Nonparametric Estimation: Smoothing and Data Visualization

Ronaldo Dias

Universidade Estadual de Campinas

1$^{\underline{o}}$ Colóquio da Região Sudeste
Abril de 2011

Preface

In recent years more and more data have been collected in order to extract information or to learn valuable characteristics about experiments, phenomena, observational facts, etc.. This is what it's been called learning from data. Due to their complexity, several datasets have been analyzed by nonparametric approaches. This field of Statistics impose minimum assumptions to get useful information from data. In fact, nonparametric procedures, usually, *"let the data speak for themselves"*. This work is a brief introduction to a few of the most useful procedures in the nonparametric estimation toward smoothing and data visualization. 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 methods. The text is designed for undergraduate students in mathematical sciences, engineering and economics. It requires at least one semester in calculus, probability and mathematical statistics.

Contents

Preface iii

List of Figures vii

1 Introduction 1

2 Kernel estimation 3
- 2.1 The Histogram 3
- 2.2 Kernel Density Estimation 4
  - 2.2.1 The Nearest Neighbor Method 7
- 2.3 Some Statistical Results for Kernel Density Estimation 8
- 2.4 Bandwidth Selection 10
  - 2.4.1 Reference to a Standard Distribution 10
  - 2.4.2 Maximum likelihood Cross-Validation 11
  - 2.4.3 Least-Squares Cross-Validation 13
- 2.5 Orthogonal series estimators 14

3 Kernel nonparametric Regression Method 17
- 3.1 k-Nearest Neighbor (k-NN) 18
- 3.2 Local Polynomial Regression: LOWESS 19
- 3.3 Penalized Maximum Likelihood Estimation 20
  - 3.3.1 Computing Penalized Log-Likelihood Density Estimates 24

4 Spline Functions 25
- 4.1 Acquiring the Taste 25
- 4.2 Logspline Density Estimation 27
- 4.3 Splines Density Estimation: A Dimensionless Approach 29

5 The thin-plate spline on $\mathbb{R}^d$ 33
- 5.1 Additive Models 35
- 5.2 Generalized Cross-Validation Method for Splines nonparametric Regression 36

6 Regression splines, P-splines and H-splines 41
- 6.1 Sequentially Adaptive H-splines 41
- 6.2 P-splines 46
- 6.3 Adaptive Regression via H-Splines Method 47
- 6.4 A Bayesian Approach to H-splines 48

7 Final Comments 55

Bibliography 57

List of Figures

2.1 Naive estimate constructed from Old faithful geyser data with $h = 0.1$. 5
2.2 Kernel density estimate constructed from Old faithful geyser data with Gaussian kernel and $h = 0.25$ 6
2.3 Bandwidth effect on kernel density estimates. The data set income was rescaled to have mean 1. 7
2.4 Effect of the smoothing parameter k on the estimates 9
2.5 Comparison of two bandwidths, $\hat{\sigma}$ (the sample standard deviation) and $\hat{R}$ (the sample interquartile) for the mixture $0.7 \times N(-2,1) + 0.3 \times N(1,1)$. 11
2.6 Effect of the smoothing parameter K on the orthogonal series method for density estimation 15

3.1 Effect of bandwiths on Nadaraya-Watson kernel 18
3.2 Effect of the smoothing parameter k on the k-NN regression estimates. . 19
3.3 Effect of the smoothing parameter using LOWESS method. 20

4.1 Basis Functions with 6 knots placed at “x” 27
4.2 Basis Functions with 6 knots placed at “x” 28
4.3 logspline density estimation for .5*N(0,1)+.5*N(5,1) 30
4.4 Histogram, SSDE, Kernel and Logspline density estimates 31

5.1 True, tensor product, gam non-adaptive and gam adaptive surfaces . . . 37
5.2 Smoothing spline fitting with smoothing parameter obtained by GCV method 39

6.1 Spline least square fittings for different values of K 42
6.2 Five thousand replicates of $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \epsilon$. . . . 43
6.3 Five thousand replicates of the affinity and the partial affinity for adaptive nonparametric regression using H-splines with the true curve. . . . 45
6.4 Density estimates of the affinity based on five thousand replicates of the curve $y_i = x_i^3 + \epsilon_i$ with $\epsilon_i \sim N(0, .5)$. Solid line is a density estimate using beta model and dotted line is a nonparametric density estimate. . 45
6.5 A comparison between smoothing splines (S-splines) and hybrid splines (H-splines) methods. 46
6.6 smooth spline and P-spline 47
6.7 H-spline fitting for airmiles data 49

6.8 Estimation results: a) Bayesian estimate with $a = 17$ and $\psi(K) = K^3$ (dotted line); b) (SS) smoothing splines estimate (dashed line). The true regression function is also plotted (solid line). The SS estimate was computed using the R function `smooth.spline` from which 4 degrees of freedom were obtained and λ was computed by GCV. 52
6.9 One hundred estimates of the curve 6.8 and a Bayesian confidence interval for the regression curve $g(t) = \exp(-t^2/2)\cos(4\pi t)$ with $t \in [0, \pi]$. . 53

Chapter 1

Introduction

Probably, the most used procedure to describe a possible relationship among variables is the statistical technique known as regression analysis. 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) be 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,

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

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 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 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 correctly specified. For most parametric estimators the risk will decay to zero at a rate of n^{-1} while nonparametric estimators decay at a 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 incorrectly specified, ad hoc, the rate n^{-1} cannot be achieved. In fact, the parametric estimator does not even converge to the true regression curve.

Chapter 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

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

where $f_{TY}(t, y)$ denotes the joint density of (T, Y) and $f_T(t)$ the marginal density of T. In order to provide an estimate $\hat{g}(t)$ of g we need to obtain estimates of $f_{TY}(t, y)$ and $f_T(t)$. Consequently, density estimation methodologies 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 of Box, Hunter and Hunter (1978) 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 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). 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,

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

and define the estimator,

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}). \tag{2.2}$$

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 (Silverman (1986)), it could give a misleading impression to a untrained observer since its 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.

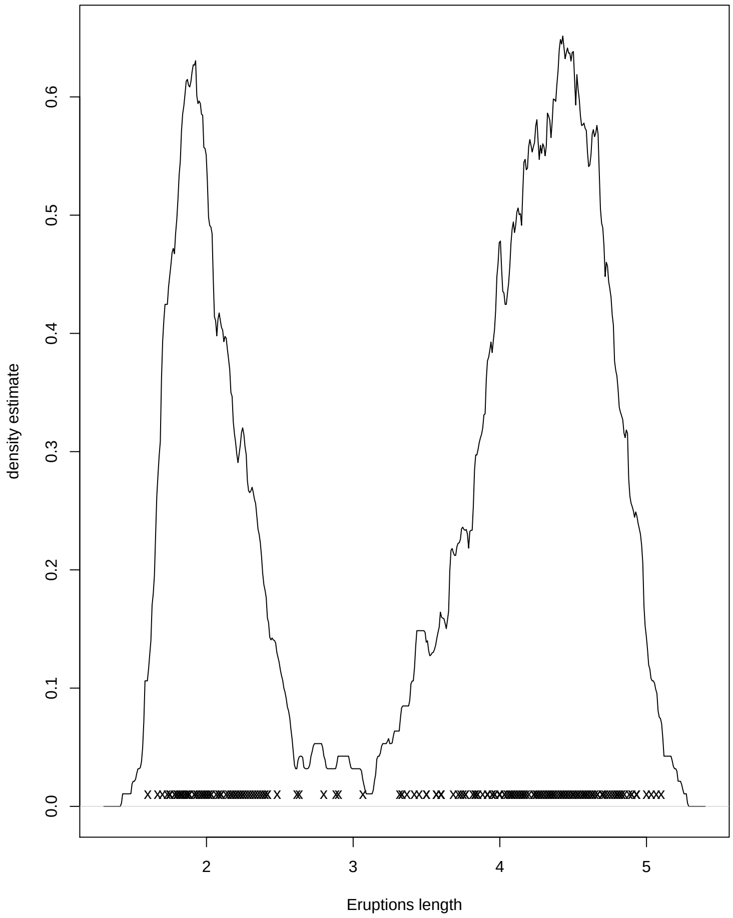

Figure 2.1: Naive estimate constructed from Old faithful geyser data with $h = 0.1$

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

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

In other words $K(x)$ is a probability density function, as for instance, the Gaussian 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 Gaussian 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 a Gaussian kernel is used.

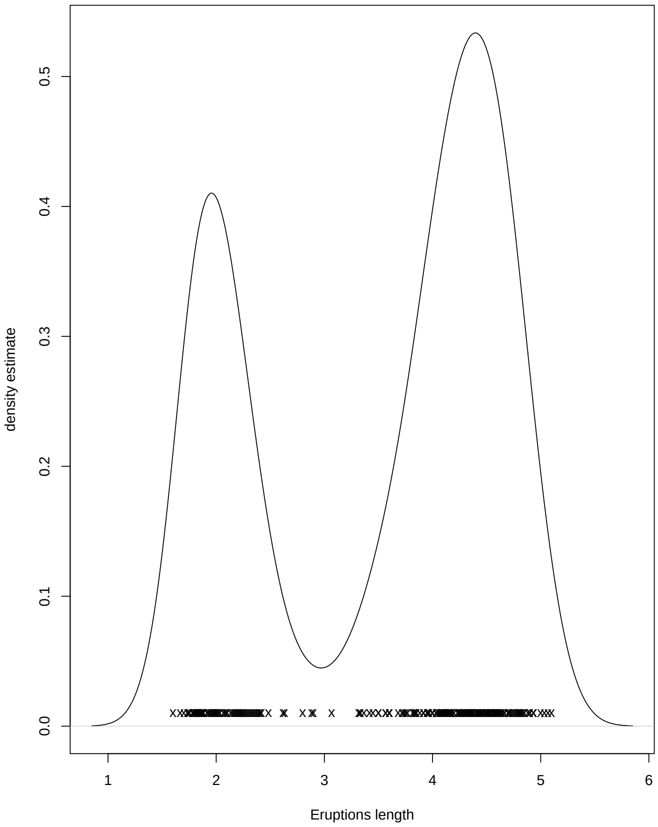

Figure 2.2: Kernel density estimate constructed from Old faithful geyser data with Gaussian kernel and $h = 0.25$

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 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 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 (Silverman (1986)) literature on kernel density estimation studying its mathematical properties and proposing several algorithms to obtain estimates based on it. This method of density estimation became, apart from the histogram, the most commonly used estimator. However it has drawbacks when the underlying

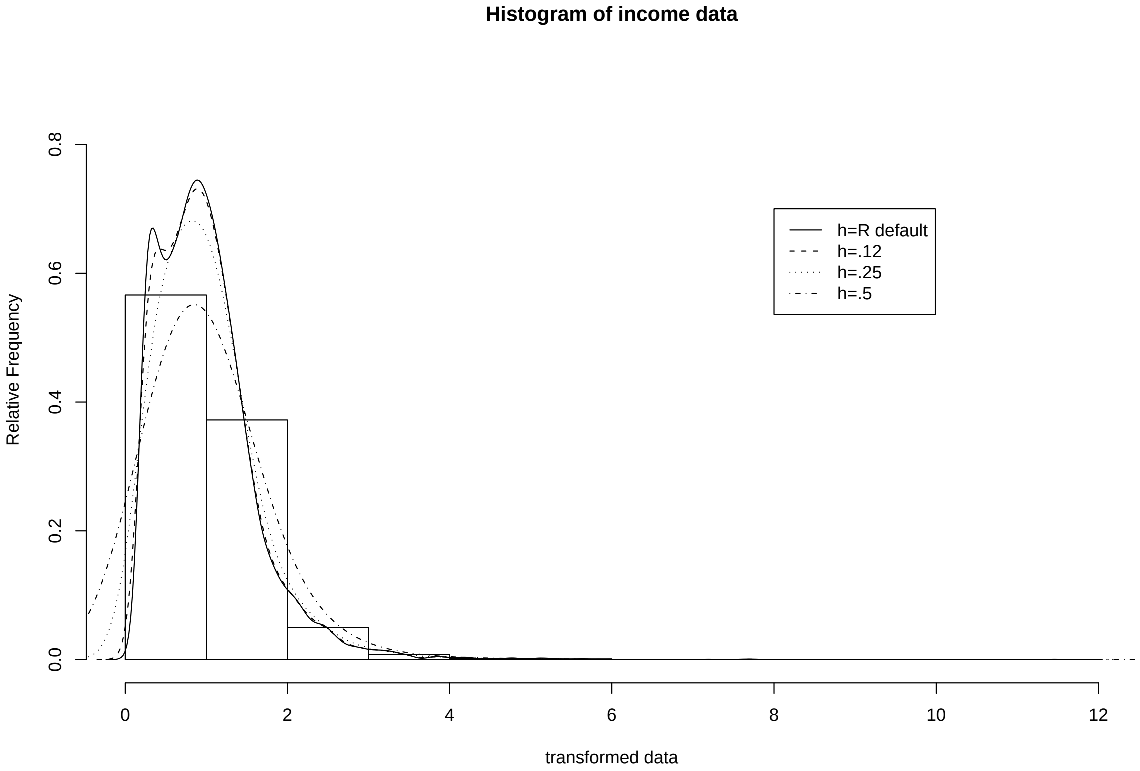

Figure 2.3: Bandwidth effect on kernel density estimates. The data set income was rescaled to have mean 1.

density has long tails Silverman (1986). 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 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.

2.3 Some Statistical Results for 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. Then we have, for a nonstochastic h

$$\begin{aligned} E[\hat{f}(x)] &= \frac{1}{nh} \sum_{i=1}^{n} E[K(\frac{x - X_i}{h})] \\ &= \frac{1}{h} E[K(\frac{x - X_i}{h})] \\ &= \frac{1}{h} \int K(\frac{x - u}{h}) f(u) du \\ &= \int K(y) f(x + yh) dy. \end{aligned} \tag{2.3}$$

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.

