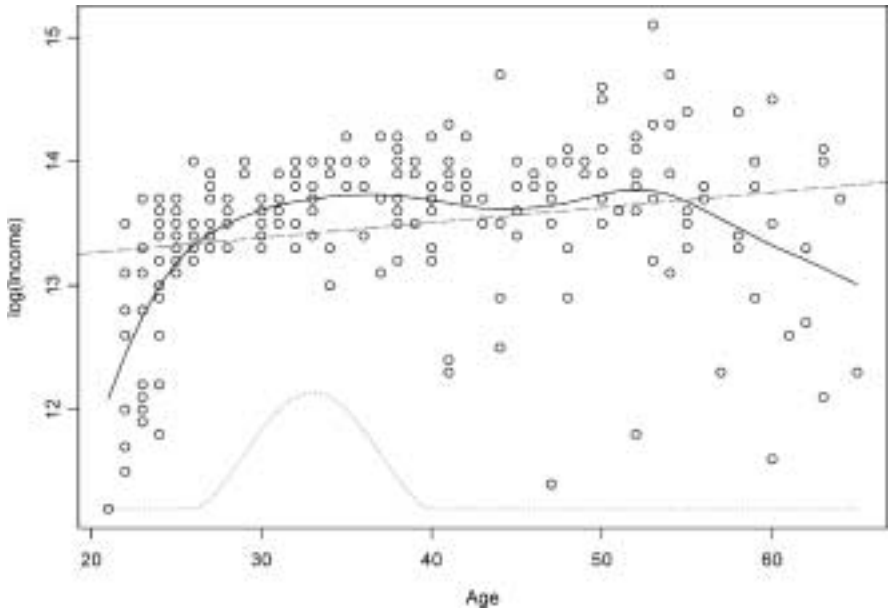

FIG. 2. *Scatterplot of log* (*income*) *versus age for* 205 *Canadian workers. The ordinary least-squares line is dashed. The solid curve is a local-linear fit with the biweight kernel* (*dotted and arbitrarily centered at* 33), *whose bandwidth* $h = 7.14$ *is optimal in a sense explained in Section* 6.

issues. These two perspectives come into focus herein in Sections 2.1 and 2.2. Extensions beyond the univariate case and others are briefly described in Section 6.

A great deal of understanding may be gleaned from simply reading books and articles. Many students absorb quite a lot in the passive comprehension of text, formulas and graphs. Even so, more is added by an activity that involves running the computer routines. In my experience such statistical procedures are more completely grasped by actually doing them. Instructors and students have greater insight into nonparametric regression after getting a real set of data, plotting the pairs, using one of the available smoothers and adding the resulting curve to the scatterplot. For the data in Figure 2 one may examine a rich variety of optional fits. In the present introduction there is a brief treatment of what, who, when and why.

1.1 What

The essential idea is *local* averaging. Thus it is sensible to restrict our attention to linear combinations of the responses. The parallel linear filters in engineering and physics provide some support for this approach. The size of the local neighborhood is called the *bandwidth*. We consider broader classes of local models in subsequent sections, as well as relaxing the view that neighborhoods are finite windows.

1.2 Elementary Illustration

The data in Figure 3 are simulated from (1) with the mean function $m_3(x) = 4.26[e^{-9.75x} - 4e^{-19.5x} + 3e^{-29.25x}]$ evaluated on an equally spaced grid of 100 x's. Such a linear combination of three exponentials as $m_3(\cdot)$ has been used to simulate a function with changing curvature since Wahba and Wold (1975). It resembles the familiar "motorcycle data" used by Fan and Gijbels (1996) to motivate the challenge of local modeling. To simplify the task of understanding analytical properties, we consider the Priestley and Chao (1972) (PC) estimator in detail in Section 2.1. There are two distinctly different versions of this elementary scatterplot smoother, known as Nadaraya–Watson (Watson, 1964; Nadaraya, 1965) and Gasser–Müller (GM; 1979). All three are asymptotically equivalent. The Nadaraya–Watson (NW) estimator is the special case of fitting a constant locally at any x_0. Here we assume without loss of generality that the x's are confined to the unit interval $x \in [0, 1]$. The NW estimate of $m(x_0)$ based on n pairs, $(x_1, y_1), \ldots, (x_n, y_n)$, is

$$\hat{m}(x_0) = \frac{\sum_{|x_i - x_0| < h(x_i)} y_i w(x_i - x_0)}{\sum_{|x_i - x_0| < h(x_i)} w(x_i - x_0)}, \tag{2}$$

where $w(x_i - x_0) = (1/h(x_i))K((x_i - x_0)/h(x_i))$. The details of the *kernels* K, the *bandwidths* h and the *design* $\{x_i\}$ are developed in Section 2. This is clearly a linear combination of the y_i. All of the smoothers in this paper have the form $\sum_{i=1}^{n} l_i y_i$. The weights l_i are determined in several ways in practice. Figure 3 displays a specific evaluation of the weighted average in (2) at $x_0 = 2/3$ using the *biweight* kernel $K(z) \propto (1 - z^2)^2$ and bandwidth $h = 0.1$.

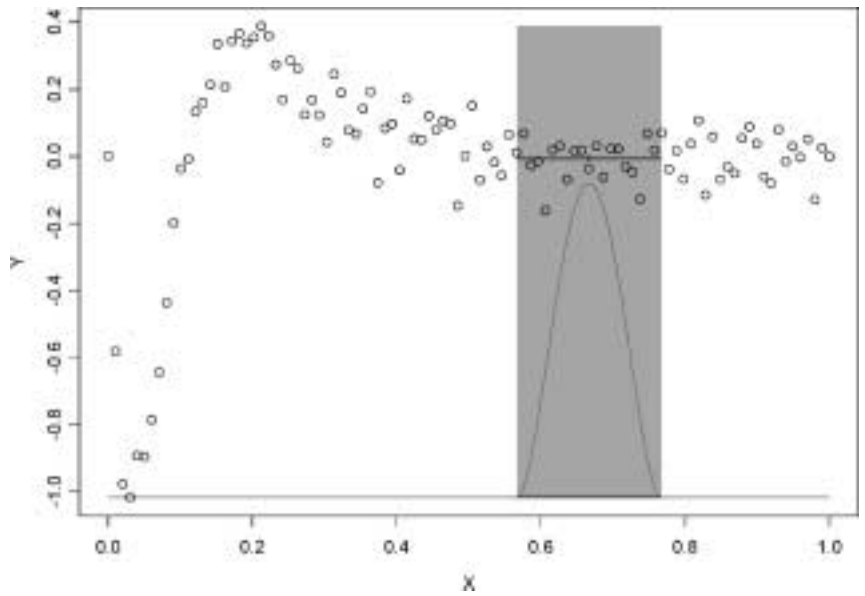

FIG. 3. *Fitting a local constant with NW in* (2). *Data from* m_3 *plus normal noise* ($\sigma = 0.065$), $n = 100$, $x_0 = 2/3$, *bandwidth* $h = 0.1$ *and a biweight kernel* (*shaded*).

