Chapter 1
Comparison of Batches

Multivariate statistical analysis is concerned with analysing and understanding data in high dimensions. We suppose that we are given a set $\{x_i\}_{i=1}^n$ of n observations of a variable vector X in $\mathbb{R}^p$. That is, we suppose that each observation x_i has p dimensions:

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}),$$

and that it is an observed value of a variable vector $X \in \mathbb{R}^p$. Therefore, X is composed of p random variables:

$$X = (X_1, X_2, \ldots, X_p)$$

where X_j, for $j = 1, \ldots, p$, is a one-dimensional random variable. How do we begin to analyse this kind of data? Before we investigate questions on what inferences we can reach from the data, we should think about how to look at the data. This involves descriptive techniques. Questions that we could answer by descriptive techniques are:

- Are there components of X that are more spread out than others?
- Are there some elements of X that indicate sub-groups of the data?
- Are there outliers in the components of X?
- How "normal" is the distribution of the data?
- Are there "low-dimensional" linear combinations of X that show "non-normal" behaviour?

One difficulty of descriptive methods for high-dimensional data is the human perceptional system. Point clouds in two dimensions are easy to understand and to interpret. With modern interactive computing techniques we have the possibility to see real time 3D rotations and thus to perceive also three-dimensional data. A "sliding technique" as described in Härdle and Scott (1992) may give insight

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-662-45171-7_1

into four-dimensional structures by presenting dynamic 3D density contours as the fourth variable is changed over its range.

A qualitative jump in presentation difficulties occurs for dimensions greater than or equal to 5, unless the high-dimensional structure can be mapped into lower-dimensional components (Klinke & Polzehl, 1995). Features like clustered sub-groups or outliers, however, can be detected using a purely graphical analysis.

In this chapter, we investigate the basic descriptive and graphical techniques allowing simple exploratory data analysis. We begin the exploration of a data set using boxplots. A boxplot is a simple univariate device that detects outliers component by component and that can compare distributions of the data among different groups. Next, several multivariate techniques are introduced (Flury faces, Andrews' curves and parallel coordinates plots (PCPs)) which provide graphical displays addressing the questions formulated above. The advantages and the disadvantages of each of these techniques are stressed.

Two basic techniques for estimating densities are also presented: histograms and kernel densities. A density estimate gives a quick insight into the shape of the distribution of the data. We show that kernel density estimates (KDEs) overcome some of the drawbacks of the histograms.

Finally, scatterplots are shown to be very useful for plotting bivariate or trivariate variables against each other: they help to understand the nature of the relationship among variables in a data set and allow for the detection of groups or clusters of points. Draftman plots or matrix plots are the visualisation of several bivariate scatterplots on the same display. They help detect structures in conditional dependencies by *brushing* across the plots. Outliers and observations that need special attention may be discovered with Andrews curves and PCPs. This chapter ends with an explanatory analysis of the Boston Housing data.

1.1 Boxplots

Example 1.1 The Swiss bank data (see Chap. 22, Sect. 22.2) consists of 200 measurements on Swiss bank notes. The first half of these measurements are from genuine bank notes, the other half are from counterfeit bank notes.

The authorities measured, as indicated in Fig. 1.1,

X_1 = length of the bill

X_2 = height of the bill (left)

X_3 = height of the bill (right)

X_4 = distance of the inner frame to the lower border

X_5 = distance of the inner frame to the upper border

X_6 = length of the diagonal of the central picture.

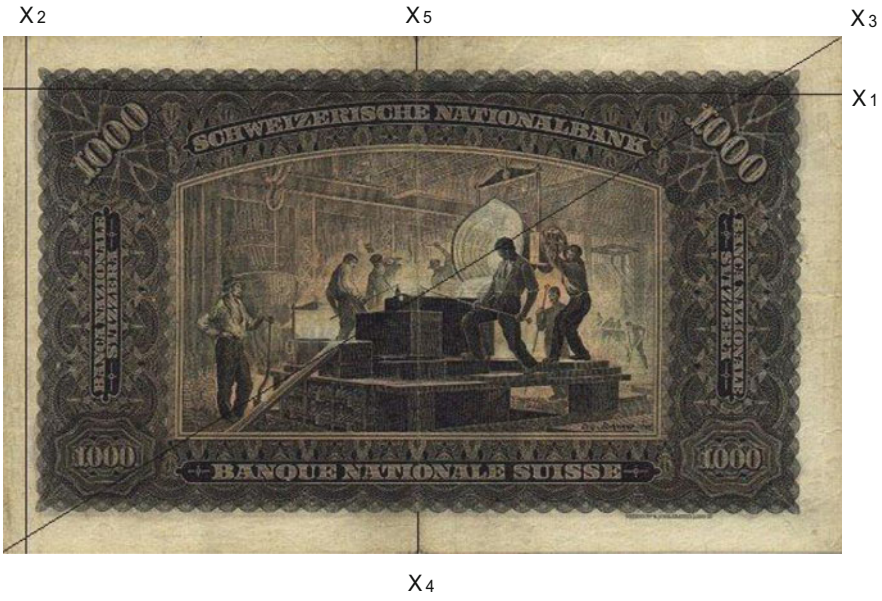

Fig. 1.1 An old Swiss 1000-franc bank note

These data are taken from Flury and Riedwyl (1988). The aim is to study how these measurements may be used in determining whether a bill is genuine or counterfeit.

The *boxplot* is a graphical technique that displays the distribution of variables. It helps us see the location, skewness, spread, tail length and outlying points.

It is particularly useful in comparing different batches. The boxplot is a graphical representation of the *Five Number Summary*. To introduce the Five Number Summary, let us consider for a moment a smaller, one-dimensional data set: the population of the 15 largest world cities in 2006 (Table 1.1).

In the Five Number Summary, we calculate the upper quartile F_U, the lower quartile F_L, the median and the extremes. Recall that order statistics $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ are a set of ordered values $x_1, x_2, \ldots, x_n$ where $x_{(1)}$ denotes the minimum and $x_{(n)}$ the maximum. The *median* M typically cuts the set of observations in two equal parts, and is defined as

$$M = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & n \text{ odd} \\ \frac{1}{2}\left\{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right\} & n \text{ even} \end{cases}. \tag{1.1}$$

Table 1.1 The 15 largest world cities in 2006

City	Country	Pop. (10,000)	Order statistics
Tokyo	Japan	3,420	$x_{(15)}$
Mexico city	Mexico	2,280	$x_{(14)}$
Seoul	South Korea	2,230	$x_{(13)}$
New York	USA	2,190	$x_{(12)}$
Sao Paulo	Brazil	2,020	$x_{(11)}$
Bombay	India	1,985	$x_{(10)}$
Delhi	India	1,970	$x_{(9)}$
Shanghai	China	1,815	$x_{(8)}$
Los Angeles	USA	1,800	$x_{(7)}$
Osaka	Japan	1,680	$x_{(6)}$
Jakarta	Indonesia	1,655	$x_{(5)}$
Calcutta	India	1,565	$x_{(4)}$
Cairo	Egypt	1,560	$x_{(3)}$
Manila	Philippines	1,495	$x_{(2)}$
Karachi	Pakistan	1,430	$x_{(1)}$

The quartiles cut the set into four equal parts, which are often called *fourths* (that is why we use the letter F). Using a definition that goes back to Hoaglin, Mosteller, and Tukey (1983) the definition of a median can be generalised to fourths, eights, etc. Considering the order statistics we can define the depth of a data value $x_{(i)}$ as $\min\{i, n - i + 1\}$. If n is odd, the depth of the median is $\frac{n+1}{2}$. If n is even, $\frac{n+1}{2}$ is a fraction. Thus, the median is determined to be the average between the two data values belonging to the next larger and smaller order statistics, i.e. $M = \frac{1}{2}\left\{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right\}$. In our example, we have $n = 15$ hence the median $M = x_{(8)} = 1{,}815$.

We proceed in the same way to get the fourths. Take the depth of the median and calculate

$$\text{depth of fourth} = \frac{[\text{depth of median}] + 1}{2}$$

with $[z]$ denoting the largest integer smaller than or equal to z. In our example this gives 4.5 and thus leads to the two fourths

$$F_L = \frac{1}{2}\left\{x_{(4)} + x_{(5)}\right\}$$

$$F_U = \frac{1}{2}\left\{x_{(11)} + x_{(12)}\right\}$$

(recalling that a depth which is a fraction corresponds to the average of the two nearest data values).

Table 1.2 Five number summary

#	15	World cities		
M	8		1,815	
F	4.5	1,610		2,105
	1	1,430		3,420

The F*-spread*, d_F, is defined as $d_F = F_U - F_L$. The *outside bars*

$$F_U + 1.5d_F \tag{1.2}$$

$$F_L - 1.5d_F \tag{1.3}$$

are the borders beyond which a point is regarded as an outlier. For the number of points outside these bars see Exercise 1.3. For the $n = 15$ data points the fourths are $1610 = \frac{1}{2}\left\{x_{(4)} + x_{(5)}\right\}$ and $2105 = \frac{1}{2}\left\{x_{(11)} + x_{(12)}\right\}$. Therefore the F*-spread* and the upper and lower *outside bars* in the above example are calculated as follows:

$$d_F = F_U - F_L = 2105 - 1610 = 495 \tag{1.4}$$

$$F_L - 1.5d_F = 1610 - 1.5 \cdot 495 = 867.5 \tag{1.5}$$

$$F_U + 1.5d_F = 2105 + 1.5 \cdot 495 = 2847.5. \tag{1.6}$$

Since Tokyo is beyond the outside bars it is considered to be an outlier. The minimum and the maximum are called the *extremes*. The *mean* is defined as

$$\overline{x} = n^{-1} \sum_{i=1}^{n} x_i,$$

which is 1,939.7 in our example. The mean is a measure of location. The median (1815), the fourths (1610;2105) and the extremes (1430;3420) constitute basic information about the data. The combination of these five numbers leads to the Five Number Summary as shown in Table 1.2. The depths of each of the five numbers have been added as an additional column.

Construction of the Boxplot

1. Draw a box with borders (edges) at F_L and F_U (i.e. 50 % of the data are in this box).
2. Draw the median as a solid line (|) and the mean as a dotted line (¦).
3. Draw "whiskers" from each end of the box to the most remote point that is NOT an outlier.
4. Show outliers as either "⋆" or "•"depending on whether they are outside of $F_{UL} \pm 1.5d_F$ or $F_{UL} \pm 3d_F$ respectively (this feather is not contained in some software). Label them if possible.

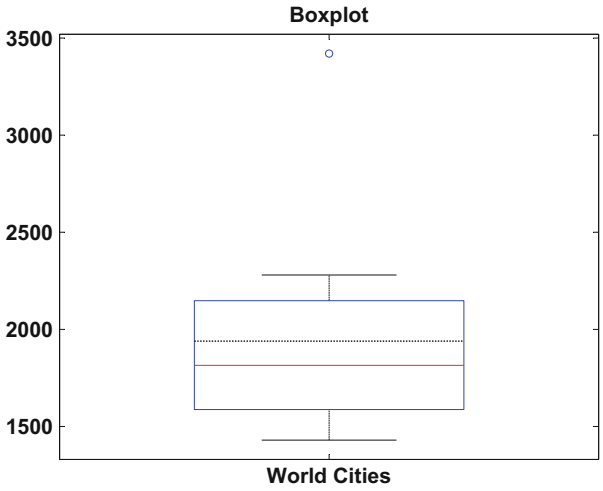

Fig. 1.2 Boxplot for world cities `MVAboxcity`

In the world cities example, the cut-off points (outside bars) are at 867.5 and 2847.5, hence we can draw whiskers to Karachi and Mexico City. We can see from Fig. 1.2 that the data are very skew: The upper half of the data (above the median) is more spread out than the lower half (below the median), the data contains one outlier marked as a circle and the mean (as a non-robust measure of location) is pulled away from the median.

Boxplots are very useful tools in comparing batches. The relative location of the distribution of different batches tells us a lot about the batches themselves. Before we come back to the Swiss bank data, let us compare the fuel economy of vehicles from different countries, see Fig. 1.3 and Table 22.3.

Example 1.2 The data are from the second column of Table 22.3 and show the mileage (miles per gallon) of American, Japanese and European cars. The five-number summaries for these data sets are $\{12, 16.8, 18.8, 22, 30\}$, $\{18, 22, 25, 30.5, 35\}$ and $\{14, 19, 23, 25, 28\}$ for American, Japanese and European cars, respectively. This reflects the information shown in Fig. 1.3. The following conclusions can be made:

- Japanese cars achieve higher fuel efficiency than US and European cars.
- There is one outlier, a very fuel-efficient car (VW-Rabbit Golf Diesel).
- The main body of the US car data (the box) lies below the Japanese car data.
- The worst Japanese car is more fuel-efficient than almost 50 % of the US cars.
- The spread of the Japanese and the US cars are almost equal.
- The median of the Japanese data is above that of the European data and the US data.