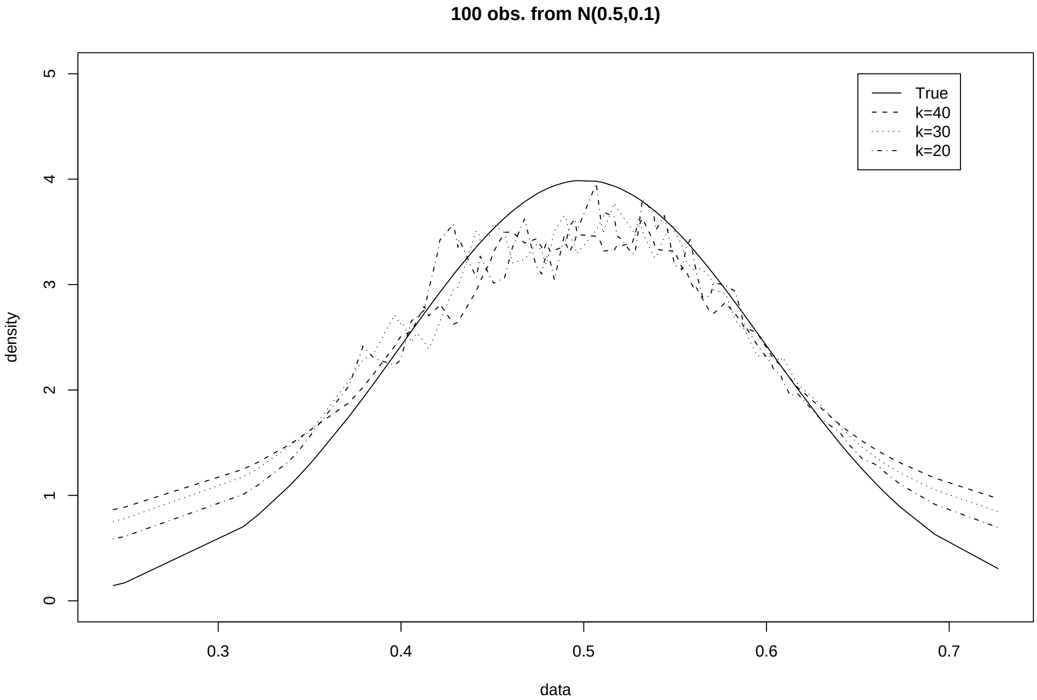

Figure 2.4: Effect of the smoothing parameter k on the estimates

To compute the bias of this estimator we have to make the assumption that the underlying density is twice differentiable and satisfies the following conditions Prakasa-Rao (1983):

- **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 by 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. Parzen (1962)

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 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.

2.4 Bandwidth Selection

It is natural to think of finding an optimal bandwidth by minimizing $MSE_f[\hat{f}]$ in $h > 0$. Härdle(1990) shows that the asymptotic approximation, say, h_* for the optimal bandwidth is

$$h_* = \Big(\frac{f(x)\|K\|_2^2}{(f''(x))^2(\int y^2K(y)dy)^2 n}\Big)^{1/5} \propto n^{-1/5}. \tag{2.4}$$

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:

$$\begin{aligned} IMSE[\hat{f}] &= \int MSE_f[\hat{f}(x)]dx \\ &= \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}(\int y^2K(y)dy)^2\|f''\|_2^2 + o(\frac{1}{nh}) + o(h^4). \end{aligned} \tag{2.5}$$

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,

$$h_{opt} = c_2^{-2/5}\Big(\int K^2(x)dx\Big)^{1/5}\Big(\|f''\|_2^2\Big)^{-1/5} n^{-1/5}, \tag{2.6}$$

where $c_2 = \int y^2K(y)dy$. Unfortunately, (2.6) still depends on the second derivative of f, which measures the speed of fluctuations in the density of f.

2.4.1 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 expression (2.6). For example, assume that a density f belongs to the Gaussian family with mean μ and variance σ^2, then

$$\begin{aligned} \int (f''(x))^2dx &= \sigma^{-5}\int(\varphi''(x))^2dx \\ &= \frac{3}{8}\pi^{-1}2\,\sigma^{-5} \approx 0.212\,\sigma^{-5}, \end{aligned} \tag{2.7}$$

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

$$\begin{aligned} h_{opt} &= (4\pi)^{-1/10}(\frac{3}{8}\pi^{-1/2})^{-1/5}\,\sigma\, n^{-1/5} \\ &= \Big(\frac{4}{3}\Big)^{1/5}\sigma\, n^{-1/5} = 1.06\,\sigma\, n^{-1/5} \end{aligned} \tag{2.8}$$

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

$$\begin{aligned} \hat{h}_{opt} &= 1.06 \min(\hat{\sigma}, \frac{\hat{R}}{(\Phi(3/4) - \Phi(1/4))})\, n^{-1/5} \\ &= 1.06 \min(\hat{\sigma}, \frac{\hat{R}}{1.349})\, n^{-1/5}, \end{aligned} \tag{2.9}$$

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.

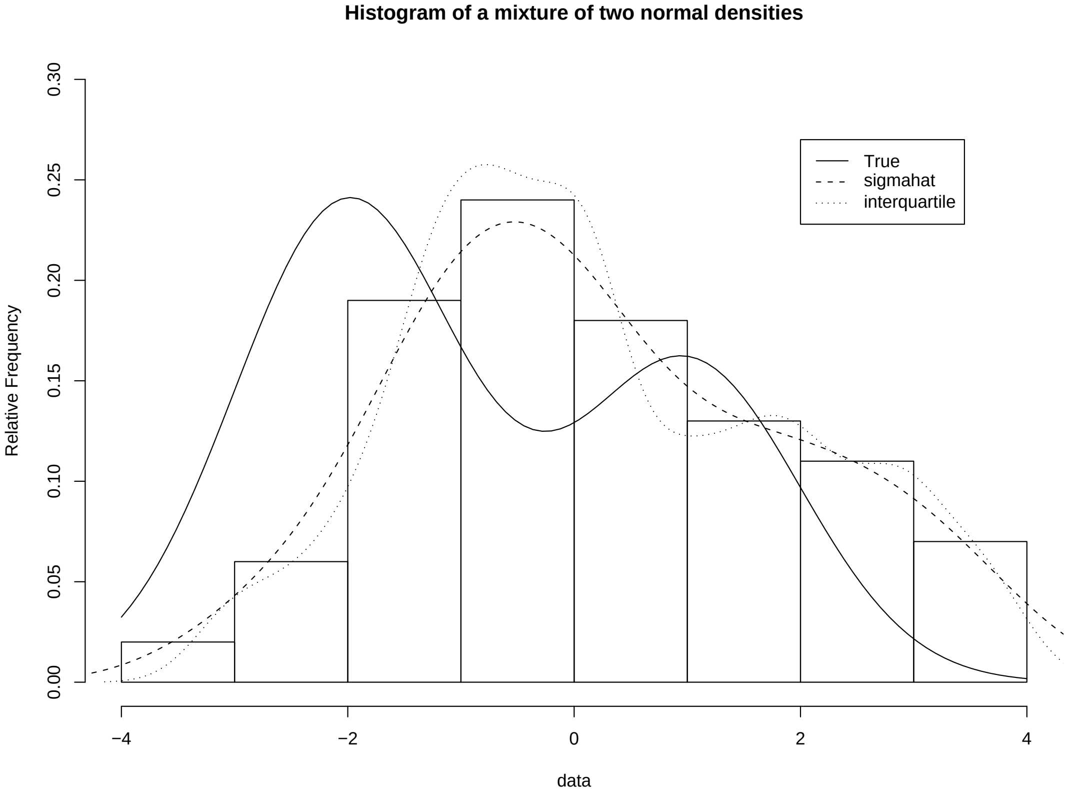

Figure 2.5: Comparison of two bandwidths, $\hat{\sigma}$ (the sample standard deviation) and $\hat{R}$ (the sample interquartile) for the mixture $0.7 \times N(-2,1) + 0.3 \times N(1,1)$.

2.4.2 Maximum likelihood Cross-Validation

Consider kernel density estimates $\hat{f}$ and suppose we want to test for a specific h the hypothesis

$$\hat{f}(x) = f(x) \quad vs. \quad \hat{f}(x) \neq f(x),$$

for a fixed x The likelihood ratio test would be based on the test statistic $f(x)/\hat{f}(x)$. For a good bandwidth this statistic should thus be close to 1. Alternatively, we would expect $\mathbb{E}[\log(\frac{f(X)}{\hat{f}(X)})]$ to be close to 0. Thus, a good bandwidth, which is minimizing this measure of accuracy, is in effect optimizing the *Kullback-Leibler* distance:

$$d_{KL}(f,\hat{f}) = \int \log\Big(\frac{f(x)}{\hat{f}(x)}\Big) f(x) dx. \tag{2.10}$$

Of course, we are not able to compute $d_{KL}(f,\hat{f})$ 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 $d_{KL}(f,\hat{f})$ is close to 0 this would give the best agreement with the hypothesis $\hat{f} = f$. Hence, we are looking for a bandwidth h, which minimizes $d_{KL}(f,\hat{f})$.

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 $\hat{f}$ in the likelihood equation we have $\prod_i \hat{f}(X_i)$ 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 $d_{KL}(f,\hat{f})$. Usually, we do not have additional observations. A way out of this dilemma is to base the estimate $\hat{f}$ on the subset $\{X_j\}_{j\neq i}$, and to calculate the likelihood for X_i. Denoting the *leave-one-out estimate*

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

Hence,

$$\prod_{i=1}^{n} \hat{f}(X_i) = (n-1)^{-n} h^{-n} \prod_{i=1}^{n} \sum_{j\neq i} K(\frac{X_i - X_j}{h}). \tag{2.11}$$

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

$$\begin{aligned} CV_{KL}(h) &= \frac{1}{n}\sum_{i=1}^{n} \log[f_{h,i}(X_i)] \\ &= \frac{1}{n}\sum_{i=1}^{n} \log\Big[\sum_{j\neq i} K(\frac{X_i - X_j}{h})\Big] - \log[(n-1)h] \qquad (2.12) \end{aligned}$$

Naturally, we choose h_{KL} such that:

$$h_{KL} = \arg\max_h CV_{KL}(h). \tag{2.13}$$

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

$$\mathbb{E}[CV_{KL}(h)] = \mathbb{E}[\log \hat{f}_i(X_i)].$$

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

$$\begin{aligned} \mathbb{E}[CV_{KL}(h)] &\approx \mathbb{E}\Big[\int \log \hat{f}(x) f(x) dx\Big] \\ &\approx -\mathbb{E}[d_{KL}(f,\hat{f})] + \int \log[f(x)] f(x) dx. \qquad (2.14) \end{aligned}$$

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 $d_{KL}(f, \hat{f})$.

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 $\hat{f}_i(X_i)$ 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.

2.4.3 Least-Squares Cross-Validation

Consider an alternative distance between f_h and f. The integrated squared error (ISE)

$$\begin{aligned} d_{ISE}(h) &= \int (f_h - f)^2(x)dx \\ &= \int f_h^2(x)dx - 2\int (f_h f)(x)dx + \int f^2(x)dx \\ d_{ISE}(h) - \int f^2(x)dx &= \int f_h^2(x)dx - 2\int (f_h f)(x)dx \qquad (2.15) \end{aligned}$$

For the last term, observe that $\int (f_h f)(x)dx = \mathbb{E}[f_h(X_i)]$ where the expectation is understood to be computed with respect to an additional and independent observation X. For estimation of this term define the leave-one-out estimate

$$\mathbb{E}_X[\hat{f_h}(X)] = \frac{1}{n}\sum_{i=1}^{n} f_{h,i}(X_i) \qquad (2.16)$$

This leads to the Least-squares cross-validation:

$$CV_{LS}(h) = \int f_h^2(x)dx - 2\sum_{i=1}^{n} f_{h,i}(X_i) \qquad (2.17)$$

The bandwidth minimizing this function is,

$$h_{LS} = \arg\min_h CV_{LS}(h).$$

This cross-validation function is called an *unbiased cross-validation* criterion, since,

$$\begin{aligned} \mathbb{E}[CV_{LS}(h)] &= \mathbb{E}[d_{ISE}(h) + 2(\mathbb{E}_X[f_h(X)] - \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(X_i)]) - \|f\|_2^2 \\ &= IMSE[f_h] - \|f\|_2^2. \qquad (2.18) \end{aligned}$$

An interesting question is, how good is the approximation of d_{ISE} by CV_{LS}. To investigate this define a sequence of bandwidths $h_n = h(X_1, \ldots, X_n)$ to be asymptotically optimal, if

$$\frac{d_{ISE}(h_n)}{\inf_{h>0} d_{ISE}(h)} \longrightarrow 1, \quad a.s. \quad \text{when} \quad n \longrightarrow \infty.$$

It can be shown that if the density f is bounded then h_{LS} is asymptotically optimal. Similarly to maximum likelihood cross-validation one can found in Härdle (1990) an algorithm to compute the least-squares cross-validation.

2.5 Orthogonal series estimators

Orthogonal series estimators approach the density estimation problem from a quite different point of view. While kernel estimators is close related to statistical thinking orthogonal series relies on the ideas of approximation theory. Without loss of generality let us assume that we are trying to estimate a density f on the interval $[0,1]$. The idea is to use the theory of orthogonal series method and then to reduce the estimation procedure by estimating the coefficients of its Fourier expansion. Define the sequence $\phi_v(x)$ by

$$\begin{cases} \phi_0(x) = 1 \\ \phi_{2r-1}(x) = \sqrt{2}\cos 2\pi r x \quad r = 1,2,\ldots \\ \phi_{2r}(x) = \sqrt{2}\sin 2\pi r x \quad r = 1,2,\ldots \end{cases}$$

It is well known that f can be represented as Fourier series $\sum_{i=0}^{\infty} a_i\phi_i$, where, for each $i \geq 0$,

$$a_i = \int f(x)\phi_i(x)dx. \tag{2.19}$$

Now, suppose that X is a random variable with density f. Then (2.19) can be written

$$a_i = \mathbb{E}\phi_i(X)$$

and so an unbiased estimator of f based on $X_1,\ldots,X_n$ is

$$\hat{a}_i = \frac{1}{n}\sum_{j=1}^{n}\phi_i(X_j).$$

Note that the $\sum_{i=1}^{\infty}\hat{a}_i\phi_i$ converges to a sum of delta functions at the observations, since

$$\omega(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - X_i) \tag{2.20}$$

where δ is the Dirac delta function. Then for each i,

$$\hat{a}_i = \int_0^1 \omega(x)\phi_i(x)dx$$

and hence the $\hat{a}_i$ are exactly the Fourier coefficients of the function ω. The easiest to way to smooth ω is to truncate the expansion $\sum \hat{a}_i\phi_i$ at some point. That is, choose K and define a density estimate $\hat{f}$ by

$$\hat{f}(x) = \sum_{i=1}^{K}\hat{a}_i\phi_i(x). \tag{2.21}$$

Note that the amount of smoothing is determined by K. Small value of K implies in over-smoothing, large value of K under-smoothing.

