A review of non-parametric curve estimation methods with application to Econometrics

Ronaldo Dias

Universidade Estadual de Campinas

`dias@ime.unicamp.br` *

Abstract

Various features of econometric data can be analyzed by non-parametric approach. This review summarizes some of the most important procedures in curve estimation that has been very useful in the field of econometrics. Specifically, it describes the theory and the applications of non-parametric density and regression estimation problems with emphases in kernel, nearest neighbor, variable kernel, orthogonal series, smoothing splines, logsplines and H-splines methods.

1   Introduction

Certainly, 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

---

*Postal address: Departamento de Estatística, IMECC, Cidade Universitária "Zeferino Vaz", Caixa Postal 6065, 13.081-970 - Campinas, SP - BRAZIL, e-mail address: `dias@ime.unicamp.br`

smoothing technique used to estimate the unknown density and hence it deserves some consideration.

Let us try to combine the data by counting how many data points fall into a small interval of length h. This kind of interval is called a *bin*. Observe that the well known dot plot (Box, Hunter and Hunter, 1978, 25–26) is a particular type of histogram where $h = 0$.

Without loss of generality, we consider a *bin* centered at 0, namely the interval $[-h/2, h/2)$. Consequently the probability for an observation of X to fall into the interval $[-h/2, h/2)$ is given by:

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

where f 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 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 $n_j = \sum_{i=1}^{n} I(X_i \in C_j)$, such that, $\sum_{j=1}^{k} n_j = n$. Then,

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

for all x. Here the function $I(x \in A)$ is defined to be :

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

Note that the density estimate $\hat{f}_h$ depends strongly upon the *histogram bandwidth* h. By varying h we can have different shapes of $\hat{f}_h$. For example, if one increases h, one is averaging over more data and the histogram appears to be smoother. The extremes of h, say, when $h \to 0$, the histogram becomes a very noisy representation of the data (needle-plot, Härdle(1990)). In opposite situation when $h \to \infty$, the histogram, now, becomes overly smooth (box-shaped, Härdle(1990)). Thus, h is the smoothing parameter of this type of density estimate, and the question of how to choose the histogram bandwidth h turns out to be an important question in representing the data via the histogram. For details of how to estimate h see Härdle (1990).

2   Kernel Density Estimation

Naturally, we can think of having a more general idea of an density estimate of the underlying density based on the method of the histogram. For this consider the weight function,

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

and define the estimator,

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

We can see that $\hat{f}_h$ 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 $\hat{f}_h$. See Silverman (1986) for a discussion of this kind of estimator. It is not difficult to verify that $\hat{f}_h$ 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 non smoothness (Silverman, 1986), it could give a misleading impression to a untrained observer.

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

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

Hence $K(x)$ is a probability density function, and usually is a symmetric density, as for instance, normal density. 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.
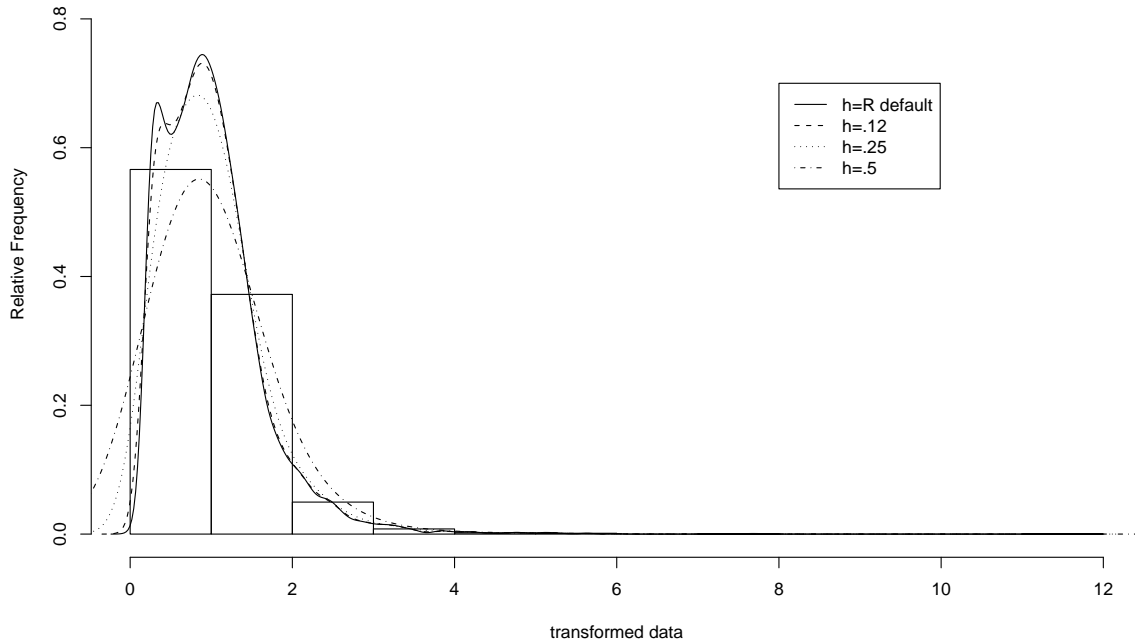
Histogram of income/mean(income)

Figure 2.1: Bandwidth effect on kernel density estimates

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

3 Statistical Results of Kernel Density Estimation

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

$$
\begin{align}
E[\hat{f}_h(x)] &= \frac{1}{nh} \sum_{i=1}^{n} E[K(\frac{x - X_i}{h})] \tag{3.1} \\
&= \frac{1}{h} E[K(\frac{x - X_i}{h})] \tag{3.2} \\
&= \frac{1}{h} \int K(\frac{x - u}{h}) f(u) du \tag{3.3} \\
&= \int K(y) f(x + yh) dy. \tag{3.4}
\end{align}
$$

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

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

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

We observe that since we assumed the kernel K symmetric around zero the term $\int yK(y)hf'(x)dy = 0$, the bias is quadratic in h. See (Parzen, 1962).

Using similar approach we obtain :

- $Var_f[\hat{f}_h(x)] = \frac{1}{nh}\|K\|_2^2 f(x) + o(\frac{1}{nh})$,

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

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

Thus the following condition $h \to 0$ and $nh \to \infty$ is usually assumed, the $MSE_f[\hat{f}_h] \to 0$, 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 estimate 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.

4   The nearest neighbor method

The concept of the nearest neighbor method is to adapt the amount of smoothing to local density of 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 that is the kth nearest to x, for suitable k, typically, $k \propto n^{1/2}$.

The kth nearest neighbor density estimate is defined as,

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

where, n is the sample size and $d_k(x)$ is the kth distance between x and the kth data point near to x.

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 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 it is required to estimate the entire density. However, it is possible to generalize the nearest neighbor estimate to provide an estimate 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.
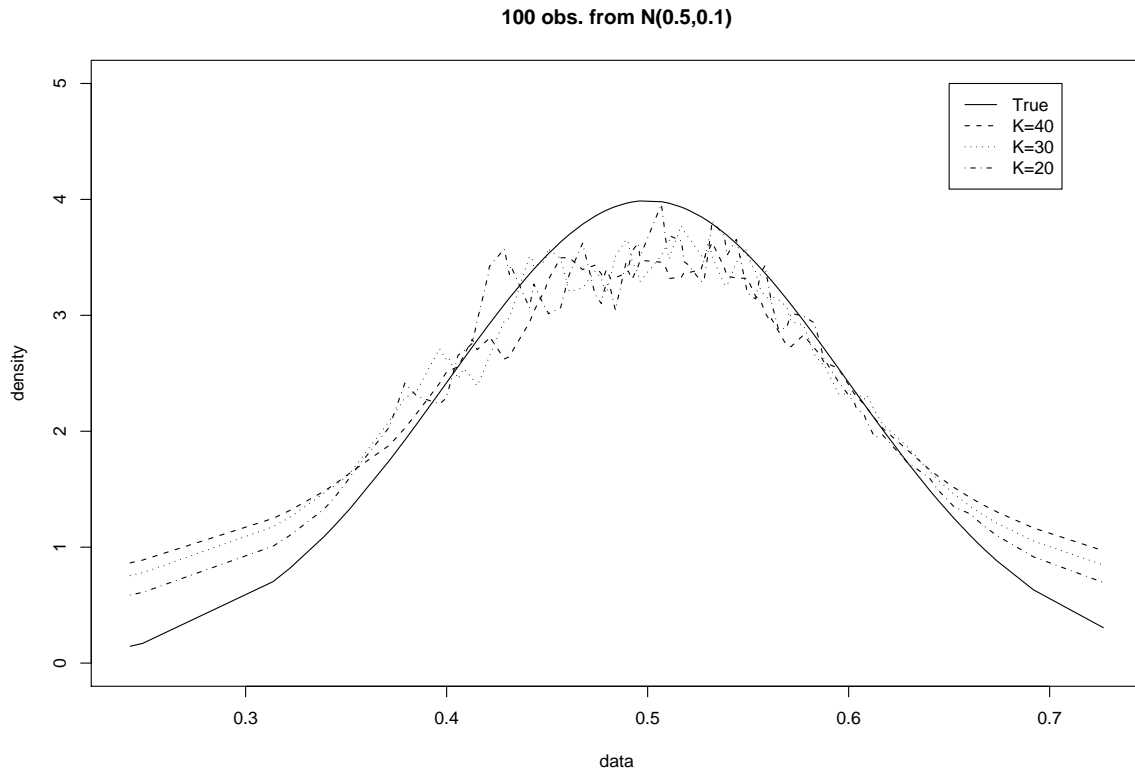
Figure 4.2: Effect of the smoothing parameter K on the estimates

5 The variable kernel method

The variable kernel method is another method which adapts the amount of smoothing to the local density of the data. The estimate is constructed similarly to the classical kernel estimate, but the scale parameter varies from one data point to another.

Let K be a kernel function and k a positive integer. Define $d_{j,k}$ to be the distance from X_j to the kth nearest point in the set containing the other $n-1$ data points. The variable kernel estimate with the smoothing parameter h is defined by

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h d_{j,k}} K\left(\frac{x - X_j}{h d_{j,k}}\right).$$

In contrast with the generalized nearest neighbor estimate, the variable kernel estimate will itself be a probability density function provided that K is.

6   Bandwidth selection: Kernel estimators

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

$$h_* = \left(\frac{f(x)\|K\|_2^2}{(f''(x))^2(\int y^2 K(y)dy)^2 n}\right)^{1/5} \propto n^{-1/5}. \tag{6.1}$$

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:

$$\begin{aligned} IMSE[\hat{f}_h] &= \int MSE_f[\hat{f}_h] \\ &= \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\left(\int y^2 K(y)dy\right)^2\|f''\|_2^2 + o(\frac{1}{nh}) + o(h^4). \end{aligned} \tag{6.2}$$

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

The optimal value of h considering the IMSE is define as

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

it can be shown that,

$$h_{opt} = c_2^{-2/5}\left(\int K^2(x)dx\right)^{1/5}\left(\|f''\|_2^2\right)^{-1/5} n^{-1/5}, \tag{6.3}$$

where $c_2 = \int y^2 K(y)dy$. Unfortunately, (6.3) depends on the second derivatives of f which measures the rapidity of fluctuations in density f.

