PART I

VISUALIZATION

COPYRIGHTED MATERIAL

CHAPTER 1

VISUALIZATION OF DATA

Our data are given as an $n \times d$ matrix of real numbers, which is called a data matrix. The rows represent the cases and the columns represent the variables: the number in the ith row and in the jth column gives the measurement of the ith object with respect to the jth variable. A data matrix is often called a "data table" in information visualization. Bertin (1981) calls the cases "objects" and the variables "characteristics". We may interpret the rows of the data matrix as points in the d-dimensional Euclidean space. Thus the elementary data visualization task is to visualize a set $\{x_1, \ldots, x_n\} \subset \mathbf{R}^d$.

Variables come in three basic types: nominal, ordinal, and quantitative. Nominal variables allow only meaningful use of the identity relation, ordinal variables allow the use of order relation, and quantitative variables allow the use of arithmetic on them. We will ignore the important types of nominal and ordinal variables and restrict ourselves to the quantitative variables. In addition we will ignore the crucial phase of transforming a raw data to a data matrix.

One can make a distinction between the raw information that is contained in a visualization and the information that can be extracted with a human perception system. For example, a scatter plot of 2D data and the corresponding data matrix of numbers contain the same raw information (up to rounding errors and overplotting);

Smoothing of Multivariate Data: Density Estimation and Visualization. By Jussi Klemelä
Copyright © 2009 John Wiley & Sons, Inc.

we can make a computer program that maps a scatter plot to a data matrix, and conversely maps a data matrix to a scatter plot. On the other hand, a human can see patterns from a 2D scatter plot, whereas looking at the numbers of a data matrix printed on a sheet of paper does not usually reveal much to a human.

Some visualizations reduce the raw information but increase the information that may be perceived by a human. For example, kernel estimates do not possess the original information; we may not recover the original data from a kernel estimate. The empirical distribution function contains exactly the same information as the original data; we may recover the original data from the distribution function. However, kernel estimates visualize local concentrations of data well, whereas distribution functions are inefficient in visualizing these local concentrations.

Pixel-based visualizations require at least one pixel to represent one observation. When the size of the data grows, at some stage the pixel based visualizations fail because a human perception system cannot separate pixels when the resolution grows too fine. However, when the size of the data is large, one may use smoothing as a visualization tool.

Thus smoothing arises in visualization in two ways: (1) as a tool that leads to enhanced visualizations taking into account the properties of human perception and (2) as a replacement of pixel based visualizations for large data sets.

Smoothing leads to the task of visualization of functions (visualization of density estimates). Visualization of functions turns out to be related with the visualization of trees. Visualization of functions is discussed in Chapter 2 and visualization of trees is discussed in Chapter 3. This chapter gives an overview of some important data visualization methods.

- Section 1.1 discusses scatter plots, projections, dynamic scatter plots, slices, prosections, and subsetting.
- Section 1.2 discusses visualization of univariate data. Covered are line plots, 1D scatter plots, index plots, time series plots, empirical distribution functions, tail plots, PP-plots, QQ-plots, Box-plots, and kernel estimates.
- Section 1.3 discusses parallel level plots. Parallel level plots are defined as a series of 1D scatter plots.
- Section 1.4 discusses graphical matrices. Graphical matrices are $n \times d$ matrices of graphical elements.
- Section 1.5 discusses visualization with graphical elements. Parallel coordinate plots, Andrew's curves, and Chernoff faces are mentioned.
- Section 1.6 discusses linking across dimensions.
- Section 1.7 discusses descriptive statistics that characterize distributions. Mean, median, mode, and dispersion matrices and vectors are defined.
- Section 1.8 discusses dimension reduction. Principal components, projection pursuit, self-organizing maps, and multidimensional scaling are mentioned.

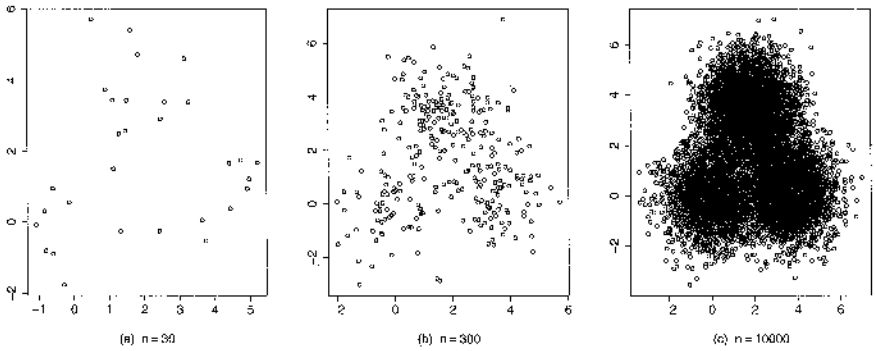

Figure 1.1 *(Scatter plots.)* Showing data of different sample sizes.

1.1 SCATTER PLOTS, PROJECTIONS, AND SLICES

1.1.1 Scatter Plots

The plot of a two-dimensional point cloud $\{x_1, \ldots, x_n\} \subset \mathbf{R}^2$ in the Cartesian coordinate system is called a scatter plot.

