2
Nonparametric Smoothing

1.1 Introduction

The problem of estimating a smooth function is one which is encountered frequently in statistics. Very often one is provided with a random sample from an unknown continuous distribution where one's interest, among other things, is in the reconstruction of the original probability density function of the distribution generating the data. In other cases one may be interested in estimating the mean of the response variable as a function of the explanatory variable. There can be a large variety of problems, including the above two, where one of the primary interests of the experimenter is to reconstruct a smooth function based on discrete data.

Depending on the nature of the data and the analysis required, a bunch of parametric methods may be available for the analysis of the same. For example, in estimating the unknown density of the data generating distribution it is often convenient to assume that the unknown distribution has a known, simple, parametric form, and one only has to estimate the set of unknown parameters to completely specify the unknown distribution. For the regression problem also, a functional parametric form for the unknown conditional mean function is frequently assumed by the statisticians. For both of the above problems, the parametric method often leads to useful results, but at the same time it puts too much structure on the data. Sometimes there are natural scientific reasons for the choice of a particular parametric model; sometimes experience of previous experiments support the choice of one; however it is frequently the case where the mathematical simplicity of the specific parametric model is the primary reason for choosing it. One often wishes for more flexibility in the analysis, especially when there is serious doubt that the parametric model may not be completely successful in explaining all the true features of the data. Thus in many cases experimenters are more inclined to let the data speak for themselves and not constrain it to a particular parametric form. In all such situations appropriate nonparametric smoothing techniques are necessary and useful.

The material presented here has borrowed from many well known books

on density estimation, nonparametric regression, other related topics, as well as some major journal papers. As there is significant overlap between the important books in this area, we have presented all our sources in the list of references, but not necessarily referred to all of them in the text. Perhaps the book by Härdle (1990a) deserves special mention, as it concentrates on the implementation of the methods using the S-Plus software. Several additional important texts have also been added to the list of references as supplementary reading. We hasten to add, however, that the literature in this area is vast and growing by the day. We do not claim that the list of references is even close to being exhaustive, but we do hope that it is perhaps adequate to have a reasonably deep overall idea of the methods.

Although the title of the chapter is nonparametric smoothing, to keep a clear focus in our presentations we have concentrated primarily on the kernel method, with only brief comments about the other methods supplemented by appropriate references.

1.2 Density Estimation

The probability density function is one of the most fundamental characteristics of a random variable. Most of the major important features of the distribution of the random variable are expressed by the probability density function (hereafter referred to as the pdf). It tells us whether the distribution is symmetric or skewed, which intervals have greater chance of concentration of the data, whether the distribution is unimodal or has several modes, etc. The simplest, the most natural and the most popular method of estimating the unknown probability density of the distribution generating a random sample $X_1, X_2, \ldots, X_n$ of univariate data is the histogram, which we describe in the next subsection.

1.2.1 The Histogram for Univariate Data

A histogram is constructed by partitioning the range of the data into a certain number b of mutually exclusive intervals $B_1, B_2, \ldots, B_b$, such that the union of these intervals cover the entire range of the data. These intervals are called the bins of the histogram. The construction of the histogram then entails the classification of the data observations among the b bins, and counting how many observations fall into each bin. For a value x, the histogram function $\hat{f}(x)$ at that point is the relative frequency of the bin containing x, divided by the binwidth (the width of the bin). Histograms are step functions, and they have the nice property that they are the maximum likelihood density estimates of the unknown density when the class of density functions are restricted to step functions on a specified bin mesh (without such a restriction the estimation will not be well posed). The penalized maximum likelihood estimator (described later) is another approach which uses likelihood ideas to estimate an unknown density.

The vast majority of histograms used in practice use a common binwidth h (also called the smoothing parameter) for all the bins, although there are situations when variable binwidths do become absolutely necessary. For the situation where a common binwidth is used, the histogram method can be described as follows. Let x_0 be an appropriate origin, and suppose that the common binwidth is h. Then the j-th bin B_j of the histogram may be defined as the interval $[x_0 + (j-1)h, x_0 + jh)$, $j = 1, 2, \ldots$, where by convention we choose the intervals to be closed on the left and open on the right. Then the histogram function at a particular point x for a specific binwidth h can be formally defined as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j=1}^{b} I(X_i \in B_j) I(x \in B_j)$$

where $I(\cdot)$ represents the indicator function. By its construction the histogram automatically normalizes the frequency counts so that the total area under the histogram function is equal to 1.

Notice also that the construction of the histogram in this case requires the specification of two parameters: the origin x_0 and the binwidth h. The former controls the positioning of the bin edges, while the latter controls the size (width) of the bins. Choice of the origin can often have a major impact on the shape and interpretation of the histogram. Silverman (1986, pp. 9–11) gives two examples where different choices of origins appear to highlight different characteristics even when the binwidths are equal. Also see Wand and Jones (1995), Figures 1.3 (c and d).

However, it is the choice of the binwidth h which is generally the more important decision. It controls the degree of smoothness of the histogram – and consequently the trade-off between the bias and the variance. Variation of the binwidth can lead to really different shapes of the histogram. As the binwidth h increases one averages over a larger interval involving more data points and as a result the histogram has a smoother appearance. A very large value of h leads to an overly smooth character, and in the limit when h approaches infinity the histogram simply has the shape of a box. On the other hand, very small values of h lead to an overly jagged and spiky appearance, possibly with several gaps in the data. Figure 1.1 presents four sets of histograms for the Old Faithful Geyser data (see Azzalini and Bowman, 1990) corresponding to binwidths of $h = 0.2, 1, 2$ and 10 respectively. The oversmoothing effect due to increasing h is clear.

Thus one requires to choose an intermediate value of h to strike an appropriate balance between the above two scenarios. A detailed discussion of the factors and considerations leading to the choice of the histogram can be found, for example, in Härdle (1990a). Here we present the most important points.

