Chapter 1

Density estimation

1.1 Introduction

1.1.1 Parametric density estimation

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

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

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, if it 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 Histogram density estimation

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 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 histogram in Figure 1.1 relates to the entire population.

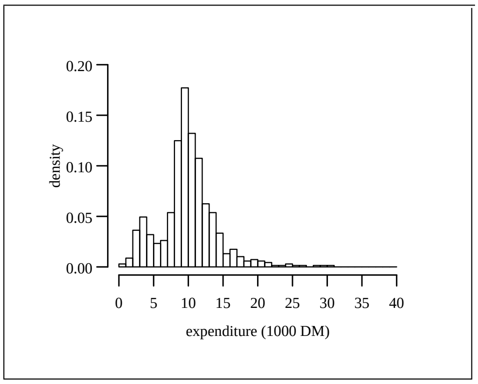

Figure 1.1: Histogram of expenditure for all cars in the population.

Figure 1.2 shows three histogram estimates of $f(x)$ for a random sample of size 10 from the population, 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 estimators in more detail in the following.

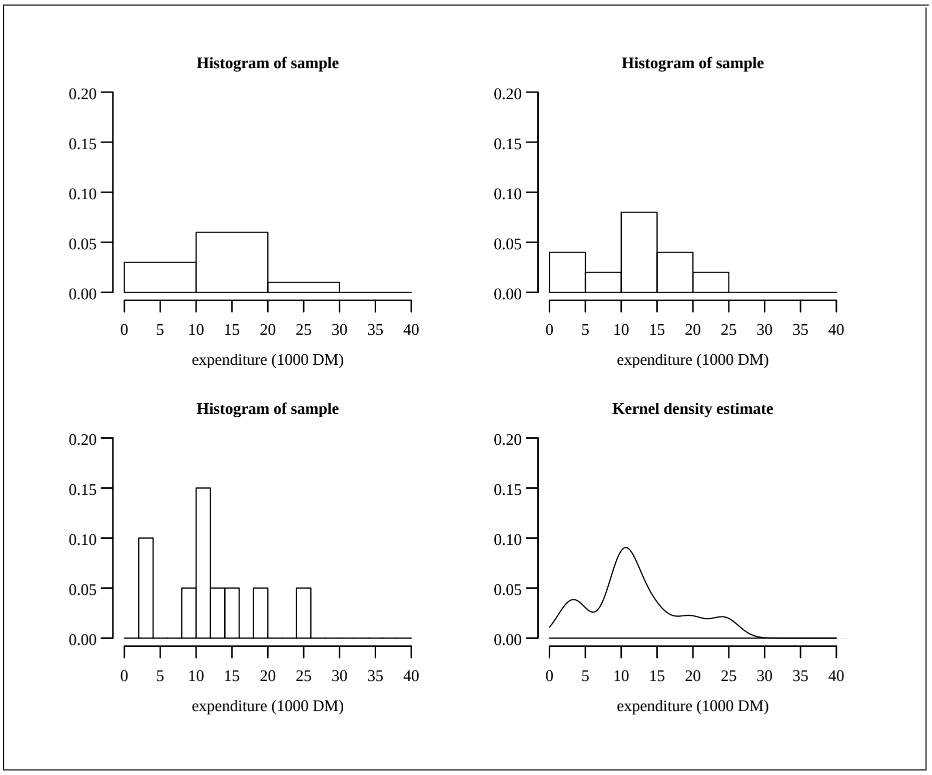

Figure 1.2: Histograms with different bin widths for the sample of size 10 and a kernel estimate of $f(x)$ for the same sample.

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

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

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, a rectangular weighting function, is defined as

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

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 $\hat{f}(x) \geq 0$ for all x, and $\int_{-\infty}^{\infty} \hat{f}(x)\, dx = 1$.

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 $\hat{f}(x)$ based on rectangular weighting functions for different values of h.

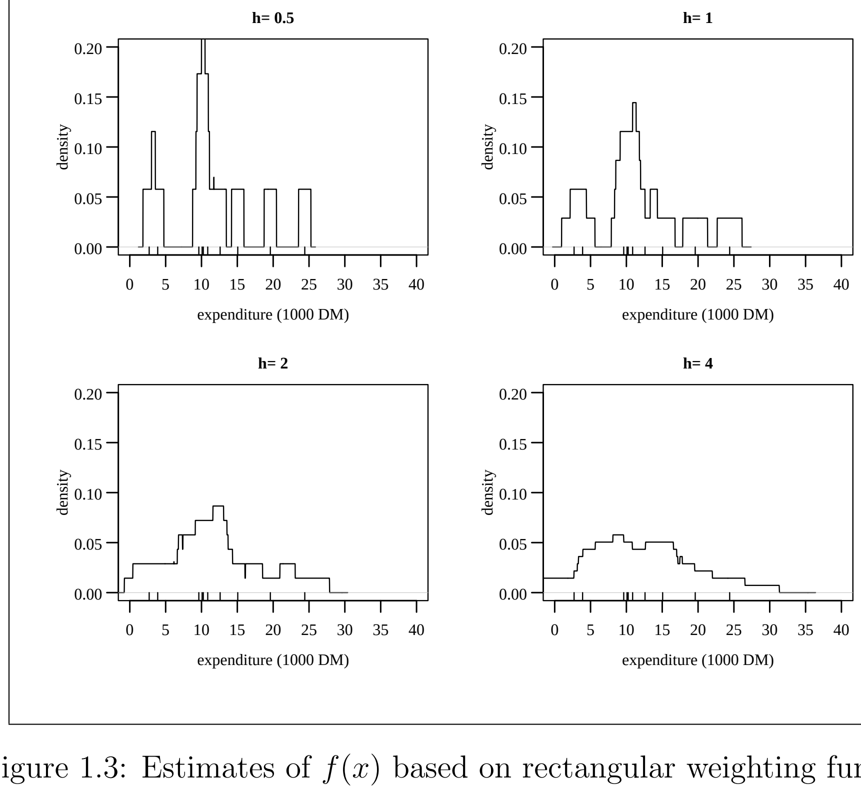

Figure 1.3: Estimates of $f(x)$ based on rectangular weighting functions.

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:

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

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 $\hat{f}(x)$.

