Chapter 2: Visualizing Multivariate Data

Contents

1. Introduction

According to Chambers, Cleveland, Kleiner, and Tukey (1983), "There is no statistical tool that is as powerful as a well-chosen graph". Certainly graphical presentation has a number of advantages over tabular displays of numerical results, not least in creating interest and attracting the attention of the viewer.

What is a graphical display?

Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading and color.

Some of the advantages of graphical methods

- In comparison with other types of presentation, well-designed charts are more effective in creating interest and in appealing to the attention of the reader.
- Visual relationships as portrayed by charts and graphs are more easily grasped and more easily remembered.
- The use of charts and graphs saves time since the essential meaning of large measures of statistical data can be visualized at a glance.
- Charts and graphs provide a comprehensive picture of a problem that makes for a more complete and better balanced understanding than could be derived from tabular or textual forms of presentation.
- Charts and graphs can bring out hidden facts and relationships and can stimulate, as well as aid, analytical thinking and investigation.

Goals for graphical displays of data

- To provide an overview;
- To tell a story;
- To suggest hypotheses;
- To criticize a model.

Do not forget to load the package “MVA”

Under R-2.15.1

Depends HSAUR2

http://www.r-project.org/

MVA: An Introduction to Applied Multivariate Analysis with R

Functions, data sets, analyses and examples from the book ‘An Introduction to Applied Multivariate Analysis with R’ (Brian S. Everitt

Version:	1.0-3
Depends:	HSAUR2
Suggests:	mvtnorm, mclust, lattice, flexclust, nlme, RLRsim, multcomp, ape, MASS, sem, KernSmooth, scatterplot3d
Published:	2012-03-07
Author:	Brian S. Everitt and Torsten Hothorn
Maintainer:	Torsten Hothorn <Torsten.Hothorn at R-project.org>
License:	GPL-2
URL:	http://dx.doi.org/10.1007/978-1-4419-9650-3
CRAN checks:	MVA results

Downloads:

Package source:	MVA_1.0-3.tar.gz
MacOS X binary:	MVA_1.0-3.tgz
Windows binary:	MVA_1.0-3.zip
Reference manual:	MVA.pdf
Old sources:	MVA archive

2. Graphics

The scatterplot

The scatterplot is the standard for representing continuous bivariate data but, as we shall see later in this chapter, it can be enhanced in a variety of ways to accommodate information about other variables.
To illustrate the use of the scatterplot and a number of other techniques to be discussed, we shall use the air pollution in US cities data introduced in the previous chapter (see Table 1.5). Let's begin our examination of the air pollution data by taking a look at a basic scatterplot of the two variables manu and popul. For later use, we first set up two character variables that contain the labels to be printed on the two axes:

```
mlab <- "Manufacturing enterprises with 20 or more workers"
plab <- "Population size (1970 census) in thousands"
plot(popul ~ manu, data = USairpollution, xlab = mlab, ylab = plab)
```

The resulting scatterplot is shown in Figure 2.1.

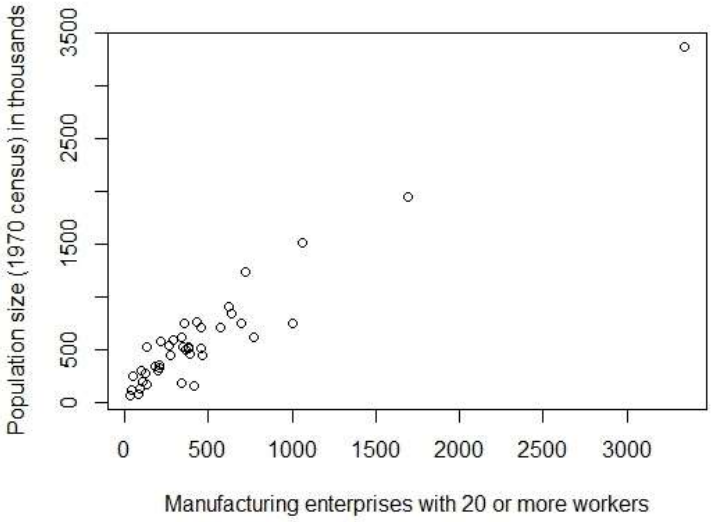


Fig. 2.1. Scatterplot of manu and popul.

The plot clearly uncovers the presence of one or more cities that are some way from the remainder, knowing as outliers. To identify the outliers we use

```
plot(popul ~ manu, data = USairpollution, xlab = mlab, ylab = plab)
text(popul ~ manu, data=USairpollution, xlab = mlab, ylab = plab , labels=cities)
```

The resulting scatterplot is shown in Figure 2.2.

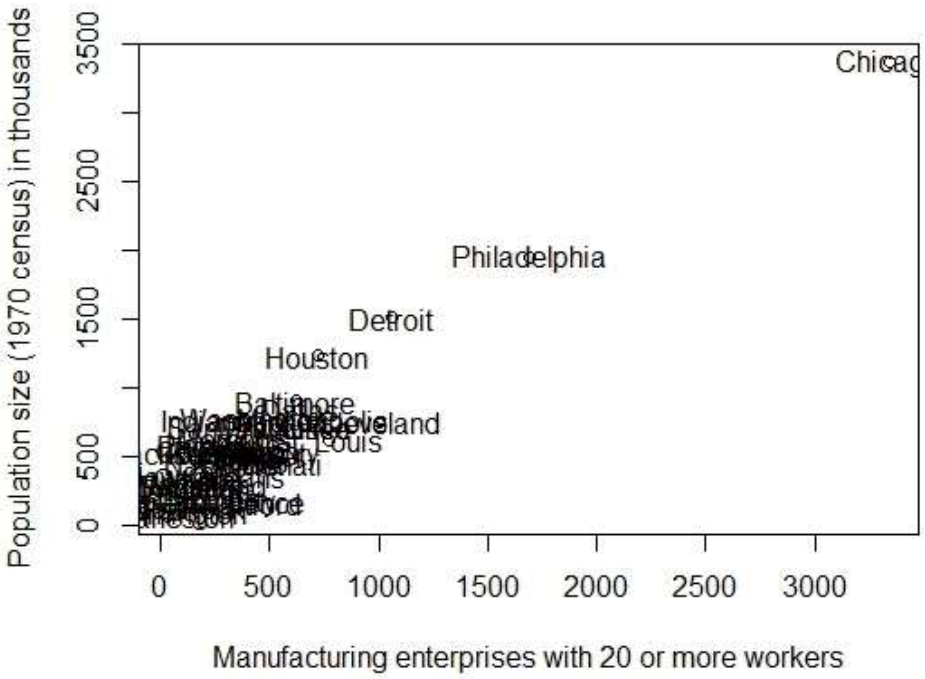


Fig. 2.2. Scatterplot of manu and popul.

From Figure 2.2, we can see that the outlying points show themselves in scatterplot of the variables. The most extreme outlier corresponds to Chicago, and other slightly less extreme outliers correspond to Philadelphia and Detroit. Each of these cities has a considerably larger population than other cities and also many more manufacturing enterprises with more than 20 workers.

In Figure 2.2, the marginal distributions are shown as rug plots on each axis (produced by rug()).

```
plot(popul ~ manu, data = USairpollution, xlab = mlab, ylab = plab)
text(popul ~ manu, data=USairpollution, xlab = mlab, ylab = plab , labels=cities)
rug(USairpollution[,c("manu")], side = 1)
rug(USairpollution[,c("popul")], side = 2)

manu<-x[,c("manu")]
hist(manu)
boxplot(manu)
```

