Published in Monographs on Statistics and Applied Probability, London: Chapman and Hall, 1986.

# DENSITY ESTIMATION FOR STATISTICS AND DATA ANALYSIS

**B.W. Silverman**

School of Mathematics University of Bath, UK

**Table of Contents**

- INTRODUCTION
  - What is density estimation?
  - Density estimates in the exploration and presentation of data
  - Further reading
- SURVEY OF EXISTING METHODS
  - Introduction
  - Histograms
  - The naive estimator
  - The kernel estimator
  - The nearest neighbour method
  - The variable kernel method
  - Orthogonal series estimators
  - Maximum penalized likelihood estimators
  - General weight function estimators
  - Bounded domains and directional data
  - Discussion and bibliography

## 1. INTROUCTION

### 1.1. What is density estimation?

The *probability density function* is a fundamental concept in statistics. Consider any random quantity $X$ that has probability density function $f$. Specifying the function $f$ gives a natural description of the distribution of $X$, and allows probabilities associated with $X$ to be found from the relation

$$P(a < X < b) = \int_a^b f(d)dx \quad \text{for all } a < b.$$

Suppose, now, that we have a set of observed data points assumed to be a sample from an unknown probability density function. *Density estimation*, as discussed in this book, is the construction of an estimate of the density function from the observed data. The two main aims of the book are to explain how to estimate a density from a given data set and to explore how density estimates can be used, both in their own right and as an ingredient of other statistical procedures.

One approach to density estimation is *parametric*. Assume that the data are drawn from one of a known parametric family of distributions, for example the normal distribution with mean $\mu$ and variance $\sigma^2$. The density $f$ underlying the data could then be estimated by finding estimates of $\mu$ and $\sigma^2$ from the data and substituting these estimates into the formula for the normal density. In this book we shall not be considering parametric estimates of this kind; the approach will be more *non parametric* in that less rigid assumptions will be made about the distribution of the observed data. Although it will be assumed that the distribution has a probability density $f$, the data will be allowed to speak for themselves in determining the estimate of $f$ more than would be the case if $f$ were constrained to fall in a given parametric family.

Density estimates of the kind discussed in this book were first proposed by Fix and Hodges (1951) as a way of freeing

discriminant analysis from rigid distributional assumptions. Since then, density estimation and related ideas have been used in a variety of contexts, some of which, including discriminant analysis, will be discussed in the final chapter of this book. The earlier chapters are mostly concerned with the question of how density estimates are constructed. In order to give a rapid feel for the idea and scope of density estimation, one of the most important applications, to the exploration and presentation of data, will be introduced in the next section and elaborated further by additional examples throughout the book. It must be stressed, however, that these valuable exploratory purposes are by no means the only setting in which density estimates can be used.

### 1.2. Density estimates in the exploration and presentation of data

A very natural use of density estimates is in the informal investigation of the properties of a given set of data. Density estimates can give valuable indication of such features as skewness and multimodality in the data. In some cases they will yield conclusions that may then be regarded as self-evidently true, while in others all they will do is to point the way to further analysis and/or data collection.

An example is given in Fig. 1.1. The curves shown in this figure were constructed by Emery and Carpenter (1974) in the course of a study of sudden infant death syndrome (also called `cot death' or `crib death'). The curve A is constructed from a particular observation, the degranulated mast cell count, made on each of 95 infants who died suddenly and apparently unaccountably, while the cases used to construct curve B were a control sample of 76 infants who died of known causes that would not affect the degranulated mast cell count. The investigators concluded tentatively from the density estimates that the density underlying the sudden infant death cases might be a mixture of the control density with a smaller proportion of a contaminating density of higher mean. Thus it appeared that in a minority (perhaps a quarter to a third) of the sudden deaths, the degranulated mast cell count was exceptionally high. In this example the conclusions could only be regarded as a cue for further clinical investigation.

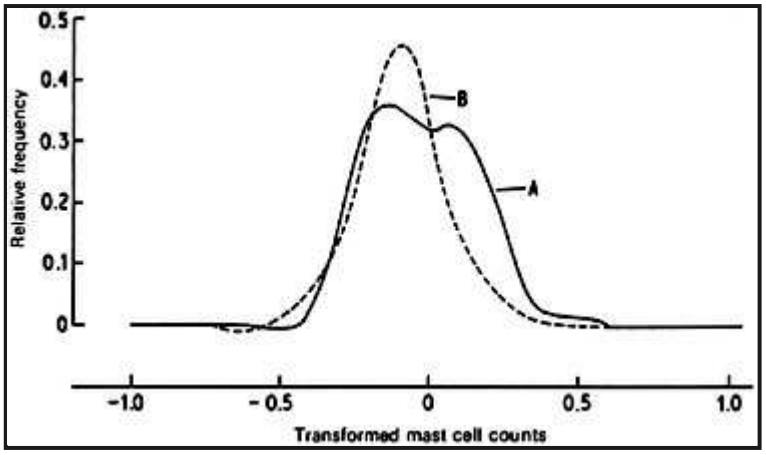


**Fig. 1.1** Density estimates constructed from transformed and corrected degranulated mast cell counts observed in a cot death study. (A, Unexpected deaths; B, Hospital deaths.) After Emery and Carpenter (1974) with the permission of the Canadian Foundation for the Study of Infant Deaths. This version reproduced from Silverman (1981a) with the permission of John Wiley & Sons Ltd.

Another example is given in Fig. 1.2. The data from which this figure was constructed were collected in an engineering experiment described by Bowyer (1980). The height of a steel surface above an arbitrary level was observed at about 15 000 points. The figure gives a density estimate constructed from the observed heights. It is clear from the figure that the distribution of height is skew and has a long lower tail. The tails of the distribution are particularly important to the engineer, because the upper tail represents the part of the surface which might come into contact with other surfaces, while the lower tail represents hollows where fatigue cracks can start and also where lubricant might gather. The non-normality of the density in Fig. 1.2 casts doubt on the Gaussian models typically used to model these surfaces, since these models would lead to a normal distribution of height. Models which allow a skew distribution of height would be more appropriate, and one such class of models was suggested for this data set by Adler and Firman (1981).

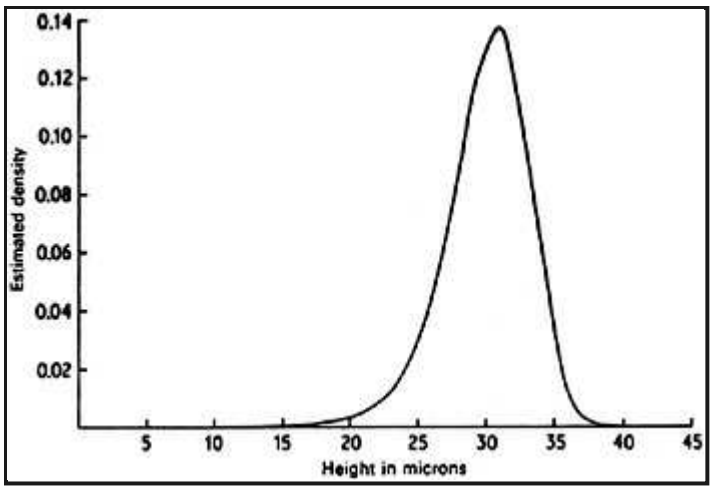


**Fig. 1.2** Density estimate constructed from observations of the height of a steel surface. After Silverman (1980) with the permission of Academic Press, Inc. This version reproduced from Silverman (1981a) with the permission of John Wiley & Sons Ltd.

A third example is given in Fig. 1.3. The data used to construct this curve are a standard directional data set and consist of the directions in which each of 76 turtles was observed to swim when released. It is clear that most of the turtles show a preference for swimming approximately in the 60° direction, while a small proportion prefer exactly the opposite direction. Although further statistical modelling of these data is possible (see Mardia 1972) the density estimate really gives all the useful conclusions to be drawn from the data set.