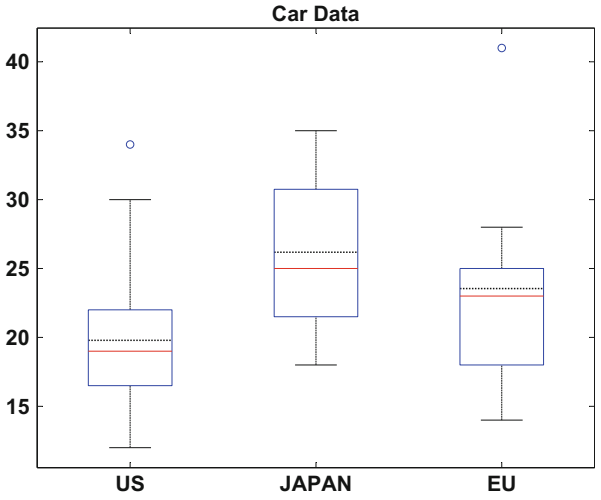

Fig. 1.3 Boxplot for the mileage of American, Japanese and European cars (from *left* to *right*) MVAboxcar

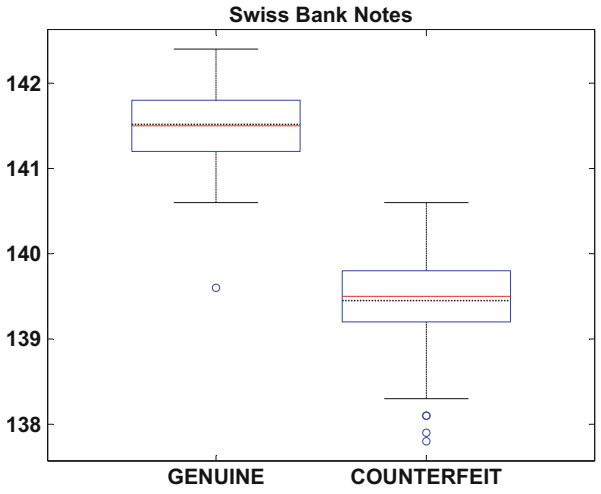

Fig. 1.4 The X_6 variable of Swiss bank data (diagonal of bank notes) MVAboxbank6

Table 1.3 Five number summary

#	100	Genuine bank notes		
M	50.5		141.5	
F	25.75	141.25		141.8
	1	140.65		142.4

Now let us apply the boxplot technique to the bank data set. In Fig. 1.4 we show the parallel boxplot of the diagonal variable X_6. On the left is the value of the genuine bank notes and on the right the value of the counterfeit bank notes. The five number summary is reported in Table 1.3 and 1.4.

Table 1.4 Five number summary

#	100	Counterfeit bank notes		
M	50.5		139.5	
F	25.75	139.2		139.8
	1	138.3		140.65

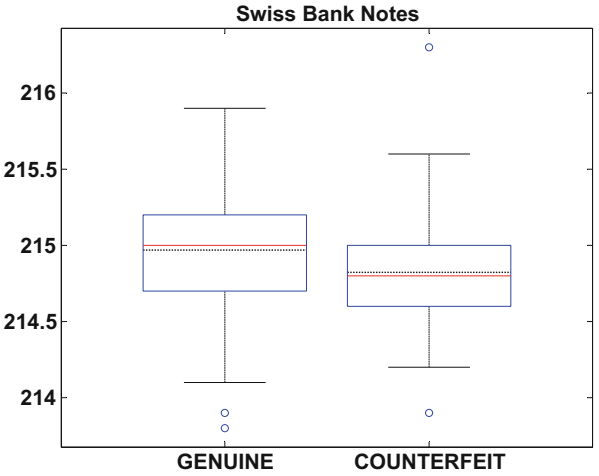

Fig. 1.5 The X_1 variable of Swiss bank data (length of bank notes) MVAboxbank1

One sees that the diagonals of the genuine bank notes tend to be larger. It is harder to see a clear distinction when comparing the length of the bank notes X_1, see Fig. 1.5. There are a few outliers in both plots. Almost all the observations of the diagonal of the genuine notes are above the ones from the counterfeit notes. There is one observation in Fig. 1.4 of the genuine notes that is almost equal to the median of the counterfeit notes. Can the parallel boxplot technique help us distinguish between the two types of bank notes?

Summary

- ↪ The median and mean bars are measures of locations.
- ↪ The relative location of the median (and the mean) in the box is a measure of how skewed it is.
- ↪ The length of the box and whiskers are a measure of spread.
- ↪ The length of the whiskers indicate the tail length of the distribution.
- ↪ The outlying points are indicated with a "⋆" or "•" depending on if they are outside of $F_{UL} \pm 1.5 d_F$ or $F_{UL} \pm 3 d_F$ respectively.

Summary (continued)
↪ The boxplots do not indicate multi-modality or clusters.
↪ If we compare the relative size and location of the boxes, we are comparing distributions.

1.2 Histograms

Histograms are density estimates. A density estimate gives a good impression of the distribution of the data. In contrast to boxplots, density estimates show possible multimodality of the data. The idea is to locally represent the data density by counting the number of observations in a sequence of consecutive intervals (bins) with origin x_0. Let $B_j(x_0, h)$ denote the *bin* of length h which is the element of a bin grid starting at x_0:

$$B_j(x_0, h) = [x_0 + (j-1)h, x_0 + jh), \quad j \in \mathbb{Z},$$

where $[.,.)$ denotes a left closed and right open interval. If $\{x_i\}_{i=1}^n$ is an i.i.d. sample with density f, the histogram is defined as follows:

$$\hat{f}_h(x) = n^{-1}h^{-1} \sum_{j \in \mathbb{Z}} \sum_{i=1}^{n} \boldsymbol{I}\{x_i \in B_j(x_0, h)\} \boldsymbol{I}\{x \in B_j(x_0, h)\}. \tag{1.7}$$

In sum (1.7) the first indicator function $\boldsymbol{I}\{x_i \in B_j(x_0, h)\}$ (see Symbols and Notation in Chap. 21) counts the number of observations falling into bin $B_j(x_0, h)$. The second indicator function is responsible for "localising" the counts around x. The parameter h is a smoothing or localising parameter and controls the width of the histogram bins. An h that is too large leads to very big blocks and thus to a very unstructured histogram. On the other hand, an h that is too small gives a very variable estimate with many unimportant peaks.

The effect of h is given in detail in Fig. 1.6. It contains the histogram (upper left) for the diagonal of the counterfeit bank notes for $x_0 = 137.8$ (the minimum of these observations) and $h = 0.1$. Increasing h to $h = 0.2$ and using the same origin, $x_0 = 137.8$, results in the histogram shown in the lower left of the figure. This density histogram is somewhat smoother due to the larger h. The binwidth is next set to $h = 0.3$ (upper right). From this histogram, one has the impression that the distribution of the diagonal is bimodal with peaks at about 138.5 and 139.9.

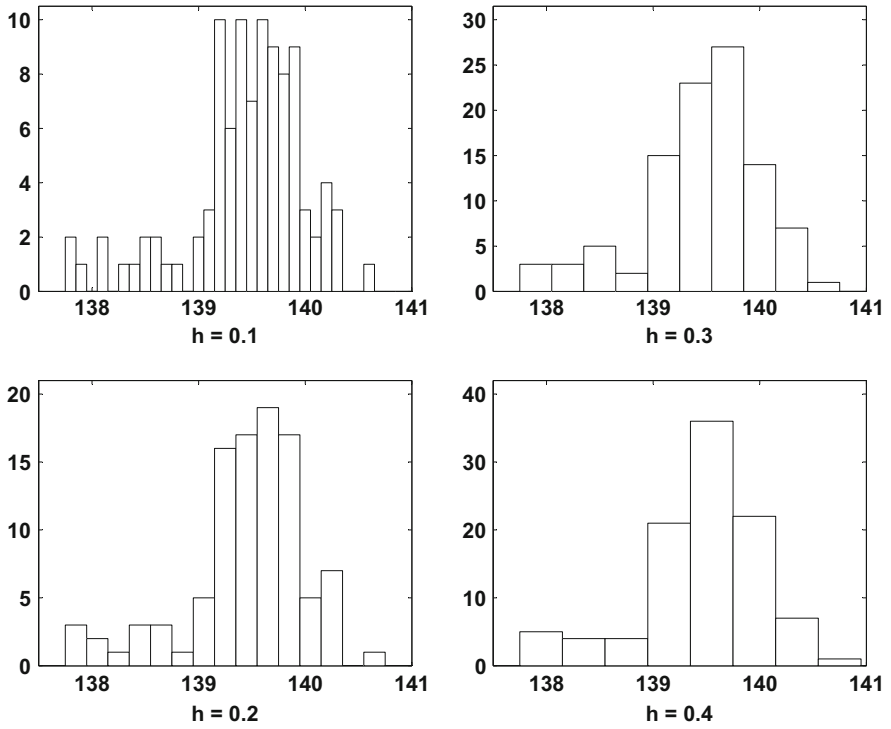

Fig. 1.6 Diagonal of counterfeit bank notes. Histograms with $x_0 = 137.8$ and $h = 0.1$ (*upper left*), $h = 0.2$ (*lower left*), $h = 0.3$ (*upper right*), $h = 0.4$ (*lower right*) `MVAhisbank1`

The detection of modes requires fine tuning of the binwidth. Using methods from smoothing methodology (Härdle, Müller, Sperlich, & Werwatz, 2004) one can find an "optimal" binwidth h for n observations:

$$h_{\text{opt}} = \left(\frac{24\sqrt{\pi}}{n} \right)^{1/3}.$$

Unfortunately, the binwidth h is not the only parameter determining the shapes of $\hat{f}$.

In Fig. 1.7, we show histograms with $x_0 = 137.65$ (upper left), $x_0 = 137.75$ (lower left), with $x_0 = 137.85$ (upper right), and $x_0 = 137.95$ (lower right). All the graphs have been scaled equally on the y-axis to allow comparison. One sees that—despite the fixed binwidth h—the interpretation is not facilitated. The shift of the origin x_0 (to four different locations) created four different histograms. This

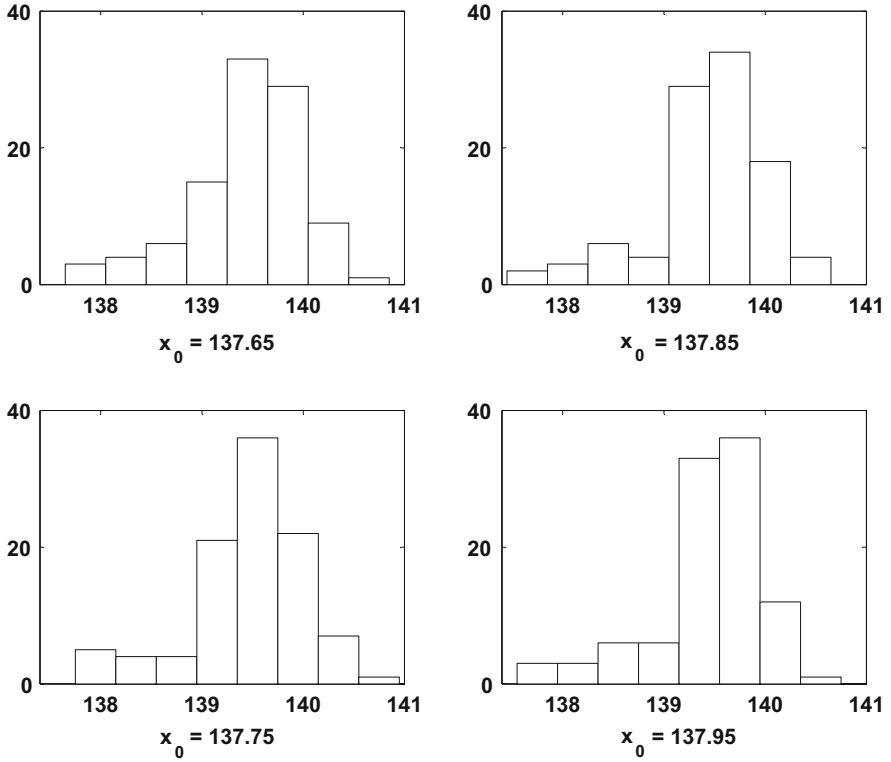

Fig. 1.7 Diagonal of counterfeit bank notes. Histogram with $h = 0.4$ and origins $x_0 = 137.65$ (*upper left*), $x_0 = 137.75$ (*lower left*), $x_0 = 137.85$ (*upper right*), $x_0 = 137.95$ (*lower right*) `MVAhisbank2`

property of histograms strongly contradicts the goal of presenting data features. Obviously, the same data are represented quite differently by the four histograms. A remedy has been proposed by Scott (1985): "Average the shifted histograms!". The result is presented in Fig. 1.8.

Here all bank note observations (genuine and counterfeit) have been used. The (so-called) averaged shifted histogram is no longer dependent on the origin and shows a clear bimodality of the diagonals of the Swiss bank notes.

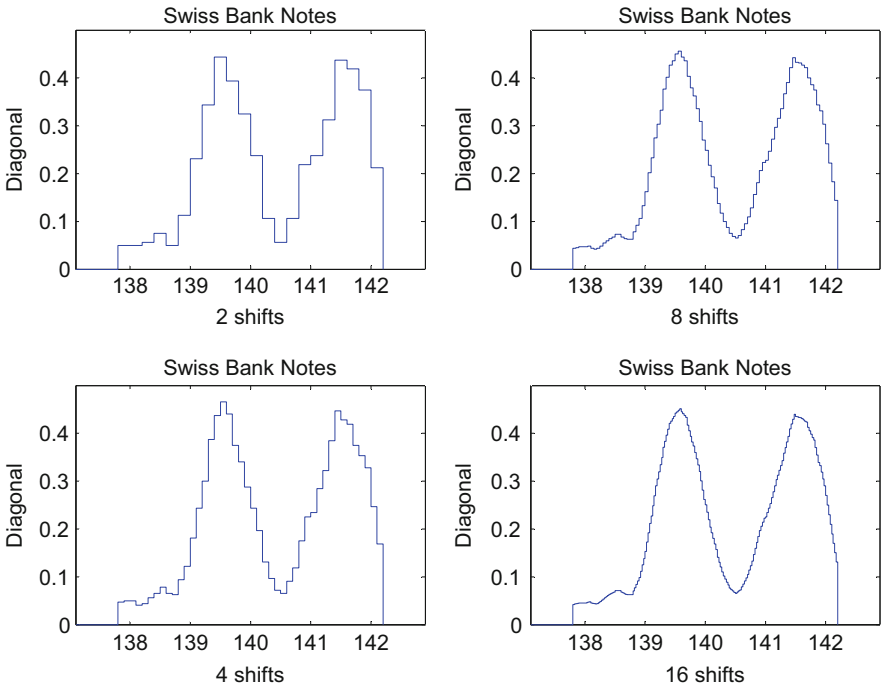

Fig. 1.8 Averaged shifted histograms based on all (counterfeit and genuine) Swiss bank notes: there are 2 shifts (*upper left*), 4 shifts (*lower left*), 8 shifts (*upper right*) and 16 shifts (*lower right*)
`MVAashbank`

Summary

- ↪ Modes of the density are detected with a histogram.
- ↪ Modes correspond to strong peaks in the histogram.
- ↪ Histograms with the same h need not be identical. They also depend on the origin x_0 of the grid.
- ↪ The influence of the origin x_0 is drastic. Changing x_0 creates different looking histograms.
- ↪ The consequence of an h that is too large is an unstructured histogram that is too flat.
- ↪ A binwidth h that is too small results in an unstable histogram.

