COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY

Box 2125, Yale Station
New Haven, Connecticut 06520

COWLES FOUNDATION DISCUSSION PAPER NO. 1069

APPLIED NONPARAMETRIC METHODS

Wolfgang Härdle and Oliver Linton

March 1994

Applied Nonparametric Methods*

Wolfgang Härdle[†]
Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin
D 10178 Berlin, Germany

Oliver Linton[‡]
Department of Economics
Yale University
New Haven, CT 06520

March 10, 1994

Abstract

We review different approaches to nonparametric density and regression estimation. Kernel estimators are motivated from local averaging and solving ill-posed problems. Kernel estimators are compared to k-NN estimators, orthogonal series and splines. Pointwise and uniform confidence bands are described, and the choice of smoothing parameter is discussed. Finally, the method is applied to nonparametric prediction of time series and to semiparametric estimation.

[*] Forthcoming in *The Handbook of Econometrics, Volume IV.* Eds R.F. Engle and D.F. McFadden. North Holland.

[†] This work was prepared while the first author was visiting CentER, KUB Tilburg, The Netherlands. It was financed, in part, by contract No 26 of the programme "Pôle d'attraction interuniversitaire" of the Belgian government.

[‡] We would like to thank Don Andrews, Roger Koenker, Jens Perch Nielsen, Tom Rothenberg and Richard Spady for helpful comments. Without the careful typewriting of Mariette Huysentruit and the skillful programming of Marlene Müller this work would not have been possible.

1 Nonparametric Estimation in Econometrics

Although economic theory generally provides only loose restrictions on the distribution of observable quantities, much econometric work is based on tightly specified parametric models and likelihood based methods of inference. Under regularity conditions, maximum likelihood estimators consistently estimate the unknown parameters of the likelihood function. Furthermore, they are asymptotically normal (at convergence rate the square root of the sample size) with limiting variance matrix that is minimal according to the Cramer-Rao theory. Hypothesis tests constructed from the likelihood ratio, Wald or Lagrange multiplier principle have therefore maximum local asymptotic power. However, when the parametric model is not true, these estimators may not be fully efficient, and in many cases – for example in regression when the functional form is misspecified – may not even be consistent. The costs of imposing the strong restrictions required for parametric estimation and testing can be considerable. Furthermore, as McFadden says in his 1985 presidential address to the Econometric society, the parametric approach

> *"interposes an untidy veil between econometric analysis and the propositions of economic theory, which are mostly abstract without specific dimensional or functional restrictions."*

Therefore, much effort has gone into developing procedures that can be used in the absence of strong *a priori* restrictions. This survey examines nonparametric smoothing methods which do not impose parametric restrictions on functional form. We put emphasis on econometric applications and implementations on currently available computer technology.

There are many examples of density estimation in econometrics. Income distributions – see Hildenbrand and Hildenbrand (1986) – are of interest with regard to welfare analysis, while the density of stock returns has long been of interest to financial economists following Mandelbrot (1963) and Fama (1965). Figure 1 shows a density estimate of the stock return data of Pagan and Schwert (1990) in comparison with a normal density. We include a bandwidth factor in the scale parameter to correct for the finite sample bias of the kernel method.

Regression smoothing methods are used frequently in demand analysis – see for example Deaton (1991), Banks, Blundell, and Lewbel (1993) and Hausman and Newey (1992). Figure 2 shows a nonparametric kernel regression estimate of the statistical Engel curve for food expenditure and total income. For comparison the (parametric) Leser curve is also included.

There are four main uses for nonparametric smoothing procedures. Firstly, they can be employed as a convenient and succinct means of displaying the features of a dataset and hence to aid practical parametric model building. Secondly, they can be used for

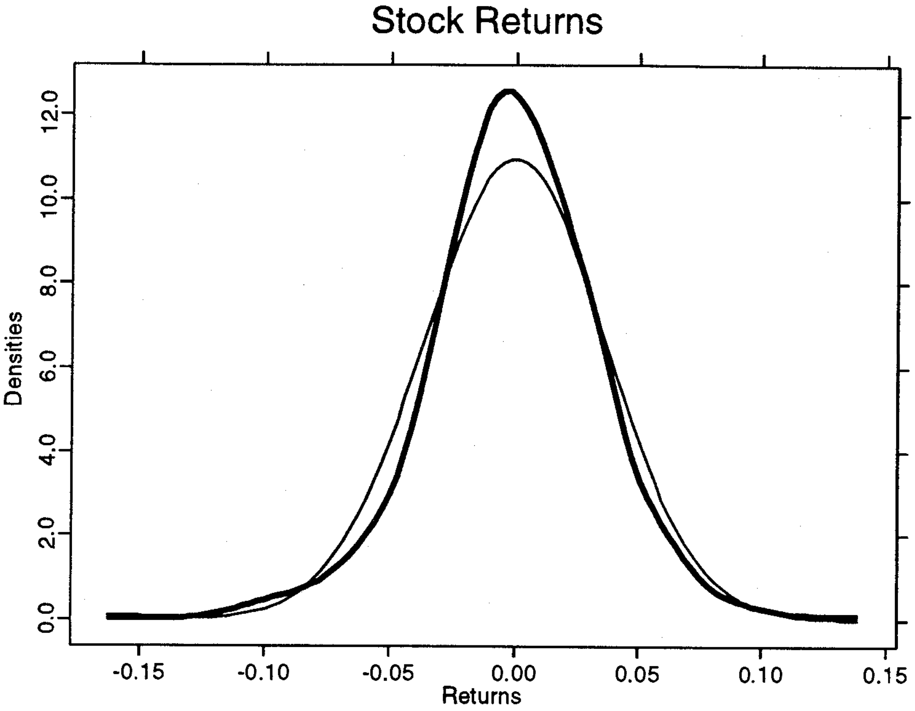

Figure 1. Density estimator of stock returns of Pagan and Schwert data compared with a mean zero normal density (thin line) with standard deviation $\sqrt{\hat{\sigma}^2 + \hat{h}^2}$, $\hat{\sigma} = 0.035$ *and* $\hat{h} = 0.009$, *both evaluated at a grid of 100 equispaced points. Sample size was 1104. The bandwidth* $\hat{h}$ *was determined by the XploRe macro* denauto *according to Silverman's rule of thumb method.*

diagnostic checking of an estimated parametric model. Thirdly, one may want to conduct inference under only the very weak restrictions imposed in fully nonparametric structures. Finally, nonparametric estimators are frequently required in the construction of estimators of Euclidean-valued quantities in semiparametric models.

By using smoothing methods one can broaden the class of structures under which the chosen procedure gives valid inference. Unfortunately, this robustness is not free. Centered nonparametric estimators converge at rate $\sqrt{nh}$, where $h \to 0$ is a smoothing parameter, which is slower than the $\sqrt{n}$ rate for parametric estimators in correctly specified models. It is also sometimes suggested that the asymptotic distributions themselves can be poor approximations in small samples. However, this problem is also found in parametric situations. The difference is quantitative rather than qualitative: typically, centered nonparametric estimators behave similarly to parametric ones in which n has been replaced by nh. The closeness of the approximation is investigated further in Hall (1992).

Smoothing techniques have a long history starting at least in 1857 when the Saxonian economist Engel found the law named after him. He analyzed Belgian data on household expenditure, using what we would now call the regressogram. Whittaker (1923) used a graduation method for regression curve estimation which one would now call spline smoothing. Nadaraya (1964) and Watson (1964) provided an extension for general random design based on kernel methods. In time series, Daniell (1946) introduced the smoothed periodogram for consistent estimation of the spectral density. Fix and Hodges (1951) extended this for the estimation of a probability density. Rosenblatt (1956) proved

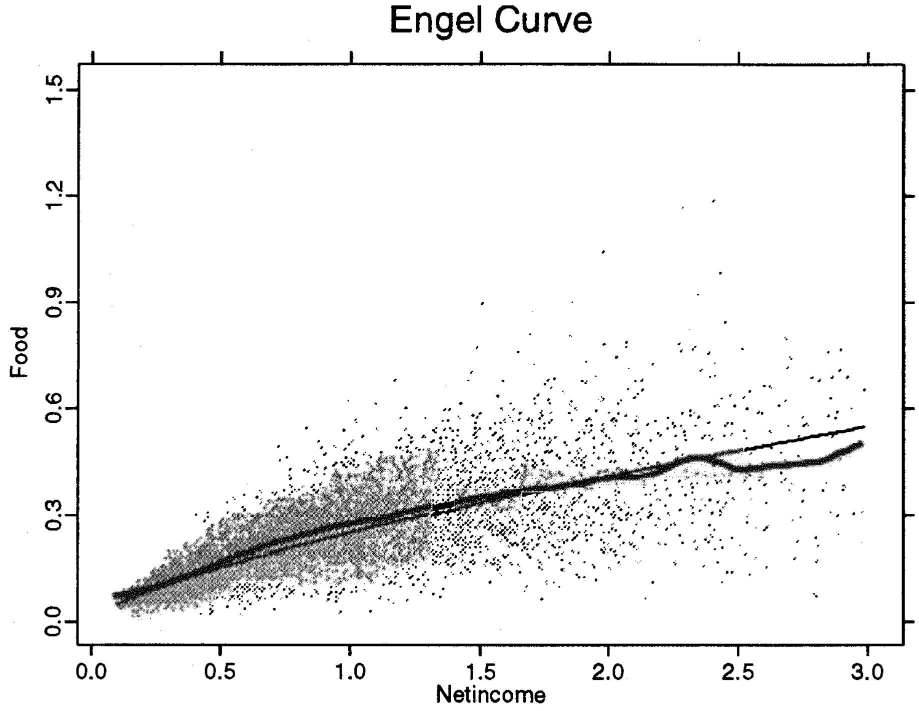

Figure 2. A kernel regression smoother applied to the food expenditure as a function of total income. Data from the Family Expenditure Survey (1968-1983), year 1973. Quartic kernel, bandwidth h=0.2. The data have been normalized by mean income. Standard deviation of net income is 0.544. The kernel has been computed using the XploRe macro `regest`.

asymptotic consistency of the kernel density estimator.

These methods have developed considerably in the last ten years, and are now frequently used by applied econometricians – see the recent survey by Deaton (1993). The massive increase in computing power as well as the increased availability of large cross-sectional and high-frequency financial time-series datasets are partly responsible for the popularity of these methods. They are typically simple to implement in software like GAUSS or XploRe (1993).

We base our survey of these methods around kernels. All the techniques we review for nonparametric regression are linear in the data, and thus can be viewed as kernel estimators with a certain equivalent weighting function. Since smoothing parameter selection methods and confidence intervals have been mostly studied for kernels, we feel obliged to concentrate on these methods as the basic unit of account in nonparametric smoothing.

2 Density Estimation

It is simplest to describe the nonparametric approach in the setting of density estimation, so we begin with that. Suppose we are given i.i.d. real-valued observations $\{X_i\}_{i=1}^{n}$ with density f. Sometimes – for the crossvalidation algorithm described in Section 4 and for semiparametric estimation – it is required to estimate f at each sample point, while on other occasions it is sufficient to estimate at a grid of points $x_1, .., x_M$ for M fixed. We shall for the most part restrict our attention to the latter situation, and in particular concentrate on estimation at a single point x.

Below we give two approaches to estimating $f(x)$.

2.1 Kernels as windows

If f is smooth in a small neighborhood $[x-h, x+h]$ of x, we can justify the following approximation:

$$2h \cdot f(x) \approx \int_{x-h}^{x+h} f(u)du = P(X \in [x-h, x+h]), \tag{1}$$

by the mean value theorem. The right-hand side of (1) can be approximated by counting the number of X_i's in this small interval of length $2h$, and then dividing by n. This is a histogram estimator with *bincenter* x and *binwidth* $2h$. Let $K(u) = \frac{1}{2}\mathrm{I}(|u| \leq 1)$, where $\mathrm{I}(\bullet)$ is the indicator function taking the value 1 when the event is true and zero otherwise. Then the histogram estimator can be written as

$$\hat{f}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i), \tag{2}$$

where $K_h(\bullet) = h^{-1}K(\bullet/h)$. This is also a kernel density estimator of $f(x)$ with kernel $K(u) = \frac{1}{2}\mathrm{I}(|u| \leq 1)$ and *bandwidth* h.

The step function kernel weights each observation inside the window equally, even though observations closer to x should possess better information than more distant ones. In addition, the step function estimator is discontinuous in x, which is unattractive given the smoothness assumption on f. Both objectives can be satisfied by choosing a smoother "window function" K as kernel, i.e. one for which $K(u) \to 0$ as $|u| \to 1$. One example is the so-called quartic kernel

$$K(u) = \frac{15}{16}(1-u^2)^2\mathrm{I}(|u| \leq 1). \tag{3}$$

In the next section we give an alternative motivation for kernel estimators. The less technically able reader may skip this section.

2.2 Kernels and ill-posed problems

An alternative approach to estimation of f is to find the best smooth approximation to the empirical distribution function and to take its derivative.

The distribution function F is related to f by

$$Af(x) = \int_{-\infty}^{\infty} \mathrm{I}(u \leq x)\ f(u)du = F(x), \tag{4}$$

which is called a Fredholm equation with integral operator $Af(x) = \int_{-\infty}^{x} f(u)du$. Recovering the density from the distribution function is the same as finding the inverse of the operator A. In practice, we must replace the distribution function by the empirical distribution function (edf) $F_n(x) = n^{-1}\sum_{i=1}^{n}\mathrm{I}(X_i \leq x)$, which converges to F at rate $\sqrt{n}$. However, this is a step function and cannot be differentiated to obtain an approximation to $f(x)$. Put another way, the Fredholm problem is ill-posed since for a sequence F_n tending to F, the solutions (satisfying $Af_n = F_n$) do not necessarily converge to f: the inverse operator in (4) is not continuous, see Vapnik (1982, p. 22).

Solutions to ill-posed problems can be obtained using the Tikhonov (1963) regularization method. Let $\Omega(f)$ be a lower semicontinuous functional called the *stabilizer*. The idea of the regularization method is to find indirectly a solution to $Af = F$ by use of the stabilizer. Note that the solution of $Af = F$ minimizes (with respect to $\hat{f}$)

$$\int_{-\infty}^{\infty} \left[\mathrm{I}(u \geq x)\ \hat{f}(u)du - F(x)\right]^2 dx.$$

The stabilizer $\Omega(\hat{f}) = \|\hat{f}\|^2$ is now added to this equation with a Lagrange parameter λ,

$$R_\lambda(\hat{f}, F) = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \mathrm{I}(x \geq u)\ \hat{f}(u)du - F(x)\right]^2 dx + \lambda \int_{-\infty}^{\infty} \hat{f}^2(u)du. \qquad (5)$$

Since we do not know $F(x)$, we replace it by the edf $F_n(x)$ and obtain the problem of minimizing the functional $R_\lambda(\hat{f}, F_n)$ with respect to $\hat{f}$.