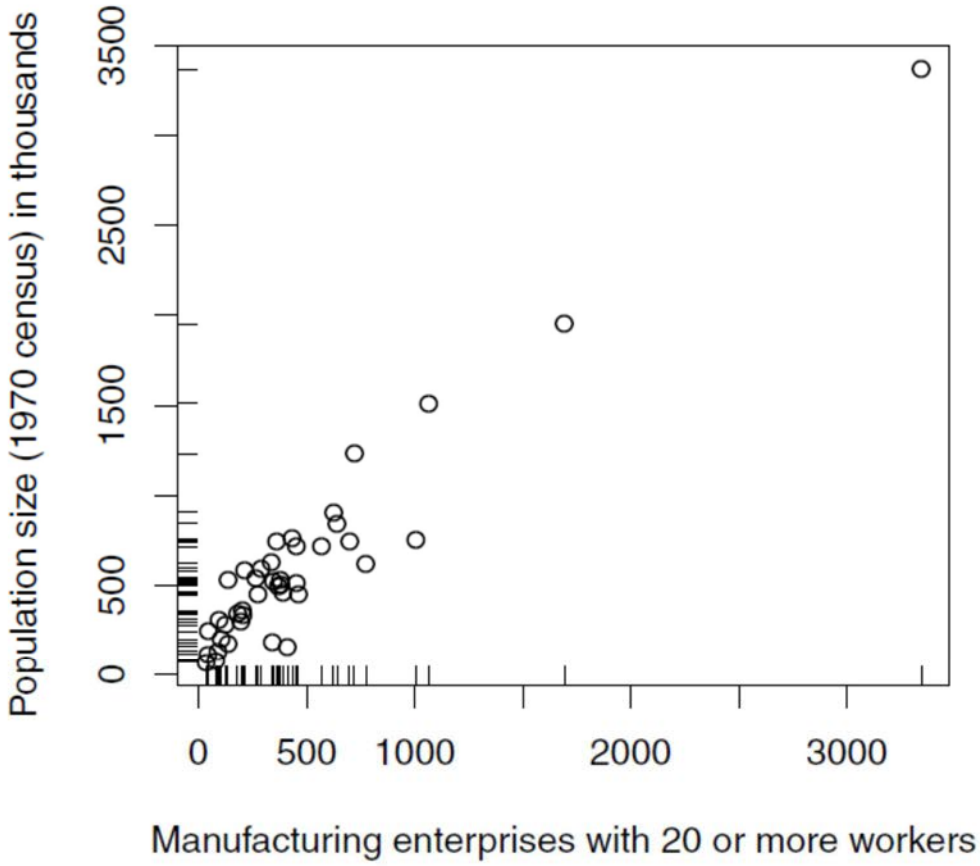


Fig. 2.2. Scatterplot of `manu` and `popul` that shows the marginal distribution in each variable as a rug plot.

The bivariate boxplot

In Figure 2.2, identifying Chicago, Philadelphia, and Detroit as outliers is unlikely to invoke much argument, but what about Houston and Cleveland? In many cases, it might be helpful to have a more formal and objective method for labeling observations as outliers, and such a method is provided by the bivariate boxplot, which is a two-dimensional analogue of the boxplot for univariate data. The bivariate boxplot is based on calculating "robust" measures of location, scale, and correlation; it consists essentially of a pair of concentric ellipses, one of which (the "hinge") includes 50% of the data and the other (called the "fence") of which delineates potentially troublesome outliers. In addition, resistant regression lines of both y on x and x on y are shown, with their intersection showing the bivariate location estimator. The acute angle between the regression lines will be small for a large absolute value of correlations and large for a small one. (Using robust measures of location, scale, etc., help to prevent the possible "masking" of multivariate outliers if the usual measures are employed when these may be distorted by the presence of the outliers in the data.) The scatterplot of manu and popul including the bivariate boxplot is shown in Figure 2.4. Figure 2.4 clearly tells us that Chicago,

Philadelphia, Detroit, and Cleveland should be regarded as outliers but not Houston, because it is on the “fence” rather than outside the “fence”.

```
require(grid) require(lattice) require(scatterplot3d) require(HSAUR2) require(MVA)
outcity<- match(lab<- c("Chicago",  "Detroit", "Cleveland", "Philadelphia"),
rownames(USairpollution))
x <- USairpollution[, c("manu", "popul")]
bvbox(x, mtitle = "", xlab = mlab, ylab = plab)
text(x[,c("manu")][outcity], x[,c("popul")][outcity], labels = lab,cex = 0.7)
```

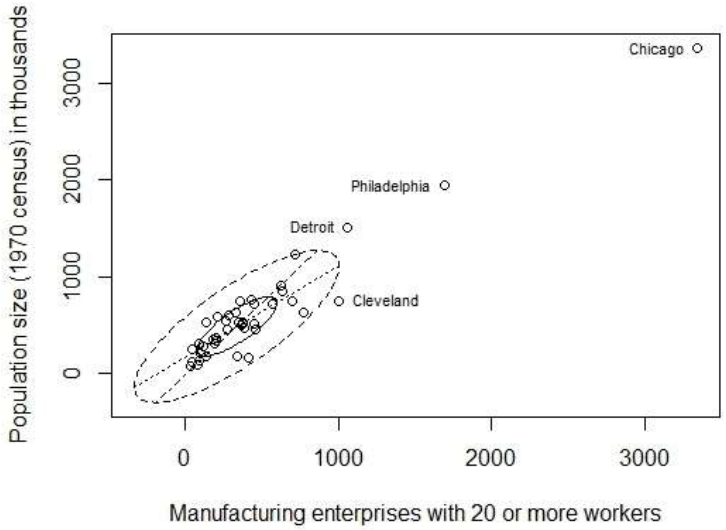


Fig. 2.4. Scatterplot of manu and popul showing the bivariate boxplot of the data.

Suppose now that we are interested in calculating the correlation between manu and popul. The observations identified as outliers may then be excluded from the calculation of the correlation coefficient. With the help of the bivariate boxplot in Figure 2.4, we have identified Chicago, Philadelphia, Detroit, and Cleveland as outliers in the scatterplot of manu and popul. The R code for finding the two correlations is

```
manu<-x[,c("manu")]
popul<-x[,c("popul")]
cor(manu, popul)
[1] 0.9553
outcity<-match(c("Chicago","Detroit","Cleveland","Philadelphia"),
rownames(USairpollution))
cor(manu[-outcity], popul[-outcity])
[1] 0.7956
```

The match() function identifies rows of the data frame USairpollution corresponding to the cities of interest, and the subset starting with a minus sign removes these units before the correlation is computed. Calculation of the correlation coefficient between the two variables using all the data gives a value of 0.96, which reduces to a value of 0.8 after excluding the four outliers.

The convex hull of bivariate data

An alternative approach to using the scatterplot combined with the bivariate boxplot to deal with the possible problem of calculating correlation coefficients without the distortion often caused by outliers in the data is convex hull trimming, which allows robust estimation of the correlation. The convex hull of a set of bivariate observations consists of the vertices of the smallest convex polyhedron in variable space within which or on which all data points lie. Removal of the points lying on the convex hull can eliminate isolated outliers without disturbing the general shape of the bivariate distribution. A robust estimate of the correlation coefficient results from using the remaining observations.
We first find the convex hull of the USairpollution data (i.e., the observations defining the convex hull) using the following R code:

```
hull <- chull(manu, popul)

[1] 9 15 41 6 2 18 16 14 7

plot(manu, popul, pch = 1, xlab = mlab, ylab = plab)
polygon(manu[hull], popul[hull], density = 15, angle = 30)
```

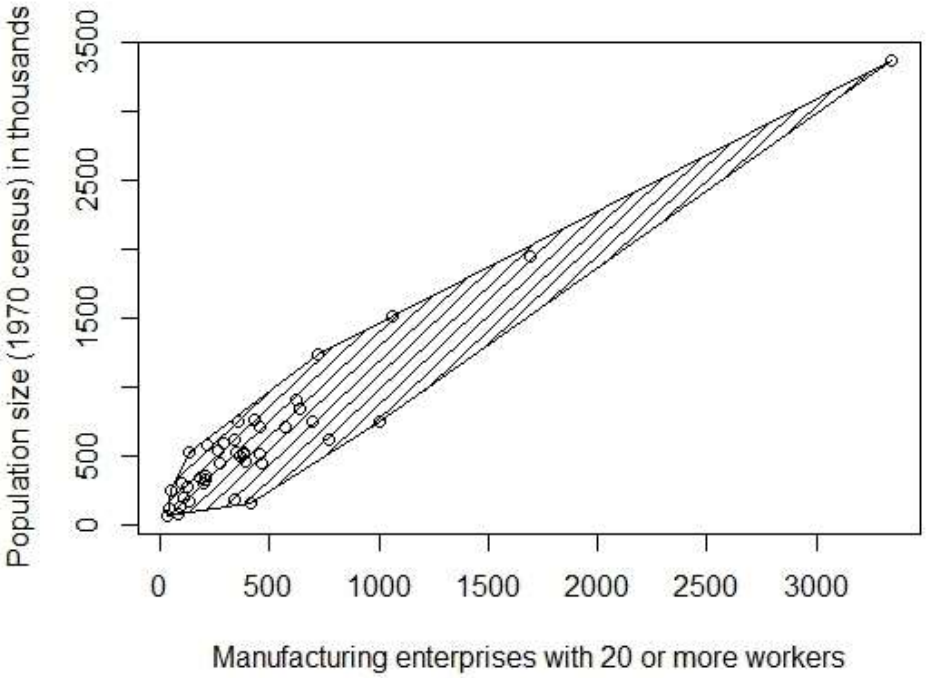


Fig. 2.5. Scatterplot of manu against popul showing the convex hull of the data.

