Applied Nonparametric Methods*

Wolfgang Härdle**

January 1992

ABSTRACT

In this note we review different approaches to nonparametric regression. Kernel estimators are motivated from local averaging, solving ill-posed problems and weighting of binned data. Kernel estimators are compared to k-NN estimators and splines. The choice of smoothing parameter is discussed and finally the method is applied for nonparametric prediction of time series.

This work was financed, in part, by contrat No 26 of the programme "Pôle d'attraction interuniversitaire" of the Belgian government.

---

* This work was prepared for Chapter 4 of the fourth *Handbook of Econometrics*, North-Holland, edited by R. Engle and D. McFadden

** CORE, Université Catholique de Louvain, and CentER, Tilburg.

ISSN-0771 3894

Applied Nonparametric Methods

Wolfgang Härdle

CORE and CentER

September 1991

Prepared for
Chapter 4 of the fourth *Handbook of Econometrics*
North-Holland
Editors: R. Engle, D. McFadden

Table of Contents

1. The Kernel Method

1.1 Kernels as windows

1.2 Kernels from averaging binned data

1.3 Kernels and ill-posed problems

1.4 Properties of kernels

1.5 Regression curve estimation

2. *k*-Nearest Neighbor Estimates

2.1 Ordinary k-NN estimates

2.2 Symmetrized k-NN estimates

3. Spline Estimates

3.1 The cubic spline

3.2 Kernels, k-NN, and splines

4. Choice of Smoothing Parameter

4.1 Crossvalidation

4.2 Other data driven selectors

4.3 Canonical kernels

5. Application to Time Series

5.1 Prediction

5.2 Correlated errors

1. The Kernel Method

There is general agreement on what we mean with a parametric econometric model: The distribution of observed data is indexed by a set of parameters. The best model "explaining" the data is found by determining the parameters that minimize an empirical distance between the data and the model. The squared deviation leads to least squares estimation of parameters. The minimization of the Kuhlback–Leibler distance is equivalent to the Maximum Likelihood method. In a nonparametric model we do not have, and thus do not estimate, parameters. We rather estimate the distribution or functionals of it directly from the data without explicit reference to a proposed (low-dimensional parametric) model.

How can this be done? If we are only interested in the distribution of the observations itself, we would take the empirical distribution function but it would not tell us a lot on the relationship between variables. For instance, it is hard to infer directly from the joint distribution function of two variables if one variable influences another in a specific way. Such an influence could be that "on the average" one variable is monotone dependent on the other, or whether a variable conditioned on the other is bigger than a certain level. In this case we are asking after the structure and behavior of the regression function, a functional of the joint distribution.

It is simplest to describe the nonparametric approach in the setting of density estimation, so we begin with that. A typical economic application for this is the estimation of income densities. Suppose we are given i.i.d. observations $\{X_i\}_{i=1}^n \in \mathbb{R}$ with density f. The functional we are interested in, for the moment, is the density $f(x)$ at a fixed point x. The distribution function of x

$$F(x) = \int_{-\infty}^{x} f(u)du$$

can be estimated by the empirical distribution function (edf)

$$F_n(x) = n^{-1} \sum_{i=1}^{n} \mathrm{I}\ (x_i \leq x).$$

This estimate is a step function and cannot be differentiated to obtain an approximation to $f(x)$. If the edf were smooth though, one could hope that also the derivative would give a good estimator for $f(x)$. Since this is not the case we need to smooth.

We present three approaches here for smoothing in density estimation. The first stems directly from the histogram, the second one from averaging histograms, the third one from considering the estimation of $f(x)$ as an ill-posed problem.

1.1 Kernels as windows

Suppose that we are interested in estimating $f(0)$. If f is smooth in a small neighborhood $[-h, h]$ of $x = 0$, we justify by the mean value theorem,

$$2h \cdot f(0) \approx \int_{-h}^{h} f(u)du = P(x \in [-\frac{h}{2}, \frac{h}{2}]). \tag{1.1}$$

The right-hand side of (1.1) can be approximated by counting the number of X_i's in this small interval of length $2h$. Let $K(u) = \frac{1}{2}$ I $(|u|) \leq 1)$, then (1.1) can be rewritten as

$$f(0) \approx (nh)^{-1} \sum_{i=1}^{n} K(|X_i| \leq h).$$

This argument can be repeated for arbitrary x. An estimator for $f(x)$ is therefore

$$\widehat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) \tag{1.2}$$

with $K_h(\bullet) = h^{-1}K(\bullet/h)$. This is precisely the kernel density estimator of $f(x)$ with kernel $K(u) = \frac{1}{2}$ I $(|u| \leq 1)$ and *bandwidth* h. A histogram is a kernel estimator evaluated only at the discrete bin midpoints with binlength h. The kernel estimator is obtained by "sliding the kernel window" continuously over the range of observations.

We motivated this particular estimator (1.2) by a smoothness assumption on f. The estimator $\widehat{f}_h$ with the step function kernel $\frac{1}{2}$ I $(|u| \leq 1)$ is by its very nature a rough approximation to f. A smoother approximation can be obtained by choosing a smoother "window function" K as kernel. One example is the so-called quartic kernel

$$K(u) = \frac{15}{16}(1 - u^2)^2 \text{ I } (|u| \leq 1). \tag{1.3}$$

Obviously, the smoothness of the kernel determines the smoothness of the density estimate $\widehat{f}_h$.

How does this work with real data?

Figure 1 shows a scatterplot of log quantities vs. log prices for sardines during a trade period of August–September in 1987 in the Marseille fish market. For the moment we are interested in estimating the two marginal densities.