One major consideration in choosing the binwidth is the trade-off between the bias and variance of the estimator $\hat{f}_h(x)$. To be specific, let us

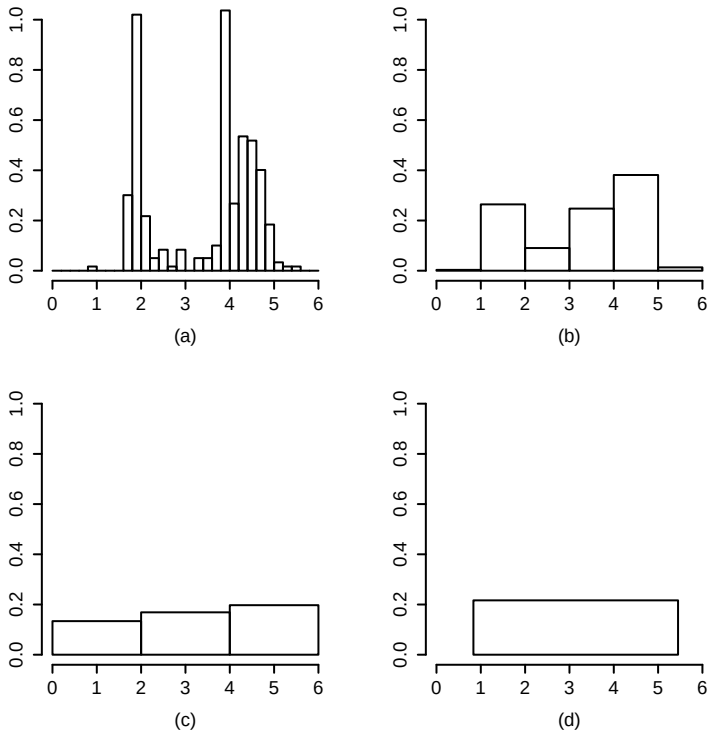

Figure 1.1: Histograms with binwidths of $h = 0.2, 1, 2$ and 10. The data set is from Azzalini and Bowman (1990), and represent durations in minutes of 299 eruptions of the Old Faithful geyser in the Yellowstone National Park, USA.

assume $x_0 = 0$, so that the j-th interval B_j is $[(j-1)h, jh)$. Then for $x \in B_j$, the bias in $\hat{f}_h(x)$ can be expressed as

$$\begin{aligned} \text{Bias}(\hat{f}_h(x)) &= ((j - \frac{1}{2})h - x)f'((j - \frac{1}{2})h) + o(h) \\ &= O(h) + o(h), \quad h \to 0. \end{aligned}$$

Thus the bias decreases (in the order of $O(h)$) as $h \to 0$. On the other hand, for any x belonging to B_j, the variance of $\hat{f}_h(x)$ is given by

$$\text{Var}(\hat{f}_h(x)) = (nh)^{-1}f(x) + o((nh)^{-1}), \quad h \to 0, \quad nh \to \infty,$$

leading to the mean squared error formula

$$\text{MSE}(\hat{f}_h(x)) = \frac{1}{nh}f(x) + ((j - \frac{1}{2})h - x)^2 f'((j - \frac{1}{2})h)^2 + o(h) + o((nh)^{-1}).$$

Notice that as $h \to 0$ and $nh \to \infty$, $\text{MSE}(\hat{f}_h(x)) \to 0$, and thus the histogram $\hat{f}_h(x)$ is a consistent estimator of $f(x)$. The conditions can be

interpreted as follows: h needs to go to zero as the sample size increases to keep the bias low, while $nh \to \infty$ is the condition necessary to ensure enough observations in the bins to keep the variance low.

Choosing a suitable value of h based on the mean squared error formula is not possible in practice, as its form involves the unknown density function f at the point x. One can alternatively define the mean integrated squared error (MISE)

$$\text{MISE}(\hat{f}_h) = E\left[\int_{-\infty}^{\infty} (\hat{f}_h - f)^2(x)dx\right]$$

as a measure of the global accuracy of $\hat{f}_h$ as an estimator for f. However, since the integrand is nonnegative, one can interchange the order of the integration and expectation in the above equation, and get

$$\text{MISE}(\hat{f}_h) = \int_{-\infty}^{\infty} E(\hat{f}_h - f)^2(x)dx = \int_{-\infty}^{\infty} \text{MSE}(\hat{f}_h(x))dx,$$

which can now be viewed as an average of the MSEs over all the values of x. A Taylor expansion of the integral up to the linear parts in the midpoints $(j - 1/2)h$ yields

$$\text{MISE}(\hat{f}_h) = (nh)^{-1} + \frac{h^2}{12}\|f'\|^2 + o(h^2) + o((nh)^{-1}), \tag{1.1}$$

where $\|\cdot\|$ represents the L_2 norm. By ignoring the asymptotically negligible terms, and by restricting ourselves to the leading terms we get the representation

$$\text{AMISE} = (nh)^{-1} + \frac{h^2}{12}\|f'\|^2 \tag{1.2}$$

of the asymptotic MISE under the assumption that f' is continuous and square integrable. One can then find the optimal bandwidth by minimizing AMISE over h, and we observe that the optimal bandwidth h_{opt} satisfies

$$h_{opt} = \left(\frac{6}{n\|f'\|^2}\right)^{1/3} \sim n^{-1/3}. \tag{1.3}$$

Applying this optimal binwidth to the MISE formula (1.1), one gets

$$\text{MISE} \sim n^{-2/3}.$$

There is a natural extension of the idea of the histogram which can be used to eliminate its dependence on the choice of the origin by constructing what is called the "naive" estimator. Apart from eliminating the issue of the choice of the origin, it also leads to a further generalization to the kernel density estimator in a natural way. The naive estimator does not count the frequency of the observations in the bins as in a regular histogram described above, but determines the density estimate at any particular point x by

suitably scaling the proportion of observations that lie in a certain interval $(x-h, x+h)$ around the given point. It can be easily seen that the above proportion, when scaled by $2h$ (the length of the interval), converges to $f(x)$ as $h \to 0$. Thus one can express the naive estimator as

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h}\right), \tag{1.4}$$

where the weight function w is given by

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{otherwise} \end{cases} \tag{1.5}$$

It is clear that the resultant function no longer requires a choice of the origin, although it still does depend on the choice of the parameter h (the half length of the intervals over which the proportions are calculated). Notice also that one can view the naive estimator as being constructed by putting rectangular (uniform) densities on $(y-h, y+h)$ around each data point $y = X_1, \ldots, X_n$, and then finding the estimate at any point x as being the sum of the contributions of the densities covering that point, divided by the sample size n.

Scott (1985) considers the average shifted histogram, another approach which reduces the dependence on the choice of the origin, but his approach can be shown to be approximating a kernel density estimator. Also see Scott (1979) for some additional comments on the properties of the histogram.

1.2.2 The Kernel Density Estimator for Univariate Data

Although the histogram and its modifications studied in the previous section give us some idea of the nature of the pdf of the distribution under study, they all lack the important property of smoothness in the sense that they are all different variations of step functions and are not differentiable at the jump points. To make up for this deficiency, we will look at a generalization of the histogram and its variants called the kernel density estimator. The basic ideas can be traced to Fix and Hodges (1951) and Akaike (1954); Rosenblatt (1956) and Parzen (1962) are among the early researchers who gave a formal shape to this line of research. It is a very useful but reasonably simple method which has led to its great popularity.