1.3 Two Kinds of Windows

The kernel weights K are calculated under two distinct approaches: (1) a fixed window width as in Figure 3 and (2) a fixed fraction of the data. In the first approach the bandwidth is typically denoted by h. When the generic kernel has compact support [e.g., uniform on $(-1, 1)$, triangular, quadratic or biweight], the estimator depends only on those pairs whose x_i are in the interval $(x_0 - h, x_0 + h)$. In this formulation the bandwidth (or smoothing parameter) is a scale parameter. When the kernel is a p.d.f. such as the standard normal, h is the standard deviation.

The second approach uses the k nearest neighbors to x_0. That is, the pairs with x_i closest to x_0 influence the estimate regardless of how distant x_i may be from x_0. These two distinct avenues yield either (1) a random number of x_i within the fixed width h or (2) a fixed number k within an interval of random width. For equally spaced x's these two are equivalent. When the spacings between the x's are not constant, the estimates and their properties differ. The symbol K for kernel functions is not to be confused with the integer k for the number of nearest neighbors. For loess, an alternative implementation of local-linear smoothing in S-Plus, the definition of span is the fraction k/n. Even though the default value (span $= 2/3$) may seem rather large, one may find that the results for $n = 100$ bivariate normals with $\rho = 0.6$ can be surprisingly nonlinear. The remainder of this paper focuses only on the first approach with a bandwidth h.

1.4 History

Loader (1999) gave a thorough coverage of the origins of local fitting, tracing it to the late 19th and early 20th century. Notably, the early contributions in actuarial science were extensive and were in widespread use. A data set and a linear smoother from Spencer (1904) addressed what was then known as the problem of graduation.

Early contributors to the kernel density estimation alternatives to histograms were Rosenblatt (1956) and Parzen (1962). For a scatterplot the parallel to the histogram is a set of piecewise constants over intervals of equal length called a regressogram by Tukey (1961). See Schlee (1988) for a brief introduction. Figures 1.3–1.5 in Eubank (1999) illustrate a regressogram fit to simulated data using a partition with seven bins. Tukey's title phrase "Curves as parameters…" anticipates the point of view that is essential to functional data analysis introduced by Ramsay and Silverman (1997).

The monograph by Wand and Jones (1995) is a comprehensive coverage of both kernel density estimation and kernel regression. They treat the entire class of kernel-type estimators of a regression function known as local polynomial kernel estimators. These estimate $m(x_0)$ by fitting a polynomial of degree p by weighted least squares. The class was introduced by Stone (1977) and studied by Cleveland (1979), and many of the properties were established by Müller (1987) and Fan (1992). The importance of the special case $p = 1$, or *local-linear kernel regression*, is due in part to its simplicity. Local-linear kernel regression has better properties than NW at the boundaries and asymptotically. Such large-sample bias comparisons are deferred to the next section.

Local-linear smoothers, derived in Section 2.2, share the advantage of being local with estimators such as NW. In Section 2 one may get some insight into the properties that make $p = 1$ the recommended polynomial degree. It may be easy to overlook the fact that one is fitting a different straight line at every point. The curve estimator $\hat{m}(x_0)$ is a continuous function of x_0. One of the elementary concepts in differential calculus asks a student to think of the smoothly progressing sequence of tangent lines. Here the smooth transition of lines fit at x_0 is for a different purpose, but the analogy may help some.

This smooth curve goes well beyond the illustration in Figure 4, which displays the smaller collection of $\hat{m}(x_1), \hat{m}(x_2), \ldots, \hat{m}(x_n)$. The simulation is an illustration patterned on a 30/70 mixture of two normal p.d.f's with scales that differ by a factor of 2. Specifically, $m_2(x) = 0.3\exp\{-64(x-0.25)^2\} + 0.7\exp\{-256(x-0.75)^2\} I_{(0,1)}(x)$. Again there are a grid of 100 equally spaced x's and additive normal noise with $\sigma \cong 0.04 = 5\%$ (range of m_2).

1.5 Motives

A natural question in the minds of many is why do we do nonparametric regression? There are several good reasons for producing such curves. One is as a *descriptive* statistic. In other words, a data analyst can accomplish something by graphing an estimate of the unknown m on the scatterplot. Another more valuable reason is for testing a simple parametric model, such as the dashed straight line in Figure 2. The comparison of the fits under the two models allows us to consider *lack-of-fit tests*, which is a compelling reason for those who appreciate George Box's saying, "All models are wrong, but some are useful." A third application is the flexible adjustment for co-