Figure 2a,b shows a kernel density estimate for the distribution of fish prices and sold quantity prices. For these estimates, a bandwidth of $h = 0.247$ was selected for the quantity data and $h = 0.085$ for the price data. These smoothing parameters correspond to the "rule of thumb" estimates of Silverman (1986).

Figure 1. Log quantities vs. log prices for sardines during the months of August–September in 1987 in the Marseille fish market. The solid line is a kernel estimate of the regression of log quantities on log price.

Figure 2a. Kernel density estimate of fish quantity data. Quartic kernel, h=0.247, n=3812. Data kindly provided by Alan Kirman.

Figure 2b. Kernel density estimate of price quantity data. Quartic kernel, h=0.085, n=3812. Data kindly provided by Alan Kirman.

The bimodal structure of the quantity density becomes evident in this plot. A parametric model like the log Normal distribution for the quantity data would not be able to capture that particular feature of the quantity distribution. The bimodal structure in fact gave reasons to investigate this data further and led to certain buyer/seller combinations.

1.2 Kernels from averaging binned data

We have already introduced the histogram with binwidth $2h$. For that definition, we have tacitly assumed that the bin mesh $B_j = [(j-1)h, jh)$ was centered at the *origin* $x_0 = 0$. The sensitivity of histograms with respect to choice of origin is well known, see e.g., Härdle (1991, Fig. 1.16). For the fish data of Figure 1, the dependence on the origin becomes evident if we compute histograms with binwidth $h = 0.5$ and origins $x_0 = 0.1, 0.2, 0.3, 0.4$. The five histograms with this same binwidth but different origins are shown in Figure 3.

Figure 3. Five histograms of Marseille fish data. Origins in x_0=0,0.1,...,0.4 and binwidth h=0.5.

Although the histograms use the same amount of smoothing, they give different impressions on the location of the peaks in the density. Note that in comparison with Figure 2, the binwidth is larger but, for ease of interpretation, only the bin centers have been connected. An ensemble of histograms, with different origins,

becomes independent of their origins if they are averaged. To this end, let

$$B_{j,\ell} = \left[\left(j - 1 + \frac{\ell}{M}\right) h, \left(j + \frac{\ell}{M}\right) h\right), \ell \in \{0, \ldots, M-1\} \tag{1.4}$$

denote a smaller bin mesh with origin $\frac{\ell}{M}$. We have now M histograms

$$\widehat{f}_{h,\ell}(x) = (nh)^{-1} \sum_{i=1}^{n} \left(\sum_j \mathrm{I}\left(x \in B_{j,\ell}\right) \mathrm{I}\left(X_i \in B_{j,\ell}\right) \right), \ell = 0, \ldots, M-1.$$

The idea is to average these M histograms to obtain independence of the origin $x_{0,\ell} = \frac{\ell}{M}$,

$$\begin{aligned} \widehat{f}_h(x) &= M^{-1} \sum_{\ell=0}^{M-1} (nh)^{-1} \sum_{i=1}^{n} \left(\sum_j \mathrm{I}\left(x \in B_{j,\ell}\right) \mathrm{I}\left(X_i \in B_{j,\ell}\right) \right) \\ &= n^{-1} \sum_{i=1}^{n} \left((Mh)^{-1} \sum_{\ell=0}^{M-1} \sum_j \mathrm{I}\left(x \in B_{j,\ell}\right) \mathrm{I}\left(X_i \in B_{j,\ell}\right) \right) \\ &= (nh)^{-1} \sum_{i=1}^{n} \left(\sum_j \mathrm{I}\left(x \in B_j^*\right) \sum_{k=1-M}^{M-1} \mathrm{I}\left(X_i \in B_{j+k}^*\right)(M - |k|) \right) \end{aligned} \tag{1.5}$$

with $B_j^* = \left[\frac{jh}{M}, \frac{(j+1)h}{M}\right)$ the smaller bins with width $\delta = \frac{h}{M}$. From (1.5) we see that averaging the shifted histograms leads again to a kernel like averaging process. Indeed (1.5) can be rewritten as

$$(nh)^{-1} \sum_j \mathrm{I}\left(x \in B_j^*\right) \sum_{k=1-M}^{M-1} w_M(k) n_{j+k}, \tag{1.6}$$

with $n_j = \sum_{i=1}^{n} \mathrm{I}\left(X_i \in B_j^*\right)$ and effective kernel weight $w_M(k) = 1 - \frac{|k|}{M}$. What we have used here is the technique of WARPing (Weighted Averaging of Rounded Points), see Härdle and Scott (1990). It consists of discretizing a kernel and then weighting the frequencies in the fine bins B_j^*. In (1.5) above the triangle kernel $K(u) = (1 - |u| \mathrm{I}(|u| \leq 1))$ was used. The kernel estimate (1.2) is obtained by letting M tend to infinity, for details, see Chapters 1, 2 in Härdle (1991). The effective weight function for the quartic kernel (1.3), for example, is given by $w_M(\ell) = \frac{15M^4}{16M^4-1}\left(1 - \left(\frac{\ell}{M}\right)^2\right)^2$, $\ell = 1 - M, \ldots, 0, \ldots, M-1$. So we see that forming a weighted average of histograms leads to a kernel estimate.

1.3 Kernels and ill-posed problems

Taking the derivative of a distribution F is a linear operation, $Af = F$. In more mathematical language one calls the equation

$$Af = \int_{-\infty}^{\infty} \mathrm{I}\ (u \leq x)\ f(u)du = F(x), \tag{1.7}$$

a Fredholm equation with the integral operator $Af = \int_{-\infty}^{x} f$. Estimating the density is the same as inverting (1.7). This 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 (1.7) is not continuous, see Vapnik (1982, p. 22).