A necessary condition for a solution $\hat{f}$ is

$$\int_{-\infty}^{\infty} \mathrm{I}(x \geq u) \left[\int_{-\infty}^{\infty} \mathrm{I}(x \geq s)\ \hat{f}(s)ds - F_n(x)\right] dx + \lambda\hat{f}(u) = 0.$$

Applying the Fourier transform for generalized functions and noting that the Fourier transform of $\mathrm{I}(u \geq 0)$ is $\frac{\mathrm{i}}{w} + \pi\delta(\omega)$ (with $\delta(\bullet)$ the delta function), we obtain

$$\left(\frac{1}{\mathrm{i}\omega}\right)\left[\left(-\frac{1}{\mathrm{i}\omega}\right)\Gamma(\omega) - n^{-1}\sum_{i=1}^{n}\left(-\frac{e^{\mathrm{i}\omega X_i}}{\mathrm{i}\omega}\right)\right] + \lambda\Gamma(\omega) = 0,$$

where Γ is the Fourier transform of $\hat{f}$. Solving this equation for Γ and then, applying the inverse Fourier transform, we obtain

$$\hat{f}(x) = n^{-1}\sum_{i=1}^{n} \frac{1}{2\sqrt{\lambda}} e^{|x - X_i|/\sqrt{\lambda}}.$$

Thus we obtain a kernel estimator with kernel $K(u) = \frac{1}{2}\exp(-|u|)$ and bandwidth $h = \sqrt{\lambda}$. More details are given in Vapnik (1982, p. 302).

2.3 Properties of kernels

In the first two sections we derived different approaches to kernel smoothing. Here we would like to collect and summarize some properties of kernels. A *kernel* is a piecewise continuous function, symmetric around zero, integrating to one:

$$K(u) = K(-u)\ ;\ \int K(u)du = 1. \qquad (6)$$

It need not have bounded support, although many commonly used kernels live on $[-1, 1]$. In most applications K is a positive probability density function, however for theoretical reasons it is sometimes useful to consider kernels that take on negative values. For any integer j, let

$$\mu_j(K) = \int u^j K(u)du\ ;\ \nu_j(K) = \int K(u)^j du.$$

Kernel	$K(u)$	$D(K_{\text{opt}}, K)$
Epanechnikov	$(3/4)(1-u^2)\ \mathrm{I}(\lvert u\rvert) \le 1)$	1
Quartic	$(15/16)(1-u^2)^2\ \mathrm{I}(\lvert u\rvert) \le 1)$	1.005
Triangular	$(1-\lvert u\rvert)\ \mathrm{I}(\lvert u\rvert \le 1)$	1.011
Gauss	$(2\pi)^{-1/2}\exp(-u^2/2)$	1.041
Uniform	$(1/2)\ \mathrm{I}(\lvert u\rvert \le 1)$	1.060

Table 1. Common kernel functions.

The order p of a kernel is defined as the first nonzero moment,

$$\mu_j = 0,\ j = 1, .., p-1\ ;\ \mu_p \neq 0. \tag{7}$$

We mostly restrict our attention to positive kernels which can be at most of order 2. An example of a higher order kernel (of order 4) is

$$K(u) = \frac{15}{32}(7u^4 - 10u^2 + 3)\mathrm{I}(|u| \le 1).$$

A list of common kernel functions is given below; we shall comment later on the values in the third column.

2.4 Properties of the Kernel Density Estimator

The kernel estimator is a sum of iid random variables, and therefore

$$E\left[\hat{f}_h(x)\right] = \int K_h(x-z)f(z)dz = K_h * f(x), \tag{8}$$

where $*$ denotes convolution, assuming the integral exists. When f is $N(0,\sigma^2)$ and K is standard normal, $E\left[\hat{f}_h(x)\right]$ is therefore the normal density with standard deviation $\sqrt{\sigma^2 + h^2}$ evaluated at x, see Silverman (1986, p37). This explains our modification to the normal density in Figure 1.

More generally, it is necessary to approximate $E\left[\hat{f}_h(x)\right]$ by a Taylor series expansion. Firstly, we change variables

$$E\left[\hat{f}_h(x)\right] = \int K(u)f(x-uh)du. \tag{9}$$

Then expanding $f(x-uh)$ about $f(x)$ gives

$$E\left[\hat{f}_h(x)\right] = f(x) + \frac{h^2}{2}\mu_2(K)f''(x) + o(h^2), \tag{10}$$

provided $f''(x)$ is continuous in a neighborhood of x. Therefore, the bias of $\hat{f}_h(x)$ is $O(h^2)$ as $h \to 0$.

By similar calculation,

$$Var\left[\hat{f}_h(x)\right] \approx \frac{1}{nh}\nu_2(K)f(x), \tag{11}$$

see Silverman (1986, p38). Therefore, provided $h \to 0$ and $nh \to \infty$, $\widehat{f}_h(x) \xrightarrow{P} f(x)$. Further asymptotic properties of the kernel density estimator are given in Prakasa Rao (1983).

The statistical properties of $\widehat{f}_h(x)$ depend closely on the bandwidth h: the bias increases and the variance decreases with h. We investigate how the estimator itself depends on the bandwidth with the income data of Figure 2. Figure 3a shows a kernel density estimate for the income data with bandwidth $h = 0.2$ computed using the quartic kernel (3) and evaluated at a grid of 100 equispaced points. There is a clear bimodal structure for this implementation. A larger bandwidth $h = 0.4$ creates a single moded structure as shown in Figure 3b, while a smaller $h = 0.05$ results in Figure 3c where in addition to the bimodal feature there is considerable small scale variation in the density.

It is therefore important to have some method of choosing h. This problem has been heavily researched – see Jones, Marron, and Sheather (1992) for a collection of recent results and discussion. We take up the issue of automatic bandwidth selection in greater detail for the regression case in Section 4.2. We mention here one method that is frequently used in practice – Silverman's rule of thumb. Let $\widehat{\sigma}^2$ be the sample variance of the data. Silverman (1986) proposed choosing the bandwidth to be

$$h = 1.364 \left\{ \frac{\nu_2(K)}{\mu_2^2(K)} \right\}^{1/5} \widehat{\sigma} n^{-1/5}.$$

This rule is optimal (according to $IMSE$ – see Section 4 below) for the normal density, and is not far from optimal for most symmetric, unimodal densities. This procedure was used to select h in Figure 1.

2.5 Estimation of Multivariate Densities, their Derivatives and Bias Reduction

A multivariate (d-dimensional) density function f can be estimated by the kernel estimator

$$\widehat{f}_H(x) = \frac{1}{n} \sum_{i=1}^{n} k_H(x - X_i), \tag{12}$$

where $k_H(\bullet) = \{\det(H)\}^{-1} k(H^{-1}\bullet)$, where $k(\bullet)$ is a d-dimensional kernel function, while H is a d by d bandwidth matrix. A convenient choice in practice is to take $H = hS^{1/2}$, where S is the sample covariance matrix and h is a scalar bandwidth sequence, and to give k a product structure, i.e. let $k(u) = \prod_{j=1}^{d} K(u_j)$, where $u = (u_1, .., u_d)^T$ and $K(\bullet)$ is a univariate kernel function.

Partial derivatives of f can be estimated by the appropriate partial derivatives of $\widehat{f}_H(x)$ (providing $k(\bullet)$ has the same number of nonzero continuous derivatives). For any d-vector $r = (r_1, .., r_d)^T$ and any function $g(\bullet)$ define

$$g^{(r)}(x) = \frac{\partial^{|r|}}{\partial^{r_1} x_1 .. \partial^{r_d} x_d} g(x),$$

where $|r| = \sum_{j=1}^{d} r_j$, then $\widehat{f}_H^{(r)}(x)$ estimates $\widehat{f}(x)$.

The properties of multivariate derivative estimators are described in Prakasa Rao (1983, pp237). In fact, when a bandwidth $H = hA$ is used, where h is scalar and A is any

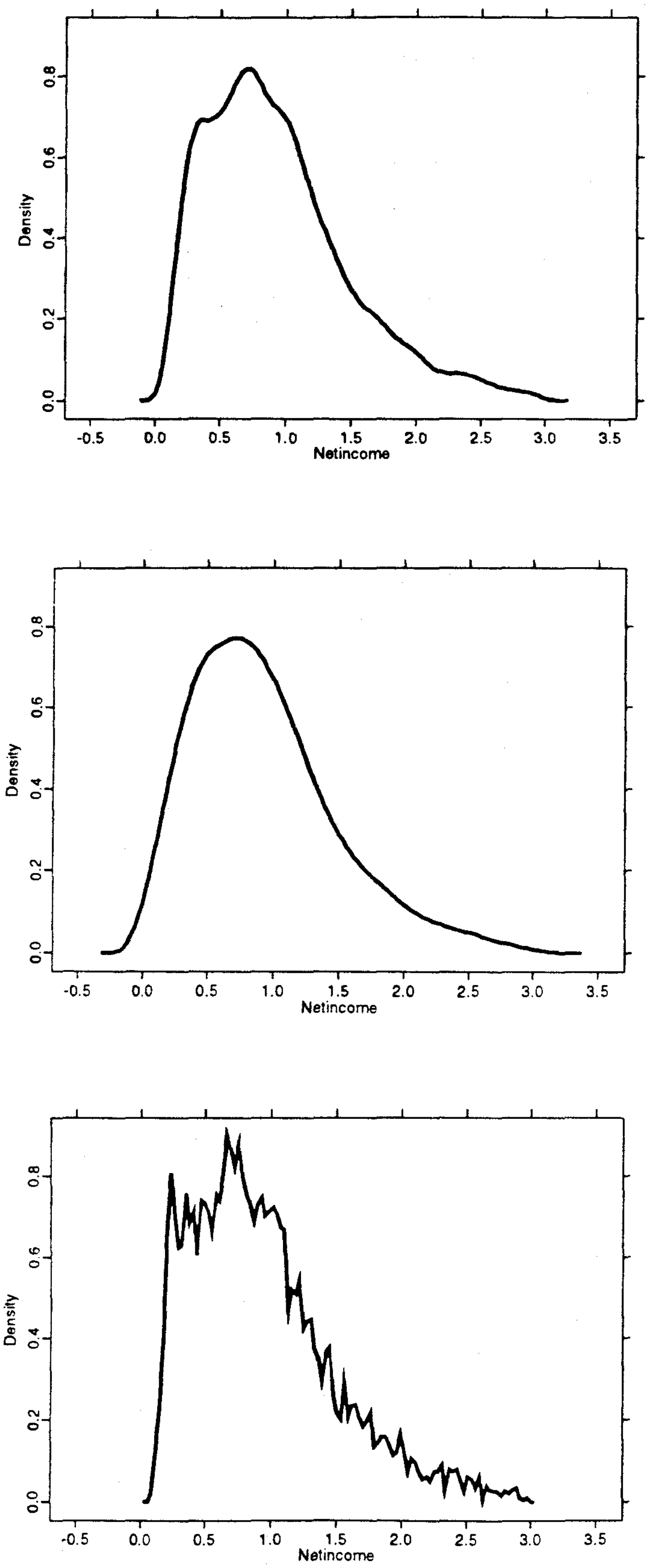

Figure 3abc. Kernel density estimates of net income distribution. Family Expenditure Survey (1968-1983). XploRe macro **denest**. *Year 1973.*

fixed positive definite d by d matrix, then $Var[\widehat{f}_H^{(r)}(x)] = O(n^{-1}h^{-(2|r|+d)})$, while the bias is $O(h^2)$. For a given bandwidth h, the variance increases with the number of derivatives being estimated and with the dimensionality of X. The latter effect is well known as the *curse of dimensionality.*

It is possible to improve the order of magnitude of the bias by using a $p'th$ order kernel, where $p > 2$. In this case, the Taylor series expansion argument shows that $E\left[\widehat{f}_h(x)\right] - f(x) = O(h^p)$, when p is an even integer. Unfortunately, with this method there is the possibility of a negative density estimate, since K must be negative somewhere. Abramson (1982) and Jones, Linton and Nielsen (1993) define *bias reduction* techniques that ensure a positive estimate. Jones and Foster (1993) review a number of other bias reduction methods.

The merits of bias reduction methods are based on asymptotic approximations. Marron and Wand (1992) derive exact expressions for the first two moments of higher order kernel estimators in a general class of mixture densities and find that unless very large samples are used, these estimators may not perform as well as the asymptotic approximations suggest. Unless otherwise stated, we restrict our attention to second order kernel estimators.

2.6 Fast Implementation of Density Estimation

Fast evaluation of (2) is especially important for optimization of the smoothing parameter which topic will be treated in Section 4.2. If the kernel density estimator has to be computed at each observation point for k different bandwidths, the number of calculations are $O(n^2hk)$ for kernels with bounded support. For the family expenditure dataset of Figure 1 with about 7000 observations this would take too long for the type of interactive data analysis we envisage. To resolve this problem we introduce the idea of discretization. The method is to map the raw data onto an equally spaced grid of smaller cardinality. All subsequent calculations are performed on this data summary which results in considerable computational saving.

Let $H_l(x;\Delta)$, $l = 0, 1, .., M-1$, be the $l'th$ histogram estimator of $f(x)$ with origin $\frac{l}{M}$ and small binwidth Δ. The sensitivity of histograms with respect to choice of origin is well known, see e.g. Härdle (1991, Fig 1.16). However, if histograms with different origins are then repeatedly averaged, the result becomes independent of the histogram origins. Let $\widehat{f}_{M,\Delta}(x) = \frac{1}{M}\sum_{l=0}^{M-1} H_l(x;\Delta)$ be the averaged histogram estimator. Then

$$\widehat{f}_{M,\Delta}(x) = \frac{1}{nh}\sum_{j\in\mathcal{Z}} \mathrm{I}(x \in B_j) \sum_{i=-M}^{M} n_{j-i}w_i, \tag{13}$$

where $\mathcal{Z} = \{.., -1, 0, 1, ..\}$, $B_j = \left[b_j - \frac{h}{2}, b_j + \frac{h}{2}\right]$ with $h = \Delta/M$ and $b_j = jh$, while $n_j = \sum_{i=1}^{n}\mathrm{I}(X_i \in B_j)$ and $w_i = \frac{M-|i|}{M}$. At the bincenters

$$\widehat{f}_{M,\Delta}(b_j) = \frac{1}{nh}\sum_{i=-M}^{M} n_{j-i}w_i.$$

Note that $\{w_i\}_{i=-M}^{M}$ is in fact a discrete approximation to the (rescaled) triangular kernel $K(u) = (1-|u|)\mathrm{I}(|u| \le 1)$. More generally, weights w_i can be used that represent the discretization of any kernel K. When K is supported on $[-1, 1]$, w_i is the rescaled evaluation of K at the points $\frac{-i}{M}$ $(i = -M, .., M)$. If a kernel with non-compact support is used, as the Gaussian for example, it is necessary to truncate the kernel function. Figure 4 shows the weights chosen from the quartic kernel with $M = 5$.

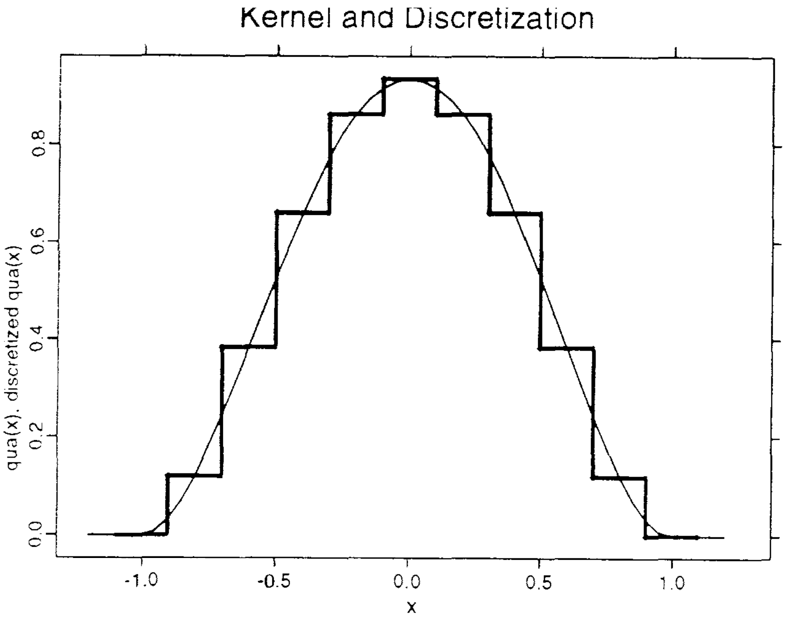

Figure 4. The quartic kernel $qua(u) = \frac{15}{16}(1-u^2)^2 I(|u| \leq 1)$. *Discretizing the kernel (without rescaling) leads to* $w_{-i} = qua(i/M)$, $i = -M, .., M$. *Here* $M = 5$ *was chosen. The weights are represented by the thick step function.*

Since (13) is essentially a convolution of the discrete kernel weights w_i with the bin-counts n_j, modern statistical languages such as GAUSS or XploRe that supply a convolution command are very convenient for computation of (13). Binning the data takes exactly n operations. If C denotes the number of nonempty bins, then evaluation of the binned estimator at the nonempty bins requires $O(MC)$ operations. In total we have a computational cost of $O(n + kM_{\max}C)$ operations for evaluating the binned estimator at k bandwidths, where $M_{\max} = Max\,\{M_j; j = 1, .., k\}$. This is a big improvement.