6.1   Reference to a standard distribution

A very natural way is to use a standard family of distributions to assign a value of the term $\|f''\|_2^2$ in the expression of the ideal bandwidth (6.3). For example, assume

that a density f belongs to a class normal family with mean μ and variance σ^2, then

$$\int (f''(x))^2 dx = \sigma^{-5} \int (\varphi''(x))^2 dx$$
$$= \frac{3}{8}\pi^{-1}2\sigma^{-5} \approx 0.212\sigma^{-5}. \tag{6.4}$$

If one uses a Gaussian kernel, then

$$h_{opt} = (4\pi)^{-1/10}(\frac{3}{8}\pi^{-1/2})^{-1/5}\sigma n^{-1/5}$$
$$= \left(\frac{4}{3}\right)^{1/5}\sigma n^{-1/5} = 1.06\sigma n^{-1/5} \tag{6.5}$$

In practice use $1.06\hat{\sigma}n^{-1/5}$!!

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

$$\hat{h}_{opt} = 1.06\min(\hat{\sigma}, \frac{\hat{R}}{(\Phi(3/4) - \Phi(1/4))})$$
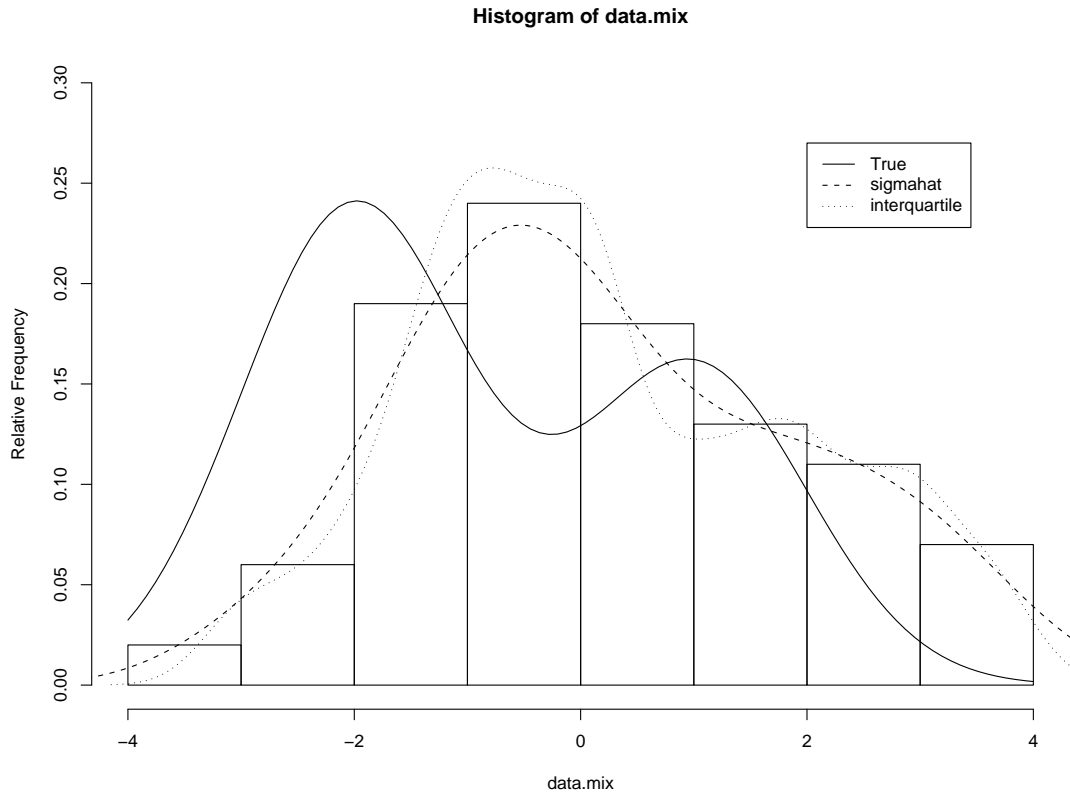$$= 1.06\min(\hat{\sigma}, \frac{\hat{R}}{1.349}). \tag{6.6}$$

Figure 6.3: Comparison of two bandwidths, $\hat{\sigma}$ and $\hat{R}$.

6.2   Maximum likelihood Cross-Validation

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

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

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

$$d_{KL}(f, f_h) = \int \log(\frac{f}{f_h})(x)f(x)dx. \tag{6.7}$$

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

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

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

Hence,

$$\prod_{i=1}^{n} f_{h,i}(X_i) = (n-1)^{-n}h^{-n}\prod_{i=1}^{n}\sum_{j \neq i} K(\frac{X_i - X_j}{h}). \tag{6.8}$$

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

$$
\begin{aligned}
CV_{KL}(h) &= \frac{1}{n}\sum_{i=1}^{n}\log[f_{h,i}(X_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\log\Big[\sum_{j \neq i} K(\frac{X_i - X_j}{h})\Big] - \log[(n-1)h] \tag{6.9}
\end{aligned}
$$

Naturally, we choose h_{KL} such that:

$$h_{KL} = \arg\max_{h} CV_{KL}(h)$$

13

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

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

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

$$
\begin{aligned}
\mathbb{E}[CV_{KL}(h) &\approx \mathbb{E}\left[\int \log f_h(x) f(x) dx\right] \\
&\approx -\mathbb{E}[d_{kl}(f, f_h)] + \int \log[f(x)] f(x) dx.
\end{aligned}
\tag{6.10}
$$

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

The Maximum likelihood cross validation has two shortcomings:

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

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

The computation of $CV_{KL}(h)$ for a set of bandwidths $h_1, \ldots, h_m$ may be based on the following algorithm (see Härdle (1990)):

Algorithm 6.1 `for k=1 to m`

```
for i=1 to n
score=0
for j=1 to n
score=score+K((X[i]-X[j])/h[k])
```

```
endfor j
cv.kl[k]=cv.kl[k]+log(score-k(0))
endfor i
cv.kl[k]=cv.kl[k]/n - log((n-1)*h[k])
endfor k
```

Note that this computation requires iterations of the order $O(m \times n^2)$, i.e., is quadratic in the number of observations. This is a great disadvantage of this technique, which forces us to look for other techniques with better numerical efficiency.

6.3   Least-squares cross- validation

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

$$
\begin{aligned}
d_{ISE}(h) &= \int (f_h - f)^2(x)dx \\
&= \int f_h^2(x)dx - 2\int (f_h f)(x)dx + \int f^2(x)dx \\
d_{ISE}(h) - \int f^2(x)dx &= \int f_h^2(x)dx - 2\int (f_h f)(x)dx
\end{aligned}
\tag{6.11}
$$

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

$$
\mathbb{E}_X[\hat{f_h}(X)] = \frac{1}{n}\sum_{i=1}^{n} f_{h,i}(X_i)
\tag{6.12}
$$

This leads to the Least-squares cross-validation:

$$
CV_{LS}(h) = \int f_h^2(x)dx - 2\sum_{i=1}^{n} f_{h,i}(X_i)
\tag{6.13}
$$

The bandwidth minimizing this function is,

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

15

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

$$
\begin{aligned}
\mathbb{E}[CV_{LS}(h)] &= \mathbb{E}[d_{ISE}(h) + 2(\mathbb{E}_X[f_h(X)] - \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} f_{h,i}(X_i)]) - \|f\|_2^2 \\
&= IMSE[f_h] - \|f\|_2^2.
\end{aligned}
\tag{6.14}
$$

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, Härdle (1990) suggests the following algorithm

Algorithm 6.2 `for k=1 to m`

```
for i=1 to n
for j=i+1 to n
cv.ls[k]=cv.ls[k] + kernel.conv((X[i]-X[j])/h[k])
cv.ls[k]=cv.ls[k] - kernel((X[i]-X[j])/h[k])*2*n/(n-1)
endfor j
endfor i
cv.ls[k]=cv.ls[k]+kernel.conv(0)*n/2
cv.ls[k]=cv.ls[k]*2/(n**2*h[k])
endfor k
```

7 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

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

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

$$a_i = \int f(x)\phi_i(x)dx. \tag{7.1}$$

Now, suppose that X is a random variable with density f. Then (7.1) 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

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

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

$$\omega(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - X_i) \tag{7.2}$$

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

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

17

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

$$\hat{f}(x) = \sum_{i=1}^{K} \hat{a}_i \phi_i(x). \tag{7.3}$$

Note that the amount of smoothing is determined by K. Small value of K implies in over-smoothing, large value of K under-smoothing.
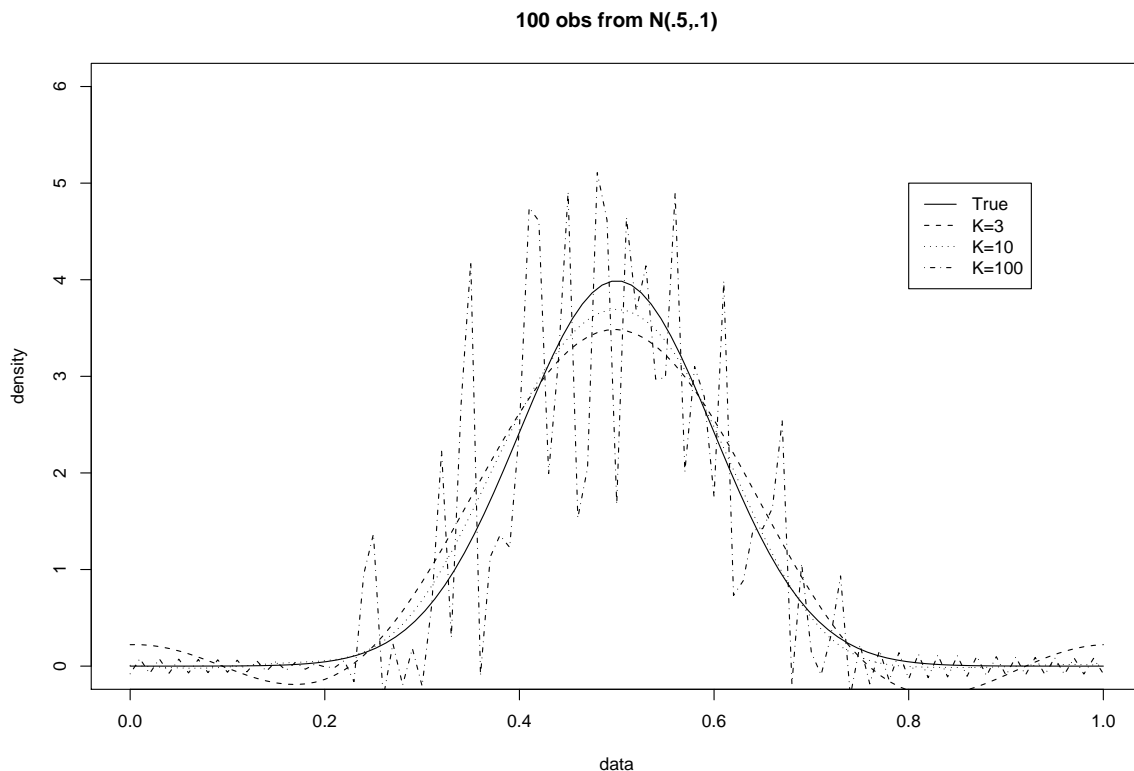
100 obs from N(.5,.1)

