Nonparametric Econometrics

Ronaldo Dias
Departamento de Estatística.
Universidade Estadual de Campinas. São Paulo, Brasil
E-mail address: dias@ime.unicamp.br

Abstract

In recent years several economic data have been analyzed by nonparametric approaches. This paper is a review of a few of the most useful procedures in the nonparametric econometric field. In particular, it describes the theory and the applications of nonparametric curve estimation (density and regression) problems with emphasis in kernel, nearest neighbor, orthogonal series, smoothing splines, logsplines and H-splines methods.

1 Introduction

It is always useful to begin the study of regression analysis by making use of simple models. For this, assume that we have collected observations from a continuous variable Y at n values of a predict variable t. Let (t_j, y_j) such that:

$$y_j = g(t_j) + \varepsilon_j, \quad j = 1, \ldots, n, \tag{1.1}$$

where the random variables ε_j are uncorrelated with mean zero and variance σ^2. Moreover, $g(t_j)$ are the values obtained from some unknown function g computed at the points $t_1, \ldots, t_n$. In general, the function g is called *regression function* or *regression curve*.

A parametric regression model assumes that the form of g is known up to a finite number of parameters. That is, we can write a parametric regression model by,

$$y_j = g(t_j, \beta_1, \ldots, \beta_p) + \varepsilon_j, \quad j = 1, \ldots, n \tag{1.2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$. Thus, to determine from the data a curve g is equivalent to determine the vector of parameters $\boldsymbol{\beta}$. One may notice that, if g has a linear form, i.e., $g(t, \boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j x_j(t)$, where $\{x_j(t)\}_{j=1}^{p}$ are the explanatory variables, e.g., as in polynomial regression $x_j(t) = t^{j-1}$, then we are dealing with a situation of a linear parametric regression model.

Certainly, there are other methods of fitting curves to data. A collection of techniques known as nonparametric regression, for example, allows great flexibility in the possible form of the regression curve. In particular, assume no parametric form for g. In fact, a nonparametric regression model makes the assumption that the regression curve belongs to some infinite collection of curves. For example, g can be in the class of functions that are differentiable with a square integrable second derivatives, etc. Consequently, in order to propose a nonparametric model one may just need to choose an appropriate space of functions where he/she believes that the regression curve lies. This choice, usually, is motivated by the degree of the smoothness of g. Then, one uses the data to determine an element of this function space that can represent the unknown regression curve. Consequently, nonparametric techniques rely more heavily on the data for information about g than their parametric counterparts. Unfortunately, nonparametric estimators have some disadvantages. In general, they are less efficient than the parametric estimators when the parametric model is appropriate. For most parametric estimators the risk will decay to zero at a rate of n^{-1} while the nonparametric estimators have rate of $n^{-\alpha}$, where the parameter $\alpha \in (0, 1)$ depends on the smoothness of g. For example, when g is twice differentiable the rate is usually, $n^{-4/5}$. However, in the case where the parametric model is incorrect, ad hoc, the rate n^{-1} cannot be achieved. In fact, the parametric estimator does not even converge to the true regression curve.

2 Kernel estimation

Suppose we have n independent measurements $\{(t_i, y_i)\}_{i=1}^n$, the regression equation is, in general, described as in (1.1). Note that the regression curve g is the conditional expectation of the independent variable Y given the predict variable T, that is, $g(t) = \mathbb{E}[Y\,|T = t]$. When we try to approximate the mean response function g, we concentrate on the average dependence of Y on $T = t$. This means that we try to estimate the conditional mean curve

$$g(t) = \mathbb{E}[Y\,|T = t] = \int y \frac{f(t, y)}{f(t)} dy, \tag{2.1}$$

where $f(t, y)$ denotes the joint density of (T, Y) and $f(t)$ the marginal density of T. In order to provide an estimate $\hat{g}(t)$ of g we need to obtain estimates of $f(t, y)$ and $f(t)$. Consequently, a density estimation methodology will be described.

2.1 The Histogram

The histogram is one of the first, and one of the most common, methods of density estimation. It is important to bear in mind that the histogram is a smoothing technique used to estimate the unknown density and hence it deserves some consideration.

Let us try to combine the data by counting how many data points fall into a small interval of length h. This kind of interval is called a *bin*. Observe that the well known dot plot (Box, Hunter and Hunter1978, 25–26) is a particular type of histogram where $h = 0$.

Without loss of generality, we consider a *bin* centered at 0, namely the interval $[-h/2, h/2)$ and let F_X be the distribution function of X such that F_X is absolutely continuous with respect to a Lesbegue measure on $\mathbb{R}$. Consequently the probability that an observation of X will fall into the interval $[-h/2, h/2)$ is given by:

$$P(X \in [-h/2, h/2)) = \int_{-h/2}^{h/2} f_X(x) dx,$$

where f_X is the density of X.

A natural estimate of this probability is the relative frequency of the observations in this interval, that is, we count the number of observations falling into the interval and divide it by the total number of observations. In other words, given the data $X_1, \ldots, X_n$, we have:

$$P(X \in [-h/2, h/2)) \approx \frac{1}{n} \#\{X_i \in [-h/2, h/2)\}.$$

Now applying the mean value theorem for continuous bounded function we obtain,

$$P(X \in [-h/2, h/2)) = \int_{-h/2}^{h/2} f(x)dx = f(\xi)h,$$

with $\xi \in [-h/2, h/2)$. Thus, we arrive at the following density estimate:

$$\hat{f}_h(x) = \frac{1}{nh} \#\{X_i \in [-h/2, h/2)\},$$

for all $x \in [-h/2, h/2)$.

Formally, suppose we observe random variables $X_1, \ldots, X_n$ whose unknown common density is f. Let k be the number of bins, and define $C_j = [x_0 + (j-1)h, x_0 + jh)$, $j = 1, \ldots, k$. Now, take $n_j = \sum_{i=1}^{n} I(X_i \in C_j)$, where the function $I(x \in A)$ is defined to be :

$$I(x \in A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise,} \end{cases}$$

and, $\sum_{j=1}^{k} n_j = n$. Then,

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{j=1}^{k} n_j I(x \in C_j),$$

for all x. Here Note that the density estimate $\hat{f}_h$ depends strongly upon the *histogram bandwidth* h. By varying h we can have different shapes of $\hat{f}_h$. For example, if one increases h, one is averaging over more data and the histogram appears to be smoother. When $h \to 0$, the histogram becomes a very noisy representation of

the data (needle-plot, Härdle(1990)). The opposite, situation when $h \to \infty$, the histogram, now, becomes overly smooth (box-shaped, Härdle(1990)). Thus, h is the smoothing parameter of this type of density estimate, and the question of how to choose the histogram bandwidth h turns out to be an important question in representing the data via the histogram. For details on how to estimate h see Härdle (1990).

2.2 Kernel Density Estimation

The motivation behind the histogram can be expanded quite naturally. For this consider a weight function,

$$K(x) = \begin{cases} \frac{1}{2}, & \text{if } \mid x \mid < 1 \\ 0, & \text{otherwise} \end{cases}$$

and define the estimator,

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

We can see that $\hat{f}$ extends the idea of the histogram. Notice that this estimate just places a "box" of side (width) $2h$ and height $(2nh)^{-1}$ on each observation and then sums to obtain $\hat{f}$. See Silverman (1986) for a discussion of this kind of estimator. It is not difficult to verify that $\hat{f}$ is not a continuous function and has zero derivatives everywhere except on the jump points $X_i \pm h$. Besides having the undesirable character of nonsmoothness (Silverman1986), it could give a misleading impression to a untrained observer since its a somewhat ragged character might suggest several different bumps.

Figure 2.1 shows the nonsmooth character of the naive estimate. The data seem to have two major modes. However, the naive estimator suggests several different small bumps.

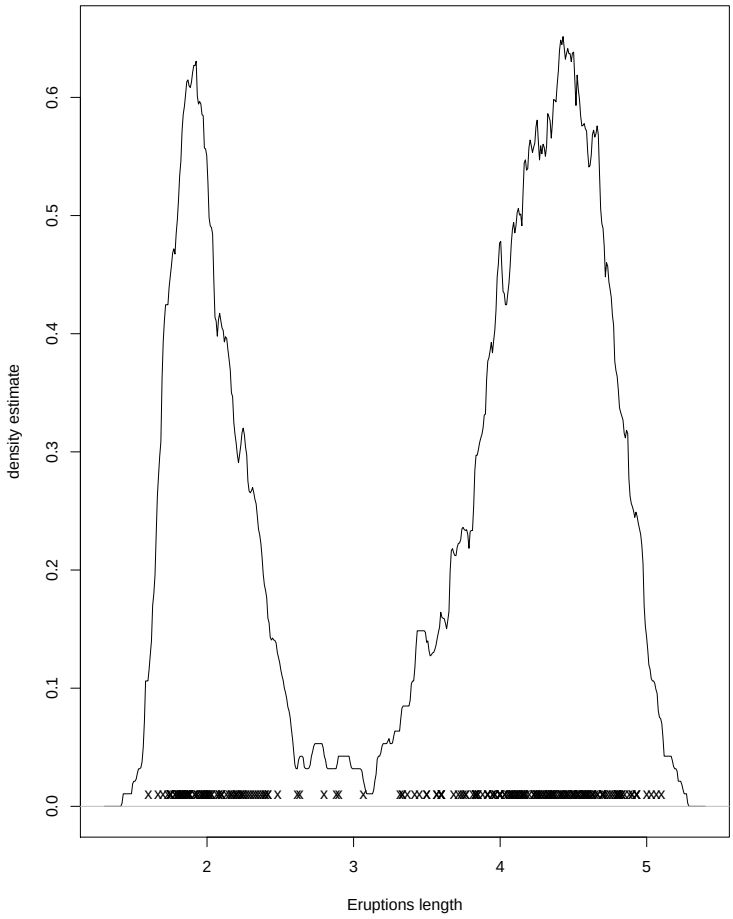

Figure 2.1: Naive estimate constructed from Old faithful geyser data with $h = 0.1$

To overcome some of these difficulties, conditions have been introduced on the function K. That is, K must be nonnegative kernel function that satisfies the following property:

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

In other words $K(x)$ is a symmetric probability density function, as for instance, the normal density, it will follow from definition that $\hat{f}$ will itself be a probability density. In addition, $\hat{f}$ will inherit all the continuity and differentiability properties of the kernel K. For example, if K is a normal density then $\hat{f}$ will be a smooth curve with derivatives of all orders.

