Chapter 1. Comparison of Batches

Applied Multivariate Statistical Analysis

Course Instructor: Shuen-Lin Jeng

Department of Statistics
National Cheng Kung University

January, 10, 2021

Introduction

Multivariate statistical analysis is concerned with analyzing and understanding data in high dimensions.

We suppose that each observation x_i has p dimensions:

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

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.

Introduction

- Are there components of X that are more spread out than others?
- Are there some elements of X that indicate subgroups 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" behavior?

Introduction

A qualitative jump in presentation difficulties occurs for dimensions greater than or equal to 5, unless high-dimensional structure can be mapped into lower dimensional components.

A boxplot is a simple univariate device that detects outliers component by component and that can compare distributions of the data among different groups.

Introduction

Two basic techniques for estimating densities are also presented: histograms and kernel densities.

Finally, scatterplots are shown to be very useful for plotting bivariate or trivariate variables against each others: 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.

1.1 Boxplots

Example 1.1 The Swiss bank data(see Appendix ,Sect. B.2) consists of 200 measurements on Swiss banknotes. The first half of these measurements are from genuine banknotes, the other half are from counterfeit banknotes.

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.

1.1 Boxplots

Fig. 1.1 An old Swiss 1000-franc bank note

Figure1.1 An old Swiss 1000-franc bank note

The aim is to study how these measurements may be used in determining whether a bill is genuine or counterfeit.

1.1 Boxplots

Five-Number Summary, we calculate the upper quartile F_U , the lower quartile F_L , the meidan, and the extremes.

The quartiles cut the set into four equal parts, which are often called *fourths* (that is why we use the letter F). 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 medians is $\frac{n+1}{2}$. If n is even, $\frac{n+1}{2}$ is a fraction.

1.1 Boxplots

Take the depth of the median and calculate

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

with $[z]$ denoting the largest integer smaller than or equal to z.
The *F-spread* , d_F , is defined as $d_F = F_U$ - F_L. The *outside bars* is

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

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

The minimum and the maximum are called the *extremes* .

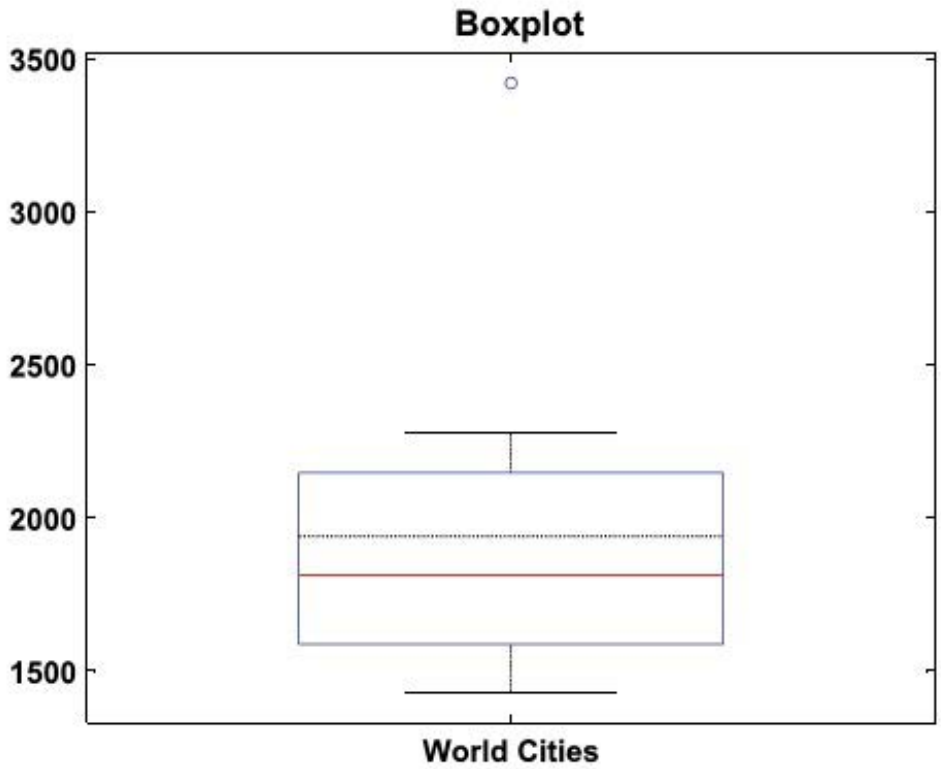
1.1 Boxplots

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 "$\star$" or "$\bullet$" depending on whether they are outside of $F_{UL} \pm 1.5 d_F$ or $F_{UL} \pm 3 d_F$ respectively (this feather is not contained in some software).Label them if possible.

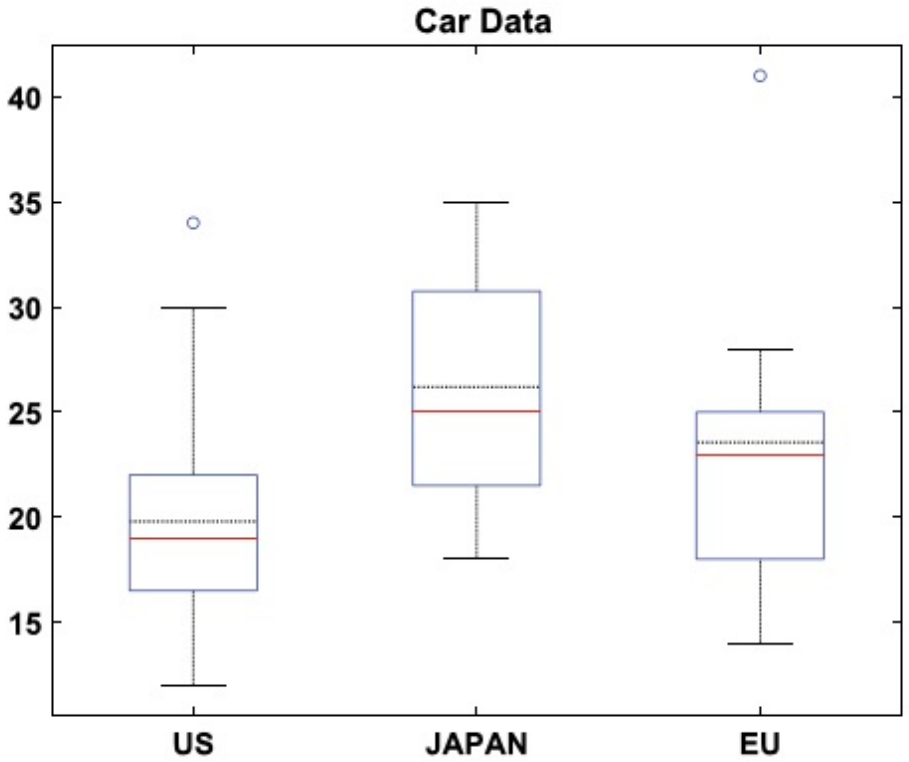
1.1 Boxplots

Fig. 1.2 Boxplot for world cities MVAboxcity



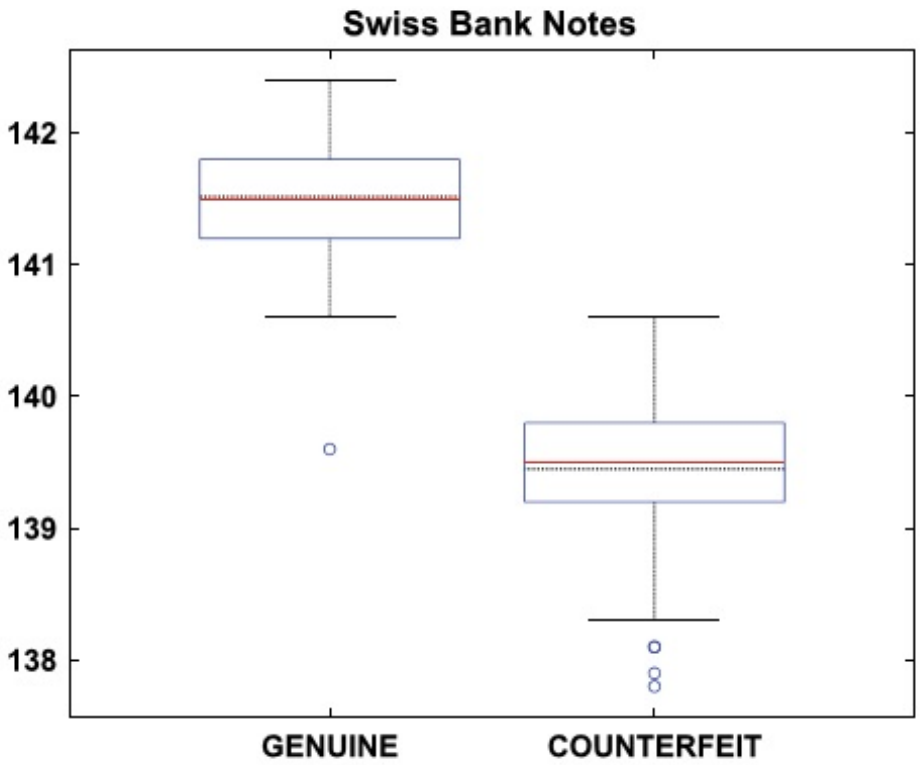
1.1 Boxplots

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



1.1 Boxplots

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



1.2 Histograms

- Histograms are density estimates.
- In contrast to boxplots, density estimates show possible multimodality of the data.
- Let $B_j(x_0, h)$ denote the *bin* of length h , which is the element of a bin grid starting at x_0 :

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

1.2 Histograms

If $\{x_i\}_{i=1}^n$ is an i.i.d. sample with density f , the histogram is defined as follows:

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

The parameter h is a smoothing or localizing 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.