Figure 7.4: 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

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

18

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.
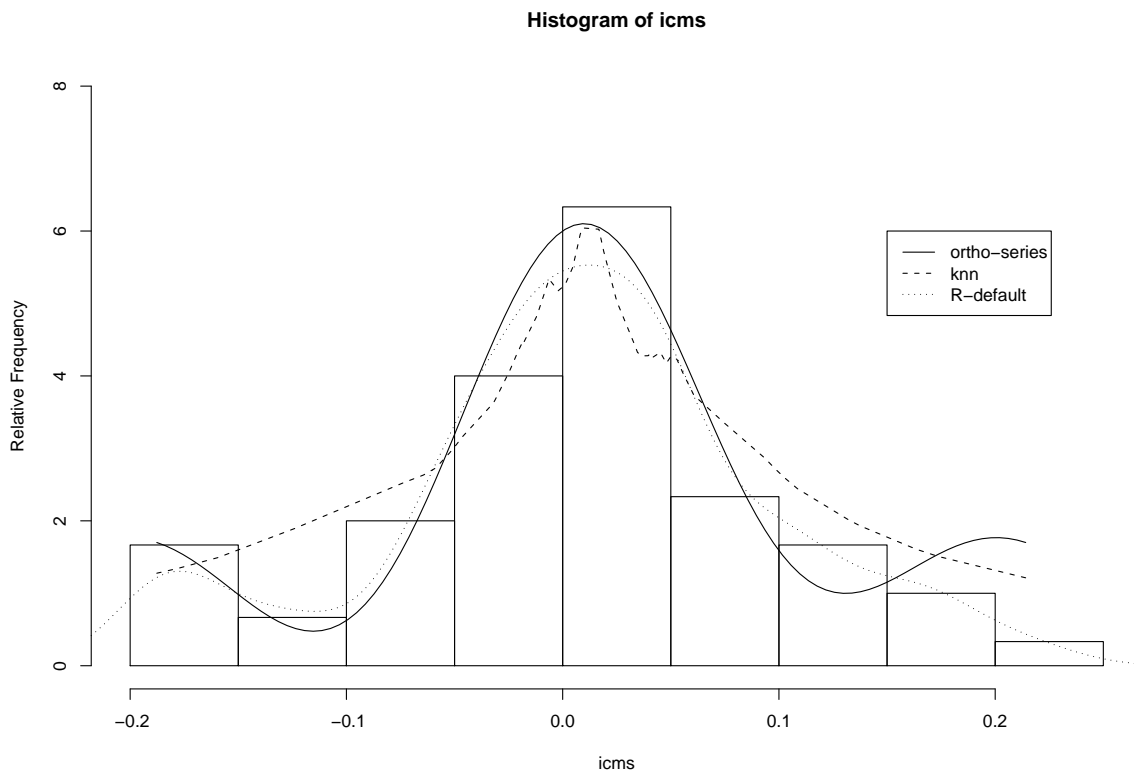
Histogram of icms

Figure 7.5: A comparison of three methods of density estimation, orthogonal series with 10 basis, k-th nearest neighbor with smoothing parameter k=5 and kernel density estimate with bandwidth h=2.5

8 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 prespecified 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,

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

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 $\left|\frac{X_i - X_i'}{h/c}\right| > M > 0$, 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,

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

Letting $h \to 0$, we have $\mathcal{L} \to \infty$.

Consequently, $\mathcal{L}(f|X_1, \ldots, X_n)$ does not have a finite maximum over the class of all densities. That is, the likelihood function can be as large as one wants it just

by taking densities with h approaching zero. Densities having this characteristic, 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 M.L. 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

$$\mathcal{L}_\lambda(g) = \frac{1}{n} \sum_{i=1}^n g(X_i) - \lambda J(g) \, , \tag{8.1}$$

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 $\gamma \in L^2(-\infty, \infty)$.

21

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,

$$2\alpha\sigma^2 + \frac{3}{4}\beta = \sigma^4, \tag{8.2}$$

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,

$$-N + \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{\sigma^2} + \frac{2\alpha}{\sigma^2} + \frac{3\beta}{4\sigma^4} = 0. \tag{8.3}$$

If we put $\sigma^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2/N - 1$, the equation (7.3) becomes,

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

Hence we have the constraint (7.2).

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.

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 S and for any fixed positive λ, let

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

and

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

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

9   Approximation by Spline functions

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

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

In the general case, the polynomial pieces $p_i(x)$ are constructed independently of each other and therefore do not constitute a continuous function $s(x)$ on $[a, b]$. This

cannot be accepted if one wants, particularly, to approximate a smooth function. Naturally, it is necessary to require the polynomial pieces $p_i(x)$ to join smoothly at knots $x_1, \ldots, x_k$, and to have all derivatives up to a certain order, coincide at knots. As a result, we get a smooth piecewise polynomial function, called a *spline function*.

Definition 9.1 *The function $s(x)$ is called a spline function (or simply "spline") of degree r with knots at $\{x_i\}_{i=1}^k$ if $-\infty =: x_0 < x_1 < \ldots < x_k < x_{k+1} := \infty$ and*

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

- *$s(x), s'(x), \ldots, s^{r-1}(x)$ are continuous functions on $(-\infty, \infty)$.*

The set of such functions, $\mathcal{S}_r(x_1, \ldots, x_k)$, is a linear space whose elements are spline functions and it is called *spline space*.

Definition 9.2 *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.

It can be shown (Schumaker, 1981) that $\mathcal{S}_\nabla(x_1, \ldots, x_k)$ is a linear space with dimension $r + k + 1$. Then we can express any spline function by a linear combination of $r+k+1$ basis functions. 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 property toward computation that they are splines which have smallest possible support. In other words, B-splines are zero on a large set. Furthermore, a stable evaluation of B-splines with aid of recurrence relation is possible.

Definition 9.3 *Let $\Omega_\infty = \{x_j\}_{\{j \in \mathbb{Z}\}}$ such that $x_j < x_{j+1}$, that is, Ω_∞ is the infinite knot set. A B-spline of degree r is defined to be*

$$B_j^r(x) = -(x_{r+j+1} - x_j)(x - x_j)_+^r [x_j, \dots, x_{r+j+1}] ,$$

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

Of special interest is the set of natural splines of order $r = 2m$ 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 (9.1), it satisfies an extra condition:

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

Precisely speaking, let's 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, $\mathcal{NS}_{2m} = \{s \in \mathcal{S}(\mathcal{P}_{2m}) : s_0 = s|_{[a,x_1)}$ 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 $\mathcal{NS}_{\in \mathbb{m}}$ to be K. Actually, it is well known that $\mathcal{NS}_{2m}$ is linear space of dimension K. See details in Schumaker (1981).

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

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

Definition 9.4 *Let $M(x,y) = (y - x)^3_+$ and let $M[x; x_1, \ldots, x_K]$ be the $(K - 1)$st divided difference of M as a function of x taken over the knot sequence $x_1 \leq x_2 \ldots \leq x_K$ with $h_{i+1} = x_{i+1} - x_i$, $i = 1, \ldots, K - 1$ Then*

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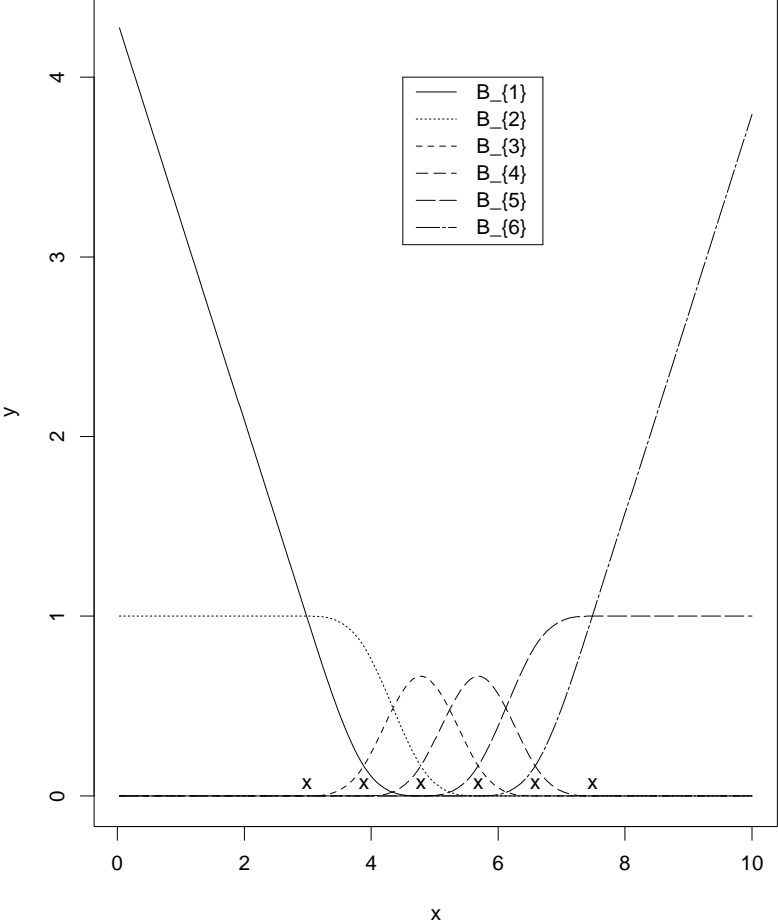\end{cases}
$$

Figure 9.6: Basis Functions with 6 knots placed at "x"

10 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 are given in the next chapter.

The basic idea of computing a density estimate provided by penalized likelihood method is to construct approximations to it. Given $x_1, \ldots, x_n$, the realizations of random variables $X_1, \ldots, X_n$, with common log density g. We are to solve a finite version of (8.1) 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

$$-\frac{1}{n}\sum_{i=1}^{n} g(x_i) + \int_a^b e^{g(s)}ds + \lambda \int_a^b (g^{(m)})^2 ds, \qquad (10.1)$$

for fixed $\lambda > 0$, and data points $x_1, \ldots, x_n$. The minimization is over a class of absolutely continuous functions on $[a, b]$ whose mth derivative is square integrable. O'Sullivan noticed that the minimizer of (10.1) satisfies a nonlinear fourth order differential equation in between distinct data points with discontinuity in the third derivatives at the data points. For details see O'Sullivan (1988). Given this behavior the exact computation has to be ruled out. His approach is to use cubic B-splines (see Chapter 3 for definitions and properties of B-splines) to approximate the estimator of (10.1).

