APPLIED SMOOTHING TECHNIQUES

Part 1: Kernel Density Estimation

Walter Zucchini

October 2003

Contents

1 Density Estimation **2**

1.1 Introduction 2

1.1.1 The probability density function 2

1.1.2 Non–parametric estimation of $f(x)$ — histograms 3

1.2 Kernel density estimation 3

1.2.1 Weighting functions 3

1.2.2 Kernels 7

1.2.3 Densities with bounded support 8

1.3 Properties of kernel estimators 12

1.3.1 Quantifying the accuracy of kernel estimators 12

1.3.2 The bias, variance and mean squared error of $\hat{f}(x)$ 12

1.3.3 Optimal bandwidth 15

1.3.4 Optimal kernels 15

1.4 Selection of the bandwidth 16

1.4.1 Subjective selection 16

1.4.2 Selection with reference to some given distribution 16

1.4.3 Cross–validation 18

1.4.4 ”Plug–in” estimator 19

1.4.5 Summary and extensions 19

Chapter 1

Density Estimation

1.1 Introduction

1.1.1 The probability density function

The probability distribution of a continuous–valued random variable X is conventionally described in terms of its probability density function (pdf), $f(x)$, from which probabilities associated with X can be determined using the relationship

$$P(a \leq X \leq b) = \int_a^b f(x)dx \ .$$

The objective of many investigations is to estimate $f(x)$ from a sample of observations $x_1, x_2, ..., x_n$. In what follows we will assume that the observations can be regarded as independent realizations of X.

The parametric approach for estimating $f(x)$ is to assume that $f(x)$ is a member of some parametric family of distributions, e.g. $N(\mu,\, \sigma^2)$, and then to estimate the parameters of the assumed distribution from the data. For example, fitting a normal distribution leads to the estimator

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi}\hat{\sigma}}\ e^{(x-\hat{\mu})/2\hat{\sigma}^2} \ , \quad x \in I\!R \ ,$$

where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$.

This approach has advantages as long as the distributional assumption is correct, or at least is not seriously wrong. It is easy to apply and it yields (relatively) stable estimates.

The main disadvantage of the parametric approach is lack of flexibility. Each parametric family of distributions imposes restrictions on the shapes that $f(x)$ can have. For example the density function of the normal distribution is symmetrical and bell–shaped, and therefore is unsuitable for representing skewed densities or bimodal densities.

1.1.2 Non–parametric estimation of $f(x)$ — histograms

The idea of the non–parametric approach is to avoid restrictive assumptions about the form of $f(x)$ and to estimate this directly from the data. A well–known non–parametric estimator of the pdf is the histogram. It has the advantage of simplicity but it also has disadvantages, such as lack of continuity. Secondly, in terms of various mathematical measures of accuracy there exist alternative non–parametric estimators that are superior to histograms.

To construct a histogram one needs to select a left bound, or starting point, x_0, and the bin width, b. The bins are of the form $[x_0+(i-1)b,\ x_0+ib)$, $i=1,2,...,m$. The estimator of $f(x)$ is then given by

$$\hat{f}(x)=\frac{1}{n}\,\frac{\text{Number of observations in the same bin as } x}{b}$$

More generally one can also use bins of different widths, in which case

$$\hat{f}(x)=\frac{1}{n}\,\frac{\text{Number of observations in the same bin as } x}{\text{Width of bin containing } x}$$

The choice of bins, especially the bin widths, has a substantial effect on the shape and other properties of $\hat{f}(x)$. This is illustrated in the example that follows.

Example 1

We consider a population of 689 of a certain model of new cars. Of interest here is the amount (in DM) paid by the customers for “optional extras”, such as radio, hubcaps, special upholstery, etc. . The top histogram in Figure 1.1 relates to the entire population; the bottom histogram is for a random sample of size 10 from the population.

Figure 1.2 shows three histogram estimates of $f(x)$ for the sample, for different bin widths. Note that the estimates are piecewise constant and that they are strongly influenced by the choice of bin width. The bottom right hand graph is an example of a so–called kernel estimator of $f(x)$. We will be examining such estimations in more detail.

1.2 Kernel density estimation

1.2.1 Weighting functions

From the definition of the pdf, $f(x)$, of a random variable, X, one has that

$$P(x-h<X<x+h)=\int_{x-h}^{x+h} f(t)dt \quad\approx\quad 2hf(x)$$

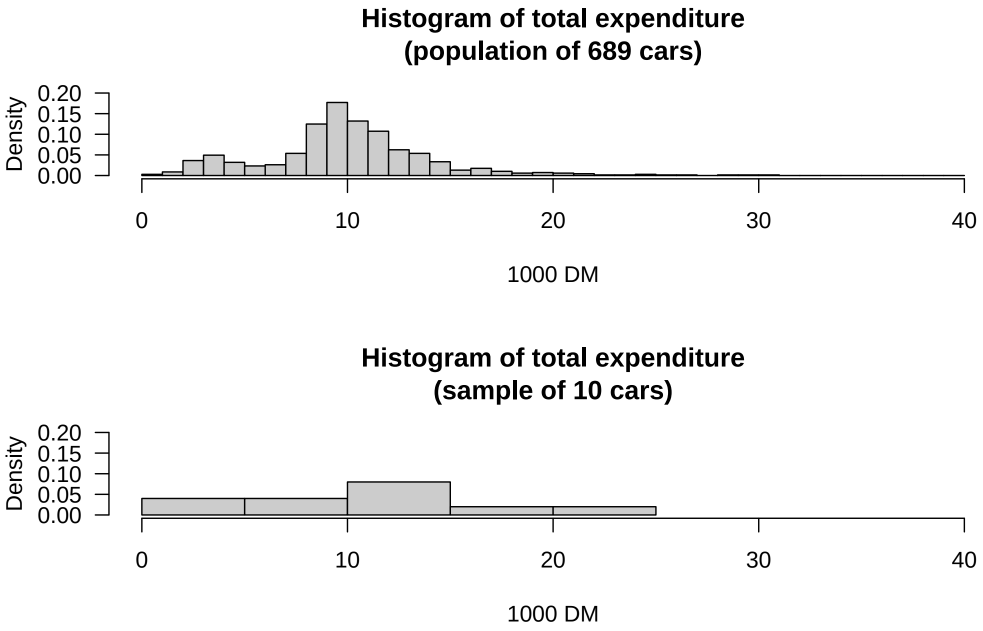

Figure 1.1: Histogram for all cars in the population and for a random sample of size $n = 10$.

