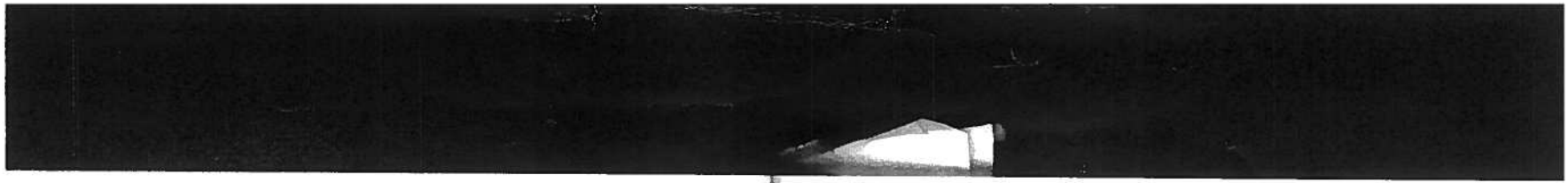


1

DENSITY ESTIMATION FOR EXPLORING DATA

1.1 Introduction

The concept of a probability density function is a central idea in statistics. Its role in statistical modelling is to encapsulate the pattern of random variation in the data which is not explained by the other structural terms of a model. In many settings this role, while important, is a secondary one, with the principal focus resting on the nature of covariate or other effects. However, there are also situations where the detailed shape of the underlying density function is itself of primary interest. In this chapter, a number of examples of this are given and the ideas behind the construction of a smooth estimate of a density function are introduced. Some of the main issues associated with using these estimates are raised.

1.2 Basic ideas

In a study of the development of aircraft technology, Saviotti and Bowman (1984) analysed data on aircraft designs. The first author subsequently collected more extensive data, mainly from Jane's (1978), on six simple characteristics of aircraft designs which have appeared during the twentieth century. These six characteristics are:

- ◇ total engine power (kW);
- ◇ wing span (m);
- ◇ length (m);
- ◇ maximum take-off weight (kg);
- ◇ maximum speed (km h^{-1});
- ◇ range (km).

The aim of this study was to explore techniques for describing the development of this technology over time, and in particular to highlight particular directions in which this development occurred. Techniques which successfully described the patterns in this well understood area might then be applied to good effect in other, less well known, areas.

Clearly, events such as two world wars have had an enormous impact on aircraft development. In view of this, the data will be considered in separate subgroups corresponding to the years 1914–1935, 1936–1955 and 1956–1984. The left panel of Fig. 1.1 displays a histogram of the data on wing span from the third time period. Since all of the six variables have markedly skewed distributions, each will be examined on a log scale.

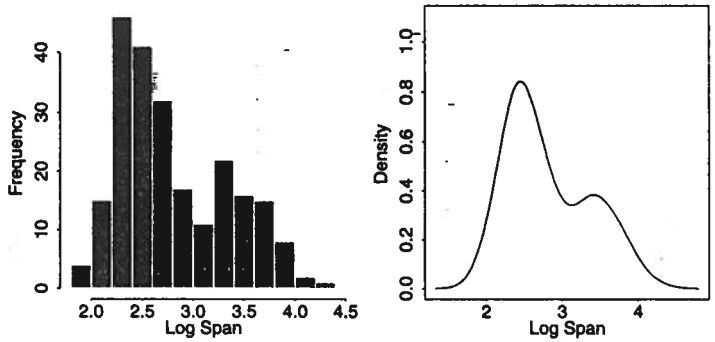


FIG. 1.1. Histogram and density estimate of the aircraft span data, on a log scale, for the third time period.

The histogram is, of course, a widely used tool for displaying the distributional shape of a set of data. More specifically, its usefulness lies in the fact that it indicates the shape of the underlying density function. For example, with the span data it is clear that some skewness exists, even on the log scale.

The right panel of Fig. 1.1 displays an alternative estimate of the density function as a smooth curve. In order to discuss the construction of estimators of this type, it is helpful to consider first the construction of a histogram. This begins by dividing the sample space into a number of intervals. Each observation contributes a 'box' which is then placed over the appropriate interval. This is illustrated in the left panel of Fig. 1.2, which uses a small subsample of the span data for the purpose of illustration. If y denotes the point at which the density $f(y)$ must be estimated, then the histogram may be written as

$$\tilde{f}(y) = \sum_{i=1}^{n} I(y - \tilde{y}_i; h),$$

where $\{y_1, \ldots, y_n\}$ denote the observed data, $\tilde{y}_i$ denotes the centre of the interval in which y_i falls and $I(z; h)$ is the indicator function of the interval $[-h, h]$. Notice that further scaling would be required to ensure that $\tilde{f}$ integrates to 1.

Viewed as a density estimate, the histogram may be criticised in three ways.

- Information has been thrown away in replacing y_i by the central point of the interval in which it falls.
- In most circumstances, the underlying density function is assumed to be smooth, but the estimator is not smooth, due to the sharp edges of the boxes from which it is built.

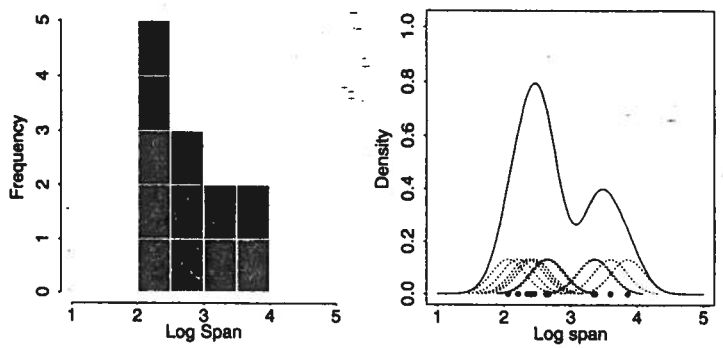


FIG. 1.2. Illustration of the construction of a histogram and density estimate from a subsample of the aircraft span data on a log scale.

- The behaviour of the estimator is dependent on the choice of width of the intervals (or equivalently boxes) used, and also to some extent on the starting position of the grid of intervals.

Rosenblatt (1956), Whittle (1958) and Parzen (1962) developed an approach to the problem which removes the first two of these difficulties. First, a smooth *kernel* function rather than a box is used as the basic building block. Second, these smooth functions are centred directly over each observation. This is illustrated in the right panel of Fig. 1.2. The kernel estimator is then of the form

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} w(y - y_i; h), \tag{1.1}$$

where w is itself a probability density, called in this context a kernel function, whose variance is controlled by the parameter h.

It is natural to adopt a function w which is symmetric with mean 0, but beyond that it is generally agreed that the exact shape is not too important. It is often convenient to use for w a normal density function, so that

$$w(y - y_i; h) = \phi(y - y_i; h),$$

where $\phi(z; h)$ denotes the normal density function in z with mean 0 and standard deviation h. Because of its role in determining the manner in which the probability associated with each observation is spread over the surrounding sample space, h is called the *smoothing parameter* or *bandwidth*. Since properties of w are inherited by $\hat{f}$, choosing w to be smooth will produce a density estimate which is also smooth.

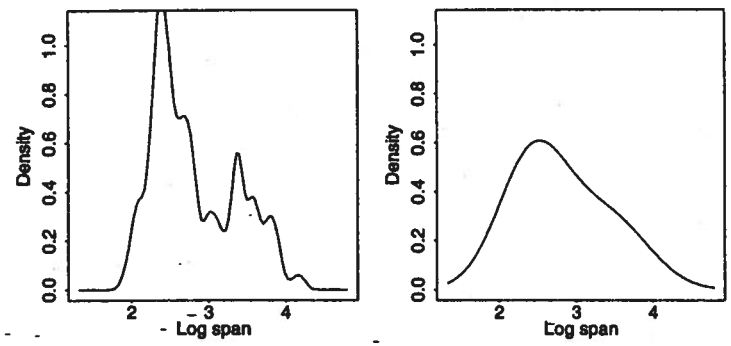


FIG. 1.3. The effect of changing the smoothing parameter on a density estimate of the aircraft span data.

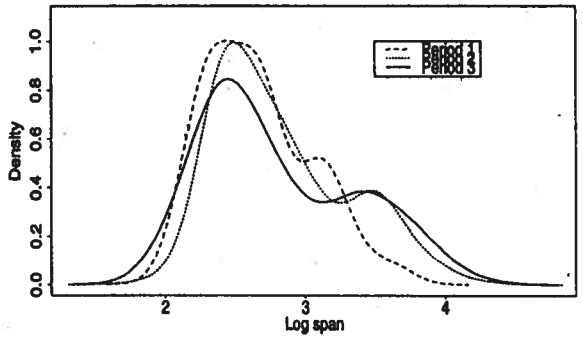


FIG. 1.4. A comparison of density estimates based on the log span data for the time periods 1914–1935, 1936–1955 and 1956–1984.

The third criticism of the histogram still applies to the smooth density estimate, namely that its behaviour is affected by the choice of the width of the kernel function. This is illustrated in Fig. 1.3. When h is small the estimate displays the variation associated with individual observations rather than the underlying structure of the whole sample. When h is large this structure is obscured by smoothing the data over too large a region. Strategies for choosing the smoothing parameter will be discussed in Chapter 2.

One advantage of a smooth density estimate is that comparisons among different groups become easier. For example, Fig. 1.4 displays estimates constructed from the three groups of data corresponding to the time periods mentioned above. It is difficult to superimpose histograms, but density estimates can be plotted together easily and the underlying shapes contrasted more effectively. In this case, the principal modes of the distributions are seen to occur at virtually identical positions, indicating that the most commonly used wing spans have changed very little throughout the century. The third time period displays a larger proportion of the distribution around a subsidiary mode at higher wing spans, as well as an increased proportion in the lower tail.

S-Plus Illustration 1.1. A density estimate of the log span data

Figure 1.1 was constructed with the following S-Plus *code. The* `provide.data` *and* `sm.density` *functions are supplied in the* `sm` *library which has been written in conjunction with this book. Details on obtaining and using this library are given in an Appendix.*

```
provide.data(aircraft)
y <- log(Span[Period==3])
par(mfrow=c(1,2))
hist(y, xlab="Log Span", ylab="Frequency")
sm.density(y, xlab="Log Span")
par(mfrow=c(1,1))
```

S-Plus Illustration 1.2. Changing the smoothing parameter

Figure 1.3 was constructed with the following S-Plus *code. The parameter* `hmult` *adjusts the default smoothing parameter by multiplying it by the stated value.*

```
provide.data(aircraft)
y <- log(Span[Period==3])
par(mfrow=c(1,2))
sm.density(y, hmult = 1/3, xlab="Log span")
sm.density(y, hmult = 2,   xlab="Log span")
par(mfrow=c(1,1))
```

An interactive exploration of the effect of changing the smoothing parameter can be obtained by adding the argument `panel=T` *to the* `sm.density` *function. This will launch a mouse-activated menu which also allows the possibility of an animated display.*

S-Plus Illustration 1.3. Comparing density estimates

Figure 1.4 was constructed with the following S-Plus *code.*

```
provide.data(aircraft)
y1 <- log(Span)[Period==1]
y2 <- log(Span)[Period==2]
```

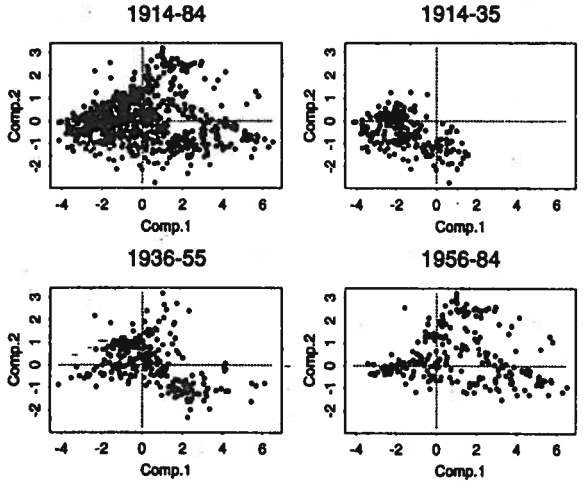


FIG. 1.5. Scatterplots of the first two principal component scores for the aircraft data. The first plot shows all the data. The remainder display the data for the time periods 1914–1935, 1936–1955 and 1956–1984.

```
y3 <- log(Span)[Period==3]
sm.density(y3, xlab="Log span")
sm.density(y2, add=T, lty=2)
sm.density(y1, add=T, lty=3)
legend(3.5, 1, c("Period 1", "Period 2", "Period 3"), lty=3:1)
```

1.3 Density estimation in two dimensions

The kernel method extends to the estimation of a density function in more than one dimension. The aircraft data are six-dimensional and Fig. 1.5 displays a plot of the first two principal components. The first component can broadly be identified with the 'size' of the aircraft as it is constructed from a mixture of all variables except speed. The second component can broadly be identified from its coefficients as 'speed adjusted for size'.

As a descriptive exercise, a two-dimensional density estimate can be constructed for these data by applying (1.1) with a two-dimensional kernel function in the form

$$\hat{f}(y_1, y_2) = \frac{1}{n}\sum_{i=1}^{n} w(y_1 - y_{1i}; h_1)\, w(y_1 - y_{2i}; h_2),$$

where $\{y_{1i}, y_{2i}; i = 1, \ldots n\}$ denote the data and (h_1, h_2) denote the joint smoothing parameters. It would be possible to use a bivariate kernel whose components are correlated but it is convenient, and usually sufficient, to employ a product of univariate components. Figure 1.6 shows the result of this with the data from the third time period, using normal kernel functions. The perspective plot shows that there are in fact three separate modes in the estimate, a feature which is not immediately clear from the simple scatterplot. The effect of the smoothing parameter is displayed in the remaining two panels, where a small value produces spurious peaks and troughs, while a large value obscures the individual modes. As in the one-dimensional case, the choice of this parameter can be important. In addition, it would clearly be very helpful to have some means of assessing which features are genuine and which are spurious when using these smoothing techniques.