Summary (continued)
↪ There is an "optimal" $h = (24\sqrt{\pi}/n)^{1/3}$.
↪ It is recommended to use averaged histograms. They are kernel densities.

1.3 Kernel Densities

The major difficulties of histogram estimation may be summarised in four critiques:

- determination of the binwidth h, which controls the shape of the histogram,
- choice of the bin origin x_0, which also influences to some extent the shape,
- loss of information since observations are replaced by the central point of the interval in which they fall,
- the underlying density function is often assumed to be smooth, but the histogram is not smooth.

Rosenblatt (1956), Whittle (1958) and Parzen (1962) developed an approach which avoids the last three difficulties. First, a smooth kernel function rather than a box is used as the basic building block. Second, the smooth function is centred directly over each observation. Let us study this refinement by supposing that x is the centre value of a bin. The histogram can in fact be rewritten as

$$\hat{f}_h(x) = n^{-1}h^{-1}\sum_{i=1}^{n} \boldsymbol{I}\left(|x - x_i| \leq \frac{h}{2}\right). \tag{1.8}$$

If we define $K(u) = \boldsymbol{I}(|u| \leq \frac{1}{2})$, then (1.8) changes to

$$\hat{f}_h(x) = n^{-1}h^{-1}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{1.9}$$

This is the general form of the kernel estimator. Allowing smoother kernel functions like the quartic kernel,

$$K(u) = \frac{15}{16}(1 - u^2)^2 \ \boldsymbol{I}(|u| \leq 1),$$

and computing x not only at bin centers gives us the kernel density estimator. Kernel estimators can also be derived via weighted averaging of rounded points (WARPing) or by averaging histograms with different origins, see Scott (1985). Table 1.5 introduces some commonly used kernels.

Table 1.5 Kernel functions

$K(\bullet)$	Kernel
$K(u) = \frac{1}{2} \boldsymbol{I}(\lvert u \rvert \leq 1)$	Uniform
$K(u) = (1 - \lvert u \rvert) \boldsymbol{I}(\lvert u \rvert \leq 1)$	Triangle
$K(u) = \frac{3}{4}(1 - u^2) \boldsymbol{I}(\lvert u \rvert \leq 1)$	Epanechnikov
$K(u) = \frac{15}{16}(1 - u^2)^2 \boldsymbol{I}(\lvert u \rvert \leq 1)$	Quartic (Biweight)
$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}) = \varphi(u)$	Gaussian

Different kernels generate different shapes of the estimated density. The most important parameter is the so-called bandwidth h, and can be optimised, for example, by cross-validation; see Härdle (1991) for details. The cross-validation method minimises the integrated squared error. This measure of discrepancy is based on the squared differences $\left\{\hat{f}_h(x) - f(x)\right\}^2$. Averaging these squared deviations over a grid of points $\{x_l\}_{l=1}^{L}$ leads to

$$L^{-1} \sum_{l=1}^{L} \left\{\hat{f}_h(x_l) - f(x_l)\right\}^2.$$

Asymptotically, if this grid size tends to zero, we obtain the integrated squared error:

$$\int \left\{\hat{f}_h(x) - f(x)\right\}^2 dx.$$

In practice, it turns out that the method consists of selecting a bandwidth that minimises the cross-validation function

$$\int \hat{f}_h^2 - 2 \sum_{i=1}^{n} \hat{f}_{h,i}(x_i),$$

where $\hat{f}_{h,i}$ is the density estimate obtained by using all datapoints except for the i-th observation. Both terms in the above function involve double sums. Computation may therefore be slow. There are many other density bandwidth selection methods. Probably the fastest way to calculate this is to refer to some reasonable reference distribution. The idea of using the Normal distribution as a reference, for example, goes back to Silverman (1986). The resulting choice of h is called the *rule of thumb*.

For the Gaussian kernel from Table 1.5 and a Normal reference distribution, the rule of thumb is to choose

$$h_G = 1.06 \, \hat{\sigma} \, n^{-1/5} \tag{1.10}$$

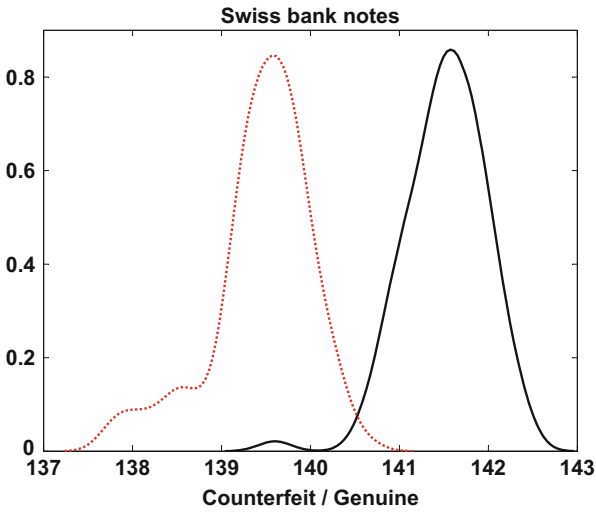

Fig. 1.9 Densities of the diagonals of genuine and counterfeit bank notes. Automatic density estimates MVAdenbank

where $\hat{\sigma} = \sqrt{n^{-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ denotes the sample standard deviation. This choice of h_G optimises the integrated squared distance between the estimator and the true density. For the quartic kernel, we need to transform (1.10). The modified rule of thumb is:

$$h_Q = 2.62 \cdot h_G. \tag{1.11}$$

Figure 1.9 shows the automatic density estimates for the diagonals of the counterfeit and genuine bank notes. The density on the left is the density corresponding to the diagonal of the counterfeit data. The separation is clearly visible, but there is also an overlap. The problem of distinguishing between the counterfeit and genuine bank notes is not solved by just looking at the diagonals of the notes. The question arises whether a better separation could be achieved using not only the diagonals, but one or two more variables of the data set. The estimation of higher dimensional densities is analogous to that of one dimensional. We show a two-dimensional density estimate for X_4 and X_5 in Fig. 1.10. The contour lines indicate the height of the density. One sees two separate distributions in this higher dimensional space, but they still overlap to some extent.

We can add one more dimension and give a graphical representation of a three-dimensional density estimate, or more precisely an estimate of the joint distribution of X_4, X_5 and X_6. Figure 1.11 shows the contour areas at three different levels of the density: 0.2 (green), 0.4 (red) and 0.6 (blue) of this three-dimensional density estimate. One can clearly recognise two "ellipsoids" (at each level), but as before, they overlap. In Chap. 14 we will learn how to separate the two ellipsoids and how to develop a discrimination rule to distinguish between these data points.

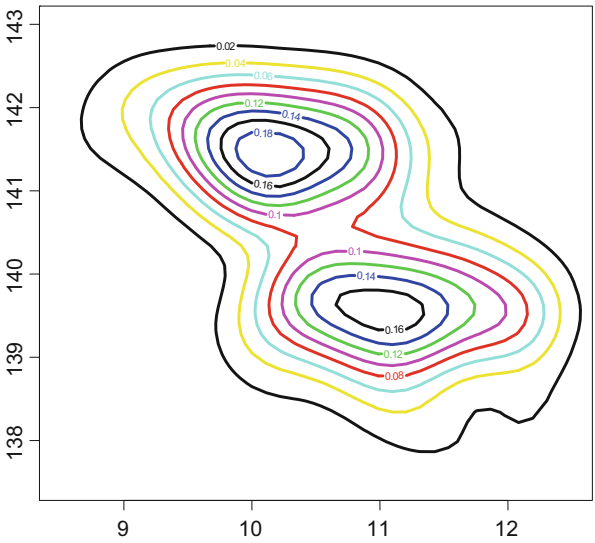

Fig. 1.10 Contours of the density of X_5 and X_6 of genuine and counterfeit bank notes `MVAcontbank2`

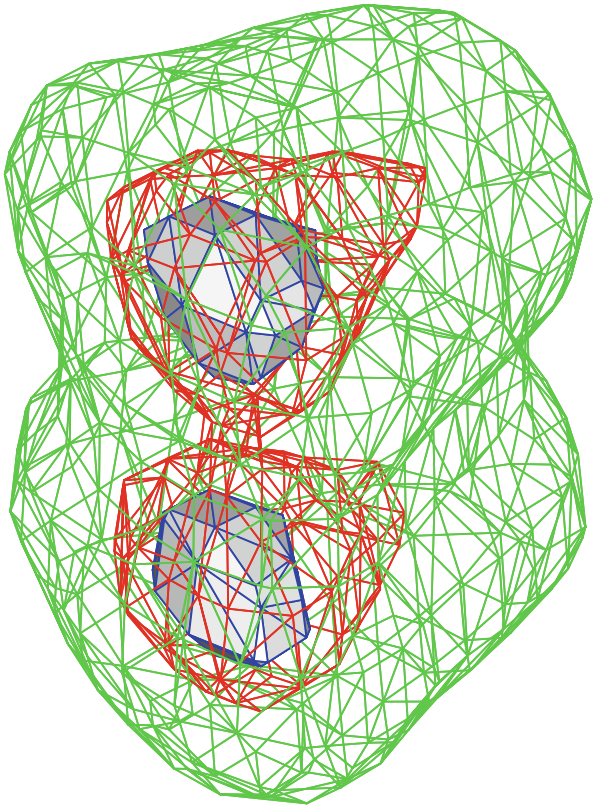

Fig. 1.11 Contours of the density of X_4, X_5, X_6 of genuine and counterfeit bank notes `MVAcontbank3`

Summary

- ↪ Kernel densities estimate distribution densities by the kernel method.
- ↪ The bandwidth h determines the degree of smoothness of the estimate $\hat{f}$.
- ↪ Kernel densities are smooth functions and they can graphically represent distributions (up to three dimensions).
- ↪ A simple (but not necessarily correct) way to find a good bandwidth is to compute the rule of thumb bandwidth $h_G = 1.06\hat{\sigma}n^{-1/5}$. This bandwidth is to be used only in combination with a Gaussian kernel φ.
- ↪ Kernel density estimates are a good descriptive tool for seeing modes, location, skewness, tails, asymmetry, etc.

1.4 Scatterplots

Scatterplots are bivariate or trivariate plots of variables against each other. They help us understand relationships among the variables of a data set. A downward-sloping scatter indicates that as we increase the variable on the horizontal axis, the variable on the vertical axis decreases. An analogous statement can be made for upward-sloping scatters.

Figure 1.12 plots the 5th column (upper inner frame) of the bank data against the 6th column (diagonal). The scatter is downward-sloping. As we already know from the previous section on marginal comparison (e.g. Fig. 1.9) a good separation between genuine and counterfeit bank notes is visible for the diagonal variable. The sub-cloud in the upper half (circles) of Fig. 1.12 corresponds to the true bank notes. As noted before, this separation is not distinct, since the two groups overlap somewhat.

This can be verified in an interactive computing environment by showing the index and coordinates of certain points in this scatterplot. In Fig. 1.12, the 70th observation in the merged data set is given as a thick circle, and it is from a genuine bank note. This observation lies well embedded in the cloud of counterfeit bank notes. One straightforward approach that could be used to tell the counterfeit from the genuine bank notes is to draw a straight line and define notes above this value as genuine. We would of course misclassify the 70th observation, but can we do better?

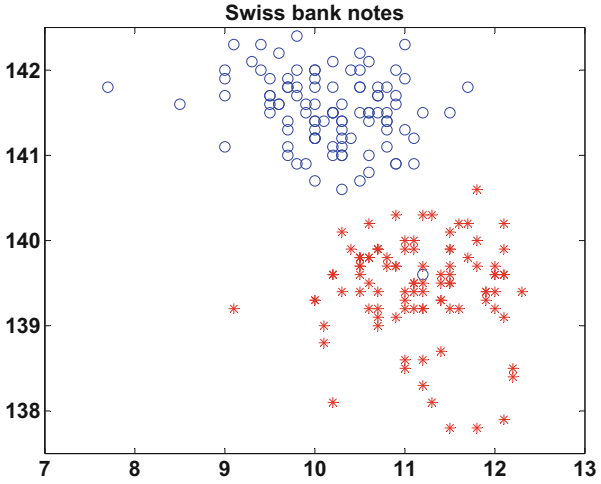

Fig. 1.12 2D scatterplot for X_5 vs. X_6 of the bank notes. Genuine notes are *circles*, counterfeit notes are *stars* `MVAscabank56`

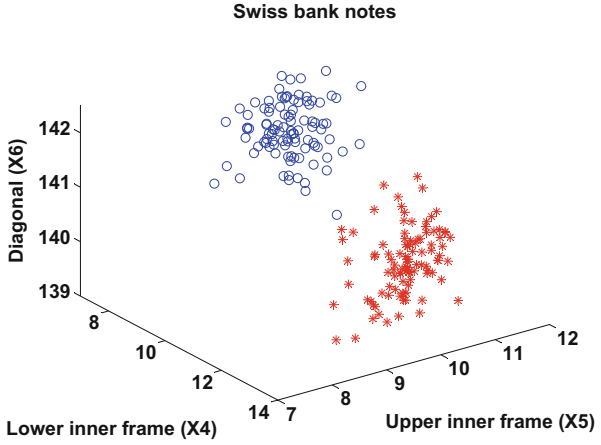

Fig. 1.13 3D scatterplot of the bank notes for (X_4, X_5, X_6). Genuine notes are *circles*, counterfeit are *stars* `MVAscabank456`

If we extend the two-dimensional scatterplot by adding a third variable, e.g. X_4 (lower distance to inner frame), we obtain the scatterplot in three dimensions as shown in Fig. 1.13. It becomes apparent from the location of the point clouds that a better separation is obtained. We have rotated the three-dimensional data until this satisfactory 3D view was obtained. Later, we will see that the rotation is the same as bundling a high-dimensional observation into one or more linear combinations