The discretization technique also works for estimating derivatives and multivariate densities, see Härdle and Scott (1990) and Turlach (1992). This method is basically a time domain version of the Fast Fourier Transform computational approach advocated in Silverman (1986), see also Jones (1989).

3 Regression Estimation

The most common method for studying the relationship between two variables X and Y is to estimate the conditional expectation function $m(x) = E(Y \mid X = x)$. Suppose that

$$Y_i = m(X_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{14}$$

where ϵ_i is an independent random error satisfying $E(\epsilon_i \mid X_i = x) = 0$, and $Var(\epsilon_i \mid X_i = x) = \sigma^2(x)$. In this section we restrict our attention to independent sampling, but some extensions to the dependent sampling case are given in Section 5. The methods we consider are appropriate for both *random design*, where (X_i, Y_i) are i.i.d, and *fixed design*, where X_i are fixed in repeated samples. In the random design case, X is an ancillary statistic, and standard statistical practice – see Cox and Hinkley (1974) – is to conduct inference conditional on the sample $\{X_i\}_{i=1}^n$. However, many papers in the literature prove theoretical properties unconditionally, and we shall, for ease of exposition, present results in this form. We also quote most results only for the case where X is scalar, although where appropriate we describe the extension to multivariate data.

In some cases, it is convenient to restrict attention to the equispaced design sequence $X_i = i/n$, $i = 1, .., n$. Although this is unsuitable for most econometric applications, there are situations where it is of interest: specifically, time itself is conveniently described in this way. Also, the relative ranks of any variable (within a given sample) are naturally equispaced – see Anand, Harris, and Linton (1993).

The estimators of $m(x)$ we describe are all of the form $\sum_{i=1}^n W_{ni}(x)Y_i$ for some weighting sequence $\{W_{ni}(x)\}_{i=1}^n$, but arise from different motivations and possess different statistical properties.

3.1 Kernel Estimators

Given the technique of kernel density estimation, a natural way to estimate $m(\bullet)$ is first to compute an estimate of the joint density $f(x, y)$ of (X, Y), and then to integrate it according to the formula

$$m(x) = \frac{\int y f(x,y)dy}{\int f(x,y)dy}. \tag{15}$$

The kernel density estimate $\widehat{f}_h(x, y)$ of $f(x, y)$ is

$$\widehat{f}_h(x,y) = n^{-1} \sum_{i=1}^n K_h(x - X_i)K_h(y - Y_i),$$

and by (6):

$$\int \widehat{f}_h(x,y)dy = n^{-1} \sum_{i=1}^n K_h(x - X_i) \ ; \int y\widehat{f}_h(x,y)dy = n^{-1} \sum_{i=1}^n K_h(x - X_i)Y_i.$$

Plugging these into numerator and denominator of (15) we obtain the Nadaraya–Watson kernel estimate

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^n K_h(x - X_i)Y_i}{\sum_{i=1}^n K_h(x - X_i)}. \tag{16}$$

The bandwidth h determines the degree of smoothness of $\widehat{m}_h$. This can be immediately seen by considering the limits for h tending to zero or to infinity, respectively. Indeed, at an observation X_i, $\widehat{m}_h(X_i) \to Y_i$, as $h \to 0$, while at an arbitrary point x, $\widehat{m}_h(x) \to \overline{Y}$, as $h \to \infty$. These two limit considerations make it clear that the smoothing parameter h in relation to the sample size n should not converge to zero too rapidly nor too slowly. Conditions for consistency of $\widehat{m}_h$ are given in the following theorem, proved in Schuster (1972):

Theorem 1. *Let $K(\bullet)$ satisfy $\int |K(u)|du \leq \infty$ and $Lim_{|u|\to\infty}\, uK(u) = 0$. Suppose also that $m(x)$, $f(x)$, and $\sigma^2(x)$ are continuous at x, and $f(x) > 0$. Then, provided $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$, we have*

$$\widehat{m}_h(x) \xrightarrow{P} m(x).$$

The kernel estimator is asymptotically normal, as was first shown in Schuster (1972).

Theorem 2. *Suppose in addition to the conditions of Theorem 1 that $\int |K(u)|^{2+\eta}du < \infty$, for some $\eta > 0$. Suppose also that $m(x)$ and $f(x)$ are twice continuously differentiable at x and that $E(|Y|^{2+\eta} \mid x)$ exists and is continuous at x. Finally, suppose that $\overline{Lim}$ $h^5 n < \infty$. Then*

$$\sqrt{nh}\left[\widehat{m}_h(x) - m(x) - h^2 B_{nw}(x)\right] \Rightarrow N(0, V_{nw}(x)),$$

where

$$B_{nw}(x) = \tfrac{1}{2}\mu_2(K)\left[m''(x) + 2m'(x)\tfrac{f'}{f}(x)\right]$$
$$V_{nw}(x) = \nu_2(K)\sigma^2(x)/f(x).$$

The Nadaraya-Watson estimator has an obvious generalization to d-dimensional explanatory variables and $p'th$ order kernels. In this case, assuming a common bandwidth h is used, the (asymptotic) bias is $O(h^p)$, when p is an even integer, while the (asymptotic) variance is $O(n^{-1}h^{-d})$.

3.2 k-Nearest Neighbor Estimators

3.2.1 Ordinary k-NN Estimators

The kernel estimate was defined as a weighted average of the response variables in a fixed neighborhood of x. The k-nearest neighbor (k-NN) estimate is defined as a weighted average of the response variables in a varying neighborhood. This neighborhood is defined through those X-variables which are among the k-nearest neighbors of a point x.

Let $\mathcal{N}(x) = \{i : X_i \text{ is one of the } k\text{-NN to } x\}$ be the set of indices of the k-nearest neighbors of x. The k-NN estimate is the average of Y's with index in $\mathcal{N}(x)$,

$$\widehat{m}_k(x) = \frac{1}{k}\sum_{i\in\mathcal{N}(x)} Y_i. \tag{17}$$

Connections to kernel smoothing can be made by considering (17) as a kernel smoother with uniform kernel $K(u) = \frac{1}{2}\mathrm{I}(|u| \leq 1)$ and variable bandwidth $h = R(k)$, the distance between x and its furthest k-NN,

$$\widehat{m}_k(x) = \frac{\sum_{i=1}^n K\left(\frac{x-X_i}{R}\right)Y_i}{\sum_{i=1}^n K\left(\frac{x-X_i}{R}\right)}. \tag{18}$$

Note that in (18), for this specific kernel, the denominator is equal to $\frac{k}{nR}$ the k-NN density estimate of $f(x)$. Formula (18) provides sensible estimators for arbitrary kernels. The bias and variance of this more general k-NN estimator is given in a theorem by Mack (1981).

Theorem 3. *Let the conditions of Theorem 2 hold, except instead that $k \to \infty$, $k/n \to 0$ and $\overline{Lim}\ k^5/n^4 < \infty$ as $n \to \infty$. Then*

$$\sqrt{k}\left[\widehat{m}_k(x) - m(x) - (k/n)^2 B_{nn}(x)\right] \Rightarrow N(0, V_{nn}(x)),$$

where,

$$B_{nn}(x) = \mu_2(K)\left[\frac{m''(x)+2m'(x)\frac{f'}{f}(x)}{8f^2(x)}\right]$$
$$V_{nn}(x) = 2\sigma^2(x)\nu_2(K).$$

In contrast to kernel smoothing, the variance of the k-NN regression smoother does not depend on f, the density of X. This makes sense since the k-NN estimator always averages over exactly k observations independently of the distribution of the X-variables. The bias constant $B_{nn}(x)$ is also different from the one for kernel estimators given in Theorem 2. An approximate identity between k-NN and kernel smoothers can be obtained by setting

$$k = 2nhf(x), \tag{19}$$

or equivalently $h = \frac{k}{\{2nf(x)\}}$. For this choice of k or h respectively, the asymptotic mean squared error formulas of Theorem 2 and Theorem 3 are identical.

3.2.2 Symmetrized k-NN Estimators

A computationally useful modification of $\widehat{m}_k$ is to restrict the k-nearest neighbors always to symmetric neighborhoods, i.e., one takes $k/2$ neighbors to the left and $k/2$ neighbors to the right. In this case, weight-updating formulas can be given, see Härdle (1990, Section 3.2). The bias formulas are slightly different, see Härdle and Carroll (1989), but (19) remains true.

3.3 Local Polynomial Estimators

The Nadaraya-Watson estimator can be regarded as the solution of the minimization problem

$$\widehat{m}_h(x) = \arg\min_{\theta} \sum_{i=1}^{n} K_h(x - X_i)\{Y_i - \theta\}^2. \tag{20}$$

This motivates the local polynomial class of estimators. Let $\widehat{\theta}_0, \ldots, \widehat{\theta}_p$ minimize

$$\sum_{i=1}^{n} K_h(x - X_i)\left\{Y_i - \theta_0 - \theta_1(X_i - x) - .. - \theta_p\frac{(X_i - x)^p}{p!}\right\}^2. \tag{21}$$

Then $\widehat{\theta}_0$ serves as an estimator of $m(x)$, while $\widehat{\theta}_j$ estimates the $j'th$ derivative of m. Clearly, $\widehat{\theta}_0$ is linear in Y. A variation on these estimators called *LOWESS* was first considered in Cleveland (1979) who employed a nearest neighbor window. Fan (1992) establishes an asymptotic approximation for the case where $p = 1$, which he calls the local linear estimator $\widehat{m}_{h,l}(x)$.

Theorem 4. *Let the conditions of Theorem 2 hold. Then*

$$\sqrt{nh}\left[\widehat{m}_{h,l}(x) - m(x) - h^2 B_l(x)\right] \Rightarrow N(0, V_l(x)),$$

where

$$B_l(x) = \tfrac{1}{2}\mu_2(K)m''(x)$$
$$V_l(x) = \nu_2(K)\sigma^2(x)/f(x).$$

The local linear estimator is unbiased when m is linear, while the Nadaraya-Watson estimator may be biased depending on the marginal density of the design.

We note here that fitting higher order polynomials can result in bias reduction, see Fan and Gijbels (1992) and Ruppert and Wand (1992) – who also extend the analysis to multidimensional explanatory variables.

The principle underlying the local polynomial estimator can be generalized in a number of ways. Tibshirani (1984) introduced the local likelihood procedure in which an arbitrary parametric regression function $g(x;\theta)$ substitutes the polynomial in (21). Fan, Heckman and Wand (1992) develop theory for a nonparametric estimator in a *GLIM* (Limited Dependent Variable) model in which, for example, a probit likelihood function replaces the polynomial in (21). An advantage of this procedure is that low bias results when the parametric model is true.

3.4 Spline Estimators

For any estimate $\widehat{m}$ of m, the residual sum of squares (RSS) is defined as $\sum_{i=1}^{n}\{Y_i - \widehat{m}(X_i)\}^2$, which is a widely used criterion, in other contexts, for generating estimators of regression functions. However, the RSS is minimized by an $\widehat{m}$ interpolating the data, assuming no ties in the $X's$. To avoid this problem it is necessary to add a stabilizer. Most work is based on the stabilizer $\Omega(\widehat{m}) = \int\{\widehat{m}''(u)\}^2 du$, although see Ansley, Kohn and Wong (1993) and Koenker, Ng and Portnoy (1993) for alternatives. The cubic spline estimator $\widehat{m}_\lambda$ is the (unique) minimizer of

$$R_\lambda(\widehat{m}, m) = \sum_{i=1}^{n}\{Y_i - \widehat{m}(X_i)\}^2 + \lambda\int\{\widehat{m}''(u)\}^2 du. \tag{22}$$

The spline $\widehat{m}_\lambda$ has the following properties: It is a cubic polynomial between two successive X-values; at the observation points $\widehat{m}_\lambda(\bullet)$ and its first two derivatives are continuous; at the boundary of the observation interval the spline is linear. This characterization of the solution to (22) allows the integral term on the right hand side to be replaced by a quadratic form, see Eubank (1988) and Wahba (1990), and computation of the estimator proceeds by standard, although computationally intensive, matrix techniques.

The smoothing parameter λ controls the degree of smoothness of the estimator $\widehat{m}_\lambda$. As $\lambda \longrightarrow 0$, $\widehat{m}_\lambda$ interpolates the observations, while if $\lambda \longrightarrow \infty$, $\widehat{m}_\lambda$ tends to a least squares regression line. Although $\widehat{m}_\lambda$ is linear in the Y data, see Härdle (1990, p58-59), its dependency on the design and on the smoothing parameter is rather complicated. This has resulted in rather less treatment of the statistical properties of these estimators, except in rather simple settings, although see Wahba (1990) – in fact, the extension to multivariate design is not straightforward. However, splines are asymptotically equivalent to kernel smoothers as Silverman (1984) showed. The equivalent kernel is

$$K(u) = \frac{1}{2}\exp\left(-\frac{|u|}{\sqrt{2}}\right)\sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right), \tag{23}$$

which is of fourth order, since its first three moments are zero, while the equivalent bandwidth $h = h(\lambda; X_i)$ is

$$h(\lambda; X_i) = \lambda^{1/4} n^{-1/4}\, f(X_i)^{-1/4}. \tag{24}$$

One advantage of spline estimators over kernels is that global inequality and equality constraints can be imposed more conveniently: for example, it may be desirable to restrict the smooth to pass through a particular point – see Jones (1985). Silverman (1985) discusses a Bayesian interpretation of the spline procedure. However, from Section 2.2 we conclude that this interpretation can also be given to kernel estimators.

3.5 Series Estimators

Series estimators have received considerable attention in the econometrics literature, following Elbadawi, Gallant and Souza (1983). This theory is very much tied to the structure of Hilbert space. Suppose that m has an expansion for all x:

$$m(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x), \tag{25}$$

in terms of the orthogonal basis functions $\{\varphi_j\}_{j=0}^{\infty}$ and their coefficients $\{\beta_j\}_{j=0}^{\infty}$. Suitable basis systems include the *Legendre* polynomials described in Härdle (1990), and the *Fourier* series used in Gallant and Souza (1991).