Figure 1.7 illustrates alternative forms of display for a two-dimensional density estimate. In the left panel the height of the estimate is indicated by grey shading in an 'imageplot'. In the right panel contours are drawn. A standard approach would be to draw these contours at equally spaced heights. However, the contours here have been carefully selected in order to contain specific proportions of the observations. The contour labelled '75' contains the 75% of the observations corresponding to the greatest density heights, and similarly for the contours labelled '50' and '25'. These contours can be chosen easily by evaluating the density estimate at each observation, ranking the density heights and locating the median and quartiles. In this way the display has a construction which is reminiscent of a boxplot, although it has the additional ability to display multimodality through the disjoint nature of some of the contours. Bowman and Foster (1993) discuss this type of display in greater detail. The term 'sliceplot' will be used to differentiate the display from a standard contour plot. With the aircraft data the contours draw clear attention to the multimodal nature of the data. This can be interpreted as a degree of specialisation, with the appearance of aircraft which are fast but not large, and large but not fast.

The main aim of the analysis of the aircraft data was to identify changes over time. An effective way to address this is to plot and compare density estimates for the three separate time periods identified above. This is done in Fig. 1.8. The shape of each density estimate is characterised in a single contour containing 75% of the data and these are superimposed to show clearly the manner in which size and speed have changed in the aircraft designs produced over the century.

S-Plus Illustration 1.4. Density estimates from the aircraft span data

Figure 1.6 was constructed with the following S-Plus *code. The* `cex` *and* `zlim` *parameters are used simply to produce more attractive axis scaling.*

```
provide.data(airpc)
pc3 <- cbind(Comp.1, Comp.2)[Period==3,]
par(mfrow=c(2,2))
```

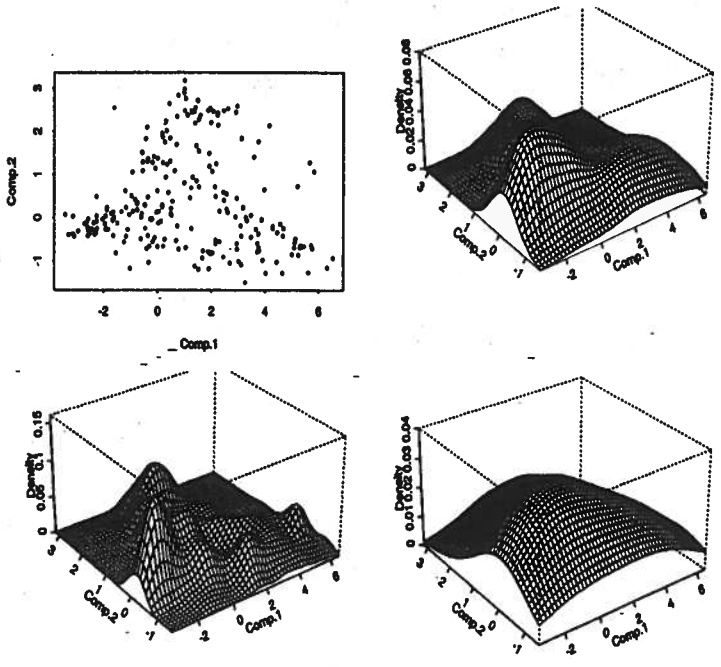


FIG. 1.6. A scatterplot and density estimates based on the log span data for the time period 1956–1984.

```
par(cex=0.6)
plot(pc3)
sm.density(pc3,             zlim=c(0,0.08))
sm.density(pc3, hmult=1/2, zlim=c(0,0.15))
sm.density(pc3, hmult=2,   zlim=c(0,0.04))
par(cex=1)
par(mfrow=c(1,1))
```

S-Plus Illustration 1.5. An imageplot and sliceplot from the aircraft span data

Figure 1.7 was constructed with the following S-Plus *code.*

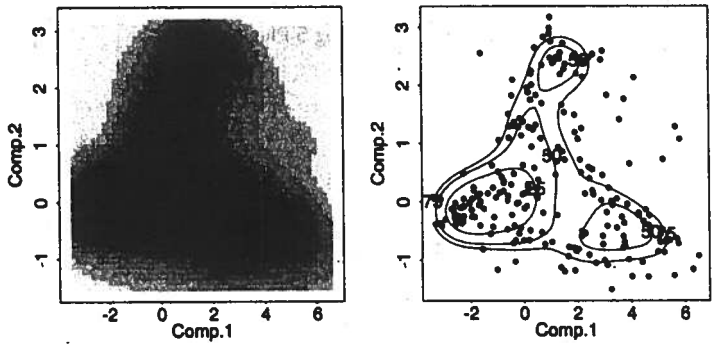


FIG. 1.7. An 'imageplot' and 'sliceplot' of density estimates based on the log span data for the time period 1956–1984.

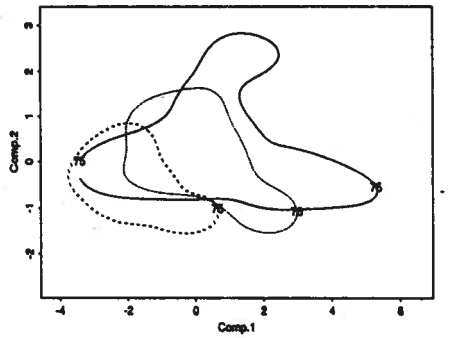


FIG. 1.8. Density estimates, represented as contour plots, based on the log span data for the time periods 1914–1935 (dashed line), 1936–1955 (dotted line) and 1956–1984 (full line).

```
provide.data(airpc)
pc3 <- cbind(Comp.1, Comp.2)[Period==3,]
par(mfrow=c(1,2))
sm.density(pc3, display="image")
sm.density(pc3, display="slice")
par(mfrow=c(1,1))
```

S-Plus Illustration 1.6. Multiple sliceplots from the aircraft span data

Figure 1.8 was constructed with the following S-Plus *code.*

```
provide.data(airpc)
pc  <- cbind(Comp.1, Comp.2)
pc1 <- pc[Period==1,]
pc2 <- pc[Period==2,]
pc3 <- pc[Period==3,]
plot(pc, type="n")
sm.density(pc1, display="slice", props=75, add=T, lty=3)
sm.density(pc2, display="slice", props=75, add=T, lty=2)
sm.density(pc3, display="slice", props=75, add=T, lty=1)
```

1.4 Density estimation in three dimensions

The Old Faithful geyser in Yellowstone National Park exhibits an unusual structure in its pattern of eruption times, and in the length of the waiting times between successive eruptions. Weisberg (1985) collected data on this and proposed a regression model to predict the waiting time. Azzalini and Bowman (1990) described similar data in time series form and identified the relationships among the three variables *waiting time*, *duration* and *subsequent waiting time* to be important in determining the structure of the series. Three clusters in the joint distribution of these variables are apparent by deduction from the marginal scatterplots.

The density estimate (1.1) can by applied with a three-dimensional kernel function in the form

$$\hat{f}(y_1, y_2, y_3) = \frac{1}{n}\sum_{i=1}^{n} w(y_1 - y_{1i}; h_1)\, w(y_2 - y_{2i}; h_2)\, w(y_3 - y_{3i}; h_3),$$

where $\{y_{1i}, y_{2i}, y_{3i}; i = 1, \ldots, n\}$ denote the data and (h_1, h_2, h_3) denote the joint smoothing parameters. It is again convenient to construct the kernel function from the product of univariate components. Displaying a density estimate as a function of three dimensions is more difficult. Contours were used to good effect with two-dimensional data, where a contour is a closed curve, or a set of closed curves if multimodality is present. A contour of a function defined in terms of three arguments is a more unusual object. In fact, it is a closed surface, or set of closed surfaces if multimodality is present. Scott (1992, Section 1.4) describes this approach and explores its use on a variety of datasets.

Figure 1.9 displays a contour plot of a three-dimensional density estimate of the geyser data, using normal kernel functions. It is represented as a 'wire frame' object, constructed in a manner similar to that described by Scott (1992, Appendix A). This contour has been chosen, as described in the previous section, to enclose exactly 75% of the observations. The space contained by the contour therefore corresponds to the upper reaches of the density estimate. This focuses

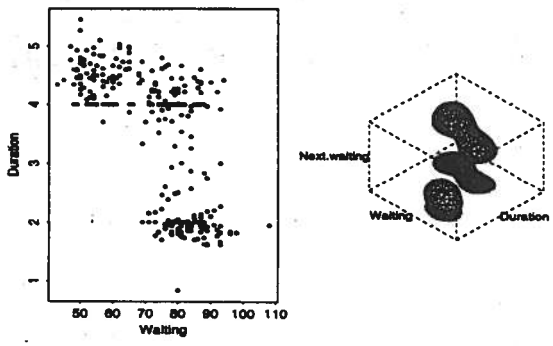


FIG. 1.9. A contour plot of the three-dimensional density estimate of the geyser data.

attention on the principal features of the density and ignores the outer regions, where behaviour is influenced by small numbers of observations.

The density contour of the geyser data is disjoint, confirming the highly clustered nature of the data. If a smaller percentage had been specified for the fraction of data contained, three disjoint surfaces would have been apparent, corresponding to the three clusters in the data. However, the contour displayed in the figure draws attention to the fact that the two upper clusters are not as well separated as the lower one is from the other two.

When surfaces are disjoint the relative locations of the separate parts in three dimensions may not always be clear. This has been resolved in Fig. 1.9 by displaying the shadow of the surfaces on the floor of the plot, as if a light were shining from the top face of the surrounding cube.

Scott (1992) describes a variety of more sophisticated techniques for constructing and displaying density contours, including the use of light sources to enhance visual perception, and the removal of strips of the surface in order to display interior contours.

Density estimation can be carried out in four, five and more dimensions. However, this should be done with care as there are increasing problems due to the 'curse of dimensionality', as described by Scott (1992, Chapter 7).

S-Plus Illustration 1.7. A three-dimensional density contour plot

Figure 1.9 was constructed with the following S-Plus *code. The geyser data are supplied in* S-Plus *as a standard dataset. However, the data used here have been organised in a manner which is more convenient for a three-dimensional analysis.*

The calculations for the density estimate may take some considerable time on some computers.

adjoining ranges $[-2\pi, 0]$ and $[2\pi, 4\pi]$ to create the circularity in the contribution of each observation, while retaining the factor $1/n$ for the original sample size, as described by Silverman (1986, Section 2.10).

S-Plus Illustration 1.8. A spherical data plot

Figure 1.10 was constructed with the following S-Plus *code.*

```
provide.data(magrem)
par(mfrow=c(1,2))
sm.sphere(maglat, maglong, theta =  60, phi =  10)
sm.sphere(maglat, maglong, theta = 240, phi = -10)
par(mfrow=c(1,1))
```

An original version of the function `sm.sphere` *for plotting spherical data was written with the help of Adrian Hines.*

S-Plus Illustration 1.9. A spherical density estimate

Figure 1.11 was constructed with the following S-Plus *code.*

```
provide.data(magrem)
par(mfrow=c(1,2))
sm.sphere(20, -30, theta=60, phi=10, sphim=T, kappa=13.6)
sm.sphere(maglat, maglong, theta=60, phi=10, sphim=T,
          kappa=13.6)
par(mfrow=c(1,1))
```

1.6 Data with bounded support

It commonly occurs that there are restrictions on the values which data can take. Directional data are a rather special case of this. A more common example is where only positive values can be recorded. All the variables in the aircraft data are of this type. In view of the strong skewness which exists in these data the underlying distributional patterns are best viewed on a log scale. However, it is also of some interest to present the density functions on the original scale.

Figure 1.12 shows a histogram and density estimate of the aircraft speed data, for all years. One problem with the density estimate is that the kernel functions centred on the observations which are very close to zero have transferred positive weight to the negative axis. This effect could be reduced by employing a smaller smoothing parameter, but this would have unwelcome effects elsewhere in the estimate, and in particular the size of the variations which already appear in the long right hand tail of the estimate would be increased.

At least two approaches to the problem are possible. In the *transformation method* the variable Y can be transformed to a new variable $t(Y)$ with unbounded support, the density of $t(Y)$ estimated and then transformed back to the original scale. In practice, the whole process can be accomplished in one operation. If

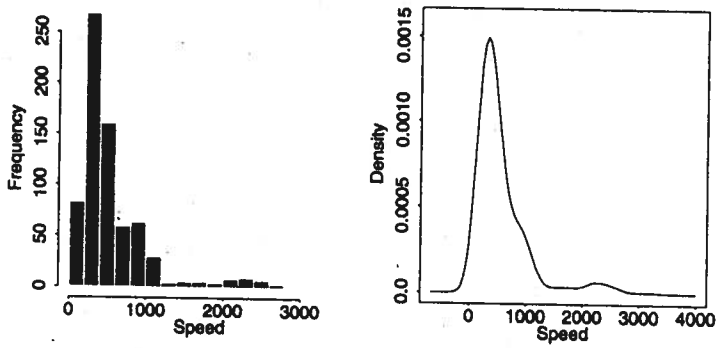


FIG. 1.12. A histogram and density estimate of the speed data on its original scale.

g denotes the probability density function of $t(Y)$, then the two densities are related by

$$f(y) = g\,(t(y))\;\; t'(y)$$

and $g(\cdot)$ can be estimated by the standard method, leading then to the estimate

$$\hat{f}(y) = \frac{1}{n}\sum_i w(t(y) - t(y_i); h)\, t'(y).$$

Wand and Jones (1995, p.43) describe this approach in some detail.

In the *modified kernel* method the kernel function w can be modified in a suitable manner. For estimation of f at a point y, a kernel function can be chosen so that its support does not exceed the support of Y. A related approach is to use standard kernel functions but to modify these near a boundary. A family of 'boundary kernels' was proposed by Gasser and Müller (1979) and is also described in Wand and Jones (1995, Section 2.11).

The effect of the transformation approach, using the log function, is illustrated in the left panel of Fig. 1.13 with the aircraft speed data. The density estimate now lies only on the positive section of the axis, and in addition the variations in the right hand tail have been eliminated. In this case the transformation method provides a satisfactory solution, and in general it does not often seem to be necessary to use other approaches, such as modified kernels. Clearly the transformation method also applies to other forms of restricted data. For example, when the support of Y is a bounded interval (a, b) then a natural transformation to consider is

$$t(x) = \log\frac{x-a}{b-x},$$

which maps (a, b) to the real line.