To calculate the correlation coefficient after removal of the points defining the convex hull requires the code

```
cor(manu[-hull],popul[-hull])
[1] 0.9225
```

The resulting value of the correlation is now 0.923 and thus is higher compared with the correlation estimated after removal of the outliers identified by using the bivariate boxplot, namely Chicago, Philadelphia, Detroit, and Cleveland.

The chi-plot

By scatterplot it is often difficult to judge whether or not the variables are independent. Consequently it is sometimes helpful to augment the scatterplot with an auxiliary display in which independence is itself manifested in a characteristic manner. The chi-plot[1] is designed to address the problem.

> Following Fisher and Switzer (1985), if y_i is a strictly increasing function of x_i, we have $\chi_i = 1$ and if y_i is a strictly decreasing function of x_i, we have $\chi_i = -1$. On the other hand, if the random variables are independent, when $n \to \infty$ the asymptotic distribution of λ_i is uniformly distributed in $\pm 4((1/(n-1)) - 0.5)$. When there is independence between x_i and y_i, χ_i is randomly distributed around zero. If y_i is increasing (decreasing) compared to x_i, we have $\lambda_i > 0$ (< 0). If Y is positively (negatively) associated with X, i.e. $Cov(Y, X) > 0$ (< 0), there is a tendency that most values of λ are larger (smaller) than zero. The χ_i is the correlation coefficient ϕ for dichotomous variables, which reduces the interpretation of χ_i to the locally Pearson correlation coefficient.

[1] Under independence, the joint distribution of two random variables ***X*** and ***Y*** can be computed from the product of the marginal distributions. The chi-plot transforms the measurements $(x_1, \dots, x_{n1})$ and $(y_1, \dots, y_n)$ into values $(\chi_1, \dots, \chi_n)$ and $(\lambda_1, \dots, \lambda_n)$, which, plotted in a scatterplot, can be used to detect deviations from independence. Under independence, these values are asymptotically normal with mean zero; i.e., the χ_i values should show a non-systematic random actuation around zero. The λ_i values measure the distance of unit i from the "center" of the bivariate distribution.

Let $(x_1, y_1), \cdots, (x_n, y_n)$ be a random sample of (X, Y) and $I(A)$ the function indicator of event A. For each data point (x_i, y_i) we have

$$\chi_i = \frac{H_i - F_i G_i}{(F_i(1-F_i)G_i(1-G_i))^{1/2}},$$

$$F_i = \frac{1}{n-1}\sum_{j \neq i} I(x_j \leq x_i),$$

$$H_i = \frac{1}{n-1}\sum_{j \neq i} I(x_j \leq x_i, y_j \leq y_i),$$

$$G_i = \frac{1}{n-1}\sum_{j \neq i} I(y_j \leq y_i),$$

$$\lambda_i = 4 * \text{sign}_i * \max\{(F_i - 0.5)^2, (G_i - 0.5)^2\},$$

and

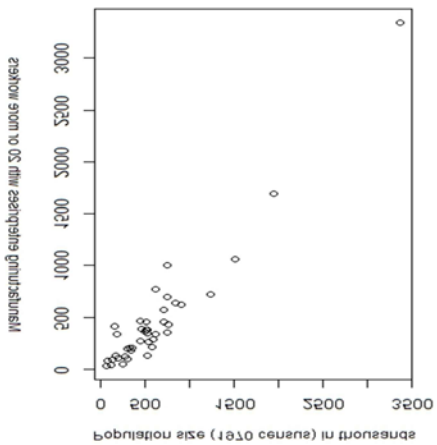
$$\text{sign}_i = signal\{(F_i - 0.5)(G_i - 0.5)\}.$$

The chi-plot is the scatter-plot of (λ_i, χ_i), for all $|\lambda_i| < 4((1/(n-1)) - 0.5)^2$,

An R function for producing chi-plots is chiplot(). To illustrate the chi-plot, we shall apply it to the manu and popul variables of the air pollution data using the code

```
plot(manu, popul, xlab = mlab, ylab = plab, cex.lab = 0.9)
chiplot(manu, popul)
```

The result is Figure 2.6, which shows the scatterplot of manu plotted against popul alongside the corresponding chi-plot. Departure from independence is indicated in the latter by a lack of points in the horizontal band indicated on the plot. Here there is a very clear departure since there are very few of the observations in this region.



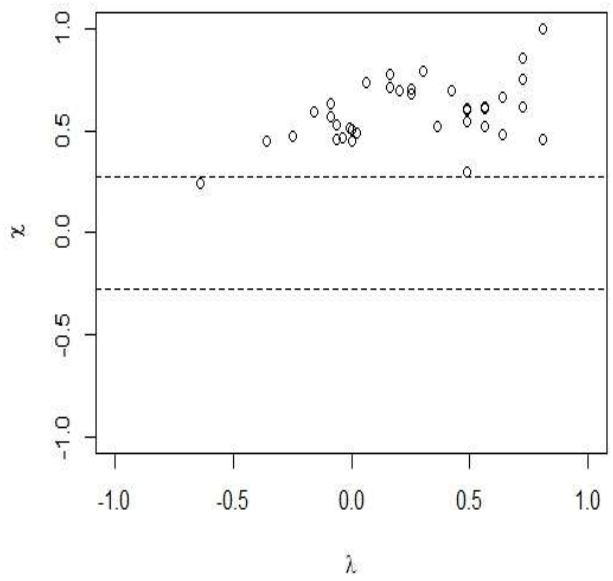


Fig. 2.6. Chi-plot for manu and popul showing a clear deviation from independence.

The scatterplot matrix

There are seven variables in the air pollution data, which between them generate 21 possible scatterplots. To aid overall comprehension of the data, the scatterplot matrix is used. A scatterplot matrix is nothing more than a square, symmetric grid of bivariate scatterplots. The grid has q rows and columns, each one corresponding to a different variable. Each of the grid's cells shows a scatterplot of two variables. Variable j is plotted against variable i in the ijth cell, and the same variables appear in cell ji, with the x- and y-axes of the scatterplots interchanged. The reason for including both the upper and lower triangles of the grid, is that it enables a row and a column to be visually scanned to see one variable against all others, with the scales for the one variable lined up along the horizontal or the vertical. As a result, we can visually link features on one scatterplot with features on another, and this ability greatly increases the power of the graphic.
The scatterplot matrix for the air pollution data is shown in Figure 2.10. The plot was produced using the function pairs(), here with slightly enlarged dot symbols, using the arguments pch = "." and cex = 1.5.

```
pairs(USairpollution, pch = ".", cex = 1.5)
```

The scatterplot matrix clearly shows the presence of possible outliers in many panels and the suggestion that the relationship between the two aspects of rainfall, namely precip, predays, and SO2 might be non-linear. In Figure 2.11, the pairs() function was customized by a small

function specified to the panel argument: in addition to plotting the x and y values, a regression line obtained via function lm() is added to each of the panels.

```
pairs(USairpollution, panel = function (x, y, ...) { points(x, y, ...); abline(lm(y ~
x), col = "grey") }, pch = ".", cex = 1.5)
```

Now the scatterplot matrix reveals that there is a strong linear relationship between SO2 and manu and between SO2 and popul, but the (3, 4) panel shows that manu and popul are themselves very highly related and thus predictive of SO2 in the same way. Figure 2.11 also underlines that assuming a linear relationship between SO2 and precip and SO2 and predays, as might be the case if a multiple linear regression model is fitted to the data with SO2 as the dependent variable, is unlikely to fully capture the relationship between each pair of variables.