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

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

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

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

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

The minimum condition for a solution $\widehat{f}$ is

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

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

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

with $\Gamma(\omega)$ the Fourier transform of $\widehat{f}(\bullet)$.

Solving this equation for Γ and then, applying the inverse Fourier transform, we obtain

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

Thus we obtain a kernel estimator with kernel

$$K(u) = \frac{1}{2} \exp(-|u|)$$

and bandwidth $h = \sqrt{\lambda}$. This approach is described in Vapnik (1982, p. 302).

1.4 Properties of kernels

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

$$\begin{aligned} K(u) &= K(-u) \\ \int K(u)du &= 1. \end{aligned} \tag{1.8}$$

In most applications K is a positive probability density function. For theoretical reasons it is sometimes useful to consider kernels that take on negative values. The order p of a kernel is defined as the first nonzero moment,

$$\begin{aligned} &\int K(u)u^j du = 0, \qquad j = 1, \ldots, p-1 \\ &\int K(u)u^p du \neq 0. \end{aligned} \tag{1.9}$$

A positive kernel can be at most of order 2. A higher order kernel (of order 4) is for example

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

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

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

Table 1.1. Common kernel functions.

1.5 Regression curve estimation

The most common method for studying the relationship between two variables x and y is to estimate the conditional expectation function $m(x) = E(y \mid x)$. Given i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$ we can then write

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

with an error term satisfying $E(\varepsilon \mid X) = 0$. Given the technique of kernel density estimation, a natural way to estimate $m(\bullet)$ is to compute first an estimate of the joint density $f(x, y)$ of (X, Y) and then to integrate it according to the formula

$$m(x) = \frac{\int y f(x, y) dy}{\int f(x, y) dy}. \tag{1.10}$$

The kernel density estimate $\widehat{f}_h(x, y)$ of $f(x, y)$ is constructed in complete analogy to the one-dimensional kernel estimate described earlier. One takes a product of two kernel functions and forms the two-dimensional density estimate

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

Note that by (1.8)

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

Plugging these into numerator and denominator of (1.10) we obtain the Nadaraya–Watson kernel estimate

$$\widehat{m}_h(x) = \frac{n^{-1}\sum_{i=1}^n K_h(x - X_i)Y_i}{n^{-1}\sum_{i=1}^n K_h(x - X_i)}. \tag{1.11}$$

Figure 1 shows a kernel regression estimate $\widehat{m}_h$ for the Marseille fish data. The bandwidth was chosen to $h = 0.5$. We shall comment later on how to choose this smoothing parameter. 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,

$$\widehat{m}_h(X_i) \to \frac{K(0)Y_i}{K(0)} = Y_i, \qquad \text{as} \quad h \to 0,$$

and at an arbitrary point x,

$$\widehat{m}_h(x) \to \frac{n^{-1}\sum_{i=1}^n K(0)Y_i}{n^{-1}\sum_{i=1}^n K(0)} = \overline{Y}, \qquad \text{as} \quad h \to \infty.$$

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 slow. Conditions for consistency of $\widehat{m}_h$ are given in the following theorem.

Theorem 1.

Let $\sigma^2(x) = var\ (Y \mid x)$ and $K(\bullet)$ satisfy $\int |K| \le \infty$. If $n \to \infty$, $h = h(n) \to 0$, $nh \to \infty$, then at every point of continuity of $m(x)$, $f(x)$, $\sigma^2(x)$,

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

The kernel estimate has asymptotic normal distribution.

Theorem 2.

Let m and f be twice differentiable, and $K(\bullet)$ satisfy $\int |K(u)|^{2+\eta} du < \infty$, for some $\eta > 0$. If $n \to \infty$, $h \sim n^{-1/5}$, then at every continuity point of $m(x)$, ($f(x)$, $\sigma^2(x)$, $E(|Y|^{2+\eta} \mid x)$, $\eta > 0$,

$$\sqrt{nh}\,(\widehat{m}_h(x) - m(x) - B(x)) \ \xrightarrow{\mathcal{L}}\ N(0, V(x))$$

where with $\mu_2(K) = \int u^2 K(u)du$, $\|K\|_2^2 = \int K^2(u)du$,

$$\begin{aligned} B(x) &= \frac{1}{2}\mu_2(K)\left[m''(x) + 2m'(x)\left(\frac{f'(x)}{f(x)}\right)\right], \\ V(x) &= \frac{\|K\|_2^2 \sigma^2(x)}{f(x)}. \end{aligned} \tag{1.12}$$

The bandwidth has been fixed here at a speed proportional to $n^{-1/5}$. The reason is that at this speed the squared bias and the variance of the kernel smoother have the same magnitude. In practice, it is desirable to have a bandwidth $h = cn^{-1/5}$ with a constant c possibly to be computed from the data. There is indeed an "optimal" data driven, estimated constant $\widehat{c}$ which is discussed later in Section 4. The optimal constant is the one balancing squared bias $B^2(x)$ and variance $V(x)$. From Theorem 2 we obtain for $h = cn^{-1/5}$ an approximate mean squared error (MSE) expansion,

$$\text{MSE}\left[\widehat{m}_h(x)\right] \approx n^{-1}h^{-1}V(x) + h^4 B(x). \tag{1.13}$$

The bandwidth minimizing this pointwise MSE is given by

$$h_0 = \left(\frac{V(x)}{4B(x)^2}\right)^{1/5}\left(\frac{\|K\|_2^2}{\mu_2^2(K)}\right)^{1/5} n^{-1/5}. \tag{1.14}$$

The constant c is therefore a function of the unknowns $V(x)$ and $B(x)$. The minimal MSE is a function of

$$T(K) = \|K\|_2^4 \ \mu_2(K). \tag{1.15}$$

This functional can be minimized with respect to K if a scale standardization of K is performed, for details see Gasser, Müller, and Mammitzsch (1985). A kernel is said to be optimal if it minimizes (1.15). The optimal kernel of order 2 is the Epanechnikov kernel given in Table 1.1. The third column of this table denotes the loss in efficiency of the other kernel with respect to this optimal one. One sees that 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.