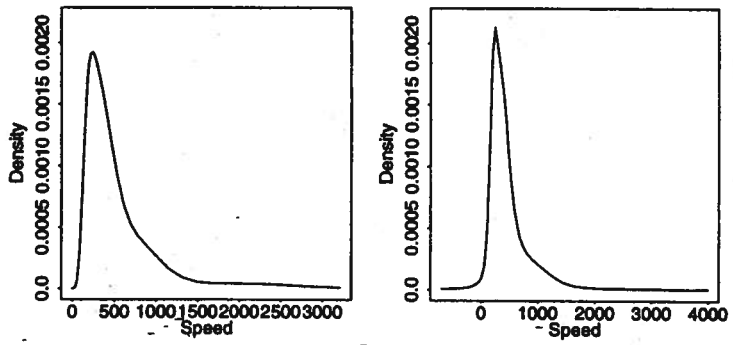


FIG. 1.13. Density estimates based on the aircraft speed data. The left panel is based on the log transformation method. The right panel is based on the variable kernel method, discussed in Section 1.7, with the nearest neighbour parameter $k = 30$.

Wand *et al.* (1991) discuss the interrelationships of the shape of the original distribution, the use of different transformation families and the selection of appropriate transformation parameters, on the performance of the final estimate.

Positive data in two dimensions can be handled in a similar way. This is explored in an exercise at the end of the chapter.

S-Plus Illustration 1.10. A density estimate from the speed data

Figure 1.12 was constructed with the following S-Plus *code. The* yht *parameter controls the height of the vertical axis in the plot of the density estimate.*

```
provide.data(aircraft)
par(mfrow=c(1,2))
hist(Speed, ylab="Frequency")
sm.density(Speed, yht=0.0016)
par(mfrow=c(1,1))
```

S-Plus Illustration 1.11. Density estimates for bounded support

Figure 1.13 was constructed with the following S-Plus *code. The function* nnbr *is provided in the* sm *library. It returns a vector consisting of the nearest neighbours from each observation to the remainder of the data. This feature of density estimation is discussed in Section 1.7.*

```
provide.data(aircraft)
hw <- nnbr(Speed, 30)
hw <- hw/exp(mean(log(hw)))
```



```
par(mfrow=c(1,2))
sm.density(Speed, yht=0.0022, positive=T)
sm.density(Speed, yht=0.0022, xlim=c(-700,4000), h.weights=hw)
par(mfrow=c(1,1))
```

1.7 Alternative forms of density estimation

Throughout this book the kernel approach is used because it is conceptually simple, deriving naturally from the histogram in the case of density estimation, and computationally straightforward. The techniques also extend naturally to the multivariate case and there are close links with techniques based on kernel functions for the smoothing of regression data. However, there are other approaches to density estimation and some of these are sketched below. The first is a variant of the kernel approach which is particularly useful for certain kinds of data. The others take rather different routes in constructing estimators and, after a brief description, will not be pursued further.

1.7.1 *Variable bandwidths*

One of the features of the speed data, illustrated in Fig. 1.12, is the presence of 'bumps' in the right hand tail of the distribution, caused by small numbers of observations. Here the data are sparse and it may be more appropriate to use a large smoothing parameter to remove these bumps in the estimate. On the other hand, where the data are clustered closely together at small values of speed it would be appropriate to use a small smoothing parameter. This idea has led to a variety of suggestions for *variable* bandwidths, where a different smoothing parameter can be used in each kernel function, of the form

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} w(y - y_i; h_i).$$

One possibility is to reflect the sparsity of the data in expressions such as $h_i = hd_k(y_i)$, where h denotes an overall smoothing parameter and $d_k(y_i)$ denotes the distance from y_i to its kth nearest neighbour among the data. Breiman *et al.* (1977) describe this approach. Silverman (1986) makes the helpful suggestion of introducing the variable element as a modification of the overall smoothing parameter, to give $h_i = hd_k(y_i)/\bar{d}$, where $\bar{d}$ denotes the geometric mean of the $d_k(y_i)$.

The right panel of Fig. 1.13 shows an estimator of this type, using $k = 30$ nearest neighbours. The behaviour in the right hand tail has been improved, as the kernel functions there are much flatter. However, this estimator has not entirely overcome the problem of the transfer of positive weight to the negative axis discussed in the previous section.

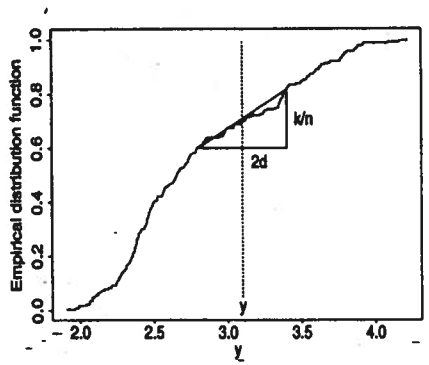


FIG. 1.14. A density estimate as the gradient of the empirical distribution function, using the aircraft span data, on a log scale, for the third time period.

1.7.2 *Nearest neighbour methods*

Figure 1.14 displays the empirical distribution function of the aircraft span data, on a log scale, for the third time period. Since a density can be obtained by finding the derivative of the distribution function, a density estimate can be constructed by measuring the gradient of the curve, as indicated graphically in this figure.

This can be done in two ways. The first is to fix a distance d on either side of the point of estimation y. This is equivalent to the use of the density estimator (1.1) with a 'box' kernel defined as $I(z;d)/(2d)$, where $I(z;d)$ is the indicator function of the interval $(-d,d)$. The density estimate is then given by $(k/n)/(2d)$, where k represents the number of observations lying within a distance d of y, and whose kernel function therefore contributes to the density estimate at that point.

An alternative approach is to fix the value of k, the number of observations which contribute to the density estimate. This then determines d, which is referred to as the kth nearest neighbour distance. The nearest neighbour form of density estimate is therefore available as $\hat{f}(y) = (k/n)/(2d_k(y))$, using the full notation for the nearest neighbour distance $d_k(y)$.

Moore and Yackel (1977) generalised work by Loftsgaarden and Quesenberry (1965) to construct a more general *nearest neighbour* density estimator of the form

$$\hat{f}(y) = \frac{1}{n}\sum w(y - y_i; hd_k(y)).$$

The distinctive feature of this is that the bandwidth changes with the point of estimation y rather than with the observation y_i. An unfortunate consequence is that the estimator does not necessarily integrate to 1. In addition, there can

be abrupt changes in nearest neighbour distances as a function of y and so the resulting density estimate can display a lack of smoothness.

Both nearest neighbour and variable bandwidth methods extend to the multivariate case since a nearest neighbour distance is easily computed in several dimensions.

1.7.3 *Orthogonal series methods*

To introduce this approach, it is useful to recall the concept of a Fourier series. A function f defined on $(-\pi,\pi)$ can be represented by the series expansion

$$f(x) = \tfrac{1}{2}a_0 + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \sin kx), \tag{1.2}$$

where the Fourier coefficients a_k, b_k are defined by

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx, \quad b_k = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx.$$

This construction and the discussion below require the technical condition of square integrability of the functions involved, but we shall not examine these mathematical aspects in further detail.

Equation (1.2) expresses $f(x)$ as a linear combination of the set of trigonometric functions

$$\{1,\ \cos x,\ \sin x,\ \cos 2x,\ \sin 2x,\ \ldots\},$$

which are then said to form a *basis* for the set of possible functions $\{f\}$. By changing the coefficients of the linear combination, different functions f are generated.

This scheme can be implemented with the trigonometric basis replaced by some other set of functions $\{\psi_k(x); k = 0,1,\ldots\}$. An extension of this kind is required to handle functions defined over an unrestricted set X, such as the real line, since the Fourier series allows only periodic functions. The generalisation of the Fourier series is then

$$f(x) = \sum_{k=0}^{\infty} c_k \psi_k(x) \tag{1.3}$$

where the coefficients are defined by

$$c_k = \int_X \psi_k(x)\, f(x)\, w(x)dx$$

and $w(x)$ is a weight function on X. However, not all sets of functions $\{\psi_k(x)\}$ are suitable for this purpose. They need to form an orthogonal basis, which means that:

1. $\int_X \psi_k(x)\psi_j(x)w(x)dx = 0$ if $k \neq j$;

and (0.4, 0.7). All the components had variance 0.03 in each direction, and covariance 0. Use a sample size of your own choice, with equal numbers in each group, to simulate data from each of these distributions. Estimate a density function for each group, using identical evaluation grids, by setting the parameters `xlim` and `ylim` in `sm.density`. The output of the `sm.density` function has a component `estimate` which can be used to extract the estimate numerically. Calculate the difference of the two estimates and superimpose the 0 contour of this difference on a plot of the data, to provide a graphical discrimination line between the two groups.

1.9 *Pole position.* Fisher *et al.* (1987) provide data on the historical positions of the South Pole, from a palaeomagnetic study of New Caledonia laterites. The measurements are recorded as latitude and longitude positions and are available in the `poles` file. Use `sm.sphere` to plot the data, using a suitable orientation. Construct a density estimate and identify its mode graphically, as an estimate of the location of the South Pole.

1.10 *Variable bandwidths.* Examine the effect of changing the nearest neighbour parameter k in a variable bandwidth estimate of the aircraft speed data, while holding the overall parameter h fixed.

1.11 *Nearest neighbour density estimate.* Write a function which will construct the nearest neighbour form of density estimate described in Section 1.7.

1.12 *Data restricted to* $[0, 1]$. Generate a sample of data from a beta distribution. Apply the logistic transformation $\log(p/(1-p))$, use `sm.density` to estimate the density function on that scale and transform the result back to the original scale. Compare your estimate with a histogram of the data.

1.13 *Censored data.* A density estimate for the case where the values of some observations in a sample are censored was described in Section 1.8. This requires weights determined by the step sizes of the Kaplan–Meier survivor function. These can be derived from the S-Plus function `survfit`. Construct a density estimate from censored data of your own choice by passing these weights to `sm.density` through the `weights` argument. Consider what happens when the largest observation in a sample is censored.

1.14 *Positive data in two dimensions.* Use a suitable distribution, such as a product of two gamma densities, to generate two-dimensional data with positive components. Use `sm.density` with the argument `positive=T` to explore the effectiveness of the transformation method, described in Section 1.6, in two dimensions.

2

DENSITY ESTIMATION FOR INFERENCE

2.1 Introduction

In Chapter 1, the role of kernel functions in constructing nonparametric density estimates was illustrated on a variety of different types of data. This led to a set of useful tools, particularly for graphical illustration of distributional shape. The uses of density estimates in reaching conclusions in a more quantitative way will now be investigated. Attention will be focused on one- and two-dimensional data on an unrestricted continuous scale.

In order to go beyond the exploratory and graphical stage it is necessary first to understand more about the behaviour of these estimators and to derive some basic properties. Although many theoretical results exist, simple expressions for means and variances of the estimators are enough to allow ideas of interval estimation and hypothesis testing to be discussed, and to motivate techniques for choosing an appropriate bandwidth to employ with a particular dataset.

2.2 Basic properties of density estimates

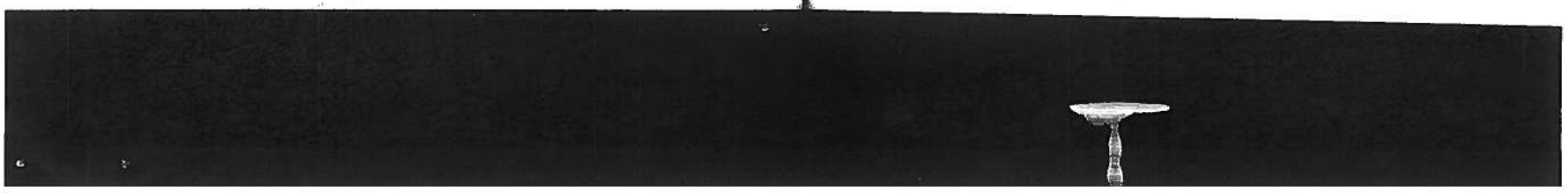
A simple manipulation shows that the mean of the density estimator (1.1) can be written as

$$\mathrm{E}\{\hat{f}(y)\} = \int w(y-z;h) f(z) dz. \tag{2.1}$$

This is a convolution of the true density function f with the kernel function w. Smoothing has therefore produced a biased estimator, whose mean is a smoothed version of the true density. Further insight can be gained through a Taylor series argument. It is convenient to use a kernel function which, with a slight change of notation, can be parametrised in the form $(1/h)w(z/h)$. A Taylor series expansion then produces the approximation

$$\mathrm{E}\{\hat{f}(y)\} \approx f(y) + \frac{h^2}{2}\sigma_w^2 f''(y), \tag{2.2}$$

where σ_w^2 denotes the variance of the kernel function, namely $\int z^2 w(z) dz$. Since $f''(y)$ measures the rate of curvature of the density function, this expresses the fact that $\hat{f}$ underestimates f at peaks in the true density and overestimates at troughs. The size of the bias is affected by the smoothing parameter h. The component σ_w^2 will reduce to 1 if the kernel function w is chosen to have unit variance.



Through another Taylor series argument, the variance of the density estimate can be approximated by

$$\text{var}\Big\{\hat{f}(y)\Big\} \approx \frac{1}{nh} f(y)\,\alpha(w), \tag{2.3}$$

where $\alpha(w) = \int w^2(z)dz$. As ever, the variance is inversely proportional to sample size. In fact, the term nh can be viewed as governing the local sample size, since h controls the number of observations whose kernel weight contributes to the estimate at y. It is also useful to note that the variance is approximately proportional to the height of the true density function.

These approximate expressions for the mean and variance of a density estimate encapsulate the effects of the smoothing parameter which were observed in Fig. 1.3. As h decreases, bias diminishes while variance increases. As h increases the opposite occurs. The combined effect of these properties is that, in order to produce an estimator which converges to the true density function f, it is necessary that both h and $1/nh$ decrease as the sample size increases. A suitable version of the central limit theorem can also be used to show that the distribution of the estimator is asymptotically normal.