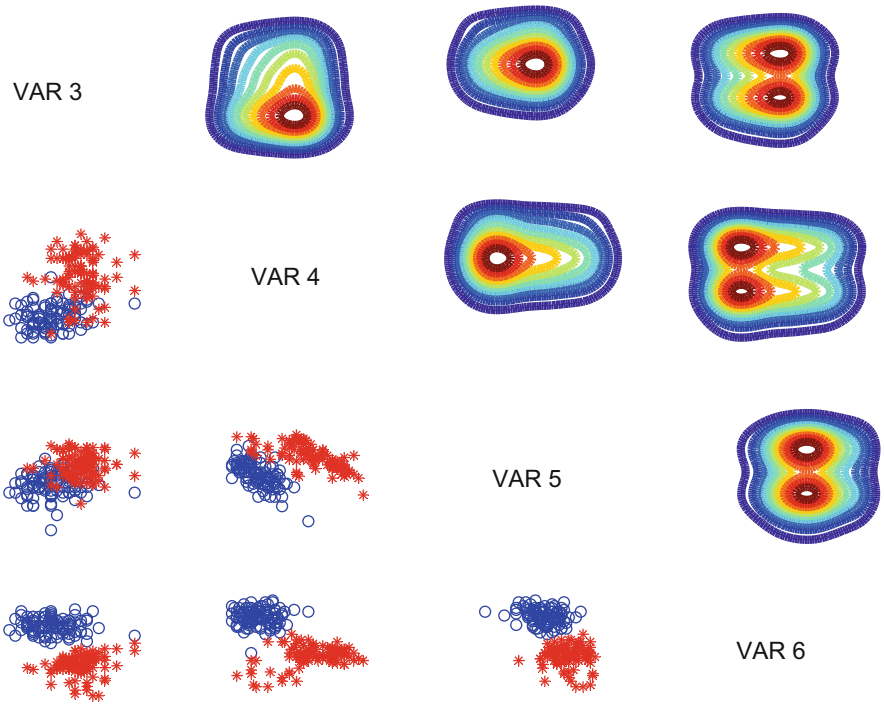

Fig. 1.14 Draftman's plot of the bank notes. The pictures in the *left-hand* column show (X_3, X_4), (X_3, X_5) and (X_3, X_6), in the *middle* we have (X_4, X_5) and (X_4, X_6), and in the *lower right* (X_5, X_6). The *upper right half* contains the corresponding density contour plots `MVAdrafbank4`

of the elements of the observation vector. In other words, the "separation line" parallel to the horizontal coordinate axis in Fig. 1.12 is, in Fig. 1.13, a plane and no longer parallel to one of the axes. The formula for such a separation plane is a linear combination of the elements of the observation vector:

$$a_1 x_1 + a_2 x_2 + \cdots + a_6 x_6 = \text{const.} \tag{1.12}$$

The algorithm that automatically finds the weights $(a_1, \ldots, a_6)$ will be investigated later on in Chap. 14.

Let us study yet another technique: the scatterplot matrix. If we want to draw all possible two-dimensional scatterplots for the variables, we can create a so-called *draftman's plot* (named after a draftman who prepares drafts for parliamentary discussions). Similar to a draftman's plot the scatterplot matrix helps in creating new ideas and in building knowledge about dependencies and structure.

Figure 1.14 shows a draftman's plot applied to the last four columns of the full bank data set. For ease of interpretation we have distinguished between the group of counterfeit and genuine bank notes by a different colour. As discussed several times

earlier, the separability of the two types of notes is different for different scatterplots. Not only is it difficult to perform this separation on, say, scatterplot X_3 vs. X_4, in addition the "separation line" is no longer parallel to one of the axes. The most obvious separation happens in the scatterplot in the lower right-hand side where indicated, as in Fig. 1.12, X_5 vs. X_6. The separation line here would be upward-sloping with an intercept at about $X_6 = 139$. The upper right half of the draftman's plot shows the density contours that we introduced in Sect. 1.3.

The power of the draftman's plot lies in its ability to show the internal connections of the scatter diagrams. Define a *brush* as a re-scalable rectangle that we can move via keyboard or mouse over the screen. Inside the brush we can highlight or colour observations. Suppose the technique is installed in such a way that as we move the brush in one scatter, the corresponding observations in the other scatters are also highlighted. By moving the brush, we can study conditional dependence.

If we brush (i.e. highlight or colour the observation with the brush), the X_5 vs. X_6 plot and move through the upper point cloud, we see that in other plots (e.g. X_3 vs. X_4), the corresponding observations are more embedded in the other sub-cloud.

Summary

- ↪ Scatterplots in two and three dimensions helps in identifying separated points, outliers or sub-clusters.
- ↪ Scatterplots help us in judging positive or negative dependencies.
- ↪ Draftman scatterplot matrices help detect structures conditioned on values of other variables.
- ↪ As the brush of a scatterplot matrix moves through a point cloud, we can study conditional dependence.

1.5 Chernoff-Flury Faces

If we are given data in numerical form, we tend to also display it numerically. This was done in the preceding sections: an observation $x_1 = (1, 2)$ was plotted as the point $(1, 2)$ in a two-dimensional coordinate system. In multivariate analysis we want to understand data in low dimensions (e.g. on a 2D computer screen) although the structures are hidden in high dimensions. The numerical display of data structures using coordinates therefore ends at dimensions greater than three.

If we are interested in condensing a structure into 2D elements, we have to consider alternative graphical techniques. The Chernoff-Flury faces, for example, provide such a condensation of high-dimensional information into a simple "face". In fact faces are a simple way of graphically displaying high-dimensional data. The size of the face elements like pupils, eyes, upper and lower hair line, etc. are assigned to certain variables. The idea of using faces goes back to Chernoff (1973) and has been further developed by Bernhard Flury. We follow the design described in Flury and Riedwyl (1988) which uses the following characteristics.

1. right eye size
2. right pupil size
3. position of right pupil
4. right eye slant
5. horizontal position of right eye
6. vertical position of right eye
7. curvature of right eyebrow
8. density of right eyebrow
9. horizontal position of right eyebrow
10. vertical position of right eyebrow
11. right upper hair line
12. right lower hair line
13. right face line
14. darkness of right hair
15. right hair slant
16. right nose line
17. right size of mouth
18. right curvature of mouth

19–36. like 1–18, only for the left side.

First, every variable that is to be coded into a characteristic face element is transformed into a (0, 1) scale, i.e. the minimum of the variable corresponds to 0 and the maximum to 1. The extreme positions of the face elements therefore correspond to a certain "grin" or "happy" face element. Dark hair might be coded as 1, and blond hair as 0 and so on.

As an example, consider the observations 91–110 of the bank data. Recall that the bank data set consists of 200 observations of dimension 6 where, for example, X_6 is the diagonal of the note. If we assign the six variables to the following face elements

$$X_1 = 1, 19 \text{ (eye sizes)}$$

$$X_2 = 2, 20 \text{ (pupil sizes)}$$

$$X_3 = 4, 22 \text{ (eye slants)}$$

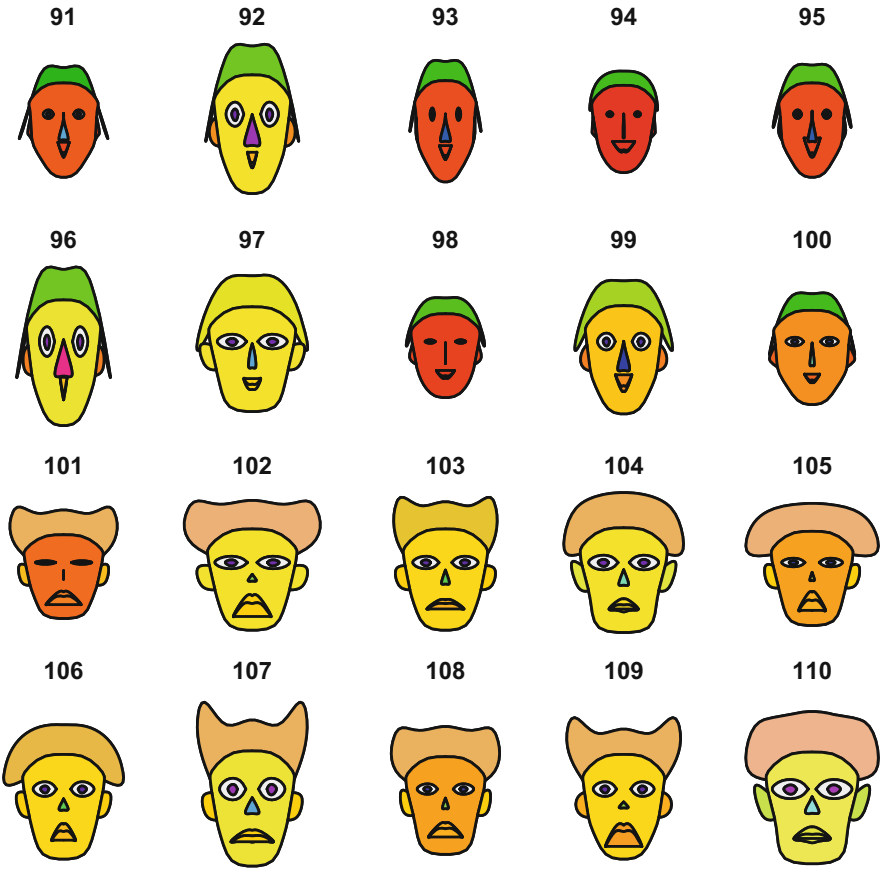

Fig. 1.15 Chernoff-Flury faces for observations 91–110 of the bank notes `MVAfacebank10`

$$X_4 = 11, 29 \text{ (upper hair lines)}$$
$$X_5 = 12, 30 \text{ (lower hair lines)}$$
$$X_6 = 13, 14, 31, 32 \text{ (face lines and darkness of hair)},$$

we obtain Fig. 1.15. Also recall that observations 1–100 correspond to the genuine notes, and that observations 101–200 correspond to the counterfeit notes. The counterfeit bank notes then correspond to the upper half of Fig. 1.15. In fact the faces for these observations look more grim and less happy. The variable X_6 (diagonal) already worked well in the boxplot in Fig. 1.4 in distinguishing between the counterfeit and genuine notes. Here, this variable is assigned to the face line and the darkness of the hair. That is why we clearly see a good separation within these 20 observations.

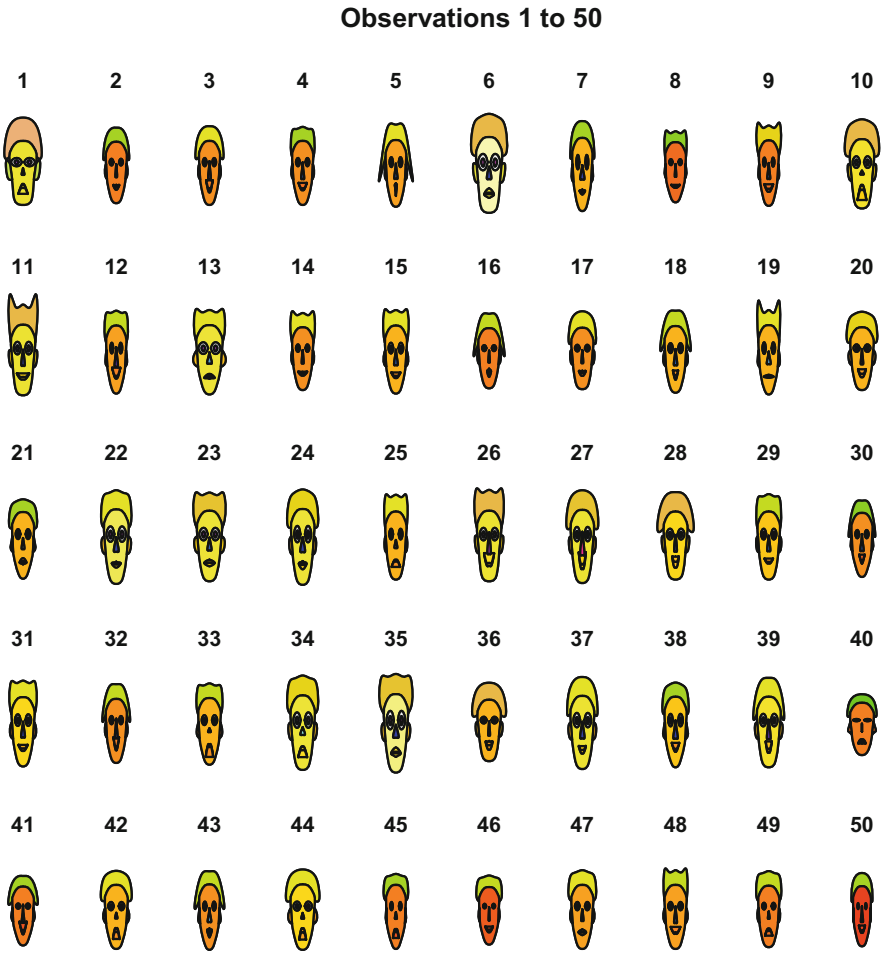

Fig. 1.16 Chernoff-Flury faces for observations 1–50 of the bank notes MVAfacebank50

What happens if we include all 100 genuine and all 100 counterfeit bank notes in the Chernoff-Flury face technique? Figures 1.16 and 1.17 show the faces of the genuine bank notes with the same assignments as used before, and Figs. 1.18 and 1.19 show the faces of the counterfeit bank notes. Comparing Figs. 1.16 and 1.18 one clearly sees that the diagonal (face line) is longer for genuine bank notes. Equivalently coded is the hair darkness (diagonal) which is lighter (shorter) for the counterfeit bank notes. One sees that the faces of the genuine bank notes have a much darker appearance and have broader face lines. The faces in Figs. 1.16 and 1.17 are obviously different from the ones in Figs. 1.18 and 1.19.