1.2 Histograms

Using methods from smoothing methodology Härdle et al.(2004), one can find an 'optional' bin-width h for n observations:

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

Unfortunately, the bin-width 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 bin-width h—the interpretation is not facilitated. The shift of the origin x_0 (to 4 different locations) created 4 different histograms.

1.2 Histograms

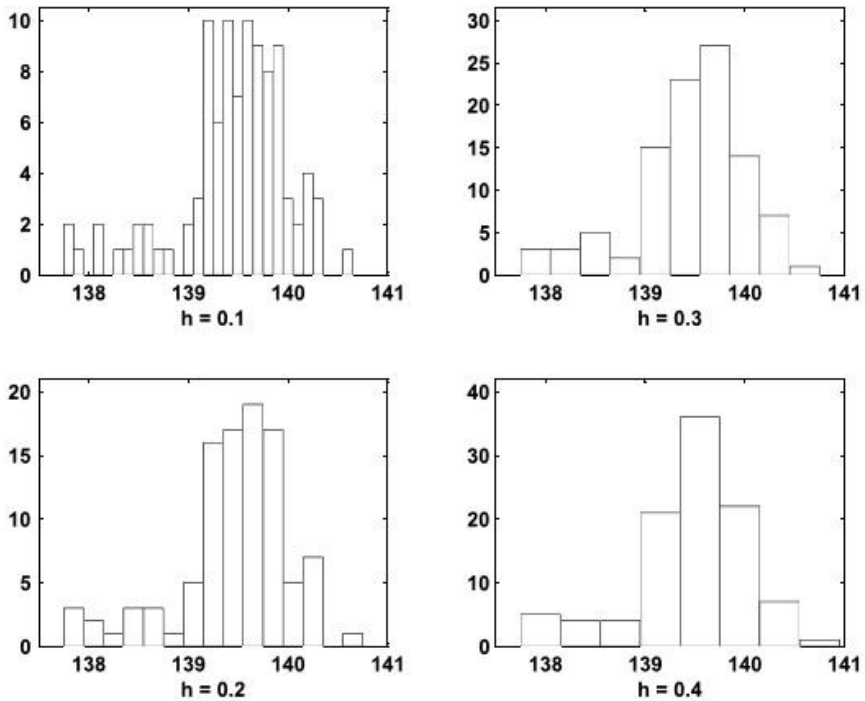


Fig. 1.6 Diagonal of counterfeit banknotes. Histograms with $x_0 = 137.8$ and $h = 0.1$ (upper left), $h = 0.2$ (lower left), $h = 0.3$ (upper right), $h = 0.4$ (lower right) MVAhisbank1

A remedy has been proposed by Scott(1985):"Average the shifted histograms!". The result is presented in Fig.1.8.

1.2 Histograms

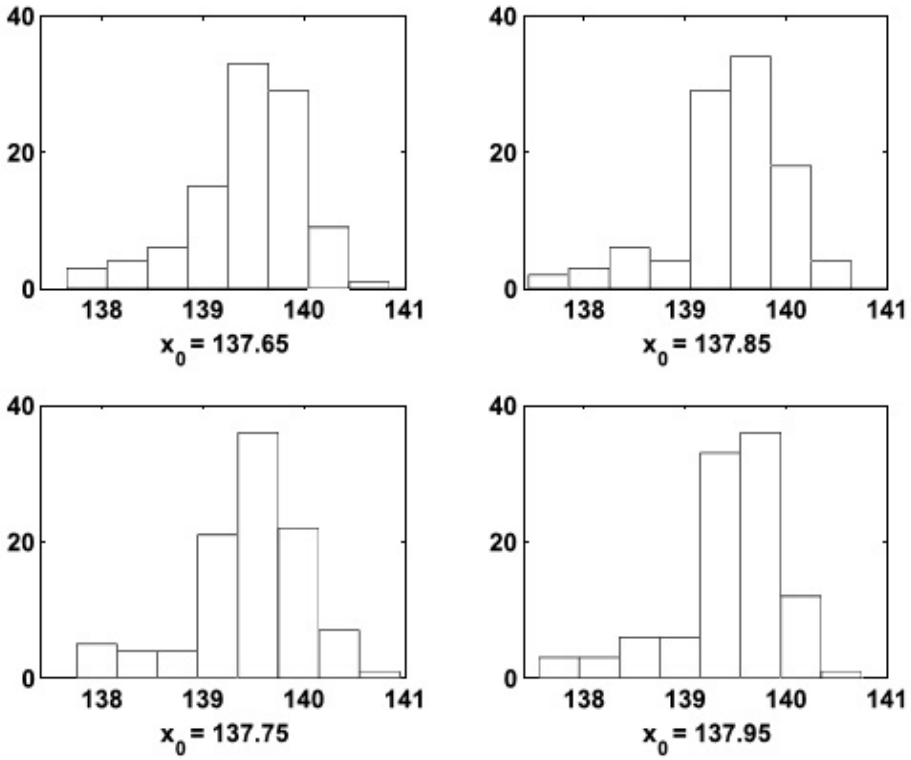


Fig. 1.7 Diagonal of counterfeit banknotes. Histogram with $h = 0.4$ and origins $x_0 = 137.65$ (upper left), $x_0 = 137.75$ (lower left), $x_0 = 137.85$ (upper right), $x_0 = 137.95$ (lower right) MVAhisbank2

1.3 Kernel Densities

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

- determination of the bin-width h
- choice of the bin origin x_0
- loss of information since observations are replaced by the central point
- the underlying density function is often assumed to be smooth, but the histogram is not smooth.

The histogram can in fact be written as

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

1.3 Kernel Densities

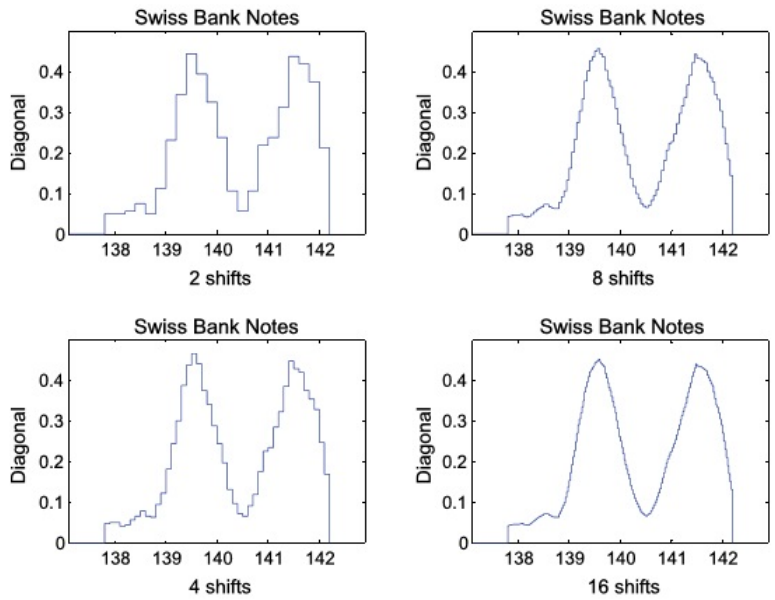


Fig. 1.8 Swiss banknotes: there are 2 shifts (upper left), 4 shifts (lower left), 8 shifts (upper right), and 16 shifts (lower right) MVAashbank

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

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

1.3 Kernel Densities

Table 1.5 Kernel functions

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

Averaging these squared deviations over a grid of $\{x_i\}_{l=1}^{L}$ leads to

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

1.3 Kernel Densities

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

$$h_G = 1.06\hat{\sigma} n^{-1.5} \tag{1.10}$$

where $\hat{\sigma} = \sqrt{n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