A similar analysis enables approximate expressions to be derived for the mean and variance of a density estimate in the multivariate case. In p dimensions, with a kernel function defined as the product of univariate components w, and with smoothing parameters $(h_1, \ldots, h_p)$, these expressions are

$$\mathbf{E}\Big\{\hat{f}(y)\Big\} \approx f(y) + \frac{1}{2}\sigma_w^2 \left[\sum_{j=1}^{p} h_j^2 \frac{\partial^2}{\partial y_j^2} f(y)\right],$$

$$\text{var}\Big\{\hat{f}(y)\Big\} \approx \frac{1}{nh_1 \cdots h_p} f(y)\,\alpha(w)^p.$$

Wand and Jones (1995, Section 4.3) derive results for more general kernel functions.

It is helpful to define an overall measure of how effective $\hat{f}$ is in estimating f. A simple choice for this is the *mean integrated squared error* (MISE) which, in the one-dimensional case, is

$$\begin{aligned}\text{MISE}(\hat{f}) &= \mathbf{E}\left\{\int [\hat{f}(y) - f(y)]^2 dy\right\} \\ &= \int \left[\mathbf{E}\Big\{\hat{f}(y)\Big\} - f(y)\right]^2 dy + \int \text{var}\Big\{\hat{f}(y)\Big\}\, dy.\end{aligned}$$

This combination of bias and variance, integrated over the sample space, has been the convenient focus of most of the theoretical work carried out on these estimates. In particular, the Taylor series approximations described in (2.2) and (2.3) allow the mean integrated squared error to be approximated as

$$\text{MISE}(\hat{f}) \approx \frac{1}{4} h^4 \sigma_w^4 \int f''(y)^2 dy + \frac{1}{nh}\alpha(w).$$

Establishing the properties of the estimators which employ variable bandwidths, as described in Chapter 1, is more complex. Here the smoothing parameter h in the kernel function over observation y_i is replaced by expressions such as $hd_k(y_i)$, where $d_k(y_i)$ denotes the distance from y_i to its kth nearest neighbour among the data. It was shown in Section 1.7 that a nearest neighbour distance is inversely proportional to a simple form of density estimate. In view of this, it is instructive to represent smoothing parameters involving variable bandwidths as $h/\tilde{f}(y_i)$, where $\tilde{f}$ denotes a density estimate. This shows that this approach is based on a *pilot* estimate of the underlying density which is then used to adjust the kernel widths locally. This representation also suggests a means of investigating the behaviour of estimators of this type by analysing estimators which use the true density f as a pilot estimator. Abramson (1982) adopted this approach and used Taylor series expansions to derive first-order asymptotic properties. An additional exploration of more general variable bandwidths of the form $h/\tilde{f}(y_i)^\alpha$ led to the proposal that the power $\alpha = 1/2$ is most appropriate, since the asymptotic arguments then suggest that the principal bias term is eliminated. This results also holds in the multivariate case.

Bowman and Foster (1993) avoided asymptotic calculations by employing numerical integration in the calculations of mean and variance. This showed that although considerable caution should be exercised with Taylor series approximations in this context, the broad conclusions of these analyses were corroborated.

As an illustration of variable bandwidths, Fig. 2.1 displays density estimates constructed from data on a tephra layer, resulting from a volcanic eruption in Iceland around 3500 years ago. Dugmore *et al.* (1992) report the geological background to the collection of these data, which refer to the chemical composition of shards of volcanic glass found in the tephra layer. These compositions can help to identify whether tephra layers have resulted from the same eruption. Figure 2.1 displays data on the percentage of aluminium oxide (Al_2O_3) in the sample from a single site. A natural model for compositional data is to apply a logistic transformation as described by Aitchison (1986).

Three density estimates have been produced in Fig. 2.1. In each case the same smoothing parameter was used, but in one of the estimates nearest neighbour weights were used, and in another these weights were calculated on a square root scale. In both the estimates involving variable kernels the weights were scaled, as described in Section 1.7, in order to allow comparisons to be made. The main effect of the variable weights is to allow the density estimate to peak more sharply. This peak is more pronounced for the nearest neighbour weights, corresponding to a pilot density estimate $\tilde{f}$. The square root scale offers a more modest modification which has some backing from the theory of Abramson (1982).

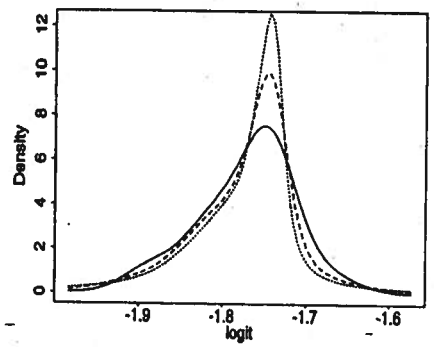


FIG. 2.1. Density estimates from the tephra data, on a logit scale, using simple kernels (full line), nearest neighbour weights (dotted line) and square root nearest neighbour weights (dashed line).

S-Plus Illustration 2.1. Square root weights for variable kernels

Figure 2.1 was constructed with the following S-Plus *code.*

```
provide.data(tephra)
logit <- log(Al203/(100-Al203))
nn    <- nnbr(logit, 7)
hw    <- nn/exp(mean(log(nn)))
sm.density(logit, h.weights = hw, lty = 2, yht = 12.2)
hw    <- sqrt(nn)
hw    <- hw/exp(mean(log(hw)))
sm.density(logit, h.weights = hw, lty = 3, add = T)
sm.density(logit, add = T)
```

Mathematical aspects: The approximate mean and variance of a one-dimensional density estimate

An asymptotic argument for the mean of a density estimate begins with (2.1). A change of variable produces the expression

$$\mathrm{E}\left\{\hat{f}(y)\right\} = \int w(z)\, f(y - hz)\, dz.$$

If $f(y - hz)$ is now replaced by the Taylor series expansion $f(y) - hzf'(y) + \frac{1}{2}h^2z^2f''(y)$, then the principal terms allow the mean to be approximated as

$$\mathrm{E}\left\{\hat{f}(y)\right\} \approx f(y) + \frac{h^2}{2}\int z^2 w(z)dz\, f''(y),$$

where the remainder term in this approximation is $o(h^2)$. The term involving h reduces to 0 because $\int zw(z)dz = 0$.

The variance can be written as

$$\begin{aligned}\mathrm{var}\left\{\hat{f}(y)\right\} &= \frac{1}{n}\mathrm{var}_z\{w(y-z\,;h)\} \\ &= \frac{1}{n}\{\mathrm{E}_z\{w(y-z\,;h)^2\} - \mathrm{E}_z\{w(y-z\,;h)\}^2\}.\end{aligned}$$

The expectations in this last expression can be expanded by a change of variable and Taylor series, as above, to yield the approximate expression (2.3).

2.3 Confidence and variability bands

One of the issues raised in Chapter 1 by the use of density estimates in exploring data was the need to assess which features of a density estimate indicate genuine underlying structure and which can be attributed to random variation. It is best to construct tools for this purpose around specific questions. For example, methods for assessing normality, and the independence of variables in a bivariate distribution, will be derived later in this chapter. However, it would also be of some value to have a general purpose method of indicating the uncertainty associated with a density estimate.

A confidence band, displaying confidence intervals for the true density at a range of values of y, would be very helpful. The immediate difficulty is the forms of the mean and variance of a density estimate, outlined in Section 2.2. The variance involves the true density $f(y)$, and the mean shows that bias is present, involving the second derivative $f''(y)$. In theory it is possible to use estimates of these unknown factors. However, this introduces an unsatisfactory degree of complexity, and further uncertainty, into the problem.

The bias component is particularly problematic, through the involvement of $f''(y)$. This leads to the idea of addressing the lesser, but still useful, aim of quantifying the variability of a density estimate without attempting to account for bias. Since the expression for variance has a relatively simple form, a variance stabilising argument can be adopted. For any transformation $t(\cdot)$, a Taylor series argument shows that

$$\mathrm{var}\left\{t(\hat{f}(y))\right\} \approx \mathrm{var}\left\{\hat{f}(y)\right\}\left[t'\left(\mathrm{E}\left\{\hat{f}(y)\right\}\right)\right]^2.$$

When $t(\cdot)$ is the square root transformation, the principal term of this expression becomes

$$\mathrm{var}\left\{\sqrt{\hat{f}(y)}\right\} \approx \frac{1}{4}\frac{1}{nh}\alpha(w),$$

which does not depend on the unknown density f. This is the analogue for density estimation of the argument employed by Tukey (1977) in the derivation of the 'hanging rootogram' for histograms, also discussed by Scott (1992, Section 3.2.1).

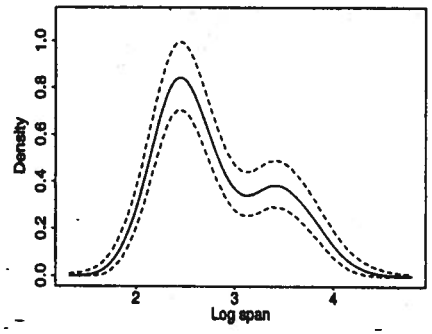


FIG. 2.2. A density estimate and variability bands from the aircraft span data.

On this square root scale the variability of the density estimate can now be expressed by placing a band of width two standard errors (se), $2\sqrt{\alpha(w)/4nh} = \sqrt{\alpha(w)/nh}$ around $\sqrt{\hat{f}}$. The edges of this band can then be translated back to the original scale. This has been carried out for the aircraft span data, for the third time period, in Fig. 2.2. For these particular data some caution needs to be exercised since the plausibility of the assumption of independent random sampling is in doubt. However, a band of this type requires careful interpretation under any circumstances. It is not a confidence band for the true density function because of the presence of bias, for which no adjustment has been made. In order to reflect the fact that only variance is being assessed, the term *variability band* will be used. This follows the terminology of Simonoff (1996) who described a similar display as a 'variability plot'.

Bands of this type are unlikely to provide a powerful means of examining particular hypotheses about a density function, particularly in view of their pointwise nature. However, the graphical display of variance structure can be helpful. For example, the width of the bands for the aircraft span data strengthens the evidence that the shoulder in the right hand tail of the estimate is a genuine feature, particularly since the effect of bias is to diminish modes and raise intervening troughs. The strength of any feature of this sort is therefore likely to be underestimated.

S-Plus Illustration 2.2. Variability bands for a density estimate

Figure 2.2 was constructed with the following S-Plus *code.*

```
provide.data(aircraft)
y <- log(Span)[Period==3]
sm.density(y, xlab = "Log span", display = "se")
```


2.4 Methods of choosing a smoothing parameter

In order to construct a density estimate from observed data it is necessary to choose a value for the smoothing parameter h. In this section the asymptotically optimal choice for h, and three of the most common, and arguably most effective, practical strategies, are described.

2.4.1 *Optimal smoothing*

An overall measure of the effectiveness of $\hat{f}$ in estimating f is provided by the mean integrated squared error described in Section 2.2. From the approximate expression given there it is straightforward to show that the value of h which minimizes MISE in an asymptotic sense is

$$h_{\text{opt}} = \left\{ \frac{\gamma(w)}{\beta(f)n} \right\}^{1/5}, \tag{2.4}$$

where $\gamma(w) = \alpha(w)/\sigma_w^4$, and $\beta(f) = \int f''(y)^2 dy$. This optimal value for h cannot immediately be used in practice since it involves the unknown density function f. However, it is very informative in showing how smoothing parameters should decrease with sample size, namely proportionately to $n^{-1/5}$, and in quantifying the effect of the curvature of f through the factor $\beta(f)$.

2.4.2 *Normal optimal smoothing*

Evaluation of the optimal formula for h when f is a normal density yields the simple formula

$$h = \left(\frac{4}{3n}\right)^{1/5} \sigma,$$

where σ denotes the standard deviation of the distribution. Clearly, the assumption of normality is potentially a self-defeating one when attempting to estimate a density nonparametrically but, for unimodal distributions at least, it gives a useful choice of smoothing parameter which requires very little calculation.

This approach to smoothing has the potential merit of being cautious and conservative. The normal is one of the smoothest possible distributions and so the optimal value of h will be large. If this is then applied to non-normal data it will tend to induce oversmoothing. The consequent reduction in variance at least has the merit of discouraging overinterpretation of features which may in fact be due to sampling variation. All of the one- and two-dimensional density estimates based on unrestricted continuous data in Chapter 1 were produced with normal optimal smoothing parameters.

In order to accommodate long tailed distributions and possible outliers, a robust estimate of σ is sometimes preferable to the usual sample standard deviation. A simple choice is the median absolute deviation estimator

$$\tilde{\sigma} = \text{median}\{|y_i - \tilde{\mu}|\}/0.6745,$$

where $\tilde{\mu}$ denotes the median of the sample; see Hogg (1979).

Normal optimal smoothing parameters can also be found in the multidimensional case. These are given by

$$h_i = \left\{ \frac{4}{(p+2)n} \right\}^{1/(p+4)} \sigma_i,$$

where p denotes the number of dimensions, h_i denotes the optimal smoothing parameter and σ_i the standard deviation in dimension i. For practical implementation the latter is replaced by a sample estimate. In the case of the aircraft data this simple form of smoothing is seen to be very effective in the two-dimensional displays of Section 1.3, even in the presence of a multimodal structure. The three-dimensional density contour produced from the geyser data, displayed in Fig. 1.9, was also constructed from normal optimal smoothing parameters.

This principle was carried to its logical extension by Terrell and Scott (1985) and Terrell (1990), who defined the 'oversmoothed' bandwidth to be the largest possible value of the optimal smoothing parameter, over all distributions with the same variance. This turns out to be fractionally larger than the normal optimal value.