Figure 2.2 exhibits the smooth properties of $\hat{f}$ when Gaussian kernel is used.

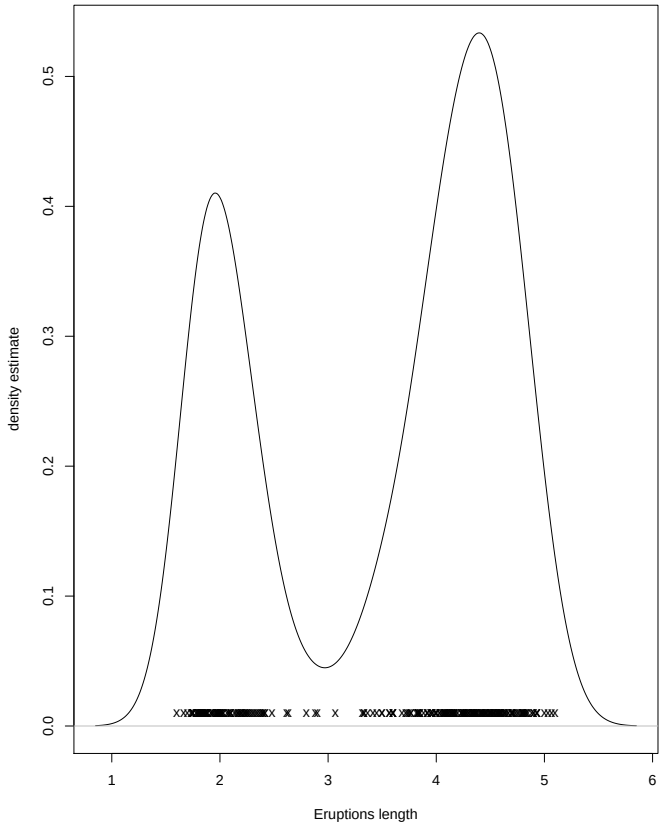

Figure 2.2: Kernel density estimate constructed from Old faithful geyser data with Gaussian kernel and $h = 0.25$

Note that an estimate based on the kernel function places "bumps" on the observations and the shape of those "bumps" is determined by the kernel function K. The bandwidth h sets the width around each observation and this bandwidth controls the degree of smoothness of a density estimate. It is possible to verify that as $h \to 0$, the estimate becomes a sum of Dirac delta functions at the observations while as $h \to \infty$, it eliminates all the local roughness and possibly important details are missed.

The data for the figure 2.3 which is labelled "income" were provided to me by Charles Kooperberg. This data set consists of 7125 random samples of yearly net income in the United Kingdom (Family Expenditure Survey, 1968-1983). The income

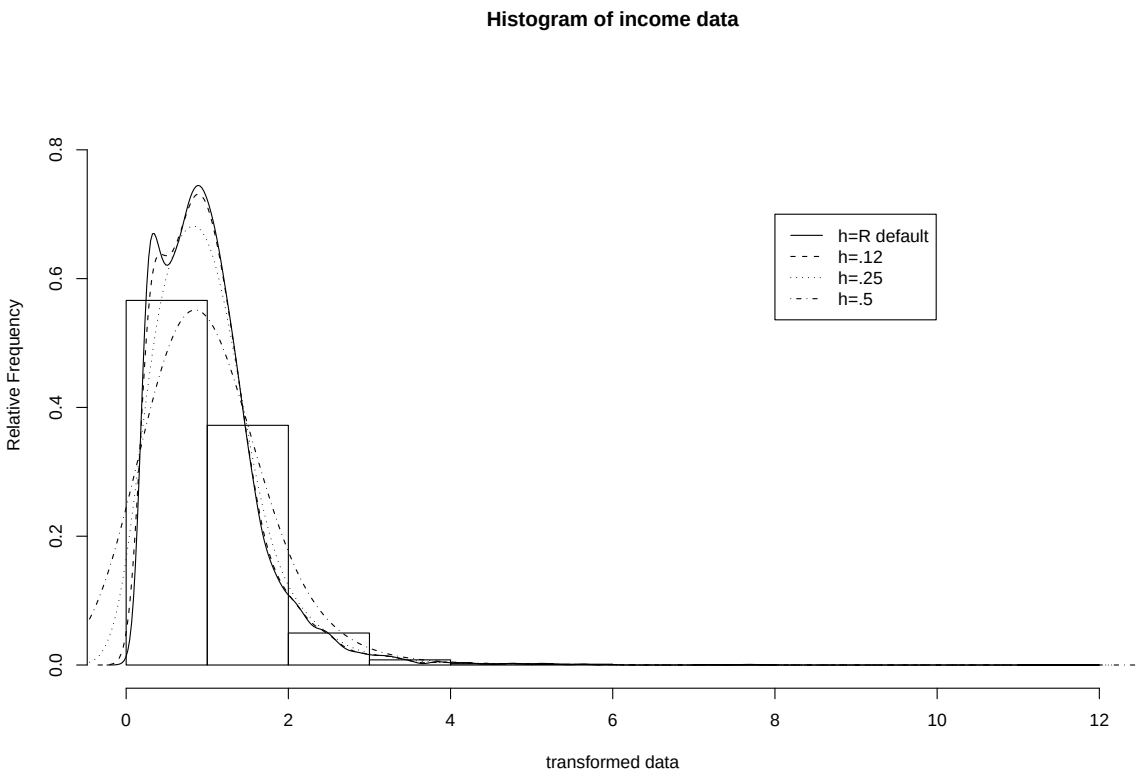

Figure 2.3: Bandwidth effect on kernel density estimates. The data set income was rescaled to have mean 1.

data is considerably large and so it is more of a challenge to computing resources and there are severe outliers. The peak at 0.24 is due to the UK old age pension, which caused many people to have nearly identical incomes. The width of the peak is about 0.02, compared to the range 11.5 of the data. The rise of the density to the left of the peak is very steep.

There is a vast (Silverman1986) literature on kernel density estimation studying its mathematical properties and proposing several algorithms to obtain an estimated based on it. This method of density estimation became, apart from histogram, the most commonly used estimator. However it has the drawbacks when the underlying density has long tails (Silverman1986). What causes this problem is the fact that the bandwidth is fixed for all observations, not considering any local characteristic of the data.

In order to solve this problem several other Kernel Density Estimation Methods were proposed such as the nearest neighbor and the variable kernel. A detailed discussion and illustration of these methods can be found in Silverman (1986).

2.2.1 The Nearest Neighbor Method

The idea behind of the nearest neighbor method is to adapt the amount of smoothing to local characteristics of the data. The degree of smoothing is then controlled by an integer k. Essentially, the nearest neighbor density estimator uses distances from x in $f(x)$ to the data point. For example, let $d(x_1, x)$ be the distance of data point x_1 from the point x, and for each x denote $d_k(x)$ as the distance from its kth nearest neighbor among the data points $x_1, \ldots, x_n$.

The kth nearest neighbor density estimate is defined as,

$$\hat{f}(x) = \frac{k}{2nd_k(x)},$$

where n is the sample size and, typically, k is chosen to be proportional to $n^{1/2}$.

In order to understand this definition, suppose that the density at x is $f(x)$. Then, one would expect about $2rnf(x)$ observations to fall in the interval $[x - r, x + r]$ for each $r > 0$. Since, by definition, exactly k observations fall in the interval $[x - d_k(x), x + d_k(x)]$, an estimate of the density at x may be obtained by putting

$$k = 2d_k(x)n\hat{f}(x).$$

Note that while estimators like histogram are based on the number of observations falling in a box of fixed width centered at the point of interest, the nearest neighbor estimate is inversely proportional to the size of the box needed to contain a given number of observations. In the tail of the distribution, the distance $d_k(x)$ will be larger than in the main part of the distribution, and so the problem of under-smoothing in the tails should be reduced. Like the histogram the nearest neighbor estimate is not a smooth curve. Moreover, the nearest neighbor estimate does not integrate to one and

the tails of $\hat{f}(x)$ die away at rate x^{-1}, in other words extremely slowly. Hence, this estimate is not appropriate if one is required to estimate the entire density. However, it is possible to generalize the nearest neighbor estimator in a manner related to the kernel estimate. The generalized kth nearest neighbor estimate is defined by,

$$\hat{f}(x) = \frac{1}{nd_k(x)} \sum_{i=1}^{n} K(\frac{x - X_i}{d_k(x)}).$$

Observe that the overall amount of smoothing is governed by the choice of k, but the bandwidth used at any particular point depends on the density of observations near that point. Again, we face the problems of discontinuity of at all the points where the function $d_k(x)$ has discontinuous derivative. The precise integrability and tail properties will depend on the exact form of the kernel.

Figure 2.4 shows the effect of the smoothing parameter k on the density estimate. Observe that as k increases rougher the density estimate becomes. This effect is equivalent when h is approaching to zero in the kernel density estimator.