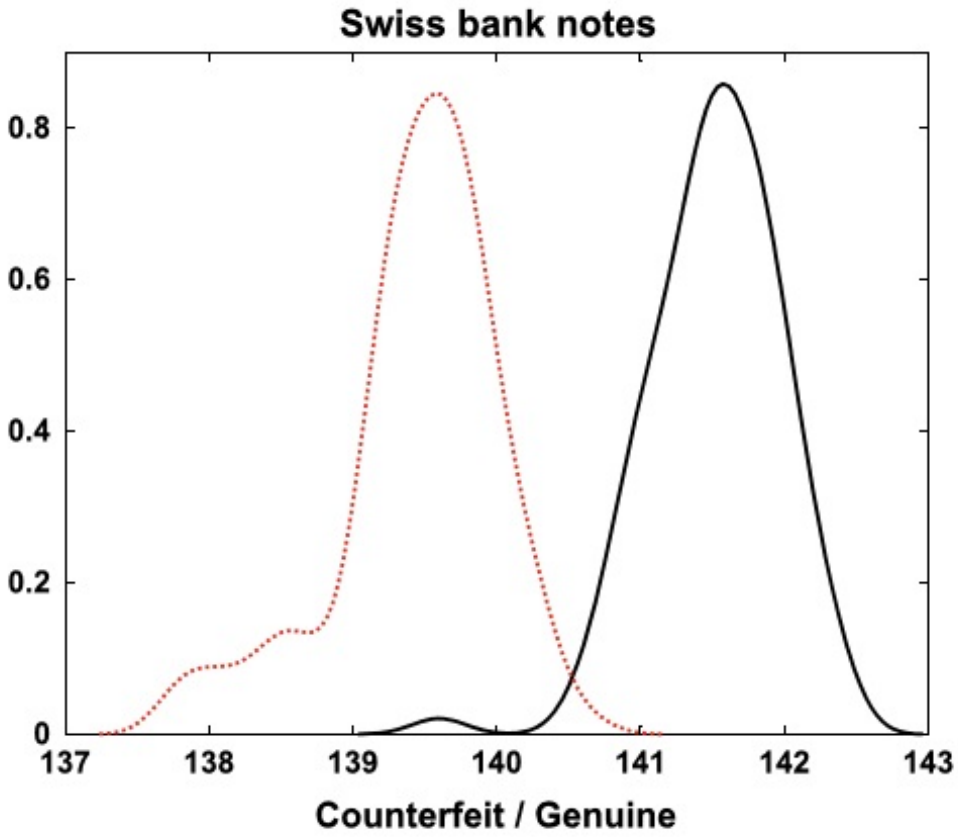
For the quartic kernel, we need to transform (1.10). The modified rule of thumb is

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

1.3 Kernel Densities

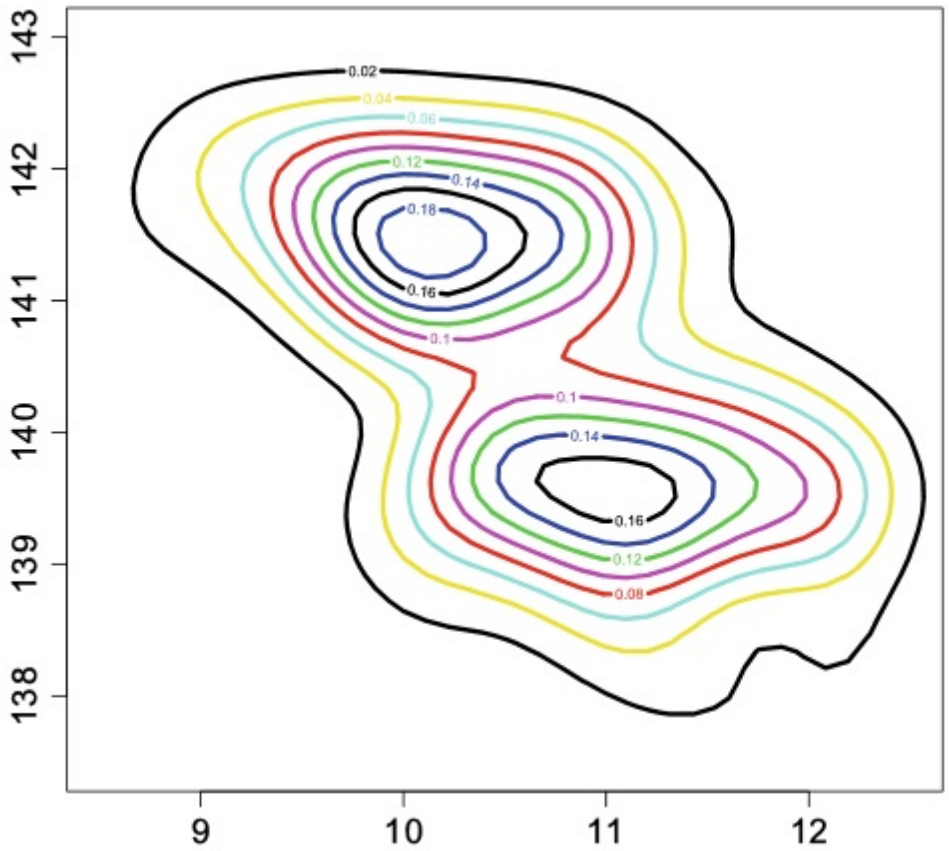
Fig. 1.9 Densities of the diagonals of genuine and counterfeit banknotes. Automatic density estimates

MVAdenbank



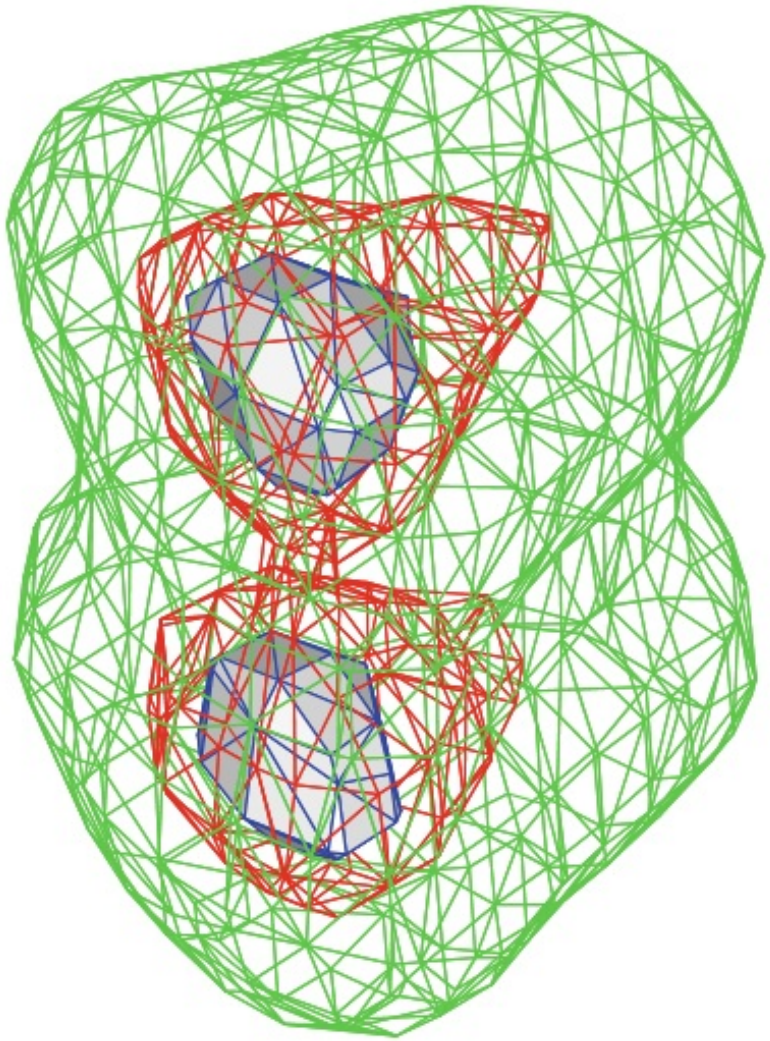
1.3 Kernel Densities

Fig. 1.10 Contours of the density of X_5 and X_6 of genuine and counterfeit banknotes MVAcontbank2



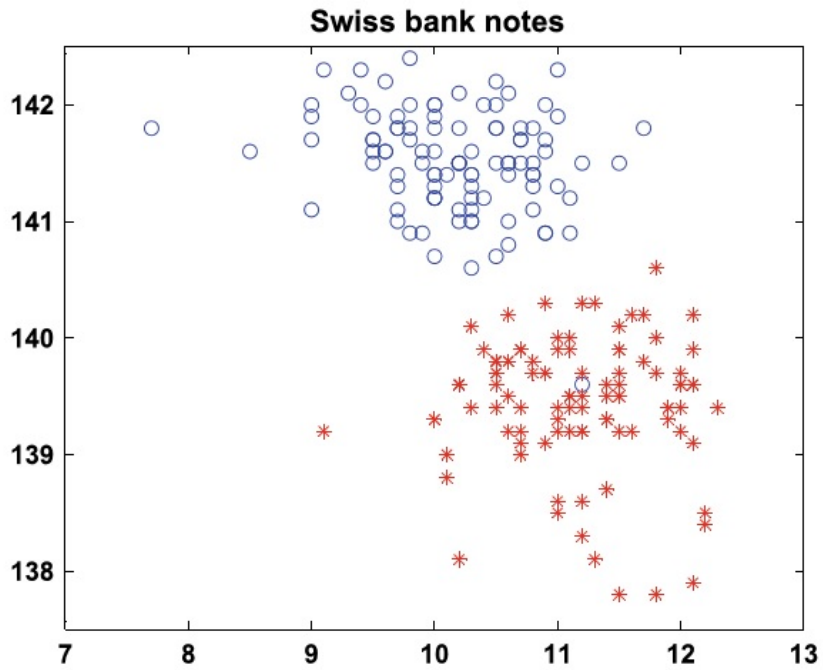
1.3 Kernel Densities

Fig. 1.11 Contours of the density of X_4, X_5, X_6 of genuine and counterfeit banknotes MVAcontbank3