2.4.3 *Cross-validation*

The ideas involved in cross-validation are given a general description by Stone (1974). In the context of density estimation, Rudemo (1982) and Bowman (1984) applied these ideas to the problem of bandwidth choice, through estimation of the integrated squared error (ISE)

$$\int \{\hat{f}(y) - f(y)\}^2 dy = \int \hat{f}(y)^2 dy - 2\int f(y)\hat{f}(y)dy + \int f(y)^2 dy.$$

The last term on the right hand side does not involve h. The other terms can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \int \hat{f}^2_{-i}(y)dy - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(y_i), \qquad (2.5)$$

where $\hat{f}_{-i}(y)$ denotes the estimator constructed from the data without the observation y_i. It is straightforward to show that the expectation of this expression is the MISE of $\hat{f}$ based on $n-1$ observations, omitting the $\int f^2$ term. The value of h which minimises this expression therefore provides an estimate of the optimal smoothing parameter. Stone (1984) derived an asymptotic optimality result for bandwidths which are chosen in this cross-validatory fashion. From a computational point of view, the use of normal kernel functions allows the integrals in (2.5) to be evaluated easily, by expanding $\hat{f}^2_{-i}(x)$ into its constituent terms and applying results on the convolution of normal densities.

The left panel of Fig. 2.3 displays the cross-validation function (2.5) for the tephra data. This shows a minimum at $h = 0.018$. The right panel of the figure shows the density estimate produced by this value of smoothing parameter.

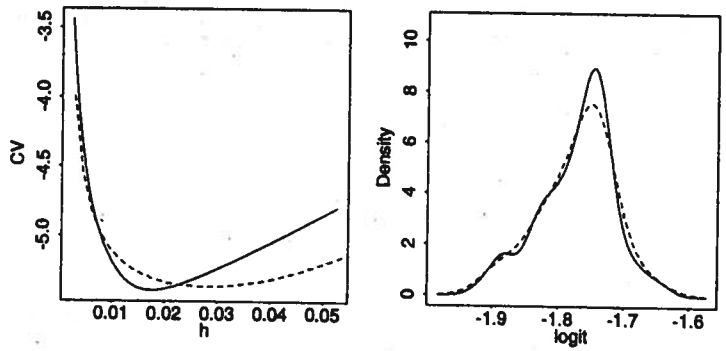


FIG. 2.3. The left panel shows the cross-validation function based on the tephra data, on a logit scale. The MISE function for a normal density has been superimposed, on a shifted vertical scale, for comparison. The right panel shows density estimates from the tephra data using the cross-validatory smoothing parameter (full line), and the normal optimal smoothing parameter (dashed line).

Cross-validation has therefore successfully chosen a suitable amount of smoothing to apply to the data. In the left panel, the MISE function for a normal distribution is also displayed. Since the vertical position is unimportant the function has been shifted to allow it to be superimposed on the graph. The normal optimal smoothing parameter in this case has the value 0.026. Cross-validation has attempted to take account of the mild skewness in the data. The smaller smoothing parameter helps to accentuate the peak of the density, but has the additional effect of producing a less smooth left hand tail.

Of the estimates produced so far from the tephra data, the variable bandwidth approach displayed in Fig. 2.1 seems the most effective choice for these data. It is an advantage of the cross-validatory approach that its general definition allows it to be applied in a wide range of settings. It can, for example, be applied to the choice of the overall smoothing parameter h in the variable bandwidth form $h_i = hd_k(y_i)$. The general computational form of the cross-validation function, using normal kernels, is given at the end of this section. Minimisation of this for a kernel estimator with normalised nearest neighbour weights ($k = 7$) on a square root scale produces a smoothing parameter of 0.022 for the tephra data. This is very close to the normal optimal value and so the resulting estimate is very similar to that displayed in Fig. 2.1.

Local minima can occur in cross-validation functions, as described by Wand and Jones (1995, Section 3.3), and so it can be wise to employ plotting, in addition to a numerical algorithm, to locate the minimising smoothing parameter.

Marron (1993) recommends using the smoothing parameter corresponding to the local minimum at the largest value of h.

Techniques known as ***biased cross-validation*** (Scott and Terrell 1987) and ***smoothed cross-validation*** (Hall *et al.* 1992) also aim to minimise ISE but use different estimates of this quantity to do so. These approaches are also strongly related to the 'plug-in' approach described in the following subsection.

Cross-validation can also be employed with multivariate density estimates since (2.5) remains a valid definition in several dimensions.

2.4.4 *Plug-in bandwidths*

Since the earliest days of density estimation, iterative procedures have been proposed in which an estimate $\hat{f}$ is used in the formula for the optimal smoothing parameter:

$$h = \left\{ \frac{\gamma(w)}{\beta(\hat{f})n} \right\}^{1/5}.$$

Scott *et al.* (1977) is an early example of this approach. If normal kernels are used, $\gamma(w)$ and $\beta(\hat{f})$ can be calculated relatively easily and the value of h which solves this equation can be found by a suitable numerical algorithm.

Within the past few years, considerable progress has been made in the development of this approach. In particular, Sheather and Jones (1991), extending the work of Park and Marron (1990), described a bandwidth selection procedure with excellent properties. This is based on a clever method of estimation of f'', using an additional smoothing parameter related to h for this purpose, guided by asymptotic theory on the estimation of derivatives of a density, which requires a larger amount of smoothing than in the estimation of f itself.

This estimator has very good finite sample, as well as asymptotic, properties. In particular, it is more stable than the cross-validation approach described above. The two techniques take different approaches to the same problem of minimising ISE. Cross-validation estimates the ISE function and locates the minimum. The plug-in approach minimises the function theoretically and then estimates this minimising value directly. The good performance of this estimation process produces a method which is subject to less variability.

With the tephra data, the Sheather–Jones method produces a smoothing parameter of 0.016. This is very close to the cross-validatory value of 0.018 and so in this particular example there is little difference between the two approaches.

2.4.5 *Discussion*

How best to choose smoothing parameters is still the subject of much research and debate. For one-dimensional data the 'plug-in' approach of Sheather and Jones (1991) seems to be very effective. In contrast, there is both empirical and theoretical evidence that cross-validation is subject to greater variability. However, it does have the advantage of providing a general approach which can be adapted to a wide variety of settings, including multivariate data and variable

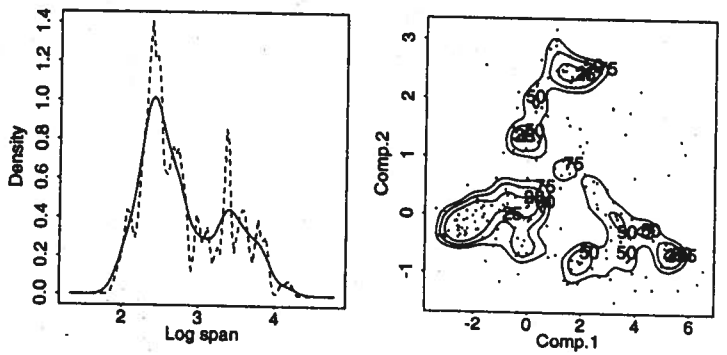


FIG. 2.4. The left panel displays density estimates produced from the log span data, using the Sheather–Jones plug-in (full line) and cross-validatory (dashed line) smoothing parameters. The right panel displays a density contour plot of the first two principal components of the aircraft data, using cross-validation to select the parameters in two dimensions.

bandwidths. Other examples of its versatility are provided by Fisher *et al.* (1987), who describe a method for spherical data, and Bowman and Prvan (1996), who show how the method can be applied to the smoothing of distribution functions.

As a further illustration, the left panel of Fig. 2.4 displays density estimates produced from the log span data, considered extensively in Chapter 1. The full line corresponds to the use of the Sheather–Jones smoothing parameter. This seems an effective choice, which captures the structure of the data, with a shoulder in the right hand tail. The dashed line corresponds to use of the cross-validatory smoothing parameter. This is very small and results in a highly variable estimate. Silverman (1986, Section 3.4.3) gives an explanation of the fact that cross-validation is rather sensitive to the presence of clustering in the data. There are several repeated values in the aircraft span data and this results in a very small smoothing parameter, as the method attempts to reproduce this granularity in the density estimate.

In two dimensions, the smoothing parameters (h_1, h_2) can be chosen to minimise the cross-validatory criterion (2.5) when $\hat{f}$ is a bivariate estimate. To reduce the complexity of this problem it is convenient to use a simple scaling for each dimension, given by $h_1 = a\hat{\sigma}_1$, $h_2 = a\hat{\sigma}_2$, and to minimise this over a. This has been done in the right hand panel of Fig. 2.4, where the contours of a density estimate, using the aircraft data from the third time period, and using cross-validatory smoothing parameters, are displayed. This plot should be contrasted with Fig. 1.7, which displays a density estimate for the same data using a normal optimal smoothing parameter. Again, cross-validation has reacted strongly

to the existence of small clusters in the data. It seems clear from the scatterplot that these clusters are present, but the normal optimal smoothing parameter arguably provides a clearer overall summary of the structure in these particular data.

The attention devoted to an appropriate choice of smoothing parameter when constructing a density estimate is motivated by the large influence of this parameter on the detailed shape of the estimate produced. In some problems this is indeed the main focus of attention. There is, however, quite a range of other problems where the substantive questions relate to the presence or absence of particular types of underlying structure. This perspective can change the importance of the role of the smoothing parameter, as the focus shifts from estimation of the density function to the comparison of different models for the data. Some problems of this type will be explored in the remainder of this chapter and, indeed, are a major focus throughout the book.

Jones *et al.* (1996) give a helpful and balanced discussion of methods of choosing the smoothing parameter in density estimation.

S-Plus Illustration 2.3. Data based choices of smoothing for the tephra data

Figure 2.3 was constructed with the following S-Plus *code. The function* `nmise` *calculates the MISE for a density estimate of a normal distribution. It is supplied with the* `sm` *library.*

```
provide.data(tephra)
logit <- log(Al203/(100-Al203))
par(mfrow=c(1,2))
h.cv <- hcv(logit, display = "lines", ngrid = 32)
n  <- length(logit)
sd <- sqrt(var(logit))
h  <- seq(0.003, 0.054, length=32)
lines(h, nmise(sd, n, h) - 5.5, lty = 3)
sm.density(logit, h.cv)
sm.density(logit, lty = 3, add = T)
par(mfrow=c(1,1))
```

S-Plus Illustration 2.4. Data based choices of smoothing for the aircraft data

Figure 2.4 was constructed with the following S-Plus *code.*

```
provide.data(aircraft)
provide.data(airpc)
y <- log(Span)[Period==3]
par(mfrow=c(1,2))
sm.density(y, h = hcv(y), xlab="Log span", lty=3, yht=1.4)
sm.density(y, h = hsj(y), add = T)
```



```
pc3 <- cbind(Comp.1, Comp.2)[Period==3,]
sm.density(pc3, h = hcv(pc3), display = "slice")
par(mfrow=c(1,1))
```

Mathematical aspects: Convolutions of normal densities

A variety of calculations with density estimates constructed from normal kernels lead to expressions involving the convolution of two normal densities. The following derivation of the computational form of the cross-validatory criterion is one example (it is convenient to give this convolution result a separate statement, for future reference):

$$\int \phi(y-\mu_1;\sigma_1)\phi(y-\mu_2;\sigma_2)dy = \phi\left(\mu_1-\mu_2;\sqrt{\sigma_1^2+\sigma_2^2}\right).$$

Here $\phi(z;\sigma)$ denotes the normal density function with mean 0 and standard deviation σ.

Mathematical aspects: The computational form of the cross-validatory criterion

The cross-validatory criterion (2.5) can be written in a more computationally convenient form when normal kernels are used. To derive a general expression, variable bandwidths h_i will be used. In the one-dimensional case, the criterion becomes

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{-i}^2(y)dy - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{(n-1)^2}\sum_{j\neq i}\sum_{k\neq i}\phi(y-y_j;h_j)\phi(y-y_k;h_k)dy \\
&\qquad -\frac{2}{n(n-1)}\sum_{i\neq j}\phi(y_i-y_j;h_j) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\phi(0;\sqrt{2}h_i) + \frac{(n-2)}{n(n-1)^2}\sum_{i\neq j}\phi\left(y_i-y_j;\sqrt{h_i^2+h_j^2}\right) \\
&\qquad -\frac{2}{n(n-1)}\sum_{i\neq j}\phi(y_i-y_j;h_j).
\end{aligned}$$

In the two-dimensional case, with product kernels $\phi(y_1-y_{1i};h_{1i})\phi(y_2-y_{2i};h_{2i})$, this becomes

$$\begin{aligned}
&\frac{1}{n(n-1)}\sum_{i=1}^{n}\phi(0;\sqrt{2}h_{1i})\phi(0;\sqrt{2}h_{2i}) \\
&+\frac{(n-2)}{n(n-1)^2}\sum_{i\neq j}\phi\left(y_{1i}-y_{1j};\sqrt{h_{1i}^2+h_{1j}^2}\right)\phi\left(y_{2i}-y_{2j};\sqrt{h_{2i}^2+h_{2j}^2}\right)
\end{aligned}$$

$$-\frac{2}{n(n-1)}\sum_{i\neq j}\phi(y_{1i}-y_{1j};h_{1j})\phi(y_{2i}-y_{2j};h_{2j}).$$

2.5 Testing normality

The tephra data exhibit some negative skewness. Before settling on a model for these data it would be helpful to assess whether this skewness is a real feature or simply the result of random variation. A probability plot of the data is a standard graphical device for checking normality. The left panel of Fig. 2.5 shows such a plot, from which it remains unclear whether there is strong evidence of non-normality. There are very many formal tests of normality; Stephens (1974) provides a comprehensive review. A simple but powerful approach is to assess the straightness of a probability plot, as described for example by Shapiro and Wilk (1965) and Filliben (1975). Applied to the tephra data this latter test does not detect significant non-normality.

One use of a nonparametric approach to modelling is in assessing the fit of a parametric model by comparing the two. The existence of density estimates raises the possibility that effective comparisons might be made on the natural density scale, rather than from a probability plot or empirical distribution function, which many other tests use. Goodness-of-fit statistics based on density functions have been proposed by Bickel and Rosenblatt (1973) and Bowman (1992) among others. The addition of new tests in this much studied area needs to be well motivated. However, density based statistics are of interest because they penalise departures from the hypothesised distribution in a different, and arguably more intuitive, way than more traditional procedures. They also have the potential to be equally easy to implement for multivariate data, which is not true of many other techniques.