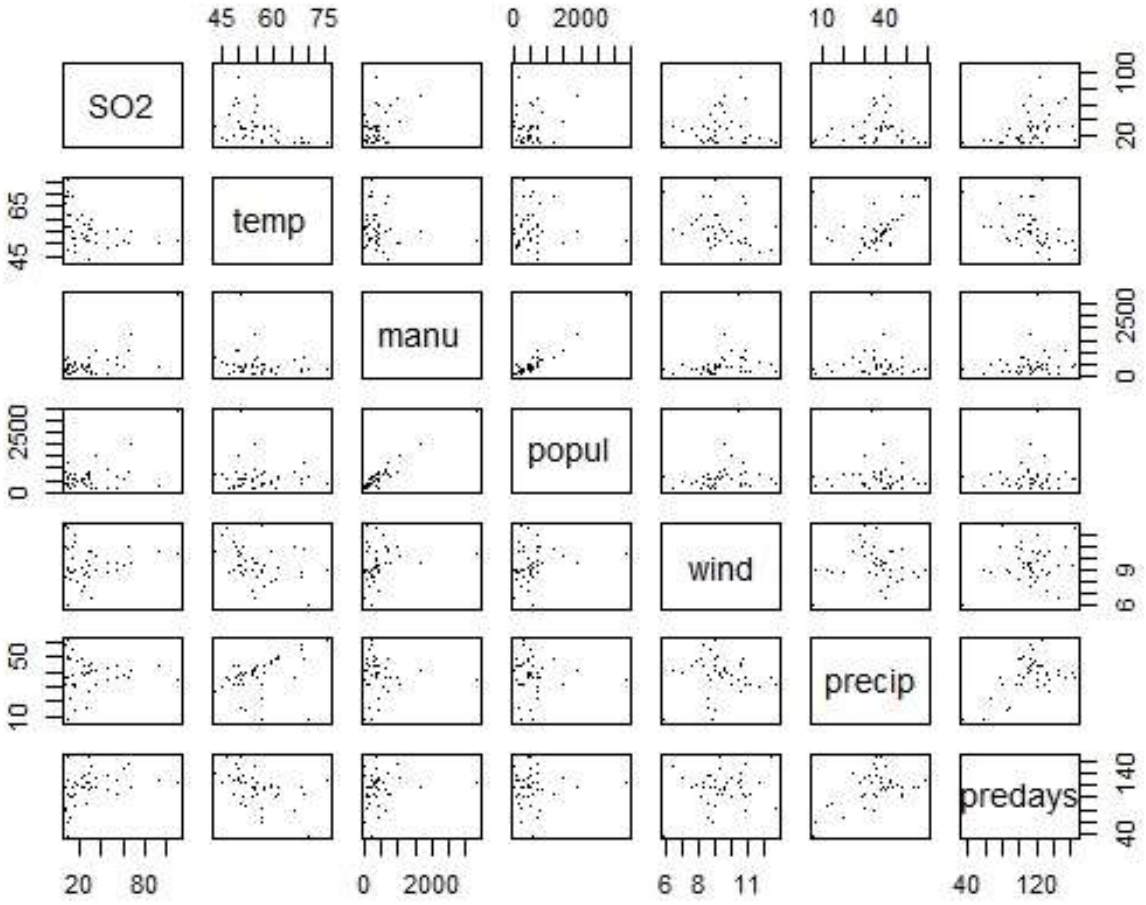


Fig. 2.10. Scatterplot matrix of the air pollution data.

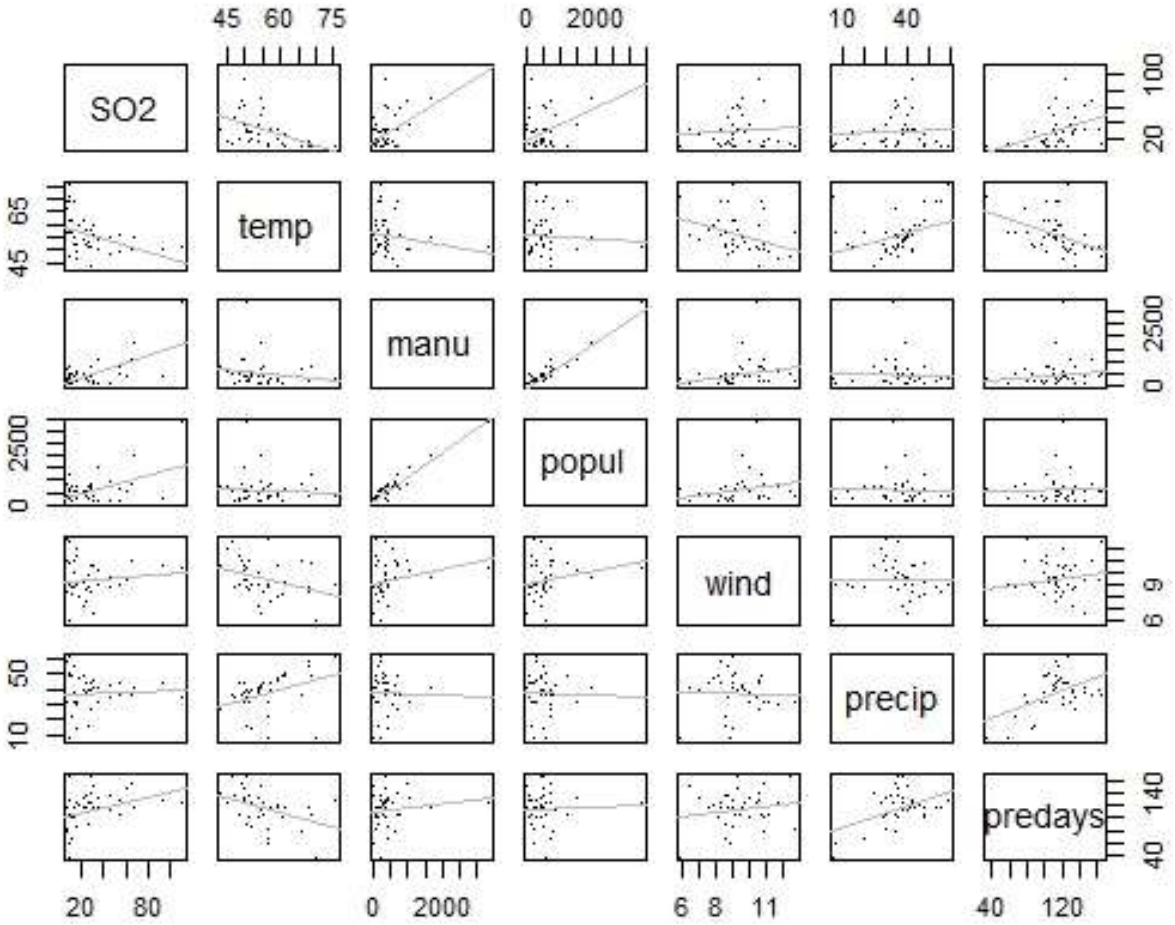


Fig. 2.11. Scatterplot matrix of the air pollution data showing the linear fit of each pair of variables.

3. Enhancing the scatterplot with estimated bivariate densities

As we have seen above, scatterplots and scatterplot matrices are good at highlighting outliers in a multivariate data set. But in many situations another aim in examining scatterplots is to identify regions in the plot where there are high or low densities of observations that may indicate the presence of distinct groups of observations; i.e., "clusters". It is often very helpful to add some type of bivariate density estimate to the scatterplot. A bivariate density estimate is simply an *approximation* to the bivariate probability density function of two variables obtained from a sample of bivariate observations of the variables. If, of course, we are willing to assume a particular form of the bivariate density of the two variables, for example the bivariate normal, then estimating the density is reduced to estimating the parameters of the assumed distribution. More commonly, however, we wish to allow the data to speak for themselves and so we need to look for a non-parametric estimation procedure. The simplest such estimator would be a two-dimensional histogram, but for small and moderately sized data sets that is not of any real use for estimating the bivariate density function simply because most of the "boxes" in the histogram will contain too few observations; and if the number of boxes is reduced, the resulting histogram will be too coarse a representation of the density function. Other non-parametric density estimators attempt to overcome the deficiencies of the simple two-dimensional histogram estimates by "smoothing" them in one way or another. A variety of non-parametric estimation procedures are proposed that here we give a brief description of just one popular class of estimators, namely kernel density estimators.

```
library(squash)
x <- rnorm(100)
y <- rnorm(100)
hist2(x, y)
```

Kernel density estimators

One Dimension

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

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

For any given h, a naive estimator of $\mathrm{P}(x - h < X < x + h)$ is the proportion of the observations $x_1, \ldots, x_n$ falling in the interval $(x - h, x + h)$,

$$\hat{f}(x) = \frac{1}{2nh} \sum_{i=1}^{n} I(x_i \in (x - h, x + h));$$

i.e., the number of $x_1, \ldots, x_n$ falling in the interval $(x - h, x + h)$ divided by $2nh$. If we introduce a weight function W given by