As remarked earlier, the kernel density estimator is a natural extension of the naive estimator considered in the previous section. Consider the weight function in equation (1.5) and the corresponding naive estimator in (1.4). We now replace the weight function with a smooth kernel function

$K(x)$ satisfying $\int K(x)dx = 1$. We will use the rescaling notation

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right),$$

which simplifies the presentation somewhat. The kernel density estimator corresponding to the kernel K and bandwidth (also called the window width or the smoothing parameter) h is defined to be

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i).$$

While one can interpret the naive estimator as one where we place boxes (rectangular densities) of equal width with the observed data points being at the center of the boxes, in kernel density estimation one overlays smooth kernel functions with the data points now representing the locations of these kernels. Generally (but not necessarily) K is chosen to be a symmetric probability density function, in which case one can think of the kernel density estimator as being generated by imposing the kernel around each data point with the latter being at the point of symmetry of the kernel function. The kernel density estimate at a point x is then obtained as the sum of the contributions at the point x of the kernels around each data point X_i, $i = 1, 2, \ldots, n$ divided by the sample size n. Thus it is the average of the n kernel ordinates at this point.

As an illustration, the construction of a kernel density estimate is demonstrated with five data points in Figure 1.2. The shape of the kernel density estimate using the Epanechnikov kernel (one of the most popular kernel functions which is defined later) is demonstrated for three different values of $h = 0.5, 1, 2$, and the actual kernel density estimate is the the composite function divided by n ($= 5$ in this case). The increase in the degree of smoothness due to an increase in the bandwidth is also clearly observed from the figure.

Notice that whenever the kernel function is a probability density function, so will be the kernel density estimate. In addition, the kernel density estimate inherits the smoothness properties of the kernel function. If the kernel function K is p times continuously differentiable, then so is the kernel density estimate $\hat{f}_h$. Like the naive estimator, the kernel density estimators do not require the choice of an origin.

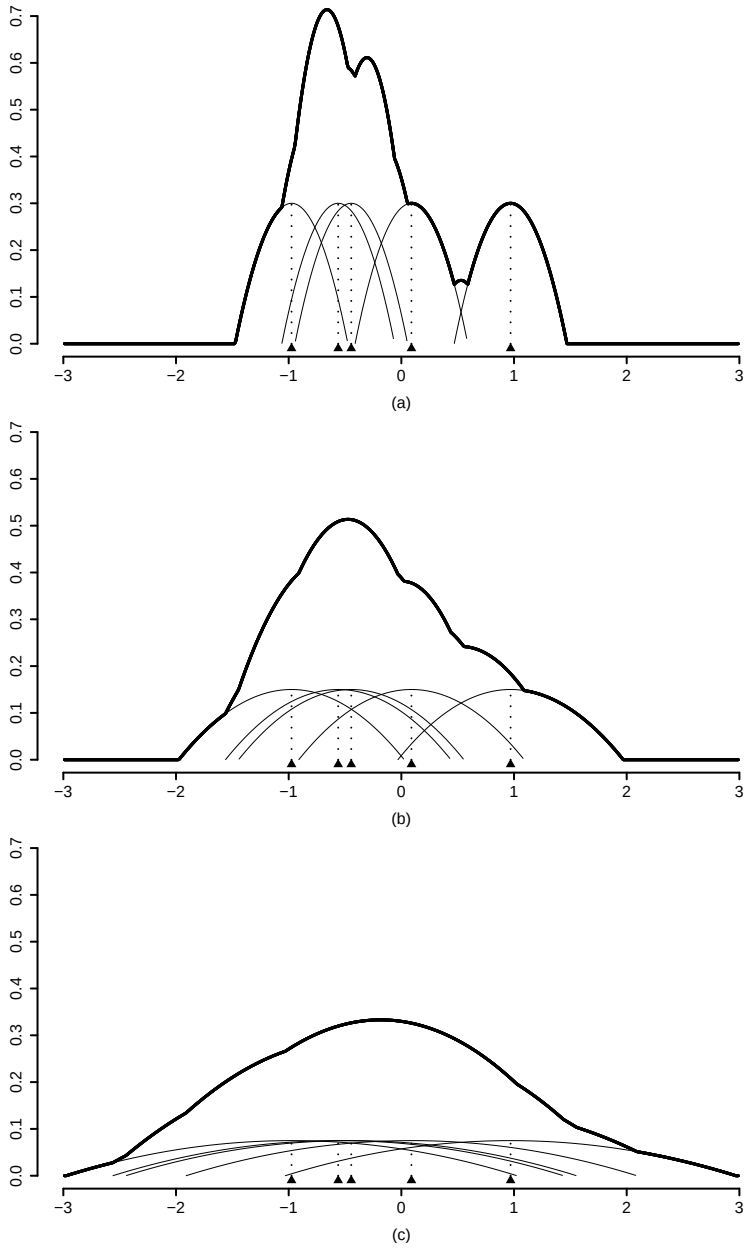

Figure 1.2: Kernel density estimates with the Epanechnikov kernel for five data points; values of the bandwidth used are 0.5, 1 and 2 respectively.

As in the case of the histogram, the MSE is still the most natural measure of the discrepancy between the density estimator $\hat{f}_h$ and the true

density f, while the MISE still remains as one of the most useful global measures for quantifying the overall discrepancy. The choice of the degree of smoothing is still governed by the trade-off between bias and variance. For a clear focus in our presentations we assume the following conditions on the kernel, which represent the most often used conditions for the application of the kernel method. We assume that the kernel K is a symmetric nonnegative probability density function. Thus

$$K \geq 0, \quad \int K(y)dy = 1, \quad \int yK(y)dy = 0. \tag{1.6}$$

We will also assume that $\int y^2 K(y)dy = \mu_2(K)$ is a finite non zero constant, which in this case is the variance of the distribution with density K.

The density estimate $\hat{f}_h(x) = \int h^{-1}K((x-y)/h)dF_n(y)$ at any point x has the form of a sample mean, as it is a sum involving independent and identically distributed random variables, divided by the sample size; here F_n represents the empirical distribution function. As a result $E\hat{f}_h(x)$ has the form of a smoothed model density $\int h^{-1}K((x-y)/h)dF(y)$, where F is the distribution function corresponding to the true unknown density f. In the particular case where we use a normal kernel with bandwidth h with f being the $N(\mu, \sigma^2)$ density, $E\hat{f}_h$ is simply the density of the $N(\mu, \sigma^2 + h^2)$ distribution, the convolution density of the above two normals. The bias

$$\text{Bias}(\hat{f}_h(x)) = \int h^{-1}K((x-y)/h)f(y)dy - f(x) = \int K(t)\{f(x-ht) - f(x)\}dt$$

has the approximate representation, using a Taylor series expansion, the assumptions on K, and ignoring the higher order terms (as will be appropriate when $h \to 0$),

$$\text{Bias}(\hat{f}_h(x)) \approx \frac{h^2}{2} f''(x)\mu_2(K). \tag{1.7}$$

(See Parzen, 1962). A similar Taylor series expansion leads to the approximate representation

$$\text{Var}(f_h(x)) \approx (nh)^{-1} f(x)\|K\|^2,$$

where $\|\cdot\|$ represents the L_2 norm, assuming h to be small, n to be large, and $nh \to \infty$. The optimal value of h can then be obtained by minimizing the asymptotic MISE formula

$$\frac{h^4}{4}(\mu_2(K))^2\|f''\|^2 + (nh)^{-1}\|K\|^2.$$

Here we have assumed that f'' is continuous and square integrable. This leads to an optimal bandwidth equal to

$$h_{opt} = (\mu_2(K))^{-2/5} \left(\frac{\|K\|^2}{\|f''\|^2}\right)^{1/5} n^{-1/5} \sim n^{-1/5}.$$

Notice that in this case the bandwidth converges to zero slower than the binwidth in the case of the histogram. Also, larger bandwidths suffice for smoother densities as can be observed from the presence of the $\|f''\|$ term. The quantity $\|f''\|$ is, in a sense, a measure of how rapidly the density fluctuates, and it seems proper that densities which fluctuate more should require smaller values for h for optimum kernel density estimation.