A simple method of estimating $m(x)$ involves firstly selecting a basis system and a truncation sequence $t(n)$, where $t(n)$ is an integer less than n, and then regressing Y_i on $\varphi_{ti} = (\varphi_0(X_i), .., \varphi_t(X_i))^T$. Let $\left\{\widehat{\beta}_j\right\}_{j=0}^{t(n)}$ be the least squares "parameter" estimates, then

$$\widehat{m}_{t(n)}(x) = \sum_{j=0}^{t(n)} \widehat{\beta}_j \varphi_j(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \tag{26}$$

where $W_n(x) = (W_{n1}, .., W_{nn})^T$, with

$$W_n(x) = \varphi_{tx}^T (\Phi_t^T \Phi_t)^{-1} \Phi_t^T, \tag{27}$$

where $\varphi_{tx} = (\varphi_0(x), .., \varphi_t(x))^T$ and $\Phi_t = (\varphi_{t1}, .., \varphi_{tn})^T$.

These estimators are typically very easy to compute. In addition, the extension to additive structures and semiparametric models is convenient, see Andrews and Whang

(1990) and Andrews (1991). Finally, provided $t(n)$ grows at a sufficiently fast rate, the optimal (given the smoothness of m) rate of convergence can be established – see Stone (1982), while fixed window kernels achieve at best a rate of convergence (of MSE) of $n^{4/5}$. However, the same effect can be achieved by using a kernel estimator, where the order of the kernel changes with n in such a way as to produce bias reduction of the desired degree, see Müller (1987). In any case, the evidence of Marron and Wand (1992) cautions against the application of bias reduction techniques unless quite large sample sizes are available. Finally, a major disadvantage with the series method is that there is relatively little theory about how to select the basis system and the smoothing parameter $t(n)$.

3.6 Kernels, k–NN, splines, and series

Splines and series are both "global" methods in the sense that they try to approximate the whole curve at once, while kernel and nearest neighbor methods work separately on each estimation point. Nevertheless, when X is uniformly distributed, kernels and nearest neighbor estimators of $m(x)$ are identical, while spline estimators are roughly equivalent to a kernel estimator of order 4. Only when the design is not equispaced, do substantial differences appear.

We apply kernel, k-NN, orthogonal series (we used the Legendre system of orthogonal polynomials), and splines to the car data set (Table 7, p. 352–355 in Chambers, Cleveland, Kleiner and Tukey (1983)).

In each plot, we give a scatterplot of the data x = price in dollars of car (in 1979) versus y = miles per US gallon of that car, and one of the nonparametric estimators. The sample size is $n = 74$ observations. In Figure 5a we have plotted together with the raw data a kernel smoother $\widehat{m}_h$ for which a quartic kernel was used with $h = 2000$. Very similar to this is the spline smoother shown in Figure 5b ($\lambda = 10^9$). In this example, the X's are not too far from uniform. The effective local bandwidth for the spline smoother from (24) is a function of $f^{-1/4}$ only, which does not vary that much. Of course at the right end with the isolated observation at $x = 15906$ and $y = 21$ (Cadillac Seville) both kernel and splines must have difficulties. Both work essentially with a window of fixed width. The series estimator (Figure 5d) with $t = 8$ is quite close to the spline estimator.

In contrast to these regression estimators stands the k-NN smoother ($k = 11$) in Figure 5c. We used the symmetrized k-NN estimator for this plot. By formula (19) the dependence of k on f is much stronger than for the spline. At the right end of the price scale no local effect from the outlier described above is visible. By contrast in the main body of the data where the density is high this k-NN smoother tends to be wiggly.

3.7 Confidence Intervals

The asymptotic distribution results contained in Theorems 2-4 can be used to calculate pointwise confidence intervals for the estimators described above. In practice it is usual to ignore the bias term, since this is rather complicated, depending on higher derivatives of the regression function and perhaps on the derivatives of the density of X. This approach can be justified when a bandwidth is chosen that makes the bias relatively small.

In this section we restrict our attention to the Nadaraya-Watson regression estimator. In this case, we suppose that $hn^{1/5} \to 0$, which ensures that the bias term does not appear

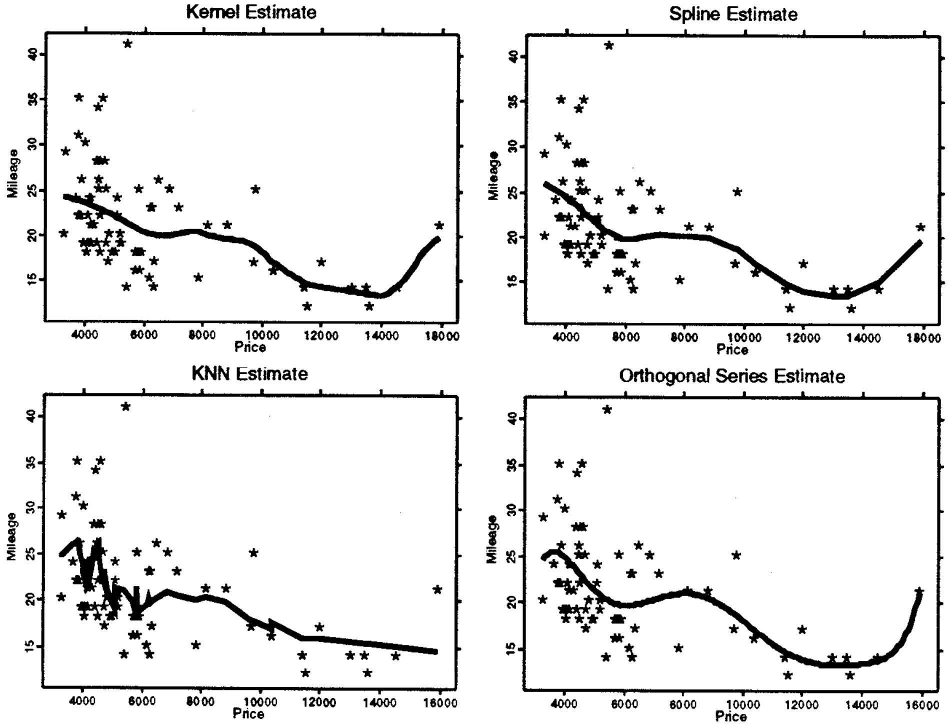

Figure 5abcd. Scatterplot of car price (x) and miles per gallon (y) with four different smooth approximations (n = 74, h = 2000, k = 11, $\lambda = 10^9$, t = 8). Standard deviation of car price is 2918.

in the limiting distribution. Let

$$CLO(x) = \widehat{m}_h(x) - c_{\alpha/2}\widehat{s}$$

$$CUP(x) = \widehat{m}_h(x) + c_{\alpha/2}\widehat{s},$$

where $\Phi(c_\alpha) = (1-\alpha)$ with $\Phi(\bullet)$ the standard normal distribution, while $\widehat{s}^2$ is a consistent estimate of the asymptotic variance of $\widehat{m}_h(x)$. Suitable estimators include

1) $\widehat{s}_1^2 = n^{-1}h^{-1}\nu_2(K)\widehat{\sigma}_h^2(x)/\widehat{f}_h(x)$

2) $\widehat{s}_2^2 = \ \widehat{\sigma}_h^2(x)\sum_{i=1}^n W_{ni}^2(x)$

3) $\widehat{s}_3^2 = \sum_{i=1}^n W_{ni}^2(x)\widehat{\epsilon}_i^2,$

where $\widehat{f}_h(x)$ is defined in (2), $\widehat{\epsilon}_i = Y_i - \widehat{m}_h(X_i)$ are the nonparametric residuals, and $\widehat{\sigma}_h^2(x) = \sum_{i=1}^n W_{ni}(x)\widehat{\epsilon}_i^2$ is a nonparametric estimator of $\sigma^2(x)$ – see Robinson (1987) and Hildenbrand and Kneip (1992) for a discussion of alternative conditional variance estimators and their application.

With the above definitions,

$$P\{m(x) \in [CLO(x), CUP(x)]\} \to 1 - \alpha. \tag{28}$$

These confidence intervals are frequently employed in econometric applications, see for example Bierens and Pott-Buter (1990), Banks, Blundell, and Lewbel (1993), and Gozalo (1989). This approach is relevant if the behavior of the regression function at a single point is under consideration. Usually, however, its behavior over an interval is under study. In this case, pointwise confidence intervals do not take account of the joint nature of the implicit null hypothesis.

We now consider uniform confidence bands for the function m, over some compact subset χ of the support of X. Without loss of generality we take $\chi = [0,1]$. We require functions $CLO^*(x)$ and $CUP^*(x)$ such that

$$P\{m(x) \in [CLO^*(x), CUP^*(x)] \ \mathit{for\ all}\ x \in \chi\} \to 1 - \alpha, \tag{29}$$

Let

$$CLO^*(x) = \widehat{m}_h(x) - \left[\tfrac{c_\alpha^*}{\delta} + \delta + \tfrac{1}{2\delta}\ln\left\{\tfrac{\nu_2(K')}{4\pi^2\nu_2(K)}\right\}\right]\widehat{s}_1$$

$$CUP^*(x) = \widehat{m}_h(x) + \left[\tfrac{c_\alpha^*}{\delta} + \delta + \tfrac{1}{2\delta}\ln\left\{\tfrac{\nu_2(K')}{4\pi^2\nu_2(K)}\right\}\right]\widehat{s}_1,$$

where $\delta = \sqrt{2\log(1/h)}$, and $exp\{\text{-}{-2}exp(-c_\alpha^*)\} = (1-\alpha)$. Then (29) is satisfied under the conditions given in Härdle (1990, Theorem 4.3.1). See also Prakasa Rao (1983, Theorem 2.1.17) for a treatment of the same problem for density estimators.

In the figure below we show the uniform confidence bands for the income data of Figure 2.

Hall (1993) advocates using the bootstrap to construct uniform confidence bands. He argues that the error in (29) is $O(\frac{1}{\log n})$, which can be improved to $O(\frac{(\log h^{-1})^3}{nh})$ by the judicious use of this resampling method in the random design case. See also Hall (1992) and Härdle (1990) for further applications of the bootstrap in nonparametric statistics.

3.8 Regression Derivatives and Quantiles

There are a number of other functionals of the conditional distribution that are of interest for applications. The first derivative of the regression function measures the strength of the relationship between Y and X, while second derivatives can quantify the concavity or convexity of the regression function. Let $\widehat{m}(x)$ be any estimator of $m(x)$ that has at least r non-zero derivatives at x. Then $m^{(r)}(x)$ can be estimated by the $r'th$ derivative of $\widehat{m}(x)$, denoted $\widehat{m}^{(r)}(x)$. Müller (1988) describes kernel estimators of $m^{(r)}(x)$ based on the convolution method of Gasser and Müller (1984); their method gives simpler bias expressions than the Nadaraya-Watson estimator. An alternative technique is to fit a local polynomial (of order r) estimator, and take the coefficient on the $r'th$ term in (21), see Ruppert and Wand (1992). In each case, the resulting estimator is linear in Y_i, with bias of order h^2 and variance of order $n^{-1}h^{-(2r+1)}$.

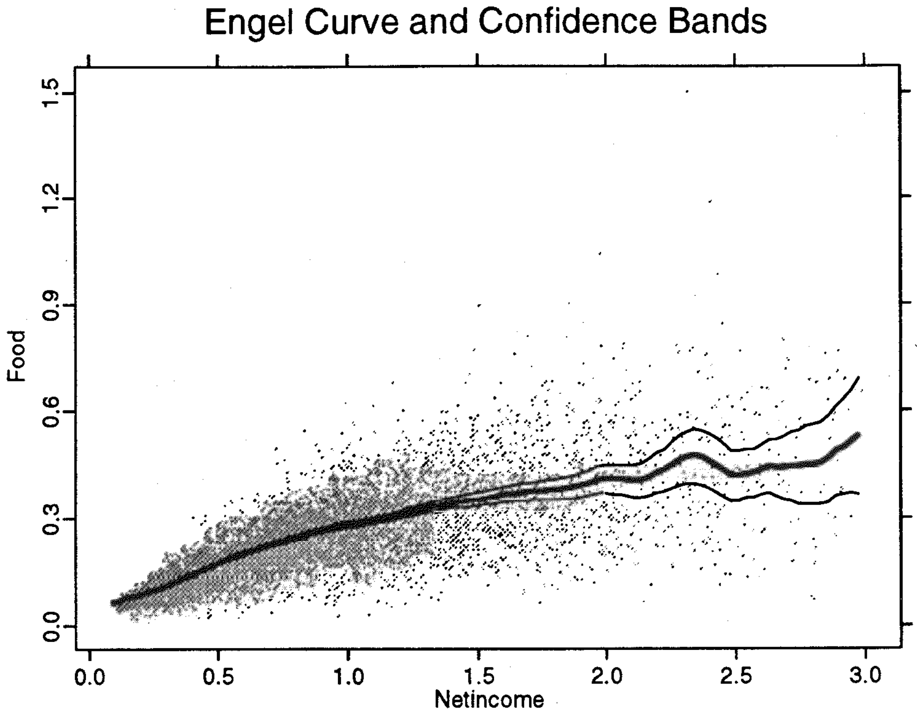

Figure 6. Uniform confidence bands for the income data. Food versus net income. Calculated using XploRe macro `reguncb`

Quantiles can also be useful. The median is an alternative – and robust – measure of location, while other quantiles can help to describe the spread of the conditional distribution. Let $f_{Y|X=x}(y)$ denote the conditional distribution of Y given $X = x$, and let $c_\alpha(x)$ be the α'th conditional quantile, i.e.

$$\alpha = \int_{-\infty}^{c_\alpha(x)} f_{Y|X=x}(y)dy, \tag{30}$$

where for simplicity we assume this is unique. There are several methods for estimating $c_\alpha(x)$.

Firstly, let $Z_j = (W_{nj}(x), Y_j)^T$, where $W_{nj}(x)$ are kernel or nearest neighbor weights. We first sort $\{Z_j\}_{j=1}^n$ on the variable Y_j, and find the largest index J such that

$$\sum_{j=1}^{J} W_{nj}(x) \leq \alpha.$$

Then let

$$\widehat{c}_\alpha(x) = Y_J. \tag{31}$$

Stute (1986) shows that $\widehat{c}_\alpha(x)$ consistently estimates $c_\alpha(x)$, with the same convergence rates as in ordinary nonparametric regression, see also Bhattacharya and Gangopadhyay (1990). When K is the uniform kernel and $\alpha = \frac{1}{2}$, this procedure corresponds to the running median discussed in Härdle (1990, p69-71). A smoother estimator is obtained by also smoothing in the y direction, i.e.

$$\widehat{c}_\alpha(x) = \frac{1}{n}\sum_{j=1}^{n} K_h(\frac{J-j}{n})Y_j.$$

Provided K has at least r non-zero derivatives, the $r'th$ derivative of $c_\alpha(x)$ can be estimated by the $r'th$ derivative of $\widehat{c}_\alpha(x)$. See Anand, Harris and Linton (1993) and Robb, McGee and Burbidge (1992) for applications.

An alternative method of estimating conditional quantiles is through minimizing an appropriate loss function, which idea originates with Koenker and Bassett (1978). In particular,

$$\widehat{c}_\alpha(x) = \arg\min_\theta \sum_{i=1}^{n} K_h(x - X_i)\rho_\alpha(Y_i - \theta), \tag{32}$$

where $\rho_\alpha(y) = |y| + (2\alpha - 1)y$, consistently estimates $c_\alpha(x)$. Computation of the estimator can be carried out by linear programming techniques. Chaudhuri (1992) provides asymptotic theory for this estimator in a general multidimensional context and for estimators of the derivatives of $c_\alpha(x)$.

In neither (31) nor (32) is the estimator linear in Y_i, although the asymptotic distribution of the estimators are determined by a linear approximation to them, i.e. the estimators are asymptotically normal.