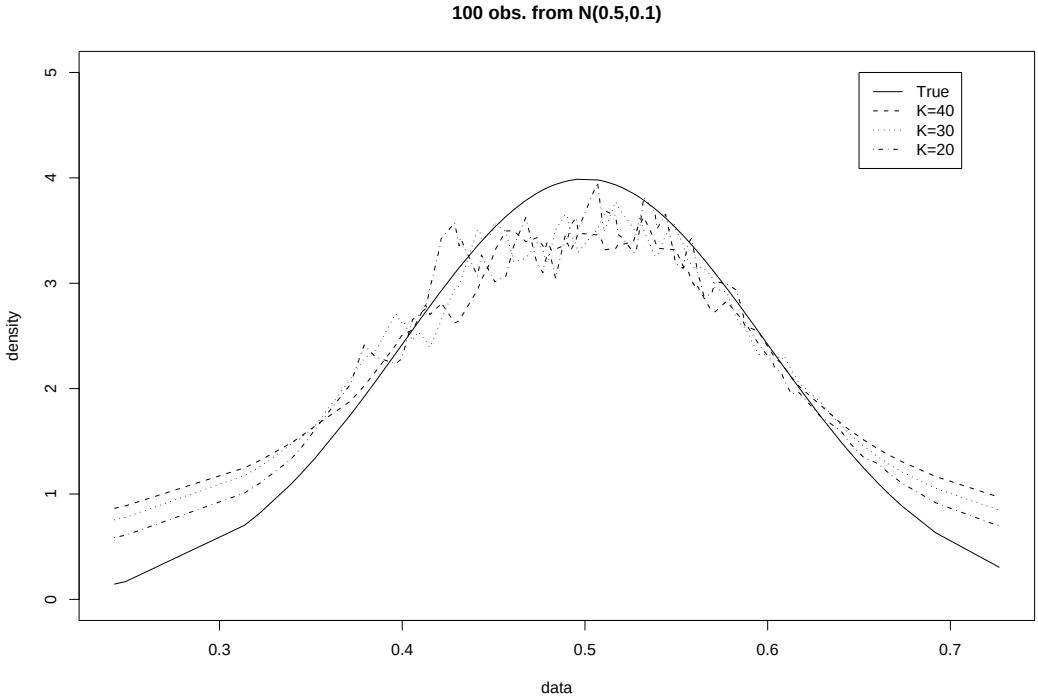

Figure 2.4: Effect of the smoothing parameter K on the estimates

2.2.2 Some Statistical Results of Kernel Density Estimation

As starting point one might want to compute the expected value of $\hat{f}$. For this, suppose we have $X_i, \ldots, X_n$ i.i.d. random variables with common density f and let $K(\cdot)$ be a probability density function defined on the real line that satisfies the following conditions (Prakasa-Rao1983):

- **Condition 1.** $\sup_x K(x) \leq M < \infty$; $|x|K(x) \to 0$ as $|x| \to \infty$.
- **Condition 2.** $K(x) = K(-x),\ x \in (-\infty, \infty)$ with $\int_{-\infty}^{\infty} x^2 K(x)dx < \infty$.

Then we have, for a nonstochastic h

$$\begin{aligned} E[\hat{f}(x)] &= \frac{1}{nh}\sum_{i=1}^{n} E[K(\frac{x - X_i}{h})] \\ &= \frac{1}{h}E[K(\frac{x - X_i}{h})] \\ &= \frac{1}{h}\int K(\frac{x-u}{h})f(u)du \\ &= \int K(y)f(x + yh)dy. \end{aligned} \tag{2.2}$$

Now, let $h \to 0$. We see that $E[\hat{f}(x)] \to f(x)\int K(y)dy = f(x)$. Thus, $\hat{f}$ is an asymptotic unbiased estimator of f.

To compute the bias of this estimator we have to make the assumption that the underlying density is twice differentiable. Using a Taylor expansion of $f(x + yh)$, the bias of $\hat{f}$ in estimating f is

$$b_f[\hat{f}(x)] = \frac{h^2}{2}f''(x)\int y^2 K(y)dy + o(h^2).$$

We observe that since we have assumed the kernel K is symmetric around zero, we have that $\int yK(y)hf'(x)dy = 0$, and the bias is quadratic in h. [1]

[1] See (Parzen1962)

Using a similar approach we obtain :

- $Var_f[\hat{f}(x)] = \frac{1}{nh}\|K\|_2^2 f(x) + o(\frac{1}{nh})$, where $\|K\|_2^2 = \int \|K(x)\|^2 dx$
- $MSE_f[\hat{f}(x)] = \frac{1}{nh} f(x)\|K\|_2^2 + \frac{h^4}{4}(f''(x)\int y^2 K(y)dy)^2 + o(\frac{1}{nh}) + o(h^4)$,

where $MSE_f[\hat{f}]$ stands for mean squared error of the estimator $\hat{f}$ of f.

Hence, when the conditions $h \to 0$ and $nh \to \infty$ are usually assumed, the $MSE_f[\hat{f}] \to 0$, which means that the kernel density estimate is a consistent estimator of the underlying density f. Moreover, MSE balances variance and squared bias of the estimator in such way that the variance term controls the *under-smoothing* and the bias term controls *over-smoothing*. In other words, an attempt to reduce the bias increases the variance, making the estimate too noisy (under-smooth). On the contrary, minimizing the variance leads to a very smooth estimate (over-smooth) with high bias.

2.2.3 Bandwidth Selection

It is natural to think of finding an optimal bandwidth, say, h_* such that $h_* = \arg\min_h MSE_f[\hat{f}]$. Härdle(1990) shows that

$$h_* = \Big(\frac{f(x)\|K\|_2^2}{(f''(x))^2(\int y^2 K(y)dy)^2 n}\Big)^{1/5} \propto n^{-1/5}. \tag{2.3}$$

The problem with this approach is that h_* depends on two unknown functions $f(\cdot)$ and $f''(\cdot)$. An approach to overcome this problem uses a global measure that can be defined as:

$$\begin{aligned} IMSE[\hat{f}] &= \int MSE_f[\hat{f}(x)]dx \\ &= \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}(\int y^2 K(y)dy)^2 \|f''\|_2^2 + o(\frac{1}{nh}) + o(h^4). \end{aligned} \tag{2.4}$$

IMSE is the well known *integrated mean squared error* of a density estimate.

The optimal value of h considering the IMSE is define as

$$h_{opt} = \arg\min_{h>0} IMSE[\hat{f}].$$

it can be shown that,

$$h_{opt} = c_2^{-2/5} \Big(\int K^2(x)dx \Big)^{1/5} \Big(\|f''\|_2^2 \Big)^{-1/5} n^{-1/5}, \tag{2.5}$$

where $c_2 = \int y^2 K(y)dy$. Unfortunately, (2.5) still depends on the second derivative of f, which measures the speed of fluctuations in the density of f.

2.2.3.1 Reference to a Standard Distribution

A very natural way to get around the problem of not knowing f'' is to use a standard family of distributions to assign a value of the term $\|f''\|_2^2$ in the expression (2.5). For example, assume that a density f belongs to a class normal family with mean μ and variance σ^2, then

$$\begin{aligned} \int (f''(x))^2 dx &= \sigma^{-5} \int (\varphi''(x))^2 dx \\ &= \frac{3}{8}\pi^{-1} 2\sigma^{-5} \approx 0.212\sigma^{-5}, \end{aligned} \tag{2.6}$$

where $\varphi(x)$ is the standard normal density. If one uses a Gaussian kernel, then

$$\begin{aligned} h_{opt} &= (4\pi)^{-1/10} (\frac{3}{8}\pi^{-1/2})^{-1/5} \sigma n^{-1/5} \\ &= \Big(\frac{4}{3}\Big)^{1/5} \sigma n^{-1/5} = 1.06\sigma n^{-1/5} \end{aligned} \tag{2.7}$$

Hence, in practice a possible choice for h_{opt} is $1.06\hat{\sigma}n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation.

If we want to make this estimate more insensitive to outliers, we have to use a more robust estimate for the scale parameter of the distribution. Let $\hat{R}$ be the sample interquartile, then one possible choice for h is

$$\begin{aligned} \hat{h}_{opt} &= 1.06 \min(\hat{\sigma}, \frac{\hat{R}}{(\Phi(3/4) - \Phi(1/4))}) \\ &= 1.06 \min(\hat{\sigma}, \frac{\hat{R}}{1.349}), \end{aligned} \tag{2.8}$$

where Φ is the standard normal distribution function.

Figure 2.5 exhibits how a robust estimate of the scale can help in choosing the bandwidth. Note that by using $\hat{R}$ we have strong evidence that the underlying density has two modes.

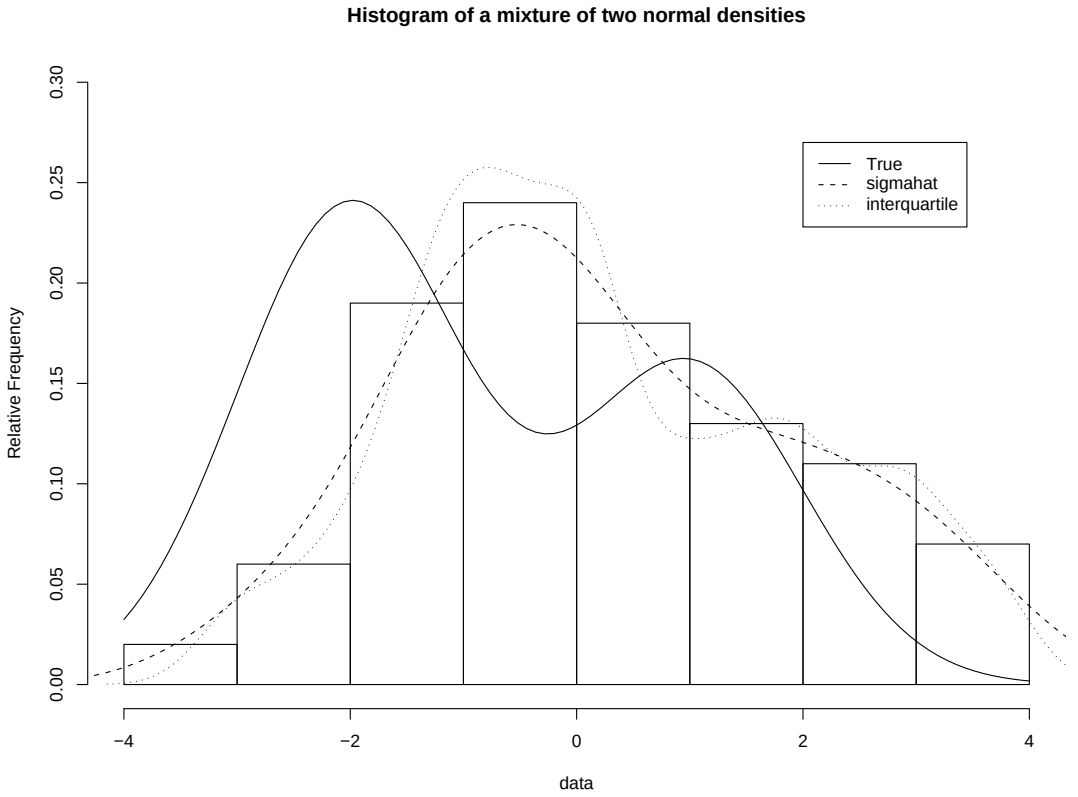

Figure 2.5: Comparison of two bandwidths, $\hat{\sigma}$ (the sample standard deviation) and $\hat{R}$ (the sample interquartile) for the mixture $0.7 \times N(-2, 1) + 0.3 \times N(1, 1)$.