2. *k*-Nearest Neighbor Estimates

2.1. Ordinary k-NN estimates

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 $\mathcal{J}_x = \{i : X_i \text{ is one of the } k\text{-NN to } x\}$ 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 $\mathcal{J}_x$,

$$\widehat{m}_k(x) = \frac{1}{k} \sum_{i \in \mathcal{J}_x} Y_i. \tag{2.1}$$

Connections to kernel smoothing can be made by considering (2.1) 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,

$$\widehat{m}_k(x) = \frac{(nR)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{R}\right) Y_i}{(nR)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{R}\right)}. \tag{2.2}$$

Note that in (2.2), for this specific kernel, the denominator is equal to $\frac{k}{nR}$ the k-NN density estimate of $f(x)$. Formula (2.2) can be generalized to arbitrary kernels. The bias and variance of this more general k-NN estimator is given in a theorem by Mack (1981).

Theorem 3.

If $n \to \infty$, $k/n \to 0$, then for (2.2), asymptotic expressions of the k-NN bias and variance are given by

$$\begin{aligned} B_k(x) &= \left(\frac{k}{n}\right)^2 \mu_2(K) \left[\frac{m''(x) + 2m'(x)\left(\frac{f'(x)}{f(x)}\right)}{8f^2(x)}\right], \\ V(x) &= 2\frac{\sigma^2(x)}{k} \|K\|_2^2. \end{aligned} \tag{2.3}$$

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 independent of distribution of the X-variables. The bias formula in (2.3) 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

$$k = 2nhf(x), \tag{2.4}$$

or equivalently

$$h = \frac{k}{(2nf(x))}.$$

For this choice of k or h respectively, the asymptotic mean squared error formulas are identical.

2.2. Symmetrized k-NN estimates

A computationally interesting 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 each neighborhood, we perform a local linear fit. 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 (2.4) remains true. In an example later, we shall use this estimator and compare it with the kernel estimator and the spline.

3. Spline Estimates

3.1 The cubic spline

For an estimate $\widehat{m}$ of m, the residual sum of squares (RSS) is defined as $\sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2$. If any curve $\widehat{m}$ is allowed as an estimator for m the RSS is minimized by an $\widehat{m}$ interpolating the data. Again this can be viewed as an ill-posed problem and so a regularization term is added in order to give reasonable estimates. The widely used cubic spline estimators are based on the stabilizer $\Omega(\widehat{m}) = \|\widehat{m}''\|_2^2$. Analog to the technique described earlier for density estimation, the spline estimator is defined as the (unique) minimizer $\widehat{m}_\lambda$ of

$$R_\lambda(\widehat{m}, m) = \sum_{i=1}^{n}(Y_i - \widehat{m}(X_i))^2 + \lambda \int (\widehat{m}''(u))^2 du. \tag{3.1}$$

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.

The smoothing parameter λ controls the degree of smoothness of the estimator $\widehat{m}_\lambda$. If λ tends to zero the stabilizer is given less weight and thus the spline is very rough and interpolates in the limit the observations. If λ tends to infinity increasing importance in (3.1) is given to the stabilizer $\|\widehat{m}''\|_2^2$ and hence $\widehat{m}_\lambda$ has almost no curvature and so in the limit is equal to a least squares regression line.

Splines are asymptotically equivalent to kernel smoothers as has been shown by Silverman (1984). The equivalent kernel is given by

$$K(u) = \frac{1}{2}\exp\left(-\frac{|u|}{\sqrt{2}}\right)\sin\left(\frac{|u|}{\sqrt{2}} + \frac{\pi}{4}\right) \tag{3.2}$$

and the equivalent bandwidth $h = h(\lambda; X_i)$ by

$$h(\lambda; X_i) = \lambda^{1/4} n^{-1/4}\ f(X_i)^{-1/4}. \tag{3.3}$$

The spline kernel is a function with negative sidelobes and thus cannot be a second-order kernel as defined in (1.9). In fact it is a fourth-order kernel since it is symmetric and has zero second moment, $\mu_2(K) = 0$.

3.2 Kernels, kNN, and splines

The similarity of spline and kernel smoothing becomes evident from the following figure where we apply kernel, k-NN and splines to the car data set (Table 7, p. 352–355 in Chambers, Cleveland, Kleiner and Tukey (1983)).

Figure 4. Scatterplot of car price (x) and miles per gallon (y) with three different smooth approximations (n=74, h=2000, k=11, λ=109).

In the upper left plot of this figure we see a scatterplot of x = price of car (in 1979) versus y = miles per gallon of that car. In total we have $n = 74$ observations. In the lower left we have plotted together with the raw data a kernel smoother $\widehat{m}_h$ with a bandwidth of $h = 2000$ and quartic kernel. Very similar to this is the spline smoother ($\lambda = 109$) although it is asymptotically equivalent to a kernel estimator with a kernel different from the quartic. The similarity comes also from the fact that the effective local bandwidth for the spline smoother from (3.3) is only a function of $f^{-1/4}$. In the scatterplot one sees that the marginal X's are nonuniform but not too far from uniform so the "local character" of spline smoothing does not really show up. 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.

In contrast to these two regression estimators stands the k-NN smoother ($k = 11$) in the upper right corner. We used the symmetrized k-NN estimator for this plot. By formula (2.4) 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 since here k is too small by (2.4).