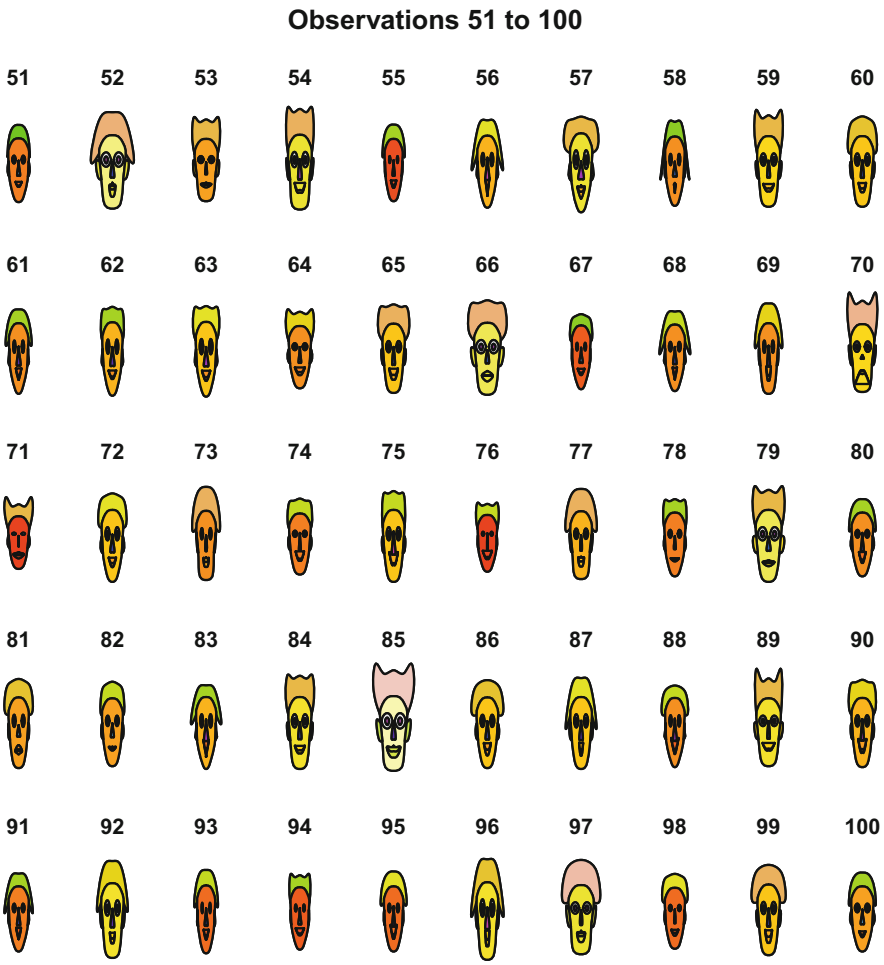

Fig. 1.17 Chernoff-Flury faces for observations 51–100 of the bank notes `MVAfacebank50`

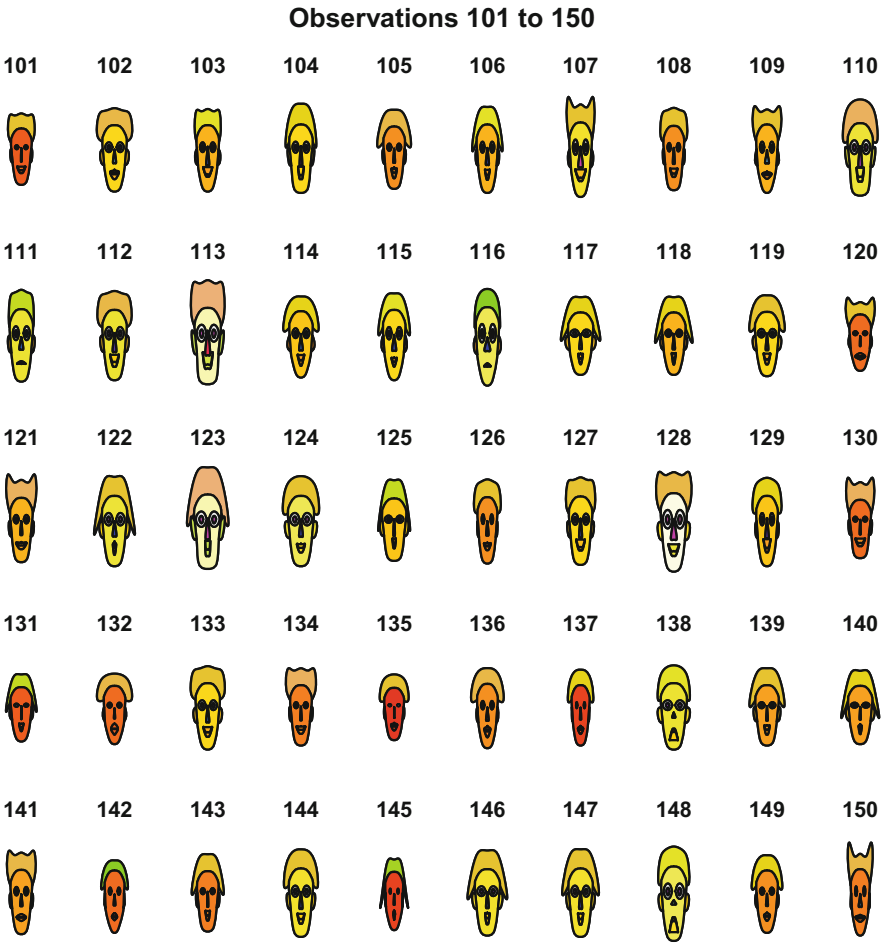

Fig. 1.18 Chernoff-Flury faces for observations 101–150 of the bank notes `MVAfacebank50`

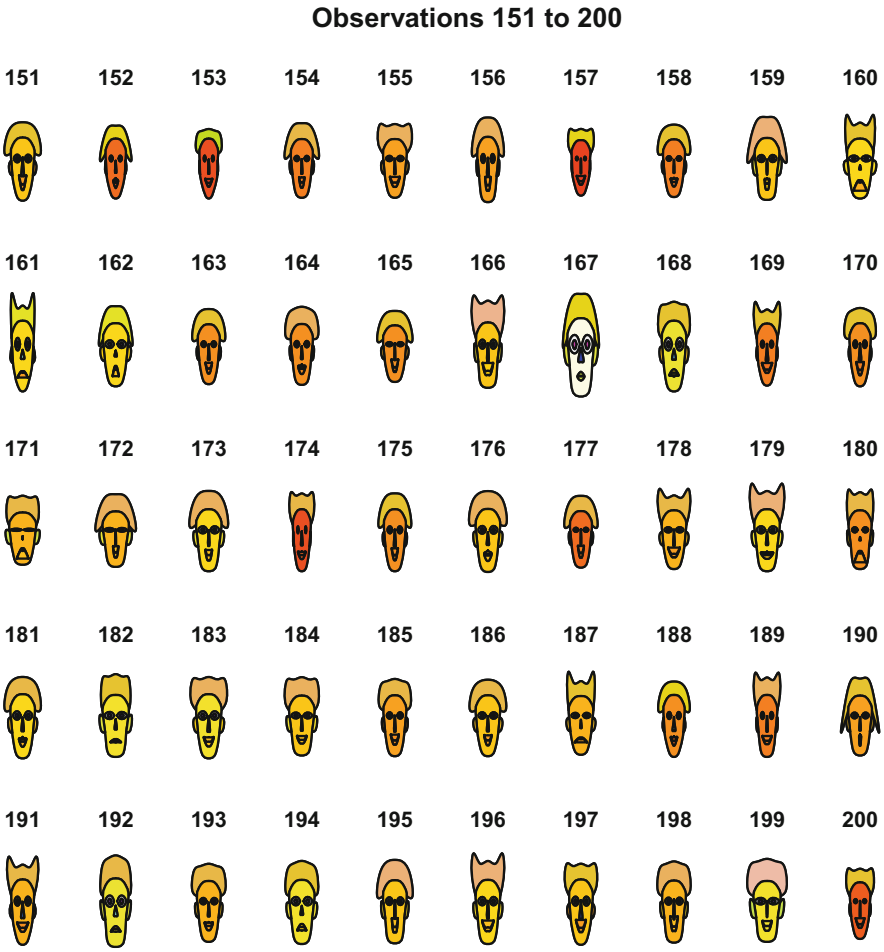

Fig. 1.19 Chernoff-Flury faces for observations 151–200 of the bank notes `MVAfacebank50`

Summary

- ↪ Faces can be used to detect sub-groups in multivariate data.
- ↪ Sub-groups are characterised by similar looking faces.
- ↪ Outliers are identified by extreme faces, e.g. dark hair, smile or a happy face.
- ↪ If one element of X is unusual, the corresponding face element significantly changes in shape.

1.6 Andrews' Curves

The basic problem of graphical displays of multivariate data is the dimensionality. Scatterplots work well up to three dimensions (if we use interactive displays). More than three dimensions have to be coded into displayable 2D or 3D structures (e.g. faces). The idea of coding and representing multivariate data by curves was suggested by Andrews (1972). Each multivariate observation $X_i = (X_{i,1}, \ldots, X_{i,p})$ is transformed into a curve as follows:

$$f_i(t) = \begin{cases} \frac{X_{i,1}}{\sqrt{2}} + X_{i,2}\sin(t) + X_{i,3}\cos(t) + \cdots \\ \quad + X_{i,p-1}\sin\left(\frac{p-1}{2}t\right) + X_{i,p}\cos\left(\frac{p-1}{2}t\right) & \text{for } p \text{ odd} \\ \frac{X_{i,1}}{\sqrt{2}} + X_{i,2}\sin(t) + X_{i,3}\cos(t) + \cdots + X_{i,p}\sin\left(\frac{p}{2}t\right) & \text{for } p \text{ even} \end{cases} \tag{1.13}$$

the observation represents the coefficients of a so-called Fourier series ($t \in [-\pi, \pi]$).

Suppose that we have three-dimensional observations: $X_1 = (0, 0, 1)$, $X_2 = (1, 0, 0)$ and $X_3 = (0, 1, 0)$. Here $p = 3$ and the following representations correspond to the Andrews' curves:

$$\begin{aligned} f_1(t) &= \cos(t) \\ f_2(t) &= \frac{1}{\sqrt{2}} \quad \text{and} \\ f_3(t) &= \sin(t). \end{aligned}$$

These curves are indeed quite distinct, since the observations X_1, X_2, and X_3 are the 3D unit vectors: each observation has mass only in one of the three dimensions. The order of the variables plays an important role.

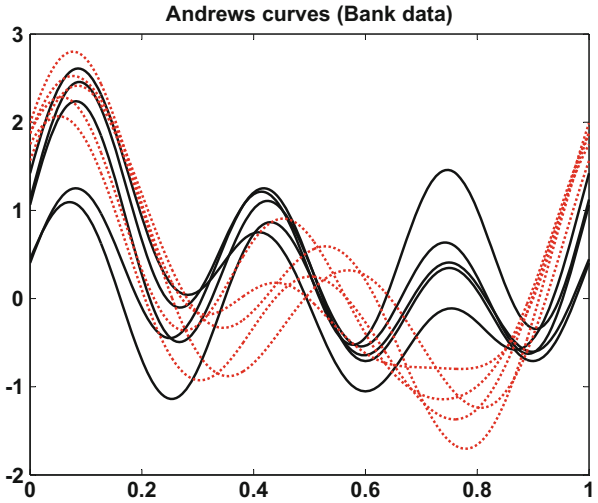

Fig. 1.20 Andrews' curves of the observations 96–105 from the Swiss bank note data. The order of the variables is 1,2,3,4,5,6 MVAandcur

Example 1.3 Let us take the 96th observation of the Swiss bank note data set,

$$X_{96} = (215.6, 129.9, 129.9, 9.0, 9.5, 141.7).$$

The Andrews' curve is by (1.13):

$$f_{96}(t) = \frac{215.6}{\sqrt{2}} + 129.9\sin(t) + 129.9\cos(t) + 9.0\sin(2t) + 9.5\cos(2t) + 141.7\sin(3t).$$

Figure 1.20 shows the Andrews' curves for observations 96–105 of the Swiss bank note data set. We already know that the observations 96–100 represent genuine bank notes, and that the observations 101–105 represent counterfeit bank notes. We see that at least four curves differ from the others, but it is hard to tell which curve belongs to which group.

We know from Fig. 1.4 that the sixth variable is an important one. Therefore, the Andrews' curves are calculated again using a reversed order of the variables.

Example 1.4 Let us consider again the 96th observation of the Swiss bank note data set,

$$X_{96} = (215.6, 129.9, 129.9, 9.0, 9.5, 141.7).$$

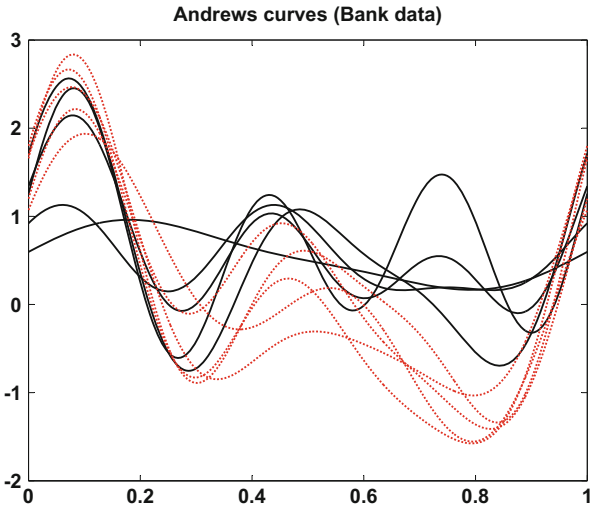

Fig. 1.21 Andrews' curves of the observations 96–105 from the Swiss bank note data. The order of the variables is 6,5,4,3,2,1 `MVAandcur2`

The Andrews' curve is computed using the reversed order of variables:

$$f_{96}(t) = \frac{141.7}{\sqrt{2}} + 9.5\sin(t) + 9.0\cos(t) + 129.9\sin(2t) + 129.9\cos(2t) + 215.6\sin(3t).$$

In Fig. 1.21 the curves f_{96}–f_{105} for observations 96–105 are plotted. Instead of a difference in high frequency, now we have a difference in the intercept, which makes it more difficult for us to see the differences in observations.

This shows that the order of the variables plays an important role in the interpretation. If X is high-dimensional, then the last variables will only have a small visible contribution to the curve: they fall into the high frequency part of the curve. To overcome this problem Andrews suggested using an order which is suggested by Principal Component Analysis. This technique will be treated in detail in Chap. 11. In fact, the sixth variable will appear there as the most important variable for discriminating between the two groups. If the number of observations is more than 20, there may be too many curves in one graph. This will result in an over plotting of curves or a bad "signal-to-ink-ratio", see Tufte (1983). It is therefore advisable to present multivariate observations via Andrews' curves only for a limited number of observations.