2.2.3.2 Maximum likelihood Cross-Validation

Consider kernel density estimates f_h and suppose we want to test for a specific h the hypothesis

$$f_h(x) = f(x) \quad vs. \quad f_h(x) \neq f(x),$$

for a fixed x The likelihood ratio test would be based on the test statistic $f(x)/f_h(x)$. For a good bandwidth this statistic should thus be close to 1. We would also say that on the average $\mathbb{E}[\log(f(X)/f_h)(X)]$ should be close to 0. Thus, a good bandwidth,

which is minimizing this measure of accuracy, is in effect optimizing the *Kullback-Leibler* distance:

$$d_{KL}(f, f_h) = \int \log\Big(\frac{f(x)}{f_h(x)}\Big) f(x)dx. \tag{2.9}$$

Of course, we are not able to compute $d_{KL}(f, f_h)$ from the data, since we do not know f. But from a theoretical point of view, we can investigate this distance for the choice of an appropriate bandwidth h. When $d_{KL}(f, f_h)$ is close to 0 this would give the best agreement with the hypothesis $f_h = f$. Hence, we are looking for a bandwidth h, which minimizes $d_{KL}(f, f_h)$.

Suppose we are given a set of additional observations X_i, independent of the others. The likelihood for these observations is $\prod_i f(X_i)$. Substituting f_h in the likelihood equation we have $\prod_i f_h(X_i)$ and the value of this statistic for different h would indicate which value of h is preferable, since the logarithm of this statistic is close to $d_{KL}(f, f_h)$. Usually, we do not have additional observations. A way out of this dilemma is to base the estimate f_h on the subset $\{X_j\}_{j\neq i}$, and to calculate the likelihood for X_i. Denoting the *leave-one-out estimate*

$$f_h(X_i) = (n-1)^{-1}h^{-1}\sum_{j\neq i} K(\frac{X_i - X_j}{h}).$$

Hence,

$$\prod_{i=1}^{n} f_{h,i}(X_i) = (n-1)^{-n}h^{-n}\prod_{i=1}^{n}\sum_{j\neq i} K(\frac{X_i - X_j}{h}). \tag{2.10}$$

However it is convenient to consider the logarithm of this statistic normalized with the factor n^{-1} to get the following procedure:

$$\begin{aligned} CV_{KL}(h) &= \frac{1}{n}\sum_{i=1}^{n}\log[f_{h,i}(X_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\Big[\sum_{j\neq i} K(\frac{X_i - X_j}{h})\Big] - \log[(n-1)h] \end{aligned} \tag{2.11}$$

Naturally, we choose h_{KL} such that:

$$h_{KL} = \arg\max_h CV_{KL}(h). \tag{2.12}$$

Since we assumed that X_i are i.i.d., the scores $\log f_{h,i}(X_i)$ are identically distributed and so,

$$\mathbb{E}[CV_{KL}(h)] = \mathbb{E}[\log f_{h,i}(X_i)].$$

Disregarding the leave-one-out effect, we can write

$$\begin{aligned} \mathbb{E}[CV_{KL}(h)] &\approx \mathbb{E}\Big[\int \log f_h(x) f(x) dx\Big] \\ &\approx -\mathbb{E}[d_{KL}(f, f_h)] + \int \log[f(x)] f(x) dx. \end{aligned} \tag{2.13}$$

The second term of the right-hand side does not depend on h. Then, we can expect that we approximate the optimal bandwidth that minimizes $d_{KL}(f, f_h)$.

The Maximum likelihood cross validation has two shortcomings:

- When we have identical observations in one point, we may obtain an infinite value if $CV_{KL}(h)$ and hence we cannot define an optimal bandwidth.

- Suppose we use a kernel function with finite support, e.g., the interval $[-1, 1]$. If an observation X_i is more separated from the other observations than the bandwidth h, the likelihood $f_{h,i}(X_i)$ becomes 0. Hence the score function reaches the value $-\infty$. Maximizing $CV_{KL}(h)$ forces us to use a large bandwidth to prevent this degenerated case. This might lead to slight over-smoothing for the other observations.

2.3 Kernel nonparametric Regression Method

Suppose we have i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$. Using equation (2.1) we know how to estimate the denominator by using the kernel density estimation method. For the numerator one can estimate the joint density using the multiplicative kernel

$$f_{h_1,h_2}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x - X_i)K_{h_2}(y - Y_i).$$

where, $K_{h_1}(x - X_i) = h_1^{-1}K((x - X_i)/h_1)$, $K_{h_2}(x - Y_i) = h_2^{-1}K((x - Y_i)/h_2)$. It is not difficult to show that

$$\int y f_{h_1,h_2}(x,y)dy = \frac{1}{n}\sum_{i=1}^{n} K_{h_1}(x - X_i)Y_i.$$

Based on the methodology of kernel density estimation Nadaraya (1964) and Watson (1964) suggested the following estimator g_h for g.

$$g_h(x) = \frac{\sum_{i=1}^n K_h(x - X_i)Y_i}{\sum_{j=1}^n K_h(x - X_j)} \tag{2.14}$$

In general, the kernel function $K_h(x) = K((x - x_j)/h)$ is taken as probability density function symmetric around zero and parameter h is called smoothing parameter or bandwidth. In addition, with conditions 2.2.2, g_h is a consistent estimator of the regression curve g and its asymptotic distribution is normal with mean zero and asymptotic variance $(g(x))^{-1}\sigma^2 \int (K(s))^2 ds$ as $h \to 0$ and $nh \to \infty$. (See details in Härdle (1990)). This approach can be extended to the multivariate regression problem by considering the multidimensional kernel density estimation method. (see, details in Scott (1992))

2.3.1 k-Nearest Neighbor (k-NN)

One may notice that regression by kernels is based on local averaging of observations Y_i in a fixed neighborhood of x. Instead of this fixed neighborhood k-NN employs varying neighborhoods in the X variable support. That is,

$$g_k(x) = \frac{1}{n}\sum_{i=1}^{N} W_{ki}(x) Y_i, \tag{2.15}$$

where,

$$W_{ki}(x) = \begin{cases} n/k & \text{if } i \in J_x \\ 0 & \text{otherwise,} \end{cases} \tag{2.16}$$

with $J_x = \{i : X_i \text{ is one of the } k \text{ nearest observations to } x\}$

It can be shown that the bias and variance of the k-NN estimator g_k with weights (2.16) are given by, for a fixed x

$$\mathbb{E}[g_k(x)] - g(x) \approx \frac{1}{24(f(x))^3}[g''(x)f(x) + 2g'(x)f'(x)](k/n)^2 \tag{2.17}$$

and

$$Var[g_k(x)] \approx \frac{\sigma^2}{k}. \tag{2.18}$$

We observe that the bias increasing and the variance is decreasing in the smoothing parameter k. To balance this trade-off one should choose $k \sim n^{4/5}$. For details, see Härdle (1990).

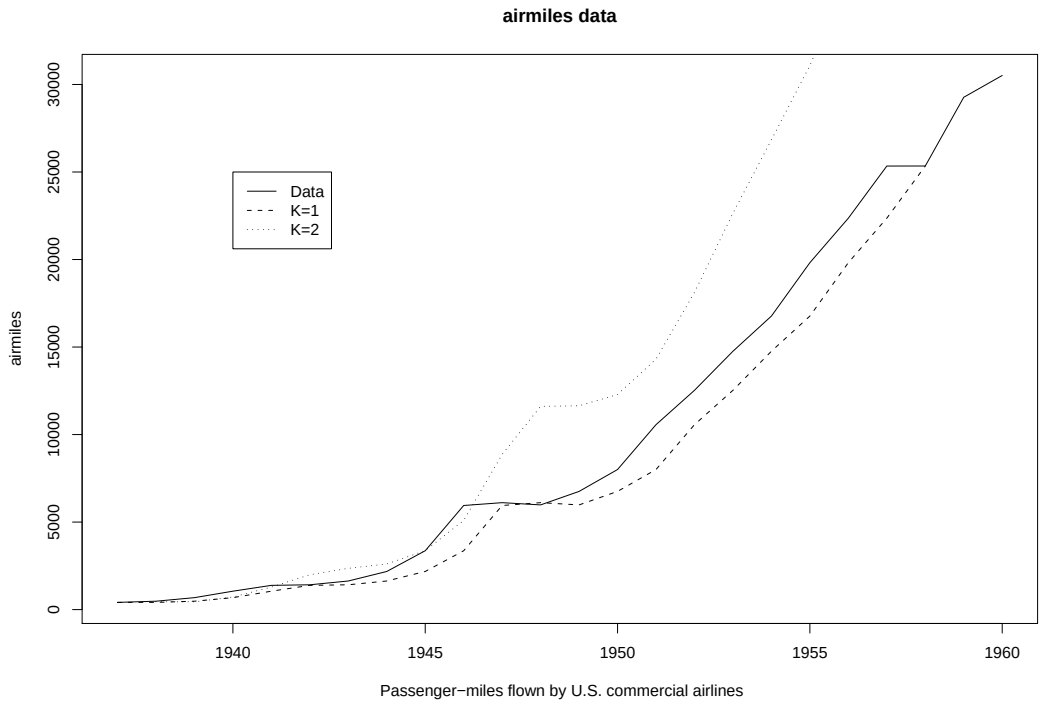

Figure 2.6: Effect of the smoothing parameter k on the k-NN regression estimates.