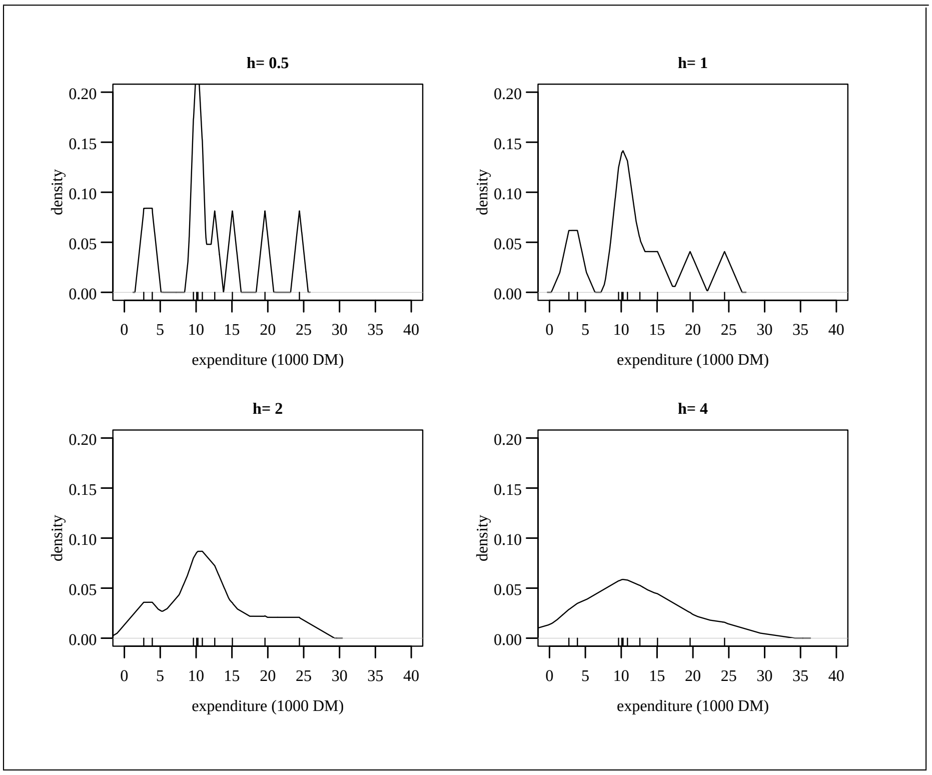

Figure 1.4: Estimates of $f(x)$ based on triangular weighting functions.

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.

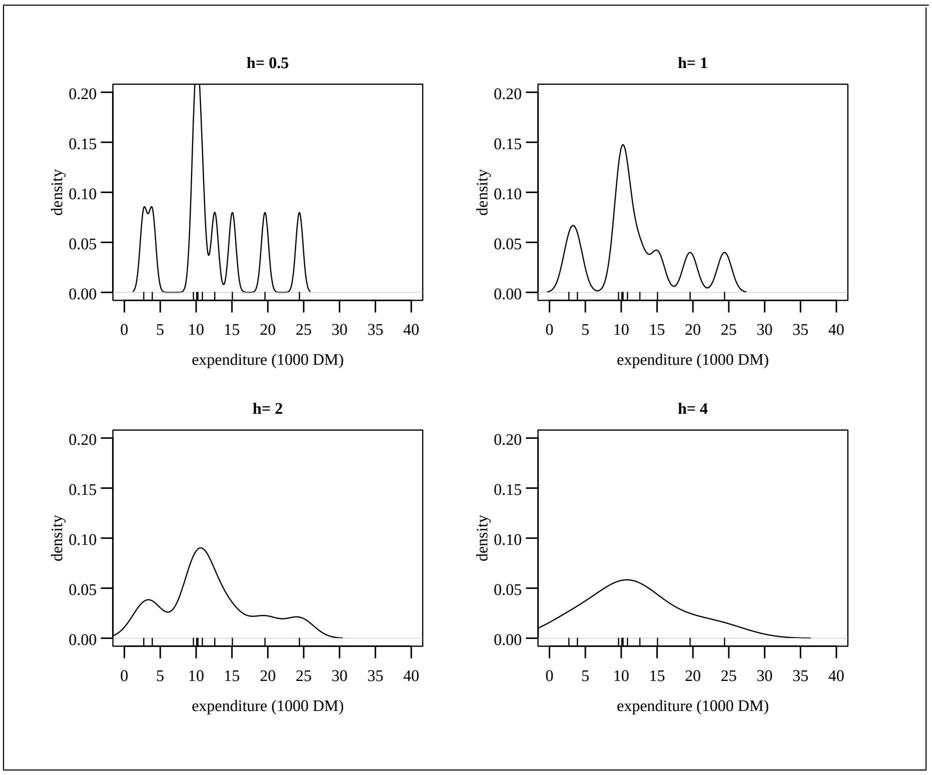

Figure 1.5: Estimates of $f(x)$ based on Gaussian weighting functions.

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	$K(z)$		Efficiency
Epanechnikov	$\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}z^2\right)$ 0	for $\lvert z\rvert < \sqrt{5}$ otherwise	1
Rectangular	$\frac{1}{2}$ 0	for $\lvert z\rvert < 1$ otherwise	$\sqrt{\frac{108}{125}} \approx 0.9295$
Triangular	$1-\lvert z\rvert$ 0	for $\lvert z\rvert < 1$ otherwise	$\sqrt{\frac{243}{250}} \approx 0.9859$
Biweight	$\frac{15}{16}(1-z^2)^2$ 0	for $\lvert z\rvert < 1$ otherwise	$\sqrt{\frac{3087}{3125}} \approx 0.9939$
Gaussian	$\frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}$	$z \in I\!R$	$\sqrt{\frac{36\pi}{125}} \approx 0.9512$

Table 1: Six kernels and their efficiencies (which are defined in Section 1.3.3).

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

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

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. Given the kernel and the bandwidth, the kernel density estimator of $f(x)$ is:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x-x_i}{h}\right) \qquad (1.6)$$

The two components K and h 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 to use a transformation. The idea is to estimate the pdf of a transformed random variable $Y = t(X)$ which has unbounded support, where t is some strictly monotone function. Suppose that the pdf of Y is given by $g(y)$. Then the relationship between f and g is given by

$$f(x) = g(t(x))t'(x)\ . \tag{1.7}$$

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 $t : [0,\ \infty) \rightarrow (-\infty,\ \infty)$ is the log-transformation, i.e. $t(x) = \log(x)$. Here $t'(x) = \frac{d \log(x)}{dx} = \frac{1}{x}$ and so

$$\hat{f}(x) = \hat{g}(\log(x))\frac{1}{x} \tag{1.8}$$

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

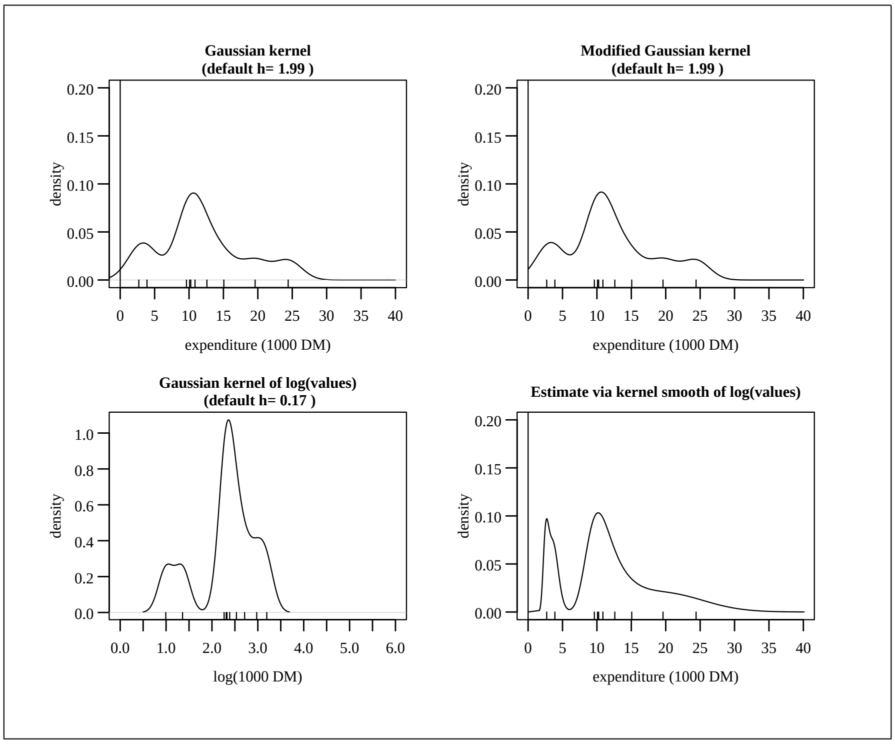

Figure 1.6: Kernel estimates of pdf with support $[0, \infty)$.