Another important consideration is the choice of the kernel itself. In this connection, note that the substitution of the optimal value of the bandwidth as determined above leads to the value of the MISE as being approximately equal to

$$\frac{5}{4}(\mu_2(K))^{2/5}(\|K\|^2)^{4/5}\|f''\|^{2/5}n^{-4/5}. \tag{1.8}$$

Thus the order of the bias, the variance, the optimal bandwidth and the optimum MISE can be expressed through the following table (see Härdle 1990a).

Estimator	Bias	Variance	Optimal h	Optimum MISE
Histogram	$\sim h$	$\sim (nh)^{-1}$	$\sim n^{-1/3}$	$\sim n^{-2/3}$
Kernel	$\sim h^2$	$\sim (nh)^{-1}$	$\sim n^{-1/5}$	$\sim n^{-4/5}$

Thus the kernel density estimator is asymptotically more efficient than the histogram in terms of the MISE criterion, since its MISE converges at a rate of $O(n^{-4/5})$, compared to the rate $O(n^{-2/3})$ of the histogram.

To continue the discussion regarding the choice of the kernel further, we notice from the MISE expression (1.8) that kernels which have small values for

$$C(K) = \mu_2(K)^{2/5}(\|K\|^2)^{4/5}$$

in equation (1.8) should be preferred. To further investigate this consideration, consider $\mu_2(K) = 1$. Wherever it is not, we can replace the kernel by its scaled version $\mu_2(K)^{-1/2}K((\mu_2(K))^{-1/2}t)$ (this does not affect the value of $C(K)$). In this case one only has to choose K so as to minimize $\left\{\int K(t)^2dt\right\}$. It has been shown by Hodges and Lehmann (1956) that this problem is solved by setting

$$K(t) = \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right), \quad -\sqrt{5} \le t \le \sqrt{5}.$$

It is called the Epanechnikov kernel (most often one uses the scaled version of the kernel where the support of the kernel is $(-1, 1)$). Denoting the $C(K)$ function for the Epanechnikov kernel by $C(K_{epn})$, one can compare the efficiency of any kernel with that of the Epanechnikov kernel by defining the efficiency of the kernel to be

$$\mathit{eff}(K) = \{C(K_{epn})/C(K)\}^{5/4}.$$

The form of some common kernels, and their efficiencies are given in the following table. Notice that even the rectangular kernel leading to the naive estimator returns an efficiency of about 0.93. Figure 1.3 exhibits the shape of some of the commonly used kernels.

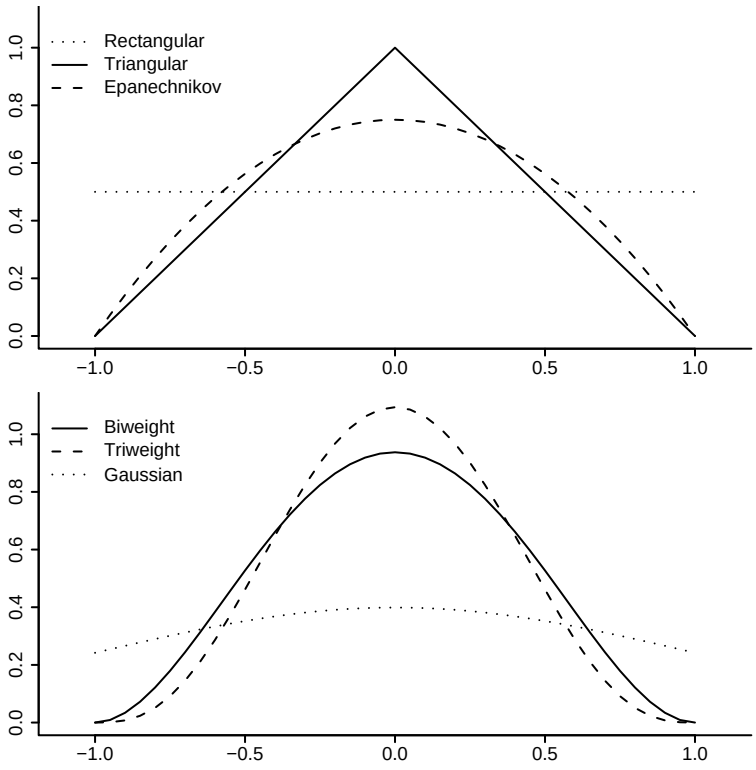

Figure 1.3: Shape of some commonly used kernels.

KERNEL	FORM	RANGE	EFFICIENCY
Rectangular	$\frac{1}{2}$	$-1 \le x \le 1$	0.9295
Triangular	$1 - \|x\|$	$-1 \le x \le 1$	0.9859
Epanechnikov	$\frac{3}{4}(1-x^2)$	$-1 \le x \le 1$	1.0000
Biweight	$\frac{15}{16}(1-x^2)^2$	$-1 \le x \le 1$	0.9939
Triweight	$\frac{35}{32}(1-x^2)^3$	$-1 \le x \le 1$	0.9867
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2)$	$-\infty < x < \infty$	0.9512

Thus the choice of the shape of the kernel is a relatively less important issue compared to the choice of the smoothing parameter.

1.2.3 Multivariate Kernel Density Estimation

Suppose now the support of the random variable is a subset of $\mathbb{R}^d$. One can easily imagine that as the dimension d of the of the data increases, the computational effort necessary to implement the kernel density estimation process adds up very fast, and one would rarely see the application of the kernel density estimation method for very high values of the dimension d.