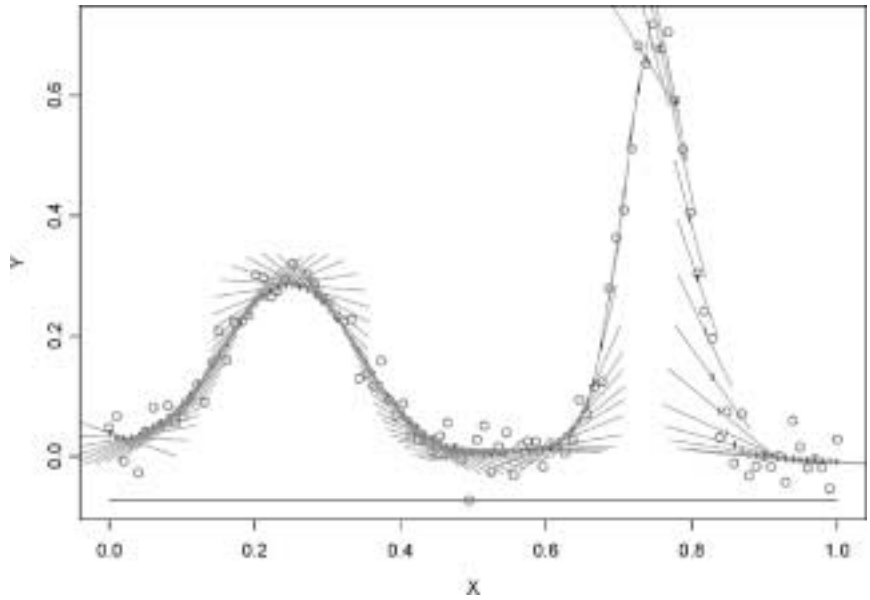

FIG. 4. *Data from $m_2(x)$ on a grid of $n = 100$ with local-linear estimates of $m(x_1), \ldots, m(x_n)$ at those x's.*

variates. This process of inference about the parameters of a model in the face of a nonparametric nuisance is known as *semiparametric*. An elementary example might involve a one-way layout of subjects' responses, $Y_{ij} = \mu + \beta_j + m(x_{ij}) + \varepsilon_{ij}$, in which the effect of age x could be controlled without imposing a linearity condition.

2. MEANS AND VARIANCES FOR LOCAL SMOOTHERS

As with kernel density estimators (see Silverman, 1986), we can produce approximations for the mean and variance of $\hat{m}(x_0)$. A Taylor series expansion of the unknown mean function is the classic approach to an analytical approximation of the large-sample properties. The leading terms of these expansions yield asymptotic expressions for the bias and the variance of kernel estimators, $\hat{m}(x_0)$. The role of K as a symmetric p.d.f. is apparent in the expansion (5).

2.1 Weighted Average

The more intuitive local average by PC aids our understanding of these properties, which were more rigorously demonstrated by Benedetti (1977). Assume here without loss of generality that $0 \le x_1 \le x_2 \le x_3 \le \cdots \le x_n \le 1$, known as a *fixed design*. The PC estimator is

$$\hat{m}_{\mathrm{PC}}(x_0) = \sum_{i=1}^{n} Y_i \frac{x_i - x_{i-1}}{h} K\left(\frac{x_0 - x_i}{h}\right), \tag{3}$$

where K is a kernel constrained to be a unimodal p.d.f. supported on $(-1, 1)$ and symmetric about zero. Approximating sums by integrals may make it more apparent that these estimators are convolving $K(\cdot)$ with $m(\cdot)$. The estimator in (3) is asymptotically normal and the convolution is consistent, provided that the scale parameter h becomes vanishingly small.

The equation that follows derives from an elementary application of the expectation of the Y_i from model (1). For these approximations to hold requires an infinitesimal h as n gets large, technically $h_n \to 0$ as $n \to \infty$, which also brings the x_i closer together. Approximating the sum over i by an integral with respect to x,

$$\begin{aligned} E[\hat{m}_{\mathrm{PC}}(x_0)] &= \sum_{i=1}^{n} E[Y_i] \frac{x_i - x_{i-1}}{h} K\left(\frac{x_0 - x_i}{h}\right) \\ &= \int_0^1 m(u) \frac{1}{h} K\left(\frac{x_0 - u}{h}\right) du + O(n^{-1}). \end{aligned}$$

A change of variable $z = (x_0 - u)/h$ so that $u = x_0 - zh$ yields

$$\begin{aligned} &E[\hat{m}_{\mathrm{PC}}(x_0)] \\ &\quad = \int_{-(1-x_0)/h}^{x_0/h} m(x_0 - zh) K(z)\, dz + O(n^{-1}). \end{aligned} \tag{4}$$

Using a Taylor series expansion of $m(\cdot)$ about x_0, the integral in (4) is approximately

$$\int_{-1}^{1} \left[m(x_0) - zhm'(x_0) + \tfrac{1}{2} z^2 h^2 m''(x_0) - \cdots\right] K(z)\, dz$$

for sufficiently small h. For the normal kernel the integral is over $(-\infty, \infty)$. Thus formally

$$E[\hat{m}_{\mathrm{PC}}(x_0)] = m(x_0)\int K(z)\,dz - hm'(x_0)\int zK(z)\,dz + (h^2/2)m''(x_0)\int z^2K(z)\,dz - \cdots. \tag{5}$$

Since K is a p.d.f. which is symmetric about zero, the leading term of the bias expansion is

$$\frac{h^2}{2}m''(x_0)\mu_2(K) + o(h^2) + O(n^{-1}) \tag{6}$$

with the obvious notation for the second moment. Similar approximations for small h lead to

$$\mathrm{Var}[\hat{m}_{\mathrm{PC}}(x_0)] = \frac{\sigma^2\int K^2(z)\,dz}{nh} + o\left(\frac{1}{nh}\right). \tag{7}$$

See Fan and Gijbels (1996, Section 3.7) for a derivation of such asymptotic bias and variance expressions for more general designs and a larger class of estimators.

Optimal bandwidths. Notice that the bias in (6) is small for small h and the variance in (7) is small for large h, so a proper choice of h involves the usual bias–variance trade-off. A reasonably standard way to accomplish this trade-off is to minimize the leading term of the expansion of the asymptotic mean squared error (AMSE) at x_0. Therefore, as $n \to \infty$ so that $h_n \to 0$ in a manner such that $nh \to \infty$, summing the variance and the square of the bias yields (introducing some obvious new notation)

$$\mathrm{AMSE}(x_0) = \frac{\sigma^2R(K)}{nh} + \frac{\mu_2^2(K)}{4}[m''(x_0)]^2h^4 = \frac{A}{nh} + \frac{Bh^4}{4}. \tag{8}$$

The large-sample approximations here hold only for values of x_0 that are not within one bandwidth of either end of the range. In the limit, h_n becomes small enough that x_0 will not be too close to 0 or 1. When that close proximity to either end of the range occurs, there is a boundary bias that is not captured by the expression in (6).

Differentiating (8) with respect to h and setting it to zero yields $-A/(nh^2) + Bh^3 = 0$, which implies $h^5 = A/Bn$. Thus the asymptotically optimal bandwidth is $h_*(x_0) = [A/Bn]^{1/5}$.