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

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

$$\hat{f}_c(x) = \begin{cases} \dfrac{\hat{f}(x)}{\int\limits_0^\infty \hat{f}(x)\,dx} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{1.9}$$

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 left 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 $t : (a, b) \rightarrow$

$(-\infty,\ \infty)$ is $t(x) = \log\left(\frac{x-a}{b-x}\right)$ (analogous to the logit–transformation in logistic regression). Here $t'(x) = \frac{1}{x-a} + \frac{1}{b-x}$ and so

$$\hat{f}(x) = \begin{cases} \hat{g}\left(\log\left(\frac{x-a}{b-x}\right)\right)\left(\frac{1}{x-a} + \frac{1}{b-x}\right) & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

Figure 1.7 provides an illustration of (1.10) for $a = 0$ and $b = 27$. The four figures shown are analogous to those in Example 2 but with

$$\hat{f}_c(x) = \begin{cases} \frac{\hat{f}(x)}{\int_a^b \hat{f}(x)\,dx} & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases} \tag{1.11}$$

for the graph on the top right.

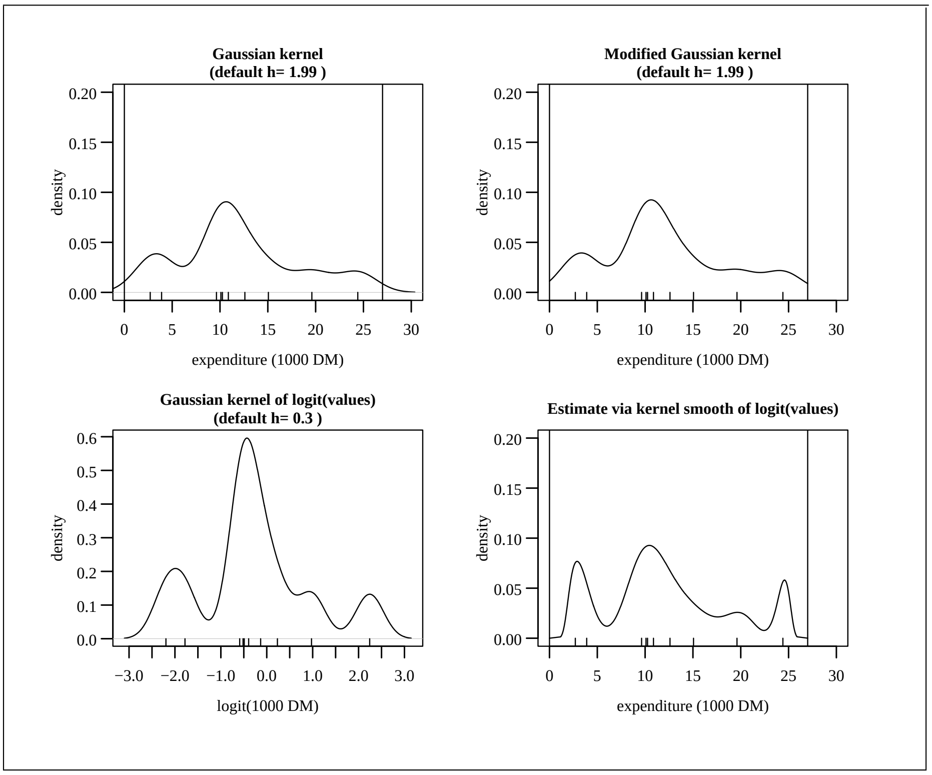

Figure 1.7: Kernel estimates of pdf with support [0,27].

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 and bandwidth at each point in the estimation of $f(x)$.

1.3 Properties 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 the bias and the variance. We note that the MSE of $\hat{f}(x)$ is a function of the argument x:

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

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

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

The IMSE is the sum of the integrated squared bias and the integrated variance. We consider each of these components in turn.

1.3.1 Bias, variance and mean squared error

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

The transformation $z = \frac{x-t}{h}$ (i.e. $t = x - hz$, $\frac{dt}{dz} = -h$, $z \to -\infty$ as $t \to \infty$, $z \to \infty$ as $t \to -\infty$) yields:

$$E\hat{f}(x) = -\int_{\infty}^{-\infty} K(z)f(x-hz)\,dz = \int_{-\infty}^{\infty} K(z)f(x-hz)\,dz \tag{1.14}$$

This formula shows that the expectation of $\hat{f}(x)$ is a weighted average of the values of the function f, centered at the point x, where the weights are determined by the kernel and the bandwidth. To illustrate this, consider again the population in Example 1, namely the amount paid by customers for optional extras when puchasing a particular model of car. The population density is given in Figure 1.8.

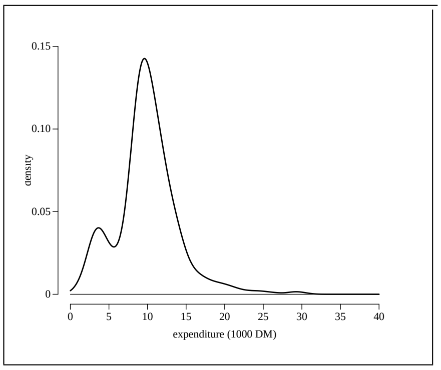

Figure 1.8: Population density $f(x)$ of amount paid for optional extras.

Note that the density shown in Figure 1.8 has been estimated from the population with a kernel density estimator too and that, in contrast to this example, in general the population density is completely unknown.

Figure 1.9 displays $E\hat{f}(x)$ and $f(x)$ for a Gaussian kernel and for four different values of the bandwidth, h. Also shown in these displays (in grey) are the weighting functions. These incorporate both the kernel and the bandwidth.