4. Choice of Smoothing Parameter

4.1 Crossvalidation

Given a certain method of nonparametric regression estimation, the choice of how much to smooth has to be made in practice. In Sections 2 and 3 we have seen that k-NN and spline estimation are asymptotically equivalent to the kernel method, so we describe here only the selection of bandwidth h for kernel regression smoothing. A convenient measure of accuracy for $\widehat{m}_h$ is the averaged squared error

$$d_A(h) = n^{-1} \sum_{j=1}^{n} (\widehat{m}_h(X_j) - m(X_j))^2 w(X_j) \tag{4.1}$$

with a weight function w. This weight function allows to control and downweight boundary effects. For a discussion of the boundary effects, see Gasser and Müller (1979).

The minimization of (4.1) with respect to h can of course only be based on an estimate of $d_A(h)$. A naive estimate would be to just replace the unknown values $m(X_j)$ by the observations Y_j. This makes in a sense use of the same observation twice, indeed the response variable Y_j is used in $\widehat{m}_h(X_j)$ to provide itself. This must in practive lead to an undersmooth function, a curve with a too small bandwidth. In theoretical terms this can be expressed via asymptotic expressions for variance and squared bias: The naive re-substitution estimate (where we use Y_j twice) generates a term of the order of the variance with negative sign. Thus the variance term of d_A is wrongly underestimated and therefore creates a too low bandwidth.

The simplest way to avoid the problems of using Y_j twice is to use it only once. Instead of evaluating $\widehat{m}_h(X_j)$ with the j-th observation ones take this observation

out,

$$\widehat{m}_{h,j}(X_j) = \frac{n^{-1}\sum_{j\neq i} K_h(X_j - X_i)Y_i}{n^{-1}\sum_{j\neq i} K_h(X_j - X_i)}.$$

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

$$CV(h) = n^{-1}\sum_{j=1}^{n}(\widehat{m}_{h,j}(X_j) - Y_j)^2 w(X_j). \tag{4.2}$$

Choosing an h that minimized $CV(h)$ is asymptotically optimal in the following sense.

Asymptotic optimality

A bandwidth selection rule $\widehat{h}$ is asymptotically optimal if

$$\frac{d_A(\widehat{h})}{\inf_h d_A(h)} \xrightarrow{a.s.} 1.$$

The infinum here is taken over a set of h's that is specified in the following theorem. In practice it is advisable to perform minimization of $CV(h)$ over a log-scale range of h's since h is really a scaling parameter.

Theorem 4. *Suppose that*

(A1) for $n = 1, 2, \ldots$, $H_n = [\underline{h}, \overline{h}]$, *where*

$$\underline{h} \geq C^{-1}n^{\delta-1}, \quad \overline{h} \leq Cn^{-\delta}$$

for some constants $C, \delta \in (0, 1/2)$*;*

(A2) K *is Hölder continuous, that is, for some* $L > 0$, $\xi \in (0, 1)$

$$|K(u) - K(v)| \leq L\,|u - v|^{\xi},$$

and also

$$\int |u|^{\xi}\,|K(u)|du < \infty;$$

(A3) the regression function m *and the marginal density* f *are Hölder continuous;*

(A4) the conditional moments of Y given $X = x$ are bounded in the sense that there are positive constants $C_1, C_2, \ldots$ such that for $k = 1, 2, ...$, $E(Y^k \mid X = x) \leq C_k$ for all x;

(A5) the marginal density $f(x)$ of X is bounded from below on the support of w;

(A6) the marginal density $f(x)$ of X is compactly supported.
Then the bandwidth selection rule, "Choose $\widehat{h}$ to minimize $CV(h)$" is asymptotically optimal.

Proof: The Hölder continuity of K, m, f ensures that it suffices to consider a discrete subset H_n' of H_n. The existence of all conditional moments of order k gives over this sufficiently dense subset H_n' of H_n:

$$\sup_{h,h' \in H_n'} \left| \frac{d_A(h) - d_A(h') - (CV(h) - CV(h'))}{d_M(h) + d_M(h')} \right| \stackrel{a.s.}{\to} 0, \tag{4.3}$$

where $d_M(h) = \int MSE(\widehat{m}_h(x)) f(x) dx$ and $MSE(x)$ is defined in (1.13). A key step in proving (4.3) is Whittle's inequality (Whittle 1960) on bounding higher moments of quadratic forms of independent random variables. Using the Hölder continuity of K, m and f and Theorem 4.1.1 of Härdle (1990) gives

$$\sup_{h,h' \in H_n} \left| \frac{d_A(h) - d_A(h') - (CV(h) - CV(h'))}{d_A(h) + d_A(h')} \right| \stackrel{a.s.}{\to} 0. \tag{4.4}$$

Now let $\varepsilon > 0$ be given and let

$$\widehat{h}_0 = \arg\min_{h \in H_n} [d_A(h)],$$

$$\widehat{h} = \arg\min_{h \in H_n} [CV(h)].$$

From (4.4) we have with probability 1,

$$\frac{d_A(\widehat{h}) - d_A(\widehat{h}_0) - (CV(\widehat{h}) - CV(\widehat{h}_0))}{d_A(\widehat{h}) + d_A(\widehat{h}_0)} \leq \varepsilon.$$

This implies

$$0 \geq CV(\widehat{h}) - CV(\widehat{h}_0) \geq (1 - \varepsilon) d_A(\widehat{h}) - (1 + \varepsilon) d_A(\widehat{h}_0),$$

which entails

$$1 \leq \frac{d_A(\widehat{h})}{d_A(\widehat{h}_0)} \leq \frac{1+\varepsilon}{1-\varepsilon}.$$

Since ε was arbitrary, so

$$P\left\{\lim_{n\to\infty}\left|\frac{d_A(\widehat{h})}{d_A(\widehat{h}_0)}-1\right|<\delta\right\}=1 \quad \forall\delta>0,$$

which means that $\widehat{h}$ is asymptotically optimal.