There is a very large number of possible ways in which a proposed and an estimated density can be compared in a test statistic. One possibility is the integrated squared error

$$\int\left\{\hat{f}(y)-\phi\left(y-\hat{\mu};\sqrt{\hat{\sigma}^2+h^2}\right)\right\}^2 dy.$$

Notice that the density estimate is not compared to the proposed normal density $\phi(y-\hat{\mu};\hat{\sigma})$. Since the smoothing operation produces bias, it is more appropriate to contrast the density estimate with its mean under the assumption of normality, which is shown at the end of this section to be $\phi(y-\mu;\sqrt{\sigma^2+h^2})$.

In order to remove the effects of location and scale the data can be standardised to have sample mean 0 and sample variance 1. The test statistic on standardised data then becomes

$$\int\left\{\hat{f}(y)-\phi\left(y;\sqrt{1+h^2}\right)\right\}^2 dy.$$

When the kernel function is a normal density a convenient computational form of the statistic can be derived from results on convolutions. The statistic has a

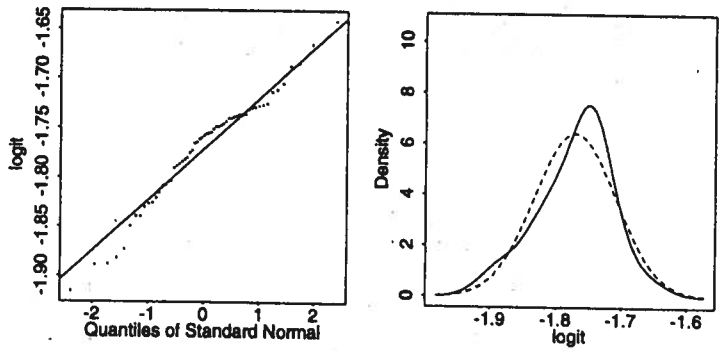


FIG. 2.5. The left panel shows a probability plot of the tephra data. The right panel shows a density estimate (full line) of the tephra data, and the mean value (dashed line) of this estimate under the assumption that the data are normally distributed.

mathematical structure which is similar to many of the well known test statistics for normality, such as those based on empirical distribution functions. However, as is the case with many goodness-of-fit statistics, it is computationally straightforward to derive percentage points by simulation.

In order to carry out the test, it is necessary to decide on the value of the smoothing parameter. Since this is a test for normality, it is natural to use normal optimal smoothing. When normality is an appropriate assumption the smoothing parameter will also be appropriate. When the data are not normally distributed, this form of smoothing will induce conservative behaviour, since the large value of the smoothing parameter will not promote the display of spurious features.

When applied to the tephra data, where the sample size is 59, the test statistic takes the value 0.00908. This is greater than the critical value in the table overleaf (divided by 1000), which was produced by simulation. The test has therefore identified significant non-normality. The different nature of the test statistic has allowed it to detect features which the probability plot did not.

Bowman (1992) investigated the power of the ISE statistic and found it to perform well in comparison with the best of the traditional procedures, such as the Shapiro–Wilk and Anderson–Darling tests. However, the strongest argument for the study of such tests is the ease with which analogous procedures can be defined in more than one dimension. Bowman and Foster (1993) investigated this for the multivariate normal distribution and provided tables of critical values. With multivariate data it will often be informative to construct Mahalanobis probability plots, or other diagnostic procedures which focus on important kinds of departures from normality. However, there is also a role for global tests of the

kind described above in cases such as discrimination, and exploratory techniques such as projection pursuit, where local values of the density function are essential components of the analysis.

The table below records the upper 5% points of the simulated distribution of the ISE statistic (multiplied by 1000 for convenience) under normality, for a variety of sample sizes and for up to three dimensions.

Sample size	Dimensions 1	2	3
25	109.0	64.2	30.3
50	76.6	50.7	23.1
100	56.7	35.5	16.7
150	45.3	27.2	14.3
200	38.0	23.1	11.8
250	33.2	20.5	10.9
300	30.1	18.9	9.64
350	27.2	16.9	8.89
400	23.2	16.1	8.50
500	20.5	13.9	7.47

S-Plus Illustration 2.5. Assessing normality for the tephra data

Figure 2.5 was constructed with the following S-Plus *code.*

```
provide.data(tephra)
logit <- log(A1203/(100-A1203))
par(mfrow=c(1,2))
qqnorm(logit)
qqline(logit)
cat("ISE statistic:", nise(logit),"\n")
sm <- sm.density(logit)
y  <- sm$eval.points
sd <- sqrt(hnorm(logit)^2 + var(logit))
lines(y, dnorm(y, mean(logit), sd), lty = 3)
par(mfrow=c(1,1))
```

Mathematical aspects: The exact mean and variance of a density estimate constructed from normal data

Using the result on convolutions of normal densities given at the end of Section 2.4, it follows that when the data have a normal distribution with mean μ and standard deviation σ, the mean of the density estimate is

$$\mathrm{E}\left\{\hat{f}(y)\right\} = \mathrm{E}_z\{\phi(y-z;h)\}$$

$$= \int \phi(y-z;h)\phi(z-\mu;\sigma)dz$$
$$= \phi\left(y-\mu;\sqrt{h^2+\sigma^2}\right).$$

For completeness, an expression is also derived for the variance:

$$\begin{aligned}\mathrm{var}\left\{\hat{f}(y)\right\} &= \frac{1}{n}\mathrm{var}_z\{\phi(y-z;h)\}\\ &= \frac{1}{n}[\mathrm{E}_z\{\phi^2(y-z;h)\} - \mathrm{E}_z\{\phi(y-z;h)\}^2]\\ &= \frac{1}{n}\left[\phi\left(0;\sqrt{2}h\right)\phi\left(y-\mu;\sqrt{\sigma^2+\tfrac{1}{2}h^2}\right) - \phi\left(y-\mu;\sqrt{\sigma^2+h^2}\right)^2\right].\end{aligned}$$

This will be used in Section 2.6.

2.6 Normal reference band

In the discussion of the test of normality it was natural to consider a graphical representation of the density estimate and a normal density curve, as in Fig. 2.5. This idea can be extended further. It was shown at the end of the previous section that when the true density function is normal with mean μ and variance σ^2, and the kernel function w is also normal, the mean and variance of the density estimate at the point y are

$$\mathrm{E}\left\{\hat{f}(y)\right\} = \phi\left(y-\mu;\sqrt{h^2+\sigma^2}\right),$$
$$\mathrm{var}\left\{\hat{f}(y)\right\} = \frac{1}{n}\left[\phi\left(0;\sqrt{2}h\right)\phi\left(y-\mu;\sqrt{\sigma^2+\tfrac{1}{2}h^2}\right) - \phi\left(y-\mu;\sqrt{\sigma^2+h^2}\right)^2\right].$$

These expressions allow the likely range of values of the density estimate to be calculated, under the assumption that the data are normally distributed. This can be expressed graphically through a *reference band*. At each point y of interest, the band is centred at $\mathrm{E}\left\{\hat{f}(y)\right\}$ and extends a distance $2se\{\hat{f}(y)\}$ above and below. The sample mean and variance can be used in place of the normal parameters μ and σ^2.

Figure 2.6 has reference bands superimposed on density estimates for the tephra data, using two different smoothing parameters. In both cases the peak of the observed curve lies just outside the reference band. As a follow-up to the goodness-of-fit test, this indicates a sharper peak than a normal density and the presence of some negative skewness. It should be emphasised that this is not intended as a replacement for the global test of normality. It does, however, provide a useful graphical follow-up procedure by indicating the likely position of a density estimate when the data are normal. This can be helpful in identifying where non-normal features might lie, or in understanding, through the displayed standard error structure, why some apparently non-normal feature has not led to a significant test result.

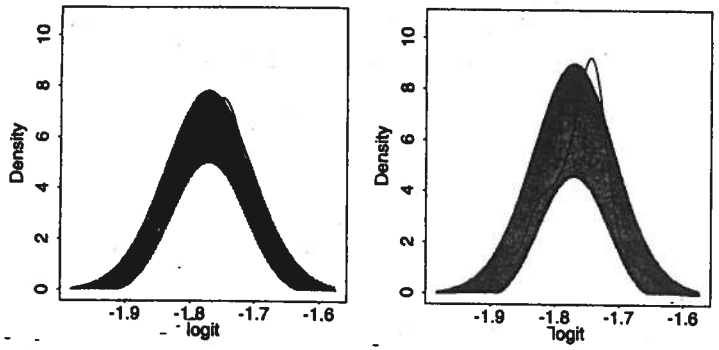


FIG. 2.6. Density estimates of the tephra data, with reference bands for a normal distribution. The left panel displays an estimate and band using the normal optimal smoothing parameter. The right panel uses the Sheather–Jones plug-in smoothing parameter.

S-Plus Illustration 2.6. Reference bands for normality with the tephra data

Figure 2.6 was constructed with the following S-Plus *code.*

```
provide.data(tephra)
logit <- log(Al203/(100-Al203))
par(mfrow=c(1,2))
sm.density(logit, model = "Normal")
sm.density(logit, h = hsj(logit), model = "Normal")
par(mfrow=c(1,1))
```

2.7 Testing independence

In two dimensions, other statistical problems become important. One is to assess the presence, or strength, of dependence between two variables. While the correlation coefficient is very useful for linear structures, there are nonlinear patterns which can be missed. An example arises with the aircraft data. Figure 2.7 displays a scatterplot of the variables span and speed, both on log scales, for the third time period. In order to assess whether these two variables are associated, a natural starting point is the correlation coefficient. For these data this takes the value 0.017, which is so close to 0 that it does not provide even a hint of linear correlation. A standard nonparametric approach is to calculate the rank correlation coefficient. This takes the value 0.023, which again gives no evidence of a relationship between the two variables.

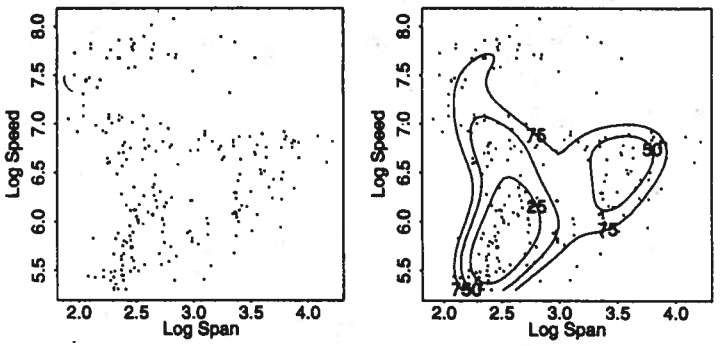


FIG. 2.7. The left panel displays the aircraft span and speed data, on log scales, for the third time period, 1956–1984. The right panel shows the same data with a density contour plot superimposed.

The existence of density estimates gives the opportunity to explore dependence in a more flexible way. A general definition of independence is that the joint density $f_{12}(y_1, y_2)$ of two variables decomposes into the product of the marginal densities $f_1(y_1)$ and $f_2(y_2)$. An assessment of independence in a very general sense can then be performed by contrasting these joint and marginal density estimates. There are many ways in which this could be done. However, a natural approach is to construct a likelihood ratio expression

$$\frac{1}{n}\sum \log\left\{\frac{\hat{f}_{12}(y_{1i}, y_{2i})}{\hat{f}_1(y_{1i})\hat{f}_2(y_{2i})}\right\}.$$

Kent (1983) and Joe (1989) considered statistics of this kind. A computational approach to the derivation of the distribution of this test statistic under the null hypothesis of independence is to apply a permutation argument, where values of y_{1i} are associated with randomly permuted values of y_{2i}. An empirical p-value can be calculated as the proportion of statistics computed from the permuted data whose values exceed that of the observed statistic from the original data. With the aircraft span and speed data, the empirical p-value is 0.00, based on a simulation size of 200.

The striking difference in the linear and nonparametric assessments of association between these variables is explained by the right hand panel of Fig. 2.7. The density contours show clear association, with span and speed generally increasing together, but with a significant cluster of observations falling in the region of low span and high speed. The presence of these observations causes the linear association between the variables to reduce almost to 0.

Bjerve and Doksum (1993), Doksum *et al.* (1994) and Jones (1996) suggest how dependence between variables can be quantified in a local way through the definition of correlation curves and local dependence functions.

S-Plus Illustration 2.7. Density contour plots for exploring independence

Figure 2.7 was constructed with the following S-Plus *code.*

```
provide.data(aircraft)
Speed3 <- log(Speed[Period==3])
Span3  <- log(Span[Period==3])
par(mfrow=c(1,2))
plot(Span3, Speed3, xlab = "Log Span", ylab = "Log Speed")
air3  <- cbind(Span3, Speed3)
sm.density(air3, display="slice",
        xlab= "Log Span", ylab = "Log Speed")
par(mfrow=c(1,1))
```

An additional script to carry out the permutation test is available.

2.8 The bootstrap and density estimation

The material of earlier sections shows that the form of the bias and variance of a density estimate complicates any attempt to carry out simple forms of inference, including even a simple confidence interval for the density estimate at a point. In such situations, the bootstrap sometimes provides an attractive tool, giving a potential means of carrying out inference by resampling methods, where analytic methods prove too complex. Davison and Hinkley (1997) give a general description of the bootstrap method.

A simple bootstrap procedure for density estimates is as follows.

1. Construct a density estimate $\hat{f}$ from the observed data $\{y_1, \ldots, y_n\}$.
2. Resample the data $\{y_1, \ldots, y_n\}$ with replacement to produce a bootstrap sample $\{y_1^*, \ldots, y_n^*\}$.
3. Construct a density estimate $\hat{f}^*$ from the bootstrap data $\{y_1^*, \ldots, y_n^*\}$.
4. Repeat steps 2 and 3 a large number of times to create a collection of bootstrap density estimates $\{\hat{f}_1^*, \ldots, \hat{f}_B^*\}$.
5. Use the distribution of $\hat{f}_i^*$ about $\hat{f}$ to mimic the distribution of $\hat{f}$ about f.