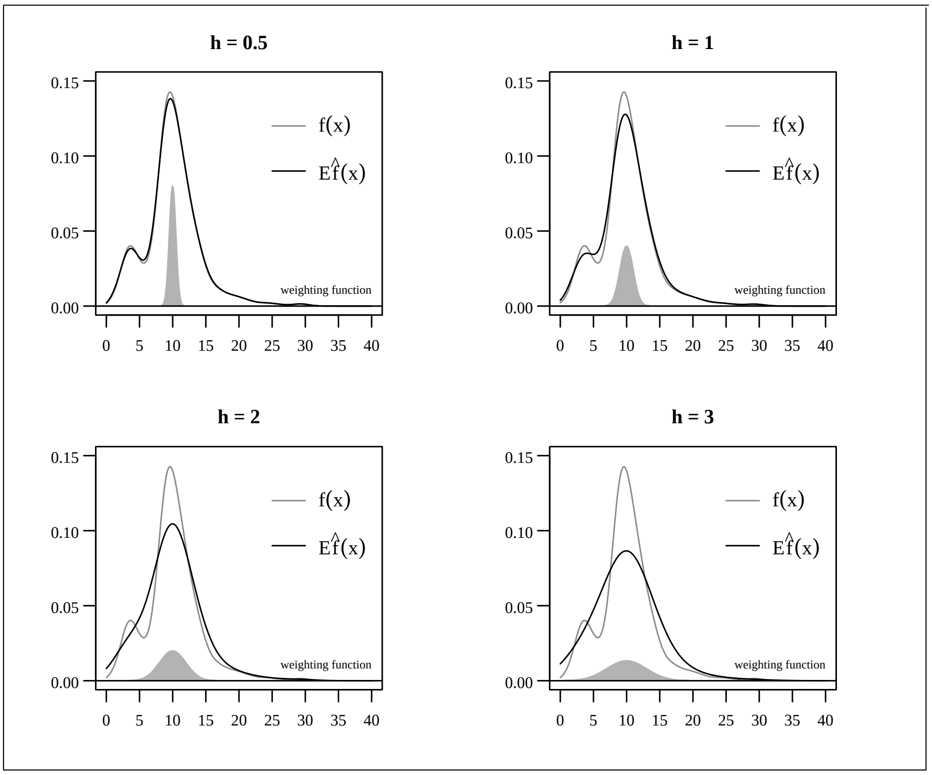

Figure 1.9: $E\hat{f}(x)$, $f(x)$ and the weighting function $w(t,h) = \frac{1}{h}K(\frac{t}{h})$ for different bandwidths.

Note that, as the bandwidth increases, so the difference between $E\hat{f}(x)$ and $f(x)$ becomes greater near the peaks and the valleys of $f(x)$; the function $E\hat{f}(x)$ becomes an increasingly smoothed version of $f(x)$. That's because the weighting function becomes more "spread out".

In general $\hat{f}(x)$ is a biased estimator of $f(x)$:

$$\text{Bias}(\hat{f}(x)) \;=\; E\hat{f}(x) - f(x) \;=\; \int_{-\infty}^{\infty} K(z)f(x-hz)\,dz \;-\; f(x) \tag{1.15}$$

The bias for our example is displayed in Figure 1.10. Note that the bias increases in absolute value as the bandwidth, h, increases whereas it is independent of the sample size n. The same is true for the integrated squared bias (ISB).

We now consider the variance of the estimator.

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

Now $x_1, x_2, \ldots, x_n$, are independently distributed, and

$$\begin{aligned}\mathrm{Var}\left(K\left(\frac{x-x_i}{h}\right)\right) &= E\left(K^2\left(\frac{x-x_i}{h}\right)\right) - \left(E\,K\left(\frac{x-x_i}{h}\right)\right)^2\\ &= \int_{-\infty}^{\infty} K^2\left(\frac{x-t}{h}\right) f(t)\,dt - \left(\int_{-\infty}^{\infty} K\left(\frac{x-t}{h}\right) f(t)\,dt\right)^2\\ &= h\int_{-\infty}^{\infty} K^2(z)f(x-hz)\,dz - \left(h\int_{-\infty}^{\infty} K(z)f(x-hz)\,dz\right)^2\\ &= h\int_{-\infty}^{\infty} K^2(z)f(x-hz)\,dz - h^2(E\hat{f}(x))^2\end{aligned}$$

The second-last step follows from the transformation $z = \frac{x-t}{h}$ (i.e. $t = x - hz$, $\frac{dt}{dz} = -h$, $z \to -\infty$ as $t \to \infty$, $z \to \infty$ as $t \to -\infty$). It follows that

$$\mathrm{Var}(\hat{f}(x)) = \frac{1}{nh}\int_{-\infty}^{\infty} K^2(z)f(x-hz)\,dz - \frac{1}{n}(E\hat{f}(x))^2 \qquad (1.16)$$

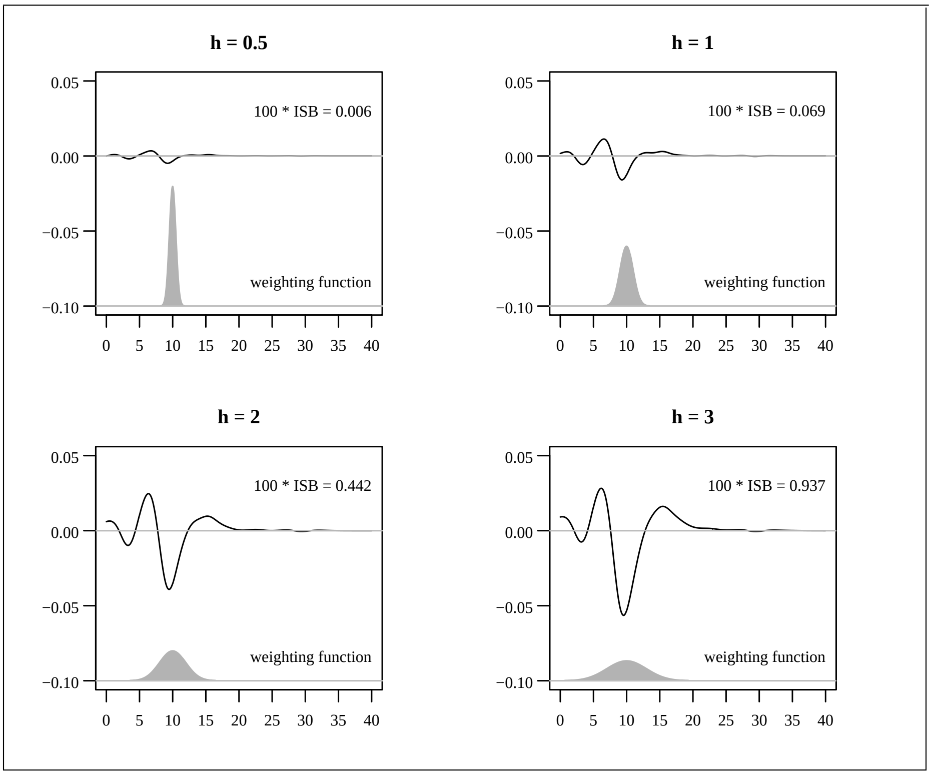

Figure 1.10: The bias of $\hat{f}(x)$, the integrated *squared* bias (ISB), and the weighting function for different bandwidths.

As is to be expected, both Var$(\hat{f}(x))$, and hence also the integrated variance, decrease with increasing sample size, n, as illustrated in Figure 1.11. They also both decrease with increasing bandwidth. Thus increasing h has desirable effect of reducing the variance, but it also has the undesirable effect of increasing (squared) bias, as was seen in Figure 1.10. Figure 1.11 displays the variance for different sample sizes and bandwidths.