Figure 1.1 shows three scatter plots of data of three different sizes, when the data is simulated from the same density. The sample sizes are $n = 30$, $n = 300$, and $n = 10,000$. The data are generated from the three-modal density later shown in Figure 2.7. When the sample size is small, then the point cloud is too sparse to reveal the three modes of the density. When the sample size is large, then the scatter plot is mostly black, so the visuality of density of the points in different regions is obscured. When the sample size is moderate, a more accurate impression of the distribution of the probability mass is possible. Smoothing provides a better way to visualize the density of the observation than a scatter plot; see Figure 1.1.

1.1.2 Projections

When $d \geq 3$, we may project data to the 2D Euclidean space and then visualize with a scatter plot this two-dimensional point cloud. We may choose to look only at the projections on the coordinate planes. A *scatter plot matrix*, or a *draftman's plot*, shows scatter plots of all two-dimensional projections to the coordinate planes. A scatter plot matrix consists of

$$\binom{d}{2} = \frac{d(d-1)}{2}$$

scatter plots. A difficulty with the use of scatter plot matrices comes from the fact that single observations cannot be identified across the different windows. A *brush* is a rectangle that can be moved with the mouse. The observations inside the rectangle are highlighted or colored in every window of the scatter plot matrix.

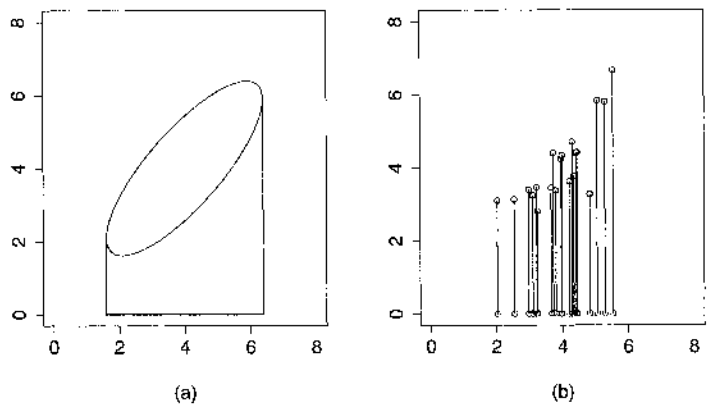

Figure 1.2 *(Projection.)* Shown are one-dimensional projections of two-dimensional sets. (a) Projection to the x_1 axis of an ellipse; (b) projection to the x_1 axis of an ellipsoidal point cloud.

Projections are defined in Appendix B.7. We may also construct a projection by first making a rotation and then projecting to the coordinate planes. See Appendix B.7 for the definition of rotations.

Figure 1.2 illustrates 1D projections of 2D sets. Panel (a) shows how an ellipse is projected on the x_1-axis. Panel (b) shows how an ellipsoidal point cloud is projected on the x_1-axis.

1.1.3 Dynamic Scatter Plots

Dynamic scatter plots show a series of linked 2D projections. Making a series of 2D projections that are close to each other conveys an impression of 3D projections. Smooth transitions between 2D projections make it possible to identify observations in different projections, and brushing may be used to additionally ease the identification of the observations across different projections.

A *grand tour* is a sequence of 2D projections that goes through all possible angles. A dense set of 2D planes is needed so that one can move smoothly through the set; a continuous space-filling path thus moves through the 2D planes in the d-dimensional space. Asimov (1985) introduces the grand tour and Wegman and Solka (2002) discusses algorithms for the grand tour.

1.1.4 Slices

A two-dimensional slice, or section, of a multivariate set $A \subset \mathbf{R}^d$, along the coordinate axes, with the conditions $x_3 = x_{30}, \ldots, x_d = x_{d0}$, is the set

$$\{(x_1, x_2) : (x_1, x_2, x_3, \ldots, x_d) \in A,\ x_3 = x_{30}, \ldots, x_d = x_{d0}\} \subset \mathbf{R}^2.$$

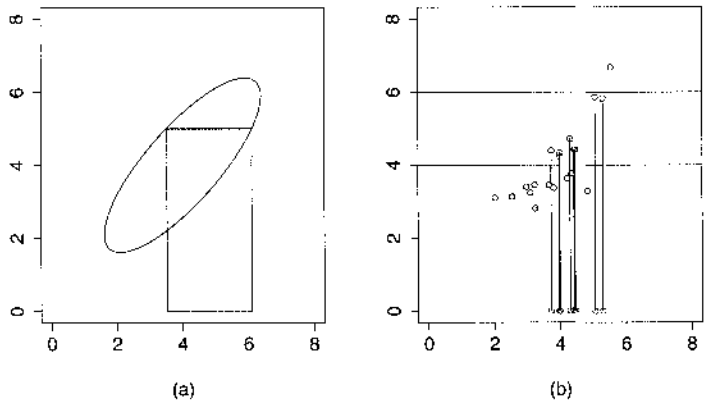

Figure 1.3 *(Slice.)* One-dimensional slices of two-dimensional sets. (a) A slice of an ellipse; (b) a thickened section of an ellipsoidal point cloud.

When set $A \subset \mathbf{R}^d$ is a point cloud, almost all slices are empty. Thus we expand the condition on the coordinates $3, \ldots, d$ to be an inclusion on an interval. This leads to a thickened section

$$\{(x_1, x_2) : (x_1, x_2, x_3, \ldots, x_d) \in A,\ (x_3, \ldots, x_d) \in [a_3, b_3] \times \cdots \times [a_d, b_d]\}\,.$$

A scatter plot of a thickened section shows two of the coordinates when the other coordinate values are restricted to an interval. A scatter plot of a thickened section is called also a *conditioning plot*, or *coplot*. A coplot visualizes conditional dependence of the two variables, conditioning on the chosen interval of the values of the other variables.