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.

11 Smoothing Spline Density Estimation

In order to provide an algorithm which is not restricted to the unidimensional case under penalized maximum likelihood estimation, Gu (1993) proposed a dimensionless fully automatic algorithm which updates the smoothing parameter jointly with the estimate in a performance oriented iteration via cross validation estimate. The performance is measured by a proxy of the symmetrized Kullback-Leibler distance between the true density and the estimate. Specifically, let $X_1, \ldots, X_n$ be i.i.d. sample from an unknown probability density function f on a finite domain $\mathcal{X}$. The goal is to estimate the density f from the data X_i. For this, assume $f > 0$ and take the logistic transformation $f = e^g / (\int e^g)$. Observe that the logistic transformation is not one-to-one. For instance take $g* = g + c$. Then $f = e^{g*} / (\int e^{g*}) = e^g / (\int e^g)$. Hence some extra condition is necessary.

Gu and Qiu (1993) proposed side conditions on g, $g(x_0) = 0$, $x_0 \in \mathcal{X}$ or $\int_{\mathcal{X}} g = 0$ and defined the smoothing spline density estimate to be the minimizer of the penalized likelihood score,

$$-\frac{1}{n} \sum_{i=1}^{n} g(X_i) + \log \int_{\mathcal{X}} e^g + \frac{\lambda}{2} J(g) \tag{11.1}$$

in a function space $\mathcal{H}$, where J is a roughness penalty and λ is the smoothing parameter. The roughness penalty J is taken as square semi-norm in $\mathcal{H}$ and $\mathcal{H}$ is a Hilbert space in which evaluation functionals are continuous so that the first term of (11.1) is continuous.

The Hilbert space with continuous evaluation functionals is called *reproducing kernel Hilbert space* (RKHS) possessing a reproducing kernel ($\mathcal{RK}$) $R(\cdot, \cdot)$, a positive definite bivariate function on $\mathcal{X} \times \mathcal{X}$, such that, for all x in $\mathcal{X}$ we have, $R(x, \cdot) = R(\cdot, x)$ in $\mathcal{H}$, and, for any function g in $\mathcal{H}$, $\langle R(x, \cdot), g(\cdot) \rangle = g(x)$ (the reproducing property). Here, the notation $\langle \cdot, \cdot \rangle$ stands for the inner product in $\mathcal{H}$. Moreover, $\mathcal{H}$ can be decomposed as:

$$\mathcal{H} = \mathcal{H}_J \bigoplus J_\perp$$

where, $\mathcal{H}_J = \{g : J(g) \in (0, \infty)\}$, is a $RKHS$ with a square norm J, and $J_\perp = \{g : J(g) = 0\}$. Denote the $\mathcal{RK}$ of $\mathcal{H}_J$ as R_J. Observe that the SSDE (smoothing splines density estimate) depends on the data X_i on the domain $\mathcal{X}$, the reproducing kernel R_J and the null space $J_\perp$.

Note that the space $\mathcal{H}$ is usually infinite dimensional and the minimizer in $\mathcal{H}$ is in general not easy to compute. An attempt to solve this problem was proposed by Gu and Qiu (1993), where the minimizer is calculated in an adaptive finite dimensional space $\mathcal{H}_n = J_\perp \bigoplus \mathcal{H}_J^n$, where $\mathcal{H}_J^n = \{R_J(X_i, \cdot), i = 1, \ldots, n\}$ with R_J the $\mathcal{RK}$ of $\mathcal{H}_J$.

Using the notation of Gu (1993), let $\xi_i = R_J(X_i, \cdot)$, and let $\{\phi\}_{n=1}^M$ be a basis for $J_\perp$. Thus, any function in $\mathcal{H}_n$ can be written as,

$$g = \sum_{i=1}^n c_i \xi_i + \sum_{k=1}^M d_k \phi_k = \boldsymbol{\xi}^T \boldsymbol{c} + \boldsymbol{\phi}^T \boldsymbol{d}$$

where $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vector functions and $\boldsymbol{c}$ and $\boldsymbol{d}$ are vector of coefficients. Consequently, the variational problem (11.1) becomes for a fixed $\lambda > 0$,

$$A_\lambda(\boldsymbol{c},\boldsymbol{d}) = \frac{-\mathbf{1}^T}{n}(Q\boldsymbol{c} + S\boldsymbol{d}) + \log \int_{\mathcal{X}} \exp(\boldsymbol{\xi}^T \boldsymbol{c} + \boldsymbol{\phi}^T \boldsymbol{d}) + \frac{\lambda}{2}\boldsymbol{c}^T Q \boldsymbol{c} \ ,$$

where Q is $n \times n$ matrix with (i,j)th entry $\xi_i(X_i) = R_J(X_i, X_j)$ and S is $n \times M$ matrix with (i,j)th entry $\phi_k(X_i)$.

A standard technique to find the minimizer of $A_\lambda(\boldsymbol{c},\boldsymbol{d})$ is to apply the Newton-Raphson iteration method (Wahba, 1990). Let $\mu_g(h) = \int he^g / \int e^g$, $V_g(f, h) = \mu_g(fh) - \mu_g(f)\mu_g(h)$ and $V_g(h) = V_g(h, h)$. Take $\tilde{g} = \boldsymbol{\xi}^T \boldsymbol{c} + \boldsymbol{\phi}^T \boldsymbol{d}$ as the current iterate of g. Then it can be shown, (Gu, 1993) that the Newton updating equation becomes:

$$\begin{pmatrix} V_{\tilde{g}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T) + \lambda Q & V_{\tilde{g}}(\boldsymbol{\xi}, \boldsymbol{\phi}^T) \\ V_{\tilde{g}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T) & V_{\tilde{g}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T) \end{pmatrix} \begin{pmatrix} \boldsymbol{c} - \tilde{\boldsymbol{c}} \\ \boldsymbol{d} - \tilde{\boldsymbol{d}} \end{pmatrix} = \begin{pmatrix} Q\frac{\mathbf{1}}{n} - \mu_{\tilde{g}}(\boldsymbol{\xi}) - \lambda Q\tilde{\boldsymbol{c}} \\ S^T\frac{\mathbf{1}}{n} - \mu_{\tilde{g}}(\boldsymbol{\phi}) \end{pmatrix} \quad (11.2)$$

The performance of a smoothing spline estimate depends strongly on the choice of the smoothing parameter λ. Gu (1993) proposed an iteration scheme that updates λ and g jointly according to a performance estimate. The performance is measured by loss function $L(g, g_0) = \mu_{g_0}(g_0 - g) + \mu_g(g - g_0)$, the symmetrized Kullback-Leibler distance between an estimate $g = \log f$ and the true density $g_0 = \log f_0$.

Since L is not easy to compute and it depends on g_0, approximation are made and the minimization goes over a proxy of L namely,

$$\hat{L}_{\tilde{g}}(g, \tilde{g}_0) = V_{\tilde{g}}(g)/2 - V_{\tilde{g}}(g, \tilde{g}) + \mu_{\tilde{g}}(g) - \mu_{\tilde{g}_0}(g),$$

with an update $\tilde{g}$ using (11.2) (with a variable λ), and $\tilde{g}_0$ is an approximation by sample means of the true density g_0. See Gu (1993) for details of this approximation.

The algorithm to carry out the performance-oriented iteration is of the order $O(n^3)$ and it can be used "easily" to estimate multivariate densities. Simulations (Dias, 1993) have shown that when the data have too much structure (multi-modal, large numbers of bumps), Gu's algorithm has a good performance, although the computational cost is extremely big for large data sets (> 200).

12   Logspline density estimation

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

Consider an increasing sequence of knots $\{t_j\}_{j=1}^{K}$, $K \geq 4$, in $\mathbb{R}$. Denote by $\mathcal{S}_0$ the set of real functions such that s is a cubic polynomial in each interval of the form $(-\infty, t_1], [t_1, t_3], \ldots, [t_K, \infty)$. Elements in $\mathcal{S}_0$ are the well-known cubic splines with knots at $\{t_j\}_{j=1}^{K}$. (Properties of splines are given in the next chapter.) Notice that $\mathcal{S}_0$ is a $(K + 4)$-dimensional linear space. Now, let $\mathcal{S} \subset \mathcal{S}_0$ such that the dimension of $\mathcal{S}$ is K with functions $s \in \mathcal{S}$ linear on $(-\infty, t_1]$ and on $[t_K, \infty)$. Thus, $\mathcal{S}$ has a basis of the form $1, B_1 \ldots, B_{K-1}$, such that B_1 is linear function with negative slope

on $(-\infty, t_1]$ and $B_2, \ldots, B_{K-1}$ are constant functions on the same interval. Similarly, B_{K-1} is linear function with positive slope on $[t_K, \infty)$ and $B_1, \ldots, B_{K-2}$ are constant on the interval $[t_K, \infty)$ (Kooperberg and Stone, 1991).

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

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

and

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

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

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

The log-likelihood function $L(\boldsymbol{\theta})$ is strictly concave and hence the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is unique, if it exists. We refer to $\hat{f} = f(\cdot, \hat{\boldsymbol{\theta}})$ as the *logspline density estimate*. Note that the estimation of $\hat{\boldsymbol{\theta}}$ makes logspline procedure not essentially non-parametric. Thus, estimation of $\boldsymbol{\theta}$ by Newton-Raphson, together with small numbers of basis function necessary to estimate a density, make the logspline algorithm extremely fast when it is compared with Gu's algorithm for smoothing spline density estimation, (Gu, 1993).

In the Logspline approach the number of knots is the smoothing parameter. That is, too many knots leads to a noisy estimate while too few knots gives a very smooth curve. Based on their experience of fitting logspline models, Kooperberg and Stone provide a table with the number of knots based on the number of observations. No indication was found that the number of knots takes in consideration the structure of the data (number of modes, bumps, asymmetry, etc.). However, an objective

34

criterion for the choice of the number of knots, *Stepwise Knot Deletion*, is included in the logspline procedure.