Substituting this $h_*(x_0)$ into (8) yields

$$\frac{A^{4/5}}{n^{4/5}}B^{1/5} + \frac{B^{1/5}}{4}\frac{A^{4/5}}{n^{4/5}}.$$

Hence the minimized value is

$$\inf_{h>0}\mathrm{AMSE}(x_0) = \tfrac{5}{4}[\mu_2^2A^4m''(x_0)^2]^{1/5}n^{-4/5}. \tag{9}$$

Therefore, the rate of convergence to zero is of order $n^{-4/5}$, slower than the rates of order n^{-1} that are typical for optimal parametric estimation. That there is a penalty for the more difficult task of nonparametric function estimation should not surprise anyone. This fraction is specific to the model assumptions.

In addition to formalizing the optimal rate of convergence, there is something quite noteworthy in the expression for the optimal bandwidth,

$$h_*(x_0) = \left[\frac{\sigma^2R(K)}{\mu_2^2(K)m''(x_0)^2n}\right]^{1/5}. \tag{10}$$

As with many optimal quantities, this depends on the unknown regression function $m(\cdot)$. Specifically, it depends on the curvature of $m(\cdot)$ at x_0 as measured by the second derivative $m''(x_0)$. The fact that, except for σ, the ingredients of $h_*(x_0)$ are known has enticed a line of research to produce estimates of $m''(x_0)$ and substitute these into $h_*(x_0)$. These so-called plug-in rules are discussed in Section 6.

Recall that the dominant term of the bias expansion is $m''(x_0)\mu_2(K)h^2/2$ for any $h > 0$. This implies that the bias is most severe near peaks and troughs, where $m''(\cdot)$ is greatest. Furthermore, it is positive in any trough and negative at any peak. This explains why these estimators tend to fill in valleys and undershoot peaks, regardless of whether one is using an optimal $h_*(x_0)$. See Figures 4 and 5 for illustrations of this.

Optimal kernels. Still another interesting thing may be learned from these asymptotic expressions for the minimized AMSE in (9). The factor that depends on the kernel K is $R^4\mu_2^2$. It follows that the kernel that is best in this sense minimizes the scale-invariant product

$$R^2(K)\mu_2(K) = \left[\int K^2(z)\,dz\right]^2\int z^2K(z)\,dz,$$

subject to $K(z) \geq 0$ for every z,

$$\int K(z)\,dz = 1, \quad \int zK(z)\,dz = 0$$

and

$$\int z^2K(z)\,dz = a^2 < \infty.$$

The solution due to Hodges and Lehmann (1956) is familiar to students of nonparametrics using ranks. Hodges and Lehmann were seeking this density as a

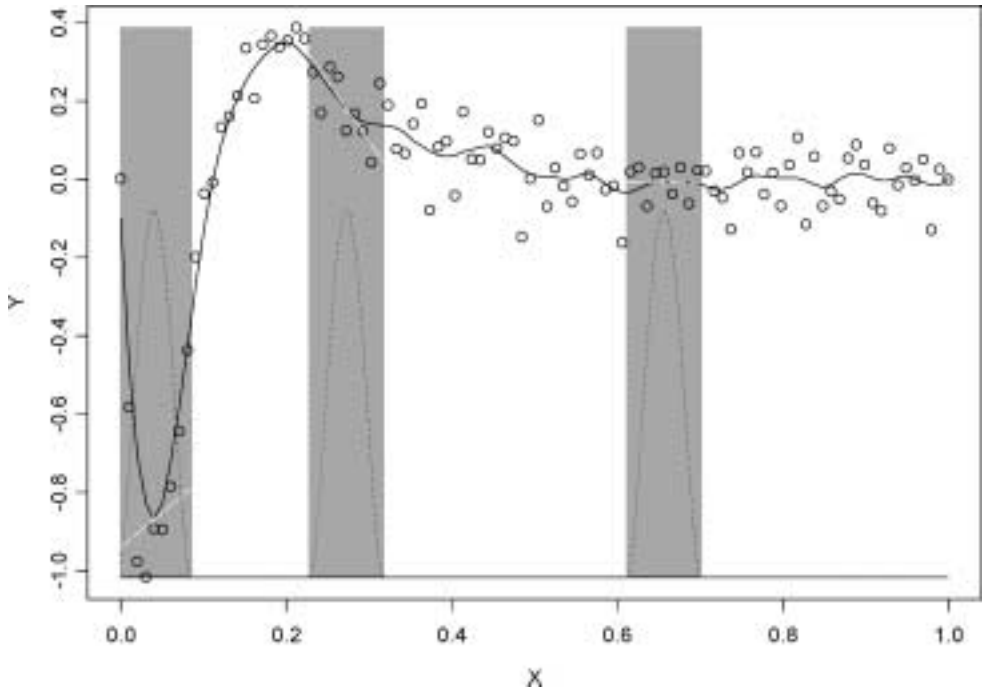

FIG. 5. *Local-linear curve estimate* (*solid line*) *for the same data as Figure* 3 *with* $h = 0.045$. *The biweight kernel* (*dotted*) *is displayed at three selected places and the fitted lines are dashed in the corresponding shaded bands.*

worst-case p.d.f. for the asymptotic relative efficiency of the Wilcoxon rank sum to the two-sample Student t. They obtained their optimal result in a somewhat different context. However, in both settings the goal is to identify functional extremes for asymptotic efficiencies. Their classical finding is

$$K_a(z) = \frac{3}{4}\left[1 - \frac{z^2}{5a^2}\right] a\sqrt{5} I_{(-a\sqrt{5}, +a\sqrt{5})}(z).$$

This best kernel function has a scale parameter a, which may be set to $a^2 = 1/5$ for convenience. The simple quadratic $K_*(z) = 0.75[1 - z^2]I_{(-1,1)}(z)$ is often called the Epanechnikov kernel due to its derivation by Epanechnikov (1969) in the density estimation context. However, earlier credit may be due to Bartlett (1963). Wand and Jones (1995) showed this holds more generally than for (9). With the concept of canonical kernels, they demonstrated a decoupling of K and h, rewriting (8) as

$$\mathrm{AMSE}(x_0) = C(K)\left[\frac{1}{nh} + \frac{h^4 m''(x_0)^2}{4}\right].$$

Design considerations. Essentially the same large-sample results hold for local-linear fitting. Explicit expressions for local polynomial estimators were given by Wand and Jones (1995). They derived details for a special case of local linear for fixed equally spaced x's in Section 5.3, which is outlined here in the next section. The more general development for local polynomials is in the book by Fan and Gijbels (1996). These latter authors investigated a breadth of material, including estimating rth derivatives, $m^{(r)}(\cdot)$, *random* designs, $(X_1, \ldots, X_n)$, bandwidth selection and effective kernels.