$W(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & o.w. \end{cases}$,

Then the naive estimator can be written as

$$\hat{f}(x) = \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{h} W\left(\frac{x - x_i}{h}\right).$$

Unfortunately, this estimator is not a continues function and is not particularly satisfactory for practical density estimation. It does, however, lead naturally to the kernel estimator defined by

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

where K is known as the kernel function and h is the bandwidth or smoothing parameter.
In Statistics, a kernel is a non-negative real-valued integrable function satisfying
$\int_{-\infty}^{\infty} K(x)dx = 1$.
Usually, but not always, the kernel function will be a symmetric density function;
for example, the normal.
In order to compute the MSE of the estimate of $f(.)$, we need the bias and variance of $\hat{f}(.)$. Let X be a random variable having density $f(.)$. Then we have for the specific point $x \in \mathbb{R}$ we have

$$E\big(\hat{f}(x)\big) = E[K_h(x - X)] = \int K_h(x - y) f(y) dy = \int K(z) f(x - hz) dz.$$

Expanding $f(x - hz)$ in a Taylor series about x we obtain

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

uniformly in z. This leads to

$$E\big(\hat{f}(x)\big) = f(x) + \frac{1}{2} h^2 f''(x) \int z^2 K(z) dz + o(h^2)$$

where we have used

$$\int K(z)dz = 1, \qquad \int zK(z)dz = 0, \qquad \int z^2 K(z)dz = \sigma_K^2 < \infty.$$

Three commonly used kernel functions are

1. rectangular,
$$K(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & o.w. \end{cases} = \frac{1}{2} 1_{\{|x|<1\}},$$
2. triangular,
$$K(x) = \begin{cases} 1 - |x| & |x| < 1 \\ 0 & o.w. \end{cases} = (1 - |x|) 1_{\{|x|<1\}},$$
3. Gaussian,
$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

The three kernel functions are implemented in R as shown in Figure 2.12. For some grid x, the kernel functions are plotted using the R statements in Figure 2.12. The kernel estimator $\hat{f}$ is a

sum of “bumps” placed at the observations. The kernel function determines the shape of the bumps, while the window width h determines their width. Figure 2.13 shows the individual bumps $n^{-1}h^{-1}K\left(\frac{x-x_i}{h}\right)$ as well as the estimate $\hat{f}$ obtained by adding them up for an artificial set of data points,

```
rec <- function(x) (abs(x) < 1) * 0.5
tri <- function(x) (abs(x) < 1) * (1 - abs(x))
gauss <- function(x) 1/sqrt(2*pi) * exp(-(x^2)/2)
x <- seq(from = -3, to = 3, by = 0.001)
plot(x, rec(x), type = "l", ylim = c(0,1), lty = 1, ylab = expression(K(x)))
lines(x, tri(x), lty = 2)
lines(x, gauss(x), lty = 3)
legend("topleft", legend = c("Rectangular", "Triangular", "Gaussian"), lty = 1:3,
title = "kernel functions", bty = "n")
```

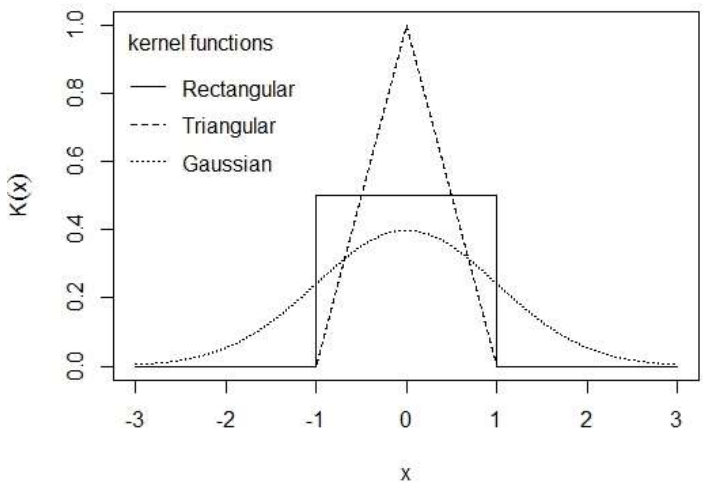


Fig. 2.12. Three commonly used kernel functions.

```
x <- c(0, 1, 1.1, 1.5, 1.9, 2.8, 2.9, 3.5)
n <- length(x)
```

For a grid

```
xgrid <- seq(from = min(x) - 1, to = max(x) + 1, by = 0.01)
```

on the real line, we can compute the contribution of each measurement in x, with $h = 0.4$, by the Gaussian kernel (defined in Figure 2.12, line 3) as follows:

```
h <- 0.4
bumps <- sapply(x, function(a) gauss((xgrid - a)/h)/(n * h))
```

A plot of the individual bumps and their sum, the kernel density estimate $\hat{f}$, is shown in Figure 2.13.

```
plot(xgrid, rowSums(bumps), ylab = expression(hat(f)(x)), type = "l", xlab = "x",
lwd = 2)
rug(x, lwd = 2)
out <- apply(bumps, 2, function(b) lines(xgrid, b))
```

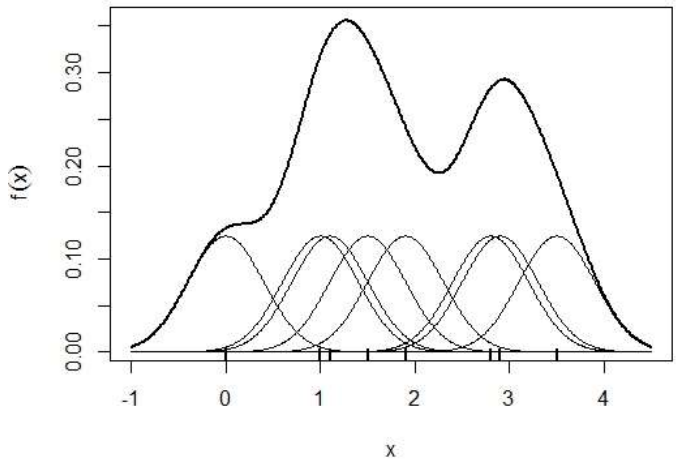


Fig. 2.13. Kernel estimate showing the contributions of Gaussian kernels evaluated for the individual observations with bandwidth h = 0:4.

Some other common kernels:

4. Epanechnikov

$K(x) = \frac{3}{4}(1 - x^2)1_{\{|x|<1\}}$,

5. Quartic (biweight)

$K(x) = \frac{15}{16}(1 - x^2)^2 1_{\{|x|<1\}}$,

6. Triweight

$K(x) = \frac{35}{32}(1 - x^2)^3 1_{\{|x|<1\}}$,

7. Tricube

$K(x) = \frac{70}{81}(1 - |x|^3)^3 1_{\{|x|<1\}}$,

8. Cosine

$K(x) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right) 1_{\{|x|<1\}}$.

Two Dimension

The kernel density estimator considered as a sum of "bumps" centered at the observations has a simple extension to two dimensions (and similarly for more than two dimensions). The bivariate estimator for data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is defined as

$$\hat{f}(x,y) = \frac{1}{nh_xh_y}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}, \frac{y-y_i}{h_y}\right).$$

In this estimator, each coordinate direction has its own smoothing parameter, h_x or h_y. An alternative is to scale the data equally for both dimensions and use a single smoothing parameter.

For bivariate density estimation, a commonly used kernel function is the standard bivariate normal density

$$K(x,y) = \frac{1}{2\pi}\, e^{-\frac{1}{2}(x^2+y^2)}.$$

Another possibility is the bivariate Epanechnikov kernel given by

$$K(x,y) = \begin{cases} \frac{2}{\pi}\,(1-x^2-y^2) & x^2+y^2 < 1 \\ 0 & o.w. \end{cases},$$

which is implemented and depicted in Figure 2.14 by using the persp function for plotting in three dimensions.

```
epa <- function(x, y) ((x^2 + y^2) < 1) * 2/pi * (1 - x^2 - y^2)
x <- seq(from = -1.1, to = 1.1, by = 0.05)
 epavals <- sapply(x, function(a) epa(a, x))
persp(x = x, y = x, z = epavals, xlab = "x", ylab = "y", zlab = expression(K(x, y)), theta = -35, axes
= TRUE, box = TRUE)
```

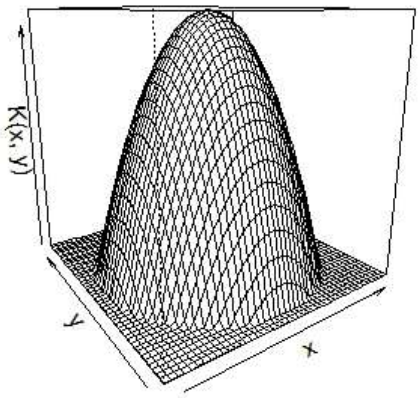


Fig. 2.14. Epanechnikov kernel for a grid between $(-1.1,-1.1)$ and (1.1,1.1).