For $1 \leq j \leq p$, let B_j be a linear combination of a truncated power basis (definition on page 39),

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

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

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

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

It would be interesting to have an algorithm which combines the low computational cost of logsplines (due to B-splines and the estimation of their coefficients) and the performance of the automatic smoothing parameter selection developed by Gu (1993).
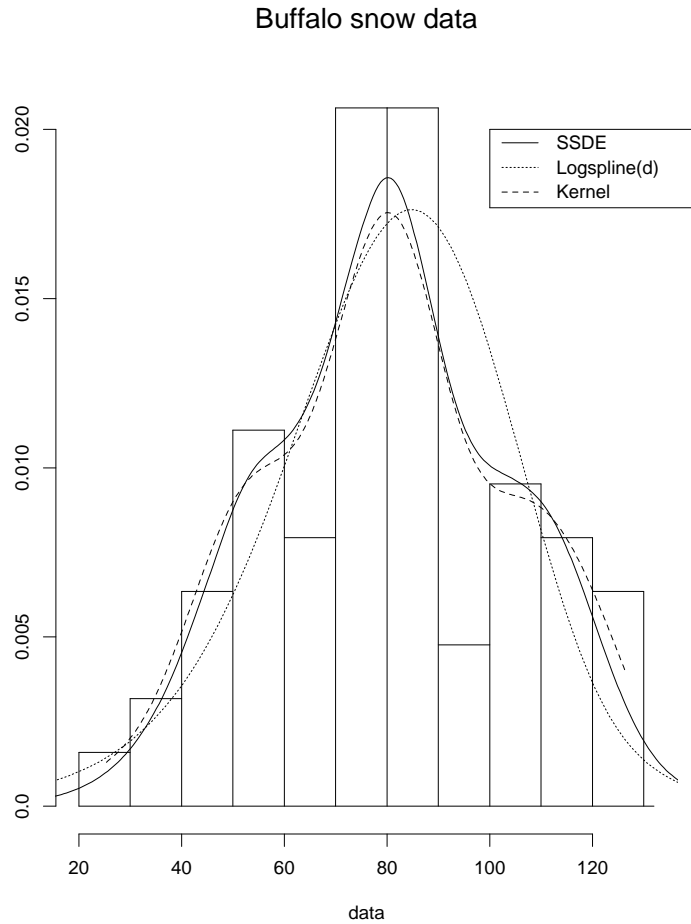
Figure 12.7: Histogram, SSDE, Kernel and Logspline density estimates

The figure (12.7) give us a comparison of four different methods of density estimation, histogram, smoothing spline density estimation(SSDE), logspline and kernel. The data is the well known Buffalo, NY, snowfall and Logspline(d) stands for logspline with deletion procedure. Even though the smoothing parameters of those four methods are not comparable we present them just as information. The smoothing parameters are, 7 knots for logspline, kernel bandwidth equal to 8 for kernel, histogram bandwidth equal to 10 for the histogram and $\lambda = 10^{-5.18}$ (Gu, 1993). Moreover, in figure (12.7), we see that logspline method, with 7 knots, the estimates provided by SSDE and kernel methods are very similar while the Logspline method

produces a smoother density estimate. Since we do not know the underlying density we are not able to say logspline provides, for this case, a poor estimate. But based on the histogram, SSDE and kernel estimates we tend to believe that logspline estimate is not the most appropriate estimate.

13 Penalized Log-likelihood for H-splines Estimation

Recall the problem of estimating an unknown density f based on the observations $X_1, \ldots, X_n$, using the penalized maximum likelihood method. Now, take the 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 solution of the variational problem in (11.1). That is, find the minimizer of

$$-\frac{1}{n} \sum_{i=1}^{n} g(X_i) + \log \int_{\mathcal{X}} e^g + \frac{\lambda}{2} J(g) \tag{13.1}$$

Now, by taking $J(g) = \int (g'')^2$ we have as solution of (13.1) a cubic spline with knots at every data point $x_1, \ldots, x_n$.

As described in the previous section we have good computational reasons to write the solution of (13.1) as a linear combination of B-splines. Actually, we are going to make use of the basis functions as in the Definition (9.4).

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 (13.1) significantly ?

- How good would such an approximation be ?

Dias (1998) gave reasonable answers to those questions by first noticing that the basis functions $B_i(x)$ given in Definition (9.4) form a basis for $\mathcal{N}\mathcal{S}_4$. It is possible to verify that if $x_0 \leq x \leq x_K$, then $0 \leq B_i(x) \leq 1$, $\sum_{i=2}^{K-2} B_i(x) \leq 1$, and $B_0(x_0) = B_1(x_0) = B_{K-1}(x_K) = B_K(x_K) = 1$. See for details Greville (1969). Observe that $(x_{i+2} - x_{i-2})M[x; x_{i-2}, \dots, x_{i+2}]$, for $i = 3, \dots, K-2$, are the usual *normalized B-splines*.

Let us assume that a density function f_0 is in $\mathcal{H}$. Then, we approximate g_0, the solution of (13.1), by linear combinations of basis functions that span $\mathcal{H}_K = \{g \in \mathcal{N}\mathcal{S}_4 : g = \sum_{i=1}^{K} c_i B_i\}$, that is:

$$g_0 \approx g = \sum_{i=1}^{K} c_i B_i,$$

where the notation $g_0 \approx g$ means g_0 is approximately equal to g. Later, we measure this approximation by taking the symmetrized Kullback-Leibler distance.

Consider $\mathcal{X}$, the domain of the density function f_0, and the logistic transformation $f_0 = e^{g_0} / \int_{\mathcal{X}} e^{g_0}$. By taking the side condition $\int_{\mathcal{X}} g_0 = 0$, (Gu and Qiu, 1993), we obtain,

$$\int_{\mathcal{X}} g_0(x)dx \approx \int_{\mathcal{X}} g(x)dx \;\; = \;\; \int_{\mathcal{X}} \sum_{j=1}^{K} c_j B_j(x) \qquad (13.2)$$

$$= \;\; \sum_{j=1}^{K} c_j \int_{\mathcal{X}} B_j(x)dx.$$

Letting $p_j = \int_{\mathcal{X}} B_j(x)dx$, we have $\sum_{j=1}^{K} c_j p_j = 0$, or $c_K = -\sum_{j=1}^{K-1} c_j p_j / p_K$.

Therefore for any function g, such that $g \in \mathcal{H}_K$ can be written as

$$
\begin{aligned}
g(x) &= \sum_{j=1}^{K-1} c_j B_j(x) + c_K B_K(x) \\
&= \sum_{j=1}^{K-1} c_j (B_j(x) - \frac{p_j}{p_k} B_K(x)) \\
&= \sum_{j=1}^{K-1} c_j R_j(x) \\
&= \boldsymbol{R}^T \boldsymbol{c}
\end{aligned}
\tag{13.3}
$$

where $\boldsymbol{R} = (R_1, \dots, R_{K-1})^T$ and $R_j = B_j - \frac{p_j}{p_k} B_K$ for $j = 1, \dots, K-1$.

Now the numerical problem becomes to minimize:

$$
A_\lambda(\boldsymbol{c}) = \frac{-\mathbf{1}^T}{n}(Q\boldsymbol{c}) + \log \int_{\mathcal{X}} \exp(\boldsymbol{R}^T \boldsymbol{c}) + (\lambda/2)\boldsymbol{c}^T \Omega \boldsymbol{c}
\tag{13.4}
$$

where Q is $n \times (K-1)$ matrix with entry $R_i(X_j) = R_{ij}$ and $\Omega = \int (R'')(R'')^T$ is the penalty, matrix with entry $\Omega_{ij} = \int R_i'' R_j''$ for $i, j = 1 \dots, K-1$, and $\text{rank}(\Omega) = K-2$, since two of the basis functions are linear.

A standard procedure to minimize $A_\lambda(\boldsymbol{c})$ is to apply Newton-Raphson iteration. For any two densities f and h, define the following:

$\mu_g(h) = \int h e^g / \int e^g$, $V_g(f, h) = \mu_g(fh) - \mu_g(f)\mu_g(h)$ with $V_g(h) = V_g(h, h)$. Write $\tilde{g} = R^T \tilde{c}$ as the current iteration of g. Then

$$
\frac{\partial A_\lambda(c)}{\partial c} \Big|_{\tilde{g}} = -Q^T \mathbf{1}/\text{n} + \mu_{\tilde{R}} + \lambda \Omega \tilde{c},
$$

where $\mu_{\tilde{g}}(R) = \mu_R$. And,

$$
\begin{aligned}
\frac{\partial^2 A_\lambda(c)}{\partial c \partial c^T} \Big|_{\tilde{g}} &= \mu_{\tilde{g}}(R, R^T) - \mu_{\tilde{g}}(R)\mu_{\tilde{g}}(R^T) + \lambda\Omega \\
&= V_{\tilde{g}}(R, R^T) + \lambda\Omega \\
&= V_{R,R} + \lambda\Omega
\end{aligned}
$$

where $V_{\tilde{g}}(R, R^T) = V_{R,R}$.

Hence, the Newton updating equation is:

$$
\begin{aligned}
(V_{R,R} + \lambda\Omega)^T(c - \tilde{c}) &= Q\mathbf{1}/n - \mu_{\tilde{g}}(R) - \lambda\Omega\tilde{c} \\
(V_{R,R} + \lambda\Omega)^T c - (V_{R,R})^T\tilde{c} - (\lambda\Omega)^T\tilde{c} &= Q\mathbf{1}/n - \mu_{\tilde{g}}(R) - \lambda\Omega\tilde{c} \\
(V_{R,R} + \lambda\Omega)^T c &= Q\mathbf{1}/n - \mu_{\tilde{g}}(R) + (V_{R,R})^T\tilde{c}.
\end{aligned}
$$

It is easy to show that $V_{R,R}\tilde{c} = V_{R,g}$ where $V_{R,g} = V_{\tilde{g}}(R, \tilde{g})$. Thus the Newton updating equation becomes:

$$
(V_{R,R} + \lambda\Omega)^T c = Q\mathbf{1}/n - \mu_{\tilde{g}}(R) + V_{R,g}. \tag{13.5}
$$

14   Performance-Oriented Iteration

Given any density f, let $g = \log f$ and consider the Kulback-Leibler measure of the difference between f and f_0, that is,

$$
\begin{aligned}
KL(f, f_0) &= \int (\log f - \log f_0) f \\
&= \int (g - g_0) e^g.
\end{aligned}
$$

Similarly,

$$
KL(f_0, f) = \int (\log f_0 - \log f) f_0 = \int (g_0 - g) e^{g_0}.
$$

Hence, the symmetrized Kulback-Leibler distance between f and f_0 is :

$$
\begin{aligned}
L(g, g_0) &= KL(f, f_0) + KL(f_0, f) \\
&= \int (g_0 - g) e^g + \int (g - g_0) e^{g_0} \\
&= \mu_g(g_0 - g) + \mu_{g_0}(g - g_0).
\end{aligned}
$$

40

Thus, given the data drawn from g_0 our objective is to find an estimate g which delivers a small $L(g, g_0)$. To compute the minimizer of (13.4) for a fixed λ, one iterates (13.5). Note that equation (13.5) defines a class of estimates with variable λ, and a better use of this equation is to start from a current $\tilde{g}$, instead of calculating the next iterate based on any prespecified λ, one may choose λ which delivers a small $L(g, g_0)$ for g among the class of estimates defined by (13.5). Now, compute an update using such a λ. This iteration scheme tries to minimize the loss function $L(g, g_0)$ that is interested, instead of the penalized likelihood score A_λ with a fixed λ which the Newton iteration is after. Gu, pointed out that when such a performance-oriented iteration converges at λ_*, and g_* is apparently the fixed point of the Newton iteration for minimizing A_{λ_*}, and hence is the minimizer of A_{λ_*}.