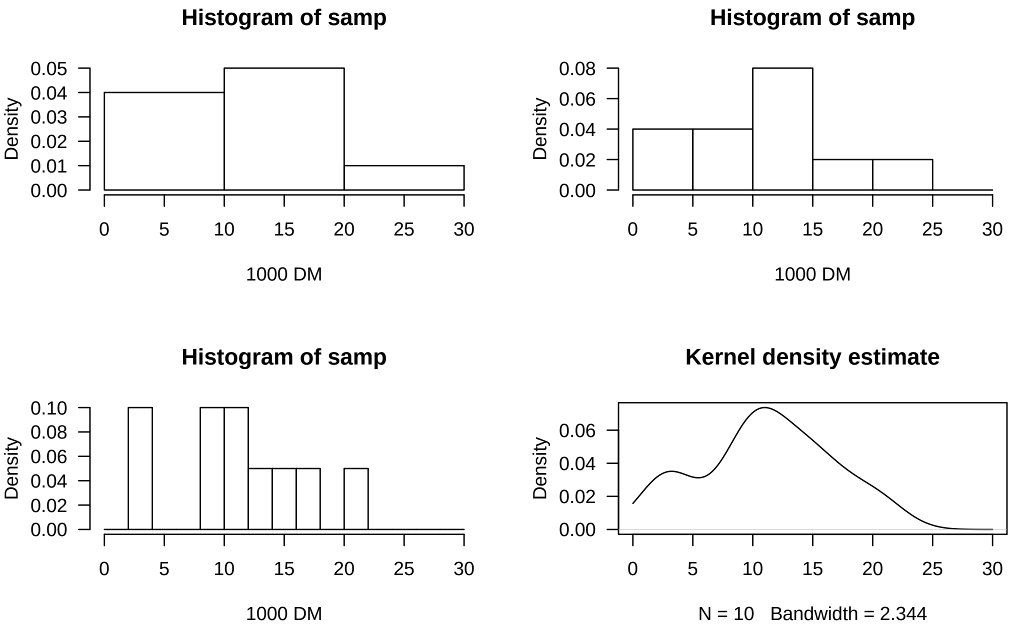

Figure 1.2: Histograms with different bin widths and a kernel estimate of $f(x)$ for the same sample.

and hence

$$f(x) \approx \frac{1}{2h} P(x - h < X < x + h) \tag{1.1}$$

The above probability can be estimated by a relative frequency in the sample, hence

$$\hat{f}(x) = \frac{1}{2h} \frac{\text{number of observations in } (x - h, x + h)}{n} \tag{1.2}$$

An alternative way to represent $\hat{f}(x)$ is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} w(x - x_i, h) \ , \tag{1.3}$$

where $x_1, x_2, ..., x_n$ are the observed values and

$$w(t, h) = \begin{cases} \frac{1}{2h} & \text{for } |t| < h \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

It is left to the reader as an exercise to show that $\hat{f}(x)$ defined in (1.3) has the properties of a pdf, that is $\hat{f}(x)$ is non–negative for all x, and the area between $\hat{f}(x)$ and the x–axis is equal to one.

One way to think about (1.3) is to imagine that a rectangle (height $\frac{1}{2h}$ and width $2h$) is placed over each observed point on the x–axis. The estimate of the pdf at a given point is $1/n$ times the sum of the heights of all the rectangles that cover the point. Figure 1.3 shows $\hat{f}(x)$ for such rectangular “weighting functions” and for different values of h.

We note that the estimates of $\hat{f}(x)$ in Figure 1.3 fluctuate less as the value of h is increased. By increasing h one increases the width of each rectangle and thereby increases the degree of “smoothing”.

Instead of using rectangles in (1.3) one could use other weighting functions, for example triangles:

$$w(t, h) = \begin{cases} \frac{1}{h}(1 - |t|/h) & \text{for } |t| < h \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Again it is left to the reader to check that the resulting $\hat{f}(x)$ is indeed a pdf. Examples of $\hat{f}(x)$ based on the triangular weighting function and four different values of h are shown in Figure 1.4. Note that here too larger values of h lead to smoother estimates $\hat{f}(x)$.

Another alternative weighting function is the Gaussian:

$$w(t, h) = \frac{1}{\sqrt{2\pi}h} \, e^{-t^2/2h^2} \ , \quad -\infty < t < \infty \ .$$

Figure 1.5 shows $\hat{f}(x)$ based on this weighting function for different values of h. Again the fluctuations in $\hat{f}(x)$ decrease with increasing h.

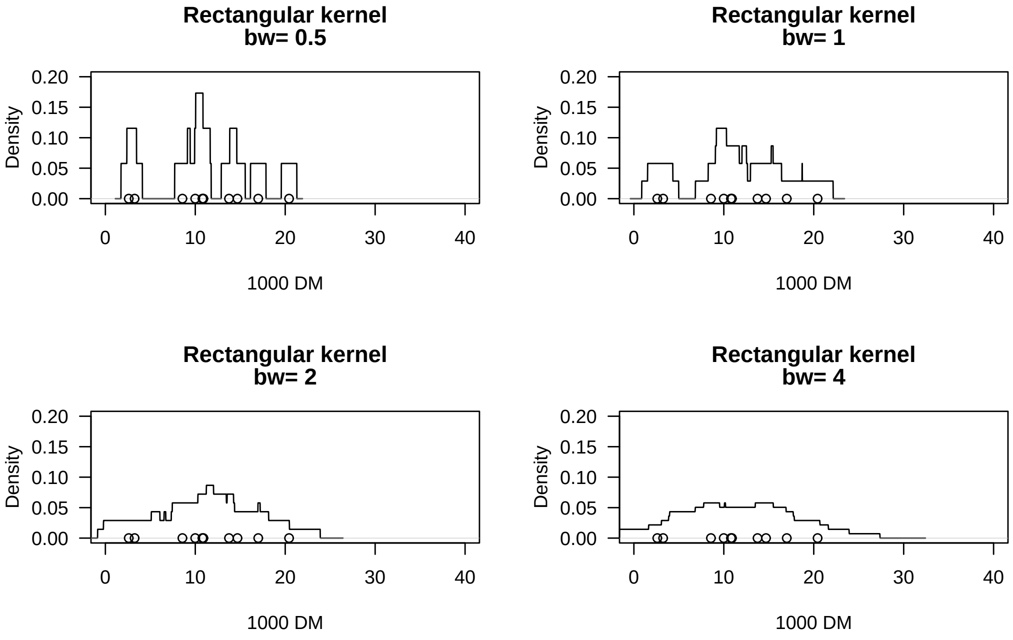

Figure 1.3: Estimates of $f(x)$ for different values of h. The abbreviation bw (short for bandwidth) is used here instead of h.