1.4 Scatterplots

Fig. 1.12 2D scatterplot for X_5 versus X_6 of the banknotes. Genuine notes are circles, counterfeit notes are stars MVAscabank56



1.4 Scatterplots

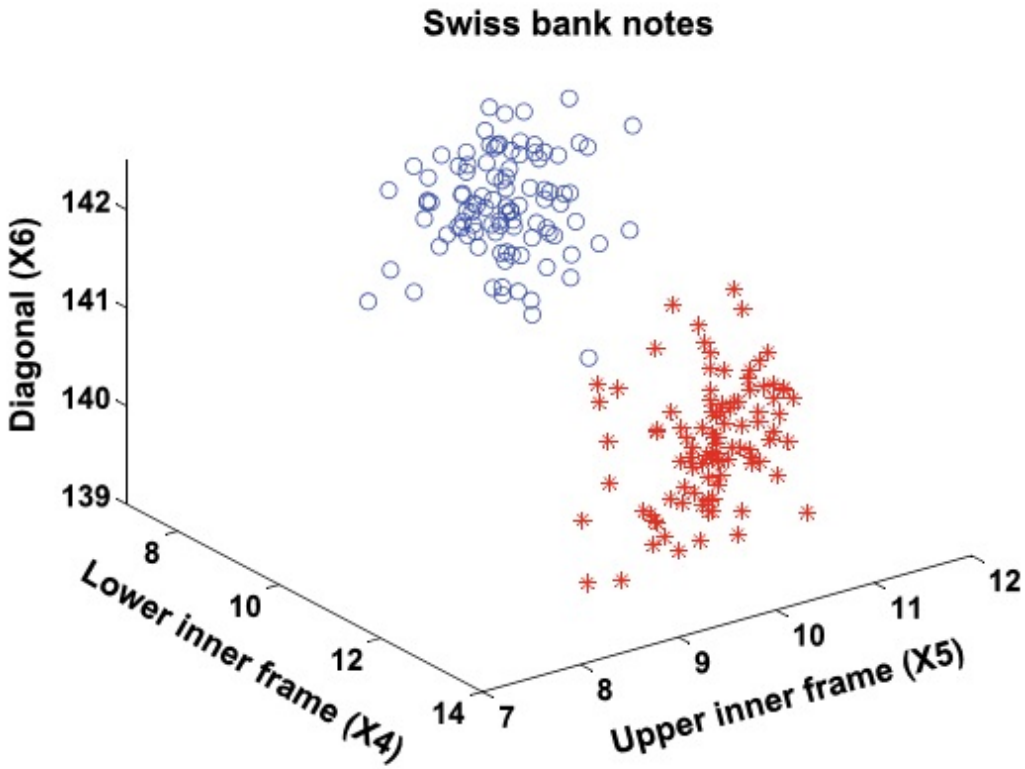


Fig. 1.13 3D Scatterplot of the banknotes for (X_4, X_5, X_6). Genuine notes are circles, counterfeit are stars MVAscabank456

1.4 Scatterplots

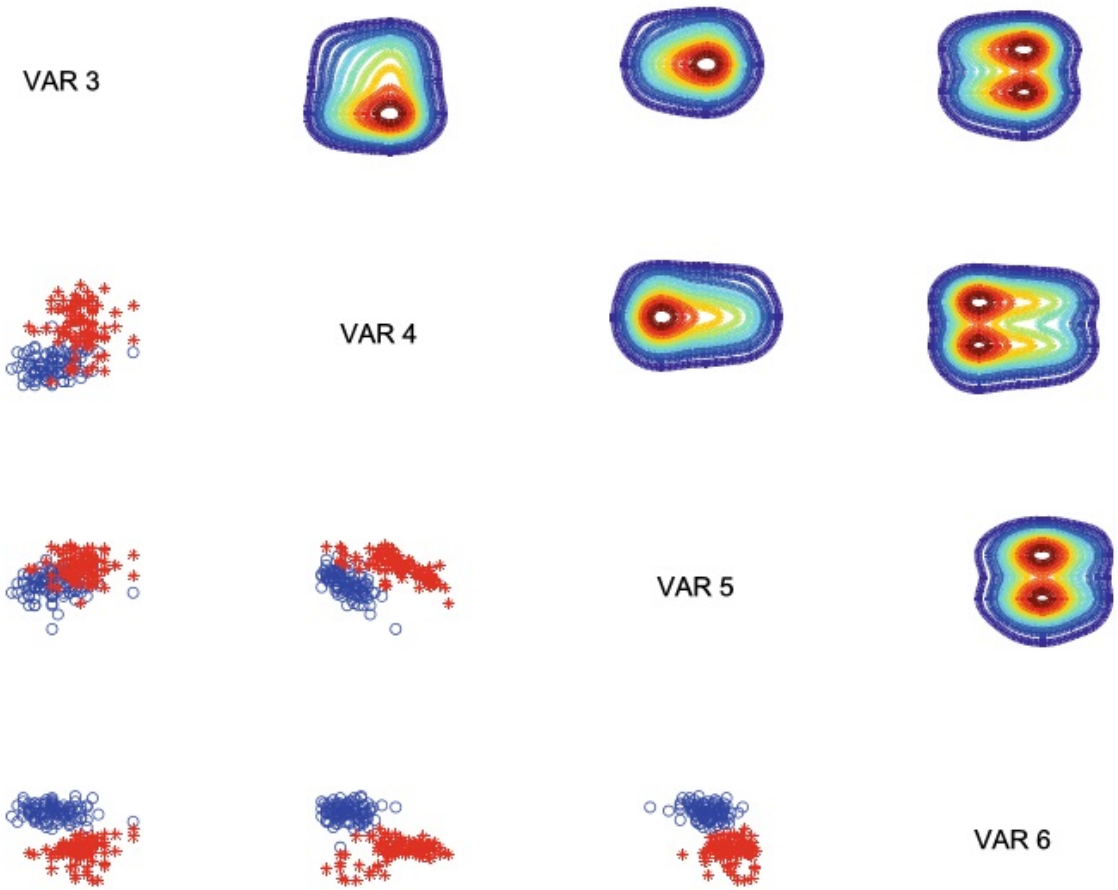


Fig. 1.14 Draftsman's plot of the banknotes. The pictures in the left-hand column show (X_3, X_4), (X_3, X_5) and (X_3, X_6), in the middle we have (X_4, X_5) and (X_4, X_6), and in the lower right (X_5, X_6). The upper right half contains the corresponding density contour plots MVAdrafbank4

1.4 Scatterplots

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 color observations.

By moving the brush, we can study conditional dependence.

1.5 Chernoff-Flury faces

The Chernoff-Flury faces, for example, provide such a condensation of high-dimensional information into a simple "face".

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 right eye
7. curvature of right eyebrow
8. density of right eyebrow

1.5 Chernoff-Flury faces

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 the 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.

1.5 Chernoff-Flury faces

First every variable that is to be coded into a characteristic face element is transformed into a (0,1) scale.

$X_1 = 1, 19$ (eye sizes)
$X_2 = 2, 20$ (pupil sizes)
$X_3 = 4, 22$ (eye slants)
$X_4 = 11, 29$ (upper hair lines)
$X_5 = 12, 30$ (lower hair lines)
$X_6 = 13, 14, 31, 32$(face lines and darkness of hair)

Recall that observations 1-100 correspond to the genuine notes, and that observations 101-200 correspond to the counterfeit notes.

1.5 Chernoff-Flury faces

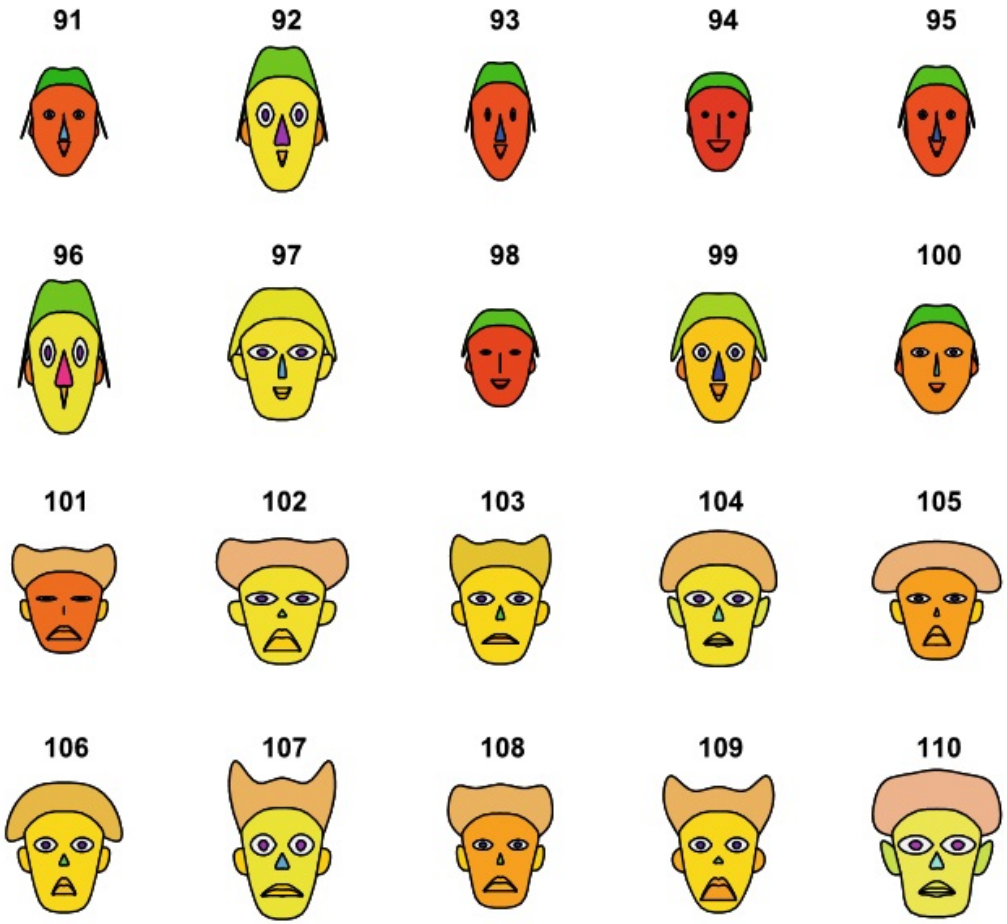


Fig. 1.15 Chernoff-Flury faces for observations 91–110 of the banknotes MVAfacebank10