Summary

- ↪ Outliers appear as single Andrews' curves that look different from the rest.
- ↪ A sub-group of data is characterised by a set of similar curves.
- ↪ The order of the variables plays an important role for interpretation.
- ↪ The order of variables may be optimised by Principal Component Analysis.
- ↪ For more than 20 observations we may obtain a bad "signal-to-ink ratio", i.e. too many curves are overlaid in one picture.

1.7 Parallel Coordinates Plots

PCP is a method for representing high-dimensional data, see Inselberg (1985). Instead of plotting observations in an orthogonal coordinate system, PCP draws coordinates in parallel axes and connects them with straight lines. This method helps in representing data with more than four dimensions.

One first scales all variables to max $= 1$ and min $= 0$. The coordinate index j is drawn onto the horizontal axis, and the scaled value of variable x_{ij} is mapped onto the vertical axis. This way of representation is very useful for high-dimensional data. It is however also sensitive to the order of the variables, since certain trends in the data can be shown more clearly in one ordering than in another.

Example 1.5 Take, once again, the observations 96–105 of the Swiss bank notes. These observations are six dimensional, so we can't show them in a six-dimensional Cartesian coordinate system. Using the PCP technique, however, they can be plotted on parallel axes. This is shown in Fig. 1.22.

PCP can also be used for detecting linear dependencies between variables: if all the lines are of almost parallel dimensions ($p = 2$), there is a positive linear dependence between them. In Fig. 1.23 we display the two variables weight and displacement for the car data set in Sect. 22.3. The correlation coefficient ρ introduced in Sect. 3.2 is 0.9. If all lines intersect visibly in the middle, there is evidence of a negative linear dependence between these two variables, see Fig. 1.24. In fact the correlation is $\rho = -0.82$ between two variables mileage and weight: The more the weight, the less the mileage.

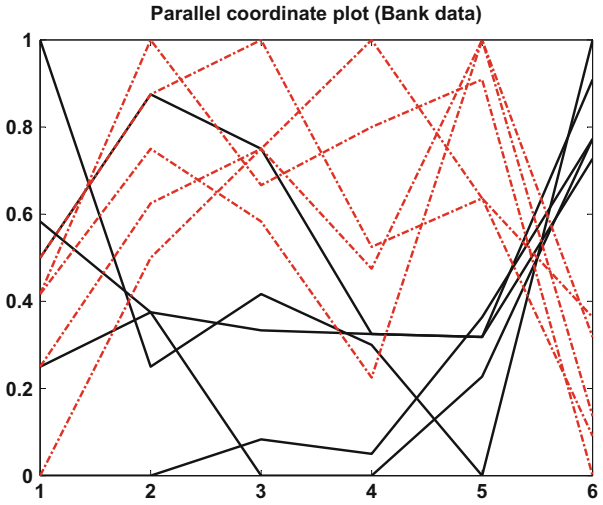

Fig. 1.22 Parallel coordinates plot of observations 96–105 MVAparcoo1

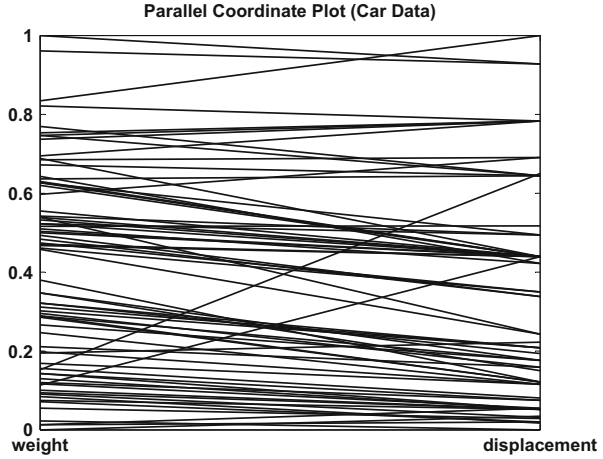

Fig. 1.23 Parallel coordinates plot indicating strong positive dependence with $\rho = 0.9$, $X_1 =$ weight, $X_2 =$ displacement MVApcp2

Another use of PCP is sub-groups detection. Lines converging to different discrete points indicate sub-groups. Figure 1.25 shows the last three variables—displacement, gear ratio for high gear and company's headquarters of the car data; we see convergence to the last variable. This last variable is the company's headquarters with three discrete values: USA, Japan and Europe. PCP can also be used for outlier detection. Figure 1.26 shows the variables headroom, rear seat

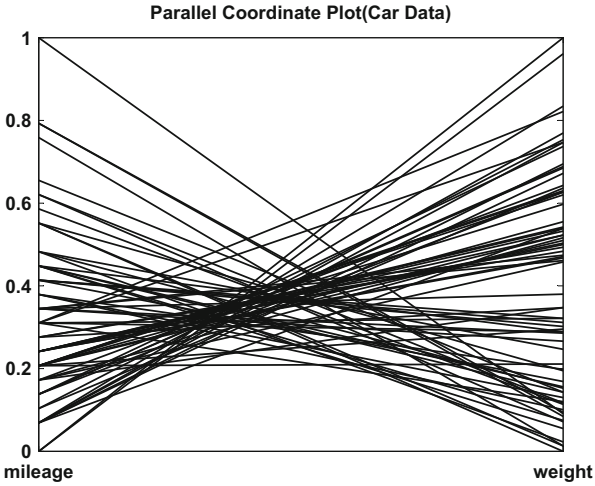

Fig. 1.24 Parallel coordinates plot showing strong negative dependence with $\rho = -0.82$, $X_1 =$ mileage, $X_2 =$ weight MVApcp3

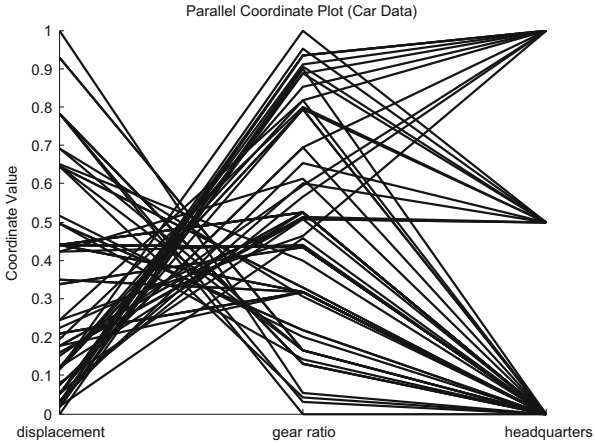

Fig. 1.25 Parallel coordinates plot with sub-groups MVApcp4

clearance and trunk (boot) space in the car data set. There are two outliers visible. The boxplot Fig. 1.27 confirms this.

PCPs have also possible shortcomings: We cannot distinguish observations when two lines cross at one point unless we distinguish them clearly (e.g. by different line style). In Fig. 1.28, observation A and B both have the same value at $j = 2$. Two lines cross at one point here. At the 3rd and 4th dimension we cannot tell which line belongs to which observation. A dotted line for A and solid line for B could have helped there.

To solve this problem one uses an interpolation curve instead of straight lines, e.g. cubic curves as in Graham and Kennedy (2003). Figure 1.29 is a variant of Fig. 1.28. In Fig. 1.29, with a natural cubic spline, it is evident how to follow the curves and distinguish the observations. The real power of PCP comes though through colouring sub-groups.

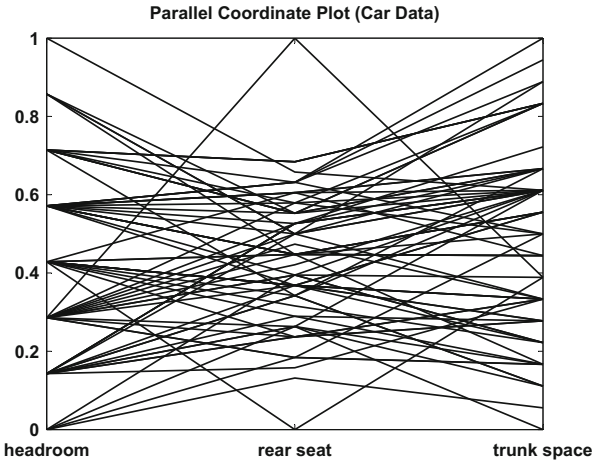

Fig. 1.26 PCP for $X_1 =$ headroom, $X_2 =$ rear seat clearance and $X_3 =$ trunk space MVApcp5

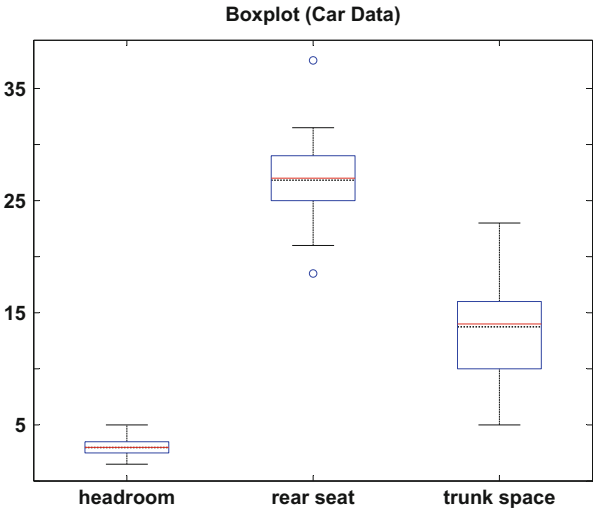

Fig. 1.27 Boxplots for headroom, rear seat clearance and trunk space MVApcp6

Example 1.6 Data in Fig. 1.30 are coloured according to X_{13}—car company's headquarters. Red stands for European car, green for Japan and black for US. This PCP with colouring can provide some information for us:

1. US cars (black) tend to have large value in X_7, X_8, X_9, X_{10}, X_{11} (trunk (boot) space, weight, length, turning diameter, displacement), which means US cars are generally larger.
2. Japanese cars (green) have large value in X_3, X_4 (both for repair record), which means Japanese cars tend to be repaired less.

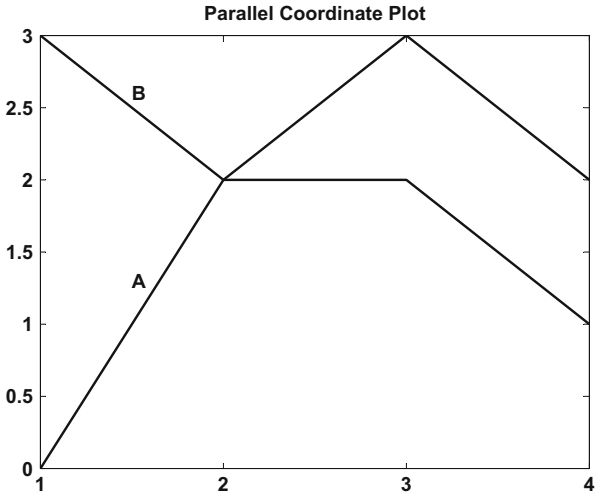

Fig. 1.28 PCP with intersection for given data points A = [0, 2, 3, 2] and B = [3, 2, 2, 1] MVApcp7

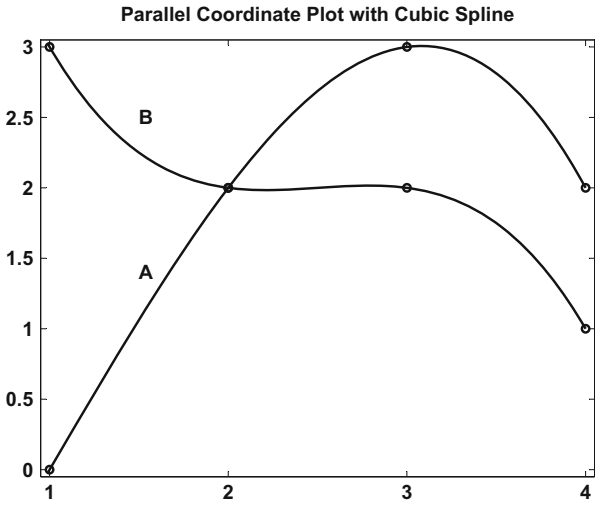

Fig. 1.29 PCP with cubic spline interpolation MVApcp8

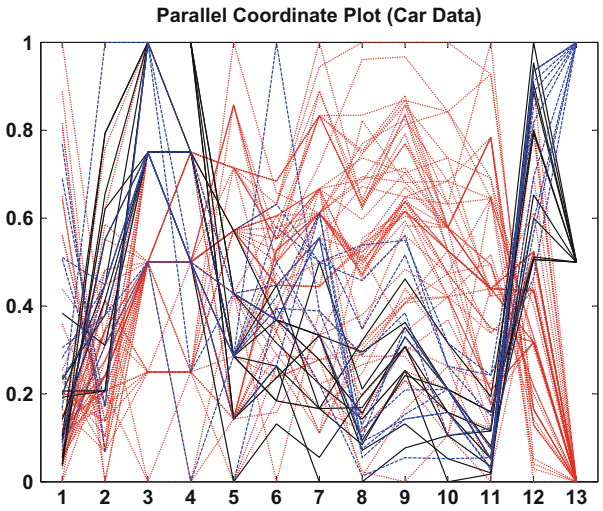

Fig. 1.30 Parallel coordinates plot for car data `MVApcp1`

Summary

- ↪ Parallel coordinates plots overcome the visualisation problem of the Cartesian coordinate system for dimensions greater than 4.
- ↪ Outliers are visible as outlying polygon curves.
- ↪ The order of variables is important, especially in the detection of sub-groups.
- ↪ Sub-groups may be screened by selective colouring.

1.8 Hexagon Plots

This section closely follows the presentation of Lewin-Koh (2006). In geometry, a hexagon is a polygon with six edges and six vertices. Hexagon binning is a type of bivariate histogram with hexagon borders. It is useful for visualising the structure

of data sets entailing a large number of observations n. The concept of hexagon binning is as follows:

1. The xy plane over the set (range(x), range(y)) is tessellated by a regular grid of hexagons.
2. The number of points falling in each hexagon is counted.
3. The hexagons with count > 0 are plotted by using a colour ramp or varying the radius of the hexagon in proportion to the counts.