Thickened sections may be useful in visualizing dependence, but there exists a huge number of possible slices to consider: for the d-dimensional data one has $d(d-1)/2$ coordinate directions, but in choosing slices, one has to consider also other than the coordinate directions in making 2D projections. Finally, for each 2D projection one needs to make a series of slices.

Figure 1.3 illustrates 1D slices of 2D sets. Panel (a) shows a slice along the line $x_2 = 5$ of an ellipse. Panel (b) shows a thickened section of an ellipsoidal point cloud when the condition is $x_2 \in [4, 6]$.

1.1.5 Prosections

A limitation of projections is that a projection of an object to a low-dimensional space can fill this space. For example, a projection of a plane in 3D space to a 2D space fills the 2D space. More generally, a projection of an m-dimensional plane in the d-dimensional space to the k-dimensional space fills the k-dimensional space when $m \geq k$.

A limitation of slices is that they may be empty with high probability. For example, a slice with a plane in the 3D space will miss a point in the 3D space almost surely. More generally, a slice of an m-dimensional plane in the d-dimensional space to the k-dimensional space misses the plane when $d - m > k$. One calls the number $d - m$ the codimension of the m-dimensional plane.

Thus 2D projections and slices can reveal planes of dimension 0, 1 and $d-2$, $d-1$, d. When $d \leq 4$, all cases are covered. When $d > 4$, a combination of a projection and a slice may be used, and these combinations are called prosections. For example, if we have an m-plane in the d-dimensional space, make an $(d - m + 1)$-dimensional section, and then project the section to 2D space, then we will get a line. This is discussed more in detail by Furnas and Buja (1994).

1.1.6 Subsetting

Scatter plots suffer from overplotting when the number of observations is large; see Figure 1.1 and Figure I.1. The same holds for all the other pixel-based visualization tools (parallel coordinate plots, graphical matrices, tail tree plots, etc.). The human perception system has a limited capability to distinguish objects at fine resolution levels. Wegman and Luo (2002) estimate that with single pixel coding less than 10^6 data items can be visualized, and this is a generous upper bound.

One may solve the problem of overplotting by switching to visualization methods based on smoothing; see Figure I.1. However, one may also visualize a large number of observations with pixel-based tools by using *subsetting*. Subsetting means that one chooses subsets of the data and visualizes only one subset at a time. Subsetting is a concept that has been used particularly in earth sciences to describe the process of retrieving just the parts of large files that may arise, for example, when collecting data with earth-observing satellites. The classical subsetting techniques include the following:

1. The subset chosen randomly.

2. The subset consists only of the central data points, with outliers removed by taking all inside a box or inside a sphere.

3. When data contain some categorical variables, these variables can be used to choose a subset.

4. When data does not contain categorical variables it is possible to define such variables using cluster analysis, for example.

5. It is possible to use one variable to define the groups (different age groups or income groups), for example, by taking all on a halfspace (thresholding).

6. With time series data subsets can be chosen temporally by date/time ranges.

7. With spatial data subsets can be chosen by latitude/longitude.

We illustrate a method that may be called *likelihood subsetting*. Likelihood subsetting is a method that lies conceptually between visualizing the complete data and smoothing the data. In likelihood subsetting one chooses the observations that are inside a level set of a density estimate. Thus we consider subsets

$$A_\lambda = \left\{ x_1, \ldots, x_n : \hat{f}(x_i) \geq \lambda \right\},$$

where $\hat{f} : \mathbf{R}^d \to \mathbf{R}$ is a density estimate and $0 \leq \lambda \leq \max\{\hat{f}(x_i) : i = 1, \ldots, n\}$. There are $n + 1$ distinct sets A_λ as λ varies, but it is reasonable to expect that for most data sets it is enough to try only few values of λ. Likelihood subsetting is typically computationally expensive, but it is much less expensive than the full program of estimating and visualizing densities, where one needs typically to evaluate density estimates on a multivariate grid.

We may implement the likelihood subsetting with a nearest neighbor density estimate, defined in Section 14.4. First we need to calculate pairwise distances between observations, which takes $O(d \cdot n^2)$ steps. Evaluation of the estimate at the data points takes $O(n^2)$ steps (for each observation, find the distance to the kth nearest neighbor).

Figure 1.4 shows scatter plots of exchange rates of Brazilian Real and Mexican new Peso between 1995-01-05 and 2007-09-26. The rates are with respect to one U.S. Dollar and transformed to returns (we use the transformation $r_i \mapsto (r_i - r_{i-1})/r_{i-1}$). Panel (a) shows 3197 observations and panel (b) shows a subset of size 1769. We applied the nearest-neighbor likelihood subsetting with the number of neighbors $k = 30$. Subsetting reveals the positive correlation between Brazilian Real and Mexican Peso. The data is provided by Federal Reserve Economic Data (http://research.stlouisfed.org).

1.2 UNIVARIATE DATA

We mention five types of visualization methods for univariate data: (1) a line plot, 1D scatter plot, index plot, time series plot; (2) empirical distribution function, tail plot; (3) PP-plot and QQ-plot; (4) box plot; and (5) kernel estimate.