There is an important feature that Fan and Gijbels (1995) called *design adaptation*. The term suggests a property of adapting to either fixed or random design. This valuable characteristic of local polynomial fits is not shared by some other kernel methods. Specifically, these local polynomial bias and variance expressions for the fixed design are identical to those for the random design. This is not true for Gasser–Müller estimators, for example, which have a variance that is greater by a factor of 1.5 for random designs. A balanced comparison of these two approaches appeared in Chu and Marron (1991). They carefully investigated properties of two distinct curve estimators, "evaluation weights" represented by NW in (2) and GM "convolution weights" represented by

$$\hat{m}_{\mathrm{GM}}(x) = \sum_{j=1}^{n} Y_j \int_{s_{j-1}}^{s_j} K(x - t)\, dt, \tag{11}$$

where $s_i = (x_i + x_{i+1})/2$, $x_0 = -\infty$ and $x_{n+1} = +\infty$. At that time the authors and discussants agreed that

kernel methods were worthy candidates for improvement. In the process they identified several distinct philosophical points of view that data analysts can bring to the task of smoothing. Fan and Gijbels (1996) made a convincing case that local polynomials eliminate these and other deficiencies of kernel fits of local constants, and they do so from an asymptotically small h perspective.

2.2 Local Polynomial Fitting

Local polynomial fitting is weighted least-squares estimation of $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, the $p+1$ coefficients of a polynomial of degree p. The objective is to minimize

$$\sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1(x_i - x_0) - \cdots - \beta_p(x_i - x_0)^p]^2 \cdot K_h(x_i - x_0),$$

where $K_h(t) = K(t/h)/h$. The standard solution is the $(p+1) \times 1$ estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_0^T \mathbf{W}_0 \mathbf{X}_0)^{-1} \mathbf{X}_0^T \mathbf{W}_0 \mathbf{Y},$$

provided that the matrix is nonsingular. Here $\mathbf{Y}$ is the $n \times 1$ vector of responses,

$$\mathbf{X}_0 = \begin{bmatrix} 1 & x_1 - x_0 & \cdots & (x_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_0 & \cdots & (x_n - x_0)^p \end{bmatrix}$$

is the $n \times (p+1)$ design matrix and $\mathbf{W}_0 = \operatorname{diag}[K_h(x_1 - x_0), \ldots, K_h(x_n - x_0)]$ is the $n \times n$ diagonal matrix of weights.

With this centering on x_0, when evaluated at x_0, the estimate $\hat{m}(x_0; p) = \mathbf{e}_1^T \hat{\boldsymbol{\beta}}$ is the intercept term, where $\mathbf{e}_1$ is the $(p+1) \times 1$ vector $(1, 0, 0, \ldots, 0)^T$. When $p = 0$, this gives NW in (2) and when $p = 1$, an explicit formula for local linear (LL) is

$$\hat{m}_{\mathrm{LL}}(x_0; 1) = \sum_{i=1}^{n} w_i Y_i \Big/ \sum_{i=1}^{n} w_i, \qquad w_i = K_h(x_i - x_0)[\hat{s}_2 - (x_i - x_0)\hat{s}_1], \tag{12}$$

where $\hat{s}_j = \sum_{i=1}^{n} (x_i - x_0)^j K_h(x_i - x_0)$, $j = 1, 2$. This form in (11) is obviously linear in Y. Loader (1999) examined LL relative to local quadratic and cubic alternatives from a nearest-neighbor finite-sample perspective including valuable advice on residual plots and effective degrees of freedom.

Large-sample normality for this entire class of linear smoothers is immediate for a fixed value of h. Even so, when we let the data guide the choice of bandwidth to be denoted by $\hat{h}$, there is still a legitimate sampling distribution for $\hat{m}(x_0)$. Clearly the normal approximation may now be quite inadequate. Nonetheless, there are ways to construct approximate confidence intervals, which are the subject of Section 7. This capability to use a data-based value $\hat{h}$ is an essential feature of kernel smoothers. For us to move beyond an arbitrarily fixed h, any reasonable choice is necessarily data dependent.

3. BANDWIDTH SELECTION

The selection of appropriate values for h is the most challenging aspect of nonparametric regression. This is true for kernel smoothing as well as for any of the other methods, which all have a smoothing parameter of some sort. There are numerous approaches to this task of adapting to the level of the noise and the amount of structure in the data set at hand. The efficiency of the estimator is far more sensitive to the value of h than it is to the choice of K. Movies by J. S. Marron, D. Ruppert, E. K. Smith and G. Conley that teach lessons about local polynomial smoothing are online at *http://www.stat.unc.edu/faculty/marron/Movies/locpoly_movies.html.*

3.1 Plug-in Estimators

This approach addresses efficiency through the asymptotic mean squared error (MSE) but attempts a direct estimate of the optimal h_* in (10). Substituting estimates of unknown quantities in that formula produces a variety of plug-in estimates that have worked well in local-linear regression in some settings. See Wand and Jones (1995) for a description of these bandwidth selectors and citations to the relevant literature. Fan and Gijbels (1996) gave the details of some of these, and implemented both *constant* (also known as *global*) and *variable* bandwidths. They elaborated on the basic concept and elucidated the more sophisticated applications of the plug-in principle.