This algorithm is extremely fast and effective for displaying the structure of data sets even for $n \geq 10^6$. If the size of the grid and the cuts in the colour ramp are chosen in a clever fashion, then the structure inherent in the data should emerge in the binned plot. The same caveats apply to hexagon binning as histograms. Variance and bias vary in opposite directions with bin width, so we have to settle for finding the value of the bin width that yields the optimal compromise between variance and bias reduction. Clearly, if we increase the size of the grid, the hexagon plot appears to be smoother, but without some reasonable criterion on hand it remains difficult to say which bin width provides the "optimal" degree of smoothness. The default number of bins suggested by standard software is 30.

Applications to some data sets are shown as follows. The data is taken from ALLBUS (2006)[ZA No.3762]. The number of respondents is 2,946. The following nine variables have been selected to analyse the relation between each pair of variables.

X_1:	Age
X_2:	Net income
X_3:	Time for television per day in minutes
X_4:	Time for work per week in hours
X_5:	Time for computer per week in hours
X_6:	Days for illness yearly
X_7:	Living space (square metres)
X_8:	Size
X_9:	Weight

Firstly, we consider two variables $X_1 =$ Age and $X_2 =$ Net income in Fig. 1.31. The top left picture is a scatter plot. The second one is a hexagon plot with borders making it easier to see the separation between hexagons. Looking at these plots one can see that almost all individuals have a net monthly income of less than 2,000 EUR. Only two individuals earn more than 10,000 EUR per month.

Figure 1.32 shows the relation between X_1 and X_5. About 40 % of respondents from 20 to 80 years old do not use a computer at least once per week. The respondent who deals with a computer 105 h each week was actually not in full-time employment.

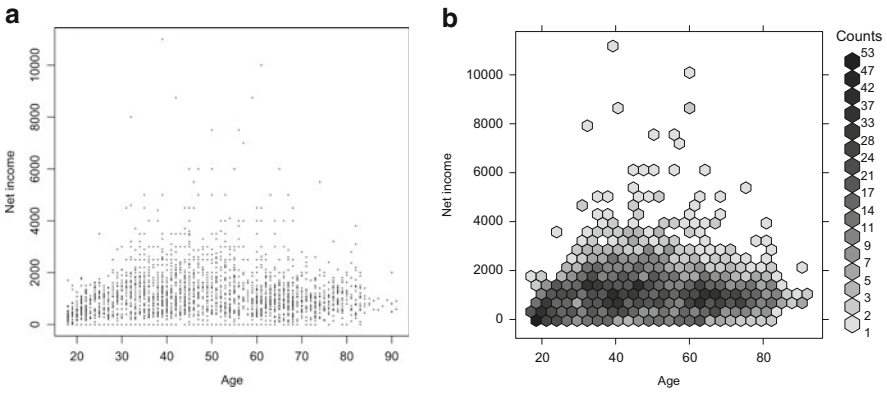

Fig. 1.31 Hexagon plots between X_1 and X_2 `MVAageIncome`

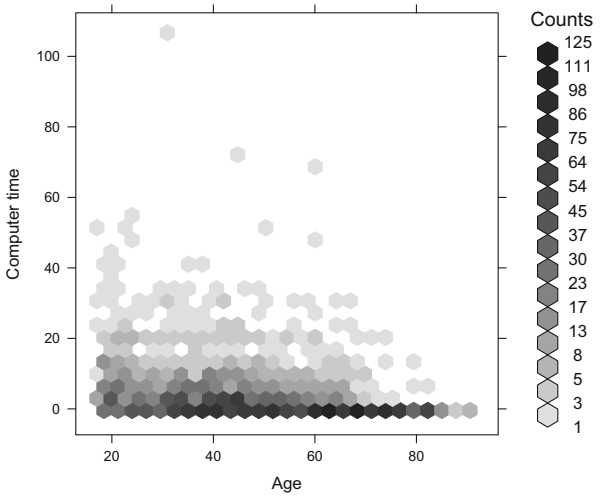

Fig. 1.32 Hexagon plot between X_1 and X_5 `MVAageCom`

Clearly, people who earn modest incomes live in smaller flats. The trend here is relatively clear in Fig. 1.33. The larger the net income, the larger the flat. A few people do however earn high incomes but live in small flats.

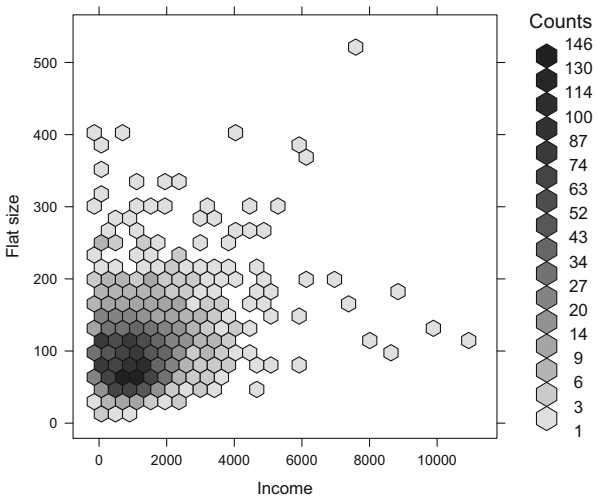

Fig. 1.33 Hexagon plot between X_2 and X_7 `MVAincomeLi`

Summary

- ↪ Hexagon binning is a type of bivariate histogram, used for visualising large data.
- ↪ Variance and bias vary in opposite directions with bin width.
- ↪ Hexagons have the property of "symmetry of the nearest neighbours" which lacks in square bins.
- ↪ Hexagons are visually less biased for displaying densities than other regular tessellations.

1.9 Boston Housing

Aim of the Analysis

The Boston Housing data set was analysed by Harrison and Rubinfeld (1978) who wanted to find out whether "clean air" had an influence on house prices. We will use this data set in this chapter and in most of the following chapters to illustrate the presented methodology. The data are described in Sect. 22.1.

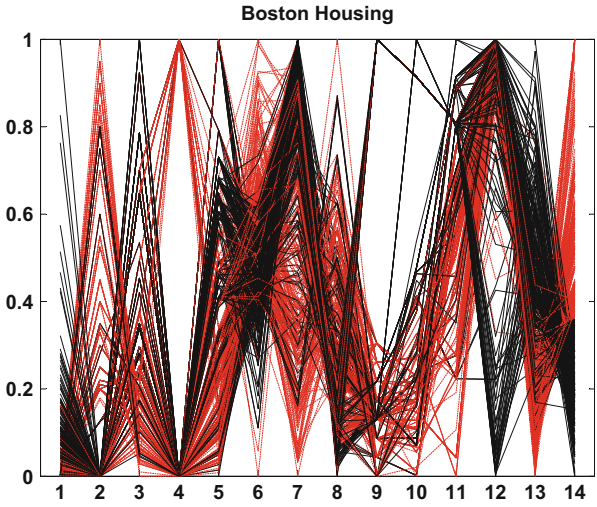

Fig. 1.34 Parallel coordinates plot for Boston housing data MVApcphousing

What Can Be Seen from the PCPs

In order to highlight the relations of X_{14} to the remaining 13 variables, we colour all of the observations with $X_{14} >$median(X_{14}) as red lines in Fig. 1.34. Some of the variables seem to be strongly related. The most obvious relation is the negative dependence between X_{13} and X_{14}. It can also be argued that a strong dependence exists between X_{12} and X_{14} since no red lines are drawn in the lower part of X_{12}. The opposite can be said about X_{11}: there are only red lines plotted in the lower part of this variable. Low values of X_{11} induce high values of X_{14}.

For the PCP, the variables have been rescaled over the interval $[0, 1]$ for better graphical representations. The PCP shows that the variables are not distributed in a symmetric manner. It can be clearly seen that the values of X_1 and X_9 are much more concentrated around 0. Therefore it makes sense to consider transformations of the original data.

The Scatterplot Matrix

One characteristic of PCPs is that many lines are drawn on top of each other. This problem is reduced by depicting the variables in pairs of scatterplots. Including all 14 variables in one large scatterplot matrix is possible, but makes it hard to see anything from the plots. Therefore, for illustratory purposes we will analyse only one such matrix from a subset of the variables in Fig. 1.35. On the basis of the PCP and the scatterplot matrix we would like to interpret each of the 13 variables and their eventual relation to the 14th variable. Included in the figure are images for

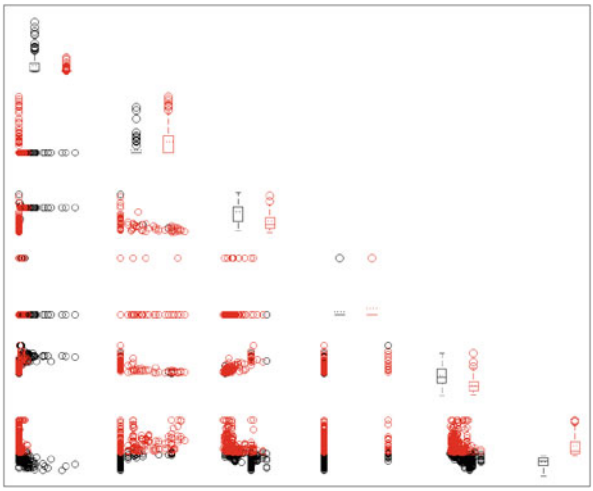

Fig. 1.35 Scatterplot matrix for variables $X_1, \ldots, X_5$ and X_{14} of the Boston housing data `MVAdrafthousing`

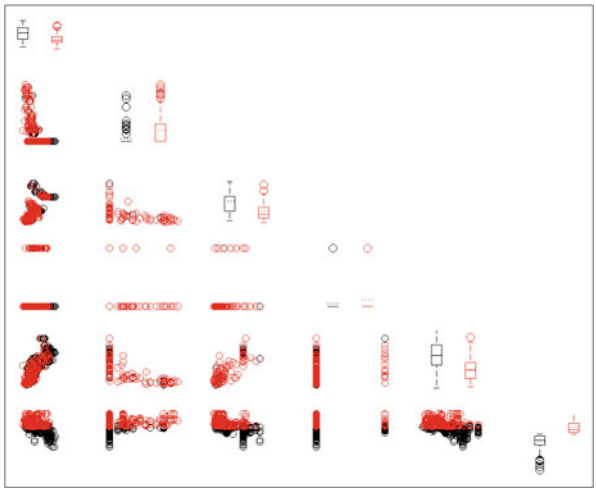

Fig. 1.36 Scatterplot matrix for variables $\tilde{X}_1, \ldots, \tilde{X}_5$ and $\tilde{X}_{14}$ of the Boston housing data `MVAdrafthousingt`

X_1–X_5 and X_{14}, although each variable is discussed in detail below. All references made to scatterplots in the following refer to Fig. 1.35.

Per-Capita Crime Rate X_1

Taking the logarithm makes the variable's distribution more symmetric. This can be seen in the boxplot of $\tilde{X}_1$ in Fig. 1.37 which shows that the median and the mean have moved closer to each other than they were for the original X_1. Plotting the KDE of $\tilde{X}_1 = \log(X_1)$ would reveal that two sub-groups might exist with different mean values. However, taking a look at the scatterplots in Fig. 1.36 of the logarithms

which include X_1 does not clearly reveal such groups. Given that the scatterplot of $\log(X_1)$ vs. $\log(X_{14})$ shows a relatively strong negative relation, it might be the case that the two sub-groups of X_1 correspond to houses with two different price levels. This is confirmed by the two boxplots shown to the right of the X_1 vs. X_2 scatterplot (in Fig. 1.35): the right boxplot's shape differs a lot from the black one's, having a much higher median and mean.

Proportion of Residential Area Zoned for Large Lots X_2

It strikes the eye in Fig. 1.35 that there is a large cluster of observations for which X_2 is equal to 0. It also strikes the eye that—as the scatterplot of X_1 vs. X_2 shows—there is a strong, though non-linear, negative relation between X_1 and X_2; almost all observations for which X_2 is high have an X_1-value close to zero, and vice versa, many observations for which X_2 is zero have quite a high per-capita crime rate X_1. This could be due to the location of the areas, e.g. urban districts might have a higher crime rate and at the same time it is unlikely that any residential land would be zoned in a generous manner.

As far as the house prices are concerned, it can be said that there seems to be no clear (linear) relation between X_2 and X_{14}, but it is obvious that the more expensive houses are situated in areas where X_2 is large (this can be seen from the two boxplots on the second position of the diagonal, where the red one has a clearly higher mean/median than the black one).

Proportion of Non-retail Business Acres X_3

The PCP (in Fig. 1.34) as well as the scatterplot of X_3 vs. X_{14} shows an obvious negative relation between X_3 and X_{14}. The relationship between the logarithms of both variables seems to be almost linear. This negative relation might be explained by the fact that non-retail business sometimes causes annoying sounds and other pollution. Therefore, it seems reasonable to use X_3 as an explanatory variable for the prediction of X_{14} in a linear-regression analysis.

As far as the distribution of X_3 is concerned, it can be said that the KDE of X_3 clearly has two peaks, which indicates that there are two sub-groups. According to the negative relation between X_3 and X_{14} it could be the case that one sub-group corresponds to the more expensive houses and the other one to the cheaper houses.

Charles River Dummy Variable X_4

The observation made from the PCP that there are more expensive houses than cheap houses situated on the banks of the Charles River is confirmed by inspecting the scatterplot matrix. Still, we might have some doubt that proximity to the river influences house prices. Looking at the original data set, it becomes clear that the

observations for which X_4 equals one are districts that are close to each other. Apparently, the Charles River does not flow through very many different districts. Thus, it may be pure coincidence that the more expensive districts are close to the Charles River—their high values might be caused by many other factors such as the pupil/teacher ratio or the proportion of non-retail business acres.