1.6 Andrews' Curves

Each multivariate observation $X_i = (X_{i,1}, \ldots, X_{i,p})$ transformed into a curve as follows:

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

the observation represents the coefficients of a so-called Fourier series ($t \in [-\pi, \pi]$). $X_1 = (0, 0, 1), X_2 = (1, 0, 0), X_3 = (0, 1, 0)$.

Here p = 3 and the following representations correspond to the Andrews' curves:

$$f_1(t) = \cos(t)$$

$$f_2(t) = \frac{1}{\sqrt{2}} \text{ and}$$

$$f_3(t) = \sin(t).$$

1.6 Andrews' Curves

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

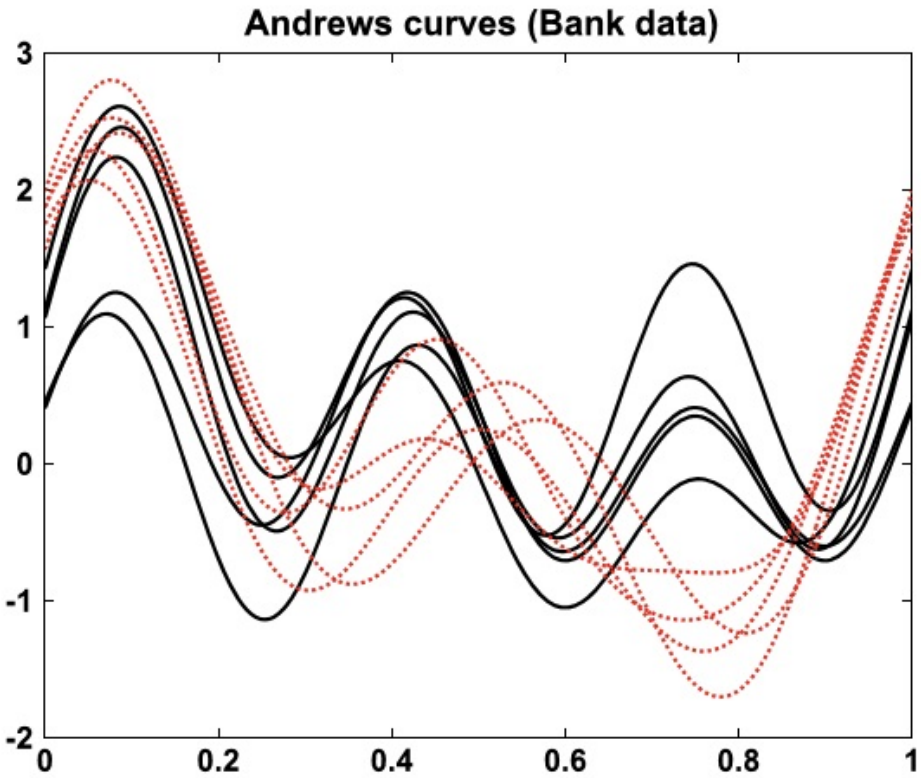
$$X_{96} = (215.6, 129.9, 129.9, 9.0, 9.5, 141.7).$$

The Andrews' curve is by (1.13):

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

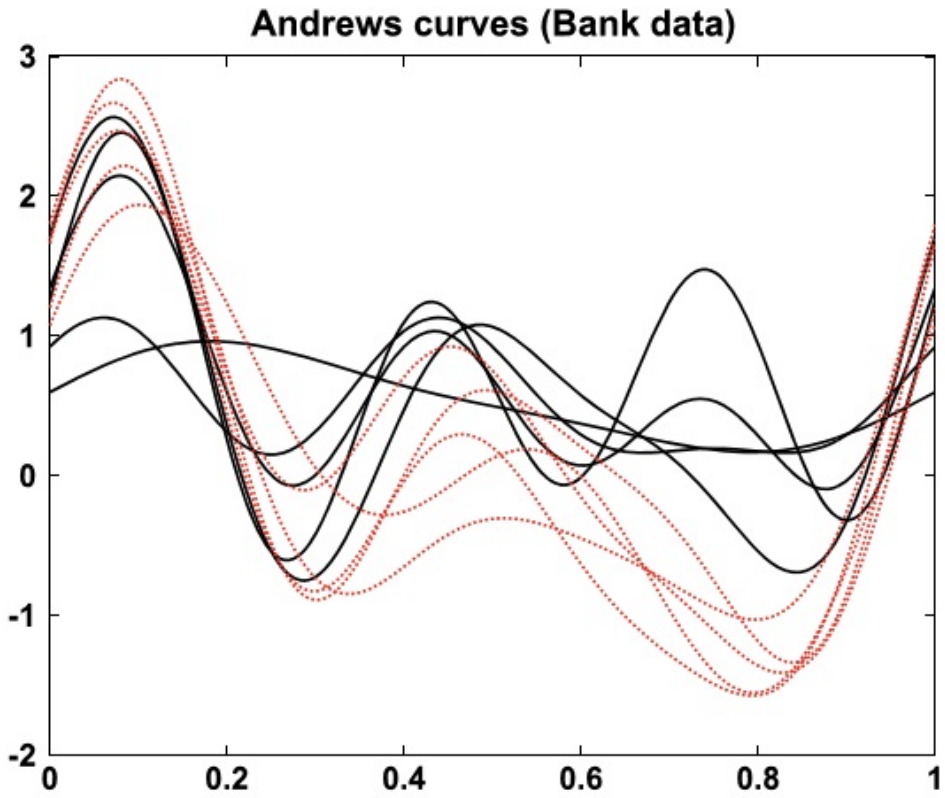
1.6 Andrews' Curves

Fig. 1.18 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



1.6 Andrews' Curves

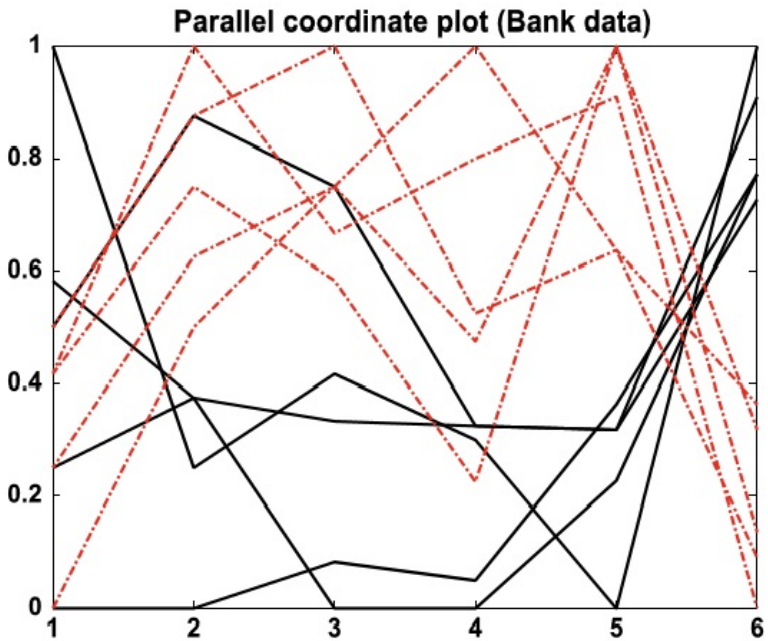
Fig. 1.19 Andrews' curves of the observations 96–105 from the Swiss bank note data.The order of the variables is 6, 5, 4, 3, 2, 1 MVAscabank56



1.7 Parallel Coordinate Plots

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.