3.2 Cross Validation

Other "classic" approaches estimate the finite sample $\mathrm{MSE}(x_0)$ or any other information measure and then minimize it. The basic idea behind cross-validation (CV) is to hold out part of the sample with which to evaluate the performance of a predictor. A common practice is to leave one out; here that is (x_i, Y_i) for each $i = 1, \ldots, n$. The predictor of the unused Y_i based on the other $n-1$ pairs may be denoted by

$\hat{m}_i(x_i; h)$. These n prediction errors are summarized in least squares CV by

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}_i(x_i; h)]^2. \tag{13}$$

There are justifications that relate the expectation of this criterion to the MSE averaged over the n design points. The CV estimate of the bandwidth that is optimal in this sense is the minimizer of (13). Alternatives that leave out nonoverlapping fifths or tenths of the sample are mainstays for a wide spectrum of nonparametric regression estimators and other prediction and classification rules in Hastie, Tibshirani and Friedman (2001).

Hart (1997) covered risk estimation and generalized cross validation (GCV) along with his "one-sided" CV. A recent article by Hart and Lee (2005) presented convincing evidence of the relatively undesirable variability of leave-one-out CV bandwidths. This paper offers both an empirical and a theoretical basis for CV's tendency to produce unrealistically small estimates of h. The relative merits of such classical methods versus plug-in rules are explored in depth in Chapter 10 of Loader (1999). Signorini and Jones (2004) provided a thorough examination of these selection methods for both NW and LL in the special case of binary responses.

3.3 Information

The Akaike information criterion (AIC) was originally designed for parametric models as an approximately unbiased estimate of the expected Kullback–Leibler information. For linear regression and time series models, Hurvich and Tsai (1989) showed that the bias of AIC can be large in small samples. This leads to overfitting, especially as the dimension of the candidate model approaches the sample size. One may think of AIC as a maximized log-likelihood plus a penalty for the number of parameters. Huvrich and Tsai proposed a corrected version, denoted by $\mathrm{AIC_C}$, which is less biased than AIC. Hurvich, Simonoff and Tsai (1998) investigated the use of $\mathrm{AIC_C}$ to choose smoothing parameters. They showed that using $\mathrm{AIC_C}$ avoids the large variability and the tendency to undersmooth (compared to the actual minimizer of average squared error) that is typical for other classical approaches such as GCV or AIC.

Consider the same simulation of $m_3(x)$ as in Figure 3. Here in Figure 5 is an evaluation of the entire curve estimate (11) with the same value of $h = 0.045$ throughout. This small h is apparently not a good choice for $x > 0.4$, because of the numerous oscillations associated with undersmoothing.

3.4 Recursive Partitioning

For nonparametric regression problems with complicated structure a single global smoothing parameter is unsatisfactory. Specifically, kernel estimators can be improved by adapting to local curvature. There has been some progress with piecewise constant bandwidths for local-linear fitting and $\mathrm{AIC_C}$. One new approach uses a recursive partitioning (RP) to simultaneously determine both the intervals in the explanatory variable and the bandwidths used throughout the intervals. The result is a regression tree with separate $\hat{h}$ values used over adaptively selected regions in the predictor variable. We denote these bandwidths by h_{RP}.

Consider local-linear regression estimates of $\mathbf{m} = (m(x_1), \ldots, m(x_n))^T$, the regression function at each of the observed predictors $\mathbf{x} = (x_1, \ldots, x_n)^T$. It was noted by Hurvich, Simonoff and Tsai (1998) that local-linear regression is a linear smoother, that is, $\hat{\mathbf{m}}(h_{\mathrm{RP}}) = H(h_{\mathrm{RP}})\mathbf{y}$, where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the response vector and $H(h_{\mathrm{RP}})$ is the *smoother* (or *hat*) *matrix* that results from evaluating (12) at each x_i. Thus we have what one may think of as expected log-likelihood plus a penalty for the effective number of parameters,

$$\mathrm{AIC_C} = \log(\hat{\sigma}^2) + \psi_n\{\mathrm{tr}(H(h_{\mathrm{RP}}))\}, \tag{14}$$

where $\hat{\sigma}^2$ is the MSE of the residuals at the observed predictors, and $\Psi_n(t) = [1 + t/n]/[1 - (t+2)/n]$ for $0 < t < n - 2$ and $= \infty$ otherwise is the penalty function applied to the trace of the smoother matrix. In an obvious parallel with linear regression, the quantity $\mathrm{tr}(H)$ can be interpreted as an effective number of parameters, a measure of model complexity. See Section 7.6 of Hastie, Tibshirani and Friedman (2001). This reflects the "roughness" of the estimated curve in the sense of a more complex basis. As the global bandwidth decreases, $\hat{\sigma}^2$ decreases and the estimates are less biased, whereas the $\mathrm{tr}(H(h))$ and $\psi_n\{\mathrm{tr}(H(h))\}$ increase as the estimated curves become less smooth. Pitblado (2000) demonstrated this behavior of $\mathrm{AIC_C}$ for global bandwidth choice. Figure 6 illustrates one such result.

Consider curve estimates using simulated data from the function m_2, which was defined in Section 3. Figure 6 shows the true regression function and the local-linear fit with a global bandwidth. The estimated curve with a global bandwidth $h = 0.040$ exhibits under-

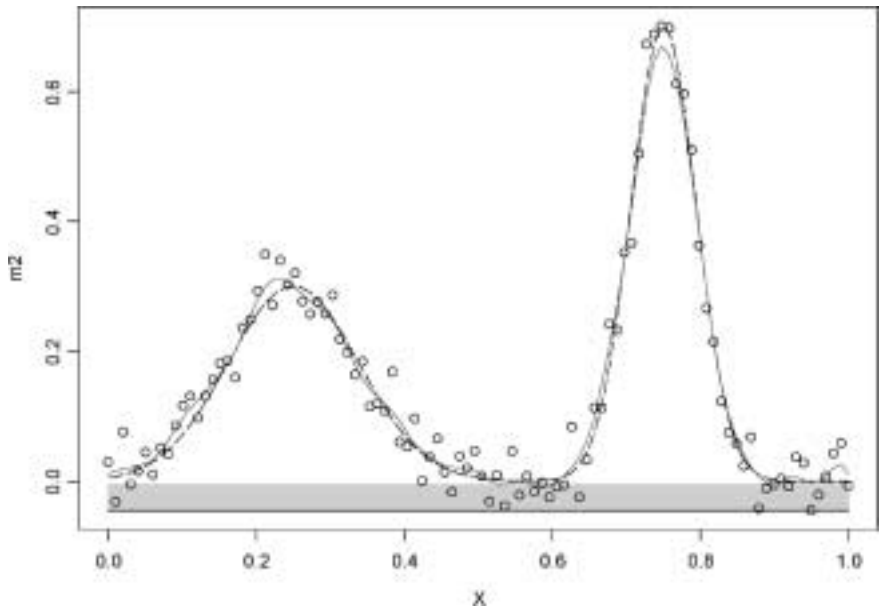

FIG. 6. *The true regression function m_2 (dashed) and the local-linear fit (solid) with the AIC_C minimizing global bandwidth (shaded region along the x axis).*

smoothing of the left half and oversmoothing of the right mode.