4 Optimality and Bandwidth Choice

4.1 Optimality

Let $Q(h)$ be a performance criterion. We say that a bandwidth sequence h^* is asymptotically optimal if

$$\frac{Q(h^*)}{\inf_{h \in H_n} Q(h)} \xrightarrow{P} 1, \tag{33}$$

as $n \to \infty$, where H_n is the range of permissible bandwidths. There are a number of alternative optimality criteria in use. Firstly, we may be interested in the quadratic loss of the estimator at a single point x, which is measured by the *Mean Squared Error*, $MSE\{\widehat{m}_h(x)\}$. Secondly, we may be only concerned with a global measure of performance. In this case, we may consider the *Integrated Mean Squared Error*, $IMSE = \int MSE\{\widehat{m}_h(x)\}\,\pi(x)f(x)dx$ for some weighting function $\pi(\bullet)$. An alternative is the in-sample version of this, the *averaged squared error*

$$d_A(h) = n^{-1}\sum_{j=1}^{n}\{\widehat{m}_h(X_j) - m(X_j)\}^2\,\pi(X_j). \tag{34}$$

The purpose of $\pi(\bullet)$ may be to downweight observations in the tail of $X's$ distribution, and thereby to eliminate *boundary effects* – see Müller (1988) for a discussion. When $h = O(n^{-1/5})$, the squared bias and the variance of the kernel smoother have the same magnitude; this is the optimal order of magnitude for h with respect to all three criteria, and the corresponding performance measures are all $O(n^{-4/5})$ in this case.

Now let $h = \gamma n^{-1/5}$, where γ is a constant. The optimal constant balances the contributions to MSE from the squared bias and the variance respectively. From Theorem 2 we obtain an approximate mean squared error expansion,

$$MSE\,[\widehat{m}_h(x)] \approx n^{-1}h^{-1}V(x) + h^4B^2(x), \tag{35}$$

and the bandwidth minimizing (35) is

$$h_0(x) = \left\{ \frac{V(x)}{4B^2(x)} \right\}^{1/5} n^{-1/5}. \tag{36}$$

Similarly, the optimal bandwidth with respect to $IMSE$ is the same as (36) with $V = \int V(x)\pi(x)f(x)dx$ and $B^2 = \int B^2(x)\pi(x)f(x)dx$ replacing $V(x)$ and $B^2(x)$. Unfortunately, in either case the optimal bandwidth depends on the unknown regression function and design density. We discuss in Section 4.2 below how one can obtain empirical versions of (36).

The optimal local bandwidths can vary considerably with x, which point is best illustrated for density estimation. Suppose that the density is standard normal and a standard normal kernel is used. In this case, as $x \to \infty$, $h_0(x) \to \infty$: when data is sparse a wider window is called for. Also at $x = \pm 1$, $h_0(x) = \infty$, which reflects the fact that $\phi'' = 0$ at these points. Elsewhere, substantially less smoothing is called for: at ± 2.236, $h_0(x) = 0.884n^{-1/5}$ (which is the minimum value of $h_0(x)$). The optimal global bandwidth is $1.06n^{-1/5}$.

Although allowing the bandwidth to vary with x dominates the strategy of throughout choosing a single bandwidth, in practice this requires considerably more computation, and is rarely used in applications.

By substituting h_0 in (35), we find that the optimal MSE and $IMSE$ depend on K only through

$$T(K) = \nu_2^2(K)\mu_2(K). \tag{37}$$

This functional can be minimized with respect to K using the calculus of variations, although it is necessary to first adopt a scale standardization of K – for details, see Gasser, Müller, and Mammitzsch (1985). A kernel is said to be optimal if it minimizes (37). The optimal kernel of order 2 is the Epanechnikov kernel given in Table 1. The third column of this table shows the loss in efficiency of other kernels in relation to this optimal one. Over a wide class of kernel estimators, the loss in efficiency is not that drastic; more important is the choice of h than the choice of K.

Any kernel can be rescaled as $K^*(\bullet) = s^{-1}K(\bullet/s)$ which of course changes the value of the kernel constants and hence h_0. In particular,

$$\nu_2(K^*) = s^{-1}\nu_2(K) \; ; \; \mu_2^2(K^*) = s^2\mu_2(K).$$

We can uncouple the scaling effect by using for each kernel K, that K^* with scale

$$s^* = \left\{ \frac{\nu_2(K^*)}{\mu_2^2(K)} \right\}^{1/5}$$

for which $\mu_2^2(K^*) = \nu_2(K^*)$. Now suppose we wish to compare two smooths with kernels K_j and bandwidths h_j respectively. This can be done by transforming both to their canonical scale, see Marron and Nolan (1989), and then comparing their s_j^*. In Table 2 we give the exchange rate between various commonly used kernels.

The bandwidth of 0.2 used with a quartic kernel in Figure 2, translates into a bandwidth of 0.133 for a uniform kernel and 0.076 for a Gaussian kernel.

s_j^* / s_i^*	Uniform	Triangle	Epanechnikov	Quartic	Gaussian
Uniform	1.000	0.715	0.786	0.663	1.740
Triangle	1.398	1.000	1.099	0.927	2.432
Epanechnikov	1.272	0.910	1.000	0.844	2.214
Quartic	1.507	1.078	1.185	1.000	2.623
Gaussian	0.575	0.411	0.452	0.381	1.000

Table 2. Kernel Exchange Rate

4.2 Choice of Smoothing Parameter

For each nonparametric regression method, one has to choose how much to smooth for the given dataset. In Section 3 we saw that k-NN, series, and spline estimation are asymptotically equivalent to the kernel method, so we describe here only the selection of bandwidth h for kernel regression smoothing.

4.2.1 Plug-in

The asymptotic approximation given in (36) can be used to determine an optimal local bandwidth. We can calculate an estimated optimal bandwidth $\widehat{h}_{pl}$ in which the consistent estimators $\widehat{m}''_{h^*}(x)$, $\widehat{\sigma}^2_{h^*}(x)$, $\widehat{f}_{h^*}(x)$ and $\widehat{f}'_{h^*}(x)$ replace the unknown functions. We then use $\widehat{m}_{\widehat{h}_{pl}}(x)$ to estimate $m(x)$. Likewise, if a globally optimal bandwidth is required, one must substitute estimators of the appropriate average functionals. This procedure is generally fast and simple to implement. Its properties are examined in Härdle, Hall, and Marron (1992).

However, this method fails to provide pointwise optimal bandwidths, when $m(x)$ possesses less than two continuous derivatives. Finally, a major disadvantage of this procedure is that a preliminary bandwidth h^* must be chosen for estimation of $m''(x)$ and the other quantities.

4.2.2 Crossvalidation

Crossvalidation is a convenient method of global bandwidth choice for many problems, and relies on the well established principle of out-of-sample predictive validation.

Suppose that optimality with respect to $d_A(h)$ is the aim. We must first replace $d_A(h)$ by a computable approximation to it. A naive estimate would be to just replace the unknown values $m(X_j)$ by the observations Y_j:

$$p(h) = n^{-1} \sum_{j=1}^{n} \{\widehat{m}_h(X_j) - Y_j\}^2 \, \pi(X_j),$$

which is called the resubstitution estimate.

However, this quantity makes use of the each observation twice – the response variable Y_j is used in $\widehat{m}_h(X_j)$ to predict itself. Therefore, $p(h)$ can be made arbitrarily small by taking $h \to 0$ (when there are no tied X observations). This fact can be expressed via asymptotic expressions for the moments of p. Conditional on $X_1, .., X_n$, we have

$$E\left[p(h)\right] = E\left[d_A(h)\right] + \frac{1}{n}\sum_{i=1}^{n}\sigma^2(X_i)\pi(X_i) - 2\frac{1}{n}\sum_{i=1}^{n}W_{ni}(X_i)\sigma^2(X_i)\pi(X_i), \tag{38}$$

and the third term is of the same order of magnitude as $E\left[d_A(h)\right]$, but with negative sign. Therefore, d_A is wrongly underestimated, and the selected bandwidth will be downward biased.

The simplest way to avoid this problem is to remove the j-th observation

$$\widehat{m}_{h,j}(X_j) = \frac{\sum_{j\neq i} K_h(X_j - X_i)Y_i}{\sum_{j\neq i} K_h(X_j - X_i)}. \tag{39}$$

This leave-one-out estimate is used to form the so-called crossvalidation function

$$CV(h) = n^{-1}\sum_{j=1}^{n}\left\{\widehat{m}_{h,j}(X_j) - Y_j\right\}^2 \pi(X_j), \tag{40}$$

which is to be minimized with respect to h. For technical reasons, the infimum must be taken only over a restricted set of bandwidths such as $H_n = [n^{-(1/5-\zeta)}, n^{-(1/5+\zeta)}]$, for some $\zeta > 0$.

Theorem 5. *Assume that the conditions given in Härdle (1990, Theorem 5.1.1) hold.*

Then the bandwidth selection rule, "Choose $\hat{h}$ to minimize $CV(h)$" is asymptotically optimal with respect to $d_A(h)$ and *$IMSE$.*

Proof: See Härdle and Marron (1985).

The conditions include the restriction that $f > 0$ on the compact support of π, moment conditions on ϵ, and a Lipschitz condition on K. However, unlike for the plug-in procedure, m and f need not be differentiable (a Lipschitz condition is required, however).

4.2.3 Other data driven selectors

There are a number of different automatic bandwidth selectors that produce asymptotically optimal kernel smoothers. They are based on various ways of correcting the downwards bias of the resubstitution estimate of $d_A(h)$. The function $p(h)$ is multiplied by a correction factor that in a sense penalizes the too small h's. The general form of this selector is

$$G(h) = n^{-1}\sum_{j=1}^{n}\left\{\widehat{m}_h(X_i) - Y_i\right\}^2 \pi(X_i)\Xi\left\{W_{ni}(X_i)\right\},$$

where Ξ is the correction function with first-order Taylor expansion

$$\Xi(u) = 1 + 2u + O(u^2), \tag{41}$$

as $u \longrightarrow 0$. Some well known example are:

(i) *Generalized Cross-validation* (Craven and Wahba 1979; Li 1985),

$$\Xi_{GCV}\ (u) = (1-u)^{-2};$$

(ii) *Akaike's Information Criterion* (Akaike 1970)

$$\Xi_{AIC}(u) = \exp\ (2u);$$

(iii) *Finite Prediction Error* (Akaike 1974),

$$\Xi_{FPE}(u) = (1+u)/(1-u);$$

(iv) *Shibata's (1981) model selector*,

$$\Xi_S(u) = 1 + 2u;$$

(v) *Rice's (1984) bandwidth selector*,

$$\Xi_T(u) = (1-2u)^{-1}.$$

Härdle, Hall, and Marron (1988) show that the general criterion $G(h)$ works in producing asymptotically optimal bandwidth selection, although they present their results for the equispaced design case only.

The method of crossvalidation was applied to the car data set to find the optimal smoothing parameter h. A plot of the crossvalidation function is given in Figure 7. The computation is for the quartic kernel using the WARPing method, see Härdle and Scott (1990). The minimal $\widehat{h} = \arg\min CV(h)$ is at 1922 which shows that in Figure 5a we used a slightly too large bandwidth.

Härdle, Hall and Marron (1988) investigate how far the crossvalidation optimal $\widehat{h}$ is from the true optimum $\widehat{h}_0$ (that minimizes $d_A(h)$). They show that for each optimization method,

$$n^{1/10}\left(\frac{\widehat{h}-\widehat{h}_0}{\widehat{h}_0}\right) \Rightarrow N(0,\sigma^2) \tag{42}$$

$$n\left\{d_A(\widehat{h}) - d_A(\widehat{h}_0)\right\} \Rightarrow C_1\chi_1^2, \tag{43}$$

where σ^2 and C_1 are both positive. To this higher order of approximation, the above methods are all asymptotically equivalent. Another interesting result is that the estimated $\widehat{h}$ and optimum $\widehat{h}_0$ are actually negatively correlated! Hall and Johnstone (1992) show how to correct for this effect in density estimation and in regression with uniform X's. It is still an open question how to improve this for the general regression setting we are considering here.

There has been considerable research into finding improved methods of bandwidth selection, that give faster rates of convergence in (42). Most of this work is in density estimation – see the recent review of Jones, Marron and Sheather (1992) for references. In this case, various $\sqrt{n}$ consistent bandwidth selectors have been suggested. The finite sample properties of these procedures are not well established, although Park and Turlach (1992) contains some preliminary simulation evidence. Härdle, Hall and Marron (1992) construct a $\sqrt{n}$ consistent bandwidth selector for regression based on a bias reduction technique.

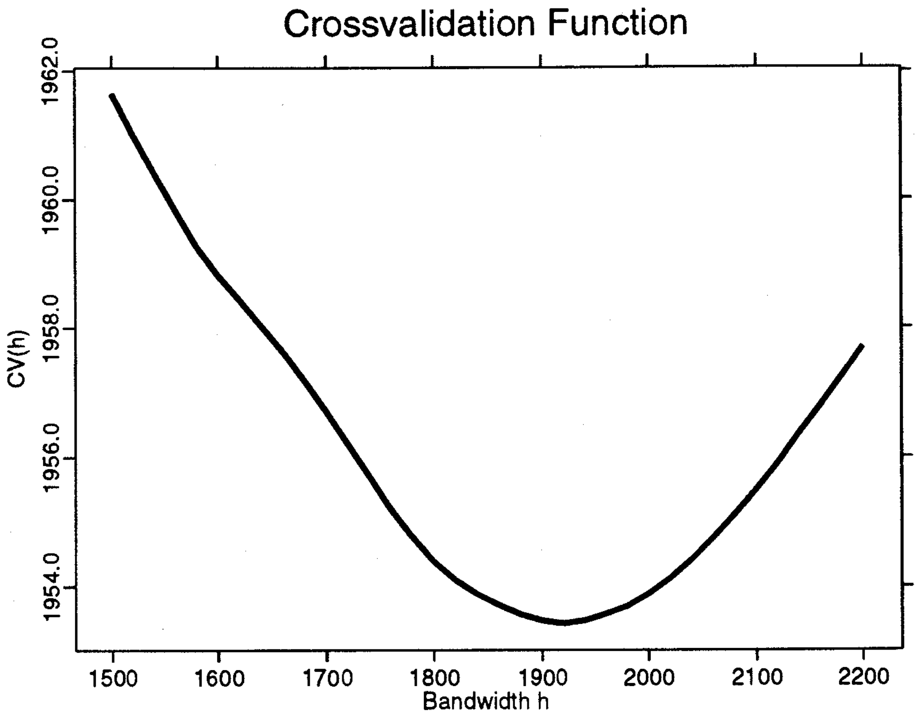

Figure 7. The crossvalidation function $CV(h)$ for the car data. Quartic kernel. Computation made with XploRe macro `regcvl`.

5 Application to Time Series

In the theoretical development described up to this point, we have restricted our attention to independent sampling. However, smoothing methods can also be applied to dependent data. Considerable resources are devoted to providing forecasts of macroeconomic entities such as GNP, unemployment and inflation, while the benefits of predicting asset prices are obvious. In many cases linear models have been the basis of econometric prediction, while more recently nonlinear models such as ARCH have become popular. Nonparametric methods can also be applied in this context, and provide a model free basis of predicting future outcomes. We focus on the issue of functional form, rather than that of correlation structure – this latter issue is treated, from a nonparametric point of view, in Brillinger (1980), see also Phillips (1991) and Robinson (1991).