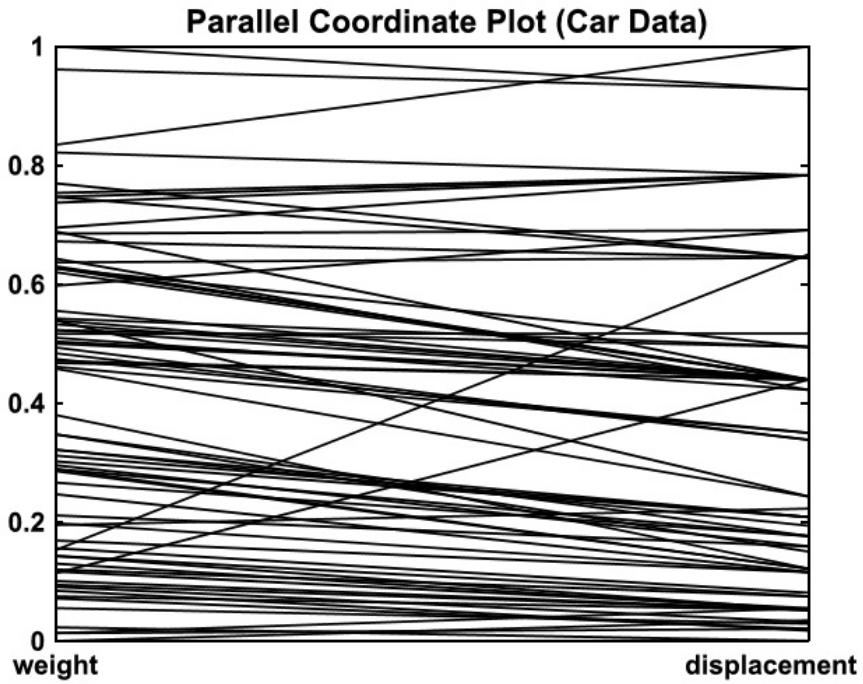
Fig. 1.20 Parallel coordinates plot of observations 96–105 MVAparcoo1



1.7 Parallel Coordinate Plots

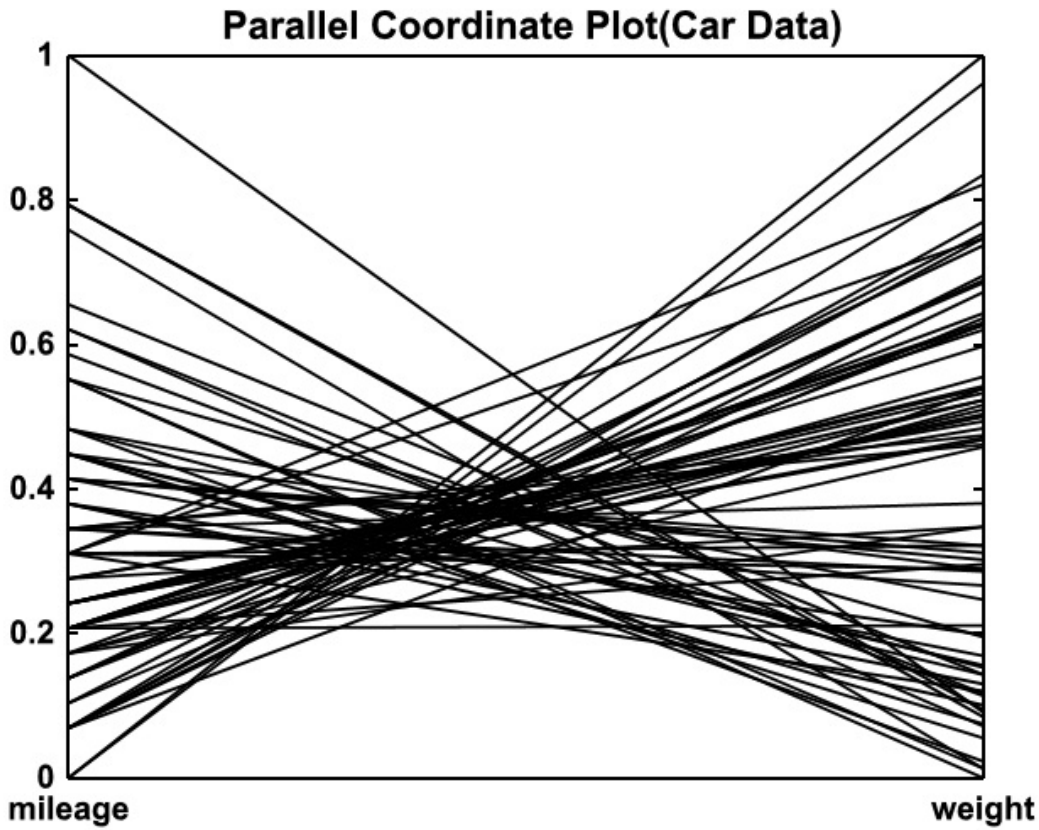
PCP can also be used to for detecting linear dependencies between variables: if all lines are of almost parallel dimensions (p = 2), there is a positive linear dependence between them.

Fig. 1.21 Coordinates Plot indicating strong positive dependence with $\rho = 0.9$, X_1 = weight, X_2 = displacement MVApcp2



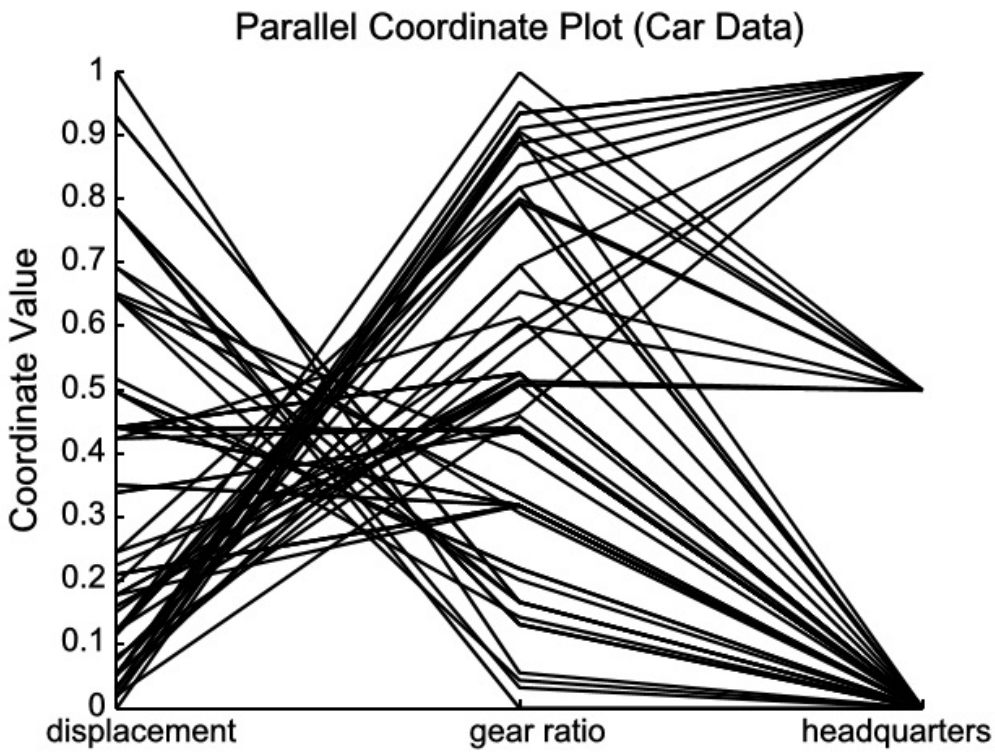
1.7 Parallel Coordinate Plots

Fig. 1.22 Coordinates Plot showing strong negative dependence with $\rho = -0.82$, X_1 = mileage, X_2 = weight MVApcp3