In principle, however, the extension of the kernel density estimation method to multivariate data is reasonably straightforward. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ represent the n multivariate observations available from the unknown true density, where

$$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{id}).$$

In this case one has to smooth over d components. Let $\boldsymbol{x} = (x_1, \ldots, x_d)$. For the case where the same bandwidth h is chosen over all the dimensions, the natural extension of the univariate kernel density estimation method has the form, under obvious notations,

$$\hat{f}_h(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} h^{-d} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right).$$

Let $\boldsymbol{a} = (a_1, \ldots, a_d)$. Define $\boldsymbol{x}^{\boldsymbol{a}} = x_1^{a_1} \ldots x_d^{a_d}$. In the multivariate setting the symmetric kernels will be defined to be those for which $\int \boldsymbol{x}^{\boldsymbol{a}} K(\boldsymbol{x}) d\boldsymbol{x} = 0$ when $\boldsymbol{a}$ satisfies $|a| = (a_1 + \ldots a_d) = 1$. By a multivariate Taylor series theorem we get the following form for the bias:

$$\text{Bias}(\hat{f}_h(\boldsymbol{x})) = C_B h^2 + o(h^2) = O(h^2) + o(h^2), \quad h \to 0$$

for an appropriate constant C_B. Thus the bias is of the same order as in the univariate case, and is thus independent of the dimension.

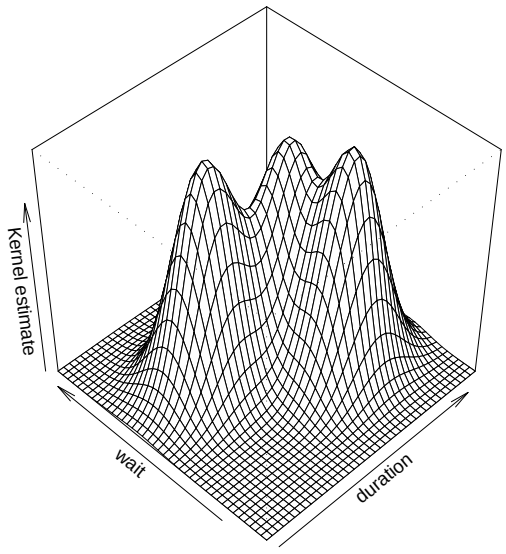

Figure 1.4: Two-dimensional kernel estimate of the Old Faith geyser data. Duration (in minutes) x_1-direction and wait (in minutes) x_2-direction with binwidths of $h_1 = 0.7$ and $h_2 = 7$.

That, however, is no longer the case for the variance, for which we get the expression

$$\mathrm{Var}(\hat{f}_h(x)) \approx C_V n^{-1} h^{-d}$$

for an appropriate constant C_V. Combining these we get the approximate MISE for the d-dimensional kernel density estimate as

$$\mathrm{MISE} = C_V n^{-1} h^{-d} + C_B^2 h^4.$$

This indicates that the optimal MISE bandwidth is $h_{opt} \sim n^{-\frac{1}{4+d}})$, with the corresponding optimal rate for the MISE itself being $n^{-\frac{4}{4+d}}$. Thus one gets poorer optimal rates in higher dimensions.

Sometimes the nature of the data will demand smoothing with different scales in different components. In that case, given $\boldsymbol{h} = (h_1, \ldots, h_d)$, the density estimate $\hat{f}_{\boldsymbol{h}}(\boldsymbol{x})$ can be constructed as

$$\hat{f}_{\boldsymbol{h}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 \times h_2 \times \ldots \times h_d} K\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_1} \right).$$

In particular the multiplicative kernel is given by

$$\hat{f}_{\boldsymbol{h}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} K_{j,h_j}(x_j - X_{ij}),$$

where $K_{j,h_j}(y) = h_j^{-1} K_j(y/h_j)$ represents a marginal one dimensional kernel for the j component.

Scott (1992) is an useful reference for multivariate kernel density estimation and visualization.

1.3 Nonparametric Regression

In regression analysis, we are generally provided with n pairs of data points $(X_1, Y_1), \ldots, (X_n, Y_n)$, and our interest, in broad terms, is in modeling the relationship between the explanatory variable X and the response variable Y. In particular, the regression function attempts to model the average value of Y conditional on the value of X being specified. Thus the regression function (or the mean response function) may be expressed as

$$Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.9}$$

with the function $m(X)$ representing the conditional mean of Y given X, and ϵ_i, $i = 1, \ldots, n$ are independently distributed errors with zero mean (but with possibly different scales). As in the case of density estimation, here one has the choice of using parametric and nonparametric methods. In nonparametric regression, which is our topic of discussion in this section, neither the regression function m, nor the error distribution are prespecified.

Fisher (1922) had viewed the regression problem as a composition of two problems: (i) the problem of determining the form of the unknown model (what he referred to as the problem of "specification"), and (ii) having decided upon the form of the model, the problem of determining the parameters which characterize the model (what he referred to as "estimation"). Fisher was primarily interested in the problem of estimation, and appeared to be inclined toward parametric models, being concerned about the relatively poor efficiency of the nonparametric approach. In this he was opposed by Pearson, who seems to have been more concerned about the specification problem. The interested reader can find more on this discussion in Tapia and Thompson (1978). Be that as it may, the nonparametric approach to regression has the advantage of flexibility eliminating the need to search for the proper parametric method which may be difficult to find.

As in the rest of this chapter, our emphasis will primarily be on the kernel smoothing approach to regression, and unless further qualified, regression smoothing will refer to the kernel smoothing version. Among others, Härdle (1990a, 1990b), Wand and Jones (1995), and Fan and Gijbels (1996) are useful references for the different methods of nonparametric regression. See also Altman (1992) for a brief general introduction to kernel and other types of nonparametric regression.

Intuitively, the basic idea of regression smoothing is the following. If the response function m is assumed to be smooth, one would expect that observations on the response variable where the predictor is close to the

target value x should contain information about m at x. This is a very important idea. In general the data will not contain repeated observations on specific values of the predictor variable, and unless information from nearby points is appropriately pooled, one would not be able to have a meaningful nonparametric estimate of the regression function. In effect, the local information idea implicitly assumes that the response curve is nearly constant (or at least representable by a polynomial of low degree) in a small interval around x, so that pooling these information in a proper way can lead to useful estimates.

Estimating the mean response curve m at x can therefore be viewed as a local averaging method for the responses on values lying in small intervals around the target value x. Formalizing this idea, one can express the estimated mean response curve as

$$\hat{m}(x) = \sum_{i=1}^{n} W_{hi}(x) Y_i, \tag{1.10}$$

where W_{hi} represents the weight attached to the response corresponding to the i-th observation on the predictor variable. There will be many considerations in choosing these weights; broadly, however, the values Y_i that correspond to X_i being close to x should get higher weights. This local averaging method can also be viewed as a form of a local weighted least squares method of estimation. This is explained in more detail later on.

The weight functions defined above control the contribution of each data point in this local averaging scheme; these weights are themselves controlled by the choice of the bandwidth (the smoothing parameter). These bandwidths determine the range of the interval around x over which to do the smoothing and subsequently the construction of the weights. As in the case of the histogram or the kernel density estimate, the value of the bandwidth controls the smoothness of the estimated mean response function $\hat{m}(x)$.