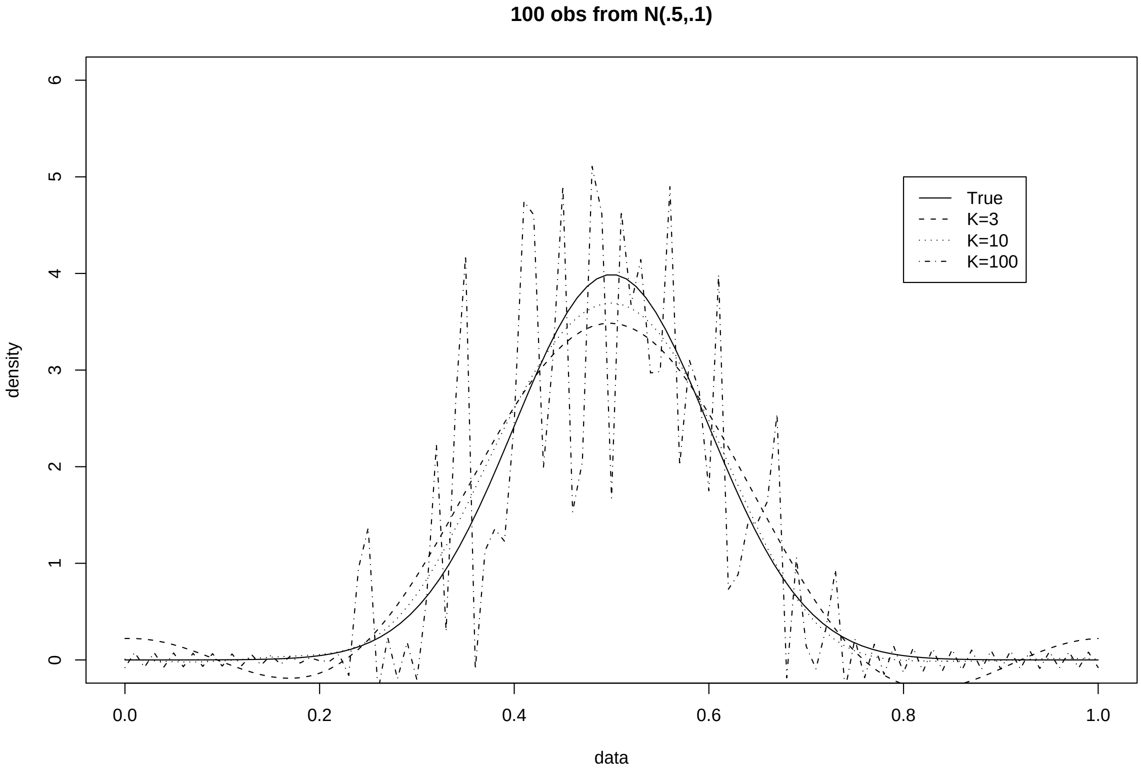

Figure 2.6: Effect of the smoothing parameter K on the orthogonal series method for density estimation

A more general approach would be, choose a sequence of weights λ_i, such that, $\lambda_i \to 0$ as $i \to \infty$. Then

$$\hat{f}(x) = \sum_{i=0}^{\infty} \lambda_i \hat{a}_i \phi_i(x).$$

The rate at which the weights λ_i converge to zero will determine the amount of smoothing. For non finite interval we can have weight functions $a(x) = e^{x^2/2}$ and orthogonal functions $\phi(x)$ proportional to Hermite polynomials.

The data in figure 2.6 were provided to me by Francisco Cribari-Neto and consists of the variation rate of ICMS (imposto sobre circulação de mercadorias e serviços) tax for the city of Brasilia, D.F., from August 1994 to July 1999.

Chapter 3

Kernel nonparametric Regression Method

Suppose we have i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^n$ and the nonparametric regression model given in equation (1.1). By 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.

$$g_h(x) = \frac{\sum_{i=1}^{n} K_h(x - X_i)Y_i}{\sum_{j=1}^{n} K_h(x - X_j)} \tag{3.1}$$

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.

Now, consider the model (1.1) and let $X_1, \ldots, X_n$ be i.i.d. random variables with density f_X such that X_i is independent of ε_i for all $i = 1, \ldots, n$. Assume the conditions given in Section 2.3 and suppose that f and g are twice continuously differentiable in neighborhood of the point x. Then, if $h \longrightarrow 0$ and $nh \longrightarrow \infty$ as $n \longrightarrow \infty$, we have $\hat{g}_h \longrightarrow g$ in probability. Moreover, suppose $\mathbb{E}[|\varepsilon_i|^{2+\delta}]$ and $\int |K(x|^{2+\delta}dx < \infty$, for some $\delta > 0$, then $\sqrt{nh}(\hat{g}_h - \mathbb{E}[\hat{g}_h]) \longrightarrow N(0, (f_X(x))^{-1}\sigma^2 \int (K(x))^2 dx)$ in distribution, where $N(\cdot, \cdot)$ stands for a Gaussian distribution, (see details in Pagan and Ullah (1999)).

As an example, figure 3.1 shows the effect of choosing h on the Nadaraya-Watson procedure. The data consist of the speed of cars and the distances taken to stop. It is important to notice that the data were recorded in the 1920s. (These datasets can be found in the software R) The Nadaraya-Watson kernel method can be extended to the multivariate regression problem by considering the multidimensional kernel density estimation method (see details in Scott (1992)).

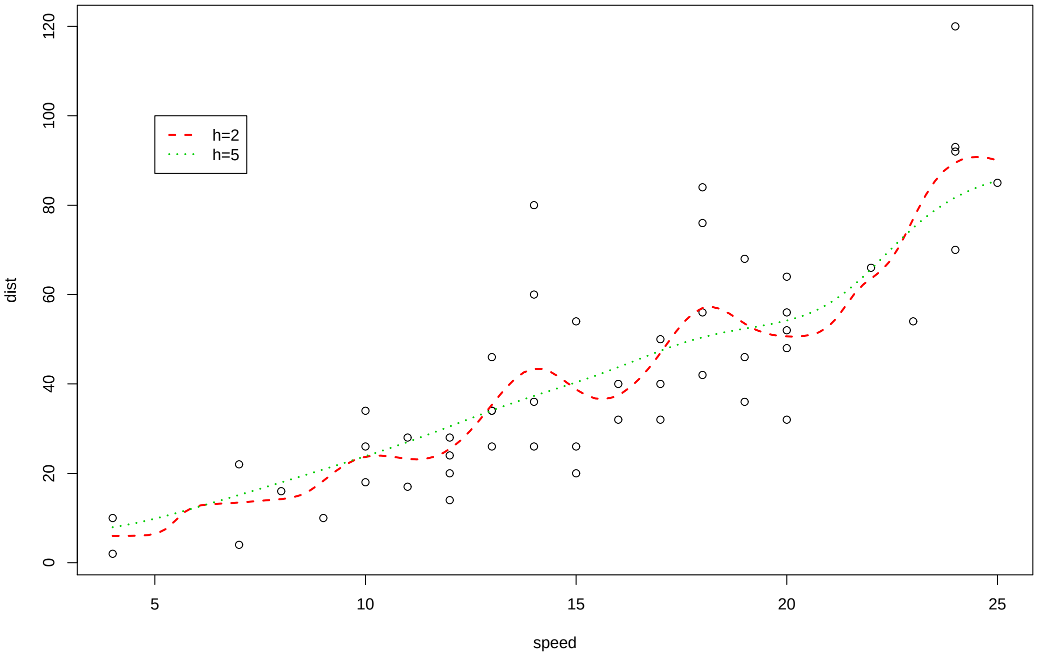

Figure 3.1: Effect of bandwiths on Nadaraya-Watson kernel

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,

$$g_k(x) = \frac{1}{n}\sum_{i=1}^{n} W_{ki}(x) Y_i, \tag{3.2}$$

where,

$$W_{ki}(x) = \begin{cases} n/k & \text{if } i \in J_x \\ 0 & \text{otherwise,} \end{cases} \tag{3.3}$$

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 (3.3) are given by, for a fixed x

$$\mathbb{E}[g_k(x)] - g(x) \approx \frac{1}{24(f(x))^3}[g''(x)f(x) + 2g'(x)f'(x)](k/n)^2 \tag{3.4}$$

and

$$Var[g_k(x)] \approx \frac{\sigma^2}{k}. \tag{3.5}$$

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 3.2 shows the effect of the parameter k on the regression curve estimates. Note that the curve estimate with $k = 2$ is less smoother than the curve estimate with $k = 1$. The data set consist of the revenue passenger miles flown by commercial airlines in the United States for each year from 1937 to 1960 and is available through R package.

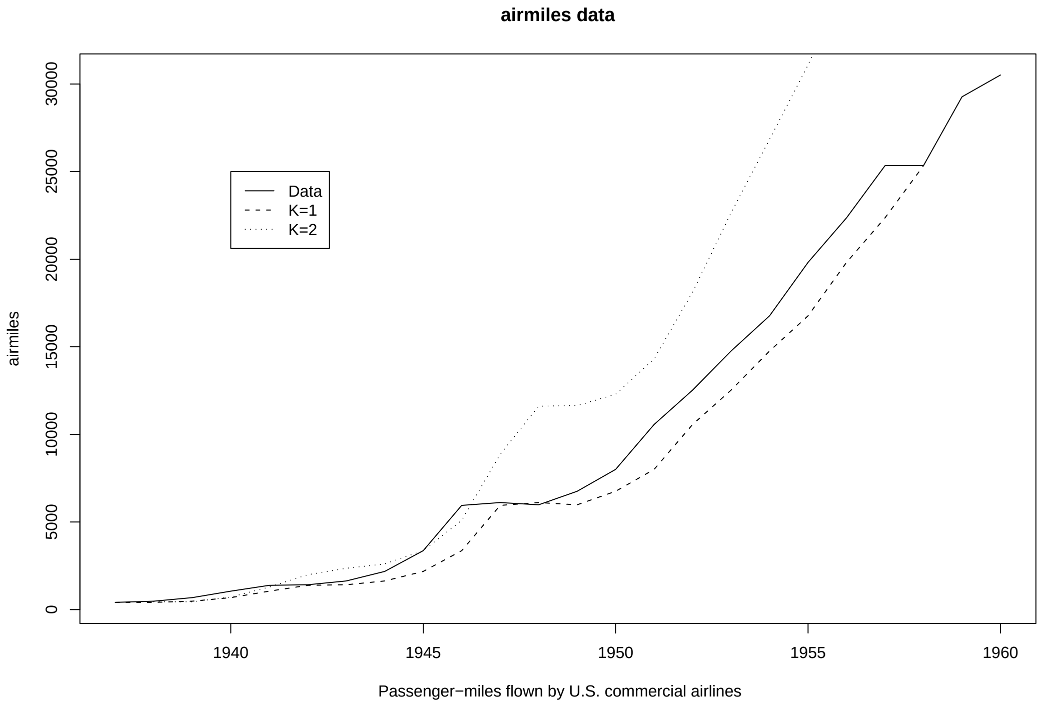

Figure 3.2: Effect of the smoothing parameter k on the k-NN regression estimates.

3.2 Local Polynomial Regression: LOWESS

Cleveland (1979) proposed the algorithm LOWESS, locally weighted scatter plot smoothing, as an outlier 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

$$n^{-1} \sum_{i=1}^{n} W_{ki}\Big(y_i - \sum_{j=0}^{p} \beta_j x^j\Big)^2, \tag{3.6}$$

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 (3.6) 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. In addition, the smoothing parameter can be determined by cross-validation as in (2.13). Note that when using the R function lowess or loess, f acts as the smoothing parameter. Its relation to the k-NN nearest neighbor is given by

$$k = \lceil n \times f \rceil, \quad f \in (0, 1),$$

where n is the sample size.

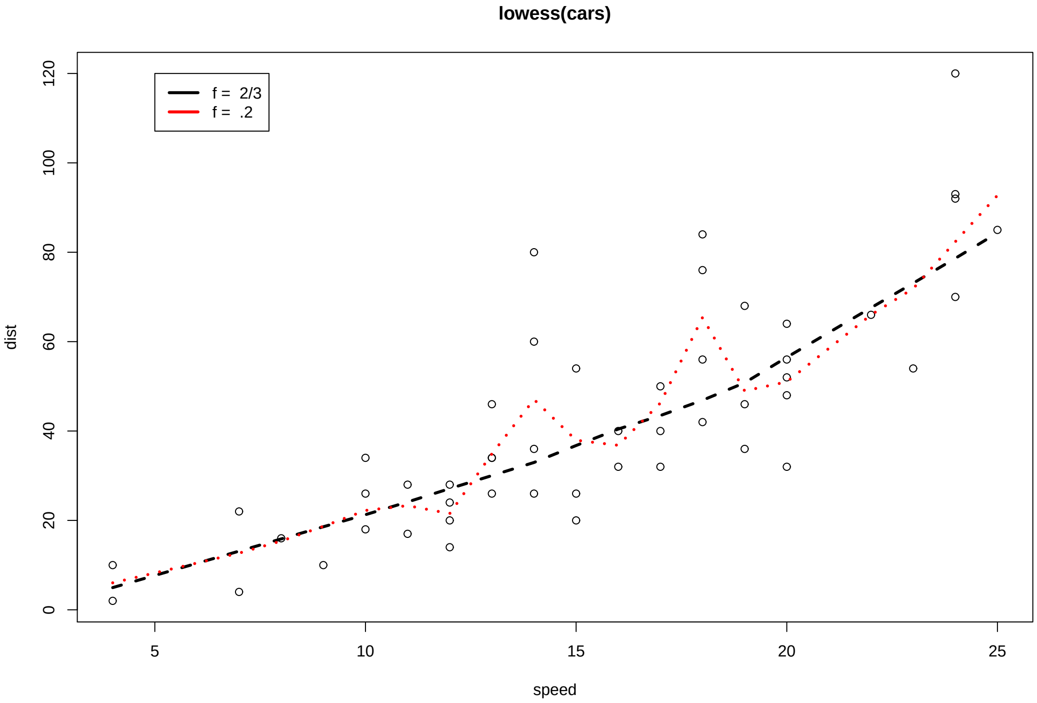

Figure 3.3: Effect of the smoothing parameter using LOWESS method.

3.3 Penalized Maximum Likelihood Estimation

The method of penalized maximum likelihood in the context of density estimation consist of estimating a density f by minimizing a penalized likelihood score $\mathcal{L}(f) + \lambda J(f)$, where $\mathcal{L}(f)$ is a goodness-of-fit measure, and $J(f)$ is a roughness penalty. This section is developed considering historical results, beginning with Good and Gaskins (1971), and ending with the most recent result given by Gu (1993).

The maximum likelihood (M.L.) method has been used as statistical standard procedure in the case where the underlying density f is known except by a finite number of parameters. It is well known the M.L. has optimal properties (asymptotically unbiased and asymptotically normal distributed) to estimate the unknown parameters. Thus, it would be interesting if such standard technique could be applied on a more general scheme where there is no assumption on the form of the underlying density by assuming f to belong to a pre-specified family of density functions.

Let $X_1, \ldots, X_n$ be i.i.d. random variables with unknown density f. The likelihood function is given by:

$$\mathcal{L}(f|X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i).$$

The problem with this approach can be described by the following example. Recall $\hat{f}_h(x)$ a kernel estimate, that is,

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

with $h' = h/c$, where c is constant greater than 0, i.e., for the moment the bandwidth is h/c. Let h be small enough such that $|\frac{X_i - X_i'}{h/c}| > M > 0$, and assume K has been chosen so that $K(u) = 0$, if $|u| > M$. Then,

$$\hat{f}_{h'}(X_i) = \frac{c}{nh} K(0).$$

If $c > \frac{1}{K(0)}$ then $\hat{f}_{h'}(X_i) > \frac{1}{nh}$. For fixed n, we can do this for all X_i simultaneously. Thus,

$$\mathcal{L} \geq (\frac{1}{nh})^n.$$

Letting $h \to 0$, we have $\mathcal{L} \to \infty$. That is, $\mathcal{L}(f|X_1, \ldots, X_n)$ does not have a finite maximum over the class of all densities. Hence, the likelihood function can be as large as one wants it just by taking densities with the smoothing parameter approaching zero. Densities having this characteristic, e.g., bandwidth $h \to 0$, approximate to delta functions and the likelihood function ends up to be a sum of spikes delta functions. Therefore, without putting constraints on the class of all densities, the maximum likelihood procedure cannot be used properly.