It is worthwhile checking the validity of step 5. To do this, consider the mean of $\hat{f}^*$. Since the distribution of y_i^* is uniform over $\{y_1, \ldots, y_n\}$, it follows that

$$\mathrm{E}\left\{\hat{f}^*(y)\right\} = \mathrm{E}\{w(y - y_i^*; h)\} = \hat{f}(y),$$

from which it is immediately apparent that the bias which is present in the distribution of $\hat{f}$ is absent in the bootstrap version. The bootstrap distribution of $\hat{f}^*$ is therefore missing an essential component. Hall (1992) gives appropriate

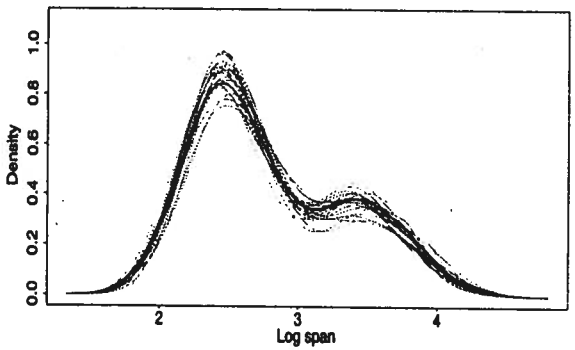


FIG. 2.8. Bootstrap density estimates from the aircraft span data, as an alternative method of constructing variability bands.

theory for this situation and shows that the bootstrap does correctly mimic the variance of $\hat{f}$. Bootstrapping can therefore be used to provide an alternative means of generating the variability bands described in Section 2.3. Figure 2.8 displays density estimates from multiple bootstrap samples of the aircraft span data. This is a bootstrap analogue of Fig. 2.2.

Bootstrapping has a very useful role to play and it can be particularly helpful in some hypothesis testing contexts. Silverman (1981) describes an interesting case where inference can be carried out on the number of modes present in a density, using a technique known as the smoothed bootstrap, which involves simulating from $\hat{f}$ rather than resampling the original data. Taylor (1989) used the smoothed bootstrap to construct an estimate of mean integrated squared error and proposed this as a means of selecting a suitable smoothing parameter. Scott (1992, Section 9.3.2) discusses and illustrates the role of the smoothed bootstrap in constructing confidence intervals.

S-Plus Illustration 2.8. Bootstrapping density estimates

Figure 2.8 was constructed with the following S-Plus *code.*

```
provide.data(aircraft)
y <- log(Span)[Period==3]
sm.density(y, xlab = "Log span")
for (i in 1:20) sm.density(sample(y, replace=T), col=6, add=T)
sm.density(y, xlab = "Log span", add=T)
```

The number of bootstrap estimates produced is controlled by the range of `i` *in the* `for` *loop. On some computers these simulations may take a long time.*

4.2 The kernel method in several dimensions

In this section the kernel method for the multivariate case will be introduced and some comparisons made with multivariate histograms and scatter plots. Throughout the chapter, bold face will be used for points in d-dimensional space. It will be assumed that $\mathbf{X}_1, \ldots \mathbf{X}_n$ is the given multivariate data set whose underlying density is to be estimated.

4.2.1 *Definition of the multivariate kernel density estimator*

The definition of the kernel estimator as a sum of 'bumps' centred at the observations is easily generalized to the multivariate case. The multivariate kernel density estimator with kernel K and window width h is defined by

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left\{\frac{1}{h}(\mathbf{x} - \mathbf{X}_i)\right\}. \tag{4.1}$$

The kernel function $K(\mathbf{x})$ is now a function, defined for d-dimensional $\mathbf{x}$, satisfying

$$\int_{R^d} K(\mathbf{x})\, d\mathbf{x} = 1. \tag{4.2}$$

Usually K will be a radially symmetric unimodal probability density function, for example the standard multivariate normal density function

$$K(\mathbf{x}) = (2\pi)^{-d/2} \exp(-\tfrac{1}{2}\mathbf{x}^T\mathbf{x}). \tag{4.3}$$

Another possible kernel is the multivariate Epanechnikov kernel

$$K_e(\mathbf{x}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1 - \mathbf{x}^T\mathbf{x}) & \text{if } \mathbf{x}^T\mathbf{x} < 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

where c_d is the volume of the unit d-dimensional sphere: $c_1 = 2, c_2 = \pi$, $c_3 = 4\pi/3$, etc. We shall see in Section 4.4 that useful kernels for the case $d = 2$ are given by

$$K_2(\mathbf{x}) = \begin{cases} 3\pi^{-1}(1 - \mathbf{x}^T\mathbf{x})^2 & \text{if } \mathbf{x}^T\mathbf{x} < 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

and

$$K_3(\mathbf{x}) = \begin{cases} 4\pi^{-1}(1 - \mathbf{x}^T\mathbf{x})^3 & \text{if } \mathbf{x}^T\mathbf{x} < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

The advantage of these kernels over the Epanechnikov kernel is that the kernels, and hence the resulting density estimates, have higher differentiability properties. In addition, they can be calculated more quickly than the normal kernel (4.3). For further discussion of computational aspects see Section 4.4.

The use of a single smoothing parameter h in (4.1) implies that the version of the kernel placed on each data point is scaled equally in all directions. In certain circumstances, it may be more appropriate to use a vector of smoothing parameters or even a matrix of shrinking coefficients. This will be the case, for example, if the spread of the data points is very much greater in one of the coordinate directions than the others. Just as for many other multivariate statistical procedures, it is probably best to pre-scale the data to avoid extreme differences of spread in the various coordinate directions. If this is done then there will generally be no need to consider more complicated forms of the kernel density estimate than the form (4.1) involving a single smoothing parameter.

An attractive intuitive approach, suggested by Fukunaga (1972, p. 175) is first to 'pre-whiten' the data by linearly transforming them to have unit covariance matrix; next to smooth using a radially

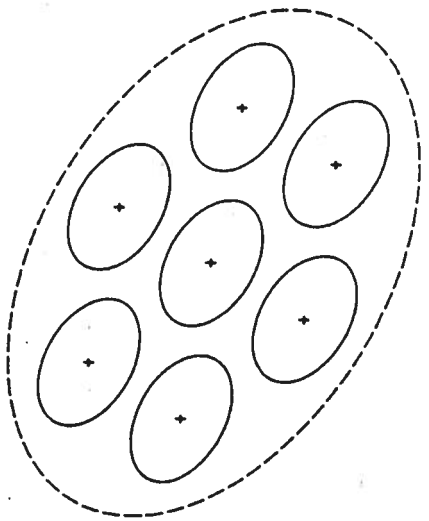


Fig. 4.1 *Shape of kernel and shape of distribution if Fukunaga method is used. After Fukunaga (1972).*

symmetric kernel; and finally to transform back. This is equivalent to using a density estimate of the form

$$\hat{f}(\mathbf{x}) = \frac{(\det S)^{-1/2}}{nh^d} \sum_{i=1}^{n} k\{h^{-2}(\mathbf{x}-\mathbf{X}_i)^{\mathrm{T}} S^{-1}(\mathbf{x}-\mathbf{X}_i)\}, \qquad (4.7)$$

where the function k is given by

$$k(\mathbf{x}^{\mathrm{T}}\mathbf{x}) = K(\mathbf{x})$$

and S is the sample covariance matrix of the data. If, for instance, K is the normal kernel, then $k(u)$ is equal to $(2\pi)^{-d/2}\exp(-\frac{1}{2}u)$. It would perhaps be advisable to use a robust version of the sample covariance for S; see the discussion of 'sphering' data given in Tukey and Tukey (1981, Section 10.2.2). The idea of (4.7), illustrated in Fig. 4.1, is to use a kernel which has the same shape as the data set itself.

4.2.2 *Multivariate histograms*

The arguments for using density estimates rather than histograms become much stronger in two or more dimensions. The construction of a multivariate histogram requires the specification not only of the size of the bins and the origin of the system of bins, but also the orientation of the bins. In addition there are severe presentational difficulties, even in the two-dimensional case. Figure 4.2, reproduced from Scott (1982), shows a typical bivariate histogram. Partly because

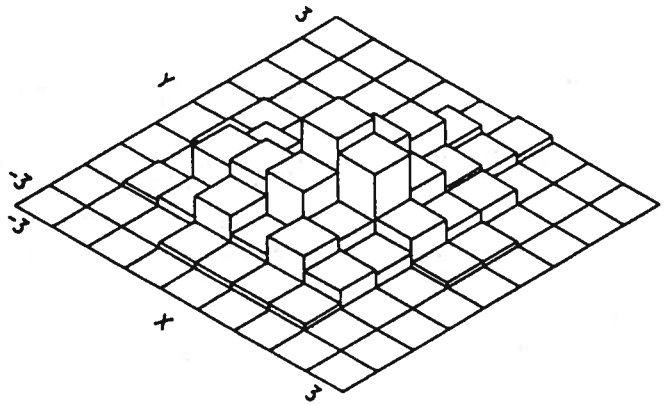


Fig. 4.2 *A typical bivariate histogram. Reproduced from Scott (1982) with the permission of the author.*

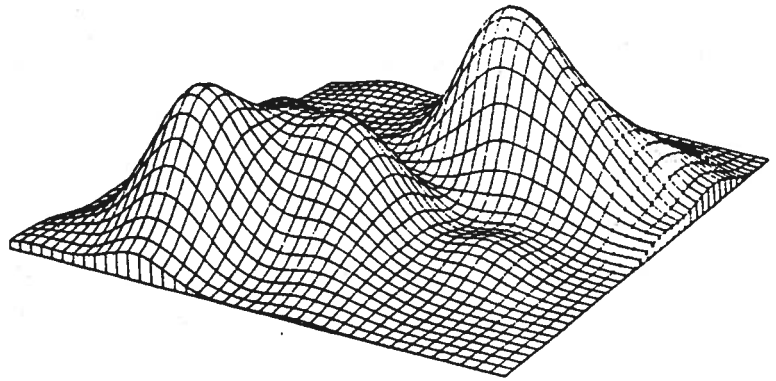


Fig. 4.3 *Density estimate, displayed on the square* $(\pm 3, \pm 3)$, *for 100 observations from bivariate normal mixture, window width 1.2.*

of the discontinuous 'block' nature of the bivariate histogram, it is difficult without some experience of looking at diagrams of this type to grasp the structure of the data.

Because they are continuous surfaces, bivariate density estimates constructed using continuous kernels are much easier to comprehend, either as perspective views or as contour plots. For example, Figs 4.3, 4.4 and 4.5, constructed from 100 data points drawn from a bivariate normal mixture, provide a clear picture of the underlying

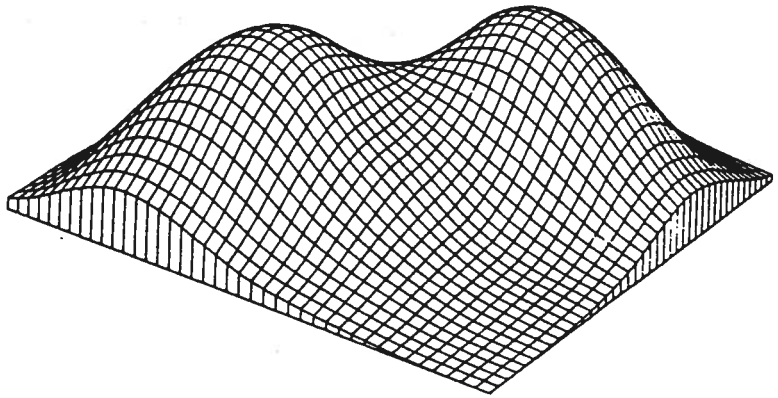


Fig. 4.4 *Density estimate for 100 observations from bivariate normal mixture, window width 2.2.*

approximately optimal window width of (4.12) converges to zero as n increases, but does so extremely slowly, at the rate $n^{-1/(d+4)}$. Furthermore, the appropriate value of h depends on the unknown density being estimated.

The value h_{opt} can be substituted back into (4.11) to give the approximate minimum possible mean integrated square error, and hence, as in Section 3.3.2, to guide the choice of kernel. The Epanechnikov kernel (4.4) is optimum among non-negative kernels in the sense of minimizing the smallest mean integrated square error achievable; however, just as before, the other kernels defined in Section 4.2.1 can achieve very similar mean integrated square errors. Again, it is appropriate to base the choice of kernel on other considerations, notably the remarks made in Section 4.4 concerning computational aspects.

4.3.2 *Choice of window width for a standard distribution*

The first step beyond choosing the smoothing parameter entirely subjectively is to use the formula (4.12) to provide the appropriate value of the window width when f is a standard density, such as the multivariate normal. If ϕ is the unit d-variate normal density, then it can be shown that

$$\int (\nabla^2 \phi)^2 = (2\sqrt{\pi})^{-d}(\tfrac{1}{2}d + \tfrac{1}{4}d^2). \tag{4.13}$$

The value given by (4.13) can then be substituted back in (4.12) to give the optimal window width for the smoothing of normally distributed data with unit variance. The window width will be given by

$$h_{opt} = A(K)n^{-1/(d+4)} \tag{4.14}$$

where the constant

$$A(K) = \left[d\beta\alpha^{-2} \left\{ \int (\nabla^2 \phi)^2 \right\}^{-1} \right]^{1/(d+4)} \tag{4.15}$$

depends on the kernel and is tabulated in Table 4.1.

Consider, now, the smoothing of a general data set with covariance matrix S, possibly robustly estimated. If the Fukunaga estimate (4.7) is being used, then the h_{opt} of (4.14) gives a directly appropriate value for the smoothing parameter. On the other hand, if the kernel is radially symmetric and the data are not transformed, the procedure indicated

Table 4.1 *The constant $A(K)$, as defined in equation (4.15), for various kernels K. Numerical values are given correct to 2 decimal places*

Kernel	*Dimensionality*	$A(K)$
Multivariate normal K as in equation (4.3)	2	0.96
	d	$\{4/(2d+1)\}^{1/(d+4)}$
K_e as in equation (4.4)	2	1.77
	3	2.78
	d	$\left\{ \dfrac{8d(d+2)(d+4)(2\sqrt{\pi})^d}{(2d+1)c_d} \right\}^{1/(d+4)}$
K_2 as in equation (4.5)	2	2.04
K_3 as in equation (4.6)	2	2.29

would be to choose a single scale parameter σ for the data and to use the value σh_{opt} for the window width. A possible choice for σ might be

$$\sigma^2 = d^{-1} \sum_i s_{ii},$$

so that σ^2 is the average marginal variance.