1.2.1 Line Plot, 1D Scatter Plot, Index Plot, Time Series Plot

Visualizing 2D data with scatter plots is a natural visualization technique. At first sight the 1D case seems more difficult: a naive 1D scatter plot would place the observations on the coordinate axis and the overplotting would make the plot useless. Figure 1.5 shows a *line plot* of a sample of size 100 from the two-modal density, whose density is shown in Figure 1.9 and in Figure 1.13(a) as a solid line. Each observation is plotted as a vertical line, placed at the location of the observation in the coordinate axis. A line plot suffers heavily from overplotting.

Lifting is the the main idea for making useful 1D scatter plots. Lifting means that one adds a new variable to the data and makes a 2D scatter plot of the original

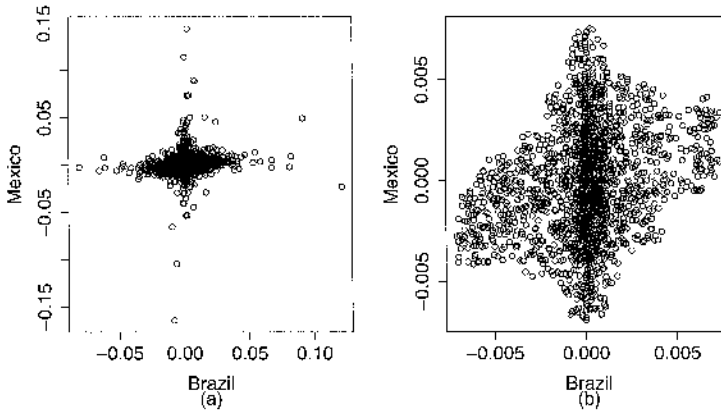

Figure 1.4 *(Likelihood subsetting.)* (a) A scatter plot of exchange rates for Brazilian Real and Mexican Peso ($n = 3197$); (b) a subset of data (size 1769) obtained by a likelihood subsetting.

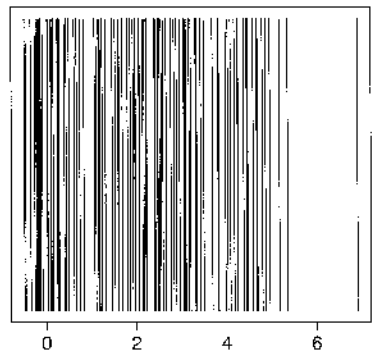

Figure 1.5 *(Line plot.)* A line plot of data of size 100.

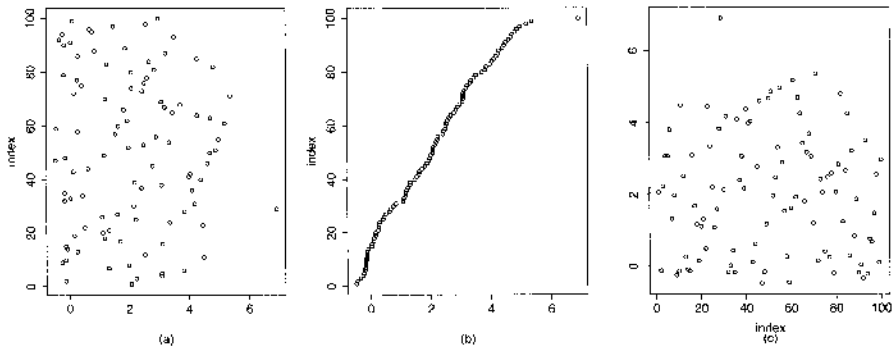

Figure 1.6 *(1D scatter plots.)* 1D scatter plots of data in Figure 1.5 when three different methods of lifting are applied.

observations together with the synthetic measurements. The new variable may be called level because its role is to spread the observations in the direction of the new coordinate. In the simplest case the level may be chosen to be the index of the data points: $\text{level}(x_i) = i$, $i = 1, \ldots, n$. In the case of time series data it is natural to choose the level to be the time index of the observations. In many cases one is free to choose the level so that the plot reveals as much information as possible. We will generalize 1D scatter plots to the multivariate case and call them *parallel level plots*. In the multivariate case one makes a 1D scatter plot from each coordinate and the level identifies the observations among the d scatter plots. Parallel level plots are defined in Definition 1.2.

Figure 1.6 shows three 1D scatter plots. In panel (a) the level is chosen to be the index. In panel (b) the level is chosen to be the number of observations which are smaller or equal to this observation,

$$\text{level}(x_i) = \#\{x_j : x_j \leq x_i,\ j = 1, \ldots, n\}.$$

In panel (c) the level is chosen to be the index, but the level is plotted as an x-coordinate and the observed values as a y-coordinate, unlike in panel (a).