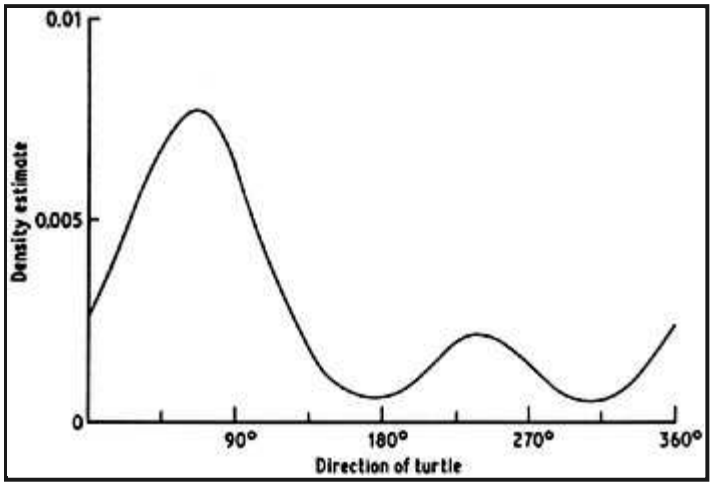


**Fig. 1.3** Density estimate constructed from turtle data. After Silverman (1978a) with the permission of the Biometrika Trustees. This version reproduced from Silverman (1981a) with the permission of John Wiley & Sons Ltd.

An important aspect of statistics, often neglected nowadays, is the presentation of data back to the client in order to provide explanation and illustration of conclusions that may possibly have been obtained by other means. Density estimates are ideal for this purpose, for the simple reason that they are fairly easily comprehensible to non-mathematicians. Even those statisticians who are sceptical about estimating densities would no doubt explain a normal distribution by drawing a bell-shaped curve rather than by one of the other methods illustrated in Fig. 1.4. In all the examples given in this section, the density estimates are as valuable for explaining conclusions as for drawing these conclusions in the first place. More examples illustrating the use of density estimates for exploratory and presentational purposes, including the important case of bivariate data, will be given in later chapters.

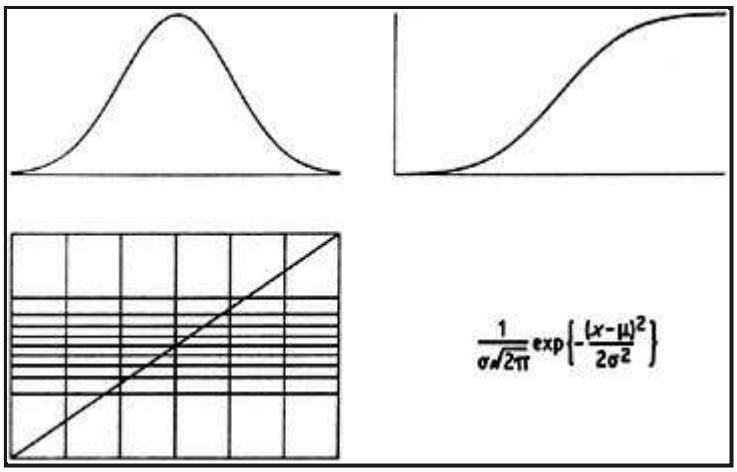


**Fig. 1.4** Four ways of explaining the normal distribution: a graph of the density function; a graph of the cumulative distribution function; a straight line on probability paper, the formula for the density function.

### 1.3. Further reading

There is a vast literature on density estimation, much of it concerned with asymptotic results not covered in any detail in this book.

Prakasa Rao's (1983) book offers a comprehensive treatment of the theoretical aspects of the subject. Journal papers providing surveys and bibliography include Rosenblatt (1971), Fryer (1977), Wertz and Schneider (1979), and Bean and Tsokos (1980). Tapia and Thompson (1978) give an interesting perspective paying particular attention to their own version of the penalized likelihood approach described in Sections 2.8 and 5.4 below. A thorough treatment, rather technical in nature, of a particular question and its ramifications is given by Devroye and Gyorfi (1985). Other texts on the subject are Wertz (1978) and Delecroix (1983). Further references relevant to specific topics will be given, as they arise, later in this book.

## 2. SURVEY OF EXISTING METHODS

### 2.1. Introduction

In this chapter a brief summary is given of the main methods available for univariate density estimation. Some of the methods will be discussed in greater detail in later chapters, but it is helpful to have a general view of the subject before examining any particular method in detail. Many of the important applications of density estimation are to multivariate data, but since all the multivariate methods are generalizations of univariate methods, it is worth getting a feel for the univariate case first.

Two data sets will be used to help illustrate some of the methods. The first comprises the lengths of 86 spells of psychiatric treatment undergone by patients used as controls in a study of suicide risks reported by Copas and Fryer (1980). The data are given in Table 2.1. The second data set, observations of eruptions of Old Faithful geyser in Yellowstone National Park, USA, is taken from Weisberg (1980), and is reproduced in Table 2.2. I am most grateful to John Copas and to Sanford Weisberg for making these data sets available to me.

**Table 2.1 Lengths of treatment spells (in days) of control patients in suicide study.**

| | | | | | |
|---|---|---|---|---|---|
| 1 | 25 | 40 | 83 | 123 | 256 |
| 1 | 27 | 49 | 84 | 126 | 257 |
| 1 | 27 | 49 | 84 | 129 | 311 |
| 5 | 30 | 54 | 84 | 134 | 314 |
| 7 | 30 | 56 | 90 | 144 | 322 |
| 8 | 31 | 56 | 91 | 147 | 369 |
| 8 | 31 | 62 | 92 | 153 | 415 |
| 13 | 32 | 63 | 93 | 163 | 573 |
| 14 | 34 | 65 | 93 | 167 | 609 |
| 14 | 35 | 65 | 103 | 175 | 640 |
| 17 | 36 | 67 | 103 | 228 | 737 |
| 18 | 37 | 75 | 111 | 231 | |
| 21 | 38 | 76 | 112 | 235 | |
| 21 | 39 | 79 | 119 | 242 | |
| 22 | 39 | 82 | 122 | 256 | |

**Table 2.2 Eruption lengths (in minutes) of 107 eruptions of Old Faithful geyser.**

| | | | | | | |
|---|---|---|---|---|---|---|
| 4.37 | 3.87 | 4.00 | 4.03 | 3.50 | 4.08 | 2.25 |
| 4.70 | 1.73 | 4.93 | 1.73 | 4.62 | 3.43 | 4.25 |
| 1.68 | 3.92 | 3.68 | 3.10 | 4.03 | 1.77 | 4.08 |
| 1.75 | 3.20 | 1.85 | 4.62 | 1.97 | 4.50 | 3.92 |
| 4.35 | 2.33 | 3.83 | 1.88 | 4.60 | 1.80 | 4.73 |
| 1.77 | 4.57 | 1.85 | 3.52 | 4.00 | 3.70 | 3.72 |
| 4.25 | 3.58 | 3.80 | 3.77 | 3.75 | 2.50 | 4.50 |
| 4.10 | 3.70 | 3.80 | 3.43 | 4.00 | 2.27 | 4.40 |
| 4.05 | 4.25 | 3.33 | 2.00 | 4.33 | 2.93 | 4.58 |
| 1.90 | 3.58 | 3.73 | 3.73 | 1.82 | 4.63 | 3.50 |
| 4.00 | 3.67 | 1.67 | 4.60 | 1.67 | 4.00 | 1.80 |
| 4.42 | 1.90 | 4.63 | 2.93 | 3.50 | 1.97 | 4.28 |
| 1.83 | 4.13 | 1.83 | 4.65 | 4.20 | 3.93 | 4.33 |
| 1.83 | 4.53 | 2.03 | 4.18 | 4.43 | 4.07 | 4.13 |
| 3.95 | 4.10 | 2.72 | 4.58 | 1.90 | 4.50 | 1.95 |
| 4.83 | 4.12 | | | | | |

It is convenient to define some standard notation. Except where otherwise stated, it will be assumed that we are given a sample of n real observations $X_1, \ldots, X_n$ whose underlying density is to be estimated. The symbol $\hat{f}$ will be used to denote whatever density estimator is currently being considered.

### 2.2. Histograms

The oldest and most widely used density estimator is the histogram. Given an *origin* $x_0$ and *a bin width* $h$, we define the *bins* of the histogram to be the intervals $[x_0 + mh, x_0 + (m + 1)h]$ for positive and negative integers $m$. The intervals have been chosen closed on the left and open on the right for definiteness.