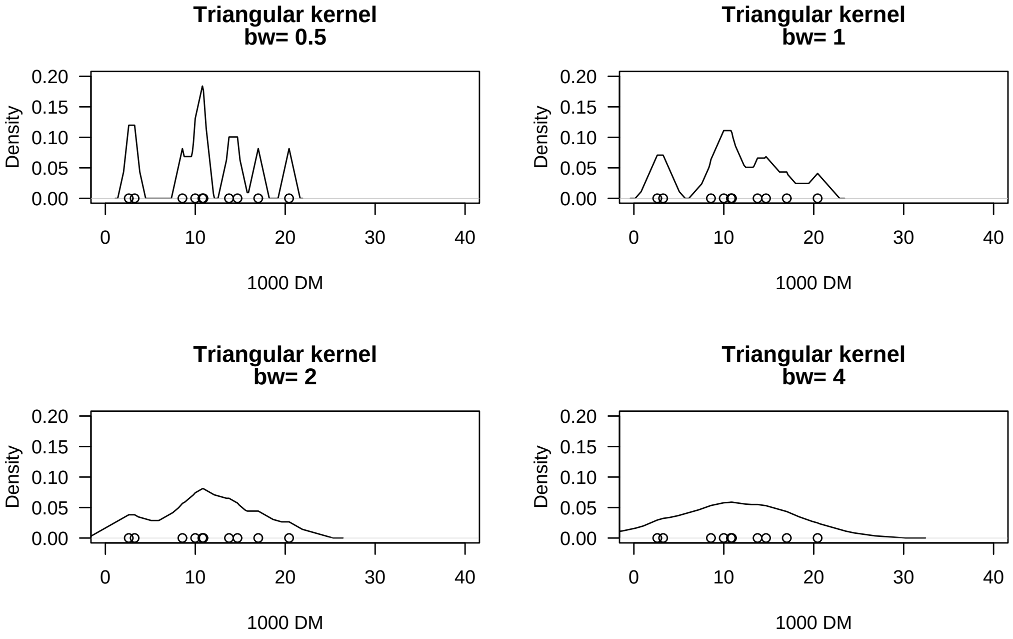

Figure 1.4: Estimates of $f(x)$ based on triangular weighting functions.

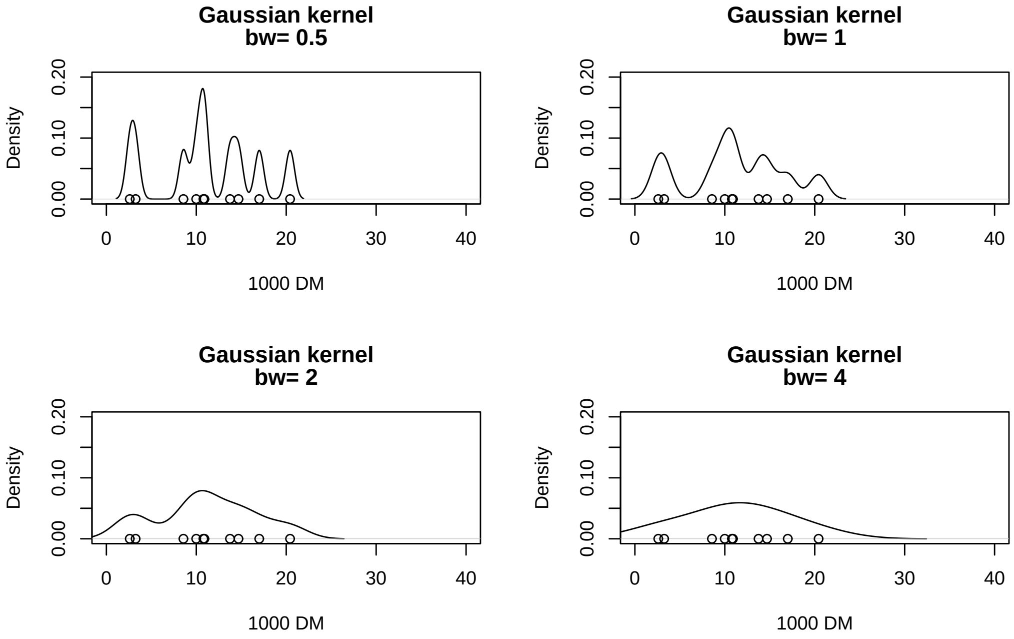

Figure 1.5: Estimates of $f(x)$ based on Gaussian weighting functions.

1.2.2 Kernels

The above weighting functions, $w(t, h)$, are all of the form

$$w(t, h) = \frac{1}{h} \, K\left(\frac{t}{h}\right) \ , \tag{1.4}$$

where K is a function of a single variable called the *kernel*.

A kernel is a standardized weighting function, namely the weighting function with $h = 1$. The kernel determines the *shape* of the weighting function. The parameter h is called the *bandwidth* or *smoothing constant*. It determines the amount of smoothing applied in estimating $f(x)$. Six examples of kernels are given in Table 1.