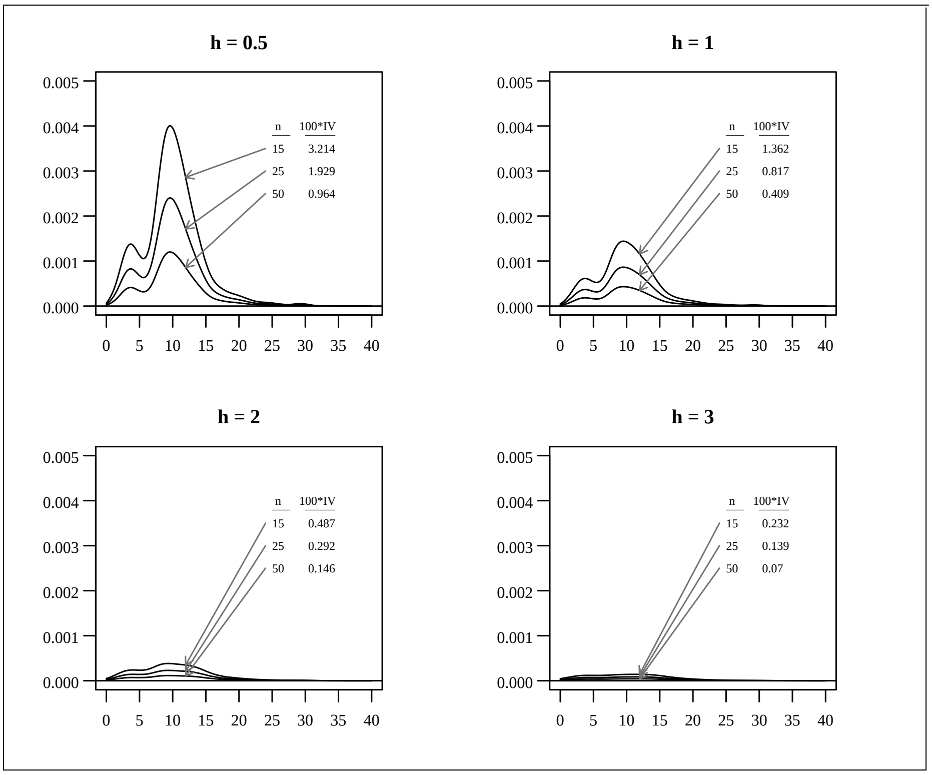

Figure 1.11: The variance $Var(\hat{f}(x))$, the integrated variance (IV) and the weighting function for different bandwidths.

The MSE, namely the sum of the variance and the squared bias, is displayed in Figure 1.12. Also given are values of the IMSE. As is to be expected, both the MSE and IMSE decrease with increasing sample size. Now consider the behaviour of MSE and IMSE as functions of h for a fixed sample size. Compare the values of MSE($\hat{f}(x)$), for any fixed value of x (e.g. $x = 10$) as h increases, i.e. compare the corresponding points of the curves in the four panels. As h increases MSE($\hat{f}(x)$) initially decreases and then increases. The same goes for the values of the IMSE. This illustrates the point that the value of h that minimizes the MSE($\hat{f}(x)$) should neither be too small (otherwise the variance becomes large) nor too large (otherwise the squared bias becomes large). We will regard the best bandwidth for estimating $f(x)$ as that which minimizes the MSE($\hat{f}(x)$).

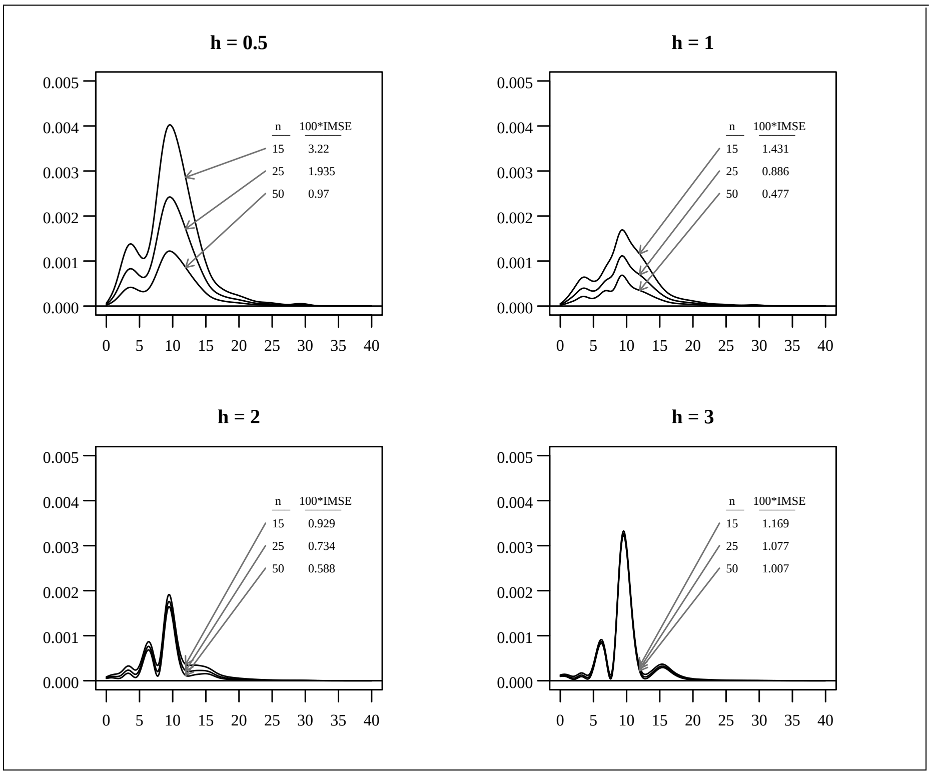

Figure 1.12: The mean squared error (MSE) and values of the integrated mean squared error (IMSE) for different sample sizes and bandwidths.

Figure 1.13 displays the IMSE and its two components, the integrated squared bias and the integrated variance, as functions of h for different sample sizes. This provides a clearer image of the effect noticeable in Figure 1.12, namely that the IMSE initally decreases and then increases. The optimal bandwith changes for different n. For $n = 15, 25, 50, 100$ it is $h_{opt} = 1.9, 1.6, 1.3, 1.05$, respectively. Thus, as n increases the optimal bandwidth decreases. To see why this makes sense note, first, that the integrated squared bias does not depend on the sample size. (The curve for the ISB is the same in all four panels.) However, the integrated variance (IV) decreases as n increases, thus contributing less to the IMSE. Thus the minimum of the IMSE occurs at a smaller value of h.