2.4 Local Polynomial Regression: LOWESS

Cleveland (1979) proposed the algorithm LOWESS, locally weighted scatter plot smoothing, as a resistant method based on local polynomial fits. The basic idea is to start with a local polynomial (a k-NN type fitting) least squares fit and then to use robust methods to obtain the final fit. Specifically, one can first fit a polynomial regression in a neighborhood of x, that is, find $\beta \in \mathbb{R}^{p+1}$ which minimize

$$n^{-1} \sum_{i=1}^{n} W_{ki} \Big(y_i - \sum_{j=0}^{p} \beta_j x^j \Big)^2, \tag{2.19}$$

where W_{ki} denote k-NN weights. Compute the residuals $\hat{\epsilon}_i$ and the scale parameter $\hat{\sigma} = median(\hat{\epsilon}_i)$. Define robustness weights $\delta_i = K(\hat{\epsilon}_i / 6\hat{\sigma})$, where $K(u) = (15/16)(1-u)^2$, if $|u| \leq 1$ and $K(u) = 0$, if otherwise. Then, fit a polynomial regression as in (2.19) but with weights $(\delta_i W_{ki}(x))$. Cleveland suggests that $p = 1$ provides good balance between computational ease and the need for flexibility to reproduce patterns in the data. The smoothing parameter can be determined by cross-validation as

similar to (2.12)

3 Spline Functions

Due to their simple structure and good approximation properties, polynomials are widely used in practice for approximating functions. For this propose, one usually divides the interval $[a, b]$ in the function support into sufficiently small subintervals of the form $[x_0, x_1], \ldots, [x_k, x_{k+1}]$ and then uses a low degree polynomial p_i for approximation over each interval $[x_i, x_{i+1}]$, $i = 0, \ldots, k$. This procedure produces a piecewise polynomial approximating function $s(\cdot)$;

$$s(x) = p_i(x) \text{ on } [x_i, x_{i+1}],\ i = 0, \ldots, k.$$

In the general case, the polynomial pieces $p_i(x)$ are constructed independently of each other and therefore do not constitute a continuous function $s(x)$ on $[a, b]$. This is not desirable if the interest is on approximating a smooth function. Naturally, it is necessary to require the polynomial pieces $p_i(x)$ to join smoothly at knots $x_1, \ldots, x_k$, and to have all derivatives up to a certain order, coincide at knots. As a result, we get a smooth piecewise polynomial function, called a *spline function*.

Definition 3.1 *The function $s(x)$ is called a spline function (or simply "spline") of degree r with knots at $\{x_i\}_{i=1}^k$ if $-\infty =: x_0 < x_1 < \ldots < x_k < x_{k+1} := \infty$, where $-\infty =: x_0$ and $x_{k+1} := \infty$ are set by definition,*

- *for each $i = 0, \ldots, k$, $s(x)$ coincides on $[x_i, x_{i+1}]$ with a polynomial of degree not greater than r;*
- *$s(x), s'(x), \ldots, s^{r-1}(x)$ are continuous functions on $(-\infty, \infty)$.*

The set $\mathcal{S}_r(x_1, \ldots, x_k)$ of spline functions is called *spline space*. Moreover, the spline space is a linear space with dimension $r + k + 1$ (Schumaker1981).

Definition 3.2 *For a given point $x \in (a, b)$ the function*

$$(t-x)_+^r = \begin{cases} (t-x)^r & \text{if } t > x \\ 0 & \text{if } t \leq x \end{cases}$$

is called the truncated power function of degree r with knot x.

Hence, we can express any spline function as a linear combination of $r+k+1$ basis functions. For this, consider a set of interior knots $\{x_1, \ldots, x_k\}$ and the basis functions $\{1, t, t^2, \ldots, t^r, (t-x_1)_+^r, \ldots, (x-t_k)_+^r\}$. Thus, a spline function is given by,

$$s(t) = \sum_{i=0}^{r} \theta_i t^i + \sum_{j=r+1}^{k} \theta_j (t - x_{j-r})_+^r$$

It would be interesting if we could have basis functions that make it easy to compute the spline functions. It can be shown that B-splines form a basis of spline spaces (Schumaker1981). Also, B-splines have an important property toward computation, they are splines which have smallest possible support. In other words, B-splines are zero on a large set. Furthermore, a stable evaluation of B-splines with aid of a recurrence relation is possible.

Definition 3.3 *Let $\Omega_\infty = \{x_j\}_{\{j \in \mathbb{Z}\}}$ be a nondecreasing sequence of knots. The i-th B-spline of order k for the knot sequence Ω_∞ is defined by*

$$B_j^k(t) = -(x_{k+j} - x_j)[x_j, \ldots, x_{k+j}](t - x_j)_+^{k-1} \quad \text{for all} \quad t \in \mathbb{R},$$

where, $[x_j, \ldots, x_{k+j}](t-x_j)_+^{k-1}$ is $(k-1)$th divided difference of the function $(x-x_j)_+^k$ evaluated at points $x_j, \ldots, x_{k+j}$.

From the definition (3.3) we notice that $B_j^k(t) = 0$ for all $t \notin [x_j, x_{j+k}]$. It follows that only k B-splines have any particular interval $[x_j, x_{j+1}]$ in their support, i.e., of all the B-splines of order k for the knot sequence Ω_∞, only the k B-splines $B_{j-k+1}^k, B_{j-k+2}^k, \ldots, B_j^k$ might be nonzero on the interval $[x_j, x_{j+1}]$. (See de Boor (1978) for details). Moreover, $B_j^k(t) > 0$ for all $x \in (x_j, x_{j+k})$ and $\sum_{j \in \mathbb{Z}} B_j^k(t) = 1$,

that is, the B-spline sequence B_j^k consists of nonnegative functions which sum up to 1 and provides a partition of unity. Thus, a spline function can be written as linear combination of B-splines,

$$s(t) = \sum_{j \in \mathbb{Z}} \beta_j B_j^k(t).$$

The value of the function s at point t is simply the value of the function $\sum_{j \in \mathbb{Z}} \beta_j B_j^k(t)$ which makes good sense since the latter sum has at most k nonzero terms.

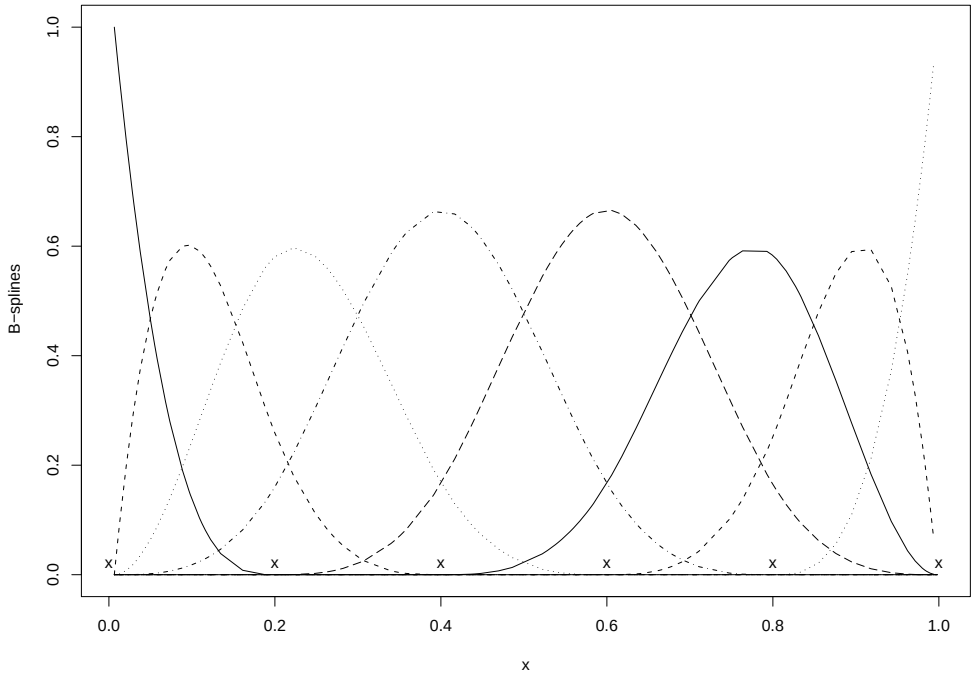

Figure 3.7: Basis Functions with 6 knots placed at "x"

Figure 3.7 shows an example of B-splines basis and their compact support property. This property makes the computation of B-splines easier and numerically stable.

Of special interest is the set of natural splines of order $2m$, $m \in \mathbb{N}$, with k knots at x_j. A spline function is a *natural spline* of order $2m$ with knots at $x_1, \ldots, x_k$, if, in addition to the properties implied by definition (3.1), it satisfies an extra condition:

- s is polynomial of order m outside of $[x_1, x_k]$.

Consider the interval $[a,b] \subset \mathbb{R}$ and the knot sequence $a := x_0 < x_1 < \ldots < x_k < x_{k+1} := b$. Then, $\mathcal{NS}_{2m} = \{s \in \mathcal{S}(\mathcal{P}_{2m}) : s_0 = s|_{[a,x_1)} \quad \text{and} \quad s_k = s|_{[x_k,b)} \in \mathcal{P}_m\}$, is the *natural polynomial spline space of order 2m with knots at* $x_1, \ldots, x_k$. The name "natural spline" stems from the fact that, as a result of this extra condition, s satisfies the so called natural boundary conditions $s^j(a) = s^j(b) = 0$, $j = m, \ldots, 2m-1$.

Now, since the dimension of $\mathcal{S}(\mathcal{P}_{2m})$ is $2m+k$ and we have enforced $2m$ extra conditions to define $\mathcal{NS}_{2m}$, it is natural to expect the dimension of $\mathcal{NS}_{2m}$ to be k. Actually, it is well known that $\mathcal{NS}_{2m}$ is linear space of dimension k. See details in Schumaker (1981).

In some applications it may be possible to deal with natural splines by using a basis for $\mathcal{S}(\mathcal{P}_{2m})$ and enforcing the end conditions. For other applications it is desirable to have a basis for $\mathcal{NS}_{2m}$ itself. To construct such a basis consisting of splines with small supports we just need functions based on the usual B-splines. Particularly, when $m = 2$, we will be constructing basis functions for the *Natural Cubic Spline Space*, $\mathcal{NS}_4$.