One possible way to overcome the problem described above is to consider a penalized log-likelihood function. The idea is to introduce a penalty term on the log-likelihood function such that this penalty term quantifies the smoothness of $g = \log f$.

Let us take, for instance, the functional $J(g) = \int (g'')^2$ as a penalty term. Then define the *penalized log-likelihood function* by

$$\mathcal{L}_\lambda(g) = \frac{1}{n} \sum_{i=1}^{n} g(X_i) - \lambda J(g) \,, \tag{3.7}$$

where λ is the smoothing parameter which controls two conflicting goals, the fidelity to the data given by $\sum_{i=1}^{n} g(X_i)$ and the smoothness, given by the penalty term $J(g)$.

The pioneer work on penalized log-likelihood method is due to Good and Gaskins (1971), who suggested a Bayesian scheme with penalized log-likelihood (using their notation) becomes:

$$\omega = \omega(f) = \mathcal{L}(f) - \Phi(f) \,,$$

where $\mathcal{L} = \sum_{i=1}^{n} g(X_i)$ and Φ is the smoothness penalty.

In order to simplify the notation, let $\int h$ have the same meaning as $\int_{-\infty}^{\infty} h(x)dx$. Now, consider the number of bumps in the density as the measure of roughness or smoothness. The first approach was to take the penalty term proportional to Fisher's information, that is,

$$\Phi(f) = \int (f')^2/f.$$

Now by setting $f = \gamma^2$, $\Phi(f)$ becomes $\int(\gamma')^2$, and then replace f by γ in the penalized likelihood equation. Doing that the constraint $f \geq 0$ is eliminated and the other constraint, $\int f = 1$, turns out to be equivalent to $\int \gamma^2 = 1$, with $\gamma \in L^2(-\infty, \infty)$.

Good and Gaskins(1971) verified that when the penalty $4\alpha \int(\gamma')^2$ yielded density curves having portions that looked "too straight". This fact can be explained noting that the curvature depends also on the second derivatives. Thus $(\gamma'')^2$ should be included on the penalty term. The final roughness functional proposed was:

$$\Phi(f) = 4\alpha \int (\gamma')^2 + \beta \int (\gamma'')^2 ,$$

with α, β satisfying,

$$2\alpha\sigma^2 + \frac{3}{4}\beta = \sigma^4, \tag{3.8}$$

where σ^2 is either an initially guessed value of the variance or it can be estimated the sample variance based on the data. According to Good and Gaskins (1971), the basis for this constraint is the feeling that the class of normal distributions form the smoothest class of distributions, the improper uniform distribution being limiting form. Moreover, they pointed out that some justification for this feeling is that a normal distribution is the distribution of maximum entropy for a given mean and variance. The integral $\int(\gamma')^2$ is also minimized for a given variance when f is normal (Good and Gaskins, 1971). They thought was reasonable to give the normal distribution special consideration and decided to choose α, β such that $\omega(\alpha, \beta; f)$ is maximized by taking the mean equal to $\bar{x}$ and variance as $\sum_{i=1}^{N}(x_i - \bar{x})^2/N - 1$. That is, if $f(x) \sim \mathcal{N}(\mu, \sigma^2)$ then $\int(\gamma')^2 = \frac{1}{4\sigma^2}$, $\int(\gamma'')^2 = \frac{3}{16\sigma^4}$ and hence we have,

$$\omega(\alpha, \beta; f) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu)^2 - \frac{\alpha}{\sigma^2} - \frac{3\beta}{16\sigma^4}.$$

The score function $\omega(\alpha, \beta; f)$ is maximized when $\mu = \bar{x}$ and σ is such that,

$$-N + \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{\sigma^2} + \frac{2\alpha}{\sigma^2} + \frac{3\beta}{4\sigma^4} = 0. \tag{3.9}$$

If we put $\sigma^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2/N - 1$, the equation (3.9) becomes,

$$\sigma^4(N-1) + 2\alpha\sigma^2 + \frac{3\beta}{4} = \sigma^4 N,$$

and so we have the constraint (3.8).

Pursuing the idea of Good and Gaskins, Silverman (1982) proposed a similar method where the log density is estimated instead of the density itself. An advantage of Silverman's approach is that using the logarithm of the density and the augmented Penalized likelihood functional, any density estimates obtained will automatically be positive and integrate to one. Specifically, Let $(m_1, \ldots, m_k)$ be a sequence of natural numbers so that $1 \leq \sum_{i=1}^{k} m_i \leq m$, where $m > 0$ is such that $g^{(m-1)}$ exists and is continuous. Define a linear differential operator D as:

$$D(g) = \sum c(m_1, \ldots, m_k)(\frac{\partial}{\partial x_1})^{m_1} \ldots (\frac{\partial}{\partial x_k})^{m_k}(g) .$$

Now assume that at least one of the coefficients $c(m_1, \ldots, m_k) \neq 0$ for $\sum m_i = m$. Using this linear differential operator define a bilinear functional $\langle \cdot, \cdot \rangle$ by

$$\langle g_1, g_2 \rangle = \int D(g_1)D(g_2) .$$

where the integral is taken over a open set Ω with respect to Lebesgue measure.

Let S be the set of real functions g on Ω for which:

- the $(m-1)$th derivatives of g exist everywhere and are piecewise differentiable,
- $\langle g, g \rangle < \infty$,
- $\int e^g < \infty$.

Given the data $X_1, \ldots, X_n$ i.i.d. with common density f, such that $g = \log f$, $\hat{g}$ is the solution, if it exists, of the optimization problem

$$\max\{\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \frac{\lambda}{2}\langle g, g \rangle\} ,$$

subject to $\int e^g = 1$. And the density estimate $\hat{f} = e^{\hat{g}}$, where the the null space of the penalty term is the set $\{g \in S : \langle g, g \rangle = 0\}$.

Note that the null space of $\langle g, g \rangle$ is an exponential family with at most $(m-1)$ parameters, for example, if $\langle g, g \rangle = \int (g^{(3)})^2$ then $g = \log f$ is in an exponential family with 2 parameters. See Silverman (1982).

Silverman presented an important result which makes the computation of the constrained optimization problem a "relatively" easy computational scheme of finding the minimum of an unconstrained variational problem. Precisely, for any g in a class of smooth functions (see details in Silverman (1982)) and for any fixed positive λ, let

$$\omega_0(g) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \frac{\lambda}{2}\int (g'')^2$$

and

$$\omega(g) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int e^g + \frac{\lambda}{2}\int (g'')^2.$$

Silverman proved that unconstrained minimum of $\omega(g)$ is identical with the constrained minimum of ω_0, if such a minimizer exists.

3.3.1 Computing Penalized Log-Likelihood Density Estimates

Based on Silverman's approach, O'Sullivan(1988) developed an algorithm which is a fully automatic, data driven version of Silverman's estimator. Furthermore, the estimators obtained by O'Sullivan's algorithm are approximated by linear combination of basis functions. Similarly to the estimators given by Good and Gaskins(1971), O'Sullivan proposed that cubic B-splines with knots at data points should be used as the basis functions. A summary of definitions and properties of B-splines were given in the section 4.

The basic idea of computing a density estimate provided by penalized likelihood method is to construct approximations to it. Given $x_1, \dots, x_n$, the realizations of random variables $X_1, \dots, X_n$, with common log density g. We are to solve a finite version of (3.7) which are reasonable approximations to the infinite dimensional problem (Thompson and Tapia, 1990, 121–145). Good and Gaskins (1971) based their computational scheme on the fact that since $\gamma \in L^2(-\infty, \infty)$ then for a given orthonormal system of functions $\{\phi_n\}$,

$$\sum_{n=0}^{\infty} a_n \phi_n \xrightarrow{m.s.} g \in L^2 \,,$$

with $\sum_{n=0}^{\infty} |a_n| < \infty$ and $\{a_n\} \in \mathbb{R}$. That is, γ in L^2 can be arbitrarily approximated by a linear combination of basis functions. In their paper, Hermite polynomials were used as basis functions. Specifically:

$$\phi_n(x) = e^{-x^2/2} H_n(x) 2^{-n/2} \pi^{-1/4} (n!)^{1/2} \,,$$

where,

$$H_n(x) = (-1)^n e^{x^2} \left(\frac{d^n}{dx^n} e^{-x^2}\right) .$$

The *log density* estimator proposed by O'Sullivan (1988) is defined as the minimizer of

$$-\frac{1}{n} \sum_{i=1}^{n} g(x_i) + \int_a^b e^{g(s)} ds + \lambda \int_a^b (g^{(m)})^2 ds, \tag{3.10}$$

for fixed $\lambda > 0$, and data points $x_1, \dots, x_n$. The minimization is over a class of absolutely continuous functions on $[a, b]$ whose mth derivative is square integrable.

Computational advantages of this log density estimators using approximations by cubic B-splines are:

- It is a fully automatic procedure for selecting an appropriate value of the smoothing parameter λ, based on the AIC type criteria.
- The banded structures induced by B-splines leads to an algorithm where the computational cost is linear in the number of observations (data points).
- It provides approximate pointwise Bayesian confidence intervals for the estimator.

A disadvantage of O'Sullivan's work is that it does not provide any comparison of performance with other available techniques.

We see that the previous computational framework is unidimensional, although Silverman's approach can be extended to higher dimensions.

Chapter 4

Spline Functions

4.1 Acquiring the Taste

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 $[x_0, x_1], \dots, [x_k, x_{k+1}]$ and then uses a low degree polynomial p_i for approximation over each interval $[x_i, x_{i+1}]$, $i = 0, \dots, k$. This procedure produces a piecewise polynomial approximating function $s(\cdot)$;

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

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 $x_1, \dots, x_k$, 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 4.1 *The function $s(x)$ is called a spline function (or simply "spline") of degree r with knots at $\{x_i\}_{i=1}^{k}$ if $-\infty =: x_0 < x_1 < \dots < x_k < x_{k+1} := \infty$, where $-\infty =: x_0$ and $x_{k+1} := \infty$ are set by definition,*

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

The set $\mathcal{S}_r(x_1, \dots, x_k)$ of spline functions is called *spline space*. Moreover, the spline space is a linear space with dimension $r + k + 1$ (Schumaker (1981)).

Definition 4.2 *For a given point $x \in (a, b)$ the function*

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

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 $\{x_1, \dots, x_k\}$ and the basis functions $\{1, t, t^2, \dots, t^r, (t - x_1)_+^r, \dots, (t - x_k)_+^r\}$. 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 Schumaker (1981). Also, B-splines have an important computational property, 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 the aid of a recurrence relation is possible.

Definition 4.3 *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*

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

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

From the Definition 4.3 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. That is, of all the B-splines of order k for the knot sequence Ω_∞, only the k B-splines $B_{j-k+1}^k, B_{j-k+2}^k, \dots, B_j^k$ 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 4.1 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 $x_1, \dots, x_k$, if, in addition to the properties implied by definition (4.1), 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 $a := x_0 < x_1 < \dots < x_k < x_{k+1} := b$. Then, $\mathcal{NS}_{2m} = \{s \in \mathcal{S}(\mathcal{P}_{2m}) : s_0 = s|_{[a,x_1)} \text{ and } s_k = s|_{[x_k,b)} \in \mathcal{P}_m\}$, is the *natural polynomial spline space of order 2m with knots at* $x_1, \dots, x_k$. The name "natural spline" stems from the fact that, as a result of this extra condition, s satisfies the so called natural boundary conditions $s^j(a) = s^j(b) = 0, j = m, \dots, 2m - 1$.

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.

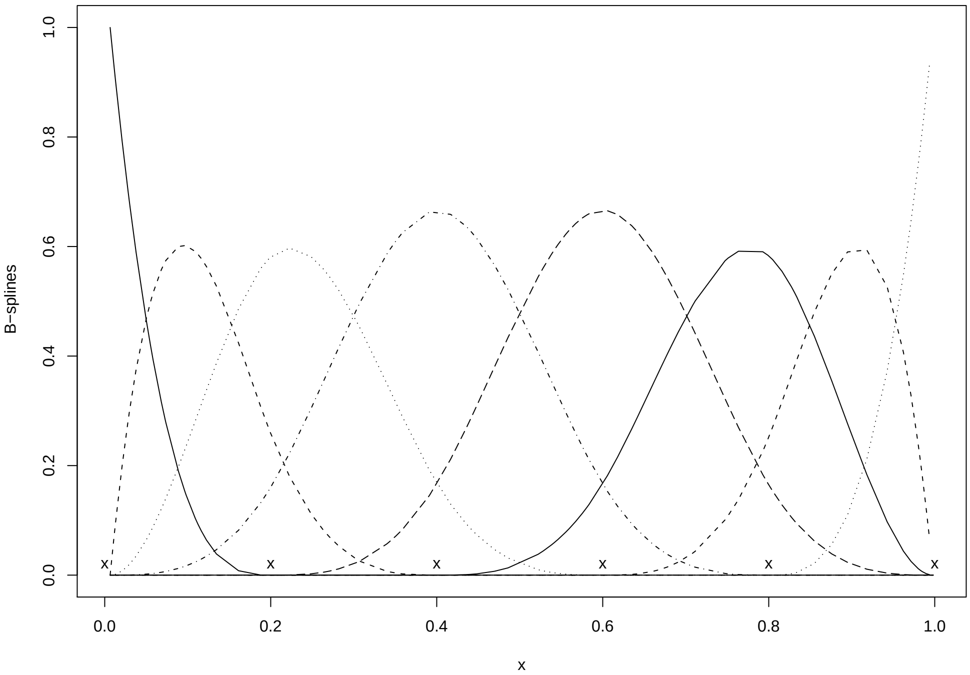

Figure 4.1: Basis Functions with 6 knots placed at "x"

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*, $\mathcal{NS}_4$. Figure 4.2 show an example of the natural splines basis.

4.2 Logspline Density Estimation

Kooperberg and Stone (1991) 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 $[t_K, \infty)$.

Let Θ be the parametric space of dimension $p = K - 1$, such that $\theta_1 < 0$ and $\theta_p > 0$

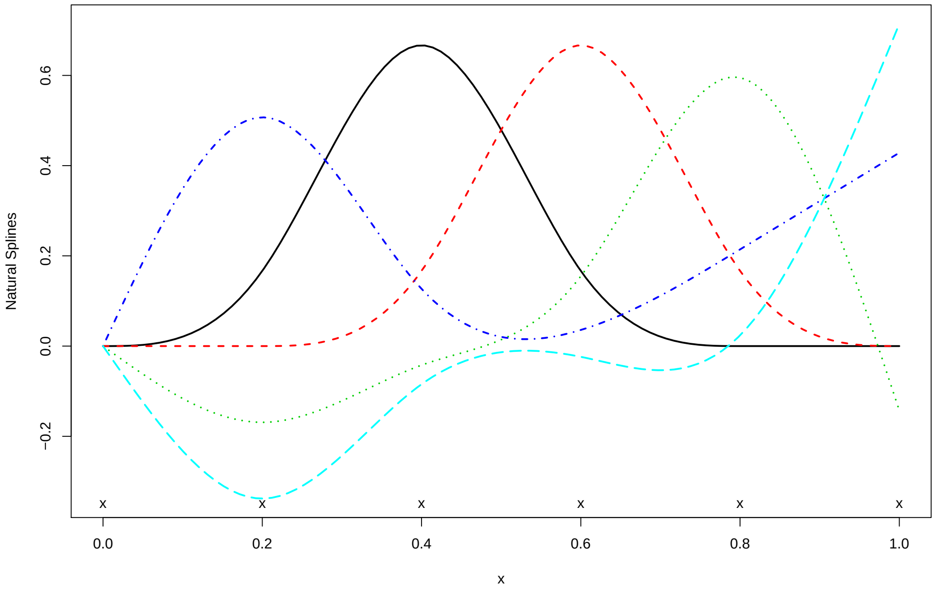

Figure 4.2: Basis Functions with 6 knots placed at "x"

for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$. 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 (1993) algorithm for smoothing spline density estimation.