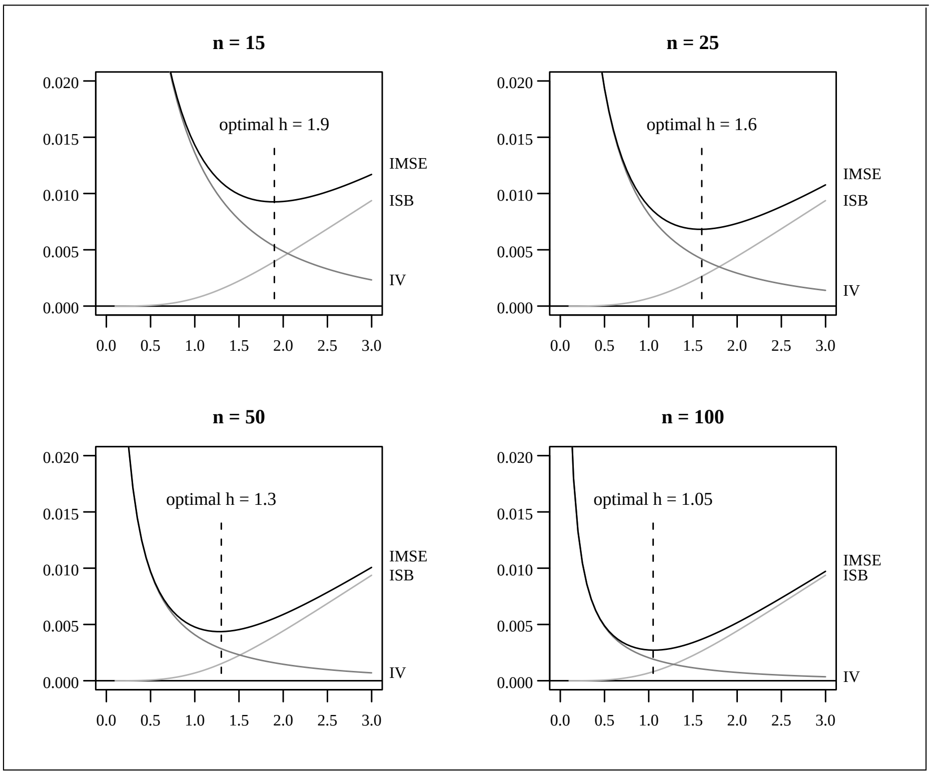

Figure 1.13: Integrated squared bias (ISB), integrated variance (IV), and integrated mean squared error (IMSE) as functions of the bandwith for different sample sizes. In each case the optimal bandwidth is shown.

Intuitively, increasing the sample size increases the amount of information available to estimate f; it enables us to estimate f in greater detail. Thus when n is very large we can distinguish the smaller dips and bumps of f with confidence, i.e. we can use a smaller bandwidth, a "greater magnification". If n is small then we have to make do with "less magnification", i.e. a larger bandwidth.

The expectation, variance and mean squared error of $\hat{f}(x)$, as well as the integrated squared bias, integrated variance and IMSE, depend not only on the kernel, K and the bandwidth, h, but also on the pdf $f(x)$. We are free to choose K and h but not, of course, $f(x)$. The example that was used to illustrate these dependencies is atypical, in that $f(x)$ was known, whereas in practice $f(x)$ is unknown. (If it were known then we wouldn't need to estimate it.) So, although there exists an optimal bandwidth, it can't be determined in practice; it can only be estimated. To develop estimators it is useful to examine the asymptotic behaviour of $\hat{f}(x)$, i.e. to study its bias, variance and MSE as n becomes large.

1.3.2 Asymptotic properties

It was illustrated in the previous section that the optimal bandwidth of a kernel estimator decreases as the sample size increases. In examining the asymptotic properties of $\hat{f}(x)$ we will therefore investigate the case in which h is decreased as n increases. We will assume that h decreases more slowly than $\frac{1}{n}$, as n becomes large. Thus

$$\lim_{n\to\infty} h = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \frac{1}{nh} = 0 \tag{1.17}$$

We will also assume that f is sufficently often differentiable, and that the kernel satisfies conditions (1.5).

We start by examining the behaviour of $\mathrm{E}\hat{f}(x) = \int_{-\infty}^{\infty} K(z) f(x-hz)\,dz$ as $n \to \infty$. Expanding $f(x-hz)$ in a Taylor series (see appendix B for details) yields

$$f(x-hz) = f(x) - hzf'(x) + \frac{1}{2}(hz)^2 f''(x) + o(h^2)\ , \tag{1.18}$$

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

$$\begin{aligned}
\mathrm{E}\hat{f}(x) &= \int_{-\infty}^{\infty} K(z)f(x)\,dz - \int_{-\infty}^{\infty} K(z)hzf'(x)\,dz + \int_{-\infty}^{\infty} K(z)\frac{(hz)^2}{2}f''(x)\,dz + o(h^2) \\
&= f(x)\int_{-\infty}^{\infty} K(z)\,dz - hf'(x)\int_{-\infty}^{\infty} zK(z)\,dz + \frac{h^2}{2}f''(x)\int_{-\infty}^{\infty} z^2K(z)\,dz + o(h^2) \\
&= f(x) + \frac{h^2}{2}k_2 f''(x) + o(h^2) \qquad (1.19)
\end{aligned}$$

where k_2, the variance of the kernel, is defined in (1.5). Thus, for small values of h,

$$\mathrm{Bias}\,\hat{f}(x) \approx \frac{h^2}{2}\ k_2 f''(x) \tag{1.20}$$

The asymptotic bias depends on

(a) h, where the bias approaches zero as h becomes small,

(b) k_2, the variance of the kernel,

(c) $f''(x)$, the curvature of f at the point x.

The curvature of f is negative near the peaks of f, leading to negative bias, and positive near the valleys, leading to positive bias. On average $\hat{f}$ "erodes the hills and fills in the valleys" of f. This effect, which was illustrated in Figure 1.10, can be diminished by decreasing h.

We now consider the behaviour of $\mathrm{Var}(\hat{f}(x))$ as n increases (and h decreases). Note that, from appendix B and (1.19) follows that

$$\begin{aligned} f(x-hx) &= f(x)+o(1) \\ E\hat{f}(x) &= f(x)+o(1) \end{aligned}$$

and hence expression (1.16) can be written in the form

$$\begin{aligned} \mathrm{Var}(\hat{f}(x)) &= \frac{1}{nh}\int\limits_{-\infty}^{\infty} K^2(z)(f(x)+o(1))\,dz - \frac{1}{n}\left(f(x)+o(1)\right)^2 \\ &= \frac{1}{nh}\int\limits_{-\infty}^{\infty} K^2(z)f(x)\,dz + o\left(\frac{1}{nh}\right). \end{aligned}$$

Now, by assumption $\lim_{n\to\infty}\frac{1}{nh}=0$ and so for large n, we have:

$$\begin{aligned} \mathrm{Var}(\hat{f}(x)) &\approx \frac{1}{nh}\,f(x)\int\limits_{-\infty}^{\infty} K^2(z)\,dz \\ &= \frac{1}{nh}f(x)j_2 \qquad (1.21) \end{aligned}$$

The asymptotic variance depends on

(a) n; it approaches zero as n becomes large,

(b) h; it increases as h is reduced,

(c) $j_2 = \int\limits_{-\infty}^{\infty} K^2(z)\,dz$, a property of the kernel,

(d) $f(x)$, the density at point x.