Suppose that we observe the vector time series $\{Z_i\}_{i=1}^n$, where $Z_i = (Y_i, X_i)$, and X_i is strictly exogenous in the sense of Engle et al. (1983). It is convenient to assume that the process is stationary and mixing as defined in Gallant and White (1988), which includes most linear processes for example, although extensions to certain types of nonstationarity can also be permitted. We consider two distinct problems. Firstly, we want to predict Y_i from its own past which we call autoregression. Secondly, when we want to predict Y_i from X_i which problem we call regression with correlated errors.

5.1 Autoregression

For convenience we restrict attention to the problem of predicting the scalar Y_{i+k} given Y_i for some $k > 0$. The best predictor is provided by the autoregression function

$$M_k(y) = E(Y_{i+k} \mid Y_i = y). \tag{44}$$

More generally, one may wish to estimate the conditional variance of Y_{i+k} from lagged values,

$$V_k(y) = Var(Y_{i+k} \mid Y_i = y),$$

and even the predictive density $f_{Y_{i+k}|Y_i}$. These quantities can be estimated using any of the smoothing methods described in this chapter. See Robinson (1983) and Bierens (1987) for some theoretical results including convergence rates and asymptotic distributions.

Diebold and Nason (1990), Meese and Rose (1991), and Mizrach (1992) estimate $M(\bullet)$ for use in predicting asset prices over short horizons. In each case a locally weighted regression estimator was employed with a nearest neighbor type window, while bandwidth was chosen subjectively (except in Mizrach (1992) where crossvalidation was used). Not surprisingly, their published results concluded that there was little gain in predictive accuracy over a simple random walk. Pagan and Hong (1991), Pagan and Schwert (1990), and Pagan and Ullah (1988) estimate $V(\bullet)$, thereby to evaluate the risk premium of asset returns. They used a variety of nonparametric methods including Fourier series and kernels. Their focus was on estimation rather than prediction, and their procedures relied on some parametric estimation, see also Whistler (1988) and Gallant, Hsieh and Tauchen (1991).

A scientific basis can also be made for choosing bandwidth in this sampling scheme. Härdle and Vieu (1991) showed that crossvalidation also works in the autoregression problem – "choose" $\widehat{h} = \arg\min CV(h)$ gives asymptotically optimal estimates.

To illustrate this result we simulated an autoregressive process $Y_i = M(Y_{i-1}) + \epsilon_i$ with

$$M(y) = y\exp(-y^2), \tag{45}$$

where the innovations ϵ_i were uniformly distributed over the interval $(-1/2, 1/2)$. Such a process is α-mixing with geometrically decreasing $\alpha(n)$ as shown by Doukhan and Ghindès (1980) and Györfi et al. (1990, Section III.4.4). The sample size investigated was $n = 100$. The quartic kernel function (3) was used. The minimum of $CV(h)$ was $\widehat{h} = 0.43$, while the optimum of $d_A(h)$ is at $h = 0.52$. The curve $d_A(h)$ is very flat for this example, since there is very little bias present. In Figure 8 we compare the estimated curve with the autoregression function and find good coincidence.

5.2 Correlated Errors

We now consider the regression model

$$Y_i = m(X_i) + \epsilon_i,$$

where X_i is fixed in repeated samples and the errors ϵ_i satisfy $E(\epsilon_i|X_i) = 0$, but are autocorrelated. The kernel estimator $\widehat{m}_h(x)$ of $m(x)$ is consistent under quite general conditions. In fact, its bias is the same as when ϵ_i are independent. However, the variance is generally affected by the dependency structure. Suppose that the error process is $MA(1)$, i.e.

$$\epsilon_i = u_i + \theta u_{i-1},$$

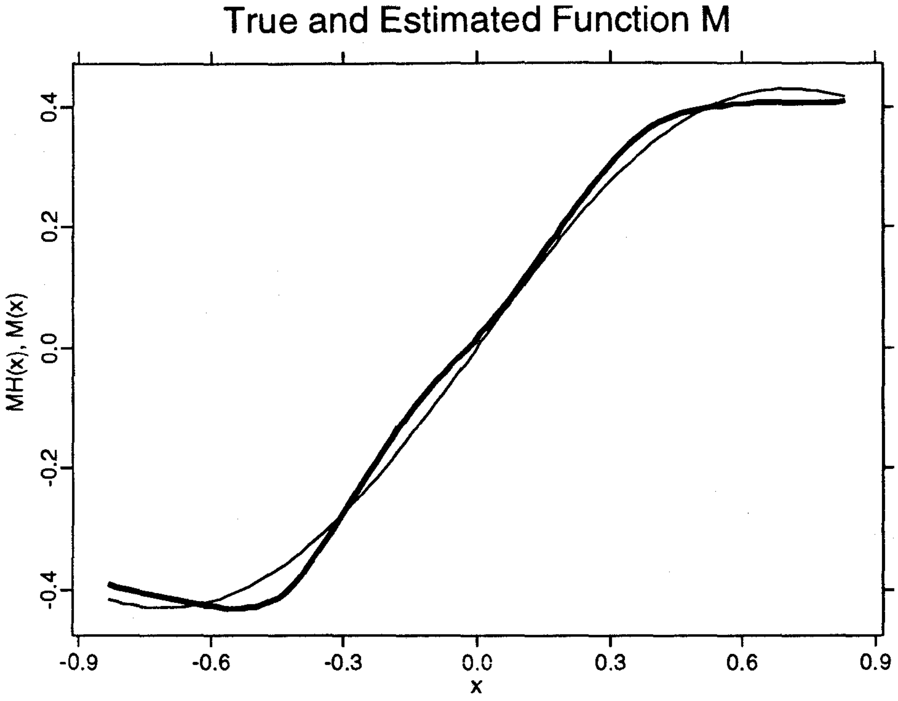

Figure 8. The time regression function $M(y) = y\exp(-y^2)$ for the simulated example (thin line) and the kernel smoother (thick line).

where u_i are i.i.d with zero mean and variance σ^2. In this case,

$$Var\,[\widehat{m}_h(x)] = \sigma^2 \left\{ (1+\theta^2) \sum_{i=1}^{n} W_{ni}^2 + 2\theta \sum_{i=1}^{n-1} W_{ni} W_{ni+1} \right\} \tag{46}$$

which is $O(n^{-1}h^{-1})$, but differs from Theorem 2. If the explanatory variable were time itself (i.e. $X_i = i/n$, $i = 1, .., n$), then a further approximation is possible:

$$Var\,[\widehat{m}_h(x)] \approx \frac{1}{nh}\sigma^2(1+\theta^2+2\theta)\nu_2(K).$$

Hart and Wehrly (1986) develop MSE approximations in a regression model in which the error correlation is a general function $\rho(\bullet)$ of the time between observations.

Unfortunately, crossvalidation fails in this case. Suppose that the errors are $AR(1)$ with autoregression parameter close to one. The effect on the crossvalidation technique described in Section 4 must be drastic. The error process stays a long time on one side of the mean curve. Therefore, the bandwidth selection procedure gives undersmoothed estimates, since it interprets the little bumps of the error process as part of the regression curve. An example is given in Härdle (1990, Figures 7.6, 7.7).

The effect of correlation on the crossvalidation criterion may be mitigated by leaving out more than just one observation. For the $MA(1)$ process, leaving out the 3 contiguous (in time) observations works. This "leave-out-some" technique is sometimes appealing also in the independent setting, see the discussion of Härdle, Hall and Marron (1988), and Hart and Vieu (1991). It may also be possible to correct for this effect by "whitening" the residuals in (40), although this has yet to be shown.

6 Applications to Semiparametric Estimation

Semiparametric models offer a compromise between parametric modeling and the nonparametric approaches we have discussed. When data are high dimensional, or if it is necessary to account for both functional form and correlation of general nature, fully nonparametric methods may not perform well. In this case, semiparametric models may be preferred.

By a semiparametric model we mean that the density of the observable data, conditional on any ancillary information, is completely specified by a finite dimensional parameter θ and an unknown function $G(\bullet)$. The exhaustive monograph of Bickel, Klaassen, Ritov, and Wellner (1992) develops a comprehensive theory of inference for a large number of semiparametric models, although mostly within iid sampling. There are a number of reviews for econometricians including: Robinson (1988b), Newey (1990) and Powell (this volume).

In many cases, θ is of primary interest. Andrews (1989) provides asymptotic theory for a general procedure designed to estimate θ when a preliminary estimate $\widehat{G}$ of G is available. The method involves substituting $\widehat{G}$ for G in an estimating equation derived perhaps from a likelihood function. Typically, the dependence of the estimated parameters $\widehat{\theta}$ on the nonparametric estimators disappears asymptotically, and

$$\sqrt{n}(\widehat{\theta}-\theta) \Rightarrow N(0,\Omega_0), \tag{47}$$

where $\Omega_0 > 0$.

Nevertheless, the small sample properties of $\widehat{\theta}$ can depend quite closely on the way in which this preliminary step is carried out – see the monte carlo evidence contained in Engle and Gardner (1976), Hsieh and Manski (1987), Stock (1989) and Delgado (1992). Some recent work has investigated analytically the small sample properties of semiparametric estimators. Carroll and Härdle (1989), Cavanagh (1989), Härdle et al. (1992), Linton (1991,1992,1993), and Powell and Stoker (1991) develop asymptotic expansions of the form

$$MSE\left[\sqrt{n}(\widehat{\theta}-\theta)\right] \approx \Omega_0 + \frac{\Omega_1}{q_1(n,h)} + \frac{\Omega_2}{q_2(n,h)}, \tag{48}$$

where q_1 and q_2 both increase with n under restrictions on $h(n)$. These expansions yield a formula for the optimal bandwidth similar to (36). An important finding is that different amounts of smoothing are required for $\widehat{\theta}$ and for $\widehat{G}$; in particular, it is often optimal to undersmooth $\widehat{G}$ (by an order of magnitude) when the properties of $\widehat{\theta}$ are at stake.

The MSE expansions can be used to define a plug-in method of bandwidth choice for $\widehat{\theta}$ that is based on second order optimality considerations.

6.1 The Partially Linear Model

Consider

$$Y_i = \beta^T X_i + \phi(Z_i) + \epsilon_i \ ; \ X_i = g(Z_i) + \eta_i, \ i = 1,2,..,n \tag{49}$$

where $\phi(\bullet)$ and $g(\bullet)$ are of unknown functional form, while $E(\epsilon_i|Z_i) = E(\eta_i|Z_i) = 0$. If an inappropriate parametric model is fit to $\phi(\bullet)$, the resulting MLE of β may be inconsistent.

This motivates using nonparametric methods that allow a more general functional form, when it is needed. Engle et al (1986) use this model to estimate the effects of temperature on electricity demand, while Stock (1991) models the effect of the proximity of toxic waste on house prices. In both cases, the effect is highly nonlinear, while the large number of covariates make a fully nonparametric analysis infeasible. See also Olley and Pakes (1991). This specification also arises from various sample selection models, see Ahn and Powell (1990) and Newey, Powell and Walker (1990).

Notice that

$$Y_i - E(Y_i|Z_i) = \beta^T [X_i - E(X_i|Z_i)] + \epsilon_i.$$

Robinson (1988a) constructed a semiparametric estimator of β by replacing $g(Z_i) = E(X_i|Z_i)$ and $m(Z_i) = E(Y_i|Z_i)$ by nonparametric kernel estimators $\widehat{g}_h(Z_i)$ and $\widehat{m}_h(Z_i)$ and then letting

$$\widehat{\beta} = \left[\sum_{i=1}^{n} \{X_i - \widehat{g}_h(Z_i)\} \{X_i - \widehat{g}_h(Z_i)\}^T\right]^{-1} \sum_{i=1}^{n} \{X_i - \widehat{g}_h(Z_i)\} \{Y_i - \widehat{m}_h(Z_i)\}.$$

In fact, Robinson modified this estimator by trimming out observations for which the marginal density of Z was small. Robinson's estimator satisfies (47), provided the dimensions of Z are not too high relative to the order of the kernel being used (provided m and g are sufficiently smooth).

Linton (1992) establishes that the optimal bandwidth for $\widehat{\beta}$ is $O(n^{-2/9})$, when Z is scalar, and the resulting correction to (asymptotic) MSE of the standardised estimator is $O(n^{-7/9})$.

6.2 Heteroskedastic Non-Linear Regression

Consider the following nonlinear regression model:

$$Y_i = \tau(X_i; \beta) + \epsilon_i, \; i = 1, 2, .., n, \tag{50}$$

where $\tau(\bullet; \beta)$ is known, while $E(\epsilon_i|X_i) = 0$ and $Var(\epsilon_i|X_i) = \sigma^2(X_i)$, where $\sigma^2(\bullet)$ is of unknown functional form. Efficient estimation of β can be carried out using the *pseudo-likelihood* principle. Assuming that ϵ_i are iid normally distributed, the sample log-likelihood function is proportional to

$$\mathcal{L}\left\{\beta; \sigma^2(\bullet)\right\} = \sum_{i=1}^{n} [Y_i - \tau(X_i; \beta)]^2 \sigma^2(X_i)^{-1}, \tag{51}$$

when $\sigma^2(\bullet)$ is known. In the semiparametric situation we replace $\sigma^2(X_i)$ by a nonparametric estimator $\widehat{\sigma}^2(X_i)$, and then let $\widehat{\beta}$ minimize $\mathcal{L}\{\beta; \widehat{\sigma}^2(\bullet)\}$.

Carroll (1982) and Robinson (1987) examine the situation where $\tau(X; \beta) = \beta^T X$ in which case

$$\widehat{\beta} = \left\{\sum_{i=1}^{n} X_i X_i^T \widehat{\sigma}^2(X_i)^{-1}\right\}^{-1} \sum_{i=1}^{n} X_i Y_i \widehat{\sigma}^2(X_i)^{-1}. \tag{52}$$

They establish (under iid sampling) that $\widehat{\beta}$ is asymptotically equivalent to the infeasible GLS estimator based on (51) Remarkably, Robinson allows X to have unbounded support, yet did not need to trim out contributions from its tails: he used nearest neighbor estimators of $\sigma^2(\bullet)$ that always average over the same number of observations. Extensions of this model to the multivariate nonlinear $\tau(\bullet;\beta)$ case are considered in Delgado (1992), while Hidalgo (1992) allows both heteroskedasticity and serial correlation of unknown form. Applications include Melenberg and van Soest (1991) and Altug and Miller (1992), and Whistler (1988).

Carroll and Härdle (1989), Cavanagh (1989) and Linton (1993) develop second order theory for these estimators. In this case, the optimal bandwidth is $O(n^{-1/5})$ when X is scalar in which case the correction to the (asymptotic) MSE is $O(n^{-4/5})$.

6.3 Single Index Models

When the conditional distribution of a scalar variable Y given the d-dimensional predictor variable X depends on X only through the index $\beta^T X$, we say that this is a single index model.

One example is the single index regression model in which $E[Y|X = x] = m(x) = g(x^T\beta)$, but no other restrictions are imposed. Define the vector of average derivatives

$$\delta = E[m'(X)] = E[g'(X^T\beta)]\beta, \tag{53}$$

and note that δ determines β up to scale – as shown by Stoker (1986). Let $f(x)$ denote the density of X and l its vector of the negative log-derivatives (partial), $l = -\frac{\partial \log f}{\partial x} = -\frac{f'}{f}$ (l is also called the *score vector*). Under assumptions on f given in Powell, Stock and Stoker (1989), we can write

$$\delta = E[m'(X)] = E[l(X)Y], \tag{54}$$

and we estimate δ by $\widehat{\delta} = n^{-1}\sum_{i=1}^{n} \widehat{l}_H(X_i)Y_i$, where $\widehat{l}_H(x) = -\frac{\widehat{f}'_H}{\widehat{f}_H}(x)$ is an estimator of $l(x)$ based on a kernel density smoother with bandwidth matrix H. Furthermore, $g(\bullet)$ is estimated by a kernel estimator $\widehat{g}_h(\bullet)$ for which $\left\{\widehat{\delta}^T X_i\right\}_{i=1}^{n}$ is the right-hand side data.