1.3.1 Kernel Regression

Nonparametric regression can be studied both under the fixed and random designs. In the univariate fixed design case the design consists of $x_1, \ldots, x_n$ which are ordered, non random numbers. In the random design model we observe a bivariate sample of random pairs. In this case the model can also be expressed as in (1.9), but now $m(x) = E(Y|X = x)$ and $v(x) = Var(Y|X = x)$ represent, respectively, the conditional mean and variance of Y given $X = x$. Except where noted otherwise, we will present the results based on the random design approach. Even though the stochastic mechanism is different, the basic idea of smoothing is the same for both random and fixed designs.

Suppose that we are given independent identically distributed random variables $(X_i, Y_i), i = 1, 2 \ldots, n$, where both X_i and Y_i are scalar random

variables, and we are interested in estimating the conditional response function $m(x) = E(Y|X = x)$. Functionally, we have

$$m(x) = E(Y|X = x) = \int yg(x, y)dy/f(x), \tag{1.11}$$

where $f(\cdot)$ is the density of X and $g(\cdot, \cdot)$ is the joint density of X and Y.

We will assume the same conditions on the kernel as in (1.6) throughout the rest of the section. Technical requirements may necessitate more restrictions on the kernel in some of the future developments (eg. some applications may require that K has a bounded first derivative; see Wand and Jones, 1995, Section 5.3).

Our discussion in section (1.2.2) has equipped us to estimate the denominator in equation (1.11), so we need to estimate the numerator only. Estimating the joint density in the numerator with the multiplicative kernel (section 1.2.3) we get

$$\hat{g}_{h_1,h_2}(x, y) = n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i)K_{h_2}(y - Y_i). \tag{1.12}$$

The entire numerator can then be estimated as

$$\begin{aligned}
\int y\hat{g}_{h_1,h_2}(x, y)dy &= n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i) \int yK_{h_2}(y - Y_i)dy \\
&= n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i) \int \frac{y}{h_2} K\left(\frac{y - Y_i}{h_2}\right) dy \\
&= n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i) \int (th_2 + Y_i)K(t)dt \\
&= n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i)Y_i
\end{aligned}$$

This leads to an estimated response function as

$$\hat{m}_{h_1}(x) = \frac{n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i)Y_i}{n^{-1} \sum_{i=1}^{n} K_{h_1}(x - X_i)}.$$

This is called the Nadaraya-Watson estimator, as it was proposed by Nadaraya (1964) and Watson (1964) at around the same time. In the context of the general weighted average form (equation 1.10) the weights are

$$W_{hi} = \frac{(nh)^{-1}K\left(\frac{x - X_i}{h}\right)}{\hat{f}_h(x)} = \frac{K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}. \tag{1.13}$$

Note that $\sum_{i=1}^{n} W_{hi}(x) = 1$ for all x. Also notice that the Y_i get more weight for regions where the corresponding X_i are sparse. As $h \to 0$, the weights W_{hi} converge to 1 if $x = X_i$, and zero otherwise. Thus the estimate corresponding to X_i converges to Y_i, and one essentially gets an interpolation of the data. On the other hand when $h \to \infty$, the weight function converges to $1/n$, so that the estimate $m(x)$ converges to the constant $\bar{Y}$. As in the kernel density estimate case, the smoothness of the curve increases with the bandwidth.

The Nadaraya-Watson estimator is in the form of a ratio estimator. For performing the appropriate analysis for determining the bias and variance, and for addressing the bandwidth selection problem, let us consider the numerator and the denominator separately. Let

$$r(x) = \int y g(x, y) dy = m(x) f(x),$$

which is estimated by

$$\hat{r}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i.$$

As in the case of the estimation of the bias of the kernel density estimator (equation (1.7)), we get

$$E[\hat{r}_h(x)] = r(x) + \frac{h^2}{2} r''(x) \mu_2(K) + o(h^2), \quad h \to 0.$$

and thus $\hat{r}_h(x)$ is asymptotically unbiased. Similar calculations with the variance of $\hat{r}_h(x)$ show

$$Var[\hat{r}_h(x)] = \frac{1}{nh} f(x) s^2(x) \|K\|^2 + o((nh)^{-1}),$$

where $s^2(x) = E[Y^2 | X = x]$. Thus $\hat{r}_h(x)$ is a consistent estimator of $r(x)$, and thus $\hat{m}_h(x)$ is a consistent estimator of $m(x)$, as $h \to 0$, and $nh \to \infty$.

It can be shown (eg. Härdle 1990a, page 134) that the leading term in the distribution of $\hat{m}_h(x) - m(x)$ is $f^{-1}(x)(\hat{r}_h(x) - m(x)\hat{f}_h(x))$ when one chooses $h \sim n^{-1/5}$, which is the bandwidth that balances the variance and the squared bias in this case. Calculations of the mean squared error based on the above representation lead to the approximation

$$\begin{aligned} \text{MSE}[\hat{m}_h(x)] &= \frac{1}{nh} \frac{v(x)}{f(x)} \|K\|^2 + \frac{h^4}{4} \left(m''(x) + 2 \frac{m'(x) f'(x)}{f(x)} \right)^2 \mu_2^2(K) \\ &+ o((nh)^{-1}) + o(h^4). \end{aligned}$$

The first term in the right hand of the above equality is the asymptotic variance of $\hat{m}_h(x)$, while the second term corresponds to the squared bias. Thus the MSE is of the order $O(n^{-4/5})$, when we choose $h \sim n^{-1/5}$.

Priestly and Chao (1972), and Gasser and Müller (1979, 1984) proposed some alternative kernel regression estimators. Notice that the presence of the random term in the denominator of the Nadaraya-Watson estimator is a source of inconvenience in the derivation of its asymptotic properties, and causes other practical difficulties as well. Gasser and Müller (1979, 1984) proposed the estimator

$$\hat{m}_h(x) = \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} K_h(t-x)dt \right] Y_i,$$

where $s_i = (X_i + X_{i+1})/2$, $X_0 = -\infty$, $X_{n+1} = +\infty$. Unlike the Nadaraya-Watson case, no normalizing denominator is necessary for the weights in this case. See Müller (1988) for more details on the Gasser-Müller method. Also see Mack and Müller (1989), and Chu and Marron (1991). Although it was originally intended for equispaced designs, the method works for non-equispaced designs as well.

The Gasser-Müller estimator can be viewed as a modification of the Priestly and Chao (1972) approach. In the fixed design model, corresponding to nonrandom and equispaced observations $x_1, \ldots, x_n$ on the interval $[0, 1]$, Priestly and Chao considered the weight function

$$n(x_i - x_{i-1})K_h(x - x_i). \tag{1.14}$$

Also see Benedetti (1977), and Cheng and Lin (1981). For definiteness we assume $x_0 = 0$. One can view $(x_i - x_{i-1})$ as an estimate of $n^{-1}f^{-1}$, which makes (1.14) correspond to (1.13).