Gu noticed that such a L is not easily computable for a class of estimates defined by (13.5), even with a known g_0. Thus, approximation is needed. In the same paper Gu showed that a proxy of L is given by

$$L_{\tilde{g}}(g, g_0) = V_{\tilde{g}}(g)/2 - V_{\tilde{g}}(g, \tilde{g}) + \mu_{\tilde{g}}(g) - \mu_{g_0}(g)$$

and then tried to minimize it by calculating an update from $\tilde{g}$ using (13.5) with variable λ. Observe that the first three terms of $L_{\tilde{g}}$ are readily computable, but the fourth needs estimation. For that, define:

$$H = V_{R,R} + \lambda\Omega,$$

with $u_R = Q\mathbf{1}/n - \mu_R + V_{R,g}$ and $v_R = V_{R,g} - \mu_R$.

It can be shown that $c = H^{-1}u_R$ and that the estimate defined by (13.5) has an expression:

$$
\begin{aligned}
g = R^T c &= R^T(H^{-1}u_R) \\
&= R^T H^{-1}(Q\mathbf{1}/n - \mu_R + V_{R,g}) \\
&= R^T H^{-1} Q\mathbf{1}/n + R^T H^{-1} v_R.
\end{aligned}
\tag{14.1}
$$

41

Straightforward calculations give,

$$\mu_{\tilde{g}}(g) - V_{\tilde{g}}(g, \tilde{g}) = -v_R^T H^{-1} u_R \qquad (14.2)$$

$$V_{\tilde{g}}(g) = u_R^T H^{-1} u_R - \lambda u_R^T H^{-1} Q H^{-1} u_R. \qquad (14.3)$$

To estimate $\mu_{g_0}(g)$, the only source of information is the empirical distribution of the data. For the last term of (14.1), sample means simply give (See Gu eq. 3.4), $Q\mathbf{1}/nH^{-1}v_R$. For the first term,

$$\mu_{g_0}(R^T H^{-1} Q\mathbf{1}/n) = \frac{1}{n} \sum_{i=1}^{n} \mu_{g_0}(R^T(X_i) H^{-1} R(X_i)).$$

It can be seen that the problem with the naive sample mean is the use of X_i itself in the estimation of $\mu_{g_0}(R^T(X_i)H^{-1}R(X_i))$. Using the empirical distribution of the remaining $n - 1$ data to estimate $\mu_{g_0}(R^T(X_i)H^{-1}R(X_i))$, we have:

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} R^T(X_j) H^{-1} R(X_i) =$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} R^T(X_j) H^{-1} R(X_i)$$

$$- \frac{1}{n(n-1)} \sum_{i=1}^{n} R^T(X_j) H^{-1} R(X_i)$$

$$= \frac{n}{n-1} (Q\mathbf{1}/n)^T H^{-1} Q\mathbf{1}/n$$

$$- \frac{1}{n(n-1)} trace(QH^{-1}Q). \qquad (14.4)$$

Letting $tr(\cdot) = trace(\cdot)$ and noting that

$$\frac{V_{\tilde{g}}(g)}{2} - \mu_{g_0}(g) = \frac{tr(QH^{-1}Q)}{n(n-1)} - \frac{(Q\mathbf{1}/n)^T H^{-1} Q\mathbf{1}/n}{n-1} - \frac{u_R^T H^{-1} u_R}{2} + v_R^T H^{-1} u_R,$$

we have a proxy of L as:

$$\hat{L}_\lambda(g, g_0) = \frac{tr(QH^{-1}Q)}{n(n-1)} - \frac{(Q\mathbf{1}/n)^T H^{-1} Q\mathbf{1}/n}{n-1} - \frac{u_R^T H^{-1} u_R}{2} - \frac{\lambda u_R^T H^{-1} Q H^{-1} u_R}{2}.$$

$$(14.5)$$

The last equation is similar to the cross-validation score given by (Gu, 1993). And since (15.4) is a cross-validation of $\mu_{g_0}(R^T H^{-1} Q \mathbf{1}/n)$, is natural to think of it as cross-validation score under the context of Hybrid splines. Moreover, the performance-oriented iteration can then be conducted by minimizing (14.5) in each iteration.

15 Examples and Comparisons

In this section we present various examples of the performance of Hybrid Splines when it is compared with other related methods such as Logspline density estimation and Kernel density estimation. The data sets were obtained by simulations either by using ratfor programs based on algorithms found in (Kennedy, Jr and Gentle, 1980) or by making use of S-language (S-plus and R) (Becker, Chambers and Wilks, 1988) routines.
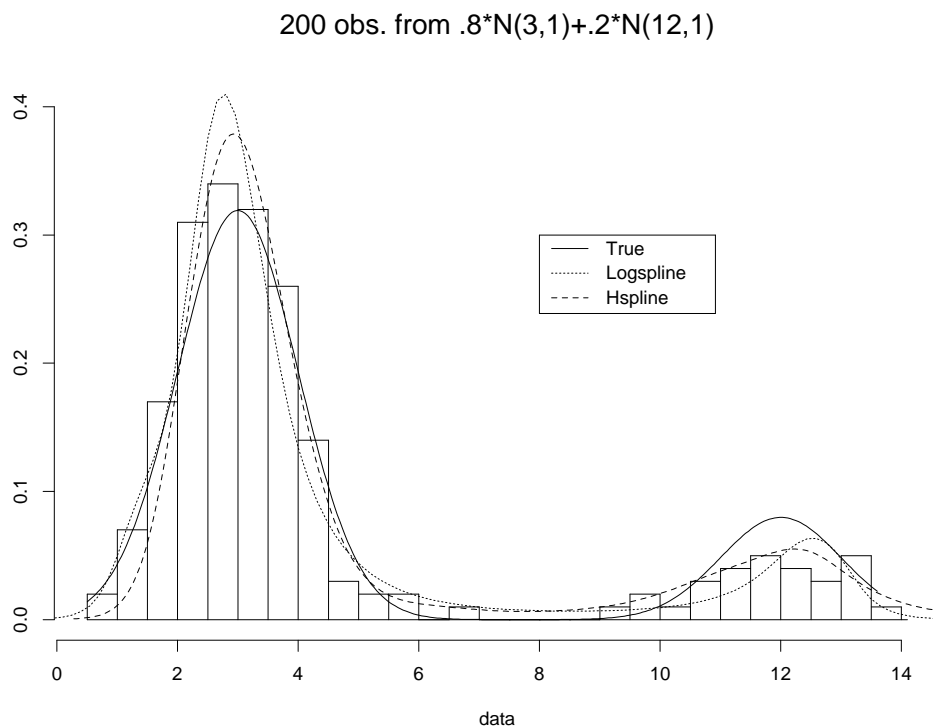
Figure 15.8: True density, Logspline and Hybrid Spline methods are compared

The plot above show us that even under obvious case when two normal distributions with different means are combined, Hybrid Spline, logspline and kernel estimate methods have a very satisfactory performance.
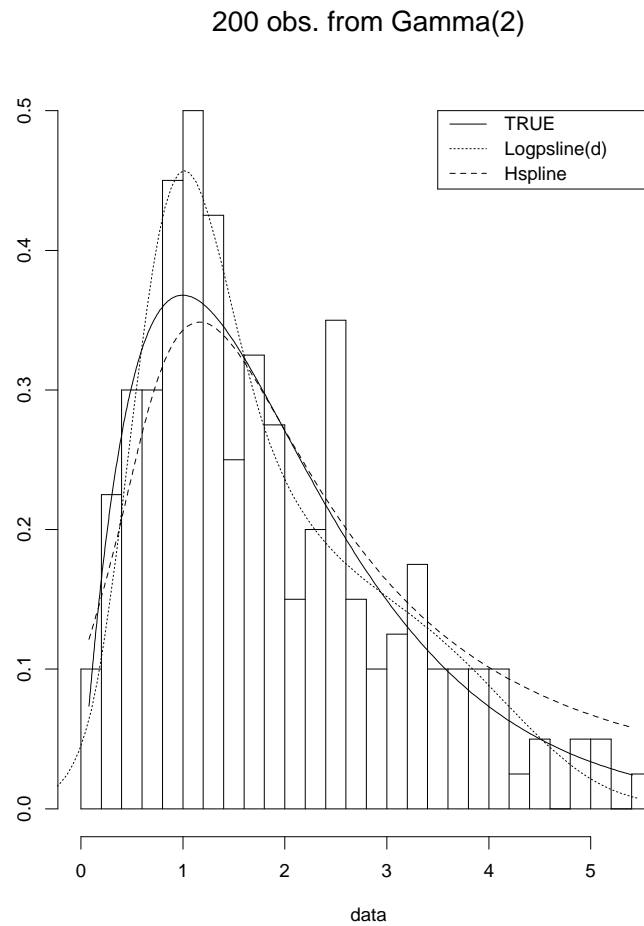
200 obs. from Gamma(2)

Figure 15.9: True density of Gamma(2), estimates given by Logspline, Hybrid Splines

Comparing Hybrid Spline method with Logspline using deletion procedure we see in Figure 4.3 that Hybrid Spline does a better job identifying peaks and valleys of the underlying density. Deletion procedure of Logspline allows it to have the position of the knots that is more significant for the fitting and hence provides the best estimate that Logspline method can produce. Note that logspline adapts using deletion procedure while H-spline adapts using addition. A new version of logspline code incorporates the addition procedure.
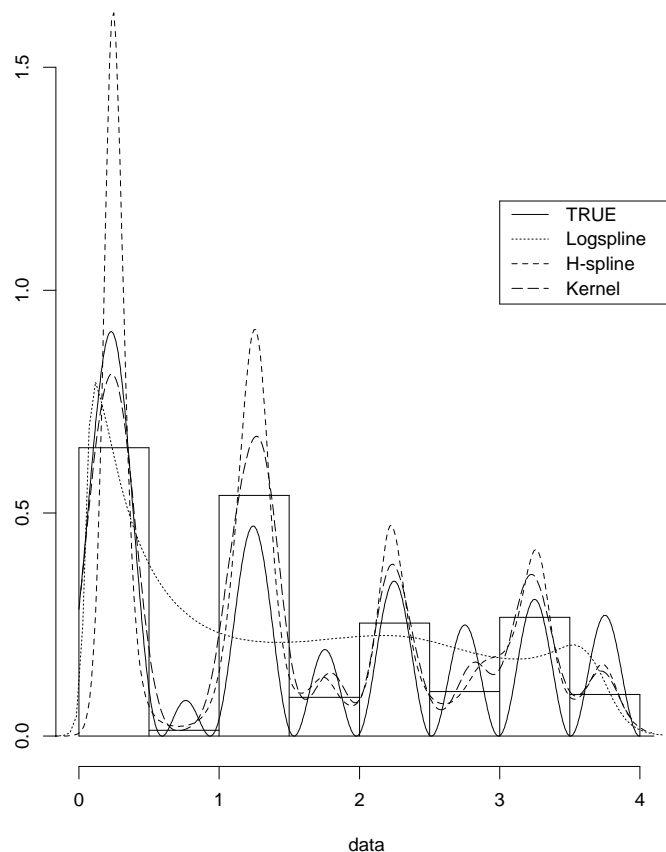
Figure 15.10: True density and estimates given by Logspline, Hybrid spline and kernel methods

From the figure (15.10) clearly kernel method produces the best and fastest estimate of the underlying density. The bandwidth was chosen by eyeball and since that one might have a prior information about the true density, this choice can be well accepted. We notice that Hybrid Spline, again, has a superior performance over the other methods when identifying 5 of the 6 peaks and 6 out of 7 valleys. Hybrid spline has shown in simulations that it is suitable in situation where the data have a lot of structure.