The histogram is then defined by

$$\hat{f}(x) = \frac{1}{nh}(\text{no. of } X_i \text{ in the same bin as } x).$$

Note that, to construct the histogram, we have to choose both an origin and a bin width; it is the choice of bin width which, primarily, controls the amount of smoothing inherent in the procedure.

The histogram can be generalized by allowing the bin widths to vary. Formally, suppose that we have any dissection of the real line into bins; then the estimate will be defined by

$$\hat{f}(x) = \frac{1}{n} \times \frac{(\text{no. of } X_i \text{ in the same bin as } x)}{(\text{width of bin containing } x)}.$$

The dissection into bins can either be carried out *a priori* or else in some way which depends on the observations themselves.

Those who are sceptical about density estimation often ask why it is ever necessary to use methods more sophisticated than the simple histogram. The case for such methods and the drawbacks of the histogram depend quite substantially on the context. In terms of various mathematical descriptions of accuracy, the histogram can be quite substantially improved upon, and this mathematical drawback translates itself into inefficient use of the data if histograms are used as density estimates in procedures like cluster analysis and nonparametric discriminant analysis. The discontinuity of histograms causes extreme difficulty if derivatives of the estimates are required. When density estimates are needed as intermediate components of other methods, the case for using alternatives to histograms is quite strong.

For the presentation and exploration of data, histograms are of course an extremely useful class of density estimates, particularly in the univariate case. However, even in one dimension, the choice of origin can have quite an effect. Figure 2.1 shows histograms of the Old Faithful eruption lengths constructed with the same bin width but different origins. Though the general message is the same in both cases, a non-statistician, particularly, might well get different impressions of, for example, the width of the left-hand peak and the separation of the two modes. Another example is given in Fig. 2.2; leaving aside the differences near the origin, one estimate suggests some structure near 250 which is completely obscured in the other. An experienced statistician would probably dismiss this as random error, but it is unfortunate that the occurrence or absence of this secondary peak in the presentation of the data is a consequence of the choice of origin, not of any choice of degree of smoothing or of treatment of the tails of the sample.

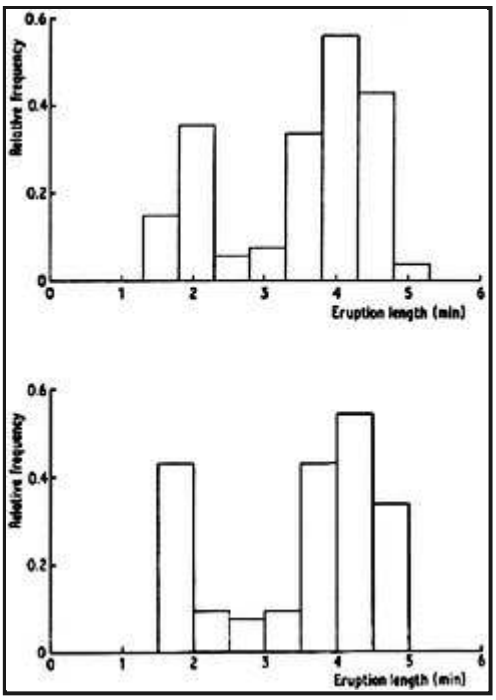


**Fig. 2.1** Histograms of eruption lengths of Old Faithful geyser.

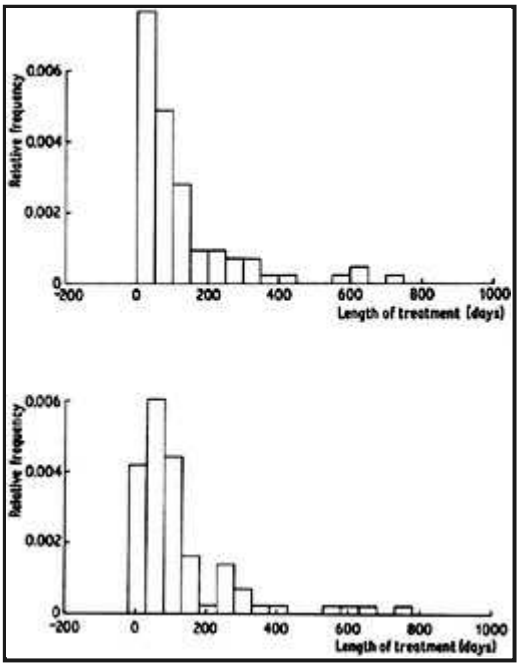


**Fig. 2.2** Histograms of lengths of treatment of control patients in suicide study.

Histograms for the graphical presentation of bivariate or trivariate data present several difficulties; for example, one cannot easily draw contour diagrams to represent the data, and the problems raised in the univariate case are exacerbated by the dependence of the estimates on the choice not only of an origin but also of the coordinate direction(s) of the grid of cells. Finally, it should be stressed that, in all cases, the histogram still requires a choice of the amount of smoothing.

Though the histogram remains an excellent tool for data presentation, it is worth at least considering the various alternative density estimates that are available.

### 2.3. The naive estimator

From the definition of a probability density, if the random variable $X$ has density $f$, then

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h).$$

For any given $h$, we can of course estimate $P(x - h < X < x + h)$ by the proportion of the sample falling in the interval $(x - h, x + h)$. Thus a natural estimator $\hat{f}$ of the density is given by choosing a small number $h$ and setting

$$\hat{f}(x) = \frac{1}{2hn} \quad [\text{no. of } X_1, \ldots, X_n \text{ falling in } (x - h, x + h)];$$

we shall call this the naive estimator.

To express the estimator more transparently, define the weight function $w$ by

$$w(x) = \begin{cases} \frac{1}{2} & \text{if } |x| < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

Then it is easy to see that the naive estimator can be written

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

It follows from (2.1) that the estimate is constructed by placing a `box' of width $2h$ and height $(2n\,h)^{-1}$ on each observation and then summing to obtain the estimate. We shall return to this interpretation below, but it is instructive first to consider a connection with histograms.

Consider the histogram constructed from the data using bins of width $2h$. Assume that no observations lie exactly at the edge of a bin. If $x$ happens to be at the centre of one of the histogram bins, it follows at once from (2.1) that the naive estimate $\hat{f}(x)$ will be exactly the ordinate of the histogram at $x$. Thus the naive estimate can be seen to be an attempt to construct a histogram where every point is the centre of a sampling interval, thus freeing the histogram from a particular choice of bin positions. The choice of bin width still remains and is governed by the parameter $h$, which controls the amount by which the data are smoothed to produce the estimate.

The naive estimator is not wholly satisfactory from the point of view of using density estimates for presentation. It follows from the definition that $\hat{f}$ is not a continuous function, but has jumps at the points $X_i \pm h$ and has zero derivative everywhere else. This gives the estimates a somewhat ragged character which is not only aesthetically undesirable, but, more seriously, could provide the untrained observer with a misleading impression. Partly to overcome this difficulty, and partly for other technical reasons given later, it is of interest to consider the generalization of the naive estimator given in the following section.

A density estimated using the naive estimator is given in Fig. 2.3. The `stepwise' nature of the estimate is clear. The boxes used to construct the estimate have the same width as the histogram bins in Fig. 2.1.