Kernel	$K(t)$	*Efficiency (exact and to 4 d.p.)*
Epanechnikov	$\frac{3}{4}(1-\frac{1}{5}t^2)/\sqrt{5}$ for $\|t\|<\sqrt{5}$ 0 otherwise	1
Biweight	$\frac{15}{16}(1-t^2)^2$ for $\|t\|<1$ 0 otherwise	$\left(\frac{3087}{3125}\right)^{1/2} \approx 0.9939$
Triangular	$1-\|t\|$ for $\|t\|<1$, 0 otherwise	$\left(\frac{243}{250}\right)^{1/2} \approx 0.9859$
Gaussian	$\frac{1}{\sqrt{2\pi}}\, e^{-(1/2)t^2}$	$\left(\frac{36\pi}{125}\right)^{1/2} \approx 0.9512$
Rectangular	$\frac{1}{2}$ for $\|t\|<1$, 0 otherwise	$\left(\frac{108}{125}\right)^{1/2} \approx 0.9295$

Table 1.1: Six kernels and their efficiencies (that will be discussed in section 1.3.4).

In general any function having the following properties can be used as a kernel:

$$\text{(a)} \int K(z)dz = 1 \qquad \text{(b)} \int zK(z)dz = 0 \qquad \text{(c)} \int z^2K(z)dz := k_2 < \infty \qquad (1.5)$$

It follows that any symmetric pdf is a kernel. However, non–pdf kernels can also be used, e.g. kernels for which $K(z) < 0$ for some values of z. The latter type of kernels have the disadvantage that $\hat{f}(x)$ may be negative for some values of x.

Kernel estimation of pdfs is charactized by the kernel, K, which determines the shape of the weighting function, and the bandwidth, h, which determines the "width" of the weighting function and hence the amount of smoothing. The two components determine the properties of $\hat{f}(x)$. Considerable research has been carried out (and continues to be carried out) on the question of how one should select K and h in order to optimize the properties of $\hat{f}(x)$. This issue will be discussed in the sections that follow.

1.2.3 Densities with bounded support

In many situations the values that a random variable, X, can take on is restricted, for example to the interval $[0,\ \infty)$, that is $f(x) = 0$ for $x < 0$. We say that the *support* of $f(x)$ is $[0,\ \infty)$. Similarly if X can only take on values in the interval (a, b) then $f(x) = 0$ for $x \notin (a, b)$; the support of $f(x)$ is (a, b).

In such situations it is clearly desirable that the estimator $\hat{f}(x)$ has the same support as $f(x)$. Direct application of kernel smoothing methods does not guarantee this property and so they need to be modified when $f(x)$ has bounded support. The simplest method of solving this problem is use a transformation. The idea is to estimate the pdf of a transformed random variable $Y = t(X)$ which has unbounded support. Suppose that the pdf of Y is given by $g(y)$. Then the relationship between f and g is given by

$$f(x) = g(t(x))t'(x) \ . \tag{1.6}$$

One carries out the following steps:

(a) Transform the observations $y_i = t(x_i), \ i = 1, 2, ..., n.$

(b) Apply the kernel method to estimate the pdf $g(y)$.

(c) Estimate $f(x)$ using $\hat{f}(x) = \hat{g}(t(x))t'(x)$.

Example 2

Suppose that $f(x)$ has support $[0, \ \infty)$.
A simple transformation $t : [0, \ \infty) \to (-\infty, \ \infty)$ is the log–transformation, i.e. $f(x) = \log(x)$. Here $t'(x) = \frac{d\log(x)}{dx} = \frac{1}{x}$ and so

$$\hat{t}(x) = \hat{g}(\log(x))\frac{1}{x} \tag{1.7}$$

The resulting estimator has support $[0, \ \infty)$. Figure 1.6 provides an illustration for this case for the sample considered in Example 1.

(a) The graph on the top left gives the estimated density $\hat{f}(x)$ obtained without restrictions on the support. Note that $\hat{f}(x) \geq 0$ for some $x < 0$.

(b) The graph on the top right shows a modified version of $\hat{f}(x)$ obtained in (a), namely

$$\hat{f}_c(x) = \begin{cases} \frac{\hat{f}(x)}{\int\limits_0^\infty \hat{f}(x)dx} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{1.8}$$

Here $\hat{f}_c(x)$ is set equal to zero for $x < 0$ and the $\hat{f}(x)$ is rescaled so that the area under the estimated density equals one.

(c) The bottom graph shows a kernel estimator of $g(y)$, that is the density of $Y = \log(X)$.

(d) The bottom right graph shows the transformed estimator $\hat{f}(x)$ obtained via $\hat{g}(y)$.

Example 3

Suppose that the support of $f(x)$ is (a, b). Then a simple transformation $t : (a, b) \to (-\infty, \ \infty)$ is $f(x) = \log\left(\frac{x-a}{b-x}\right)$. Here $t'(x) = \frac{1}{x-a} + \frac{1}{b-x}$ and so

$$\hat{t}(x) = \hat{g}\left(\log\left(\frac{x-a}{b-x}\right)\right)\left(\frac{1}{x-a} + \frac{1}{b-x}\right) \tag{1.9}$$

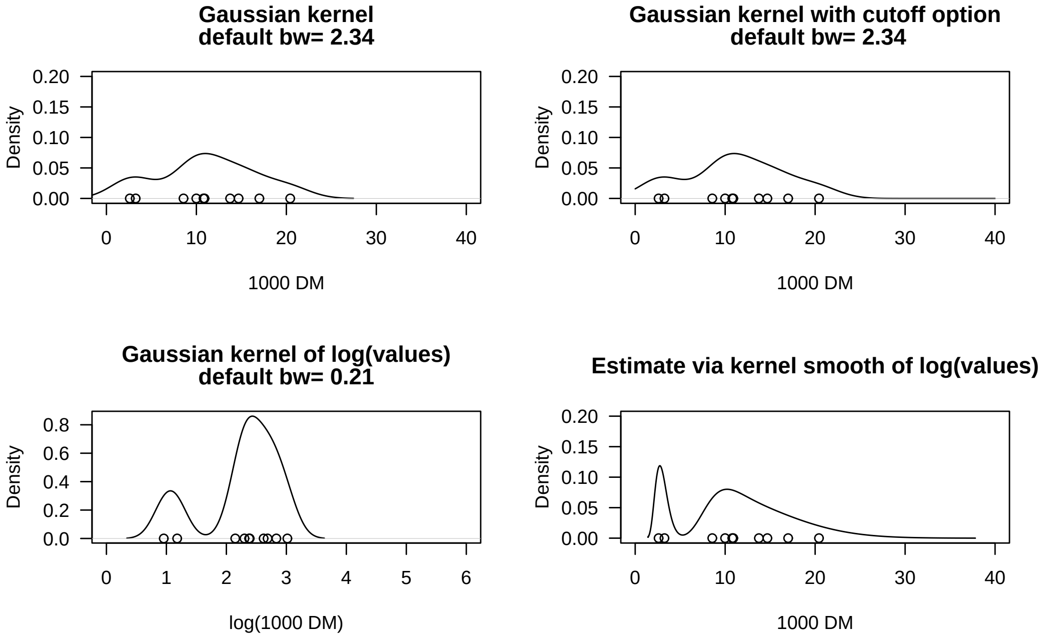

Figure 1.6: Kernel estimates of pdf with support $[0, \infty)$.