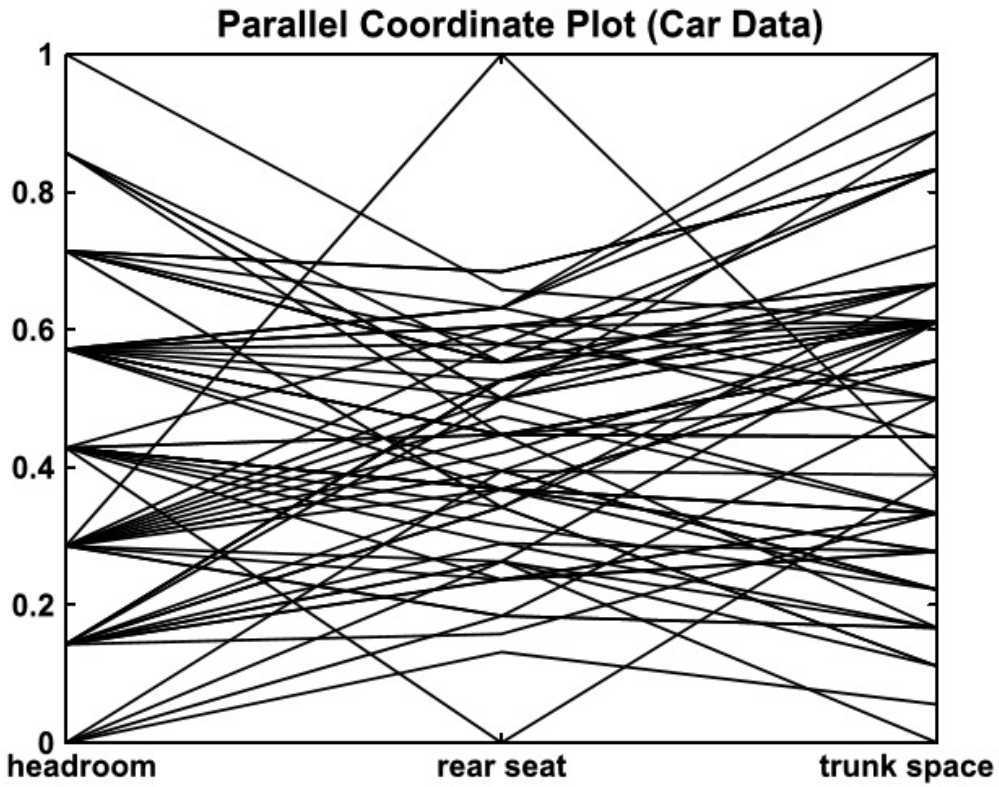
1.7 Parallel Coordinate Plots

Fig. 1.23 Parallel Coordinates Plot with subgroups MVApcp4



1.7 Parallel Coordinate Plots

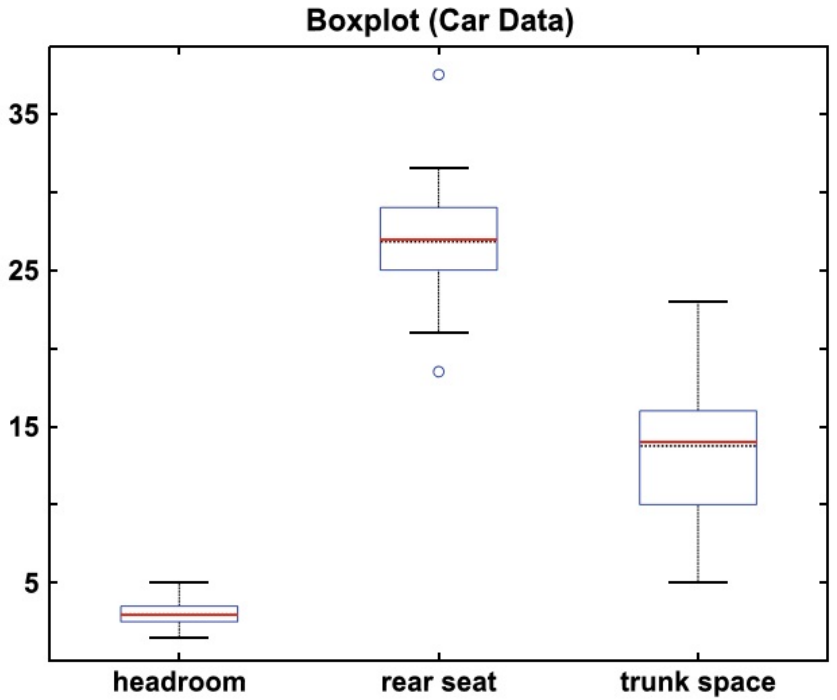
Fig. 1.24 PCP for $X_1 =$ headroom, $X_2 =$ rear seat clearance, and $X_3 =$ trunk space MVApcp5



1.7 Parallel Coordinate Plots

Fig. 1.25 Boxplots for headroom, rear seat clearance, and trunk space

MVApcp6



The real power of PCP comes though through coloring subgroups.

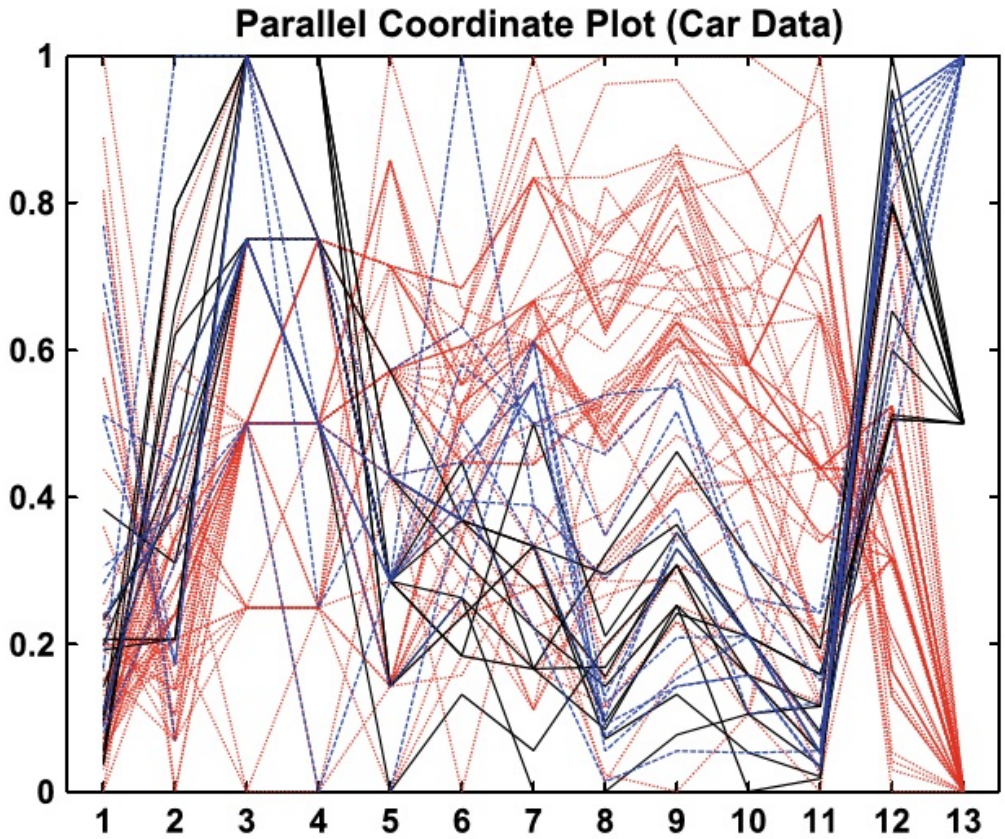
1.7 Parallel Coordinate Plots

Example 1.6 Data in Fig. 1.28 are colored according to X_{13} — car company's headquarters. Red stands for European car, green for Japan, and black for U.S. This PCP with coloring can provide some information for us :

1. U.S. cars (black) tend to have large value in $X_7, X_8, X_9, X_{10}, X_{11}$ (trunk (boot) spaces, weight, length, turning diameter, displacement), which means U.S. 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.

1.7 Parallel Coordinate Plots

Fig. 1.28 Parallel Coordinates Plot for car data
MVApcp1



1.8 Hexagon Plots

It is useful for visualizing the structure of data sets entailing a large number of observations n.

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 color ramp or varying the radius of the hexagon in proportion to the counts.

1.8 Hexagon Plots

The data is taken from ALLBUS(2006) [*ZANo*.3762].The number of respondents is 2946.