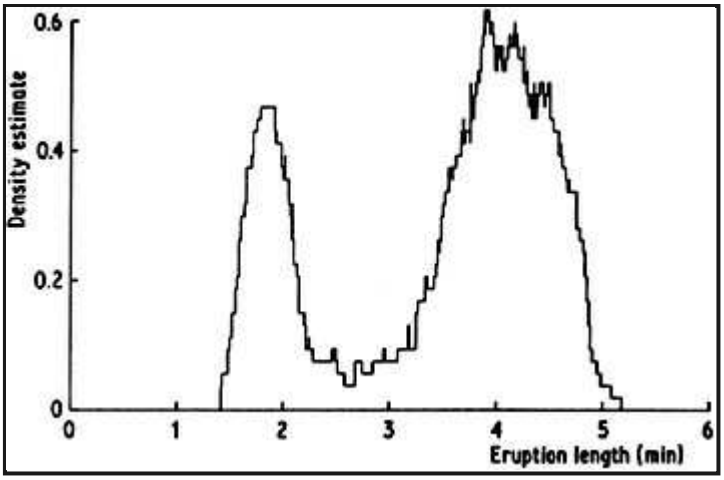


**Fig. 2.3** Naive estimate constructed from Old Faithful geyser data, $h = 0.25$.

### 2.4. The kernel estimator

It is easy to generalize the naive estimator to overcome some of the difficulties discussed above. Replace the weight function w by a kernel function $K$ which satisfies the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1. \tag{2.2}$$

Usually, but not always, $K$ will be a symmetric probability density function, the normal density, for instance, or the weight function $w$ used in the definition of the naive estimator. By analogy with the definition of the naive estimator, the *kernel estimator* with kernel $K$ is defined by

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

where $h$ is the *window width*, also called the *smoothing parameter* or *bandwidth* by some authors. We shall consider some mathematical properties of the kernel estimator later, but first of all an intuitive discussion with some examples may be helpful.

Just as the naive estimator can be considered as a sum of `boxes' centred at the observations, the kernel estimator is a sum of `bumps' placed at the observations. The kernel function $K$ determines the shape of the bumps while the window width $h$ determines their width. An illustration is given in Fig. 2.4, where the individual bumps $n^{-1} h^{-1} K\{(x - X_i)/h\}$ are shown as well as the estimate $\hat{f}$ constructed by adding them up. It should be stressed that it is not usually appropriate to construct a density estimate from such a small sample, but that a sample of size 7 has been used here for the sake of clarity.

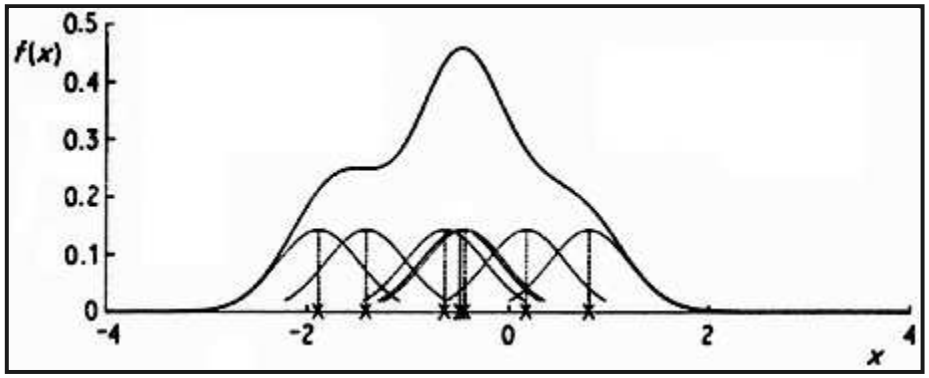


**Fig. 2.4** Kernel estimate showing individual kernels. Window width 0.4.

The effect of varying the window width is illustrated in Fig. 2.5. The limit as $h$ tends to zero is (in a sense) a sum of Dirac delta function spikes at the observations, while as $h$ becomes large, all detail, spurious or otherwise, is obscured.

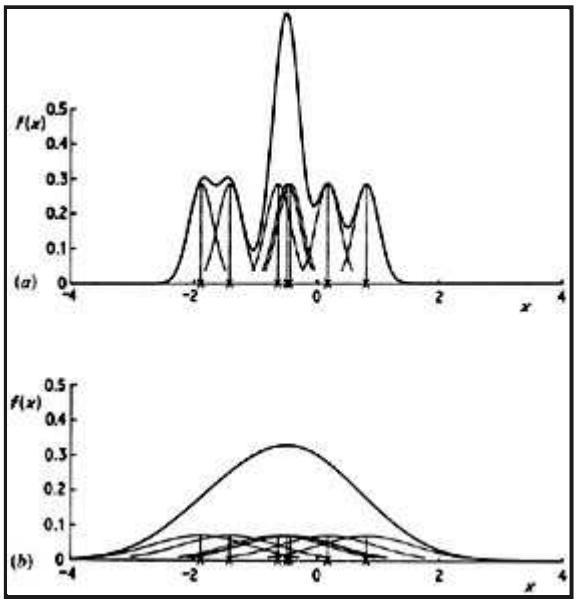


**Fig. 2.5** Kernel estimates showing individual kernels.
Window widths: (a) 0.2; (b) 0.8.

Another illustration of the effect of varying the window width is given in Fig. 2.6. The estimates here have been constructed from a pseudo-random sample of size 200 drawn from the bimodal density given in Fig. 2.7. A normal kernel has been used to construct the estimates. Again it should be noted that if $h$ is chosen too small then spurious fine structure becomes visible, while if $h$ is too large then the bimodal nature of the distribution is obscured. A kernel estimate for the Old Faithful data is given in Fig. 2.8. Note that the same broad features are visible as in Fig. 2.3 but the local roughness has been eliminated.

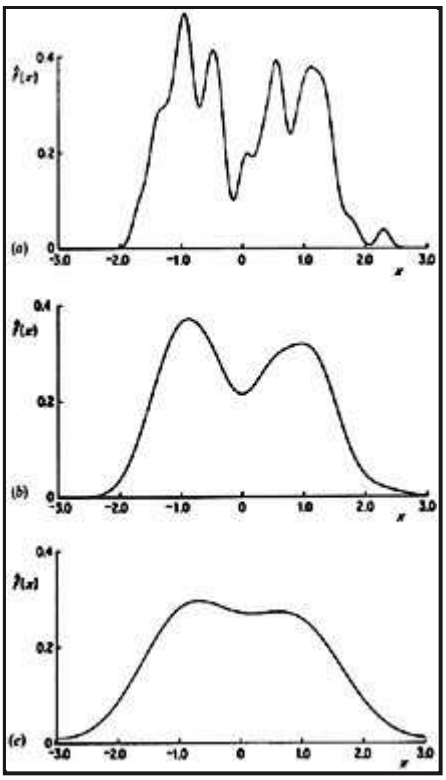


**Fig. 2.6** Kernel estimates for 200 simulated data points drawn from a bimodal density. Window widths: (a) 0.1; (b) 0.3; (c) 0.6.