4.2 Other data driven selectors

There are a number of different automatic bandwidth selectors that produce asymptotically optimal kernel smoothers. They are based on the idea to correct the downwards bias of the resubstitution estimate of $d_A(h)$. This method is most easily described in the setting of uniformly spaced X's. Suppose that $X_i = i/n$ and let us compute the downwards bias of the (nonoptimal) resubstitution estimate

$$p(h) = n^{-1}\sum_{j=:1}^{n}(\widehat{m}_h(X_i) - Y_i)^2 w(X_i).$$

The expected value is approximated as

$$\begin{aligned} E[p(h)] \approx & Ed_A(h) + \int \sigma^2(x)w(x)dx \\ & - 2n^{-1}h^{-1}K(0)\int \sigma^2(x)w(x)dx. \end{aligned} \tag{4.5}$$

The idea is now to correct for this third term in (4.5) which is the reason for the above mentioned downward bias. 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

$$G(h) = p(h)\ \Xi(n^{-1}h^{-1}K(0)),$$

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

$$\Xi(u) = 1 + 2u + O(u^2), \qquad u \to 0. \tag{4.6}$$

Simple examples are:

(i) *Generalized Cross-validation* (Craven and Whaba 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*,

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

Note that the correction function is different for the random design case where X has an arbitrary and unknown distribution. The correction works for the uniform design case since by (4.5), (4.6),

$$\begin{aligned} EG(h) &\approx \left[Ed_A(h) + \int \sigma^2(x)w(x)dx \right. \\ &\qquad \left. -2n^{-1}h^{-1}K(0)\int \sigma^2(x)w(x)dx \right] \Xi(n^{-1}h^{-1}K(0)) \qquad (4.7) \\ &\approx Ed_A(h) + \int \sigma^2(x)w(x)dx + O(n^{-2}h^{-2}). \end{aligned}$$

The constant term $\int \sigma^2(x)w(x)dx$ is independent of h, so minimizing $G(h)$ gives in the limit also asymptotically optimal smoothing parameters.

It is interesting to note that the above penalty method does not apply immediately to the case of random X's and heteroscedastic error distribution. The reason is in the fact that the downwards bias is not correctly cancelled out. In order to see this note that in the random design case the formula (4.7) changes to

$$\begin{aligned} EG(h) \approx &\left[Ed_A(h) + \int \sigma^2(x)f(x)w(x)dx \right. \\ &\left. -2n^{-1}h^{-1}K(0)\int \sigma^2(x)w(x)dx \right] \Xi(n^{-1}h^{-1}K(0)). \end{aligned} \qquad (4.8)$$

Hence the bias term does not cancel:A modification of the penalty method is necessary. If one uses the correction factor $\Xi(n^{-1}h^{-1}K(0)/\widehat{f}_h(X_j))$ inside the sum of

$p(h)$, this cancellation will still work. For this form of correction the Generalized Crossvalidation criterion is actually formally equivalent to ordinary crossvalidation. Note that this notation of generalized crossvalidation GCV stands in contrast to notation used in the spline literature. If $\widehat{m}_\lambda = A(\lambda)h$ denotes the spline smoothing operator, the spline GCV is defined as $\mathrm{GCV}(h) = \frac{RSS}{(1-n^{-1}\mathrm{tr}(A(\lambda)))^2}$. This cannot give optimal estimates though for random X and heteroscedastic errors!

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 5. The computation is for the quartic kernel using the WARPing method, see Härdle (1991). The minimal $\widehat{h} = \arg\min CV(h)$ is at 1800 which shows that in Figure 5 we used a slightly too large bandwidth.

Figure 5. The crossvalidation function $CV(h)$ for the car data. Quartic kernel. Computation made with XploRe (1991).

The question of how far the crossvalidation optimal $\widehat{h}$ is from the true optimum $\widehat{h}_0$ that minimized $d_A(h)$ has been investigated by Härdle, Hall and Marron (1988). One of the main results of this paper is that the random variables

$$n^{1/10}\left(\frac{\widehat{h}-\widehat{h}_0}{\widehat{h}_0}\right) \tag{4.9}$$

have an asymptotic Normal distribution with mean zero and variance independent of the actually used optimisation method. It does not matter whether one used Shibata's, Akaike's or any other optimizer, they are asymptotically equivalent. Another interesting result is that the estimated $\widehat{h}$ and optimum $\widehat{h}_0$ are actually negatively correlated ! It has been very recently that Hall and Johnstone (1992) corrected for this effect in density estimation and regression with uniform X's. It is still open how to improve this for the general regression setting we are considering here.

4.3 Canonical kernels

A comparison of smoothers for different kernels can only be made if the estimators are brought to the "same scale". Indeed a kernel can be rescaled as $K^*(\bullet) = s^{-1}K(\bullet s)$ which of course changes the value of the optimal bandwidth.

Note however that the kernel constants are

$$\|K^*\|_2^2 = s^{-1}\|K\|_2^2$$
$$\mu_2^2(K^*) = s^2\mu_2(K).$$

So we can uncouple the scaling effect by using for each kernel K that K^* with

$$s = s^* = \left(\frac{\|K^*\|_2^2}{\mu_2^2(K)} \right)^{1/5}.$$

Thus across kernels we shall have

$$\mu_2^2(K^*) = \|K^*\|_2^2 = \mu_2^{2/5}(K)\|K^*\|_2^{8/5}$$

So if we need to decide whether one curve (with kernel 1) is smoother than the other (kernel 2), we have to transform both bandwidths to the canonical scale $h_j^* = h_j / s_j^*$, $j = 1, 2$.