One may enhance 1D scatter plots by adding connections to the plot. Figure 1.7 shows the scatter plots of Figure 1.6 enhanced. In the scatter plot of panel (a) each observation is connected with a line to the arithmetic mean of the data. We may call this plot an *index plot*. The scatter plot of panel (b) is the same as the empirical distribution function of Figure 1.9, up to the normalizing factor n^{-1}. Figure 1.7(c) shows a *time series plot*; now the observations are joined to each other with lines, and the level is shown at the vertical axis

Time series plots may be used to visualize the spread of the data. Figure 1.8 shows time series plots of three simulated data of size 500. Panel (a) shows data whose distribution is the Student distribution with degrees of freedom 3, panel (b) shows

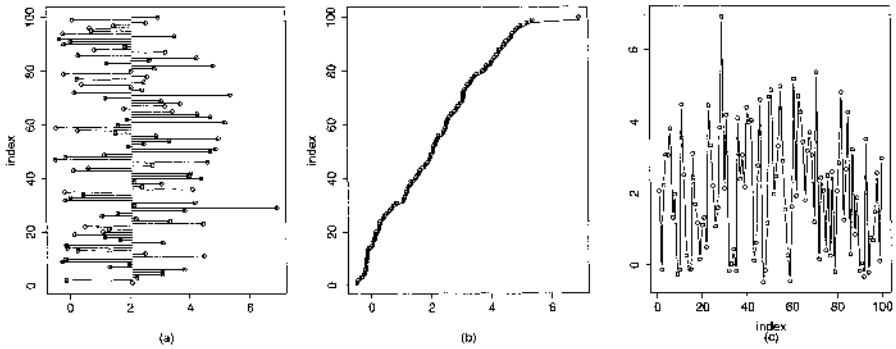

Figure 1.7 *(Enhanced 1D scatter plots.)* Shown are an index plot, an empirical distribution function, and a time series plot corresponding to the 1D scatter plots in Figure 1.6.

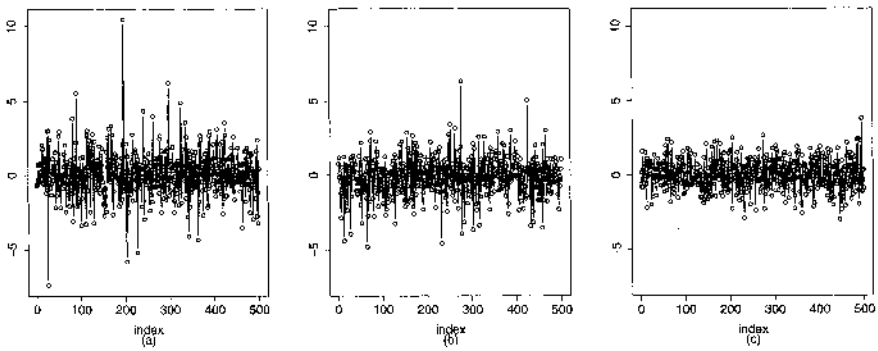

Figure 1.8 *(Time series plots.)* Time series plots used in visualizing the spread of data. (a) A sample from the Student distribution with degrees of freedom 3; (b) a sample from the standard double exponential distribution; (c) a sample from the standard Gaussian distribution. Sample sizes are 500.

data from the standard double exponential distribution, and panel c) shows data from the standard Gaussian distribution.

1.2.2 Empirical Distribution Function and Tail Plot

Empirical distribution functions may be used in visualizing the heaviness of the tails. In contrast, kernel estimates are better in mode detection.

The empirical distribution function is defined as

$$F_n(t) = n^{-1}\#\{i : i = 1, \ldots, n,\ x_i \le t\}, \qquad t \in \mathbf{R}.$$

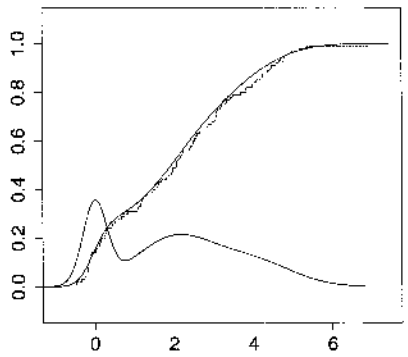

Figure 1.9 *(Distribution function versus density function.)* Empirical distribution function compared with the true distribution function and the density function. The empirical distribution function is the step function, the true distribution function is the smooth monotonically increasing function, and the density function is the two-modal function.

The empirical distribution function may be thought of as a modification of the 1D scatter plot. The level of an observation is equal to the number of observations that are smaller or equal to this observation, and we normalize the level by dividing by the sample size:

$$\text{level}(x_i) = n^{-1}\#\{x_j : x_j \leq x_i,\ j = 1, \ldots, n\}. \tag{1.1}$$

In addition the observations are joined with lines to make a step function from the scatter plot.

Figure 1.9 shows the empirical distribution function of the data in Figures 1.5 through 1.7. The true distribution function and the density function are included in the same plot. The true distribution function is the smooth monotonically increasing function and the density function is the two-modal function. The figure illustrates the difficulty of detecting modes with a distribution function.

When we want to visualize the heaviness of the tails of the underlying distribution, it useful to use *tail plots* of the data. We divide the data to the left tail and to the right tail, and visualize separately the two tails. The right tail is $\mathcal{R} = \{x_i : x_i > \mu,\ i = 1, \ldots, n\}$, where μ may be chosen to be the median of the observations. Now we choose the level

$$\text{level}(x_i) = \#\{x_j : x_j \geq x_i, x_j \in \mathcal{R}\}, \qquad x_i \in \mathcal{R}.$$

The scatter plot of $\mathcal{R}$ with this choice of level visualizes the heaviness of the right tail. It is useful to use a logarithmic axis for the y-axis. When the data are centered around 0, we may choose $\mu = 0$, and we may use a logarithmic scale also for the x-axis.