Härdle and Stoker (1989) show that

$$\sqrt{n}(\widehat{\delta} - \delta) \Rightarrow N(0, \Sigma_\delta),$$

where $\Sigma_\delta = Var[l(X)\{Y - m(X)\} + m'(X)]$, while $\widehat{g}_h$ converges at rate $\sqrt{nh}$ – i.e. like a one dimensional function. Stoker (1991) proposed alternative estimators for δ based on first estimating the partial derivatives $m'(x)$ and then averaging over the observations. A Monte Carlo comparison of these methods is presented in Stoker and Villas-Boas (1992). Härdle, Hart, Marron, and Tsybakov (1992) develop a second order theory for $\widehat{\delta}$: in the scalar case, the optimal bandwidth h is $O(n^{-2/7})$ and the resulting correction to MSE is $O(n^{-1/7})$.

Another example is the *binary choice* model

$$Y_i = \mathrm{I}(\beta^T X_i + u_i \geq 0), \tag{55}$$

where (X, u) are iid. There are many treatments of this specification following the seminal paper of Manski (1975) – in which a slightly more general specification was considered. We assume also that u is independent of X with unknown distribution function $F(\bullet)$, in which case $\Pr[Y_i = 1|X_i] = F(\beta^T X_i) = E(Y_i|\beta^T X_i)$, i.e. $F(\bullet)$ is a regression function. In fact, (55) is a special case of (53). Applications include Das (1990), Horowitz (1991), and Melenberg and van Soest (1991).

Klein and Spady (1993) use the *profile likelihood* principle (see also Ichimura and Lee (1991)) to obtain (semiparametric) efficient estimates of β. When F is known, the sample log-likelihood function is

$$\mathcal{L}\{F(\beta)\} = \sum_{i=1}^{n} \left[Y_i \ln\left\{F(\beta^T X_i)\right\} + (1 - Y_i)\ln\left\{1 - F(\beta^T X_i)\right\}\right]. \tag{56}$$

For given β, let $\widehat{F}(\beta^T X)$ be the nonparametric regression estimator of $E(Y|\beta^T X)$. A feasible estimator $\widehat{\beta}$ of β is obtained as the minimizer of

$$\mathcal{L}\left\{\widehat{F}(\beta)\right\} = \sum_{i=1}^{n} \left[Y_i \ln\left\{\widehat{F}(\beta^T X_i)\right\} + (1 - Y_i)\ln\left\{1 - \widehat{F}(\beta^T X_i)\right\}\right]. \tag{57}$$

This can be carried out by standard numerical optimization techniques. The average derivative estimator can be used to provide initial consistent estimators of β, although it is not in general efficient, see Cosslett (1987). Note that to establish $\sqrt{n}$-consistency, it is necessary to employ bias reduction techniques such as higher order kernels as well as to trim out contributions from sparse regions. Note also that $\widehat{\beta}$ is not as efficient as the MLE obtained from (56).

We examined the performance of the average derivative estimator on a simulated dataset, where

$$\begin{aligned}
&X \sim N(0, I_2)\\
&\Pr(Y = 1|X = x) = \Lambda(\beta^T x) + 0.6\phi'(\beta^T x)\\
&\beta = (1, 1)^T,
\end{aligned}$$

while Λ and ϕ are the standard logit and normal density functions respectively. A sample of size $n = 200$ was generated, and the bivariate density function was estimated using a *Nadaraya-Watson* estimator with bandwidth matrix $H = \text{diag}\{0.99, 0.78\}$. This example is taken from Härdle and Turlach (1992). The estimation of δ and its asymptotic covariance matrix $\widehat{\Sigma}_\delta$ was done with XploRe macro `adefit`. For this example $\delta = (0.135, 0.135)^T$, and

$$\widehat{\delta} = \begin{pmatrix} 0.124 \\ 0.118 \end{pmatrix}, \qquad \widehat{\Sigma}_\delta = \begin{pmatrix} 0.188 & 0.036 \\ 0.036 & 0.206 \end{pmatrix}.$$

Figure 9 shows the estimated regression function $\widehat{g}_h(\widehat{\delta}^T X_i)$.

These results allow us to test some hypotheses formally using a Wald statistic (see Stoker (1992), pp. 53–54). In particular, to test the restriction $R\delta = r_0$, the Wald statistic

$$W = n(R\widehat{\delta} - r_0)^T (R\widehat{\Sigma}_\delta R^T)^{-1} (R\widehat{\delta} - r_0)$$

is compared to a $\chi^2(\text{rank } R)$ critical value. Table 3 gives some examples for this technique.

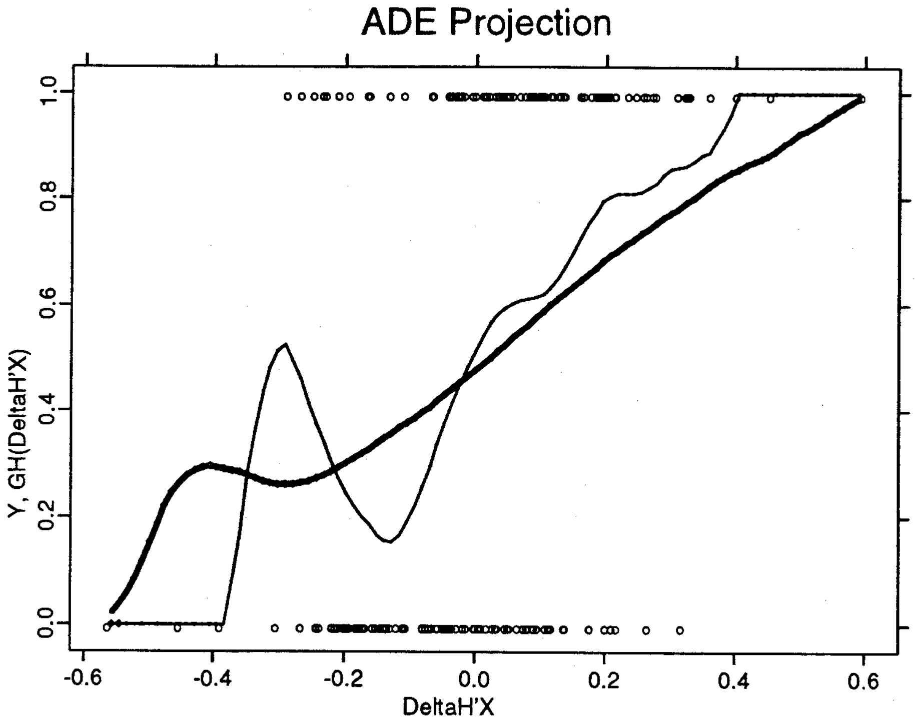

Figure 9. For the simulated dataset: $\widehat{\delta}^T X$ versus Y and two estimates of $g(\widehat{\delta}^T X_i)$ are shown. The thick line shows the Nadaraya-Watson estimator with a bandwidth $h = 0.3$, while for the thin line $h = 0.1$ was chosen.

Restriction	Value W	d.f.	$P[\chi^2(\text{d.f.}) > W]$
$\delta^1 = \delta^2 = 0$	25.25	2	0
$\delta^1 = \delta^2 = 0.135$	0.365	2	0.83
$\delta^1 = \delta^2$	0.027	1	0.869

Table 3. Wald Statistics for some restrictions on δ.

7 Conclusions

The nonparametric methods we have examined are especially useful when the variable over which the smoothing takes place is one dimensional. In this case, the relationship can be plotted and evaluated, while the estimators converge at rate $\sqrt{nh}$.

In high dimensions these methods are less attractive due to the slower rate of convergence and the lack of simple but comprehensive graphs. In this case, there are a number of restricted structures that can be employed including the nonparametric additive models of Hastie and Tibshirani (1990), or semiparametric models like the partially linear and index models examined in Section 6.

References

[1] ABRAMSON, I. (1982): "On bandwidth variation in kernel estimates – a square root law," *Annals of Statistics* 10, 1217-1223.

[2] AHN, H., and J.L. POWELL (1990): "Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." Unpublished Manuscript, University of Wisconsin.

[3] AKAIKE, H. (1970): "Statistical predictor information," *Annals of the Institute of Statistical Mathematics* 22, 203–17.

[4] AKAIKE, H. (1974): "A new look at the statistical model identification." *IEEE Transactions of Automatic Control AC* 19, 716–23.

[5] ALTUG, S. and R.A. MILLER (1992): "Human Capital, Aggregate shocks and panel data estimation" Unpublished manuscript, University of Minnesota.

[6] ANAND, S., C.J. HARRIS, and O. LINTON (1993): "On the concept of ultrapoverty," Harvard Center for Population Studies Working paper 93-02.

[7] ANDREWS, D.W.K. (1989): "Semiparametric Econometric Models: I Estimation." Cowles Foundation Discussion paper 908.

[8] ANDREWS, D.W.K. (1991): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametiic Regression Models." *Econometrica* 59, 307-346.

[9] ANDREWS, D.W.K., and Y.-J. WHANG (1990): "Additive and Interactive Regression Models: Circumvention of the Curse of Dimensionality," *Econometric Theory* 6, 466-479.

[10] ANSLEY, C.F., R. KOHN, and C. WONG (1993): "Nonparametric spline regression with prior information," *Biometrika* 80, 75-88.

[11] BANKS, J., R. BLUNDELL, and A. LEWBEL (1993): "Quadratic Engel curves, welfare measurement and consumer demand," Institute for Fiscal Studies 92-14.

[12] BHATTACHARYA, P.K., and A.K. GANGOPADHYAY (1990): "Kernel and Nearest-Neighbor Estimation of a Conditional Quantile," *Annals of Statistics* 18, 1400-15.

[13] BICKEL, P.J., C.A.J. KLAASSEN, Y. RITOV, and J.A. WELLNER (1992): *Efficient and Adaptive Inference in Semiparametric Models.* Forthcoming Monograph. Baltimore: Johns Hopkins University Press.

[14] BIERENS, H.J. (1987): "Kernel Estimators of Regression Functions." in *Advances in Econometrics: Fifth World Congress,* Vol 1, ed. by T.F. Bewley. Cambridge University Press.

[15] BIERENS, H.J., and H.A. POTT-BUTER (1990): "Specification of household Engel curves by nonparametric regression," *Econometric Reviews* 9, 123-184.

[16] BRILLINGER, D.R. (1980): *Time Series, Data analysiś and Theory.* Holden-Day.

[17] CARROLL, R.J. (1982): "Adapting for Heteroscedasticity in Linear Models," *Annals of Statistics* 10, 1224-1233.

[18] CARROLL, R.J., and W. HÄRDLE (1989): "Second Order Effects in Semiparametric Weighted Least Squares Regression." *Statistics* 20, 179-186.

[19] CAVANAGH, C.L. (1989): "The cost of adapting for heteroskedasticity in linear models," Unpublished manuscript, Harvard University.

[20] CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Semiparametric models with censoring," *Journal of Econometrics* 32, 189-218.

[21] CHAMBERS, J.M., W.S. CLEVELAND, B. KLEINER, and P.A. TUKEY (1983): *Graphical Methods for Data Analysis.* Duxburry Press.

[22] CHANDA, K.C. (1974): "Strong mixing properties of linear stochastic process." *Journal of Applied Probabilities* 11, 401–408.

[23] CHAUDHURI, P. (1991): "Global nonparametric estimation of conditional quantile functions and their derivatives," *Journal of Multivariate Analysis* 39, 246-269.

[24] CLEVELAND, W.S. (1979): "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74, 829-836.

[25] COSSLETT, S.R. (1987): "Efficiency bounds for Distribution-free estimators of the Binary Choice and the Censored Regression model," *Econometrica* 55, 559-587.

[26] COX, D.R., and D.V. HINKLEY (1974): *Theoretical Statistics.* Chapman and Hall.

[27] CRAVEN, P., and WAHBA, G. (1979): "Smoothing noisy data with spline functions," *Numer. Math.* 31, 377–403.

[28] DANIELL, P.J. (1946): "Discussion of paper by M.S.Bartlett," *Journal of the Royal Statistical Society Supplement* 8:27.

[29] DAS, S. (1990): "A Semiparametric Structural Analysis of the Idling of Cement Kilns." *Journal of Econometrics* 50, 235-256.

[30] DEATON, A.S. (1991): "Rice-prices and income distribution in Thailand: a nonparametric analysis," *Economic Journal* 99, 1-37.

[31] DEATON, A.S. (1993): "Data and econometric tools for development economics," Forthcoming in *The Handbook of Development Economics,* Volume III, Eds J.Behrman and T.N.Srinavasan.

[32] DELGADO, M. (1992): "Semiparametric Generalised Least Squares in the Multivariate Nonlinear Regression Model." *Econometric Theory* 8, 203-222.

[33] DIEBOLD, F., and J. NASON (1990): "Nonparametric Exchange rate prediction?" *Journal of International Economics* 28, 315-332.

[34] DOUKHAN, P. and GHINDES, M. (1980): "Estimation dans le processus $X_n = f(X_{n-1}) + \epsilon_n$," *Comptes Rendus, Académie des Sciences de Paris* 297, Série A, 61–4.

[35] ELBADAWI, I., A.R. GALLANT, and G. SOUZA (1983): "An elasticity can be estimated consistently without a priori knowledge of functional form," *Econometrica* 51, 1731-1751.

[36] ENGLE, R.F., and R. GARDINER (1976): "Some Finite Sample Properties of Spectral Estimators of a Linear Regression." *Econometrica* 44, 149-165.

[37] ENGLE, R.F., C.W.J. GRANGER, J. RICE, and A. WEISS (1986): "Semiparametric Estimates of the Relationship Between Weather and Electricity Sales," *Journal of the American Statistical Association* 81, 310-320.

[38] ENGLE, R.F., D.F. HENDRY, and J.F. RICHARD (1983): "Exogeneity," *Econometrica* 51, 277-304.

[39] EUBANK, R.L. (1988): *Smoothing Splines and Nonparametric Regression.* Marcel Dekker.

[40] FAMA, E.F. (1965): "The behavior of stock prices," *Journal of Business* 38, 34-105.

[41] FAMILY EXPENDITURE SURVEY, Annual Base Tapes (1968-1983). Department of Employment, Statistics Division, Her Majesty's Stationary Office, London 1968-1983.

[42] FAN, J. (1992): "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association* 87, 998-1004.

[43] FAN, J., N.E. HECKMAN, and M.P. WAND (1992): "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-Likelihood Functions," University of British Columbia Working paper 92-028.

[44] FAN, J., and I. GIJBELS (1992): "Spatial and Design Adaptation: Variable order approximation in function estimation," *Institute of Statistics Mimeo Series*, no 2080, University of North Carolina at Chapel Hill.

[45] FIX, E., and J.L. HODGES (1951): "Discriminatory analysis, nonparametric estimation: consistency properties," *Report no 4, Project no 21-49-004,* USAF School of Aviation Medicine, Randolph Field, Texas.

[46] GALLANT, A.R., D.A. HSIEH, and G.E. TAUCHEN (1991): "On Fitting a Recalcitrant Series: The pound/dollar Exchange Rate, 1974-1983," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen.

[47] GALLANT, A.R., and G. SOUZA (1991): "On the asymptotic normality of Fourier flexible form estimates," *Journal of Econometrics* 50, 329-353.