Schumaker (1972) showed that the basis obtained by Greville (1969) (except for a normalization constant!) and recently used by Kooperberg and Stone (1991)is a basis for $\mathcal{NS}_4$.

Definition 3.4 *Let* $M(x,y) = (y-x)_+^3$ *and let* $M[x; x_1, \ldots, x_k]$ *be the* $(k-1)$*st divided difference of* M *as a function of* x *taken over the knot sequence* $x_1 \leq x_2 \ldots \leq x_k$ *with* $h_{i+1} = x_{i+1} - x_i$, $i = 1, \ldots, k-1$ *Then*

$$B_i(x) = \begin{cases} M[x; x_1, x_2, x_3]/(h_3 + 2h_2) & \text{if } i = 1 \\ M[x; x_1, x_2, x_3, x_4] & \text{if } i = 2 \\ (x_{i+2} - x_{i-2})M[x; x_{i-2}, \ldots, x_{i+2}] & \text{if } i = 3, \ldots, k-2 \\ M[x; x_{k-3}, x_{k-2}, x_{k-1}, x_k] & \text{if } i = k-1 \\ M[x; x_{k-2}, x_{k-1}, x_k](h_{k-1} + 2h_k) & \text{if } i = k \end{cases}$$

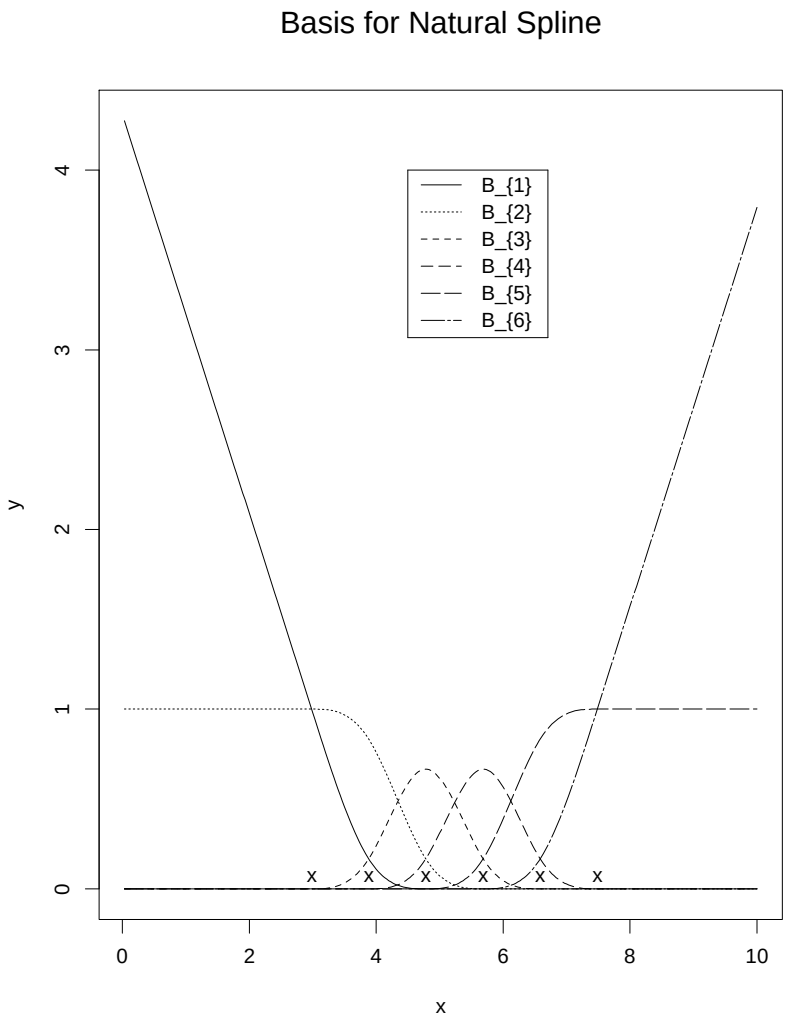

Figure 3.8: Basis Functions with 6 knots placed at "x"

3.1 Logspline Density Estimation

In 1991, Kooperberg and Stone introduced another type of algorithm to estimate an univariate density. This algorithm was based on the work of Stone (1990) and Stone and Koo (1985) where the theory of the logspline family of functions was developed.

Consider an increasing sequence of knots $\{t_j\}_{j=1}^{K}$, $K \geq 4$, in $\mathbb{R}$. Denote by $\mathcal{S}_0$ the set of real functions such that s is a cubic polynomial in each interval of the form $(-\infty, t_1], [t_1, t_2], \ldots, [t_K, \infty)$. Elements in $\mathcal{S}_0$ are the well-known cubic splines

with knots at $\{t_j\}_{j=1}^{K}$. Notice that $\mathcal{S}_0$ is a $(K+4)$-dimensional linear space. Now, let $\mathcal{S} \subset \mathcal{S}_0$ such that the dimension of $\mathcal{S}$ is K with functions $s \in \mathcal{S}$ linear on $(-\infty, t_1]$ and on $[t_K, \infty)$. Thus, $\mathcal{S}$ has a basis of the form $1, B_1 \ldots, B_{K-1}$, such that B_1 is linear function with negative slope on $(-\infty, t_1]$ and $B_2, \ldots, B_{K-1}$ are constant functions on the same interval. Similarly, B_{K-1} is linear function with positive slope on $[t_K, \infty)$ and $B_1, \ldots, B_{K-2}$ are constant on the interval $[t_K, \infty)$ (Kooperberg and Stone1991).

Let Θ be the parametric space of dimension $p = K - 1$, such that for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$, $\theta_1 < 0$ and $\theta_p > 0$. Then, define

$$c(\boldsymbol{\theta}) = \log(\int_{\mathbb{R}} \exp(\sum_{j=1}^{K-1} \theta_j B_j(x))dx$$

and

$$f(x; \boldsymbol{\theta}) = \exp\{\sum_{j=1}^{K-1} \theta_j B_j(x) - c(\boldsymbol{\theta})\}.$$

The p-parametric exponential family $f(\cdot, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$ of positive twice differentiable density function on $\mathbb{R}$ is called logspline family and the corresponding log-likelihood function is given by

$$L(\boldsymbol{\theta}) = \sum \log f(x; \boldsymbol{\theta}) \quad ; \boldsymbol{\theta} \in \Theta \ .$$

The log-likelihood function $L(\boldsymbol{\theta})$ is strictly concave and hence the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is unique, if it exists. We refer to $\hat{f} = f(\cdot, \hat{\boldsymbol{\theta}})$ as the *logspline density estimate*. Note that the estimation of $\hat{\boldsymbol{\theta}}$ makes logspline procedure not essentially nonparametric. Thus, estimation of $\boldsymbol{\theta}$ by Newton-Raphson, together with small numbers of basis function necessary to estimate a density, make the logspline algorithm extremely fast when it is compared with Gu's algorithm for smoothing spline density estimation, (Gu1993).

In the Logspline approach the number of knots is the smoothing parameter. That is, too many knots leads to a noisy estimate while too few knots gives a very smooth curve. Based on their experience of fitting logspline models, Kooperberg and Stone

provide a table with the number of knots based on the number of observations. No indication was found that the number of knots takes in consideration the structure of the data (number of modes, bumps, asymmetry, etc.). However, an objective criterion for the choice of the number of knots, *Stepwise Knot Deletion and Stepwise knot Addition*, are included in the logspline procedure.

For $1 \leq j \leq p$, let B_j be a linear combination of a truncated power basis $(x-t_k)_+^3$ for the a knot sequence $t_1, \ldots, t_p$, that is,

$$B_j(x) = \beta_j + \beta_{j0}x + \sum_k \beta_{jk}(x-t_k)_+^3.$$

Then

$$\sum_j \theta_j B_j(x) = \sum \theta_j \beta_{j0} + \sum_j \sum_k \beta_{jk}\theta_j (x-t_k)_+^3 \ .$$

Let $\sum_j \hat{\theta_j}\beta_{jk} = \boldsymbol{\beta_k^T}\ \hat{\boldsymbol{\theta}}$. Then, for $1 \leq k \leq K$ Kooperberg and Stone (1991),

$$SE(\boldsymbol{\beta_k^T}\hat{\boldsymbol{\theta}}) = \sqrt{\boldsymbol{\beta_k^T}(\boldsymbol{I}(\hat{\boldsymbol{\theta}}))^{-1}\boldsymbol{\beta_k}})$$

where $\boldsymbol{I(\theta)}$ is the Fisher information matrix obtained from the log-likelihood function.

The knots t_1 and t_K are considered permanent knots, and t_k, $2 \leq k \leq K$, are nonpermanent knots. Then at any step delete (similarly for addition step) that knot which has the smallest value of $|\boldsymbol{\beta_k^T}\hat{\boldsymbol{\theta}}|/SE(\boldsymbol{\beta_k^T}\hat{\boldsymbol{\theta}})$. In this matter, we have a sequence of models which ranges from 2 to $p-1$ knots. Now, denote by $\hat{L}_m$ the log-likelihood function of the mth model ($2 \leq m+2 \leq p-1$) evaluated at the maximum likelihood estimate for that model. To specify a stop criteria, Kooperberg and Stone make use of the Akaike Information Criterion (AIC), that is, $AIC_{\alpha,m} = -2\hat{L}_m + \alpha(p-m)$ and choose $\hat{m}$ that minimizes $AIC_{3,m}$. There is no theoretical justification for choosing $\alpha = 3$. The choice was made, according to them, because this value of α makes the probability that $\hat{f}$ is bimodal when f is $Gamma(5)$ to be about .1.

It would be interesting to have an algorithm which combines the low computational cost of logsplines (due to B-splines and the estimation of their coefficients) and the performance of the automatic smoothing parameter selection developed by Gu (1993).