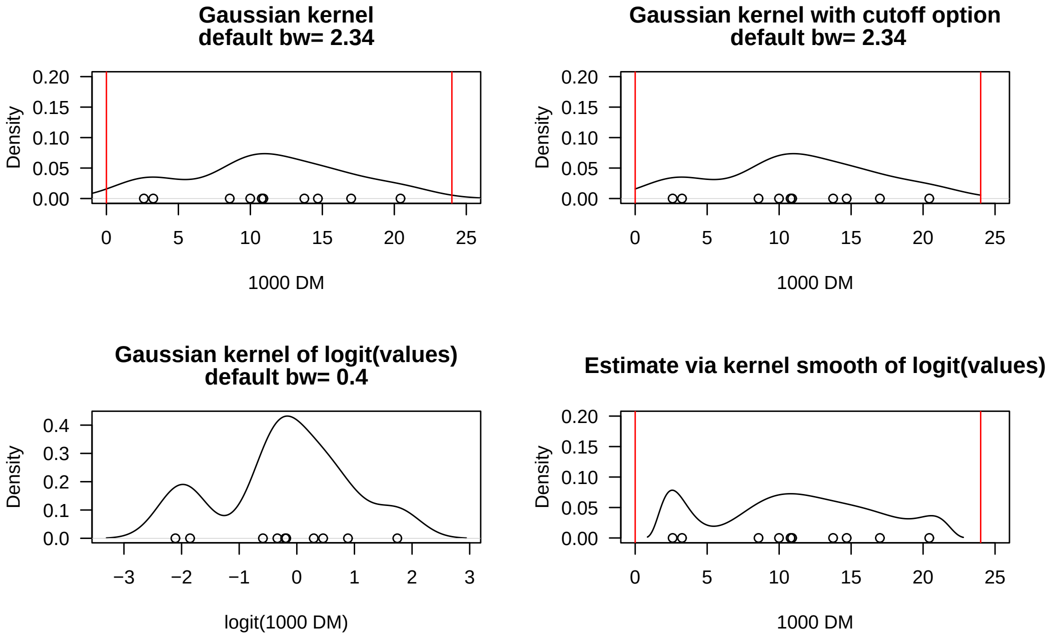

Figure 1.7: Kernel estimates of a pdf with support (0,25).

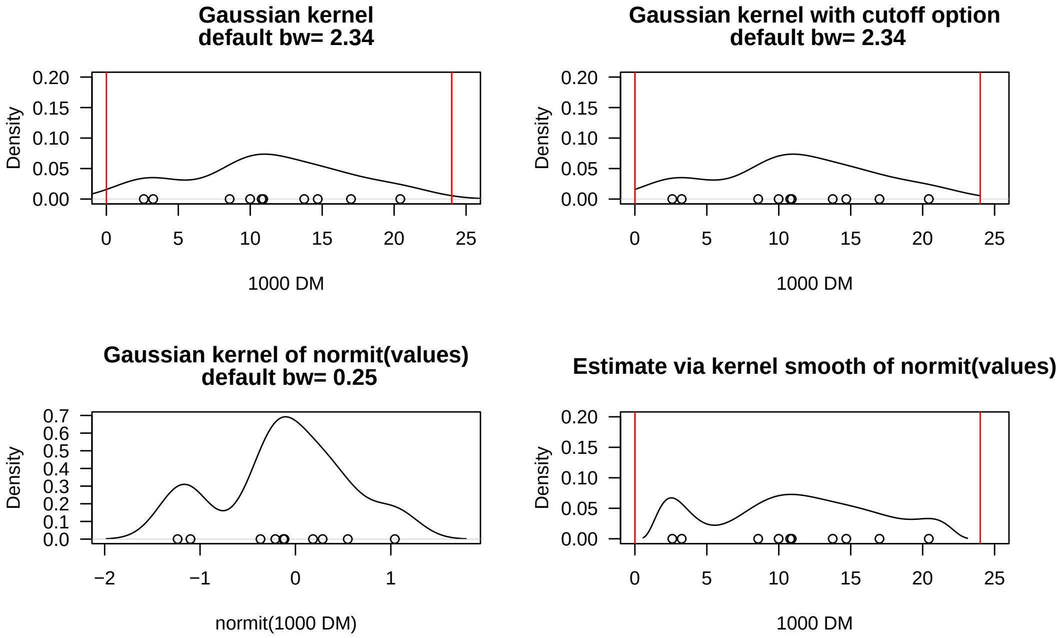

Figure 1.8: Application of the normit transformation for $\hat{f}(x)$ with support (0,25).

Figure 1.7 provides an illustration of this case with $a = 0$ and $b = 25$.

The four figures shown are analogous to those in Example 2 but with

$$\hat{f}_c(x) = \begin{cases} \frac{\hat{f}(x)}{\int_a^b \hat{f}(x)dx} & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

for the graph on the top right.

Example 4

As an alternative to the transformation in Example 2 one can use the inverse of some probability distribution function, such as the normal; e.g. $f(x) = \Phi^{-1}\left(\frac{x-a}{b-a}\right)$, where Φ is the distribution function of the standard normal. Here too $t : (a, b) \to (-\infty,\ \infty)$ and the estimator is

$$\hat{f}(x) = \begin{cases} \hat{g}\left(\Phi^{-1}\left(\frac{x-a}{b-a}\right)\right) \frac{b-a}{\varphi\left(\frac{x-a}{b-a}\right)} & \text{for } a < x < b\ , \\ 0 & \text{otherwise}\ , \end{cases} \tag{1.11}$$

where $\varphi(x)$ is the pdf of a standard normal distribution.