In the Logspline approach the number of knots is the smoothing parameter. That is, too many knots lead to a noisy estimate while too few knots give 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), show that

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

where $\boldsymbol{I}(\boldsymbol{\theta})$ 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 non-permanent 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 0.1. Figure 4.3 shows an example of logspline density estimation for a mixture of two normal densities.

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).

4.3 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

$$-\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \log \int_{\mathcal{X}} e^g + \frac{\lambda}{2} J(g) \tag{4.1}$$

in a Hilbert space $\mathcal{H}$, where J is a roughness penalty and λ is the smoothing parameter. The space $\mathcal{H}$ is such that the evaluation is continuous so that the first term in (4.1) 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

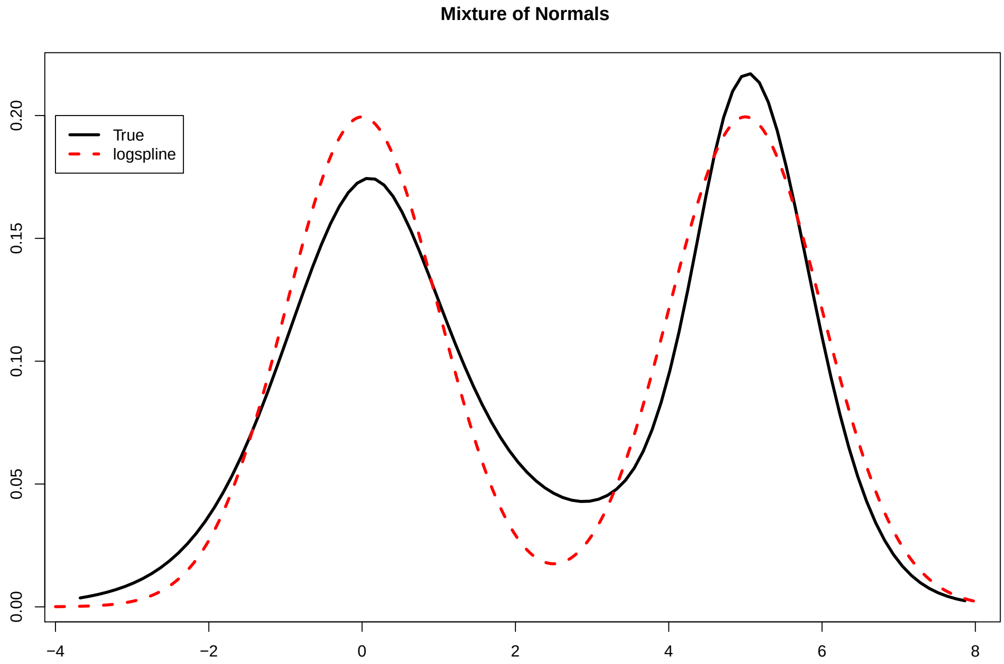

Figure 4.3: logspline density estimation for .5*N(0,1)+.5*N(5,1)

variational problem (4.1). Since, $\mathcal{H}$ is an infinite dimensional space, the minimizer of (4.1) is, in general, not computable. Thus, Gu and Qiu (1993) 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 (4.1) 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.

Figure 4.4 exhibits the performance of SSDE for Buffalo Snow data. (This data set can be found in R.)

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 (4.1) significantly? How good would such an approximation be?

Dias (1998) provided some answers to those questions by using the basis functions $B_i(x)$ given in Definition (4.3) that can be easily extend to a multivariate case by a tensor product.

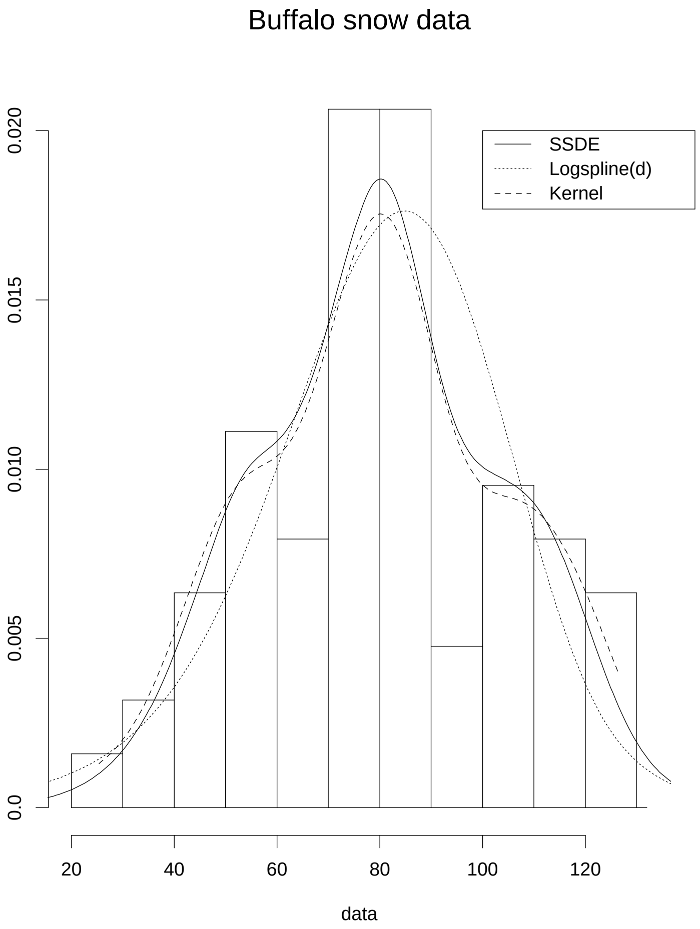

Figure 4.4: Histogram, SSDE, Kernel and Logspline density estimates

Chapter 5

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:

$$y_i = \mathcal{L}_i g + \epsilon_i \tag{5.1}$$

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 and normally 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 (5.1) becomes,

$$y_i = g(x_1(i), \ldots, x_d(i)) + \epsilon_i. \tag{5.2}$$

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

$$L_\lambda(g) = \frac{1}{n} \sum_{i=1}^{n} (y_i - g(t_i))^2 + \lambda J_m^d(g) \tag{5.3}$$

where λ is the smoothing parameter which controls the trade off between fidelity to the data and smoothness with penalty term J_m^d. Note that, when λ is large a premium is being placed on smoothness and functions with large mth derivatives are penalized. In fact, $\lambda \longrightarrow \infty$ gives an mth order polynomial regression fit to the data. Conversely, for small values of λ more emphasis is put on goodness-of-fit and the limit case of $\lambda \longrightarrow 0$, we have interpolation. In general, in smoothing spline nonparametric regression 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 $j = 1, \ldots, m$.

Wahba (1990) has shown that, if $t_1, \ldots, t_n$ are such that least squares regression on $\phi_1, \ldots, \phi_M$ is unique, then (5.3) has a unique minimizer g_λ, with representation

$$\begin{aligned} g_\lambda(t) &= \sum_{i=1}^{n} c_i E_m(t, t_i) + \sum_{j=1}^{M} b_j \phi_j(t) \\ &= Qc + Tb \end{aligned} \tag{5.4}$$

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 Wahba (1990), Silverman and Green (1994). 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 (5.4) into (5.3). Thus, the optimization problem (5.3) 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 (See, Wahba (1990)). Efforts have been done in order to reduce substantially the computational cost of solving smoothing splines fitting by introducing the concept of H-splines (Luo and Wahba (1997) and Dias (1999)), 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 (5.3) 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

$$g_\lambda(s) = \sum_{i=1}^{n} y_i H_\lambda(s, t_i). \tag{5.5}$$

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 $\mathbb{R}$. Assuming the following conditions, (Silverman (1984)),

1. $g \in \mathcal{H}[a, b]$.

2. There exists an absolutely continuous distribution function F on $[a, b]$ such that $F_n \longrightarrow F$ uniformly as $n \longrightarrow \infty$.

3. $f = F'$, $0 < \inf_{[a,b]} f \leq \sup_{[a,b]} f < \infty$.

4. The density has bounded first derivative on $[a, b]$.

5. $a(n) = \sup_{[a,b]} |F_n - F|$, the smoothing parameter λ depends on n in such a way that $\lambda \longrightarrow 0$ and $\lambda^{-1/4} a(n) \longrightarrow 0$ as $n \longrightarrow \infty$.

In particular, one can assume that the design points are regularly distributed with density f; that is, $t_i = F^{-1}((i-1/2)/n)$. Then, $\sup|F_n - F| = (1/2)n^{-1}$ so that $n^4\lambda \longrightarrow \infty$ and $\lambda \longrightarrow 0$ for (5) to hold. Thus, 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.

5.1 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 are also linear in the predictor variables. An additive model is defined by

$$y_i = \alpha + \sum_{j=1}^{p} g_j(t_j) + \epsilon_i \tag{5.6}$$

where t_j are the predictor variables and as defined before in section 5, 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 5 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

$$\mathcal{H} = \mathcal{H}_0 + \sum_{k=1}^{p} \mathcal{H}_k \tag{5.7}$$

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 5. The space $\mathcal{H}_0$ is the space of functions that are not to be penalized in the optimization. For example, recall equation (5.3) 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

$$\sum_{i=1}^{n}\{y_i - \sum_{k=0}^{p} g_k(t_i)\}^2 + \sum_{k=1}^{k} \lambda_k ||g_k||^2_{\mathcal{H}_k}, \tag{5.8}$$

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

$$\hat{g} = \sum_{k=1}^{p} Q_k c + Tb, \tag{5.9}$$

where Q_k and T are given in equation (5.4) and the vectors c and b are found by minimizing the finite dimensional penalized least square criterion

$$||y - Tb - \sum_{k=1}^{p} Q_k c|| + \sum_{k=1}^{p} \lambda_k c_k^T Q_k c^2. \tag{5.10}$$

This general problem (5.9) can potentially be solved by a backfitting type algorithm as in Hastie and Tibshirani (1990).

Algorithm 5.1 *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 (5.6) does not change the resulting estimates.

An example of gam method is given in Figure 5.1

5.2 Generalized Cross-Validation Method for Splines nonparametric Regression

Without loss of generality, take $d = 1$ and $m = 2$. The solution of (5.3) 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

$$V_0(\lambda) = \frac{1}{n} \sum_{k=1}^{n} (y_k - g_\lambda^{[k]}(t_k))^2, \tag{5.11}$$

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 Wahba (1990)) that,

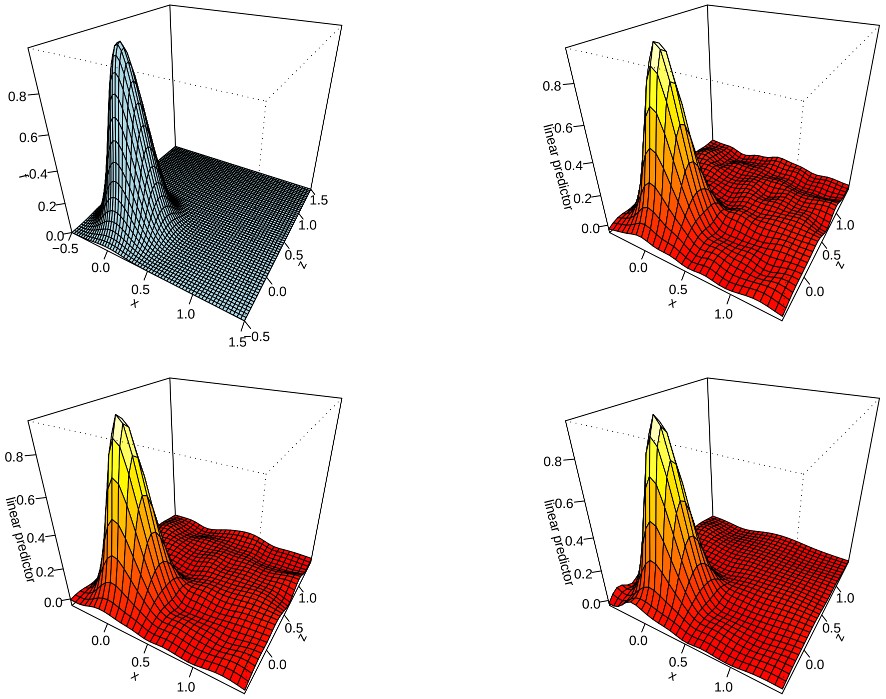

Figure 5.1: True, tensor product, gam non-adaptive and gam adaptive surfaces

for fixed λ we have by (5.5) 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

$$(y_k - g_\lambda^{[k]}(t_k)) = (y_k - g_\lambda(t_k))/(1 - h_{kk}(\lambda), \tag{5.12}$$

where $h_{kk}(\lambda)$ is the kth entry of H_λ. By substituting (5.12) into (5.11) we obtain a simplified form of V_0, that is,

$$V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - g_\lambda(t_k))^2/(1 - h_{kk}(\lambda))^2 \tag{5.13}$$

The right hand of (5.13) is easier to compute than (5.11), 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

$$V(\lambda) = \frac{1}{n}\sum_{k=1}^{n}(y_k - g_\lambda(t_k))^2/(1 - \bar{h}_{kk}(\lambda))^2 = \frac{\frac{1}{n}||(I - H_\lambda)\mathbf{y}||^2}{[\frac{1}{n}tr(I - H_\lambda]^2}, \tag{5.14}$$

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

$$T(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\mathcal{L}_i g_\lambda - \mathcal{L}_i g)^2 \tag{5.15}$$

where, $\mathcal{L}_i$ is the evaluation functional defined in section 4.3, the GCV estimate of λ is the minimizer of (5.15). Consider the expected value of $T(\lambda)$,

$$\mathbb{E}[T(\lambda)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\mathcal{L}_i g_\lambda - \mathcal{L}_i g)^2]. \tag{5.16}$$

The GCV theorem Wahba (1990) says that if g is in a reproducing kernel Hilbert 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 5.2 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()`.