16   Non-parametric regression

The goal of regression curve fitting is to find a relationship between the response variable Y and the predict variable X. If we have n independent measurements $\{(X, Y)\}_{i=1}^{n}$, the regression equation is, in general, described as

$$Y_i = g(x_i) + \varepsilon_i \quad i = 1, \ldots, n,$$

where ε's are uncorrelated random variables with mean zero and independent of Y_i and $g(x_i) = \mathbb{E}[Y_i | X = x_i]$. As an example, let's consider 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 R)
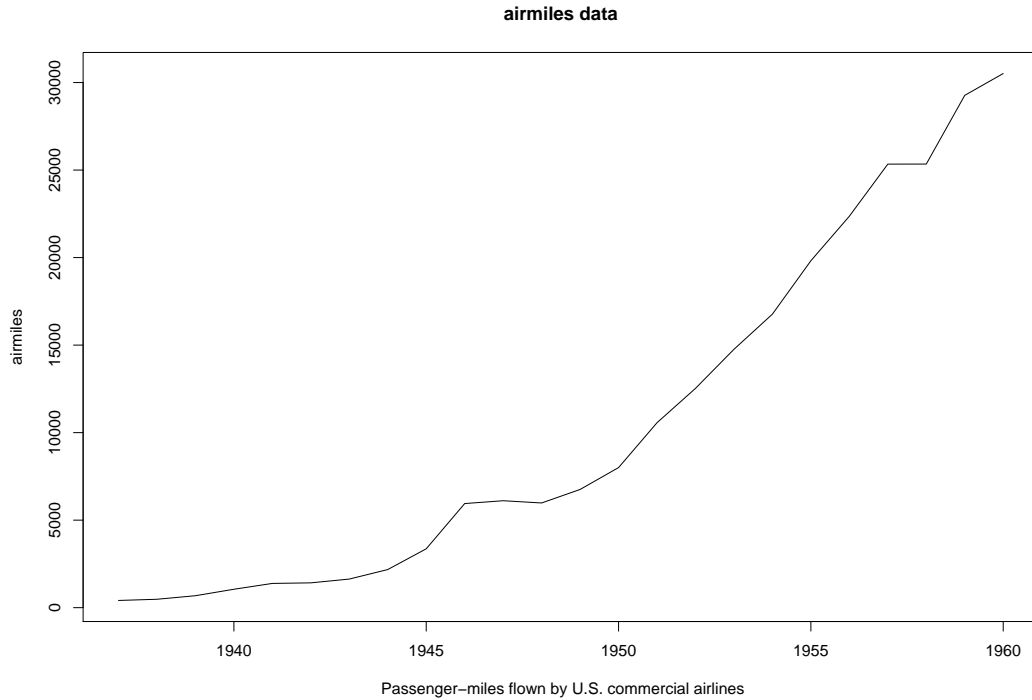
Figure 16.11: A time-series of 24 observations; yearly, 1937-1960.

When we try to approximate the mean response function g, we concentrate on the average dependence of Y on $X = x$. This means that we try to estimate the conditional mean curve

$$g(x) = \mathbb{E}[Y \,|\, X = x] = \frac{\int y f(x,y) dy}{f(x)}, \tag{16.1}$$

where $f(x,y)$ denotes the jointly density of (X,Y) and $f(x)$ the marginal density of X.

16.1 Kernel regression smoothing

Suppose we have i.i.d. observations $\{(X,Y)\}_{i=1}^n$. Recall equation (16.1), we know how to estimate the numerator by using the kernel density estimation method. For the numerator we can estimate the joint density using the multiplicative kernel

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

where, $K_{h_j}(x - \cdot) = h_j^{-1} K(x - \cdot)$, with $j = 1, 2$. It is not difficult to show that

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

Hence a natural estimate of the conditional expectation $g_h(x)$ where $h = h_1$ is the well known Nadaraya-Watson estimate

$$g_h(x) = n^{-1} \frac{\sum_{i=1}^{n} K_h(x - X_i) Y_i}{n^{-1} \sum_{j=1}^{n} K_h(x - X_j)} \tag{16.2}$$

The Nadaraya-Watson $g_h(x)$ is a consistent estimate of the regression curve $g(x)$ if $h \to 0$ and $nh \to \infty$. (See details in Härdle (1990))
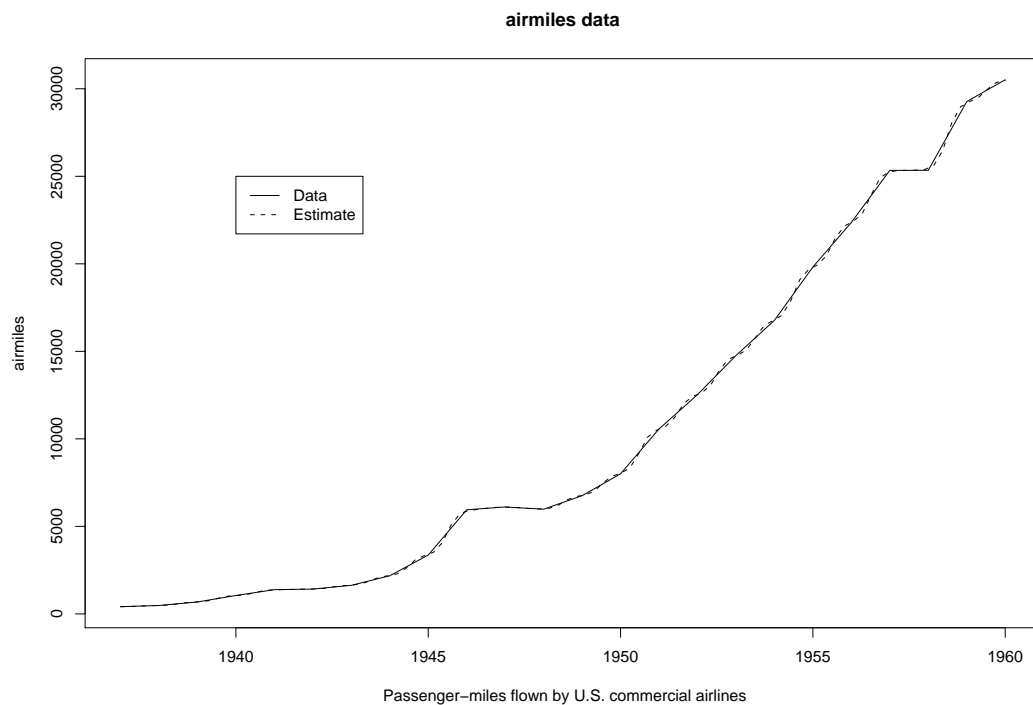

Figure 16.12: kernel smoothing method with bandwidth=1 for airmiles data.

16.2 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. That is,

$$g_K(x) = \frac{1}{n} \sum_{i=1}^{N} W_{Ki}(x) Y_i, \tag{16.3}$$

where,

$$W_{Ki}(x) = \begin{cases} n/K & \text{if } i \in J_x \\ 0 & \text{otherwise,} \end{cases} \tag{16.4}$$

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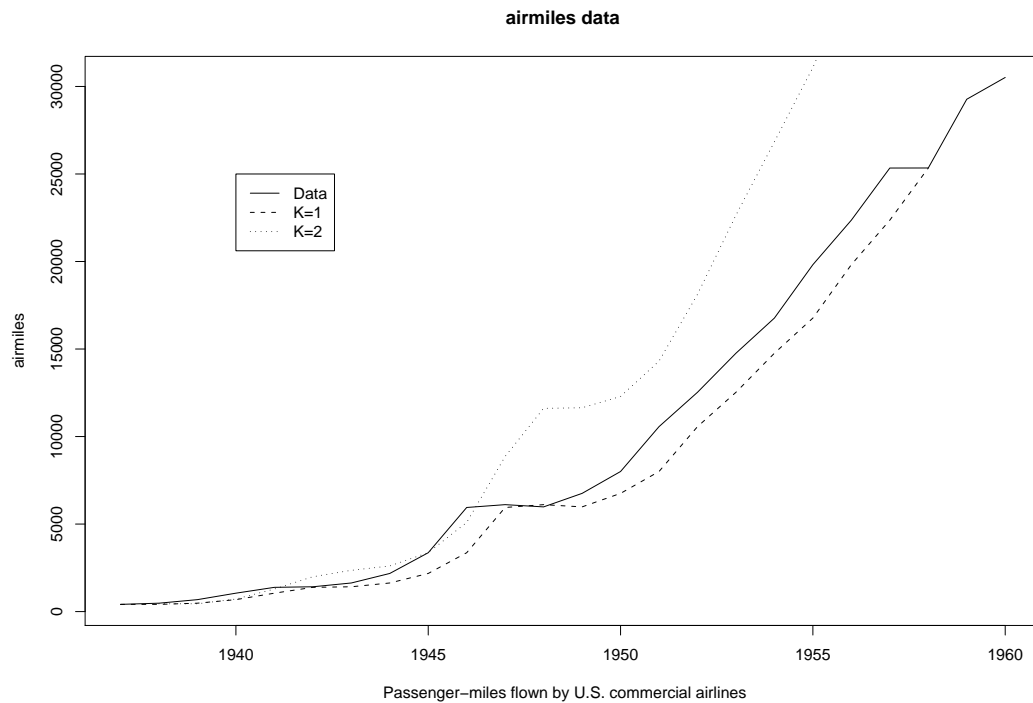

Figure 16.13: Effect of the smoothing parameter K on the K-NN regression estimates.

16.3 Smoothing splines techniques for non-parametric regression

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

$$y_i = L_i g + \epsilon_i$$

where $L_1, \ldots, L_n$ are linear functionals defined on some linear space $\mathcal{H}$ containing g, and $\epsilon_1, \ldots, \epsilon_n$ are measurement errors usually assumed to be independently identically normal distributed with mean zero and unknown variance σ^2. Typically, the 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. Such criterion is the well known penalized least square problem defined as the following: Finding the minimizer of the penalized least square equation which is,

$$A_\lambda(g) = \sum_{i=1}^{n}(y_i - L_i g)^2 + \lambda J(g), \tag{16.5}$$

where $J(g)$ is the penalty term usually taken as $\int (g'')^2$ and λ is the smoothing parameter which controls the trade off between fidelity to the data and smoothness.