How do variable bandwidths, in particular piecewise constant bandwidths, help us in this respect? Some improvements are obvious in the estimated curve with variable bandwidths in Figure 7. The new method found a partition of two with a split at $x = 0.66$ and a variable bandwidth $h_{\mathrm{RP}} = (h_1, h_2)$, where $h_1 = 0.078$ is for the left subinterval and $h_2 = 0.033$ is for the right subinterval. Regression trees are a recognized feature of the nonparametric landscape. Recursive partitioning for appropriate variable bandwidths is a successful new branch. For a full description of the methodology, see Jia and Schucany (2004), who also reported the re-

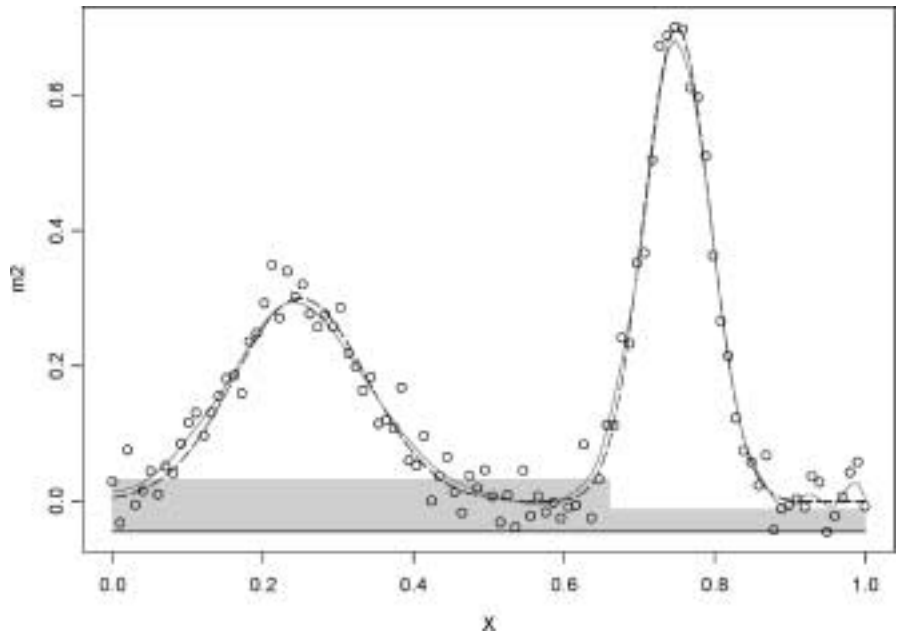

FIG. 7. *The true regression function m_2 (dashed) and the local-linear fit (solid) with two bandwidths, h_{RP} (shaded regions along the x axis), determined by recursive partitioning.*

sults of a Monte Carlo study of the effects of curvature change, sample size and signal-to-noise ratio.

4. CONFIDENCE INTERVALS

All of the linear statistics that we have considered [NW in (2), PC in (3) and LL in (12)] can be put in the form $\sum_i u_i(x_0, h)Y_i$. With nh sufficiently large, $\hat{m}(x_0)$ is approximately normal with mean $E[\hat{m}(x_0)]$ and $\mathrm{Var}[\hat{m}(x_0)] = \sigma^2 \sum_i u_i^2$. In some settings this would be enough to produce approximate $1-\alpha$ confidence intervals. Unfortunately for nonparametric curve estimation, the large-sample correctness of this does not hold.

Suppose that model (1) holds with constant variance σ^2. The naive interval

$$\hat{m}(x_0) \pm z_{\alpha/2}\hat{\sigma}\left[\sum_{i=1}^{n} u_i^2(x_0)\right]^{1/2} \tag{15}$$

with $\hat{\sigma}$ consistent for σ may not have the asymptotically correct coverage. That is, even though $\hat{\sigma} \to \sigma$ and $\hat{m}(x_0) \to m(x_0)$ when both $n \to \infty$ and $nh \to \infty$, the coverage of (15) need not converge to $1-\alpha$ as a confidence interval for $m(x_0)$. Hart (1997, Section 3.5) presented a formula for the limiting coverage for an interval based on GM in (11). The culprit is the large-sample behavior of the bias, which is not degenerate and offsets the proper normal interval. There have been several proposals to correct intervals based on a broad class of linear fits. Section 9.2 in Loader (1999) addressed these corrections to obtain both approximate pointwise confidence intervals and approximate simultaneous confidence bands.

5. OTHER WAYS TO DO THIS

There are some parallel and some different challenges in other approaches to fitting smooth models for the relationship of y to x. These others include splines and expansions in terms of basis functions, for example, wavelets or Fourier series; see Ramsay and Silverman (1997) and Hastie, Tibshirani and Friedman (2001). These alternative techniques involve selecting smoothing parameters, whether these are the number of basis functions in an expansion, the number of knots or explicit weights in the bias–variance trade-off. They all have their strengths and weaknesses in different settings, depending on the objective, the curvatures present in the unknown model, the signal-to-noise ratio, the sample size, the arrangement of the design points and so forth. However, local-linear kernel regression has the most direct interpretation in terms of familiar, intuitive, simple functions.

6. EXTENSIONS FOR KERNELS

Derivatives. Estimation of the νth derivative $m^{(\nu)}(x_0)$ was presented by Fan and Gijbels (1996) as one of the advantages of using a local polynomial of degree p. In Section 3.3 they analyzed asymptotic variance as a function of ν and p, and recommend that $p-\nu$ be odd. This is consistent with the general desirability of local linear ($p=1$) for the function $m(\cdot)$ for which $\nu = 0$.