3.2 Splines Density Estimation: A Dimensionless Approach

Let $X_1, \ldots, X_n$ a random sample from a probability density f on a finite domain $\mathcal{X}$. Assuming that $f > 0$ on $\mathcal{X}$, one can make a logistic transformation $f = e^g/(\int e^g)$. We know that this transformation is not one-to-one and Gu and Qiu (1993) proposed side conditions on g such that $g(x_0) = 0, x_0 \in \mathcal{X}$ or $\int_{\mathcal{X}} g = 0$. Given those conditions we have to find the minimizer of the penalized log-likelihood

$$-\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \log \int_{\mathcal{X}} e^g + \frac{\lambda}{2} J(g) \tag{3.5}$$

in a Hilbert space $\mathcal{H}$, where J is a roughness penalty and λ is the smoothing parameter. The space $\mathcal{H}$ is a Hilbert space where the evaluation is continuous so that the first term in (3.5) is continuous. The penalty term J is a seminorm in $\mathcal{H}$ with a null space $J_\perp$ of finite dimension $M \geq 1$. By taking a finite dimensional $J_\perp$ one prevents interpolation (i.e. the empirical distribution) and a quadratic J makes easier the numerical solution of the variational problem (3.5). Since, $\mathcal{H}$ is an infinite dimensional space, the minimizer of (3.5) is, in general, not computable. Thus, (Gu and Qiu1993) propose calculating the solution of the variational problem in finite dimensional space, say, $\mathcal{H}_n$, where n is the sample size.

The performance of the smoothing spline estimator depends upon the choice of the smoothing parameter λ. Gu (1993), suggested a performance-oriented iteration procedure (GCV-like procedure) which updates g and λ jointly according to a performance estimate. The performance is measured by a loss function which was taken as a symmetrized Kullback-Leibler distance between $e^g/\int e^g$ and $e^{g_0}/\int e^{g_0}$. Specifically, if one solves the variational problem (3.5) in $\mathcal{H}_n$ by a standard Newton-Raphson procedure, then by starting from a current iterate $\tilde{g}$, instead of calculating the next iterate with a fixed λ, one may choose a λ that minimizes the loss function.

Under this approach, one might ask the following questions:

- Is it possible to estimate a density using $K \leq n$ basis functions instead of the

original n such that it reduces the computational cost of getting the solution (3.5) significantly ?

- How good would such an approximation be ?

Dias (1998) gave reasonable answers to those questions by using the basis functions $B_i(x)$ given in Definition (3.4) that can be easily extend to a multivariate case by a tensor product.

4 Splines nonparametric Regression: The thin-plate spline on $\mathbb{R}^d$

There are many applications where a unknown function g of one or more variables and a set of measurements are given such that:

$$y_i = \mathcal{L}_i g + \epsilon_i \tag{4.1}$$

where $\mathcal{L}_1, \ldots, \mathcal{L}_n$ are linear functionals defined on some linear space $\mathcal{H}$ containing g, and $\epsilon_1, \ldots, \epsilon_n$ are measurement errors usually assumed to be independently identically normal distributed with mean zero and unknown variance σ^2. Typically, the $\mathcal{L}_i$ will be point evaluation of the function g.

Straight forward least square fitting is often appropriate but it produces a function which is not sufficiently smooth for some data fitting problems. In such cases, it may be better to look for a function which minimizes a criterion that involves a combination of goodness of fit and an appropriate measure of smoothness. Let $t = (x_1, \ldots, x_d)$, $t_i = (x_1(i), \ldots, x_d(i))$ for $i = 1, \ldots, n$ and the evaluation functionals $\mathcal{L}_i g = g(t_i)$, then the regression model (4.1) becomes,

$$y_i = g(x_1(i), \ldots, x_d(i)) + \epsilon_i. \tag{4.2}$$

The thin-plate smoothing spline is the solution to the following variational problem. Find $g \in \mathcal{H}$ to minimize

$$L_\lambda(g) = \frac{1}{n}\sum_{i=1}^{n}(y_i - g(t_i))^2 + \lambda J_m^d(g) \tag{4.3}$$

where λ is the smoothing parameter which controls the trade off between fidelity to the data and smoothness and the penalty term J_m^d is given by

$$J_m^d(g) = \sum_{\alpha_1+\ldots+\alpha_d=m} \frac{m!}{\alpha_1!\ldots,\alpha_d!}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\Big(\frac{\partial^m g}{\partial x_1^{\alpha_1}\ldots\partial x_d^{\alpha_d}}\Big)^2\prod_j dx_j.$$

The condition $2m - d > 0$ is necessary and sufficient in order to have bounded evaluation functionals in $\mathcal{H}$, i.e., $\mathcal{H}$ is a reproducing kernel in Hilbert space. Moreover, the null space of the penalty term J_m^d is the M-dimensional space spanned by polynomials $\phi_1,\ldots,\phi_M$ of degree less or equal to $m-1$, e.g., $\phi_i(t) = t^{j-1}/(j-1)!$, for $j = 1,\ldots,m$.

It can be shown that (see Wahba (1990)), if $t_1,\ldots,t_n$ are such that least squares regression on $\phi_1,\ldots,\phi_M$ is unique, then (4.3) has a unique minimizer g_λ, with representation

$$\begin{aligned} g_\lambda(t) &= \sum_{i=1}^{n} c_i E_m(t,t_i) + \sum_{j=1}^{M} b_j\phi_j(t) \\ &= Qc + Tb \end{aligned} \tag{4.4}$$

where, T is a $n \times M$ matrix with entries $\phi_j(t_l)$ for $j = 1,\ldots,M$, $l = 1,\ldots,n$ and Q is a $n \times n$ matrix with entries $E_m(t_l,t_i)$, for $i = 1,\ldots,n$. The function E_m is a Green's function for the m-iterate Laplacian ((Wahba1990)). For example, when $d = 1$, $E_m(t,t_i) = (t-t_i)_+^{m-1}/(m-1)!$. The coefficients c and b can be determined by substituting (4.4) into (4.3). Thus, the optimization problem (4.3) subject to $T'c = 0$, is reduced to a linear system of equations which is solved by standard matrix decomposition such as QR decomposition. The constraint $T'c = 0$ is necessary to guarantee that when computing the penalty term at g_λ, $J_m^d(g_\lambda)$ is conditionally positive definite. [2] Efforts have been done in order to reduce substantially the

[2] See, (Wahba1990 Silverman and Green1994)

computational cost of solving smoothing splines fitting by introducing the concept of H-splines ((Luo and Wahba1997) and (Dias1999)), where the number of basis functions and λ act as the smoothing parameters.

A major conceptual problem with spline smoothing is that it is defined implicitly as the solution to a variational problem rather than as an explicit formula involving the data values. This difficulty can be resolved, at least approximately, by considering how the estimate behaves on large data sets. It can be shown from the quadratic nature of (4.3) that g_λ is linear in the observations y_i, in the sense that there exists a weight function $H_\lambda(s,t)$ such that

$$g_\lambda(s) = \sum_{i=1}^{n} y_i H_\lambda(s,t_i). \tag{4.5}$$

It is possible to obtain the asymptotic form of the weight function, and hence an approximate explicit form of the estimate. For the sake of simplicity consider $d = 1$, $m = 2$ and suppose that the design points have local density $f(t)$ with respect to a Lesbegue measure on $\mathbb{R}$. Under mild conditions (see, Silverman (1984)), we have as $n \to \infty$,

$$H_\lambda(s,t) = \frac{1}{f(t)}\frac{1}{h(t)}K(\frac{s-t}{h(t)}),$$

where the kernel function K is given by

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

and the bandwidth $h(t)$ satisfies

$$h(t) = \lambda^{1/4} n^{-1/4} f(t)^{-1/4}.$$

Based on these formulas, we can see that the spline smoother is approximately a convolution smoothing method but the data are not convolved with a kernel with fixed bandwidth, in fact, h varies across the sample.

4.1 Additive Models

The additive model is a generalization of the usual linear regression model and what has made it so popular for statistical inference is that the linear model is linear in the predictor variables (explanatory variables). Once we have fitted the linear model we can examine the predictor variables separately, in the absence of interactions. Additive models also are linear in their predictor variables. An additive model is defined by

$$y_i = \alpha + \sum_{j=1}^{p} g_j(t_j) + \epsilon_i \tag{4.6}$$

where t_j are the predictor variables and as defined before in section 4, ϵ_i are uncorrelated error measurements with $\mathbb{E}[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$. The functions g_j are unknown but assumed to be smooth functions lying in some metric space. Section 4 describes a general framework for defining and estimating general nonparametric regression models which includes additive models as a special case. For this, suppose that Ω is the space of the vector predictor t and assume the $\mathcal{H}$ is reproducing kernel in Hilbert space. Hence $\mathcal{H}$ has the decomposition

$$\mathcal{H} = \mathcal{H}_0 + \sum_{k=1}^{p} \mathcal{H}_k \tag{4.7}$$

where $\mathcal{H}_0$ is spanned by $\phi_1, \ldots, \phi_M$ and $\mathcal{H}_k$ has the reproducing kernel $E_k(\cdot, \cdot)$, defined in section 4. The space $\mathcal{H}_0$ is the space of functions that are not to be penalized in the optimization. For example, recall equation (4.3) and let $m = 2$ then $\mathcal{H}_0$ is the space of linear functions in t.

The optimization problem becomes: For a given set of predictors $t_1, \ldots, t_n$, find the minimizer of

$$\sum_{i=1}^{n} \{y_i - \sum_{k=0}^{p} g_k(t_i)\}^2 + \sum_{k=1}^{k} \lambda_k ||g_k||^2_{\mathcal{H}_k}, \tag{4.8}$$

with $g_k \in \mathcal{H}_k$. Then, the theory of reproducing kernel guarantees that a minimizer

exists and has the form

$$\hat{g} = \sum_{k=1}^{p} Q_k c + Tb, \tag{4.9}$$

where Q_k and T are given in equation (4.4) and the vectors c and b are found by minimizing the finite dimensional penalized least square criterion

$$||y - Tb - \sum_{k=1}^{p} Q_k c|| + \sum_{k=1}^{p} \lambda_k c_k^T Q_k c^2. \tag{4.10}$$

This general problem (4.9) can potentially be solved by a backfitting type algorithm (Hastie and Tibshirani1990).

Algorithm 4.1 *1. Initialize* $g_j = g_j^{(0)}$ *for* $j = 0, \ldots, p$.