1.3.2 Local Polynomial Kernel Regression Estimators

The Nadaraya-Watson estimator (as well as the Gasser-Müller estimator) can be extended to a larger class of estimators called the local polynomial kernel regression estimators. For a good account of kernel regression of this type Stone (1977), Müller (1987) and Fan (1992) are useful references, as is the book by Fan and Gijbels (1996). The book by Wand and Jones (1995) also provides an useful overview.

As a function estimation approach, both the Nadaraya-Watson and the Gasser-Müller estimators use local constant fits, implemented by a weighted least squares method. Thus, in approximating $m(\cdot)$ locally by a constant θ, either method obtains the estimate of θ as

$$\hat{\theta} = \text{argmin}_\theta \sum_{i=1}^{n} (Y_i - \theta)^2 w_i = \sum_{i=1}^{n} w_i Y_i / \sum_{i=1}^{n} w_i.$$

The Nadaraya-Watson estimator uses $w_i = K_h(X_i - x)$, which leads to

$$W_{hi} = \frac{K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}$$

in equation (1.13). The Gasser-Müller estimator, on the other hand, employs $w_i = \int_{s_{i-1}}^{s_i} K_h(t-x)dt$. The local polynomial kernel regression approach generalizes the above to the local fitting of a p-th degree polynomial using weighted least squares. Many of the results developed in the kernel density estimation context carry over to the regression context as well.

We assume the regression model in equation (1.9), with $Var(Y|X = x) = v(x)$. The local polynomial kernel regression estimator of order p at a particular point x is then estimated by fitting a p-th degree polynomial using weighted least squares where the weights are usually chosen according to the height of the kernel function centered about that point. Using the notation K_h for the kernel function scaled by a bandwidth h, the weight assigned to a particular point Y_i is $K_h(x-X_i)$. The estimated value $\hat{m}_{p,h}(x)$ of the conditional mean response function is the height of the fit $\hat{\beta}_0$ where $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ minimizes

$$\sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1(X_i - x) - \ldots - \beta_p(X_i - x)^p]^2 K_h(x - X_i). \qquad (1.15)$$

Let

$$\boldsymbol{W}_x = \text{diag}\{K_h(x - X_1), \ldots, K_h(x - X_n)\}$$

be the $n \times n$ diagonal weight matrix for the implementation of our weighted least squares. Letting $Y = (Y_1, \ldots, Y_n)^T$ denote the response vector and

$$\boldsymbol{X}_x = \begin{bmatrix} 1 & X_1 - x & \ldots & (X_1 - x)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_1 - x & \ldots & (X_1 - x)^p \end{bmatrix}$$

represent the $n \times (p+1)$ dimensional design matrix, we get the $(p+1) \times 1$ dimensional vector of solutions

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{X}_x)^{-1} \boldsymbol{X}_x^T \boldsymbol{W}_x \boldsymbol{Y}.$$

Our required estimate is then the first component of this vector.

For the case $p = 0$, a simple explicit expression can be written down for the estimated mean response function given by

$$m_{0,h}(x) = \sum_{i=1}^{n} K_h(x - X_i) Y_i / \sum_{i=1}^{n} K_h(x - X_i),$$

which in this case turns out to be the Nadaraya-Watson estimator. For the local linear estimator ($p = 1$), the corresponding expression is

$$\hat{m}_{1,h}(x) = n^{-1} \sum_{i=1}^{n} \frac{[\hat{s}_2(x;h) - \hat{s}_1(x;h)(X_i - x)] K_h(x - X_i) Y_i}{\hat{s}_2(x;h)\hat{s}_0(x;h) - \hat{s}_1(x;h)^2}$$

where

$$\hat{s}_r(x;h) = n^{-1}\sum_{i=1}^{n}(X_i - x)^r K_h(x - X_i).$$

Let $b_n = \dfrac{h^2}{2}\displaystyle\int_{-\infty}^{\infty} t^2 K(t)dt$, and $V_n = \dfrac{v(x)}{f(x)nh}\displaystyle\int_{-\infty}^{\infty} K^2(t)dt$. The basic asymptotic properties of the Nadaraya-Watson estimator, the Gasser-Müller estimator, and the local linear estimator, for a random design, are presented in the following table (taken from Fan, 1992), which illustrate the pointwise asymptotic bias and variance of the different kernel regression smoothers at an interior point of the support of the design density.

METHOD	BIAS	VARIANCE
Nadaraya-Watson	$\left(m''(x) + 2\dfrac{m'(x)f'(x)}{f(x)}\right)b_n$	V_n
Gasser-Müller	$m''(x)b_n$	$1.5V_n$
Local linear	$m''(x)b_n$	V_n

The form of the asymptotic bias of the Nadaraya-Watson estimator shows that it can have a large bias even when the true mean response curve $m(x)$ is linear (for example, when $f'(x)/f(x)$ is large). It has other deficiencies as well, such as zero minimax efficiency. The Gasser-Müller estimator corrects the bias of the Nadaraya-Watson estimator, but does so at the cost of increasing the variance. Both the Nadaraya Watson estimator and the Gasser-Müller estimator have a large order of the bias at the boundary. Corrections have been suggested but they are still less efficient than the automatic boundary correction of the local linear fit. The local linear fit is efficient in correcting boundary bias in an asymptotic minimax sense (Cheng, Fan and Marron, 1993).

We restrict the discussion of the comparison of the traditional estimators such as Nadaraya-Watson and Gasser-Müller estimators to the local linear fit method with the local polynomial modeling fold; the local linear method is a very simple methods which is able to take care of the difficulties of the classical methods without introducing a large amount of complicated machinery. Comparisons between the local linear fit and the local constant fit have been discussed in detail by Chu and Marron (1991), Fan (1992) and Hastie and Loader (1993) among others. Also see Fan and Gijbels (1996) for a general discussion of local polynomial regression, including those in higher dimensions.

Clevelend (1979) also considers local polynomial fitting based on nearest neighbor weights rather than kernel weights; his method goes by the name of LOESS.