The application of this transformation is illustrated in Figure 1.8 in which the four figures are analogous to those in the previous two examples.

The above three examples illustrate that the transformation procedure can lead to a considerable change in the appearance of the estimate $\hat{f}(x)$. By applying kernel smoothing to the transformed values one is, in effect, applying a different kernel at each point in the estimation of $f(x)$.

1.3 Properties of kernel estimators

1.3.1 Quantifying the accuracy of kernel estimators

There are various ways to quantify the accuracy of a density estimator. We will focus here on the mean squared error (MSE) and its two components, namely bias and standard error (or variance). We note that the MSE of $\hat{f}(x)$ is a function of the argument x:

$$\begin{aligned}\mathrm{MSE}(\hat{f}(x)) &= E(\hat{f}(x)-f(x))^2 \qquad (1.12)\\ &= (E\,\hat{f}(x)-f(x))^2 + E(\hat{f}(x)-E\hat{f}(x))^2\\ &= \mathrm{Bias}^2(\hat{f}(x)) + Var(\hat{f}(x))\end{aligned}$$

A measure of the global accuracy of $\hat{f}(x)$ is the mean integrated squared error (MISE)

$$\begin{aligned}\mathrm{MISE}(\hat{f}) &= E\int_{-\infty}^{\infty}(\hat{f}(x)-f(x))^2dx \qquad (1.13)\\ &= \int_{-\infty}^{\infty}\mathrm{MSE}(\hat{f}(x))dx\\ &= \int_{-\infty}^{\infty}\mathrm{Bias}^2(\hat{f}(x))dx + \int_{-\infty}^{\infty}Var(\hat{f}(x))dx\end{aligned}$$

We consider each of these components in term.

1.3.2 The bias, variance and mean squared error of $\hat{f}(x)$

$$\begin{aligned}E(\hat{f}(x)) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}E\,K\left(\frac{x-x_i}{h}\right) \qquad (1.14)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\int_{-\infty}^{\infty}K\left(\frac{x-t}{h}\right)f(t)dt\\ &= \frac{1}{h}\int_{-\infty}^{\infty}K\left(\frac{x-t}{h}\right)f(t)dt\end{aligned}$$

The transformation $z = \frac{x-t}{h}$, i.e. $t = hz + x$, $|\frac{dz}{dt}| = \frac{1}{h}$ yields

$$E(\hat{f}(x)) = \int_{-\infty}^{\infty} K(z) f(x - hz) dz$$

Expanding $f(x - hz)$ in a Taylor series yields

$$f(x - hz) = f(x) - hzf'(x) + \frac{1}{2}(hz)^2 f''(x) + o(h^2) \ ,$$

where $o(h^2)$ represents terms that converge to zero faster than h^2 as h approaches zero. Thus

$$\begin{aligned} E(\hat{f}(x)) &= \int_{-\infty}^{\infty} K(z) f(x) dz - \int_{-\infty}^{\infty} K(z) hzf'(x) dz \qquad (1.15) \\ &\quad + \int_{-\infty}^{\infty} K(z) \frac{(hz)^2}{2} f''(z) dz + o(h^2) \\ &= f(x) \int_{-\infty}^{\infty} K(z) dz - hf'(x) \int_{-\infty}^{\infty} zK(z) dz \\ &\quad + \frac{h^2}{2} f''(x) \int_{-\infty}^{\infty} z^2 K(z) dz + o(h^2) \\ &= f(x) + \frac{h^2}{2} k_2 f''(x) + o(h^2) \qquad (1.16) \end{aligned}$$

$$\text{Bias}(\hat{f}(x)) \approx \frac{h^2}{2} \ k_2 f''(x) \qquad (1.17)$$

$$\text{This depends on} \begin{cases} h & \text{Bias } (\hat{f}(x)) \xrightarrow[0]{h} 0 \ , \\ k_2 & \text{the “variance” of the kernel ,} \\ f''(x) & \text{the curvature of the density at the point } x \ . \end{cases}$$

The variance of $\hat{f}(x)$ is given by

$$\begin{aligned} Var(\hat{f}(x)) &= Var\left(\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\right) \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n} Var\left(K\left(\frac{x - x_i}{h}\right)\right) \end{aligned}$$

because the $x_i,\ i = 1, 2, ..., n$, are independently distributed. Now

$$\begin{aligned} Var\left(K\left(\frac{x-x_i}{h}\right)\right) &= E\left(K\left(\frac{x-x_i}{h}\right)^2\right) - \left(EK\left(\frac{x-x_i}{h}\right)\right)^2 \\ &= \int K\left(\frac{x-t}{h}\right)^2 f(t)dt - \left(\int K\left(\frac{x-t}{h}\right) f(t)dt\right)^2 \end{aligned}$$

$$\begin{aligned} Var(\hat{f}(x)) &= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-t}{h}\right)^2 f(t)dt - \frac{1}{n}\left(\frac{1}{h}\int K\left(\frac{x-t}{h}\right) f(t)dt\right)^2 \\ &= \frac{1}{n}\int \frac{1}{h^2}K\left(\frac{x-t}{h}\right)^2 f(t)dt - \frac{1}{n}\left(f(x) + \text{Bias}(\hat{f}(x))\right)^2 \end{aligned}$$

Substituting $z = \frac{x-t}{h}$ one obtains

$$Var(\hat{f}(x)) = \frac{1}{nh}\int K(z)^2 f(x-hz)dz - \frac{1}{n}\left(f(x) + o(h^2)\right)^2$$

Applying a Taylor approximation yields

$$Var(\hat{f}(x)) = \frac{1}{nh}\int K(z)^2(f(x) - hzf'(x) + o(h))dz - \frac{1}{n}\left(f(x) + o(h^2)\right)^2$$

Note that if n becomes large <u>and</u> h becomes small then the above expression becomes approximately:

$$Var(\hat{f}(x)) \approx \frac{1}{nh}f(x)\int K^2(z)dz \tag{1.18}$$

We note that the variance *decreases* as h increases.