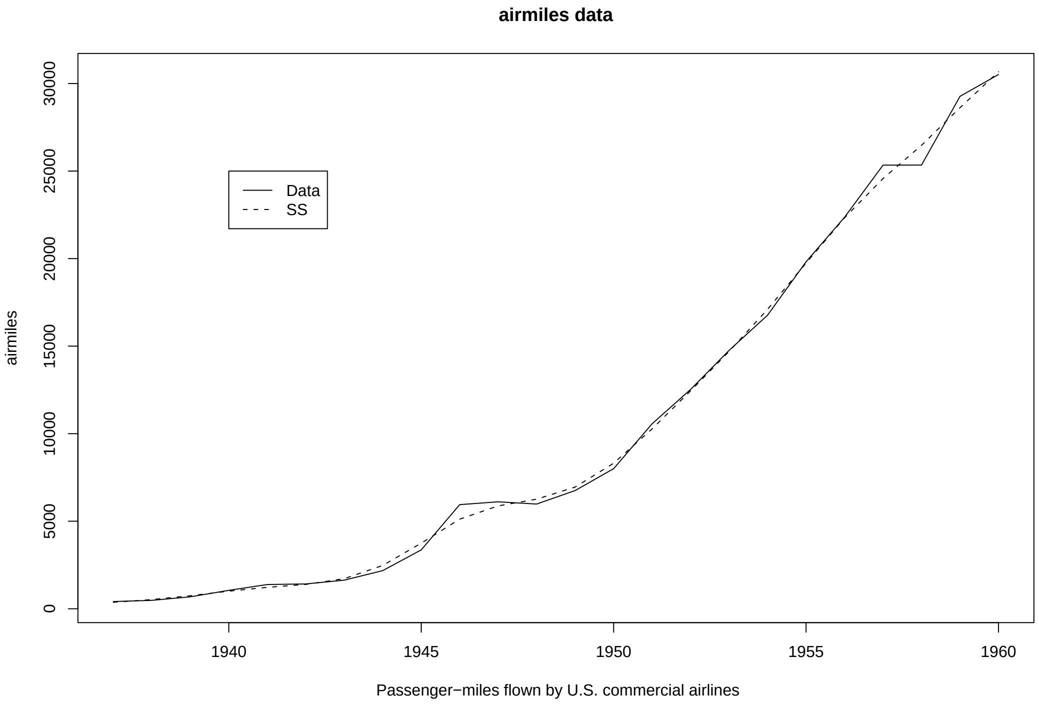

Figure 5.2: Smoothing spline fitting with smoothing parameter obtained by GCV method

Chapter 6

Regression splines, P-splines and H-splines

6.1 Sequentially Adaptive H-splines

In regression splines, the idea is to approximate g by a finite dimensional subspace of $\mathcal{W}$ spanned by basis functions $B_1, \ldots, B_K$, $K \leq n$. That is,

$$g \approx g_K = \sum_{j=1}^{K} c_j B_j,$$

where the parameter K controls the flexibility of the fitting. A very common choice for basis functions is the set of cubic B-splines (de Boor, 1978). The B-splines basis functions provide numerically superior scheme of computation and have the main feature that each B_j has compact support. In practice, it means that we obtain a stable evaluation of the resulting matrix with entries $B_{i,j} = B_j(t_i)$, for $j = 1, \ldots, K$ and $i = 1, \ldots, n$ is banded. Unfortunately, the main difficulty when working with regression splines is to select the number and the positions of a sequence of breakpoints called knots where the piecewise cubic polynomials are tied to enforce continuity and lower order continuous derivatives. (See Schumaker (1972) for details.) Regression splines are attractive because of their computational scheme where standard linear model techniques can be applied. But smoothness of the estimate cannot easily be varied continuously as functions of a single smoothing parameter (Hastie and Tibshirani, 1990). In particular, when $\lambda = 0$ we have the regression spline case, where K is the parameter that controls the flexibility of the fitting. To exemplify the action of K on the estimated curve, let us consider an example by simulation with $y(t) = \exp(-t)\sin(\pi t/2)\cos(\pi t) + \varepsilon$ with $\varepsilon \sim N(0, .05)$. The curve estimates were obtained by least square method with four different numbers of basis functions which are the cubic B-splines. Figure 6.1 shows the effect of varying the number of basis functions on the estimation of the true curve. Note that the number of basis functions is the same as the number of knots since it is assumed that we are dealing with natural cubic splines space. Observe that small values of K make smoother the estimate and hence over smoothing may occur. Large values of K may cause under-smoothing.

In the smoothing techniques the number of basis functions is chosen to be as large as the number of observations and then let the choice of the smoothing parameter

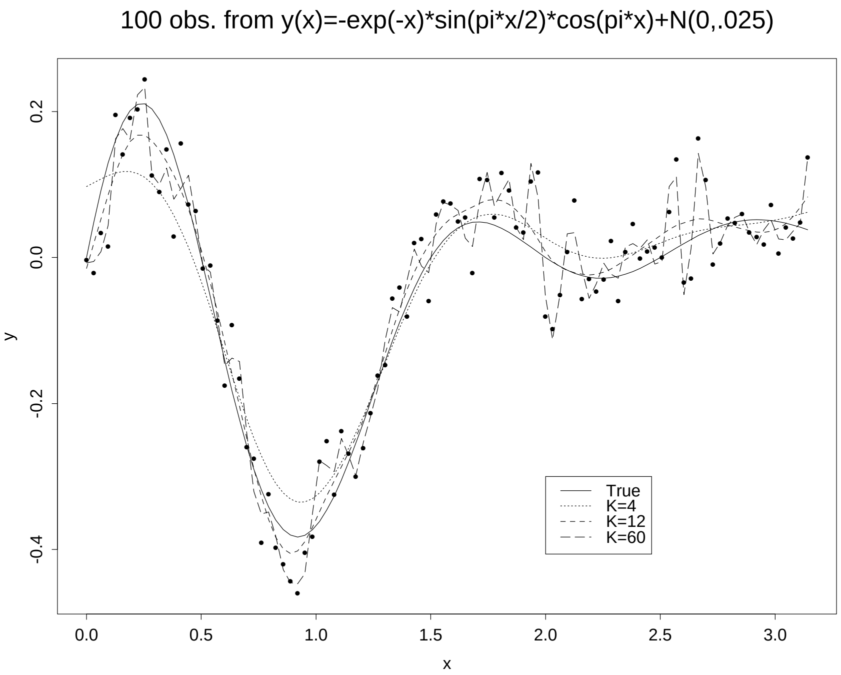

Figure 6.1: Spline least square fittings for different values of K

contolling the smoothing (Bates and Wahba, 1982). Here a different approach is to be taken. The H-splines method introduced by Dias (1994) in the case of nonparametric density estimation, combines ideas from regression splines and smoothing splines methods by finding the number of basis functions and the smoothing parameter iteratively according to a criterion that is described below. With the point evaluation functionals $L_i g = g(t_i)$ the equation (6.4) becomes,

$$A_\lambda(g) = \sum_{i=1}^{n}(y_i - g(t_i))^2 + \lambda \int (g'')^2. \tag{6.1}$$

Assume that $g \approx g_K = \sum_{i=1}^{K} c_i B_i = Xc$ so that $g_K \in \mathcal{H}_K$, where $\mathcal{H}_K$ denotes the space of natural cubic splines (NCS) spanned by the basis functions $\{B_i\}_{i=1}^{K}$ and X is a $n \times K$ matrix with entries $X_{ij} = B_i(t_j)$, for $i = 1, \ldots, K$ and $j = 1, \ldots, n$. Then, the numerical problem is to find a vector $c = (c_1, \ldots, c_K)^T$ that minimizes,

$$A_\lambda^*(c) = \|y - Xc\|_2^2 + \lambda c^T \Omega c,$$

where Ω is $K \times K$ matrix with entries $\Omega_{ij} = \int B_i''(t) B_j''(t) dt$ and $y = (y_1, \ldots, y_n)^T$. Standard calculations (de Boor, 1978) provide c as a solution of the following linear system $(X^T X + \lambda\Omega)c_\lambda = X^T y$. Note that the linear system now involves $K \times K$ matrices instead of using $n \times n$ matrices which is the case of smoothing splines. Both K and λ controls the trade off between smoothness and fidelity to the data.

Figure 6.2 shows, for $\lambda > 0$, an example of the relationship between K and λ. Note that when the number of basis functions increases, the smoothing parameter decreases to a point and then it increases with K. That is, for large values of K, the smoothing parameter λ becomes larger in order to enforce smoothness.

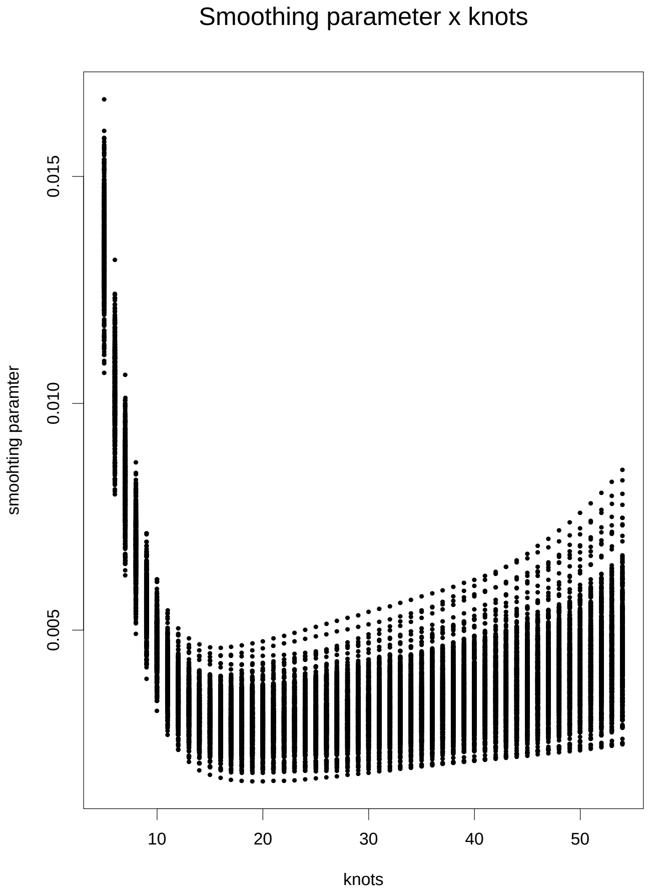

Figure 6.2: Five thousand replicates of $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \epsilon$.

Based on the facts described previously, the idea is to provide a procedure that estimates the smoothing parameter and the number of basis functions iteratively. Consider the following algorithm.

Algorithm 6.1

(1) Let K_0 be the initial number of basis functions and fix λ_0.

(2) Compute c_{λ_0} by solving $(X^TX + \lambda_0\Omega)c_{\lambda_0} = X^Ty$.

(3) Find $\hat{\lambda}$ which minimizes,

$$GCV(\lambda) = \frac{n^{-1}\sum_{i=1}^{n}(y_i - g_{K_0}(t_i))^2}{1 - n^{-1}tr(A(\lambda))},$$

where $A(\lambda) = X(X^TX + \lambda\Omega)^{-1}X^T$.

(4) Compute $g_{K_0,\hat{\lambda}} = A(\hat{\lambda})y$.

(5) Increment the number of basis functions by one and repeat steps (2) to (4) in order to get $g_{K_0+1,\hat{\lambda}}$.

(6) For a real number $\delta > 0$, if a distance $d(g_{K_0,\hat{\lambda}}, g_{K_0+1,\hat{\lambda}}) < \delta$, stop the procedure. The number δ can be determined empirically according to the particular distance $d(\cdot,\cdot)$.

Note that each time the number of basis functions K is incremented by one the numerator of GCV changes and hence this procedure provides an optimal smoothing parameter λ for the estimate g_K based on K basis functions.

The aim is to find a criterion able to tell when to stop increasing the number of basis functions. That is, to find the dimension of the natural cubic spline space where one is looking for the approximation of the solution of (6.4). For this, let us define the following transformation. Given any function in $\mathcal{W}_2^2[a,b]$, take

$$t_g = \frac{g^2}{\int g^2},$$

then $t_g \geq 0$ and $\int t_g = 1$. For any functions $f, g \in \mathcal{W}_2^2[a,b]$, define a pseudo distance closely related to the square of the Hellinger distance,

$$d^2(f,g) = \int \left(\sqrt{t_f} - \sqrt{t_g}\right)^2 = 2(1 - \rho(f,g)),$$

where

$$\rho(f,g) = \int \sqrt{t_f t_g} = \int \sqrt{\frac{f^2 g^2}{\int f^2 \int g^2}} = \int \frac{|fg|}{\sqrt{\int f^2 \int g^2}},$$

is the affinity between f and g. It is not difficult to see that $0 \leq \rho(f,g) \leq 1, \forall f, g \in \mathcal{W}_2^2[a,b]$. Note that $d^2(f,g)$ is minimum when $\rho(f,g) = 1$, i.e., $(\int f^2 g^2)^{1/2} = \int |fg|$ only if $\alpha|f| + |g| = 0$ for some α.

Increasing the number of basis functions K by one, the procedure will stop when $g_{K,\hat{\lambda}} \approx g_{K+1,\hat{\lambda}}$ in the sense of the partial affinity,

$$\rho(g_{K,\hat{\lambda}}, g_{K+1,\hat{\lambda}}) = \frac{\int |g_{K,\hat{\lambda}} g_{K+1,\hat{\lambda}}|}{\sqrt{\int g_{K,\hat{\lambda}}^2 g_{K+1,\hat{\lambda}}^2}} \approx 1,$$

where the dependence of λ on K is omitted for the sake of simplicity. Simulations were performed in order to verify the behavior of the affinity and the partial affinity. Figure 6.3 shows a typical example given by the underlying function $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \epsilon$.

One may notice that the affinity is a concave function of the number of basis functions (knots) and the partial affinity approaches one quickly. Moreover, numerical experiments have shown that the maximum of the affinity and the stabilization of the partial affinity coincide. That means, increasing the K arbitrarily not only increases the computational cost but also does not provide the best fitted curve (in the pseudo Hellinger norm).

It would be useful to have the distribution of the affinity between the true curve and the estimate produced by the adaptive H-splines method. A previous study (Dias, 1996) showed an empirical unimodal density with support on $[0,1]$ skewed to the left suggesting a beta model. To illustrate, five thousand replicates with sample size 20,100,200 and 500 were taken from a test function $y_i = x_i^3 + \epsilon_i$, where and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0,5)$.

Figure 6.4 shows that the empirical affinity distribution (unimodal, skewed to the left with range between 0 and 1), a nonparametric density estimate using kernel method and a parametric one using a beta model whose parameters were estimated using method of the moments. Similar results were obtained for several other test functions and some of them are exhibited on Figure 2.5 which brings more evidences to support a beta model.