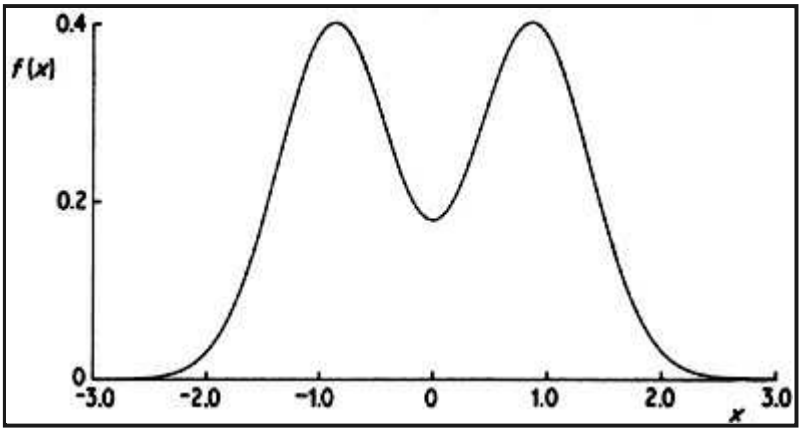


**Fig. 2.7** True bimodal density underlying data used in Fig. 2.6.

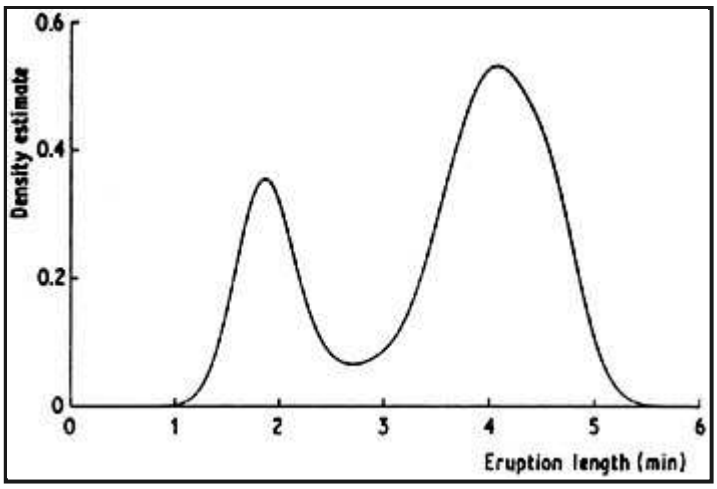


**Fig. 2.8** Kernel estimate for Old Faithful geyser data, window width 0.25.

Some elementary properties of kernel estimates follow at once from the definition. Provided the kernel $K$ is everywhere non-negative and satisfies the condition (2.2) - in other words is a probability density function - it will follow at once from the definition that $\hat{f}$ will itself be a probability density. Furthermore, $\hat{f}$ will inherit all the continuity and differentiability properties of the kernel $K$, so that if, for example, $K$ is the normal density function, then $\hat{f}$ will be a smooth curve having derivatives of all orders. There are arguments for sometimes using kernels which take negative as well as positive values, and these will be discussed in Section 3.6. If such a kernel is used, then the estimate may itself be negative in places. However, for most practical purposes non-negative kernels are used.

Apart from the histogram, the kernel estimator is probably the most commonly used estimator and is certainly the most studied mathematically. It does, however, suffer from a slight drawback when applied to data from long-tailed distributions. Because the window width is fixed across the entire sample, there is a tendency for spurious noise to appear in the tails of the estimates; if the estimates are smoothed sufficiently to deal with this, then essential detail in the main part of the distribution is masked. An example of this behaviour is given by disregarding the fact that the suicide data are naturally non-negative and estimating their density treating them as observations on $(-\infty, \infty)$. The estimate shown in Fig. 2.9(a) with window width 20 is noisy in the right-hand tail, while the estimate (b) with window width 60 still shows a slight bump in the tail and yet exaggerates the width of the main bulge of the distribution. In order to deal with this difficulty, various adaptive methods have been proposed, and these are discussed in the next two sections. A detailed consideration of the kernel method for univariate data will be given in Chapter 3, while Chapter 4 concentrates on the generalization to the multivariate case.

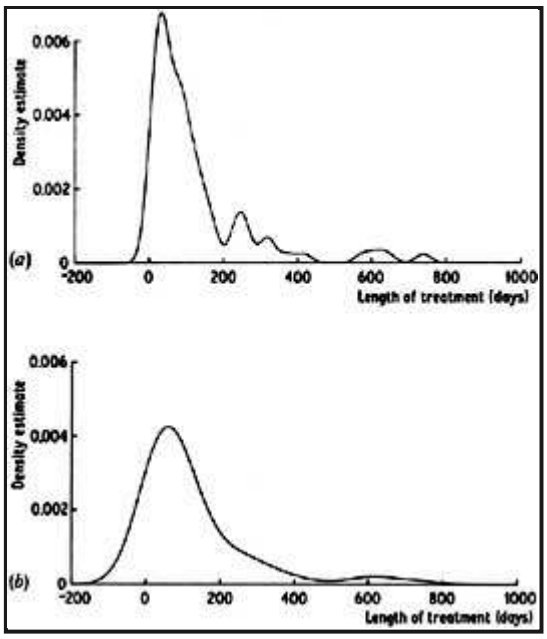


**Fig. 2.9** Kernel estimates for suicide study data. Window widths: (a) 20; (b) 60.

### 2.5. The nearest neighbour method

The nearest neighbour class of estimators represents an attempt to adapt the amount of smoothing to the `local' density of data. The degree of smoothing is controlled by an integer $k$, chosen to be considerably smaller than the sample size; typically $k \approx n^{1/2}$. Define the distance $d(x, y)$ between two points on the line to be $|x - y|$ in the usual way, and for each $t$ define

$$d_1(t) \leq d_2(t) \leq \ldots \leq d_n(t)$$

to be the distances, arranged in ascending order, from $t$ to the points of the sample.

The *kth nearest neighbour density estimate* is then defined by

$$\hat{f}(t) = \frac{k}{2nd_k(t)} \tag{2.3}$$

In order to understand this definition, suppose that the density at $t$ is $f(t)$. Then, of a sample of size $n$, one would expect about $2r\,n\,f(t)$ observations to fall in the interval $[t - r, t + r]$ for each $r > 0$; see the discussion of the naive estimator in Section 2.3 above. Since, by definition, exactly $k$ observations fall in the interval $[t - d_k(t), t + d_k(t)]$, an estimate of the density at $t$ may be obtained by putting

$$k = 2d_k(t)n\hat{f}(t);$$

this can be rearranged to give the definition of the $k$th nearest neighbour estimate.

While the naive estimator is based on the number of observations falling in a box of fixed width centred at the point of interest, the nearest neighbour estimate is inversely proportional to the size of the box needed to contain a given number of observations. In the tails of the distribution, the distance $d_k(t)$ will be larger than in the main part of the distribution, and so the problem of undersmoothing in the tails should be reduced.

Like the naive estimator, to which it is related, the nearest choice neighbour estimate as defined in (2.3) is not a smooth curve. The function $d_k(t)$ can easily be seen to be continuous, but its derivative will have a discontinuity at every point of the form $1/2(X_{(j)} + X_{(j+k)})$, where $X_{(j)}$ are the order statistics of the sample. It follows at once from these remarks and from the definition

that $\hat{f}$ will be positive and continuous everywhere, but will have discontinuous derivative at all the same points as $d_k$. In contrast to the kernel estimate, the nearest neighbour estimate will not itself be a probability density, since it will not integrate to unity. For $t$ less than the smallest data point, we will have $d_k(t) = X_{(n-k+1)}$ and for $t > X_{(n)}$ we will have $d_k(t) = t - X_{(n-k+1)}$. Substituting into (2.3), it follows that $\int_{-\infty}^{\infty} (t)dt$ is infinite and that the tails of $\hat{f}$ die away at rate $t^{-1}$, in other words extremely slowly. Thus the nearest neighbour estimate is unlikely to be appropriate if an estimate of the entire density is required. Figure 2.10 gives a nearest neighbour density estimate for the Old Faithful data. The heavy tails and the discontinuities in the derivative are clear.

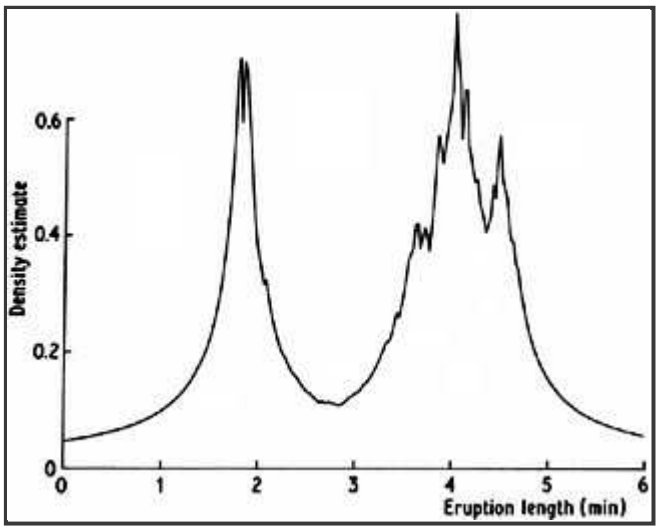


**Fig. 2.10** Nearest neighbour estimate for Old Faithful geyser data, $k = 20$.

It is possible to generalize the nearest neighbour estimate to provide an estimate related to the kernel estimate. As in Section 2.4, let $K(x)$ be a kernel function integrating to one. Then the *generalized kth nearest neighbour estimate* is defined by

$$\hat{f}(t) = \frac{1}{nd_k(t)} \sum_{i=1}^{n} K\left(\frac{t - X_i}{d_k(t)}\right). \tag{2.4}$$

It can be seen at once that $\hat{f}(t)$ is precisely the kernel estimate evaluated at t with window width $d_k(t)$. Thus the overall amount of smoothing is governed by the choice of the integer $k$, but the window width used at any particular point depends on the density of observations near that point.