The cautionary remarks made in Section 3.4.2 about the use of the window width $1.06\sigma n^{-1/5}$ for the univariate case apply more strongly for the corresponding multivariate value given by (4.14), and in particular it will often be appropriate to use a slightly smaller value. Nevertheless, the method described above does give a quick and easy choice of at least an initial value for the window width.

4.3.3 *More sophisticated ways of choosing the window width*

Both least-squares cross-validation and likelihood cross-validation carry over to the multivariate case with no essential modification at all. The least-squares cross-validation score $M_1(h)$ of (3.39) becomes

$$M_1(h) = n^{-2}h^{-d} \sum_i \sum_j K^*\{h^{-1}(\mathbf{X}_i - \mathbf{X}_j)\} + 2n^{-1}h^{-d}K(0). \tag{4.16}$$

It is interesting to note that the computational effort required to calculate $M_1(h)$ from (4.16) depends on the dimensionality d only in

various standard distributions will be assumed; for details of how these are generated in practice, see, for example, Ripley (1983) and the references cited there.

6.4.1 *Simulating from density estimates*

An idealized form of the problem under consideration in this section is easily formulated mathematically. Given a sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ from an unknown density f, we are required to construct a sequence of independent observations $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ from f. Of course, this will be impossible to achieve exactly in practice because full information about f is not available. The observations in the sample $\{\mathbf{X}_i\}$ and the required realizations $\{\mathbf{Y}_j\}$ will be assumed to be d-dimensional vectors.

The usual approach to the problem is one of two extremes. A parametric form for f could be assumed, such as the normal distribution with unknown parameters; the sample $\{\mathbf{X}_i\}$ is used to estimate the unknown parameters, and a standard simulation method is then used to generate the required simulated observations. Alternatively, the realizations $\{\mathbf{Y}_j\}$ are generated directly by successive random sampling, with replacement, from the sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$. This latter approach has the advantage of freeing the procedure from parametric assumptions but the serious disadvantage, in some contexts, of making it impossible for any value to occur in the simulated data that has not occurred exactly in the original sample $\{\mathbf{X}_i\}$. This behaviour may be unacceptable because spurious very fine structure in the original data may well be faithfully reproduced in the simulated samples. This may not present a problem in practice, depending on the context. However, difficulties will arise if the ideal aim of the stimulation is to produce samples that have the underlying 'true' structure of the observed data without sharing spurious details that have arisen from random effects.

It is very natural to consider an intermediate approach based on density estimation. The observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ can be used to construct a nonparametric estimate $\hat{f}$ of the density f, and then as many independent realizations as required can be drawn from $\hat{f}$. Depending on the context, it may well be desirable to simulate not from $\hat{f}$ itself but from a version transformed to have the same mean vector and covariance matrix as the observed data; we shall return to this refinement below.

If $\hat{f}$ is constructed by the kernel method or the adaptive kernel method, then it is very easy to find independent realizations from $\hat{f}$, provided a non-negative kernel is used. Indeed, it is not even necessary to find $\hat{f}$ explicitly in the simulation procedure. Consider the univariate case, and suppose $\hat{f}$ has been constructed by the ordinary kernel method with kernel K and window width h. Realizations Y from $\hat{f}$ can be generated as follows:

Step 1 Choose I uniformly with replacement from $\{1, \ldots, n\}$
Step 2 Generate ε to have probability density function K
Step 3 Set $Y = X_I + h\varepsilon$. (6.15)

It is necessary in Step 2 to generate a random observation from the kernel K. If K is the normal density, there are several standard ways of doing this, such as the Box–Muller and Marsaglia approaches discussed, for example, in Ripley (1983, Section 4.2). Devroye and Györfi (1985, p. 236) give a very fast algorithm for simulation from the rescaled Epanechnikov kernel

$$K(x) = \tfrac{3}{4}(1 - x^2) \text{ for } |x| \leqslant 1$$

as follows:

Step 2a Generate three uniform $[-1, 1]$ random variates V_1, V_2, V_3.

If $|V_3| \geqslant |V_2|$ and $|V_3| \geqslant |V_1|$, set $\varepsilon = V_2$;

otherwise set $\varepsilon = V_3$.

The algorithm (6.15) can be repeated as often as necessary to give independent realizations Y_j from $\hat{f}$. It is fast and easy to program. Proofs that (6.15) really does give an observation from $\hat{f}$, and that Step 2a has the desired effect, are exercises in elementary probability theory and are omitted. A multivariate version of (6.15) is easily constructed; unequal smoothing parameters in the various coordinate directions, or the use of a matrix of shrinking coefficients, are easily coped with by using the corresponding transformation in Step 3.

If the realizations Y are required to have first and second moment properties the same as those observed in the sample $\{X_1, \ldots, X_n\}$, then Step 3 in (6.15) should be replaced by

$$\textit{Step 3}'\quad Y = \bar{X} + (X - \bar{X} + h\varepsilon)/(1 + h^2\sigma_K^2/\sigma_X^2)^{1/2} \tag{6.16}$$

where $\bar{X}$ and σ_X^2 are the sample mean and variance of $\{X_i\}$ and σ_K^2 is

the variance of the kernel K. In the multivariate case, the step corresponding to Step 3′ is simplest if the kernel is scaled to have variance matrix the same as the data; it then becomes

Step 3″ $\mathbf{Y} = \bar{\mathbf{X}} + (\mathbf{X} - \bar{\mathbf{X}} + h\varepsilon)/(1 + h^2)^{1/2}$. (6.17)

Notice that the original algorithm (6.15) will yield realizations with expected value equal to $\bar{X}$, but that the smoothing will increase the variance unless a correction like Step 3′ or 3″ is used.

Modifying (6.15) to give realizations from an adaptive kernel estimate is straightforward. The last step is replaced by

Step 3* Set $Y = X_I + h\lambda_I\varepsilon$ (6.18)

where h and $\lambda_1, \ldots, \lambda_n$ are defined as in (5.7) and (5.8) above.

Some further discussion of the material of this section is given in Devroye and Györfi (1985, Section 8). They give some theoretical background that shows that, perhaps not surprisingly, very large sample sizes n are required if one is to be confident that moderately long sequences $Y_1, \ldots, Y_m$ are to be practically indistinguishable, in all respects, from sequences generated from the true density f. They also give some additional algorithms for simulation from density estimates of various kinds.

6.4.2 *The bootstrap and the smoothed bootstrap*

The bootstrap is an appealing approach to the assessment of errors and related quantities in statistical estimation. The method is described and explored in detail by Efron (1982), and only a brief explanation will be given here.

Let $\rho(F)$ be some interesting property of a distribution F that depends in some complicated way on F. Typically, even if F is known, ρ can most easily be estimated by repeatedly simulating samples from F. For example, ρ might be the sampling variance of the upper quartile of samples of size 39 drawn from F; a way of finding ρ to within reasonable accuracy is to simulate 1000 samples of size 39 from F, to find the upper quartile of each of these samples, and then to calculate the sample variance of these 1000 values.

In many statistical problems, F itself is unknown but a sample $X_1, \ldots, X_n$ of observations from F is available. The standard *bootstrap* approach is to estimate $\rho(F)$ using the procedure just described, but to simulate the samples not from F itself but from the empirical

distribution function F_n of the observed data $X_1, \ldots, X_n$. A sample from F_n is generated by successively selecting uniformly with replacement from $\{X_1, \ldots, X_n\}$. This approach approximates $\rho(F)$ by $\rho(F_n)$, and an example will be given below.

The samples constructed from F_n in the bootstrap simulations will have some rather peculiar properties. All the values taken by members of these samples will be drawn from the original values $X_1, \ldots, X_n$ and nearly every sample will contain repeated values. If n is at all large, most samples will contain some values repeated several times.

An approach that does not lead to samples with these properties is the *smoothed bootstrap*. Here the simulations are constructed not from F_n but by using an algorithm like (6.15) to simulate from a smoothed version of F_n. If $\hat{F}$ is the distribution function of the density estimate $\hat{f}$, then the effect of the smoothed bootstrap will be to estimate $\rho(F)$ by $\rho(\hat{F})$. Whether $\rho(F)$ is better estimated by $\rho(F_n)$ or $\rho(\hat{F})$ will depend on the context

An example where the smoothed bootstrap does well is given by Efron (1981). Let F be a bivariate distribution of random vectors $\mathbf{X} = (\xi, \eta)$, and let $\phi(F)$ be Fisher's variance-stabilized transformed correlation coefficient

$$\phi(F) = \tanh^{-1} \operatorname{corr}(\xi, \eta). \tag{6.19}$$

Given a sample $\mathbf{X}_1, \ldots, \mathbf{X}_{14}$, an estimate $\hat{\phi}(\mathbf{X}_1, \ldots, \mathbf{X}_{14})$ is constructed by substituting the usual sample correlation coefficient into (6.19). Let $\rho(F)$ be the sampling standard deviation of $\hat{\phi}(\mathbf{X}_1, \ldots, \mathbf{X}_{14})$ if $\mathbf{X}_1, \ldots, \mathbf{X}_{14}$ are drawn from F. Efron (1981) considered the case where F is a bivariate normal distribution with $\operatorname{var}(\xi) = \operatorname{var}(\eta) = 1$ and $\operatorname{cov}(\xi, \eta) = \frac{1}{2}$. For this case, the true value of $\rho(F)$ is 0.299. In 200 Monte Carlo trials, the bootstrap estimated $\rho(F)$ with root-mean-square error 0.065, while the smoothed bootstrap had a root-mean-square error of only 0.041, a substantial improvement. The smoothed bootstrap was implemented using the algorithm (6.15) with a normal kernel with the same covariance matrix as the data; Step 3″ of (6.17) was used to ensure that $\hat{F}$ had the same second-order characteristics as F_n. The smoothing parameter h in (6.17) was set to 0.5.

There has, as yet, been very little systematic investigation of the circumstances under which the smoothed bootstrap will give better results than the ordinary bootstrap. In addition, the choice of smoothing parameter in the smoothed bootstrap is usually made

completely arbitrarily; detailed work on this choice, making use of known properties of density estimates, would no doubt improve the performance of the smoothed bootstrap. A perhaps somewhat controversial problem where the smoothing parameter is chosen in a natural way is discussed in the next section.

6.4.3 *A smoothed bootstrap test for multimodality*

The smoothed bootstrap can be used to construct a test for multimodality based on the critical smoothing idea discussed in Section 6.3.3 above. Given a data set $X_1,\ldots,X_n$ yielding a value h_0 for the critical window width h_{crit}, one needs to decide whether h_0 is a surprisingly large value for the statistic h_{crit}. To do this, h_0 has to be assessed against some suitable unimodal null density f_0 for the data. An approach suggested in Section 6.3.3 was to use a member of a standard parametric family for f_0, but the smoothed bootstrap makes a more nonparametric approach possible.

A suitable choice of f_0 would have various desirable properties, as follows:

(a) The density f_0 must be unimodal, since f_0 must be a representative of the compound null hypothesis of unimodality.
(b) Subject to (a), f_0 should be a plausible density underlying the data; testing against all possible unimodal densities is a hopeless task, since, for example, large values of h_{crit} would be obtained from unimodal densities with very large variances.
(c) In order to give unimodality a fair chance of explaining the data, f_0 should be, in some sense, the most nearly bimodal among those densities satisfying (a) and (b).

A very natural way of constructing f_0 to satisfy these requirements is to set f_0 equal to the density estimate $\tilde{f}_{crit}$ constructed from the original data with window width h_0. It can be shown (see Silverman, 1981b) that $\tilde{f}_{crit}$ is unimodal; as a density estimate constructed from the data it is a plausible density for the data; and, by the definition of h_{crit}, reducing the window width any further would make $\tilde{f}$ multimodal, so $\tilde{f}_{crit}$ is an 'extreme' unimodal density. In fact it can be shown that, as the sample size increases, the asymptotic behaviour of h_{crit} is such as to yield good estimates of the true density f if f is indeed unimodal; see Silverman (1983).

Sampling from $\tilde{f}_{crit}$ corresponds precisely to the smoothed bootstrap algorithm (6.15) with h equal to h_0. In assessing the significance value of h_0, it is necessary only to ascertain the proportion of samples of size n from $\tilde{f}_{crit}$ that lead to values of h_{crit} greater than h_0. A given sample will have $h_{crit} > h_0$ if and only if the density estimate constructed from the sample with window width h_0 is multimodal. Thus an algorithm for assessing the p-value of h_0 is to generate a large number of samples from $\tilde{f}_{crit}$ and to count the proportion of samples which yield a multimodal density estimate using window width h_0. There is no need to find h_{crit} for each sample.

An example of the application of this technique is given in Silverman (1981b). A sample of 22 observations on chondrite meteors was investigated for multimodality. The original data are given in Good and Gaskins (1980, Table 2). The estimated significance value of h_{crit} for unimodality was 8%, calculated by the procedure discussed in this section. For further details the reader is referred to Silverman (1981b), but warned that the p-values printed in Table 1 of that paper are incorrect and should all be subtracted from 1.

6.5 Estimating quantities that depend on the density

Often it is not so much the density itself that is of interest but some curve or single numerical quantity that can be derived from the density. An important example of a curve that depends on the density is the hazard rate, the estimation of which will be discussed in Section 6.5.1. A single numerical quantity that depends on the density is called a *functional* of the density; a general discussion of the estimation of functionals of the density will be given in Section 6.5.2, and a particular application will be developed in Section 6.5.3.

6.5.1 *Hazard rates*

Given a distribution with probability density function f, the *hazard rate* $z(t)$ is defined by

$$z(t) = \frac{f(t)}{\int_t^\infty f(u)\mathrm{d}u} \tag{6.20}$$

The hazard rate is also called the age-specific or conditional failure rate. It is useful particularly in the contexts of reliability theory and survival analysis and hence in fields as diverse as engineering and medical statistics. Cox and Oakes (1984), for example, give a