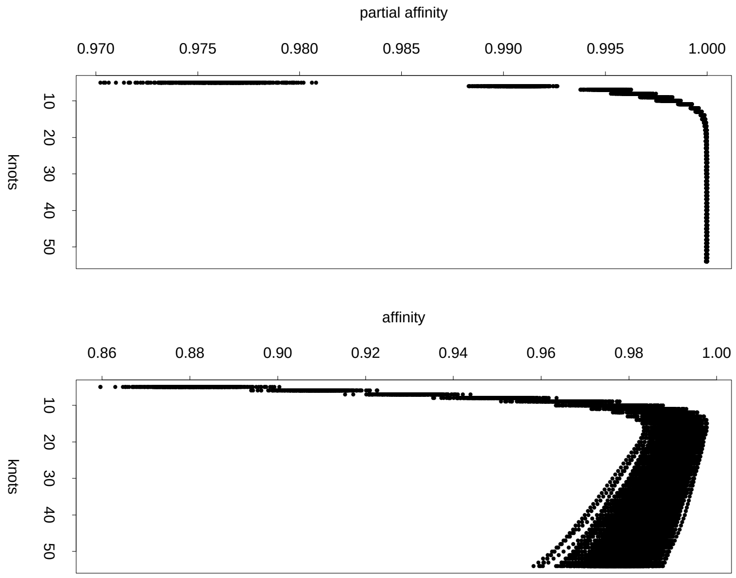

Figure 6.3: Five thousand replicates of the affinity and the partial affinity for adaptive nonparametric regression using H-splines with the true curve.

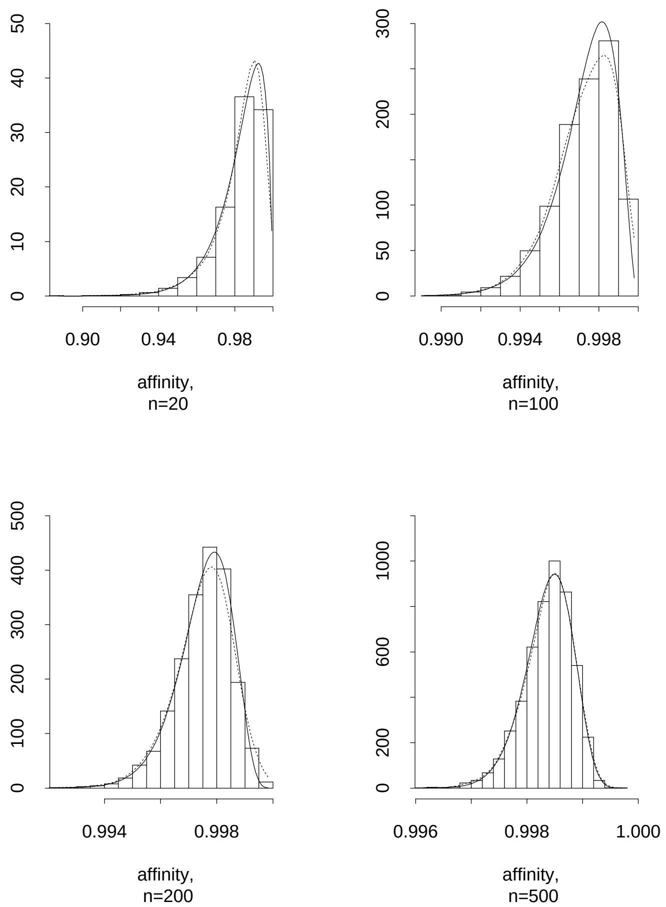

Figure 6.4: Density estimates of the affinity based on five thousand replicates of the curve $y_i = x_i^3 + \epsilon_i$ with $\epsilon_i \sim N(0, .5)$. Solid line is a density estimate using beta model and dotted line is a nonparametric density estimate.

Figure 6.5 shows that, in general, H-splines method has similar performance as smoothing splines. But as mentioned before the H-splines approach solves a linear system of order K while smoothing splines must have to solve a linear system of order $n \geq K$.

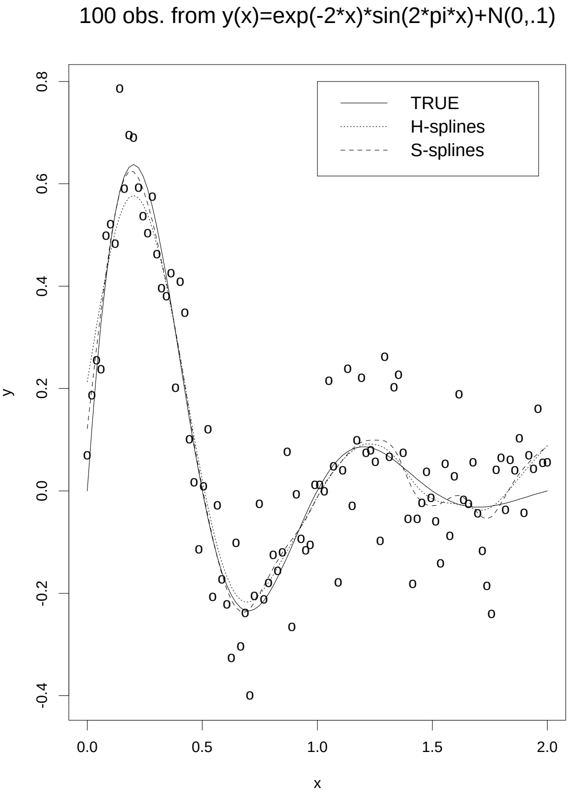

Figure 6.5: A comparison between smoothing splines (S-splines) and hybrid splines (H-splines) methods.

6.2 P-splines

The basic idea of P-splines proposed by Eilers and Marx (1996) is to use a considerable number of knots and to control the smoothness through a difference penalty on coefficients of adjacent B-splines. For this, let's consider a simple regression model $y(x) = g(x) + \varepsilon$, where ε is a random variable with symmetric distribution with mean zero and finite variance. Assume that the regression curve g can be well approximate by a linear combination of, without loss of generality, cubic B-splines, denoted by $B(x) = B(x;3)$. Specifically, Given n data points (x_i, y_i) on a set of K B-splines $B_j(.)$, we take, $g(x_i) = \sum_{j=1}^{K} a_j B_j(x_i)$. Now, the penalized least square problem becomes to find a vector of coefficients $\mathbf{a} = (a_1 \ldots, a_K)$ that minimizes:

$$PLS(\mathbf{a}) = \sum_{i=1}^{n} \Big\{ Y_i - \sum_{j=1}^{K} a_j B_j(x_i) \Big\}^2 + \lambda \int \Big\{ \sum_{j=1}^{K} a_j B''{}_j{}^{(}x_i) \Big\}.$$

Following, de Boor (1978), we have that the second derivative

$$\sum_{j=1}^{K} a_j B''_j{}^{(}x_i;3) = h^2 \sum_{j=1}^{K} \Delta^2 a_j B_j(x_i;1)$$

where h is the distance between knots and $\Delta^2 a_j = \Delta(\Delta a_j) = \Delta(a_j - a_{j-1})$ The P-splines method penalizes the higher-order of the finite differences of the coefficients of adjacent B-splines. That is,

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{K} a_j B_j(x_i) \right\}^2 + \lambda \sum_{j=m+1}^{K} (\Delta^m(a_j))^2.$$

Eilers and Marx (1996) shows that the difference penalty is a good discrete approximation to the integrated square of the kth derivative and with this this penalty moments of the data are conserved and polynomial regression models occur as limits for large values of λ.

Figure 6.6 shows a comparison of smooth.spline and P-spline estimates on simulated example.

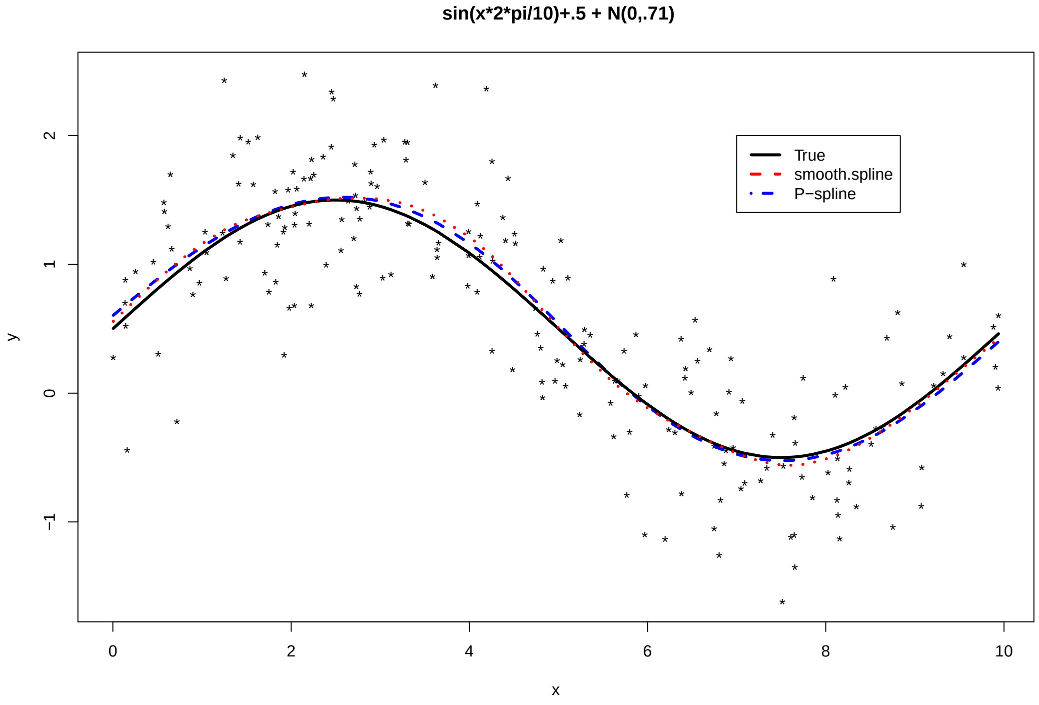

Figure 6.6: smooth spline and P-spline

6.3 Adaptive Regression via H-Splines Method

In smoothing techniques, the number of basis functions is chosen to be as large as the number of observations and then the smoothing parameter is chosen to control the flexibility of the fitting (Bates and Wahba, 1982). The h-splines method for nonparametric regression (Luo and Wahba (1997), Dias (1998) and Dias (1999)) combines some features of regression splines and of traditional smoothing splines to obtain a hybrid smoothing procedure which is usually implemented with large data sets and displays a desirable form of spatial adaptability when the underlying function is spatially inhomogeneous in its degree of complexity. Basically, choosing the number of basis functions, for instance by GCV criterion, will do most of the work for balancing between bias and variance. But there is a more important reason why we want to do a penalized regression, namely numerical stability. It is well known that as the number of basis functions increases, the regression problem becomes more ill-conditioned, which makes the numerical computation less stable. The basis functions used, in general, are the cubic spline basis which have larger correlations among them than linear spline basis, hence the ill-conditioning problem is more serious. The penalized regression step acts as a remedy for this. Similarly to smoothing splines, take the penalty term $J(g)$ as $\int (g'')^2$, the point evaluation functionals $\mathcal{L}_i g = g(t_i)$ $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{g} = (g(t_1), \ldots, g(t_n))^T$, and assume that $g \approx g_{K,\theta} = \sum_{i=1}^{K} \theta_i B_i = X_K \theta$ so that $g_{K,\theta} \in \mathcal{H}_K$, where $\mathcal{H}_K$ denotes the space of natural cubic splines (NCS) spanned by the basis functions $\{B_i\}_{i=1}^{K}$ and X_K is a $n \times K$ matrix with entries $(X_K)_{\{i,j\}} = B_i(t_j)$, for $i = 1, \ldots, K$ and $j = 1, \ldots, n$. Then, the numerical problem is to find a vector $\theta = (\theta_1, \ldots, \theta_K)^T$ that minimizes the equation (5.3),

$$L^*_\lambda(\theta) = \|\mathbf{y} - X_K \theta\|_2^2 + \lambda \theta^T \Omega \theta, \tag{6.2}$$

where now the matrix Ω is $K \times K$ matrix with entries $\Omega_{ij} = \int B_i''(t) B_j''(t) dt$. Standard calculations (de Boor, 1978) provide θ as a solution of the following linear system $(X^T X + \lambda \Omega)\theta_\lambda = X^T y$. Note that the linear system now involves $K \times K$ matrices instead of using $n \times n$ matrices which is the case of smoothing splines. Both K and λ controls the trade off between smoothness and fidelity to the data. By construction H-splines is more adaptive than the regular smoothing splines method. Simulations (see Dias (1999)) show that H-splines method has better performance even for small data sets (50 observations) and relatively large variance in the measurement errors.

6.4 A Bayesian Approach to H-splines

We have seen that there are several methods to estimate non-parametrically an unknown regression curve g by using splines since the pioneer work of Craven and Wahba (1979). Kimeldorf and Wahba (1970) and Wahba (1983) gave an attractive Bayesian interpretation for an estimate $\hat{g}$ of the unknown curve g. They showed that $\hat{g}$ can be viewed as a Bayes estimate of g with respect to a certain prior on the class of all smooth functions. The Bayesian approach allows one not only to estimate the unknown function, but also to provide error bounds by constructing the corresponding, Bayesian confidence intervals (Wahba, 1983). In this section we allow two smoothing parameters as Luo and Wahba (1997) and Dias (1999) did. However, instead of going through the difficulties of specifying them precisely in an ad-hoc manner, they are allowed to vary according to prior information. In this way, the procedure becomes more capable of providing an adequate fit.