Our first illustration of enhancing a scatterplot with an estimated bivariate density will involve data from the Hertzsprung-Russell (H-R) diagram of the star cluster CYG OB1, calibrated according to Vanisma and De Greve (1972). The H-R diagram is the basis of the theory of stellar evolution and is essentially a plot of the energy output of stars as measured by the logarithm of

their light intensity plotted against the logarithm of their surface temperature. Part of the data is shown in Table 2.1.

Table 2.1: CYGOB1 data. Energy output and surface temperature of star cluster CYG OB1.

Table 2.1: CYGOB1 data (continued).

logst	logli	logst	logli	logst	logli
4.56	5.74	4.42	4.18	3.49	6.29
4.26	4.93	4.23	4.18	4.23	4.34
4.56	5.74	3.49	5.89	4.62	5.62
4.30	5.19	4.29	4.38	4.53	5.10
4.46	5.46	4.29	4.22	4.45	5.22
3.84	4.65	4.42	4.42	4.53	5.18
4.57	5.27	4.49	4.85	4.43	5.57
4.26	5.57	4.38	5.02	4.38	4.62
4.37	5.12	4.42	4.66	4.45	5.06
3.49	5.73	4.29	4.66	4.50	5.34
4.43	5.45	4.38	4.90	4.45	5.34
4.48	5.42	4.22	4.39	4.55	5.54
4.01	4.05	3.48	6.05	4.45	4.98
4.29	4.26	4.38	4.42	4.42	4.50
4.42	4.58	4.56	5.10		

logst	logli	logst	logli	logst	logli
4.37	5.23	4.23	3.94	4.45	5.22

```
logst<-c(4.37, 4.56, 4.26, 4.56, 4.30, 4.46, 3.84, 4.57, 4.26, 4.37, 3.49, 4.43, 4.48, 4.01, 4.29,
4.42, 4.23, 4.42, 4.23, 3.49, 4.29, 4.29, 4.42, 4.49, 4.38, 4.42, 4.29, 4.38, 4.22, 3.48, 4.38,
4.56, 4.45, 3.49, 4.23, 4.62, 4.53, 4.45, 4.53, 4.43, 4.38, 4.45, 4.50, 4.45, 4.55, 4.45, 4.42)
logli<-c(5.23, 5.74, 4.93, 5.74, 5.19, 5.46, 4.65, 5.27, 5.57, 5.12, 5.73, 5.45, 5.42, 4.05, 4.26,
4.58, 3.94, 4.18, 4.18, 5.89, 4.38, 4.22, 4.42, 4.85, 5.02, 4.66, 4.66, 4.90, 4.39, 6.05, 4.42,
5.10, 5.22, 6.29, 4.34, 5.62, 5.10, 5.22, 5.18, 5.57, 4.62, 5.06, 5.34, 5.34, 5.54, 4.98, 4.50)
CYGOB1<-data.frame(logst,logli)
```

A scatterplot of the data enhanced by the contours of the estimated bivariate density, obtained with the function bkde2D()[2] from the package KernSmooth, is shown in Figure 2.15.

```
library("KernSmooth")
CYGOB1d <- bkde2D(CYGOB1, bandwidth = sapply(CYGOB1, dpik))
plot(CYGOB1, xlab = "log surface temperature", ylab = "log light intensity")
contour(x = CYGOB1d$x1, y = CYGOB1d$x2, z = CYGOB1d$fhat, add = TRUE)
```

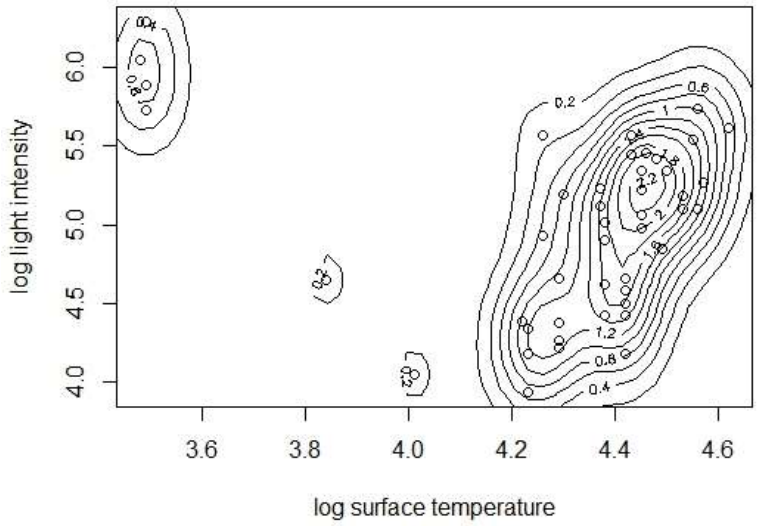


Fig. 2.15. Scatterplot of the log of light intensity and log of surface temperature for the stars in star cluster CYG OB1 showing the estimated bivariate density.

The plot shows the presence of two distinct clusters of stars: the larger cluster consists of stars that have high surface temperatures and a range of light intensities, and the smaller cluster contains stars with low surface temperatures and high light intensities. The bivariate density estimate can also be displayed by means of a perspective plot rather than a contour plot, and this is shown in Figure 2.16.

```
persp(x = CYGOB1d$x1, y = CYGOB1d$x2, z = CYGOB1d$fhat,
 xlab = "log surface temperature",
 ylab = "log light intensity",
 zlab = "density")
```

[2] The kernel is the standard bivariate normal density

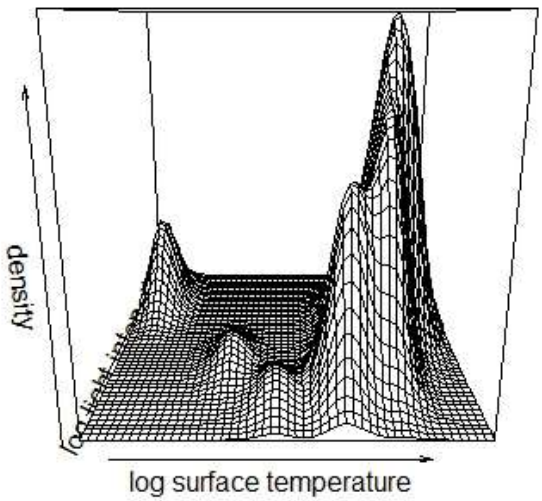


Fig. 2.16. Perspective plot of estimated bivariate density.

This again demonstrates that there are two groups of stars.

For our next example of adding estimated bivariate densities to scatterplots, we will use the body measurement data introduced in Chapter 1 (see Table 1.2), although there are rather too few observations on which to base the estimation. (The gender of each individual will not be used.) And in this case we will add the appropriate density estimate to each panel of the scatterplot matrix of the chest, waist, and hips measurements. The resulting plot is shown in Figure 2.17.