The ordinary $k$th nearest neighbour estimate is the special case of (2.4) when $K$ is the uniform kernel $w$ of (2.1); thus (2.4) stands in the to same relation to (2.3) as the kernel estimator does to the naive ed. estimator. However, the derivative of the generalized nearest neighbour estimate will be discontinuous at all the points where the function $d_k(t)$ has discontinuous derivative. The precise integrability and tail properties will depend on the exact form of the kernel, and to will not be discussed further here.

Further discussion of the nearest neighbour approach will be given in Section 5.2.

**2.6. The variable kernel method**

The variable kernel method is somewhat related to the nearest neighbour approach and is another method which adapts the amount of smoothing to the local density of data. The estimate is constructed similarly to the classical kernel estimate, but the scale parameter of the 'bumps' placed on the data points is allowed to vary 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 comprising the other $n$ - 1 data points. Then the *variable kernel estimate* with smoothing parameter $h$ is defined by

$$\hat{f}(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{hd_{j,k}} K\left(\frac{t - X_j}{hd_{j,k}}\right). \tag{2.5}$$

The window width of the kernel placed on the point $X_j$ is proportional to $d_{j,k}$, so that data points in regions where the data are sparse will have flatter kernels associated with them. For any fixed $k$, the overall degree of smoothing will depend on the parameter $h$. The choice of $k$ determines how responsive the window width choice will be to very local detail.

Some comparison of the variable kernel estimate with the generalized nearest neighbour estimate (2.4) may be instructive. In (2.4) the window width used to construct the estimate at $t$ depends on the distances from $t$ to the data points; in (2.5) the window widths are independent of the point $t$ at which the density is being estimated, and depend only on the distances between the data points.

In contrast with the generalized nearest neighbour estimate, the variable kernel estimate will itself be a probability density function provided the kernel $K$ is; that is an immediate consequence of the definition. Furthermore, as with the ordinary kernel estimator, all the local smoothness properties of the kernel will be inherited by the estimate. In Fig. 2.11 the method is used to obtain an estimate for the suicide data. The noise in the tail of the curve has been eliminated, but it is interesting to note that the method exposes some structure in the main part of the distribution which is not really visible even in the undersmoothed curve in Figure 2.9.

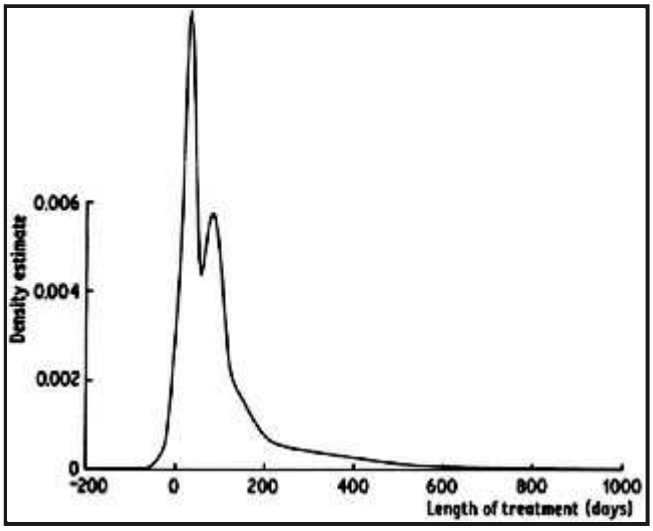


**Fig. 2.11** Variable kernel estimate for suicide study data, $k = 8$, $h = 5$.

In Section 5.3, the variable kernel method will be considered in greater detail, and in particular an important generalization, called the adaptive kernel method, will be introduced.

### 2.7. Orthogonal series estimators

Orthogonal series estimators approach the density estimation problem from quite a different point of view. They are best explained by a specific example. Suppose that we are trying to estimate a density $f$ on the unit interval [0, 1]. The idea of the orthogonal series method is then to estimate $f$ by estimating the coefficients of its Fourier expansion.

Define the sequence $\phi_\nu(x)$ by

$$\left.\begin{aligned} \phi_0(x) &= 1 \\ \phi_{2r-1}(x) &= \sqrt{2}\cos 2\pi r x \\ \phi_{2r}(x) &= \sqrt{2}\sin 2\pi r x \end{aligned}\right\} r = 1, 2, \ldots$$

Then, by standard mathematical analysis, $f$ can be represented as the Fourier series $\sum_{\nu=0}^{\infty} f_\nu \phi_\nu$, where, for each $\nu \geq 0$,

$$f_\nu = \int_0^1 f(x)\phi_\nu(x)dx. \qquad (2.6)$$

For a discussion of the sense in which *f* is represented by the series, see, for example, Kreider et al. (1966).

Suppose *X* is a random variable with density *f*. Then (2.6) can be written

$$f_\nu = E\phi_\nu(X)$$

and hence a natural, and unbiased, estimator of $f_\nu$ based on a sample $X_1$,..., $X_n$ from *f* is

$$\hat{f}_\nu = \frac{1}{n}\sum_{i=1}^{n}\phi_\nu(X_i).$$

Unfortunately, the sum $\Sigma_{\nu=0}^{\infty}\ \hat{f}_\nu\ \phi_\nu$ will not be a good estimate of *f*, but will `converge' to a sum of delta functions at the observations; to see this, let

$$\omega(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - X_i) \qquad (2.7)$$

where $\delta$ is the Dirac delta function. Then, for each $\nu$,

$$\hat{f}_\nu = \int_0^1 \omega(x)\phi_\nu(x)dx$$

and so the $\hat{f}_\nu$ are exactly the Fourier coefficients of the function $\omega$.

In order to obtain a useful estimate of the density *f*, it is necessary to smooth $\omega$ by applying a low-pass filter to the sequence of coefficients $\hat{f}_\nu$. The easiest way to do this is to truncate the expansion $\Sigma\ \hat{f}_\nu\ \phi_\nu$ at some point. Choose an integer *K* and define the density estimate $\hat{f}$ by

$$\hat{f}(x) = \sum_{\nu=0}^{K}\hat{f}_\nu\phi_\nu(x). \qquad (2.8)$$

The choice of the cutoff point *K* determines the amount of smoothing.

A more general approach is to taper the series by a sequence of weights $\lambda_\nu$, which satisfy $\lambda_\nu \to 0$ as $\nu \to \infty$, to obtain the estimate

$$\hat{f}(x) = \sum_{\nu=0}^{\infty}\lambda_\nu\hat{f}_\nu\phi_\nu(x).$$

The rate at which the weights $\lambda_\nu$ converge to zero will determine the amount of smoothing.

Other orthogonal series estimates, no longer necessarily confined to data lying on a finite interval, can be obtained by using different orthonormal sequences of functions. Suppose $a(x)$ is a weighting function and $(\psi_\nu)$ is a series satisfying, for $\mu$ and $\nu \geq 0$,

$$\int_{-\infty}^{\infty}\psi_\mu(x)\psi_\nu(x)a(x)dx = \begin{cases} 1 & \mu = \nu \\ 0 & \text{otherwise.} \end{cases}$$

For instance, for data resealed to have zero mean and unit variance, $a(x)$ might be the function $e^{-x^2/2}$ and the $\psi_\nu$ multiples of the Hermite polynomials; for details see Kreider et al. (1966).

The sample coefficients will then be defined by

$$\hat{f}_\nu = \frac{1}{n}\sum_i \psi_\nu(X_i)a(X_i)$$

but otherwise the estimates will be defined as above; possible estimates are

$$\hat{f}(x) = \sum_{\nu=0}^{K} \hat{f}_\nu \psi_\nu(x) \tag{2.9}$$

or

$$\hat{f}(x) = \sum_{\nu=0}^{\infty} \lambda_\nu \hat{f}_\nu \psi_\nu(x). \tag{2.10}$$

The properties of estimates obtained by the orthogonal series method depend on the details of the series being used and on the system of weights. The Fourier series estimates will integrate to unity, provided $\lambda_0 = 1$, since

$$\int_0^1 \phi_\nu(x)dx = 0 \quad \text{for all } \nu > 0$$

and $\hat{f}_0$ will always be equal to one. However, except for rather special choices of the weights $\lambda_\nu$, $\hat{f}$ cannot be guaranteed to be non-negative. The local smoothness properties of the estimates will again depend on the particular case; estimates obtained from (2.8) will have derivatives of all orders.