The above approximations for the bias and variance of $\hat{f}(x)$ lead to

$$\begin{aligned} MSE(\hat{f}(x)) &= Bias^2(\hat{f}(x)) + Var(\hat{f}(x)) \\ &\approx \frac{1}{4}h^4k_2^2f''(x)^2 + \frac{1}{nh}f(x)j_2 \end{aligned} \tag{1.19}$$

where $k_2 := \int z^2K(z)dz$ and $j_2 := \int K(z)^2dz$.

Integrating (1.19) with respect to x yields

$$MISE(\hat{f}) \approx \frac{1}{4}h^4k_2^2\beta(f) + \frac{1}{nh}j_2\ , \tag{1.20}$$

where $\beta(f) := \int f''(x)^2dx$.

Of central importance is the way in which MISE($\hat{f}$) changes as a function of the bandwidth h. For very small values of h the second term in (1.20) becomes large but as h gets larger so the first term in (1.20) increases. There is an optimal value of h which minimizes MISE($\hat{f}$).

1.3.3 Optimal bandwidth

Expression (1.20) is the measure that we use to quantify the performance of the estimator. We can find the optimal bandwidth by minimizing (1.20) with respect to h. The first derivative is given by

$$\frac{d\ MISE(\hat{f})}{dh} = h^3 k_2^2 \beta(f) - \frac{1}{nh^2}\, j_2 \ .$$

Setting this equal to zero yields the optimal bandwidth, h_{opt}, for the given pdf and kernel:

$$h_{opt} = \left(\frac{1}{n}\ \frac{\gamma(K)}{\beta(f)} \right)^{1/5} , \tag{1.21}$$

where $\gamma(K) := j_2 k_2^{-2}$. Substituting (1.21) for h in (1.20) gives the minimal MISE for the given pdf and kernel. After some manipulation this can be shown to be

$$MISE_{opt}(\hat{f}) = \frac{5}{4} \left(\frac{\beta(f) j_2^4 k_2^2}{n^4} \right)^{1/5} . \tag{1.22}$$

We note that h_{opt} depends on the sample size, n, and the kernel, K. However, it also depends on the unknown pdf, f, through the functional $\beta(f)$. Thus as it stands expression (1.21) is not applicable in practice. However, the "plug–in" estimator of h_{opt}, to be discussed later, is simply expression (1.21) with $\beta(f)$ replaced by an estimator.

1.3.4 Optimal kernels

The MISE($\hat{f}$) can also be minimized with respect to the kernel used. It can be shown (see, e.g., Wand and Jones, 1995) that Epanechnikov kernel is optimal in this respect.

$$K(z) = \begin{cases} \frac{3}{4\sqrt{5}}\ (1 - \frac{1}{5}z^2) & \text{for } |z| < \sqrt{5} \\ 0 & \text{otherwise .} \end{cases}$$

This result together with (1.22) enables one to examine the impact of kernel choice on MISE$_{opt}(\hat{f})$. The efficiency of a kernel, K, relative to the optimal Epanechnikov kernel, K_{EP}, is defined as

$$\text{Eff}(K) = \left(\frac{MISE_{opt}(\hat{f}) \text{ using } K_{EP}}{MISE_{opt}(f) \text{ using } K} \right)^{5/4} = \left(\frac{k_2^2 j_2^4 \text{ using } K_{EP}}{k_2^2 j_2^4 \text{ using } K} \right)^{5/4} \tag{1.23}$$

The efficiencies for a number of well–known kernels are given in Table 1. It is clear that the selection of kernel has rather limited impact on the efficiency.

The rectangular kernel, for example, has an efficiency of approximately 93%. This can be interpreted as follows: The MISE$_{opt}(\hat{f})$ obtained using an Epanechnikov kernel with $n = 93$ is approximately equal to the MISE$_{opt}(\hat{f})$ obtained using a rectangular kernel with $n = 100$.

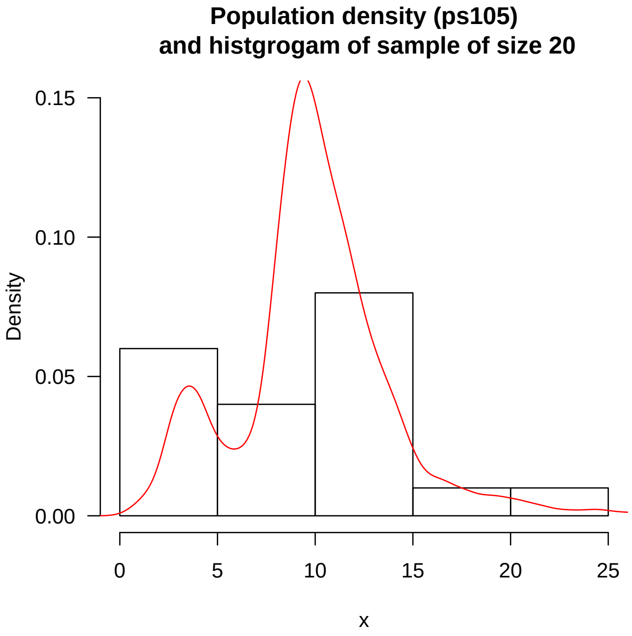

Figure 1.9: The pdf for the car example and a histogram for a sample of size 20.

1.4 Selection of the bandwidth

Selection of the bandwidth of kernel estimator is a subject of considerable research. We will outline four popular methods.

1.4.1 Subjective selection

One can experiment by using different bandwidths and simply select one that “looks right” for the type of data under investigation. Figure 1.9 shows the pdf (for the car example) and a histogram for random sample of size $n = 20$. Figure 1.10 shows kernel density estimation (based on a Gaussian kernel) of $f(x)$ using 4 different bandwidths.

Also shown is the density of the population. The latter is usually unknown in practice (otherwise we wouldn’t need to estimate it using a sample). Clearly $h = 0.5$ is too small, and $h = 3$ is too large. Appropriate here is a bandwidth greater than 1 but less than 3.