2. Cycle $j = 0, \ldots, p, \ldots, j = 0, \ldots, p, \ldots$

$$\hat{g}_j = S_j(\mathbf{y} - \sum_{j \neq k} g_j(t_j))$$

3. Continue (ii) until the individual functions do not change.

where $\mathbf{y} = (y_1, \ldots, y_n)$, $S_j = Q_k(Q_k + \lambda_k I)^{-1}$, for $j = 1, \ldots, p$, and $S_0 = T(T^T T)^{-1}$. One may observe that omitting the constant term α in (4.6) does not change the resulting estimates.

4.2 Generalized Cross-Validation Method for Splines non-parametric Regression

Without loss of generality, let's take $d = 1$ and $m = 2$. The solution of (4.3) depends strongly on the smoothing parameter. Craven and Wahba (1979) provide an automatic data-driven procedure to estimate λ. For this, let $g_\lambda^{[k]}$ be the minimizer of

$$\frac{1}{n} \sum_{i \neq k} (y_i - g(t_i))^2 + \lambda \int (g''(u))^2 du,$$

the optimization problem with the kth data point left out. Then following Wahba's notation, the ordinary cross-validation function $V_0(\lambda)$ is defined as

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - g_\lambda^{[k]}(t_k))^2, \tag{4.11}$$

and the *leave-one-out* estimate of λ is the minimizer of $V_0(\lambda)$. To proceed, we need to describe the influence matrix. It is not difficult to show (see (Wahba1990)) that, for fixed λ we have by (4.5) that g_λ is linear in the observations y_i , that is, in matrix notation

$$\mathbf{g}_\lambda = H_\lambda \mathbf{y}.$$

At this stage, one may think that the computation of this problem is prohibitive but Craven and Wahba (1979) give us a very useful mathematical identity, which will not be proved here, but is

$$(y_k - g_\lambda^{[k]}(t_k)) = (y_k - g_\lambda(t_k))/(1 - h_{kk}(\lambda), \tag{4.12}$$

where $h_{kk}(\lambda)$ is the kth entry of H_λ. By substituting (4.12) into (4.11) we obtain a simplified form of V_0, that is,

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - g_\lambda(t_k))^2/(1 - h_{kk}(\lambda))^2 \tag{4.13}$$

The right hand of (4.13) is easier to compute than (4.11), however the GCV is even easier. The generalized cross-validation (GCV) is method for choosing the smoothing parameter λ, which is based on *leaving-one-out*, but it has two advantages. It is easy to compute and it posses some important theoretical properties the would be impossible to prove for *leaving-one-out*, although, as pointed out by Wahba, in many cases the GCV and leaving-one-out estimates will give similar answers. The GCV function is defined by

$$V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - g_\lambda(t_k))^2/(1 - \bar{h}_{kk}(\lambda))^2 = \frac{\frac{1}{n}||(I - H_\lambda)\mathbf{y}||^2}{[\frac{1}{n}tr(I - H_\lambda]^2}, \tag{4.14}$$

where $\bar{h}_{kk}(\lambda) = (1/n)tr(H_\lambda)$, with $tr(H_\lambda)$ standing for the trace of H_λ. Note that $V(\lambda)$ is a weighted version of $V_0(\lambda)$. In addition, if $h_{kk}(\lambda)$ does not depend on k, then $V_0(\lambda) = V(\lambda)$ for all $\lambda > 0$.

It is important to observe that GCV is a predictive mean square error criteria. Note that by defining the predictive mean square error $T(\lambda)$ as

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\mathcal{L}_i g_\lambda - \mathcal{L}_i g)^2 \tag{4.15}$$

where, $\mathcal{L}_i$ is the evaluation functional defined in section 3.2, the GCV estimate of λ is the minimizer of (4.15). Consider the expected value of $T(\lambda)$,

$$\mathbb{E}[T(\lambda)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\mathcal{L}_i g_\lambda - \mathcal{L}_i g)^2]. \tag{4.16}$$

The GCV theorem (Wahba1990) says that if g is in a reproducing kernel Hiblert space then there is a sequence of minimizers $\tilde{\lambda}(n)$ of $\mathbb{E}V(\lambda)$ that comes close to achieving the minimum possible value of the expected mean square error, $\mathbb{E}[T(\lambda)]$, using $\tilde{\lambda}(n)$, as $n \to \infty$. That is, let the expectation inefficiency I_n^* be defined as

$$I_n^* = \frac{\mathbb{E}[T(\tilde{\lambda}(n))]}{\mathbb{E}[T(\lambda^*)]},$$

where λ^* is the minimizer of $\mathbb{E}[T(\lambda)]$. Then, under mild conditions as such the ones described and discussed by Golub, Heath and Wahba (1979) and Craven and Wahba (1979), we have $I_n^* \downarrow 1$ as $n \to \infty$.

Figure 4.9 shows the scatter plot of the revenue passenger miles flown by commercial airlines in the United States for each year from 1937 to 1960. (This data can be found in the software). The smoothing parameter λ was computed by GCV method through the R function `smooth.spline()`.

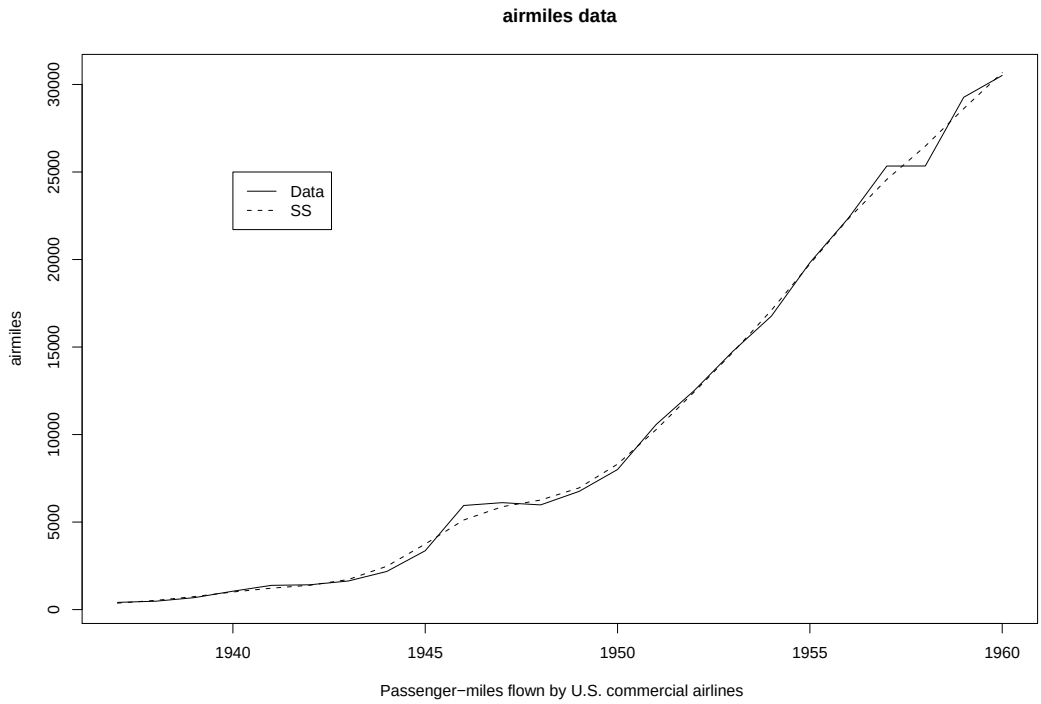

Figure 4.9: Smoothing spline fitting with smoothing parameter obtained by GCV method

5 Final Comments

Comparing with parametric techniques we have, for the nonparametric approach, more flexibility since it allows one to choose from the infinite dimensional class of functions where the underlying regression curve is assumed to belong. In general, this type of choice depends on the unknown smoothness of the true curve. But for most of the cases one can assume mild restrictions such that a regression curve has an absolutely continuous first derivative and a square integrable second derivative. Nevertheless, nonparametric estimators are less efficient than the parametric ones when a parametric model is valid. For many parametric estimators the mean square error goes to zero with rate of n^{-1}, while nonparametric estimators have rate of $n^{-\alpha}$, $\alpha \in [0, 1]$, and α depends on the smoothness of the underlying curve. When the postulate parametric model is not valid, many parametric estimators cannot have, *ad hoc*, rate n^{-1}. In fact, those estimators will not converge to the true curve. One of the

advantages of the adaptive basis functions procedures, e.g., H-splines methods is the ability to vary the amount of smoothing in response to the inhomogeneous curvature of the true functions at different locations. Those methods have been very successful in capturing the structure of the unknown function. In general, nonparametric estimators are good candidates when one does not know the form of the underlying curve.

References

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experiments: An Introduction to Design, Data Analysis, and Model Building*, John Wiley and Sons (New York, Chichester).

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *J. Amer. Statist. Assoc.* **74**(368): 829–836.

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

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

Dias, R. (1998). Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation* **60**: 277–294.

Dias, R. (1999). Sequential adaptive non parametric regression via h-splines, *Communications in Statistics: Computations and Simulations* **28**: 501–515.

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics* **21**(2): 215–223.

Greville, T. N. (1969). *Theory and Applications of Spline Functions*, Academic Press, New York.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm, *J. of the Amer. Stat'l. Assn.* **88**: 495–504.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation:theory, *Ann. of Statistics* **21**: 217–234.

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

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, Chapman and Hall.

Kooperberg, C. and Stone, C. J. (1991). A study of logspline density estimation, *Computational Statistics and Data Analyis* **12**: 327–347.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines, *Journal of the American Statistical Association* **92**: 107–116.

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

Parzen, E. (1962). On estimation of a probability density function and mode, *Ann. of Mathematical Stat.* **33**: 1065–1076.

Prakasa-Rao, B. L. S. (1983). *Nonparametric Functional Estimation*, Academic Press (Duluth, London).

Schumaker, L. L. (1972). *Spline Functions and Aproximation theory*, Birkhauser.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*, WileyISci:NJ.

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

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method, *Ann. of Statistics* **12**: 898–916.

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

Silverman, B. W. and Green, P. J. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall (London).

Stone, C. J. (1990). Large-sample inference for log-spline models, *Ann. of Statistics* **18**: 717–741.

Stone, C. J. and Koo, C.-Y. (1985). Logspline density estimation, *Contemporary Mathematics* pp. 1–158.

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

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