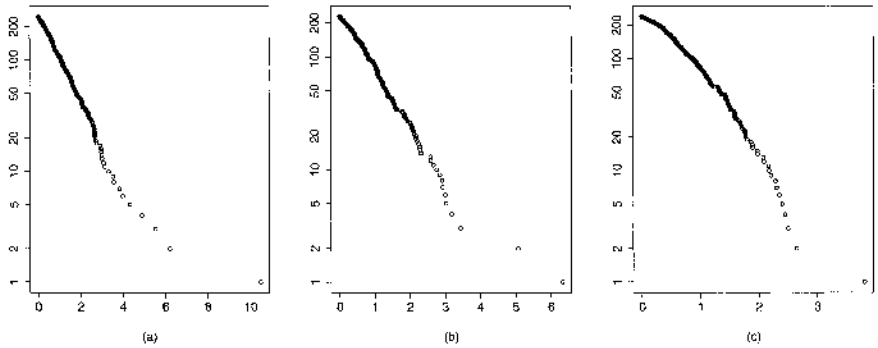

Figure 1.10 *(Right-tail plot.)* Right-tail plots of the data in Figure 1.8: (a) Sample from the Student distribution with degrees of freedom 3; (b) sample from the standard double exponential distribution; (c) sample from the standard Gaussian distribution. Sample sizes are 500.

Figure 1.10 shows right tail plots of the data in Figure 1.8. We us a logarithmic scale only for the y-axis. Tail plots have been used by Mandelbrot (1963), Bouchaud and Potters (2003), Sornette (2003).

1.2.3 PP-Plot and QQ-Plot

PP-plots and QQ-plots visualize the empirical distribution function or its inverse (quantile function) by comparing them to a reference distribution. The reference distribution might represent a null-hypothesis, so the PP-plot and QQ-plot are used to assess the goodness of fit. The reference distribution may have a parametric form and the parameters are estimated from the data. One may also scan through a collection of reference distributions in the spirit of exploratory data analysis.

We denote with $F : \mathbf{R} \to [0,1]$ a reference distribution function and with $F^{-1} : [0,1] \to \mathbf{R}$ the generalized inverse of F:

$$F^{-1}(p) = \inf\{t \in \mathbf{R} : F(t) \geq p\}, \qquad p \in [0,1].$$

Definition 1.1 (Probability plot, Quantile plot.) *A* PP-plot, *or a* probability plot, *of the data $x_1, \ldots, x_n \in \mathbf{R}$, associated to a reference distribution function $F : \mathbf{R} \to [0,1]$, is a plot of points*

$$(F_n(x_i), F(x_i)), \qquad i = 1, \ldots, n,$$

where F_n is the empirical distribution function. A QQ-plot, *or a* quantile plot, *of the data $x_1, \ldots, x_n \in \mathbf{R}$, associated to a reference distribution function $F : \mathbf{R} \to [0,1]$, is a plot of points*

$$\left(F^{-1}(p_i), F_n^{-1}(p_i)\right), \qquad i = 1, \ldots, n.$$

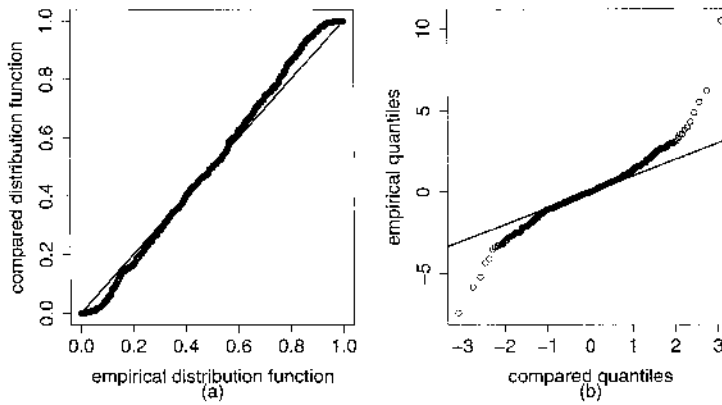

Figure 1.11 *(PP-plot, QQ-plot.)* (a) A PP-plot of a sample from a Student distribution with degrees of freedom 3; (b) a QQ-plot of the sample. The reference distribution is the standard Gaussian distribution.

where $p_i = (i - 1/2)/n$. The graph of the line $y = x$ ($x_1 = x_2$) is also included in the plots.

Note that $F_n(x_{(i)}) = i/n$, where $x_{(1)} < \cdots < x_{(n)}$ is the ordered sample, and thus the PP-plot is the plot of the points

$$\left(i/n, F(x_{(i)})\right), \qquad i = 1, \ldots, n.$$

Note also that

$$F_n^{-1}(p) = \inf\{t \in \mathbf{R} : F_n(t) \geq p\} = \begin{cases} x_{(1)}, & 0 < p \leq 1/n, \\ x_{(2)}, & 1/n < p \leq 2/n, \\ \vdots \end{cases}$$

Thus the QQ-plot is the plot of the points

$$\left(F^{-1}(p_i), x_{(i)}\right), \qquad i = 1, \ldots, n.$$

Figure 1.11 shows a PP-plot and a QQ-plot for a sample of size 500 from the Student density with degrees of freedom 3, when the reference distribution is the standard Gaussian distribution. Panel (a) shows a PP-plot and panel (b) shows a QQ-plot.