It is of interest to estimate the curve g. For this assume that the points $t_1 < t_2 < \ldots, < t_n$ are in the interval $[a, b]$ such that $L_i g = g(t_i)$ and the function $g \in \mathcal{W}_2^2[a, b] = \{g : g' \text{ abs. continuous and } \int (g'')^2 < \infty\}$. Define $\hat{g}$ as the estimate of the curve g so that:

$$\hat{g} = \arg \min_{g \in \mathcal{W}_2^2[a,b]} A_\lambda(g).$$

It is well known that $\hat{g}$ is necessarily a natural cubic spline with knots at t_i (see, for example, Silverman and Green (1994), Wahba (1981) and Craven and Wahba

(1979)). Note that the roughness penalty $\int_a^b (g''(t))^2 dt$ has the property of reducing the problem of choosing g from an infinite-dimensional class of functions to a finite class of functions since $\hat{g}$ can be written as linear combination of basis functions. Although this fact might lead someone to think that the non-parametric regression problem becomes a parametric problem, one notices that the number of parameters can be as large as the number of observations, and there may be difficulties in interpreting a curve or surface g. Moreover, if the number of observations is large, the system of linear equations for exact solution is too expensive to solve.
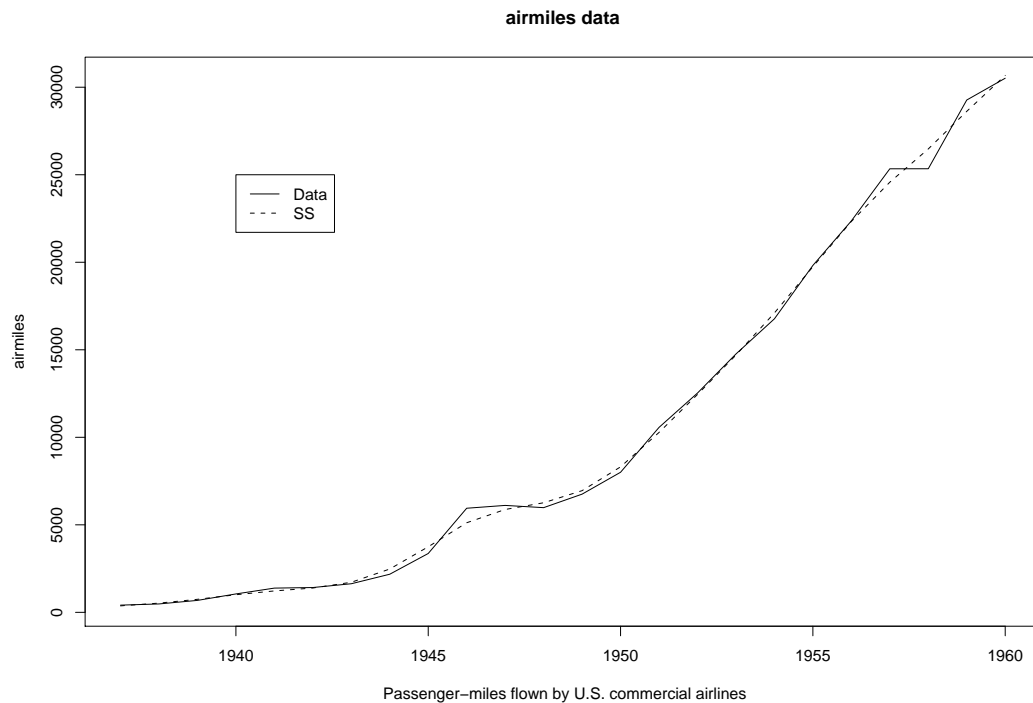

Figure 16.14: Smoothing spline fitting with smoothing parameter obtained by GCV method

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,

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

51

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 $B_{i,j} = B_j(x_i)$, 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 $y(x) = \exp(-x)\sin(\pi x/2)\cos(\pi x) + \varepsilon$ 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 16.15 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.

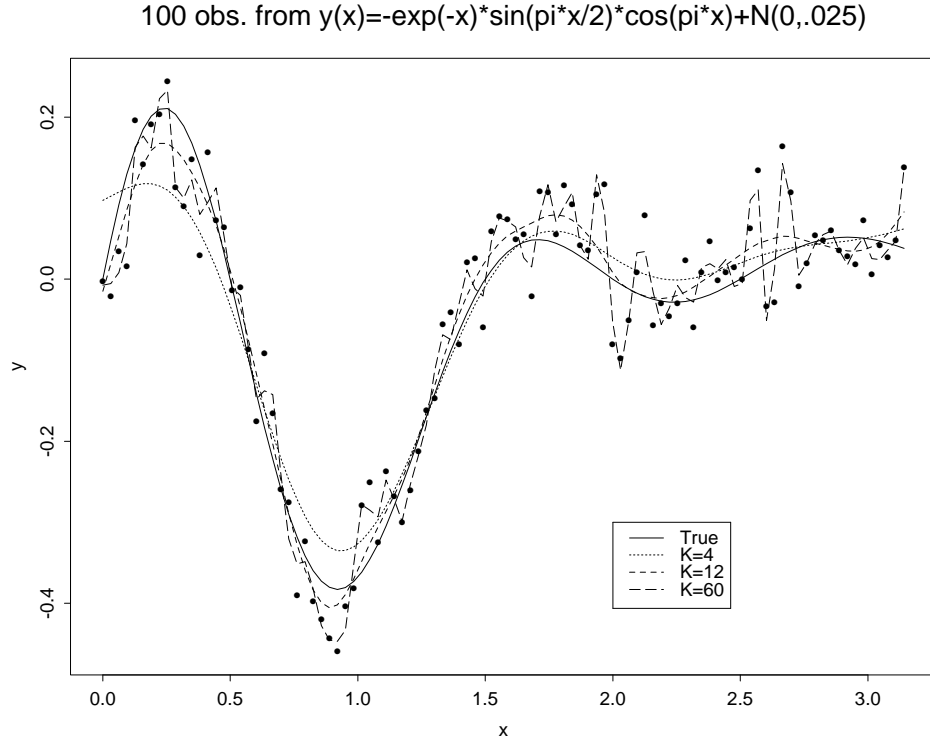100 obs. from y(x)=-exp(-x)*sin(pi*x/2)*cos(pi*x)+N(0,.025)

Figure 16.15: Spline least square fittings for different values of K

16.3.1 The hybrid 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 (Luo and Wahba (1997), Dias (1998) and Dias (1999)) combines ideas from regression splines and smoothing splines methods by finding the number of basis functions and the smoothing parameter iteratively. By taking the penalty term $J(g)$ as $\int (g'')^2$,the point evaluation functionals $\mathcal{L}_i g = g(t_i)$ $\mathbf{y} = (y_1, \dots, y_n)^T$ and $\mathbf{g} = (g(t_1), \dots, g(t_n))^T$, the penalized least square criterion (16.5) becomes,

$$ L_\lambda(g) = ||\mathbf{y} - \mathbf{g}||^2 + \lambda \int (g'')^2, \qquad (16.6) $$

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 $\{B_i\}_{i=1}^{K}$

53

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,

$$L_\lambda^*(\theta) = \|\mathbf{y} - X_K\theta\|_2^2 + \lambda\theta^T\Omega\theta, \tag{16.7}$$

where 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 $(X^TX + \lambda\Omega)\theta_\lambda = X^Ty$. 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.
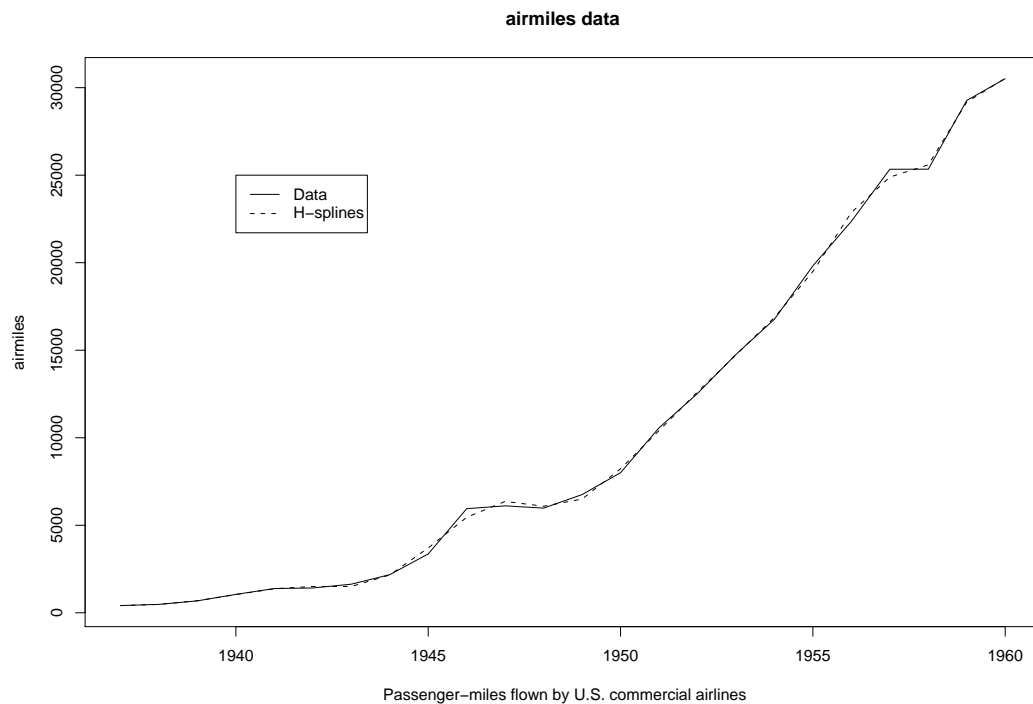

Figure 16.16: H-spline fitting exhibits for airmiles data

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 for small data sets (50 observations) and relatively large variance in the measurement errors.

Comparing with parametric techniques we have, for the non-parametric approach, more flexibility since it allows one to choose the infinity dimensional class of functions that the underlying density belongs. In general, this type of choice depends on the unknown smoothness of the true density. But for the most of the cases one can assume mild restrictions such that a density has an absolutely continuous first derivative and a square integrable second derivative. Nevertheless, non-parametric estimators are less efficient than the parametric ones when a parametric model is valid. For many parametric estimators the mean square error goes to zero with rate of n^{-1}, while non-parametric 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. Consequently, non-parametric estimators are good candidates when one does not know the form of the underlying curve.

References

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

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics*, Wadsworth Publishing Co (Belmont CA).

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).

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

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

Dias, R. (1993). The hybrid spline method for density estimation, *University of Wisconsin-Madison*. Unpublished manuscript.

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.

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

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

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

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

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

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

Kennedy, W. J., Jr and Gentle, J. E. (1980). *Statistical Computing*, Dekker:NY.

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.

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.

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

Rao, B. L. S. P. (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.

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. (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. (1981). Data-based optimal smoothing of orthogonal series density estimates, *Ann. of Statistics* **9**: 146–156.

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