5. Application to Time Series

In the theoretical development described up to this point, one important assumption about the stochastic nature of the observations was the independence. The smoothing methods can also be applied to correlated data, in particular to nonparametric prediction of time series. We first consider the nonparametric prediction problem, then we turn to the analysis of regression curve estimation with correlated errors.

5.1 Prediction

We relax the assumption on the independence of the sequence of observations $(X_1, Y_1), (X_2, Y_2), \ldots$. We assume the process is α-mixing,

$$|P(A \cap B) - P(A)P(B)| \leq \alpha(k) \tag{5.1}$$

holds for all n, $k \in I\!N$ and any set A [resp. B] which is $\sigma((X_1, Y_i), \ldots, (X_n, Y_n))$ [resp. $\sigma((X_{n+k}, Y_{n+k}, \ldots]$ measurable, the sequence $\alpha(k)$ tending to zero for $k \to \infty$. If the process is stationary the best predictor (in a quadratic sense) for Y given $X = x$ is the conditional expectation

$$m(x) = E(Y \mid X = x).$$

Our aim is to estimate $m(\bullet)$ from data $\{(X_i, Y_i)\}_{i=1}^n$. This nonparametric estimation technique is also good for processes like $\{Z_i : i \geq 1\}$, and that one is interested in predicting Z_{n+s} from Z_n for some $s > 0$. The predictor is provided by the autoregression function

$$M(z) = E(Z_{n+s} \mid Z_n = z) \quad \forall\ n \geq 1. \tag{5.2}$$

The autoregression function M can then be interpreted as a regression curve of Y on X if we define $X_i = Z_i$, $Y_i = Z_{i+s}$, $\forall\ i \geq 1$. Clearly $\{(X_i, Y_i), i \geq 1\}$ is α-mixing when $\{Z_i, i \geq 1\}$ has this property.

For which concerns examples of processes satisfying this α-mixing condition we refer to Györfi et al. (1990), Chapters II.2 and III.4. For instance any Markov process satisfying Doeblin's condition is α-mixing with coefficients that verify (5.1) above. Also linear process of the form

$$Z_n = \sum_{i=0}^{\infty} \gamma_i \ T_{n-i},$$

where $(T_j)_{j \in I\!N}$ is a sequence of i.i.d. variables, can be shown to be α-mixing under appropriate summability conditions on (γ_i) (see Chanda (1974) and Garodetskii (1977). Härdle and Vieu (1991) showed that crossvalidation also works in this case, "choose" $\widehat{h} = \arg\min CV(h)$ gives asymptotically optimal estimates.

To give some insight into this process we simulated an autoregressive process $Z_i = M(Z_{i-1}) + \varepsilon_i$ with

$$M(x) = x \exp(-x^2),$$

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 (1.3) was used.

A plot of the generated time series (Z_0-uniform in $(-1/2, 1/2)$) is given in Figure 6 as a function of the time index. We are interested in finding the dependence structure between Z_{n-1} and Z_n.

Figure 6. The simulated time series with $M(x)$=$x \exp(-x^2)$, $\varepsilon \sim U(-1/2,1/2)$.

When we plot Z_{n-1} versus Z_n we obtain Figure 7. The (uniform) error structure become quite visible here but the shape of $M(x)$ may be guessed as linear from this figure. Only at the far ends we seem to see a curved structure of this point cloud. As an aid to interpret this picture we have added the true curve M and have plotted the two-dimensional time path. The starting and the end points are given as bullets.

Figure 7. The simulated series from Figure 6 plotted as Z_{n-1} versus Z_n. The solid curve is the true function M.

Since this is a simulated example we can also compute the distance $d_A(h)$. The cross-validation function $CV(h)$ and $d_A(h)$ are shown in Figure 8. The minimum of $CV(h)$ was $\widehat{h} = 1.5$, the optimum of $d_A(h)$ is at 2.1. The curve $d_A(h)$ is very flat for this example since we recall that there is almost no bias present.

Figure 8. The functions $d_A(h)$ (dashed line) and $CV(h)$ (solid line) for the simulated example.

The comparison of the estimated curve with the time regression function gives an impression of how well the smoothing method works. This comparison is displayed in Figure 9 where we find good coincidence with the time regression curve.

Figure 9. The time regression function $M(x)=x\exp(-x^2)$ for the simulated example (thick line) and the kernel smoother (thin line).

It might be reasonable to leave out more than just one observation, especially when the time series is strongly correlated. Such a leave-out estimator where we, in fact, sum over indexes $|i-j| > \rho_n$ for a slowly increasing sequence ρ_n is also covered by our theory. This "leave-out-more" technique is sometimes appealing also in the independent setting, see the discussion of Härdle, Hall and Marron (1988). The examples treated in Hart and Vieu (1991) in the setting of density estimation discuss also this point.

5.2 Correlated errors

Let us now consider the case of fixed design $X_i = i/n$ and correlated errors, i.e., $Y_i = m(X_i) + \varepsilon_i$, ε_i nonindependent. It is obvious that methods designed for i.i.d. errors must fail for this case. Imagine that the errors ε_i follow an autoregression of order 1,

$$\varepsilon_{i+1} = \rho\varepsilon_i + u_i, \qquad u_i \quad \text{white noise}$$

with ρ close to 1. 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 and hence the leave-one-out technique must give undersmooth estimates since the leave-out estimate of d_A "interprets" the little bumps of the error process as elements of the regression curve. An example is given in Härdle (1990, Figure 7.6, 7.7).