When the sample is originating from the reference distribution F, the points in the PP-plot lie near the line $x_1 = x_2$. Location differences will lead to plots where the points lie above or below this line. Scale differences will lead to plots where the points lie first on the one side of the line $x_1 = x_2$, next intersect this line, and then lie on the other side of the line.

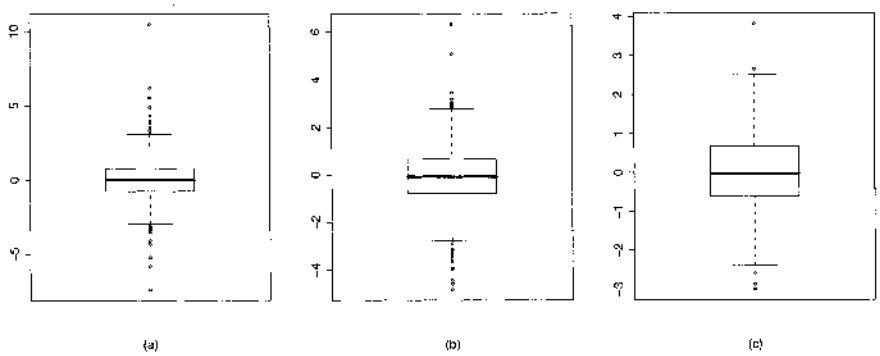

Figure 1.12 *(Box plot.)* Box plots of the data in Figure 1.8. (a) A sample from the Student distribution with degrees of freedom 3; (b) a sample from the standard double exponential distribution; (c) a sample from the standard Gaussian distribution. Sample sizes are 500.

A PP-plot is related to the Kolmogorov-Smirnov test statistic

$$KS_n(G) = \sup_{x \in \mathbf{R}} |G(x) - F_n(x)|,$$

which is used to test the null-hypothesis $F = G$, where F is the true distribution function. The asymptotic distribution of the normalized test statistic $\sqrt{n}KS_n(F)$ is the distribution of $\sup_{p \in [0,1]} |B(p)|$, where B is the standard Brownian bridge. Thus the asymptotic distribution, when the null-hypothesis holds, is independent of the true distribution F. The formal testing procedure is useful when we accompany it with the PP-plot: for the case where the null-hypothesis is rejected, we need a graphical tool that indicates the direction of the deviation from the null-hypothesis.

1.2.4 Box Plot

A *box plot* shows a box with a center line at the median, the upper bound of the box at the 75% quantile, and the lower bound of the box at the 25% quantile. Thus the box shows the interval where 50% of the data lies. In addition lines are drawn at 1.5 times the inter-quartile range, and the extreme observations beyond these lines are drawn as circles. Figure 1.12 shows box plots of the data in Figure 1.8. The previous plots did not reduce the information content of the sample (except that a right-tail plot and a left-tail plot visualize only half of the data); one may reconstruct the data up to the effects of overplotting. A box plot reduces the information in the sample to the 5 numbers: median, 2 quartiles, and 2 extremes. A box plot may be more suited for the presentation of the data than for the exploration of the data. Note, however, that the extreme observations are plotted individually in a box plot, and often only the extreme observations need to be identified, whereas the central observations may be aggregated in visualizations.

1.2.5 Kernel Estimates

The previous visualization methods are not effective for the visualization of multimodality. We may use smoothing to visualize the modes. Thus we consider density estimates, like kernel estimates, as a tool for visualizing data. Kernel estimators are defined in Section 14.3.

Unlike 1D scatter plots, empirical distribution functions, and many other visualization methods, a kernel estimate looses some information: it is not possible to recover the original data when we have available only a kernel estimate. However, less information is lost when the smoothing parameter becomes smaller. In the limit, as the smoothing parameter of kernel estimates approaches zero, a kernel estimate shows delta peaks at each observation. This leads to the idea of visualizing the data by showing a scale of kernel estimates corresponding to a sequence of smoothing parameter values. A scale of kernel estimates shows simultaneously a summary of the data (when the smoothing parameter is large or moderate) and idiosyncratic features (when the smoothing parameter is small). We do not have to choose a single best kernel estimate but we may visualize the data with a scale of density estimates. The scale space view of curve estimation is developed in Chaudhuri and Marron (2000); Chaudhuri and Marron (1999) introduced SiZer as an exploratory tool in the spirit of the scale space view.

Figure 1.13(a) shows a scale of kernel estimates. The true density is shown in panel (b) as the solid line, and also in Figure 1.9. The sample size is 100 and the smoothing parameter is in the interval $h \in [0.2, 1]$. The Gaussian kernel is used. Figure 1.13(b) shows the kernel estimates with $h = 0.1$ and $h = 1$ as dashed lines, in addition to showing the true density with the solid line. We discuss in Chapter 19 a stagewise minimization estimator which estimates this target density better; see Figure 19.1(a).

We can combine a kernel estimate and a 1D scatter plot when the level of an observation is chosen to be the value of the kernel estimate at the observation:

$$\text{level}(x_i) = \hat{f}(x_i), \qquad i = 1, \ldots, n,$$

where $\hat{f}$ is a kernel estimate. Figure 1.14 shows two 1D scatter plots with the level determined by kernel estimates with $h = 0.2$ and $h = 1$. The standard Gaussian kernel function is used in the kernel estimates. The data are shown also in Figures 1.5 through 1.7.