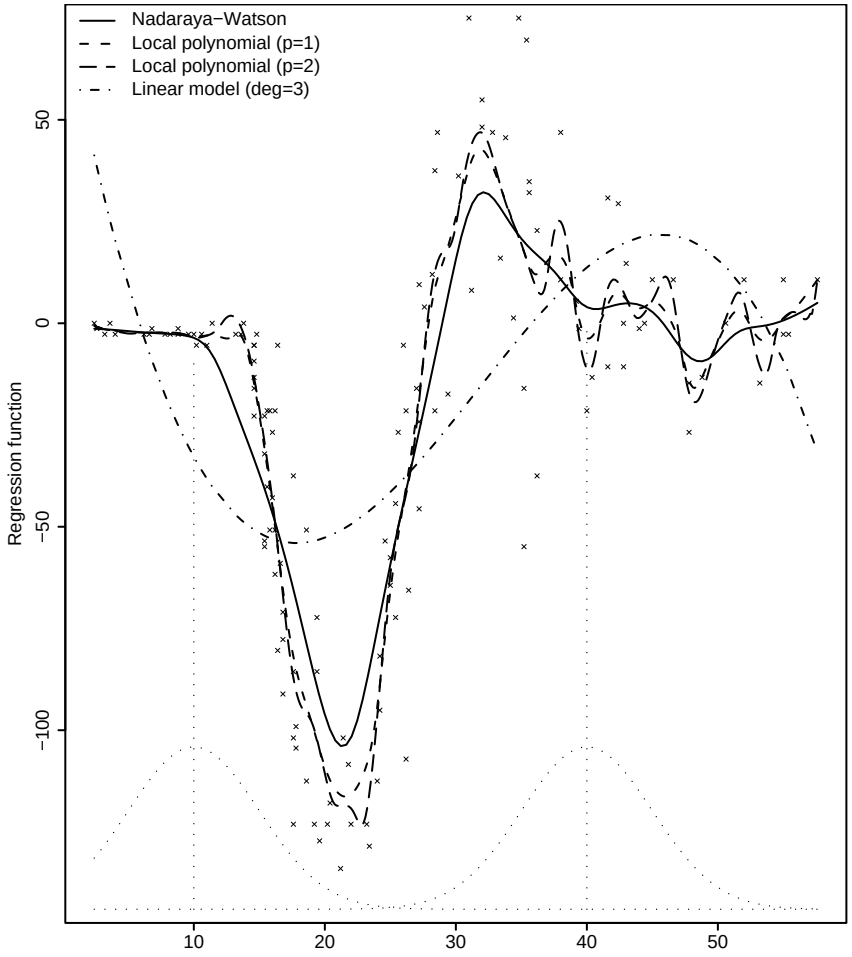

Figure 1.5: The motor-cycle impact data with regression estimators. The data set is from Silverman (1985).

1.3.3 Spline Smoothing

The quantity $\sum_{i=1}^{n}(Y_i - g(X_i))^2$ can serve as a measure of the closeness to the data for a curve g. However, the measure vanishes for any estimate where the estimated function $\hat{g}$ function which satisfies $\hat{g}(X_i) = Y_i$, $i = 1, 2, \ldots, n$, however poor be its match to g at the other points. Spline smoothing is a way to get around this problem. Here we use the square of the L_2 norm of the second derivative of g as a term which minimizes too much local variation, much in the spirit of penalized maximum likelihood estimation. Thus spline smoothing is defined as the minimization of $S_\lambda(g) = \sum_{i=1}^{n}(Y_i - g(X_i))^2 + \lambda\|g''\|^2$. For the class of all twice differ-

entiable functions on the interval $[X_{(1)}, X_{(n)}]$ results in the cubic spline, which consists of cubic polynomials between adjacent order statistics.

In the spline smoothing set up, the weighting parameter λ acts as the smoothing parameter in that decreasing λ leads to rougher estimates. As $\lambda \to 0$ for the parameter λ in $S_\lambda(g)$, one approaches a situation where one gets an interpolation of the observed values of Y, while in the limit $\lambda \to \infty$ one approaches a linear function in x.

See Schimek (2000) for detailed discussions on the latest developments of different aspects of nonparametric smoothing.

References

Akaike, H. (1954). An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, **6**, 127–132.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, **46**, 175–185.

Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, **39**, 357–365.

Benedetti, J. K. (1977). On the nonparametric estimation of regression functions. *Journal of Royal Statistical Society*, **B**, **39**, 248–253.

Cheng, K. F. and Lin, P. E. (1981). Nonparametric estimation of a regression function. *Zietschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **57**, 223–233.

Cheng, M.-Y., Fan, J. and Marron, J.S. (1993). Minimax efficiency of local polynomial fit estimators at boundaries. Mimeo Series #2098, Institute of Statistics, University of North Carolina, Chapel Hill.

Chu, C.-K. and Marron, J. S. (1991). Choosing a kernel regression estimator (with discussion). *Statistical Science*, **6**, 404–436.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer-Verlag, New York.

Devroye, L. and Gyorfi, L. (1985). *Nonparametric Density Estimation: The L_1 view*. John Wiley, New York.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, **87**, 998–1004.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and its Applications*. Chapman & Hall, London.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal statistical Society of London*, Series **A**, **222**, 309–368.

Fix, E. and Hodges, J. L. (1951). Discriminatory analysis - nonparametric discrimination: consistency properties. *Report No. 4, Project no. 21-29-004*, USAF School of Aviation medicine, Randolph Field, Texas.

Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for curve Estimation*, T. Gasser and M. Rosenblatt ed., Springer Verlag, Heidelberg, 23–68.

Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, **11**, 171–185.

Härdle, W. (1990a). *Smoothing Techniques with Implementations in S*. Springer Verlag, New York.

Härdle, W. (1990b). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, UK.

Hastie, T. J. and Loader, C. (1993). Local regression: automatic kernel carpentry (with discussion). *Statistical Science*, **8**, 120–143.

Hodges, J. L. and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics*, **27**, 324–335.

Mack, Y. P. and Müller, H.-G. (1989). Convolution type estimators for nonparametric regression. *Statistics and Probability Letters*, **7**, 229–239.

Müller, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *Journal of the American Statistical Association*, **82**, 231–238.

Müller, H.-G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Lecture Notes in Statistics, **46**, Springer-Verlag, Berlin.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its applications*, **10**, 186–190.

Parzen (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, **33**, 1065–1076.

Priestly, M. B. and Chao, M. T. (1972). Nonparametric function fitting. *Journal of the Royal Statistical Society*, **B 34**, 385–392.

Rosenblatt, M. (1956). Remarks on some non-parametric estimates of a density function. *Annals of Mathematical Statistics*, **27**, 832–837.

Schimek, M. G., Editor, (2000). *Smoothing and Regression: Approaches, Computation and Application*. John Wiley and Sons, New York.

Scott, D. W. (1979). On optimal and data based histograms. *Biometrika*, **66**, 605–610.

Scott, D. W. (1985). Average shifted histograms: effective nonparametric density estimators in several dimensions. *Annals of Statistics*, **13**, 1024–1040.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley, New York.

Schmidt, G. Mattern, R. and Schueler, F. (1981). Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under the effects of impact. EEC Research Program on Biomechanics of Impacts, Final report, Phase III, Project G5, Institut für Rechtsmedizin, University of Heidelberg, West Germany.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society B*, **47**, 1–52

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

Stone, C.J. (1977). Consistent nonparametric regression (with discussion). *Annals of Statistics*, **5**, 595–645.

Tapia, D. and Thompson, J. (1978). *Nonparametric Probability Density Estimation*. The Johns Hopkins University Press, Baltimore, MD.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, **A 26**, 359–372.