Nitric Oxides Concentration X_5

The scatterplot of X_5 vs. X_{14} and the separate boxplots of X_5 for more and less expensive houses reveal a clear negative relation between the two variables. As it was the main aim of the authors of the original study to determine whether pollution had an influence on housing prices, it should be considered very carefully whether X_5 can serve as an explanatory variable for price X_{14}. A possible reason against it being an explanatory variable is that people might not like to live in areas where the emissions of nitric oxides are high. Nitric oxides are emitted mainly by automobiles, by factories and from heating private homes. However, as one can imagine there are many good reasons besides nitric oxides not to live in urban or industrial areas. Noise pollution, for example, might be a much better explanatory variable for the price of housing units. As the emission of nitric oxides is usually accompanied by noise pollution, using X_5 as an explanatory variable for X_{14} might lead to the false conclusion that people run away from nitric oxides, whereas in reality it is noise pollution that they are trying to escape.

Average Number of Rooms per Dwelling X_6

The number of rooms per dwelling is a possible measure of the size of the houses. Thus we expect X_6 to be strongly correlated with X_{14} (the houses' median price). Indeed—apart from some outliers—the scatterplot of X_6 vs. X_{14} shows a point cloud which is clearly upward-sloping and which seems to be a realisation of a linear dependence of X_{14} on X_6. The two boxplots of X_6 confirm this notion by showing that the quartiles, the mean and the median are all much higher for the red than for the black boxplot.

Proportion of Owner-Occupied Units Built Prior to 1940 X_7

There is no clear connection visible between X_7 and X_{14}. There could be a weak negative correlation between the two variables, since the (red) boxplot of X_7 for the districts whose price is above the median price indicates a lower mean and median than the (black) boxplot for the district whose price is below the median price. The fact that the correlation is not so clear could be explained by two opposing effects. On the one hand, house prices should decrease if the older houses are not in a good shape. On the other hand, prices could increase, because people often like older

houses better than newer houses, preferring their atmosphere of space and tradition. Nevertheless, it seems reasonable that the age of the houses has an influence on their price X_{14}.

Raising X_7 to the power of 2.5 reveals again that the data set might consist of two sub-groups. But in this case it is not obvious that the sub-groups correspond to more expensive or cheaper houses. One can furthermore observe a negative relation between X_7 and X_8. This could reflect the way the Boston metropolitan area developed over time; the districts with the newer buildings are further away from employment centers and industrial facilities.

Weighted Distance to Five Boston Employment Centers X_8

Since most people like to live close to their place of work, we expect a negative relation between the distances to the employment centers and house prices. The scatterplot hardly reveals any dependence, but the boxplots of X_8 indicate that there might be a slightly positive relation as the red boxplot's median and mean are higher than the black ones. Again, there might be two effects in opposite directions at work here. The first is that living too close to an employment centre might not provide enough shelter from the pollution created there. The second, as mentioned above, is that people do not travel very far to their workplace.

Index of Accessibility to Radial Highways X_9

The first obvious thing one can observe from the scatterplots, as well in the histograms and the KDEs, is that there are two sub-groups of districts containing X_9 values which are close to the respective group's mean. The scatterplots deliver no hint as to what might explain the occurrence of these two sub-groups. The boxplots indicate that for the cheaper and for the more expensive houses the average of X_9 is almost the same.

Full-Value Property Tax X_{10}

X_{10} shows behaviour similar to that of X_9: two sub-groups exist. A downward-sloping curve seems to underlie the relation of X_{10} and X_{14}. This is confirmed by the two boxplots drawn for X_{10}: the red one has a lower mean and median than the black one.

Pupil/Teacher Ratio X_{11}

The red and black boxplots of X_{11} indicate a negative relation between X_{11} and X_{14}. This is confirmed by inspection of the scatterplot of X_{11} vs. X_{14}: The point cloud is

downward sloping, i.e. the less teachers there are per pupil, the less people pay on median for their dwellings.

Proportion of African-American B, $X_{12} = 1000(B - 0.63)^2 I(B < 0.63)$

Interestingly, X_{12} is negatively—though not linearly—correlated with X_3, X_7 and X_{11}, whereas it is positively related with X_{14}. Looking at the data set reveals that for almost all districts X_{12} takes on a value around 390. Since B cannot be larger than 0.63, such values can only be caused by B close to zero. Therefore, the higher X_{12} is, the lower the actual proportion of African-Americans is. Among observations 405–470 there are quite a few that have a X_{12} that is much lower than 390. This means that in these districts the proportion of African-Americans is above zero. We can observe two clusters of points in the scatterplots of $\log(X_{12})$: one cluster for which X_{12} is close to 390 and a second one for which X_{12} is between 3 and 100. When X_{12} is positively related with another variable, the actual proportion of African-Americans is negatively correlated with this variable and vice versa. This means that African-Americans live in areas where there is a high proportion of non-retail business land, where there are older houses and where there is a high (i.e. bad) pupil/teacher ratio. It can be observed that districts with housing prices above the median can only be found where the proportion of African-Americans is virtually zero.

Proportion of Lower Status of the Population X_{13}

Of all the variables X_{13} exhibits the clearest negative relation with X_{14}—hardly any outliers show up. Taking the square root of X_{13} and the logarithm of X_{14} transforms the relation into a linear one.

Transformations

Since most of the variables exhibit an asymmetry with a higher density on the left-hand side, the following transformations are proposed:

$$
\begin{aligned}
\widetilde{X_1} &= \log(X_1) \\
\widetilde{X_2} &= X_2/10 \\
\widetilde{X_3} &= \log(X_3) \\
\widetilde{X_4} &\quad \text{none, since } X_4 \text{ is binary} \\
\widetilde{X_5} &= \log(X_5)
\end{aligned}
$$

$$\widetilde{X_6} = \log(X_6)$$
$$\widetilde{X_7} = {X_7}^{2.5}/10000$$
$$\widetilde{X_8} = \log(X_8)$$
$$\widetilde{X_9} = \log(X_9)$$
$$\widetilde{X_{10}} = \log(X_{10})$$
$$\widetilde{X_{11}} = \exp(0.4 \times X_{11})/1000$$
$$\widetilde{X_{12}} = X_{12}/100$$
$$\widetilde{X_{13}} = \sqrt{X_{13}}$$
$$\widetilde{X_{14}} = \log(X_{14})$$

Taking the logarithm or raising the variables to the power of something smaller than one helps to reduce the asymmetry. This is due to the fact that lower values move further away from each other, whereas the distance between greater values is reduced by these transformations.

Figure 1.37 displays boxplots for the original mean variance scaled variables as well as for the proposed transformed variables. The transformed variables' boxplots are more symmetric and have less outliers than the original variables' boxplots.

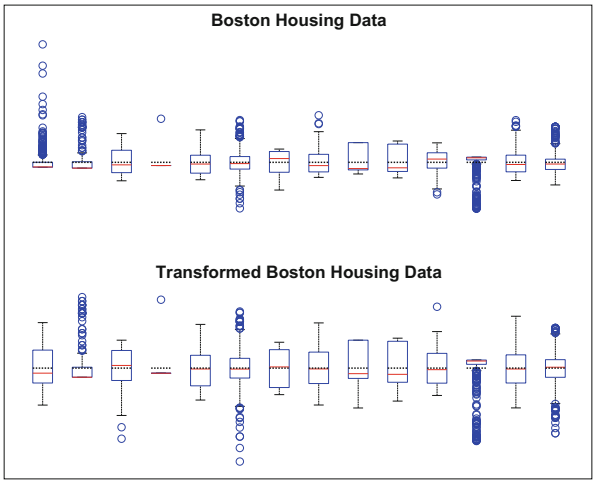

Fig. 1.37 Boxplots for all of the variables from the Boston housing data before and after the proposed transformations
MVAboxbhd

1.10 Exercises

Exercise 1.1 *Is the upper extreme always an outlier?*

Exercise 1.2 *Is it possible for the mean or the median to lie outside of the fourths or even outside of the outside bars?*

Exercise 1.3 *Assume that the data are normally distributed* $N(0,1)$. *What percentage of the data do you expect to lie outside the outside bars?*

Exercise 1.4 *What percentage of the data do you expect to lie outside the outside bars if we assume that the data are normally distributed* $N(0,\sigma^2)$ *with unknown variance* σ^2*?*

Exercise 1.5 *How would the five-number summary of the 15 largest US cities differ from that of the 50 largest US cities? How would the five-number summary of 15 observations of* $N(0,1)$*-distributed data differ from that of 50 observations from the same distribution?*

Exercise 1.6 *Is it possible that all five numbers of the five-number summary could be equal? If so, under what conditions?*

Exercise 1.7 *Suppose we have* 50 *observations of* $X \sim N(0,1)$ *and another* 50 *observations of* $Y \sim N(2,1)$. *What would the* 100 *Flury faces look like if you had defined as face elements the face line and the darkness of hair? Do you expect any similar faces? How many faces do you think should look like observations of* Y *even though they are* X *observations?*

Exercise 1.8 *Draw a histogram for the mileage variable of the car data (Sect. 22.3). Do the same for the three groups (USA, Japan, and Europe). Do you obtain a similar conclusion as in the parallel boxplot in Fig. 1.3 for these data?*

Exercise 1.9 *Use some bandwidth selection criterion to calculate the optimally chosen bandwidth h for the diagonal variable of the bank notes. Would it be better to have one bandwidth for the two groups?*

Exercise 1.10 *In Fig. 1.9 the densities overlap in the region of diagonal* ≈ 140.4. *We partially observed this in the boxplot of Fig. 1.4. Our aim is to separate the two groups. Will we be able to do this effectively on the basis of this diagonal variable alone?*

Exercise 1.11 *Draw a parallel coordinates plot for the car data.*

Exercise 1.12 *How would you identify discrete variables (variables with only a limited number of possible outcomes) on a parallel coordinates plot?*

Exercise 1.13 *True or false: the height of the bars of a histogram are equal to the relative frequency with which observations fall into the respective bins.*

Exercise 1.14 *True or false: kernel density estimates must always take on a value between 0 and 1. (Hint: Which quantity connected with the density function has to*

be equal to 1? Does this property imply that the density function has to always be less than 1?)

Exercise 1.15 *Let the following data set represent the heights of 13 students taking the Applied Multivariate Statistical Analysis course:*

$$1.72, 1.83, 1.74, 1.79, 1.94, 1.81, 1.66, 1.60, 1.78, 1.77, 1.85, 1.70, 1.76.$$

1. *Find the corresponding five-number summary.*
2. *Construct the boxplot.*
3. *Draw a histogram for this data set.*

Exercise 1.16 *Describe the unemployment data (see Table 22.19) that contain unemployment rates of all German Federal States using various descriptive techniques.*

Exercise 1.17 *Using yearly population data (see Sect. 22.20), generate*

1. *a boxplot (choose one of variables)*
2. *an Andrew's Curve (choose ten data points)*
3. *a scatterplot*
4. *a histogram (choose one of the variables)*

What do these graphs tell you about the data and their structure?

Exercise 1.18 *Make a draftman plot for the car data with the variables*

$$X_1 = price,$$
$$X_2 = mileage,$$
$$X_8 = weight,$$
$$X_9 = length.$$

Move the brush into the region of heavy cars. What can you say about price, mileage and length? Move the brush onto high fuel economy. Mark the Japanese, European and American cars. You should find the same condition as in boxplot Fig. 1.3.

Exercise 1.19 *What is the form of a scatterplot of two independent random variables X_1 and X_2 with standard normal distribution?*

Exercise 1.20 *Rotate a three-dimensional standard normal point cloud in 3D space. Does it "almost look the same from all sides"? Can you explain why or why not?*

Exercise 1.21 *There are many reasons for using hexagons to visualise the structure of data.*

1. *Hexagons have the property of "symmetry of nearest neighbours" which lacks in square bins.*

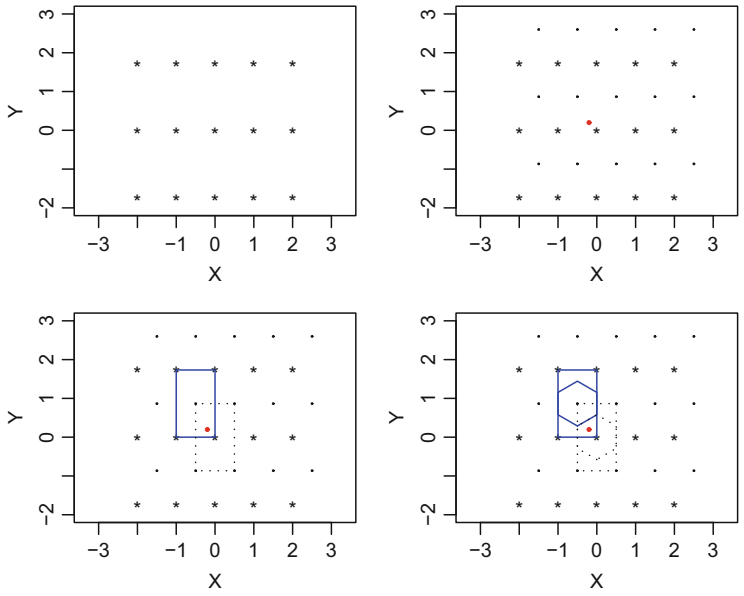

Fig. 1.38 Hexagon binning algorithm `MVAhexaAl`

2. *Hexagons have the maximum number of sides that a polygon can have for a regular tessellation of the plane.*
3. *Hexagons are visually less biased for displaying densities than other regular tessellations.*

The hexagon binning algorithm is as follows:

1. *Decrease y-axis variable by a factor of* $\sqrt{3}$ *(making the calculation more quickly)*
2. *Create a dual lattice (circle and star lines in Fig. 1.38)*
3. *Bin each point into a pair of near neighbour rectangles*
4. *Choose the closest of the rectangle centers (adjusting for* $\sqrt{3}$*)*

The rectangles created from dual lattice have length h_x *(bin width of hexagons) and height* $h_y = \sqrt{3}h_x$*. From these rectangles we can get hexagons with bin width* h_x*. The first point of the star lattice has coordinates* x_0 *and* y_0*. The other star points will have coordinates* $x_0 + k_1 h_x$ *and* $y_0 + l_1 h_y$*, where* $k_1, l_1 = 1, 2, \ldots$ *The first point of the circle lattice has coordinates* $x_0 + \frac{h_x}{2}$ *and* $y_0 + \frac{\sqrt{3}h_x}{2}$*. Other circle points are calculated like star points. Suppose an arbitrary point with coordinates* x, y *lies in the intersection of two near neighbour rectangles. What's the distance from this point to one of two corners?*