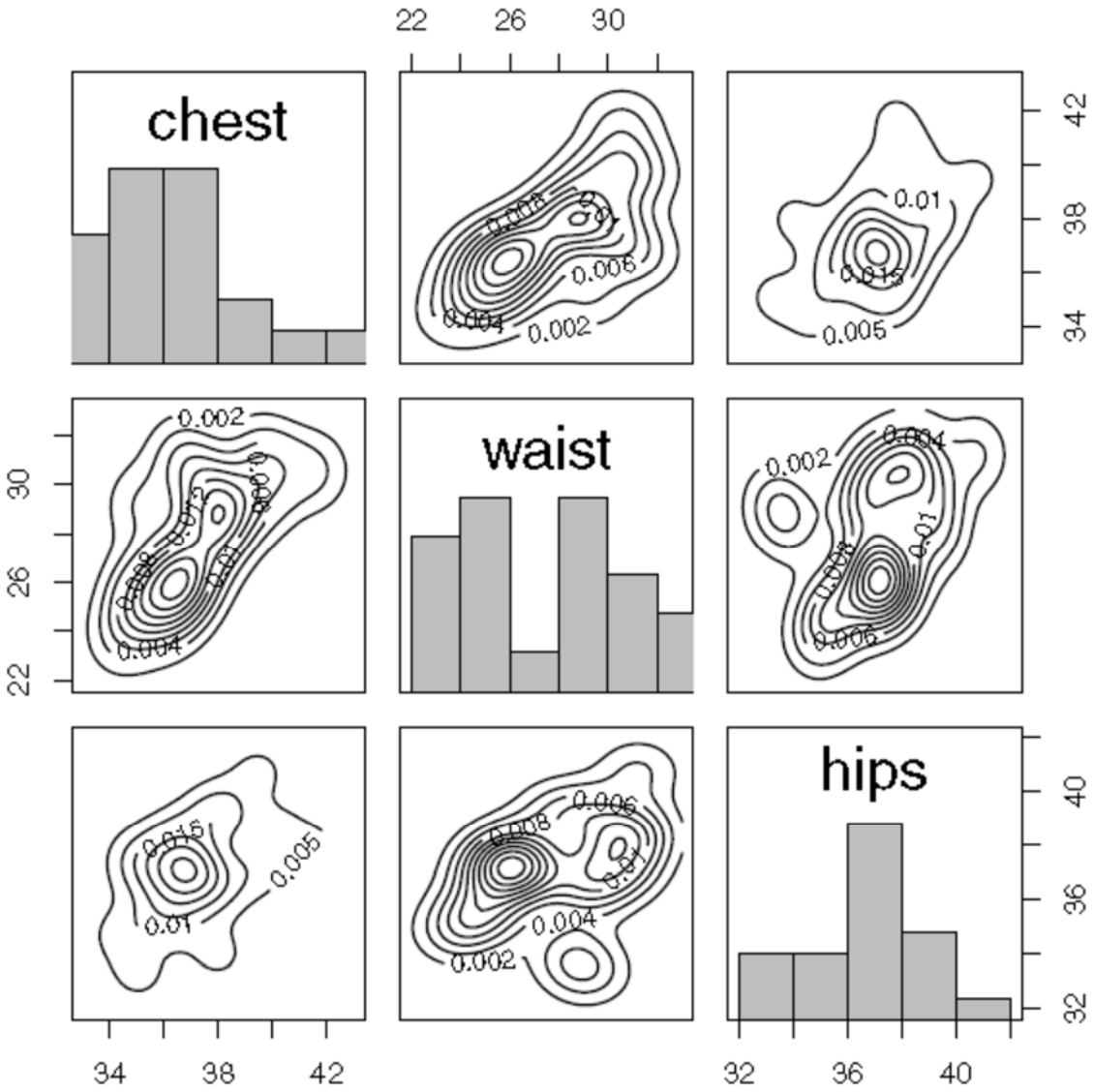


Fig. 2.17. Scatterplot matrix of body measurements data showing the estimated bivariate densities on each panel.

The waist/hips panel gives some evidence that there might be two groups in the data, which, of course, we know to be true, the groups being men and women. And the Waist histogram on the diagonal panel is also bimodal, underlining the two-group nature of the data.

Multi Dimension

Kernel density estimation can be easily generalized from univariate to multivariate data, in theory if not always in practice. The general form of the estimator is

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n \det \boldsymbol{H}} \sum_{i=1}^{n} K_q[\boldsymbol{H}^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)],$$

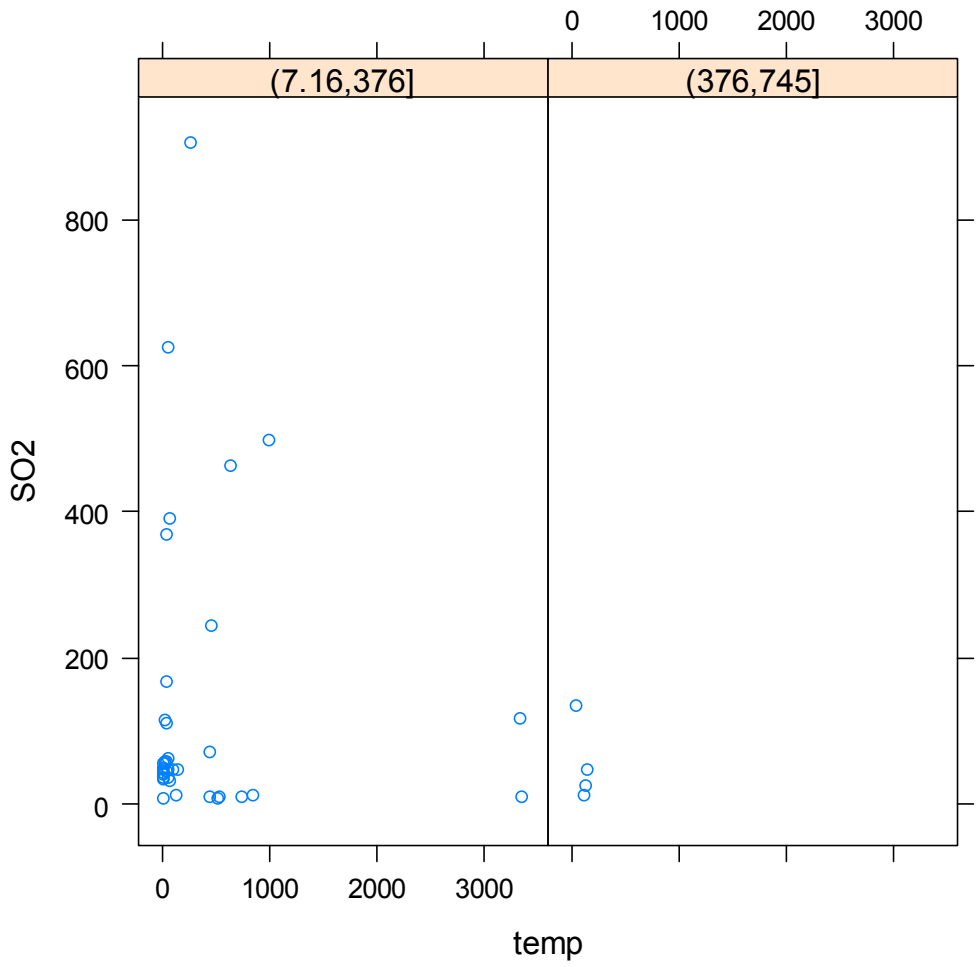
where $\boldsymbol{x} = (x_1, \ldots, x_q)'$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})'$, $i = 1, \ldots, n$ are q-vectors; $\boldsymbol{H}$ is the bandwidth (or smoothing) $q \times q$ positive definite matrix and $K_q: \mathbb{R}^q \to \mathbb{R}$ is the kernel function.
A popular technique for generating K_q from a univariate kernel K is by using a product kernel,