Multivariate predictors. Ruppert (1997) proposed a local bandwidth selector for local polynomial fits that adapt easily to multidimensional explanatory variables. Fan and Gijbels (1996) devoted all of Chapter 7 to multivariate predictors. They discussed local polynomial univariate smoothers as the building blocks for a variety of approaches. Ultimately they provided details for the extension of local-linear regression to a d-dimensional explanatory $\mathbf{X}$.

Change-point analyses. This represents a fertile area for extensions to statistical methodology. The prototype model has a jump discontinuity in the mean function m. There are obvious parallels in higher dimensions, variance functions, transition probabilities and abrupt changes in the complexity of models [e.g. ARMA(p, q)]. In the prototypical situation a smoother is designed not to respond to such jumps. Wavelets seem to be better suited to reproducing such irregular (unsmooth) features. The use of kernel fits separately on each side of a candidate discontinuity in $m(x_0)$ has been investigated by Müller (1992), Loader (1996) and Gerard and Schucany (1997).

Dependent data. Hart and Lee (2005) addressed deficiencies of CV for bandwidth selection. Excellent coverage of the issues that arise for dependent data may be found in Lin, Wang, Welsh and Carroll (2004). Welsh, Lin and Carroll (2002) established a fundamental difference between kernels and splines in this setting. Their key finding is that splines have equivalent kernel representations when the additive model (1) has independence, but not when the ε_i are dependent. The essence of the difference is their "local" behavior in which kernels are local and splines are not. Furthermore, they concluded that there are compelling reasons to recommend efficient nonlocal splines for such applications.

ACKNOWLEDGMENTS

I am grateful to Professor Ron Randles and Professor Tom Hettmansperger for inviting me to contribute

to this special issue. I thank An Jia for assistance with the computing and the figures, Matt Wand for the data from Ullah (1985) and Dick Gunst for valuable comments on several drafts of the paper. Referee comments led to a much improved presentation.

I greatly appreciate the funding from Dr. Robert Haley for research on kernel smoothing of brain images. This study was supported by the U.S. Army Medical Research and Materiel Command Grant DAMD17-01-1-0741 through a consortium agreement with the University of Texas Southwestern Medical Center at Dallas. The content of this paper does not necessarily reflect the position or the policy of the U.S. government, and no official endorsement should be inferred.

REFERENCES

ALTMAN, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Amer. Statist.* **46** 175–185.

BARTLETT, M. S. (1963). Statistical estimation of density functions. *Sankhyā Ser. A* **25** 245–254.

BENEDETTI, J. K. (1977). On the nonparametric estimation of regression functions. *J. Roy. Statist. Soc. Ser. B* **39** 248–253.

CHU, C.-K. and MARRON, J. S. (1991). Choosing a kernel regression estimator (with discussion). *Statist. Sci.* **6** 404–436.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836.

EPANECHNIKOV, V. A. (1969). Nonparametric estimation of a multivariate probability density. *Theory Probab. Appl.* **14** 153–158.

EUBANK, R. L. (1999). *Nonparametric Regression and Spline Smoothing*, 2nd ed. Dekker, New York.

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004.

FAN, J. and GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57** 371–394.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman and Hall, London.

GASSER, T. and MÜLLER, H.-G. (1979). Kernel estimation of regression functions. *Smoothing Techniques for Curve Estimation*. *Lecture Notes in Math.* **757** 23–68. Springer, Heidelberg.

GERARD, P. D. and SCHUCANY, W. R. (1997). Locating exotherms in differential thermal analysis with nonparametric regression. *J. Agric. Biol. Environ. Stat.* **2** 255–268.

HART, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.

HART, J. D. and LEE, C.-L. (2005). Robustness of one-sided cross-validation to autocorrelation. *J. Multivariate Anal.* **92** 77–96.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining, Inference, and Prediction*. Springer, New York.

HODGES, J. L., JR. and LEHMANN, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Ann. Math. Statist.* **27** 324–335.

HURVICH, C. M., SIMONOFF, J. S. and TSAI, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 271–293.

HURVICH, C. M. and TSAI, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76** 297–307.

JIA, A. and SCHUCANY, W. R. (2004). Recursive partitioning for kernel smoothers: A tree-based approach for estimating variable bandwidths in local linear regression. Unpublished manuscript.

LIN, X., WANG, N., WELSH, A. H. and CARROLL, R. J. (2004). Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika* **91** 177–193.

LOADER, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist.* **24** 1667–1678.

LOADER, C. R. (1999). *Local Regression and Likelihood*. Springer, New York.

MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231–238.

MÜLLER, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.* **20** 737–761.

NADARAYA, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theory Probab. Appl.* **10** 186–190.

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.* **33** 1065–1076.

PITBLADO, J. (2000). Estimating partially variable bandwidths in local linear regression using an information criterion. Ph.D. dissertation, Dept. Statistical Science, Southern Methodist Univ.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Nonparametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385–392.

RAMSAY, J. O. and SILVERMAN, B. W. (1997). *Functional Data Analysis*. Springer, New York.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

RUPPERT, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **92** 1049–1062.

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press.

SCHLEE, W. (1988). Regressograms. *Encyclopedia of Statistical Sciences* **8** 1–3. Wiley, New York.

SCOTT, D. W. (1992). *Multivariate Density Estimation*. Wiley, New York.

SHEATHER, S. (2005). Density estimation. *Statist. Sci.* **19** 588–597.

SIGNORINI, D. F. and JONES, M. C. (2004). Kernel estimators for univariate binary regression. *J. Amer. Statist. Assoc.* **99** 119–126.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

SPENCER, J. (1904). On the graduation of rates of sickness and mortality. *J. Institute of Actuaries* **38** 334–343.

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

TUKEY, J. W. (1961). Curves as parameters and touch estimation. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 681–694. Univ. California Press, Berkeley.

ULLAH, A. (1985). Specification analysis of econometric models. *J. Quantitative Economics* **1** 187–209.

WAHBA, G. and WOLD, S. (1975). A completely automatic French curve. *Comm. Statist.* **4** 1–17.

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.

WELSH, A. H., LIN, X. and CARROLL, R. J. (2002). Marginal longitudinal nonparametric regression: Locality and efficiency of spline and kernel methods. *J. Amer. Statist. Assoc.* **97** 482–493.