1.4.2 Selection with reference to some given distribution

Here one selects the bandwidth that would be optimal for a particular pdf. Convenient here is the normal. We note that one is not assuming that $f(x)$ is normal; rather one is

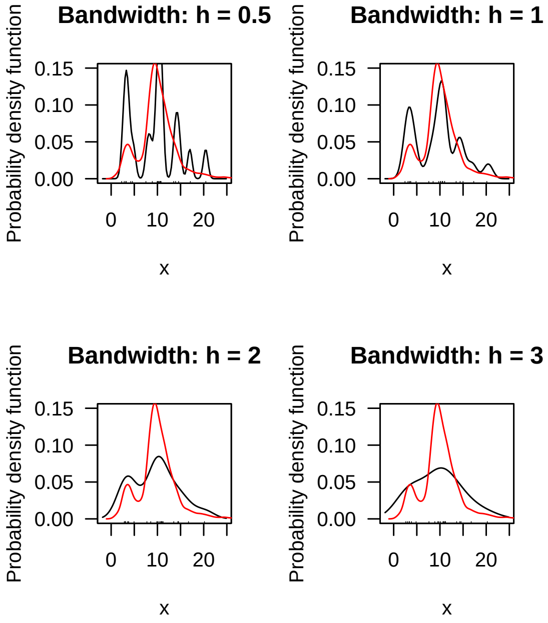

Figure 1.10: The pdf for the car example and kernel density estimates using a Gaussian kernel and different bandwidths.

selecting h which would be optimal if the pdf were normal. In this case it can be shown that

$$\beta(f) = \int f''(x)^2 dx = \frac{3\sigma^{-5}}{8\sqrt{\pi}}$$

and using a Gaussian kernel leads to

$$h_{opt} = \left(\frac{4}{3n}\right)^{1/5} \sigma \qquad \approx \qquad \frac{1.06\sigma}{n^5} \ . \tag{1.24}$$

The normal distribution is not a "wiggly" distribution; it is unimodal and bell–shaped. It is therefore to be expected that (1.24) will be too large for multimodal distributions. Secondly to apply (1.24) one has to estimate σ. The usual estimator, the sample variance, is not robust; it overestimates σ if some outliers (extreme observations) are present and thereby increases $\hat{h}_{opt}$ even more. To overcome these problems Silverman (1986) proposed the following estimator

$$\hat{h}_{opt} = \frac{0.9\hat{\sigma}}{n^5} \ , \tag{1.25}$$

where $\hat{\sigma} = min(s, R/1.34)$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ and R is the interquartile range of the data. The constant 1.34 is derived from the fact that for a $N(\mu, \sigma^2)$ random variable, X, one has $P\{|X - \mu| < 1.34 \ \sigma\} = 0.5$.

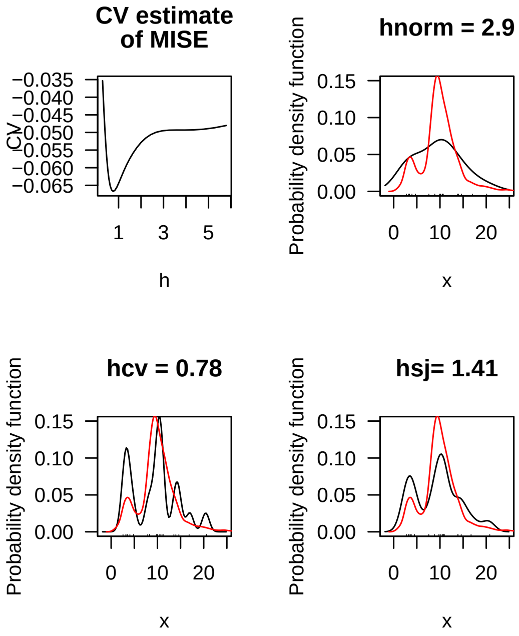

Figure 1.11: The cross-validation criterion (top left) and the estimated pdf using three different bandwidth selectors, namely cross-validation (bottom left), normal-based (top right) and plug-in (bottom right).

The expression (1.25) is used as the default option in the R function “density”. It is also used as a starting value in some more sophisticated iterative estimators for the optimal bandwidth. The top right graph in Figure 1.11 shows the estimated density, with this method of estimating h_{opt}.

1.4.3 Cross–validation

The technique of cross–validation will be discussed in more detail in the chapter on model selection. At this point we will only outline its application to the problem of estimating optimal bandwidths. By definition

$$\begin{aligned} MISE(\hat{f}) &= \int (\hat{f}(x) - f(x))^2 dx \\ &= \int \hat{f}(x)^2 dx - 2\int \hat{f}(x) f(x) dx + \int f(x)^2 dx \end{aligned}$$

The third term does not depend on the sample or on the bandwidth. An approximately unbiased estimator of the first two terms is given by

$$\widehat{MCV}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_i(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i) \ , \tag{1.26}$$

where $\hat{f}_{-i}(x)$ is the estimated density at the argument x using the original sample apart from observation x_i. One computes $\widehat{MCV}(\hat{f})$ for different values of h and estimates the optimal value, h_{opt}, using the h which minimizes $\widehat{MCV}(\hat{f})$. The top left hand graph in Figure 1.11 shows the curve $\widehat{MCV}(\hat{f})$ for the sample of car data. The bottom left hand graph shows the corresponding estimated density. In this example cross-validation has selected a bandwidth that is too small and the resulting estimated $\hat{f}$ has not been sufficiently smoothed.

1.4.4 ”Plug–in” estimator

The idea developed by Sheather and Jones (1991) is to estimate h from (1.21) by applying a separate smoothing technique to estimate $f''(x)$ and hence $\beta(f'')$. For details see, e.g. Wand and Jones (1995), section 3.6. An *R* function to carry out the computations is available in the *R* library “sm” of Bowman and Azzalini (1997).

Figure 1.11 shows that for the car example considered here the plug–in estimator yields the most sensible estimator of $\hat{f}$.

1.4.5 Summary and extensions

The above bandwidth selectors represent only a sample of the many suggestions that have been offered in the recent literature. Some alternatives are described in Wand and Jones (1995) in which the theory is given in more detail. These authors also offer recommendations regarding which estimators should be used. The plug–in estimator outlined above is one of their recommendations.

The above methodology can be extended to bivariate (and multivariate) pdfs. it is also possible to provide approximate confidence bands for the estimated pdfs. These subjects will be covered in the presentations by the participants of the course.