[48] GALLANT, A.R., and H. WHITE (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models.* Blackwell, Oxford.

[49] GARODETSKII, V.V. (1977): "On the strong mixing condition for linear process," *Theory of Probability and its Applications* 22, 411–413.

[50] GASSER, T. and H.G. MÜLLER (1984): "Estimating regression functions and their derivatives by the kernel method," *Scandinavian Journal of Statistics* 11, 171–85.

[51] GASSER, T., H.G. MÜLLER, and V. MAMMITZSCH (1985): "Kernels for nonparametric curve estimation,"*Journal of the Royal Statistical Society Series B* 47, 238–52.

[52] GOZALO, P.L., (1989): "Nonparametric analysis of Engel curves: estimation and testing of demographic effects," Brown University, Department of Economics Working paper 92-15.

[53] GYORFI, L., W. HÄRDLE, P. SARDA, and P. VIEU (1990): *Nonparametric Curve Estimation from Time Series.* Lecture Notes in Statistics, 60. Springer-Verlag, Heidelberg, New York.

[54] HALL, P. (1992): *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

[55] HALL, P. (1993): "On Edgeworth Expansion and Bootstrap Confidence Bands in Nonparametric Curve Estimation," *Journal of the Royal Statistical Society Series B* 55, 291-304.

[56] HALL, P., and I. JOHNSTONE (1992): "Empirical functional and efficient smoothing parameter selection," (with discussion). *Journal of the Royal Statistical Society Series B.* 54, 475-530.

[57] HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Econometric Society Monographs 19, Cambridge University Press.

[58] HÄRDLE, W. (1991). *Smoothing Techniques with Implementation in S.* Springer-Verlag, Heidelberg, New York, Berlin.

[59] HÄRDLE, W. and R.J. CARROLL (1990): "Biased cross-validation for a kernel regression estimator and its derivatives," *Österreichische Zeitschrift für Statistik und Informatik* 20, 53-64.

[60] HÄRDLE, W., P. HALL, and H. ICHIMURA (1993): "Optimal Smoothing in Single Index Models." *Annals of Statistics* 21, to appear.

[61] HÄRDLE, W., P. HALL and J.S. MARRON (1988): "How far are automatically chosen regression smoothing parameters from their optimum?" (with discussion). *Journal of the American Statistical Association* 83, 86–99.

[62] HÄRDLE, W., P. HALL and J.S. MARRON (1992): "Regression smoothing parameters that are not far from their optimum" *Journal of the American Statistical Association* 87, 227–233.

[63] HÄRDLE, W., J. HART, J.S. MARRON, and A.B. TSYBAKOV (1992): "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association* 87, 218-226.

[64] HÄRDLE, W., and M. JERISON (1991): "Cross Section Engel Curves over Time," *Recherches Economiques de Louvain* 57, 391-431.

[65] HÄRDLE, W., and J.S. MARRON (1985): "Optimal bandwidth selection in nonparametric regression function estimation," *Annals of Statistics* 13, 1465-81.

[66] HÄRDLE, W., and M. MÜLLER (1993): "Nichtparametrische Glättungsmethoden in der alltäglichen statistischen Praxis," *Allg. Statistiches Archiv* 77, 9-31.

[67] HÄRDLE, W., and D.W. SCOTT (1992): "Smoothing in Low and High Dimensions by Weighted Averaging Using Rounded Points," *Computational Statistics* 1, 97-128.

[68] HÄRDLE, W., and T.M. STOKER (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association* 84, 986-995.

[69] HÄRDLE, W., and B.A. TURLACH (1992): "Nonparametric Approaches t Generalized Linear Models," In: Fahrmeir, L., Francis, B., Gilchrist, R., Tutz, G. (Eds.) *Advances in GLIM and Statistical Modelling*, Lecture Notes in Statistics, 78. Springer-Verlag, New York.

[70] HÄRDLE, W., and P. VIEU (1991): "Kernel regression smoothing of time series," *Journal of Time Series Analysis* 13, 209-232.

[71] HART, J., and P. VIEU (1990): "Data-driven bandwidth choice for density estimation based on dependent data," *Annals of Statistics* 18, 873–890.

[72] HART, D., and T.E. WEHRLY (1986): "Kernel regression estimation using repeated measurements data," *Journal of the American Statistical Association* 81, 1080–8.

[73] HASTIE, T.J., and R.J. TIBSHIRANI (1990): *Generalized Additive Models* Chapman and Hall.

[74] HAUSMAN, J.A., and W.K. NEWEY (1992): "Nonparametric estimation of exact consumer surplus and deadweight loss," MIT, Department of Economics Working paper 93-2.

[75] HIDALGO, J. (1992): "Adaptive Estimation in Time Series Models with Heteroscedasticity of Unknown Form." *Econometric Theory* 8, 161-187.

[76] HILDENBRAND, K., and W. HILDENBRAND (1986): "On the mean income effect: a data analysis of the U.K. family expenditure survey," in *Contributions to Mathematical Economics,* eds W.Hildenbrand and A.Mas-Colell. North Holland.

[77] HILDENBRAND, W., and A. KNEIP (1992): "Family expenditure data, heteroscedasticity and the law of demand," Universität Bonn Discussion paper A-390.

[78] HOROWITZ, J.L. (1991): "Semiparametric estimation of a work-trip mode choice model," University of Iowa Department of Economics Working paper 91-12.

[79] HSIEH, D.A., and C.F. MANSKI (1987): "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression." *Annals of Statistics* 15, 541-551.

[80] HUSSEY, R. (1992): "Nonparametric evidence on asymmetry in business cycles using aggregate employment time series," *Journal of Econometrics* 51, 217-231.

[81] ICHIMURA, H., and L.F. LEE (1991): "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen.

[82] JONES, M.C. (1985): "Discussion of the paper by B.W.Silverman," *Journal of the Royal Statistical Society Series B* 47, 25-26.

[83] JONES, M.C. (1989): "Discretized and interpolated Kernel Density Estimates, " *Journal of the American Statistical Association* 84, 733-741.

[84] JONES, M.C., and P.J. FOSTER (1993): "Generalized jacknifing and higher order kernels," Forthcoming in *Journal of Nonparametric Statistics.*

[85] JONES, M.C., O. LINTON, and J.P. NIELSEN (1993): "A Multiplicative bias reduction method," Preprint, Nuffield College, Oxford.

[86] JONES, M.C., J.S. MARRON, and S.J. SHEATHER (1992): "Progress in databased selection for Kernel Density estimation," *Australian Graduate School of Management* Working paper no 92-014.

[87] KLEIN, R.W., and R.H. SPADY (1991): "An Efficient Semiparametric Estimator for Binary Choice Models." *Econometrica* 61, 387-421.

[88] KOENKER, R., and G. BASSETT (1978): "Regression quantiles," *Econometrica* 46, 33-50.

[89] KOENKER, R., P. NG, and S. PORTNOY (1993): "Quantile Smoothing Splines," Forthcoming in *Biometrika*

[90] LEWBEL, A. (1991): "The Rank of Demand Systems: Theory and Nonparametric Estimation," *Econometrica* 59, 711-730.

[91] LI, K.-C. (1985): "From Stein's unbiased risk estimates to the method of generalized cross-validation." *Annals of Statistics* 13, 1352–77.

[92] LINTON, O.B. (1991): "Edgeworth Approximation in Semiparametric Regression Models" PhD thesis, Department of Economics, UC Berkeley.

[93] LINTON, O.B. (1992): "Second Order Approximation in the Partially Linear Model," Cowles Foundation Discussion Paper no 1065.

[94] LINTON, O.B. (1993): "Second Order Approximation in a linear regression with heteroskedasticity of unknown form." Nuffield College Discussion paper no 75.

[95] LINTON, O.B. and J.P. NIELSEN (1993): "A Multiplicative Bias Reduction Method for Nonparametric Regression," Forthcoming in *Statistics and Probability Letters.*

[96] McFADDEN, D. (1985): "Specification of Econometric models," Econometric Society, Presidential Address.

[97] MACK, Y.P. (1981): "Local properties of k-NN regression estimates," *SIAM J. Alg. Disc. Meth.* 2, 311–23.

[98] MANDELBROT, B. (1963): "The variation of certain speculative prices," *Journal of Business* 36, 394-419.

[99] MANSKI, C.F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3, 205-228.

[100] MARRON, J.S., and D. NOLAN (1989): "Canonical kernels for density estimation," *Statistics and Probability Letters* 7, 191-195.

[101] MARRON, J.S., and M.P. WAND (1992): "Exact Mean Integrated Squared Error." *Annals of Statistics* 20, 712-736.

[102] MEESE, R.A., and A.K. ROSE (1991): "An empirical assessment of nonlinearities in models of exchange rate determination," *Review of Economic Studies* 80, 603-619.

[103] MELENBERG, B., and A. van SOEST (1991): "Parametric and semi-parametric modelling of vacation expenditures," CentER for Economic Research, Discussion paper no 9144, Tilburg, Holland.

[104] MIZRACH, B. (1992): "Multivariate nearest-neighbor forecasts of EMS exchange rates," *Journal of Applied Econometrics* 7, 151-163.

[105] MÜLLER, H.G. (1987): "On the asymptotic mean square error of L_1 kernel estimates of C_∞ functions," *Journal of Approximation Theory* 51, 193-201.

[106] MÜLLER, H.G. (1988): *Nonparametric Regression Analysis of Longitudinal Data.* Lecture Notes in Statistics, Vol. 46. Heidelberg/New York: Springer-Verlag.

[107] NADARAYA, E.A. (1964): "On estimating regression," *Theory of Probability and its Applications* 10, 186-190.

[108] NEWEY, W.K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics* 5, 99-135.

[109] NEWEY, W.K., J.L. POWELL, and J.R. WALKER (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results," *American Economic Review Papers and Proceedings* 80, 324-328.

[110] OLLEY, G.S., and A. PAKES (1991): "The Dynamics of Productivity in the Telecommunications Equipment Industry," Unpublished manuscript, Yale University.

[111] PAGAN, A.R., and Y.S. HONG, (1991): "Nonparametric Estimation and the Risk Premium," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen.

[112] PAGAN, A.R., and W. SCHWERT (1990): "Alternative models for conditional stock volatility," *Journal of Econometrics* 45, 267-290.

[113] PAGAN, A.R., and A. ULLAH (1988): "The Econometric Analysis of models with risk terms," *Journal of Applied Econometrics* 3, 87-105.

[114] PARK, B.U., and B.A. TURLACH (1992): "Practical performance of several data-driven bandwidth selectors (with discussion)," *Computational Statistics* 7, 251-271.

[115] PHILLIPS, P.C.B. (1991): "Spectral Regression for Cointegrated Time Series." in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen.

[116] POWELL, J.L., J.H. STOCK, and T.M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.

[117] POWELL, J.L., and T.M. STOKER (1991): "Optimal Bandwidth Choice for Density-weighted averages," Unpublished manuscript, Princeton University.

[118] PRAKASA RAO, B.L.S. (1983): *Nonparametric Functional Estimation.* Academic Press.

[119] RICE, J.A. (1984): "Bandwidth choice for nonparametric regression" *Annals of Statistics* 12, 1215–30.

[120] ROBB, A.L., L. MAGEE, and J.B. BURBIDGE (1992): "Kernel smoothed consumption-age quantiles," *Canadian Journal of Economics* 25, 669-680.

[121] ROBINSON, P.M. (1983): "Nonparametric Estimators for Time Series." *Journal of Time Series Analysis* 185-208.

[122] ROBINSON, P.M. (1987): "Asymptotically Efficient Estimation in the Presence of Heteroscedasticity of Unknown form." *Econometrica* 56, 875-891.

[123] ROBINSON, P.M. (1988a): "Root-N-Consistent Semiparametric Regression," *Econometrica* 56, 931-954.

[124] ROBINSON, P.M. (1988b): "Semiparametric Econometrics: A Survey," *Journal of Applied Econometrics* 3, 35-51.

[125] ROBINSON, P.M. (1991): "Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models." *Econometrica* 59, 1329-1364.

[126] ROSENBLATT, M. (1956): "Remarks on some nonparametric estimates of a density function," *Annals of Mathematical Statistics* 27, 642-669.

[127] RUPPERT, D., and M.P. WAND (1992): "Multivariate Locally Weighted Least Squares Regression," Rice University, Technical Report no 92-4.

[128] SCHUSTER, E.F. (1972): "Joint asymptotic distribution of the estimated regression function at a finite number of distinct points," *Annals of Mathematical Statistics* 43, 84-8.

[129] SENTANA, E., and S. WADHWANI (1991): "Semi-parametric Estimation and the Predictability of Stock Returns: Some Lessons from Japan," *Review of Economic Studies* 58, 547-563.

[130] SHIBATA, R. (1981): "An optimal selection of regression variables," *Biometrika*, 68, 45–54.

[131] SILVERMAN, B.W. (1984): "Spline smoothing: the equivalent variable kernel method." *Annals of Statistics* 12, 898–916.

[132] SILVERMAN, B.W. (1985): "Some aspects of the Spline Smoothing approach to Non-parametric Regression Curve Fitting," *Journal of the Royal Statistical Society Series B* 47, 1-52

[133] SILVERMAN, B.W. (1986). *Density estimation for statistics and data analysis.* London: Chapman and Hall.

[134] STOCK, J.H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association* 84, 567-576.

[135] STOCK, J.H. (1991): "Nonparametric Policy Analysis: An Application to Estimating Hazardous Waste Cleanup Benefits," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen.

[136] STOKER, T.M. (1986): "Consistent Estimation of Scaled Coefficients." *Econometrica* 54, 1461-1481.

[137] STOKER, T.M. (1991): "Equivalence of direct, indirect, and slope estimators of average derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Eds Barnett, Powell, and Tauchen.

[138] STOKER, T.M. (1992): *Lectures on Semiparametric Econometrics.* CORE Lecture Series.

[139] STOKER, T.M., and J.M. VILLAS-BOAS (1992): "Monte Carlo Simulation of Average Derivative Estimators," Unpublished manuscript, MIT.

[140] STONE, C.J. (1982): "Optimal global rates of convergence for nonparametric regression," *Annals of Statistics* 10, 1040-1053.

[141] STRAUSS, J., and D. THOMAS (1990): "The shape of the calorie-expenditure curve," Unpublished manuscript, Rand Corporation, Santa Monica.

[142] STUTE, W. (1986): "Conditional Empirical Processes," *Annals of Statistics* 14, 638-647.

[143] TIBSHIRANI, R. (1984): "Local Likelihood estimation," PhD Thesis, Stanford University.

[144] TIKHONOV, A.N. (1963): "Regularization of incorrectly posed problems," *Soviet Math.*, 4, 1624–1627.

[145] TURLACH, B.A. (1992): "On discretization methods for average derivative estimation," CORE Discussion Paper no. 9232, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

[146] VAPNIK, V. (1982): *Estimation of Dependencies Based on Empirical Data.* Heidelberg, New York, Berlin: Springer-Verlag.

[147] WAHBA, G. (1990): *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics, no 59.

[148] WATSON, G.S. (1964): "Smooth regression analysis," *Sankhya Series A* 26, 359-372.

[149] WHISTLER, D. (1988): "Semiparametric ARCH Estimation of Intra-Daily Exchange Rate Volatility," Unpublished manuscript, London School of Economics.

[150] WHITTAKER, E.T. (1923): "On a new method of graduation," *Proc. Edinburgh Math.Soc* 41, 63-75.

[151] XploRe (1993): An interactive statistical computing environment. *Available from XploRe Systems, Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultät, Humboldt-Universität zu Berlin, D 10178 Berlin, Germany*