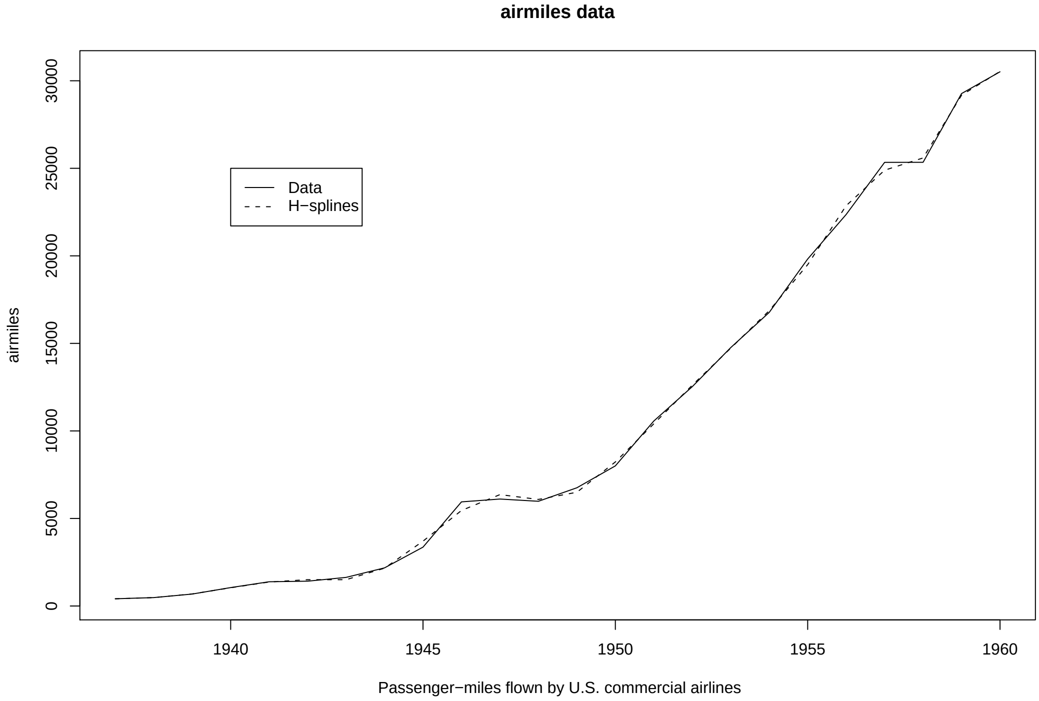

Figure 6.7: H-spline fitting for airmiles data

Suppose we have the following regression model,

$$y_i = g(t_i) + \varepsilon_i \quad i = 1, \ldots, n.$$

where ε_i's are uncorrelated with a $N(0, \sigma^2)$. Moreover, assume that the parametric form of the regression curve g is unknown. Then the likelihood of g given the observations $\mathbf{y}$ is,

$$l_{\mathbf{y}}(g) \propto (\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{g}||^2\}. \tag{6.3}$$

The Bayesian justification of penalized maximum likelihood is to place a prior density proportional to $\exp\{-\frac{\lambda}{2} \int (g'')^2\}$ over the space of all smooth functions. (see details in Silverman and Green (1994) and Kimeldorf and Wahba (1970)). However, an infinite dimensional case has a paradox alluded to by Wahba (1983). Silverman (1985) proposed a finite dimensional Bayesian formulation to avoid the paradoxes and difficulties involved in the infinite dimensional case. For this, let $g_{K,\theta} = \sum_{i=1}^{K} \theta_i B_i^x = X_K \theta_K$ with a knot sequence x placed at order statistics. A complete Bayesian approach would assign prior distribution to the coefficients of the expansion, to the knot positions, to the number of knots and for σ^2. A Bayesian approach to hybrid splines non-parametric regression assigns priors for g_K, K, λ and σ^2. Given a realization of K the interior knots are placed at order statistics. This well known procedure in non-parametric regression reduces the computational cost substantially and avoids trying to solve a difficult problem of optimizing the knot positions. Any other procedure has to take into account the fact that changes in the knot positions might cause considerable change in the function g (see details in Wahba (1982) for ill-posed problems in splines non-parametric regression). Moreover, in theory the number of basis functions (which is a linear function of the number of knots) can be as large as the sample size. But then one has to solve a system of n equations instead of K. An attempt to keep the computational cost down one might want to have K small as possible and hence over-smoothing may occur. For any K large enough λ keeps the balance between over-smoothing and under-smoothing. Thus the penalized likelihood becomes with $g = g_K$,

$$l_p \propto (\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}||\mathbf{y} - g_K||^2\} \times \exp\{-\frac{\lambda}{2}\int (g_K'')^2\}, \tag{6.4}$$

where $g_K = g_{K,\theta} = X_K\theta_K$. The reason why we suppress the subindex θ in $g_{K,\theta}$ will be explained later in this section. Note that maximization of the penalized likelihood is equivalent to minimization of (5.3). For this proposed Bayesian set up we have a prior for $(g_K, K, \lambda, \sigma^2)$ of the form,

$$\begin{aligned} p(g_K, K, \lambda, \sigma^2) &= p(g_K|K,\lambda)p(K,\lambda)p(\sigma^2) \qquad (6.5) \\ &= p(g_K|K,\lambda)p(\lambda|K)p(K)p(\sigma^2), \end{aligned}$$

where

$$p(g_K|K,\lambda) \propto \exp\{-\frac{\lambda}{2}\int (g_K'')^2\},$$

$$p(\sigma^2) \propto \frac{1}{(\sigma^2)^{(u+1)}} \exp\{-v\sigma^2\},$$

for $u > 0, v > 0$,

$$p(K) = \frac{\exp\{-a\}a^K/K!}{1 - \exp\{-a\}(1+q^*)}, \quad K = 1, \ldots, K^*$$

where $q^* = \sum_{j=K^*+1}^{\infty} a^j/j!$, and

$$p(\lambda|K) = \psi(K)\exp\{-\psi(K)\lambda\},$$

with ψ any smooth function of K. It is well known that, for $\lambda > 0$, when the number of basis functions increases the smoothing parameter decreases to a point and then it increases with K. That is, for large values of K the smoothing parameter λ becomes larger to enforce smoothness. (See details in Dias (1999).). Therefore, functions ψ that satisfy these requirements are recommended. In particular, a flexible class is given by $\psi(K) = K^{-b}\exp(-cK)$ for suitably chosen hyperparameters b and c. However, undesirably large values of K can be excluded through fixing K^* appropriately or be made unlikely by fixing a accordingly. These large values of K can be controlled by the hyperparameters a and K^* of the prior $p(K)$.

The choice of a is governed by the prior expectation of the structure of the underlying curve such as maxima, minima, inflection points etc. We suggest the reader to follow some of the rules recommended by Wegman and Wright (1983). These recommendations are based on the assumption of fitting a cubic spline, the most popular case and are summarized below.

1. Extrema should be center in intervals and inflection points should be located near knot points.

2. No more than one extremum and one inflection point should fall between knots (because a cubic could not fit more).

3. Knot points should be located at data points.

Note that $g_K = X_K\theta_K$ is completely determined if we know the coefficients θ_K. Hence, the overall parameter space Ξ can be written as countable union of subspaces $\Xi_K = \{\xi : \xi = (\theta_K, \phi) \in (\mathbb{R}^K \times \{1, \ldots, K^*\} \times [0, \infty]^2)\}$ with $\phi = (K, \lambda, \sigma^2)$. Thus, the posterior is given by

$$\pi(\xi|\mathbf{y}) \propto l_p(\xi|\mathbf{y})p(\xi). \tag{6.6}$$

In order to sample from the posterior $\pi(\xi|\mathbf{y})$ we have to consider the variation of dimensionality of this problem. Hence one has to design move types between subspaces Ξ_K. However, assigning a prior to g_K, equivalently to the coefficients θ_K, leads to a serious computational difficulty pointed out by Denison, Mallick and Smith (1998) where a comparative study was developed. They suggest that the least square estimates for the vector θ_K leads to a non-significant deterioration in performance for overall curve estimation. Similarly, given a realization of (K, λ, σ^2), we solve the penalized least square objective function (6.2) to obtain the estimates, $\hat{\theta}_K = \hat{\theta}_K(\mathbf{y})$, for the vector θ_K and consequently we have an estimate $\hat{g}_K = X_K\hat{\theta}_K$. Thus, there is no prior assigned for this vector of parameters, and so, we write $g_K = g_{K,\theta}$. Having got $\hat{\theta}_K$, we approximate the marginal posterior $\pi(\phi|\mathbf{y})$ by the conditional posterior

$$\pi(\phi|\mathbf{y}, \hat{\theta}_K) \propto l_p(\phi|\mathbf{y}, \hat{\theta}_K)p(\phi), \tag{6.7}$$

with

$$p(\phi) = p(K, \lambda, \sigma^2) = p(\lambda|K)p(K)p(\sigma^2).$$

Note that if one assigns independent normal distributions to the parameters θ_K it will not be difficult to obtain the marginal posterior $\pi(\phi|\mathbf{y})$ and approximations will not be necessary. However, the results will be very similar.

To solve the problem of sampling from the posterior, $\pi(\phi|\mathbf{y}, \hat{\theta}_K)$, Dias and Gamerman (2002) used reversible jump methodology (Green (1995)). This technique is beyond the level of this book and it will not be explained here but the interested reader will find the details of the algorithm in Dias and Gamerman (2002).

In figure 6.8 we present a simulated example to verify the estimates provided by this approach. The final estimate is,

$$\hat{y}_+(t_i) = \frac{1}{100}\sum_{j=1}^{100}\hat{y}_j(t_i).$$

Figure 6.9, exhibit approximate Bayesian confidence intervals for the true curve regression f and it was computed as following. Let $y(t_i)$ and $\hat{y}(t_i)$ be a particular model and its estimate provided by this proposed method with $i = 1, \ldots, n$ where n is the sample size. For each $i = 1, \ldots, n$ the fitting vectors $(\hat{y}_1(t_i), \hat{y}_2(t_i), \ldots, \hat{y}_{100}(t_i))^T$ form random samples and from those vectors the lower and upper quantiles were computed in order to obtain the confidence intervals.

Figure 6.8 exhibits an example of how useful a Bayesian approach to hybrid splines non-parametric regression can be. It describes a situation where a prior information tells that the underlying curve has large curvature and the variance of the error measurements is not too small and the traditional methods of smoothing, e.g. smoothing splines, might not be able to capture all the structure of the true regression curve. By using vague but proper priors to the smoothing parameters K and λ and for the variance σ^2 this Bayesian approach provides a much better fitting than the traditional smooth splines approach does.

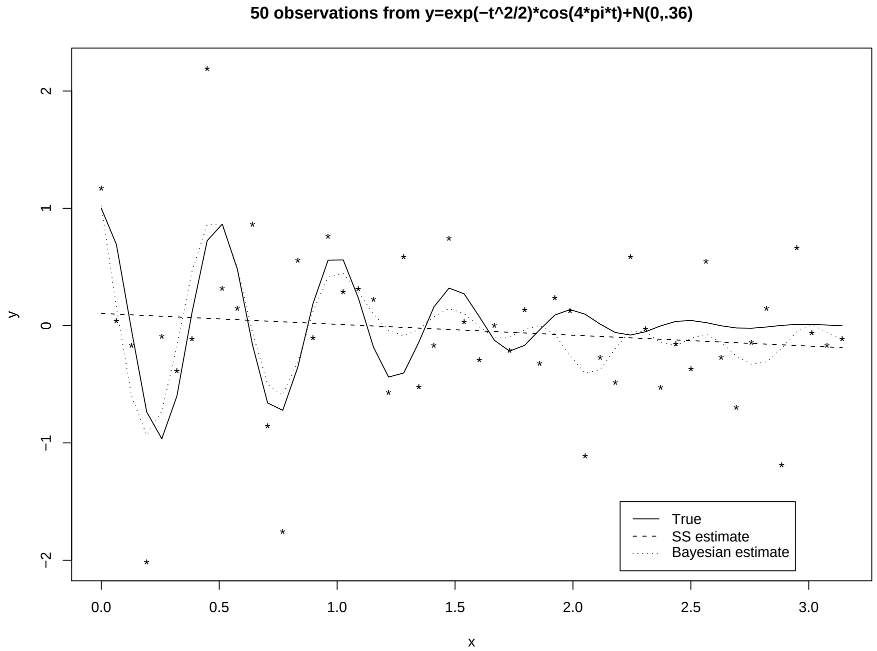

Figure 6.8: Estimation results: a) Bayesian estimate with $a = 17$ and $\psi(K) = K^3$ (dotted line); b) (SS) smoothing splines estimate (dashed line). The true regression function is also plotted (solid line). The SS estimate was computed using the R function `smooth.spline` from which 4 degrees of freedom were obtained and λ was computed by GCV.

Figure 6.9 shows one hundred curves sampling from the posterior (after burn-in) and approximate 95% Bayesian confidence interval for the regression curve $g(t) = \exp(-t^2/2)\cos(4\pi t)$ with $t \in [0, \pi]$. On the right panel of this figure we see the curve estimate which is an approximation for the posterior mean and the percentiles curves 2.5% and 97.5%.

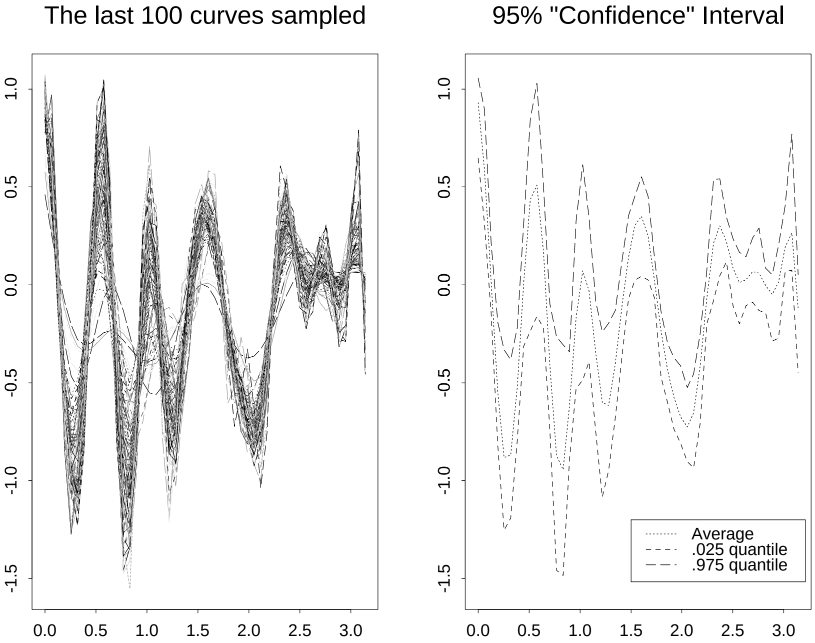

Figure 6.9: One hundred estimates of the curve 6.8 and a Bayesian confidence interval for the regression curve $g(t) = \exp(-t^2/2)\cos(4\pi t)$ with $t \in [0, \pi]$.

Chapter 7

Final Comments

Compared to parametric techniques nonparametric modeling has more flexibility since it allows one to choose from an 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 the 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.

Bibliography

Bates, D. and Wahba, G. (1982). *Computational Methods for Generalized Cross-Validation with large data sets*, Academic Press, London.

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.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic bayesian curve fitting, *Journal of the Royal Statistical Society B* **60**: 363–377.

Dias, R. (1994). Density estimation via h-splines, *University of Wisconsin-Madison*. Ph.D. dissertation.

Dias, R. (1996). Sequential adaptive nonparametric regression via H-splines. Technical Report RP 43/96, University of Campinas, June 1996. Submitted.

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.

Dias, R. and Gamerman, D. (2002). A Bayesian approach to hybrid splines nonparametric regression, *Journal of Statistical Computation and Simulation.* **72**(4): 285–297.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with *B*-splines and penalties, *Statist. Sci.* **11**(2): 89–121. With comments and a rejoinder by the authors.

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.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika* **58**: 255–277.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and bayesian model determination, *Biometrika* **82**: 711–732.

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.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines, *The Annals of Mathematical Statistics* **41**: 495–502.

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.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators, *SIAM J. on Scientific and Stat'l. Computing* **9**: 363–379.

Pagan, A. and Ullah, A. (1999). *Nonparametric econometrics*, Cambridge University Press, Cambridge.

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. (1982). On the estimation of a probability density function by the maximum penalized likelihood method, *Ann. of Statistics* **10**: 795–810.

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

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting, *Journal of the Royal Statistical Society, Series B, Methodological* **47**: 1–21.

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.

Thompson, J. R. and Tapia, R. A. (1990). *Nonparametric Function Estimation, Modeling and Simulation*, SIAM:PA.

Wahba, G. (1982). Constrained regularization for ill posed linear operator equations, with applications in meteorology and medicine, *in* S. S. Gupta and J. O. Berger (eds), *Statistical Decision Theory and Related Topics III, in two volumes*, Vol. 2, Academic:NY:Lnd, pp. 383–418.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline, *JRSS-B, Methodological* **45**: 133–150.

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

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

Wegman, E., J. and Wright, I. W. (1983). Splines in statistics, *Journal of the American Statistical Association* **78**: 351–365.