**2.8. Maximum penalized likelihood estimators**

The methods discussed so far are all derived in an *ad hoc* way from the definition of a density. It is interesting to ask whether it is possible to apply standard statistical techniques, like maximum likelihood, to density estimation. The *likelihood* of a curve $g$ as density underlying a set of independent identically distributed observations is given by

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

This likelihood has no finite maximum over the class of all densities. To see this, let $\hat{f}_h$ be the naive density estimate with window width 1/2 $h$; then, for each $i$,

$$\hat{f}_h(X_i) \geq \frac{1}{nh}$$

and so

$$\prod \hat{f}_h(X_i) \geq n^{-n}h^{-n} \to \infty \quad \text{as } h \to 0.$$

Thus the likelihood can be made arbitrarily large by taking densities approaching the sum of delta functions $\omega$ as defined in (2.7) above, and it is not possible to use maximum likelihood directly for density estimation without placing restrictions on the class of densities over which the likelihood is to be maximized.

There are, nevertheless, possible approaches related to maximum likelihood. One method is to incorporate into the likelihood a term which describes the roughness - in some sense - of the curve under consideration. Suppose *R(g)* is a functional which quantifies the roughness of *g*. One possible choice of such a functional is

$$R(g) = \int_{-\infty}^{\infty} (g'')^2. \tag{2.11}$$

Define the *penalized log likelihood* by

$$l_\alpha(g) = \sum_{i=1}^{n} \log g(X_i) - \alpha R(g) \tag{2.12}$$

where $\alpha$ is a positive smoothing parameter.

The penalized log likelihood can be seen as a way of quantifying the conflict between smoothness and goodness-of-fit to the data, since the log likelihood term $\sum \log g(X_i)$ measures how well *g* fits the data. The probability density function $\hat{f}$ is said to be a *maximum penalized likelihood density estimate* if it maximizes $l_\alpha(g)$ over the class of all curves *g* which satisfy $\int_{-\infty}^{\infty} g = 1$, $g(x) \geq 0$ for all *x*, and $R(g) < \infty$. The parameter $\alpha$ controls the amount of smoothing since it determines the `rate of exchange' between smoothness and goodness-of-fit; the smaller the value of $\alpha$, the rougher - in terms of $R(\hat{f})$ - will be the corresponding maximum penalized likelihood estimator. Estimates obtained by the maximum penalized likelihood method will, by definition, be probability densities. Further details of these estimates will be given in Section 5.4.

**2.9. General weight function estimators**

It is possible to define a general class of density estimators which includes several of the estimators discussed above. Suppose that $w(x, y)$ is a function of two arguments, which in most cases will satisfy the conditions

$$\int_{-\infty}^{\infty} w(x, y)dy = 1 \tag{2.13}$$

and

$$w(x, y) \geq 0 \quad \text{for all } x \text{ and } y. \tag{2.14}$$

We should think of *w* as being defined in such a way that most of the weight of the probability density $w(x, \cdot)$ falls near *x*. An estimate of the density underlying the data may be obtained by putting

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} w(X_i, t). \tag{2.15}$$

We shall refer to estimates of the form (2.15) as *general weight function estimates*. It is clear from (2.15) that the conditions (2.13) and (2.14) will be sufficient to ensure that $\hat{f}$ is a probability density function, and that the smoothness properties of $\hat{f}$ will be inherited from those of the functions $w(x, \cdot)$. This class of estimators can be thought of in two ways. Firstly, it is a unifying concept which makes it possible, for example, to obtain theoretical results applicable to a whole range of apparently distinct estimators. On the other hand, it is possible to define useful estimators which do not fall into any of the classes discussed in previous sections but which are nevertheless of the form (2.15). We shall discuss such an estimator later in this section.

To obtain the histogram as a special case of (2.15), set

$$w(x,y) = \begin{cases} \dfrac{1}{h(x)} & \text{if } x \text{ and } y \text{ fall in the same bin} \\ 0 & \text{otherwise,} \end{cases}$$

where $h(x)$ is the width of the bin containing $x$.

The kernel estimate can be obtained by putting

$$w(x,y) = \frac{1}{h} K\left(\frac{y-x}{h}\right). \tag{2.15a}$$

The orthogonal series estimate as defined in (2.8) above is given by putting

$$w(x,y) = \sum_{\nu=0}^{K} \phi_\nu(x)\phi_\nu(y);$$

the generalization (2.10) is obtained from

$$w(x,y) = \sum_{\nu=0}^{\infty} \lambda_\nu a(x)\psi_\nu(x)\psi_\nu(y).$$

Another example of a general weight function estimator can be obtained by considering how we would deal with data which lie naturally on the positive half-line, a topic which will be discussed at greater length in Section 2.10. One way of dealing with such data is to use a weight function which is, for each fixed $x$, a probability density which has support on the positive half-line and which has its mass concentrated near $x$. For example, one could choose $w(x, \cdot)$ to be a gamma density with mean $x$ or a log-normal density with median $x$; in both cases, the amount of smoothing would be controlled by the choice of the shape parameter. It should be stressed that the densities $w(x, \cdot)$ will become progressively more concentrated as $x$ approaches zero and hence the amount of smoothing applied near zero will be much less than in the right-hand tail. Using the log-normal weight function corresponds precisely to applying the kernel method, with normal kernel, to the logarithms of the data points, and then performing the appropriate inverse transformation.

An example for which this treatment is clearly appropriate is the suicide data discussed earlier. Figure 2.12 gives a kernel estimate of the density underlying the logarithms of the data values; the corresponding density estimate for the raw data is given in Fig. 2.13. The relative undersmoothing near the origin is made abundantly clear by the large spike in the estimate.

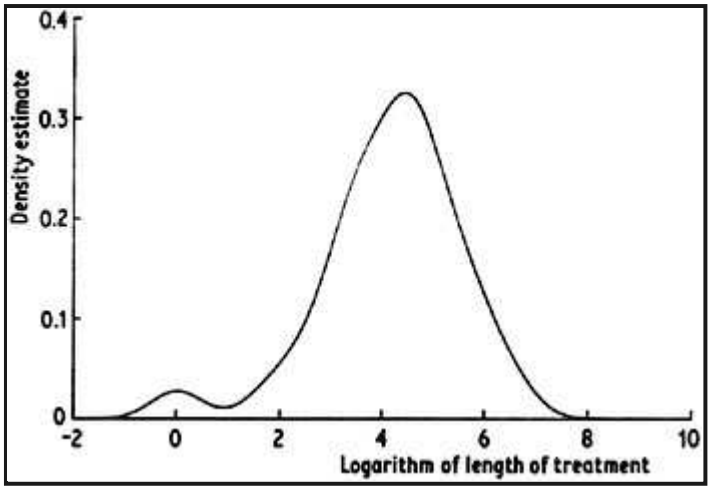


**Fig. 2.12** ernel estimate for logarithms of suicide study data, window width 0.5.

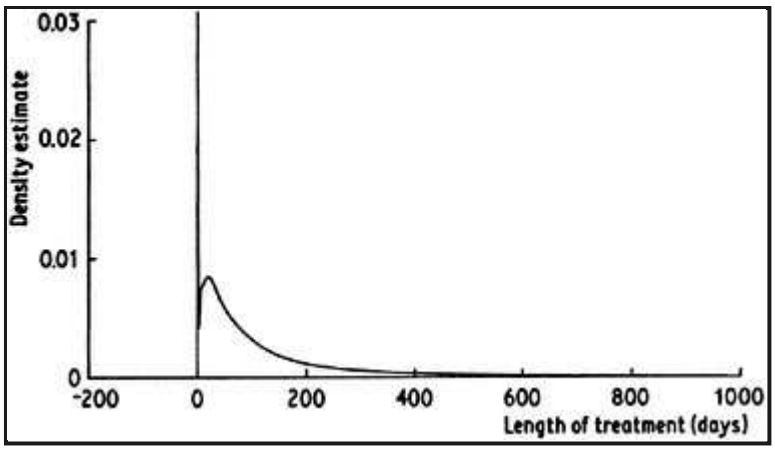


**Fig. 2.13** Log-normal weight function estimate for suicide study data, obtained by transformation of Fig. 2.12. Note that the vertical scale differs from that used in previous figures for this data set.