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

$$\begin{aligned} \text{MSE}(\hat{f}(x)) &= \text{Bias}^2(\hat{f}(x)) + \text{Var}(\hat{f}(x)) \\ &\approx \frac{1}{4}h^4 k_2^2 (f''(x))^2 + \frac{1}{nh} f(x) j_2. \end{aligned} \tag{1.22}$$

Integrating (1.22) over x yields the integrated mean squared error:

$$IMSE(\hat{f}) \approx \frac{1}{4}h^4 k_2^2 \beta(f) + \frac{1}{nh} j_2 \,. \tag{1.23}$$

where $\beta(f) = \int_{\infty}^{\infty} (f''(x))^2 \, dx$ is the integrated squared curvature of $f(x)$.

The asymptotic IMSE depends on

(a) n; it decreases as n becomes large,

(b) $\beta(f)$; it is larger for ”wiggly” densities,

(c) j_2, k_2, which are properties of the kernel, K,

(d) h, the bandwidth.

Of central importance is the behaviour of IMSE($\hat{f}$) as a function of the bandwidth h. Each of the above components is positive, and thus IMSE($\hat{f}$) is of the form $ah^4 + bh^{-1}$, with $a, b > 0$, which is a U-shaped function. Setting the derivative with respect to h equal to zero, we see that the function has a minimum as the bandwidth $h = (b/4a)^{1/5}$, i.e.

$$h_{opt} = \left(\frac{1}{n} \frac{\gamma(K)}{\beta(f)} \right)^{1/5} \,, \tag{1.24}$$

where $\gamma(K) = j_2 k_2^{-2} = \left(\int_{\infty}^{\infty} K^2(z) \, dz \right) \left(\int_{\infty}^{\infty} z^2 K(z) \, dz \right)^{-2}$. 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 (1.24) is not applicable in practice. However, the “plug-in” estimator of h_{opt}, to be discussed later, is simply expression (1.24) with $\beta(f)$ replaced by an estimator.

Substituting h_{opt} in (1.23) gives the minimum attainable value of IMSE for the given sample size, pdf and kernel:

$$IMSE_{opt}(\hat{f}) = \frac{5}{4} \left(\frac{\beta(f) j_2^4 k_2^2}{n^4} \right)^{1/5} \,. \tag{1.25}$$

1.3.3 Optimal kernels

The asymptotic IMSE($\hat{f}$) given in (1.23) can also be minimized with respect to the kernel used. It can be shown (see e.g. Hodges and Lehmann, 1956) that the 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 1.25 enables one to examine the impact of kernel choice on IMSE$_{opt}(\hat{f})$. The efficiency of a kernel, K, relative to the (optimal) Epanechnikov kernel, K_{EP}, is defined as

$$\text{Eff}(K) = \left(\frac{IMSE_{opt}(\hat{f}) \text{ using } K_{EP}}{IMSE_{opt}(\hat{f}) \text{ using } K}\right)^{5/4} = \left(\frac{k_2^2 j_2^4 \text{ using } K_{EP}}{k_2^2 j_2^4 \text{ using } K}\right)^{1/4} \qquad (1.26)$$

The reason for the power 5/4 in 1.26 is that for large n the optimal IMSE will be the same whether one uses n observations and the kernel K or whether one uses $n\,\text{Eff}(K)$ observations and the kernel K_{EP}. The efficiencies for a number of well-known kernels are given in Table 1. It is clear that the selection of the 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 IMSE$_{opt}(\hat{f})$ obtained using an Epanechnikov kernel with $n = 93$ is approximately equal to the IMSE$_{opt}(\hat{f})$ obtained using a rectangular kernel with $n = 100$.

1.4 Selection of the bandwidth

Selection of the bandwidth for kernel estimators is a subject of considerable research. We will outline four popular methods. We consider again the data in Example 1; the expenditures for optional extras for a polulation of 689 cars. Figure 1.14 shows the (estimated) pdf for the population and a histogram for a random sample of size $n = 20$.

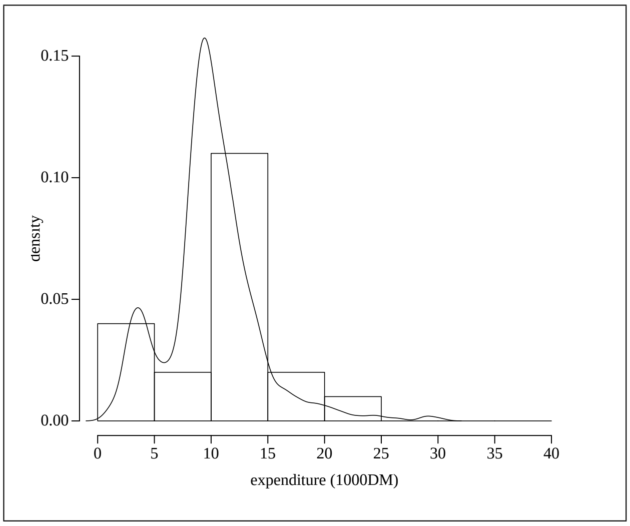

Figure 1.14: The pdf for the population of the car example and a histogram for a sample of size 20.

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 1.15 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). The bandwidth $h = 1$ seems to be too small as the two peaks at $x \approx 20$ and $x \approx 25$ correspond to single observations. On the other hand $h = 8$, and possibly even $h = 4$, are too large because there seem to be two distinct groups of observations (centered at $x \approx 3$ and $x \approx 12$) suggesting that $f(x)$ is bimodal. Therefore it would seem that $1 < h < 4$ is reasonable.

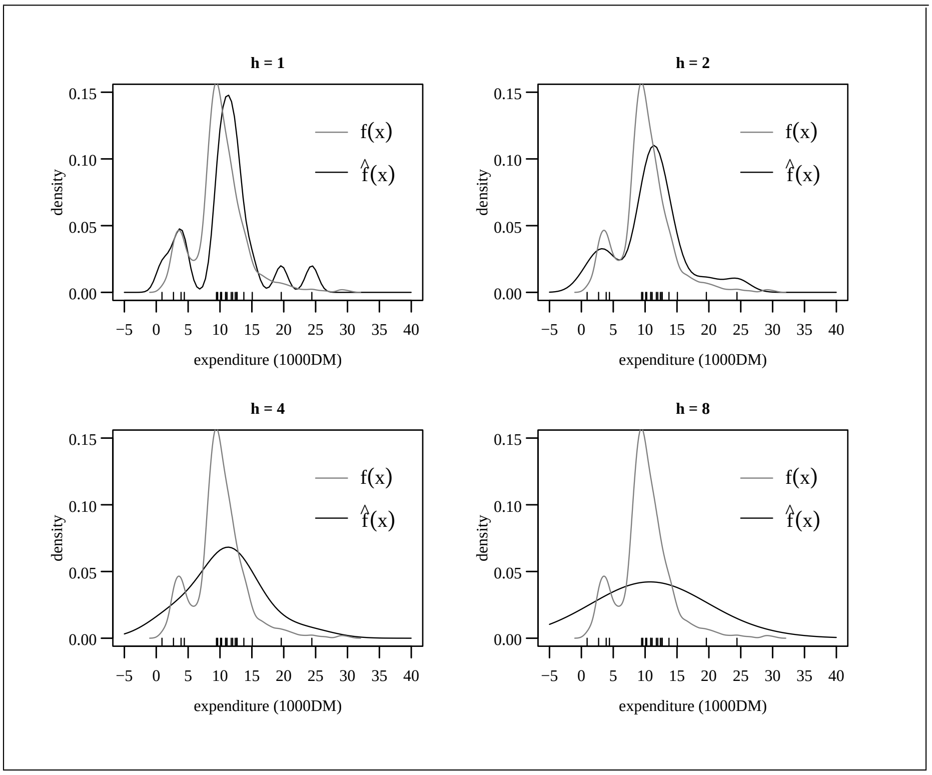

Figure 1.15: The pdf for the car example and kernel density estimates using a Gaussian kernel and different bandwidths.