For a theoretical treatment of this problem let us assume that we have N

collections of the time series,

$$Y_{ij} = m(X_i) + \varepsilon_{ij}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, N.$$

An econometrical example for this observation scheme is a collection of time series of electricity demand which we observe repeatedly over days or weeks. Suppose now that the errors have the following correlation structure

$$\text{cov } (\varepsilon_{ij}, \varepsilon_{k\ell}) = \begin{cases} \sigma^2 \rho(X_i - X_k), & \text{if} \quad j = \ell \\ 0, & \text{if} \quad j \neq \ell. \end{cases}$$

There is independence of errors over repetitions of the series but correlation only within the series. For ρ assume that $\rho(0) = 1$, $\rho(u) = \rho(-u)$, $|\rho(u)| \leq 1$ for $u \in [-1, 1]$.

Hart and Wehrly (1986) computed the variance of kernel estimators for this model and showed that the bias is the same. In fact the variance changes from $\frac{\sigma^2}{N}\|K\|_2^2$ to

$$\left(\frac{\sigma^2}{N}\right) \int_{-1}^{1} \int_{-1}^{1} \rho(h(u-v))K(u)K(v)dudv.$$

Note that the kernel estimator is applied here to the averaged data $Y_{i\bullet} = N^{-1} \sum_{j=1}^{N} Y_{ij}$.

A Taylor expansion in terms of ρ gives yet another approximation to the variance,

$$\left(\frac{\sigma^2}{N}\right)(1 + h^2 \rho''(0)\mu_2(K)).$$

Since the bias stays the same as in the independent case, we obtain the following optimal bandwidth

$$h_{0,N} = \left\{ \frac{-2\sigma^2 \rho''(0)}{\mu_2^2(K)[m''(x)]^2} \right\} N^{-1/2}, \tag{5.3}$$

which minimizes the MSE as a function of N.

In practice one has of course to estimate the correlation function $\rho(k)$. Hart and Wehrly (1986) used the canonical estimate

$$\widehat{\rho}(k) = \frac{\widehat{c}(k)}{\widehat{c}(0)}$$

where

$$\widehat{c}(k) = (nN)^{-1} \sum_{j=1}^{N} \sum_{i=1}^{n} (Y_{ij} - Y_{i\bullet})(Y_{i+k,j} - Y_{i+k,\bullet}).$$

Estimates of second derivatives in formula (5.2) have to be constructed by differentiating a regression estimate $\widehat{m}_h$ with smooth enough kernel.

REFERENCES

Akaike, H. (1970). Statistical predictor information. *Annals of the Institute of Statistical Mathematics*, 22, 203–17.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions of Automatic Control AC*, 19, 716–23.

Chambers, J.M., Cleveland, W.S., Kleiner, B., and P.A. Tukey (1983). *Graphical Methods for Data Analysis*. Duxburry Press.

Chanda, K.C. (1974). Strong mixing properties of linear stochastic process. *Journal of Applied Probabilities*, 11, 401–408.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.*, 31, 377–403.

Doukhan, P. and Ghindès, M. (1980). Estimation dans le processus $X_n = f(X_{n-1}) + \varepsilon_n$. *Comptes Rendus, Académie des Sciences de Paris*, 297, Série A, 61–4.

Garodetskii, V.V. (1977). On the strong mixing condition for linear process. *Theory of Probability and its Applications*, 22, 411–413.

Gasser, T. and H. G. Müller (1984). Estimating regression functions and their derivatives by the kernel method.*Scandanavian Journal of Statistics*, 11, 171–85.

Gasser, T., Müller, H. G., and V. Mammitzsch (1985). Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society, Series B*, 47, 238–52.

Györfi, L., Härdle, W., Sarda, P., and P. Vieu (1990). Nonparametric Curve Estimation from Time Series. *Lecture Notes in Statistics*, 60. Heidelberg, New York: Springer-Verlag.

Härdle, W. (1990). *Smoothing Techniques with Implementation in S*. Heidelberg, New York, Berlin: Springer-Verlag.

Hall, P. and I. Johnstone (1992). Empirical functional and efficient smoothing parameter selection. *Journal of the Royal Statistical Society, Series B*, in print.

Härdle, W. (1991). *Applied Nonparametric Regression*. New York: Cambridge University Press.

Härdle, W. and Carroll, R. J. (1989). Biased cross-validation for a kernel regression estimator and its derivatives. *Östereichische Zeitschrift für Statistik und Informatik*.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *Journal of the American Statistical Association*, 83, 86–99.

Härdle, W. and D.W. Scott (1990). Smoothing by weighted averaging of rounded points. CORE Discussion Paper 9040, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Härdle, W. and P. Vieu (1991). Kernel regression smoothing of time series. CORE Discussion Paper 9031, Université Catholique de Louvain, Louvain-la-Neuve, Belgium. *Journal of Time Series Analysis*, in print.

Hart, J. and P. Vieu (1990). Data-driven bandwidth choice for density estimation based on dependent data. *Annals of Statistics*, 18, 873–890.

Hart, D. and T. E. Wehrly (1986).Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association*, 81, 1080–8.

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

Mack, Y. P. (1981). Local properties of k-NN regression estimates. *SIAM J. Alg. Disc. Meth.*, 2, 311–23.

Rice, J. A. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics*, 12, 1215–30.

Shibata, R.(1981). An optimal selection of regression variables. BIOK, 68, 45–54.

Silverman, B. W. (1984). Spline smoothing: the equivalant variable kernel method. *Annals of Statistics*, 12, 898–916.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Tikhonov, A.N. (1963). Regularization of incorrectly posed problems. *Soviet Math.*, 4, 1624–1627.

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

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Applications*, 5, 302. XploRe (1991).

XploRe 3.0 – a computing environment for eXploratory Regression. Available from CORE, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.