**2.10. Bounded domains and directional data**

It is very often the case that the natural domain of definition of a density to be estimated is not the whole real line but an interval bounded on one or both sides. For example, both the suicide data and the Old Faithful eruption lengths are measurements of positive quantities, and so it will be preferable for many purposes to obtain density estimates $\hat{f}$ for which $\hat{f}(x)$ is zero for all negative $x$. In the case of the Old Faithful data, the problem is really of no practical importance, since there are no observations near zero, and so the lefthand boundary can simply be ignored. The suicide data are of course quite another matter. For exploratory purposes it will probably suffice to ignore the boundary condition, but for other applications, and for presentation of the data, estimates which give any weight to the negative numbers are likely to be unacceptable.

One possible way of ensuring that $\hat{f}(x)$ is zero for negative $x$ is simply to calculate the estimate for positive $x$ ignoring the boundary conditions, and then to set $\hat{f}(x)$ to zero for negative $x$. A drawback of this approach is that if we use a method, for example the kernel method, which usually produces estimates which are probability densities, the estimates obtained will no longer integrate to unity. To make matters worse, the contribution to $\int_0^\infty \hat{f}(x)\,dx$ of points near zero will be much less than that of points well away from the boundary, and so, even if the estimate is rescaled to make it a probability density. the weight of the distribution near zero will be underestimated.

Some of the methods can be adapted to deal directly with data on the half-line. For example, we could use an orthogonal series estimate of the form (2.9) or (2.10) with functions $\psi_\nu$ which were orthonormal with respect to a weighting function $a$ which is zero for $x < 0$. The maximum penalized likelihood method can be adapted simply by constraining $g(x)$ to be zero for negative $x$, and using a roughness penalty functional which only depends on the behaviour of $g$ on $(0, \infty)$.

Another possible approach is to transform the data, for example by taking logarithms as in the example given in Section 2.9 above. If the density estimated from the logarithms of the data is $\hat{g}$, then standard arguments lead to

$$\hat{f}(x) = \frac{1}{x}\hat{g}(\log x) \quad \text{for } x > 0.$$

It is of course the presence of the multiplier $1/x$ that gives rise to the spike in Fig. 2.13; not with standing difficulties of this kind, Copas and Fryer (1980) did find estimates based on logarithmic transforms to be very useful with some other data sets.

It is possible to use other adaptations of methods originally designed for the whole real line. Suppose we augment the data by adding the reflections of all the points in the boundary, to give the set $\{X_1, -X_1, X_2, -X_2, \ldots\}$. If a kernel estimate $f^*$ is constructed from this data set of size $2n$, then an estimate based on the original data can be given by putting

$$\hat{f}(x) = \begin{cases} 2f^*(x) & \text{for } x \geq 0 \\ 0 & \text{for } x < 0. \end{cases}$$

This estimate corresponds to a general weight function estimator with, for $x$ and $y > 0$,

$$w(x,y) = \frac{1}{h}K\left(\frac{y-x}{h}\right) + \frac{1}{h}K\left(\frac{y+x}{h}\right).$$

Provided the kernel is symmetric and differentiable, some easy manipulation shows that the estimate will always have zero derivative at the boundary. If the kernel is a symmetric probability density, the estimate will be a probability density. It is clear that it is not usually necessary to reflect the whole data set, since if $X_i/h$ is sufficiently large, the reflected point $-X_i/h$ will not be felt in the calculation of $f^*(x)$ for $x \geq 0$, and hence we need only reflect points near 0. For example, if $K$ is the normal kernel there is no practical need to reflect points $X_i > 4h$.

This reflection technique can be used in conjunction with any method for density estimation on the whole line. With most methods estimates which satisfy $\hat{f}'(0+) = 0$ will be obtained.

Another, related, technique forces $\hat{f}(0+) = 0$ rather than $\hat{f}'(0+) = 0$. Reflect the data as before, but give the reflected points weight -1 in the calculation of the estimate; thus the estimate is, for $x \geq 0$,

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x-X_i}{h}\right) - K\left(\frac{x+X_i}{h}\right)\right]. \qquad (2.16)$$

We shall call this technique *negative reflection*. Estimates constructed from (2.16) will no longer integrate to unity, and indeed the total contribution to $\int_0^\infty \hat{f}(x)dx$ from points near the boundary will be small. Whether estimates of this form are useful depends on the context.

All the remarks of this section can be extended to the case where the required support of the estimator is a finite interval $[a, b]$. Transformation methods can be based on transformations of the form

$$Y_i = H^{-1}\left(\frac{X_i - a}{b-a}\right)$$

where $H$ is any cumulative probability distribution function strictly increasing on $(-\infty, \infty)$. Generally, the estimates obtained by transformation back to the original scale will be less smoothed for points near the boundaries. The reflection methods are easily generalized. It is necessary to reflect in both boundaries and it is of course possible to use ordinary reflection in one boundary and negative reflection in the other, if the corresponding boundary conditions are required.

Another way of dealing with data on a finite interval $[a, b]$ is to impose periodic or `wrap around' boundary conditions. Of course this approach is particularly useful if the data are actually directions or angles; the turtle data considered in Section 1.2 were of this kind. For simplicity, suppose that the interval on which the data naturally lie is [0, 1]. which can be regarded as a circle of circumference l: more general intervals are dealt with analogously. If we want to use a method like the kernel method, a possible approach is to wrap the kernel round the circle. Computationally it may be simpler to augment the data set by replicating shifted copies of it on the intervals [-1,0] and [1,2], to obtain the set

$$\begin{aligned}\{&X_1 - 1, X_2 - 1, \ldots, X_n - 1, X_1, X_2, \ldots, X_n,\\ &X_1 + 1, X_2 + 1, \ldots, X_n + 1\}.\end{aligned} \qquad (2.17)$$

in principle we should continue to replicate on intervals further away from [0. 1], but that is rarely necessary in practice. Applying the kernel method or one of its variants to the augmented data set will give an estimate on [0, 1] which has the required boundary property; of course the factor 1 / $n$ should be retained in the definition of the estimate even though the augmented data set has more than $n$ members.

The orthogonal series estimates based on Fourier series will automatically impose periodic boundary conditions, because of the periodicity of the functions $\phi_\nu$ of section 2.7.

### 2.11. Discussion and bibliography

A brief survey of the kind conducted in this chapter of course asks far more questions than it answers, and some of these questions will be the subject of discussion in subsequent chapters. The overriding problems are the choice of what method to use in any given practical context and, given that a particular method is being used, how to choose the various parameters needed by the method. The remarks already made about the mathematical properties of the estimates obtained by various procedures will of course be important in making these decisions. To obtain a fuller understanding of the importance and consequences of the various choices it is essential to investigate the statistical properties of the various methods and also to consider the difficulties involved in computing the estimates.

This chapter has by no means considered all the methods available for density estimation. Generalizations and other approaches are considered in later chapters of this book, and in the other books and surveys mentioned in Section 1.3.

The naive estimator was introduced by Fix and Hodges (1951) in an unpublished report; the first published paper to deal explicitly with probability density estimation was by Rosenblatt (1956), who discussed both the naive estimator and the more general kernel estimator. Whittle (1958) formulated the general weight function class of estimators, while the orthogonal series estimator was introduced by Cencov (1962). The nearest neighbour estimate was first considered by Loftsgaarden and Quesenberry (1965), while the variable kernel method is due to Breiman, Meisel and Purcell (1977), though Wertz (1978, p. 59) refers to presumably independent but related work by Victor. The maximum penalized likelihood approach was first applied to density estimation by Good and Gaskins (1971). The reflection and replication techniques of Section 2.10 were introduced and illustrated by Boneva, Kendall and Stefanov (1971), while the transformation technique is discussed by Copas and Fryer (1980).