1.4.2 Selection with reference to some given distribution

Here one selects the bandwidth that would be optimal for a particular pdf. Convenient is the normal distribution. We note that one is not assuming that $f(x)$ is normal; rather one is selecting h which would be optimal if the pdf were normal. For the normal distribution it can be shown that

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

and using a Gaussian kernel leads to

$$h_{opt} = \sigma \left(\frac{4}{3n} \right)^{1/5} \approx \frac{1.06\sigma}{n^{1/5}} \ . \tag{1.27}$$

The normal distribution is not a "wiggly" distribution; it is unimodal and bell-shaped. It is therefore to be expected that (1.27) will be too large for multimodal distributions. Secondly to apply (1.27) 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

$$\hat{h}_{opt} \;=\; \frac{0.9\hat{\sigma}}{n^{1/5}} \;, \tag{1.28}$$

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 $P\{|X - \mu| < 1.34\;\sigma\} = 0.5$.

The estimator (1.28) 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-hand graph in Figure 1.16 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, the integrated squared error, ISE, is

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

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 the mean cross validation criterion:

$$\widehat{MCV}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\int\limits_{-\infty}^{\infty} \hat{f}_{-i}^2(x)\,dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(x_i)\;, \tag{1.29}$$

where $\hat{f}_{-i}(x)$ is the estimated density at the argument x using the original sample apart from observation x_i. One can show that $\widehat{MCV}$ is an estimator of the $IMSE(\hat{f})$ (based on $n-1$ observations and apart from the constant term $\int_{-\infty}^{\infty} f^2(x)\,dx$). In order to select the bandwidth, 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 1.16 shows the curve $\widehat{MCV}(\hat{f})$ for the sample of car data. The bottom left-hand graph shows the corresponding estimated density.

1.4.4 "Plug-in" estimator

The idea developed by Sheather and Jones (1991) is to estimate h from (1.23) 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). The resulting density estimator is shown in in the bottom right hand graph in Figure 1.16.

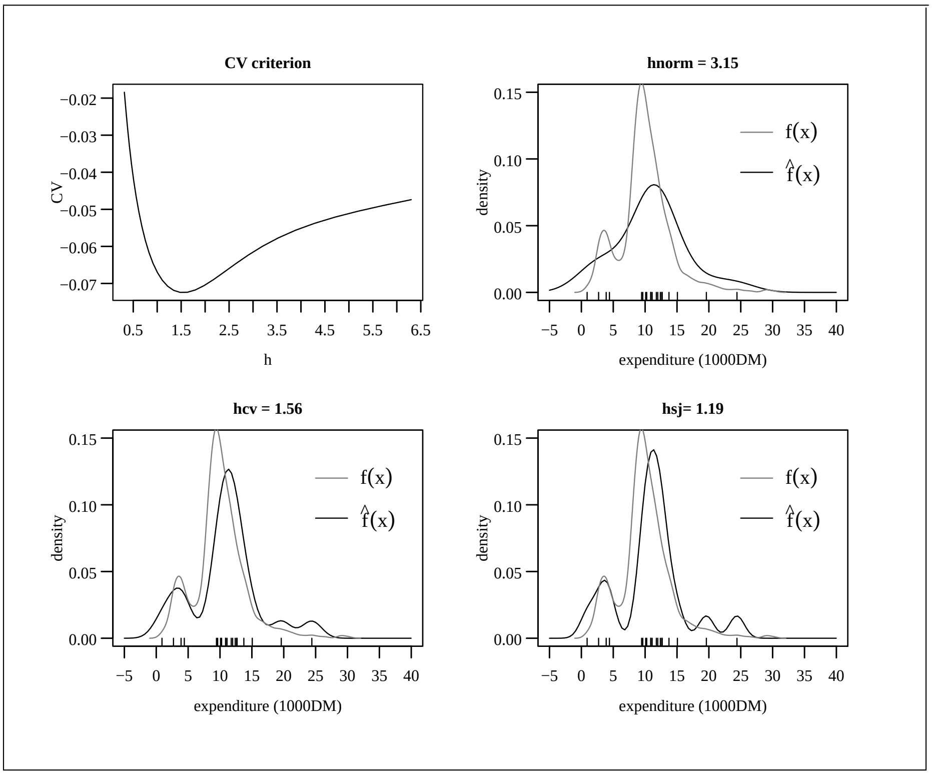

Figure 1.16: The cross-validation criterion (top left) and the estimated pdf using three different bandwidth selectors, namely cross-validation (bottom left), normal-based (top right) and plug-in (bottom right).

In this particular example, having only $n = 20$ observations, the criteria have been unable to select a bandwidth that both leads to a smooth density estimator *and* identifies the two peaks which we know are present in the population. Clearly the bandwidth with reference to the normal is inappropriately large in this example.

These 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.

1.4.5 Summary and extensions

The above discussion has repeatedly emphasized the fact that the choice of bandwidth has a strong influence on the properties of kernel density estimators. One of the main points is that, as the sample size increases so h should be decreased, and vice versa. Throughout the discussion we only considered the option of selecting a single value of h to estimate $f(x)$ for all x; we have focused on trying to minimize (1.23) with respect to h. However, we could focus on minimizing (1.22) instead, in which case h would depend on x. In other words we can use different bandwidths for different values of x, selecting a small bandwidth in intervals where the observations are plentiful, and a larger bandwidth where they are sparser. An account of such variable bandwidth (or local) kernel density estimators is given, for example, in Section 2.10 of Wand and Jones (1995), and in Section 5.3 of Silverman (1986). We will discuss one variable-bandwidth method in the the next chapter, in the context of kernel regression

Kernel density estimation is also applicable to multivariate pdfs. Although the technical and computational details for the mutivariate case are a little more complicated, the basic idea, and the properties of the estimators, are analogous to those that we have discussed for univariate pdfs. For information on this topic see, for example, Chapter 4 in Wand and Jones (1995), or Chapter 5 in Silverman (1986).

It is also possible to provide approximate confidence bands for the estimated pdfs. To emphasize their approximate nature, these are sometimes called *variability bands* rather than confidence bands. For details see, for example, Bowman and Azzalini (1997), Section 2.3.