$$K_q(\boldsymbol{u}) = \prod_{j=1}^{q} K(u_j).$$

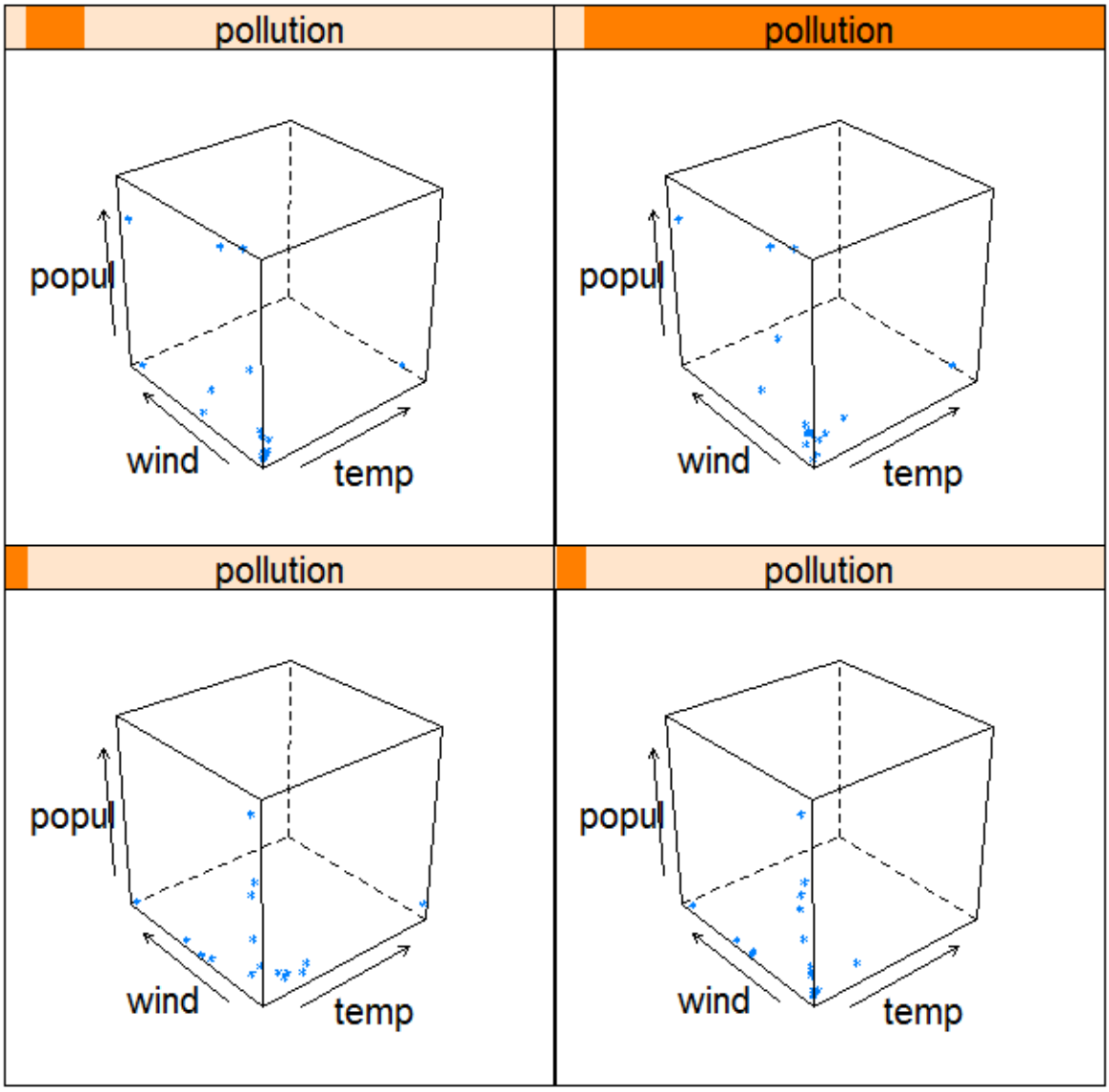
4. Stalactite plots

In this section, we will describe a multivariate graphic, the stalactite plot, specially designed for the detection and identication of multivariate outliers. Like the chi-square plot for assessing multivariate normality, described in Chapter 1, the stalactite plot is based on the generalised distances of observations from the multivariate mean of the data. But here these distances are calculated from the means and covariances estimated from increasing sized subsets of the data. As mentioned previously when describing bivariate boxplots, the aim is to reduce the masking effects that can arise due to the innocence of outliers on the estimates of means and covariances obtained from all the data. The central idea of this approach is that, given distances using, say, m observations for estimation of means and covariances, the m+1 observations to be used for this estimation in the next stage are chosen to be those with the m+1 smallest distances. Thus an observation can be included in the subset used for estimation for some value of m but can later be excluded as m increases. Initially m is chosen to take the value q +1, where q is the number of variables in the multivariate data set because this is the smallest number allowing the calculation of the required generalised distances. The cutoff distance generally employed to identify an outlier is the maximum expected value from a sample of n random variables each having a chi-squared distribution on q degrees of freedom.

plot(xyplot(SO2 ~ temp| cut(wind, 2)))




```
pollution <- equal.count(SO2,4)
plot(cloud(popul ~ temp * wind | pollution, panel.aspect = 0.9))
```



The stalactite plot, is specifically designed for the detection and identification of multivariate outliers. This plot is based on the generalized distances of observations from the multivariate mean of the data. But here these distances are calculated from the means and covariances estimated from increasing sized subsets of the data. The stalactite plot graphically illustrates the evolution of the outliers as the size of the subset of observations used for estimation increases. We will now illustrate the application of the stalactite plot on the US cities air pollution data. The plot is shown in Figure 2.25.

```
require(MVA)
stalac (USairpollution)
```

Initially most cities are indicated as outliers (a "*" in the plot), but as the number of observations on which the generalized distances are calculated is increased, the number of outliers indicated by the plot decreases. The plot clearly shows the outlying nature of a number of cities over nearly all values of m. The effect of masking is also clear; when all 41 observations are used to calculate the generalized distances, only observations Chicago, Phoenix, and Providence are indicated to be outliers.

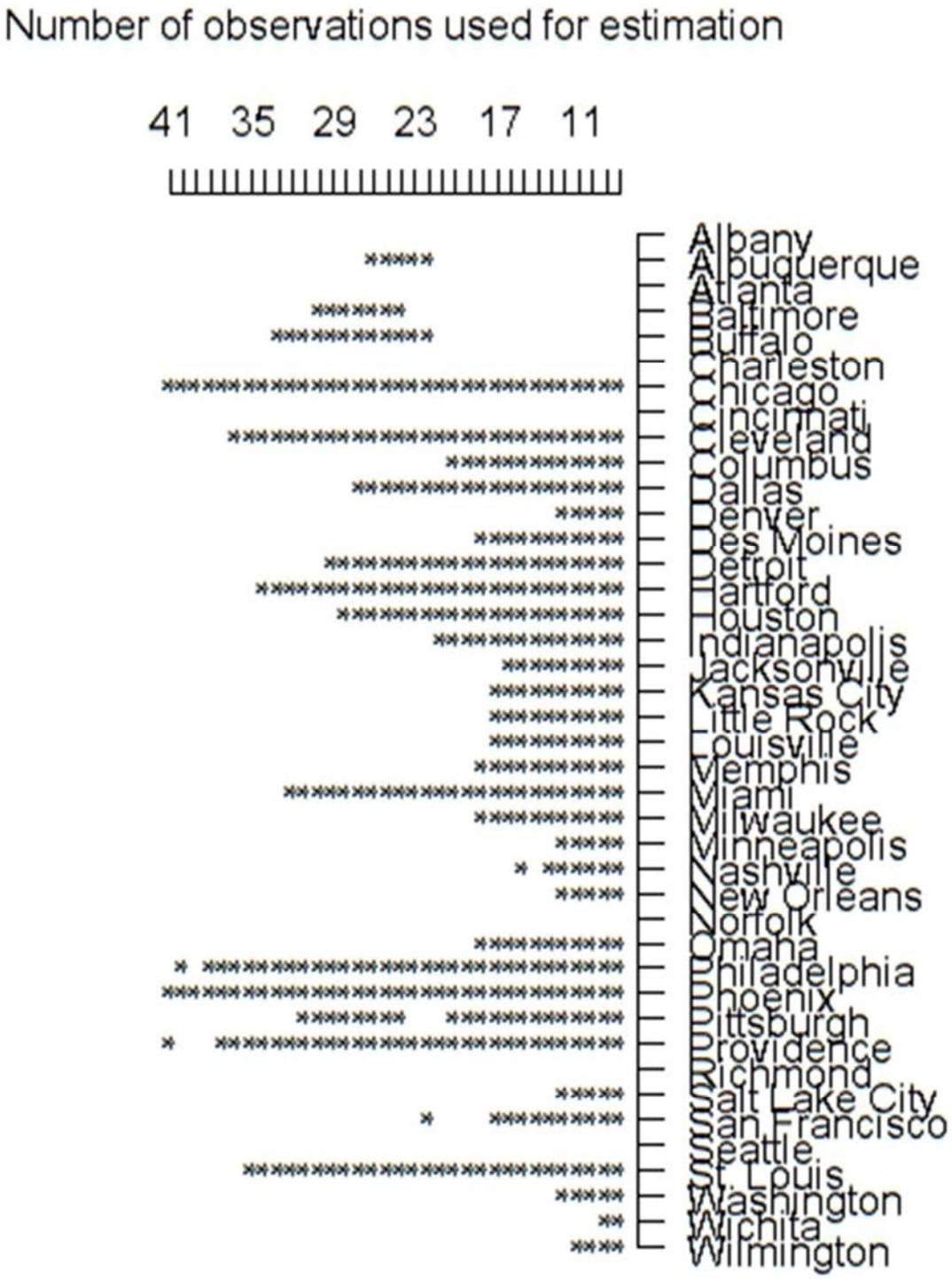


Fig. 2.25. Stalactite plot of US cities air pollution data.