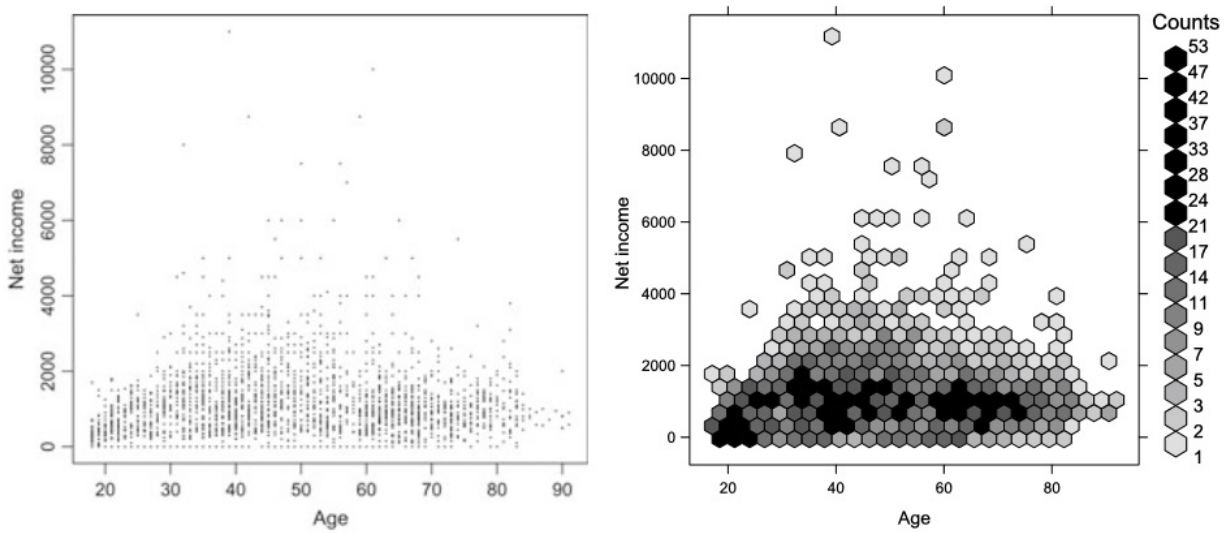


Fig. 1.29 Hexagon plots between X_1 and X_2 MVAageIncome

1.8 Hexagon Plots

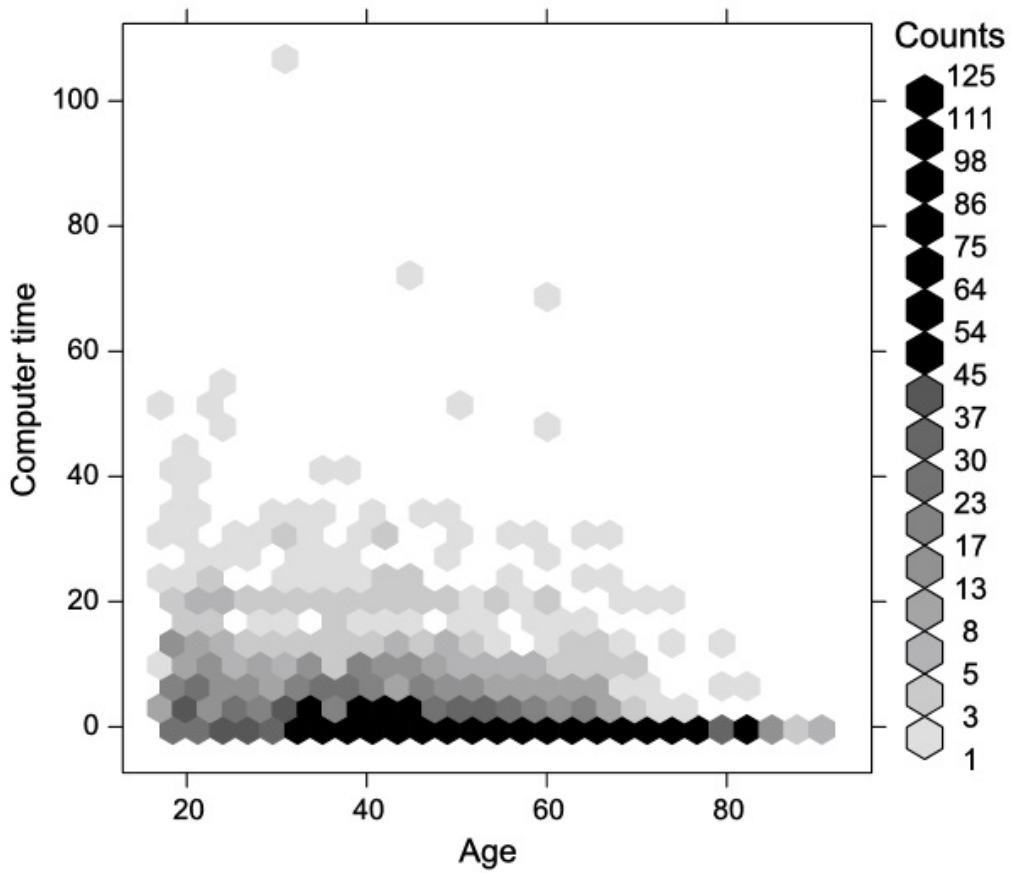


Fig. 1.30 Hexagon plot between X_1 and X_5 MVAageCom

1.8 Hexagon Plots

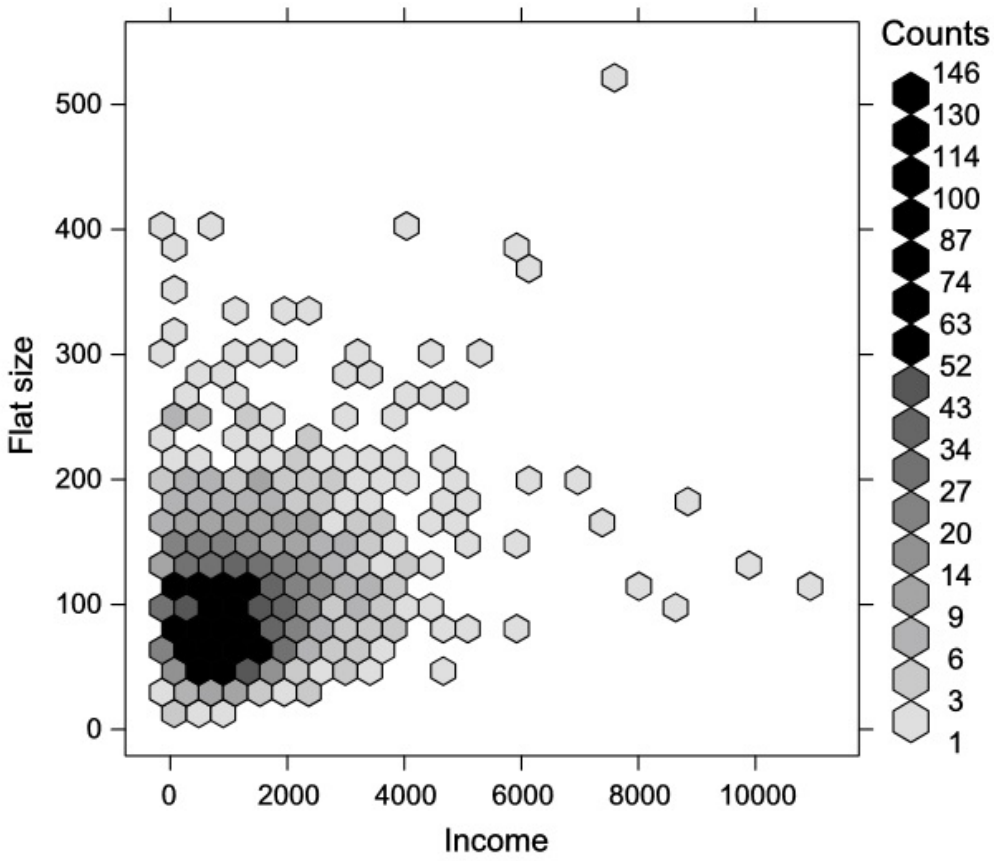
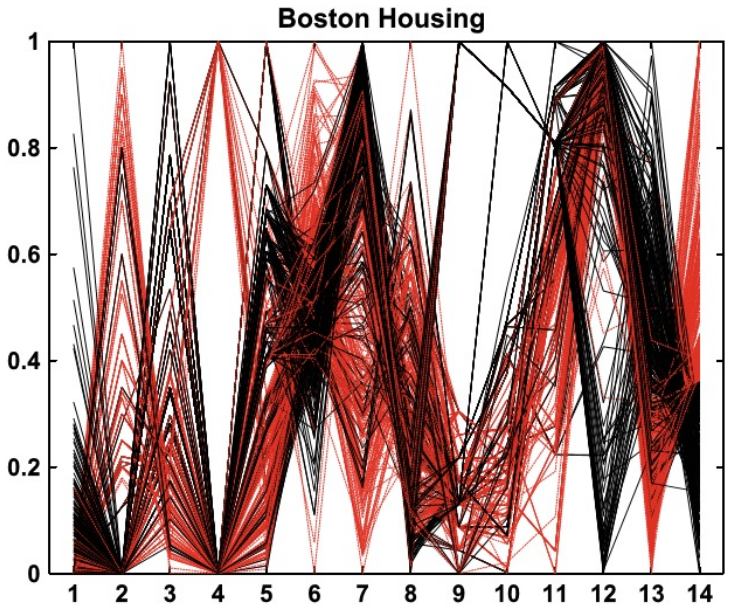


Fig. 1.31 Hexagon plot between X_2 and X_7 MVAincomeLi

1.9 Boston Housing

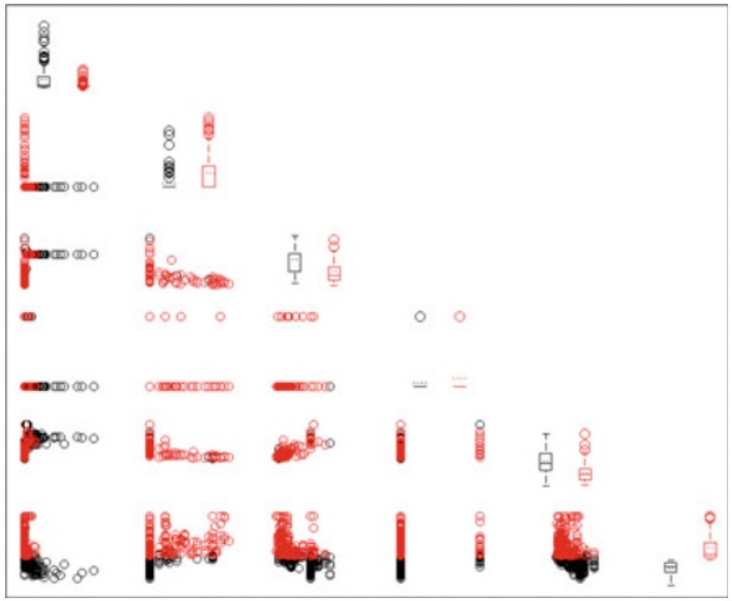
The Boston Housing data set was analyzed by Harrison and Rubinfeld (1978),who wanted to find out whether "clean air" had an influence on house prices. We color all of the observations with $X_{14} >$ median (X_{14}) as red lines in Fig. 1.32.

Fig. 1.32 Parallel coordinates plot for Boston Housing data MVApcphousing



1.9 Boston Housing

Fig. 1.33 Scatterplot matrix for variables $X_1, \ldots, X_5$ and X_{14} of the Boston Housing data MVAdrafthousing



- **Per-capita crime rate X_1**
- **Proportion of residential area zoned for large lots X_2**
- **Proportion of non-retail business acres X_3**
- **Charles River dummy variable X_4**