1.3 PARALLEL LEVEL PLOTS

In *parallel level plots* we introduce a new dimension to the point cloud, called a level. We attach each observation with a level and make d parallel scatter plots. Each coordinate gives rise to a scatter plot and the level is the y-coordinate of the scatter plots. Thus we identify the observations across scatter plots through the value of the level of the observations. Linking across dimensions is discussed also in Section 1.6 and parallel level plots will be enhanced in Chapter 6 by introducing tail tree plots.

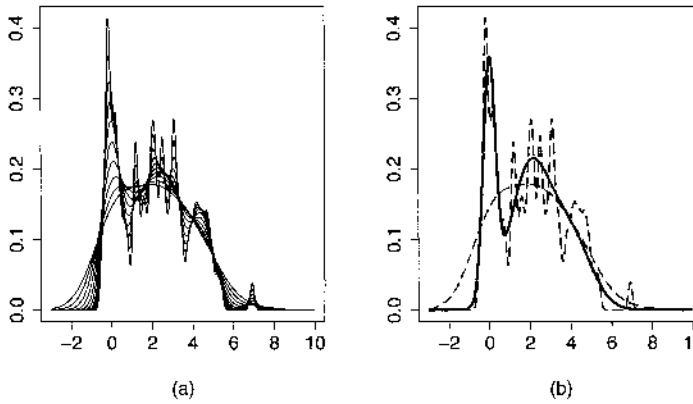

Figure 1.13 *(Scale of kernel estimates.)* (a) A scale of 10 kernel estimates with $h \in [0.2, 1]$. (b) The true density with a solid line, and the estimates with $h = 0.2$ and $h = 1$ with dashed lines. The data of size $n = 100$ is shown in Figures 1.5 through 1.7.

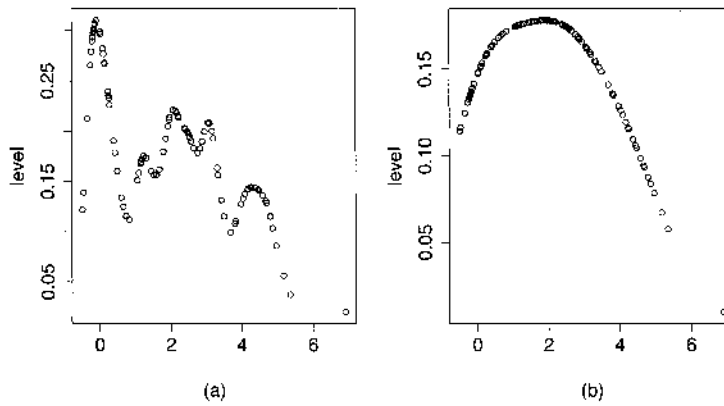

Figure 1.14 *(1D scatter plots via kernel estimates.)* Shown are 1D scatter plots of the data in Figures 1.5 through 1.7. The level of scatter plots is the value of a kernel estimate at the observations. (a) $h = 0.2$; (b) $h = 1$.

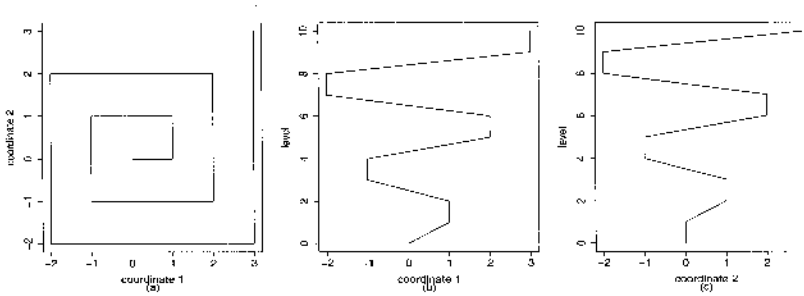

Figure 1.15 *(Parallel level plot, 2D.)* (a) A rectangular spiral in the 2D Euclidean space; (b-c) a parallel level plot of the spiral.

Definition 1.2 (Parallel level plot.) A parallel level plot *of data* $x_1, \ldots, x_n \in \mathbf{R}^d$, *associated with a level mapping* $\text{level}(x_i) \in \mathbf{R}$, *consist of* d *windows, where the kth window,* $k = 1, \ldots, d$, *shows a scatter plot of points* $(x_i^k, \text{level}(x_i))$, $i = 1, \ldots, n$. *We use the notation* $x_i = (x_i^1, \ldots, x_i^d) \in \mathbf{R}^d$.

When the level mapping satisfies $\text{level}(x_i) \neq \text{level}(x_j)$, for $i \neq j$, $i, j = 1, \ldots, n$, the unique identification of the observations across the windows is possible.

Parallel level plots generalize the 1D scatter plot, shown in Figures 1.6 and 1.7, to the multivariate case. All these plots introduce a new variable (level) to the data. In the multivariate case the level is the same for each coordinate of a single observation, and the level is used to identify the observations across the windows.

1.3.1 Multivariate Time Series

Parallel level plots have been common in visualizing multivariate time series: one takes the time index of the data points to be their level. If we have d time series, and each time series contains measurements at the same n time points, then the plot of these times series in d parallel windows is a parallel level plot. Becker, Clark and Lambert (1994) uses a visualization, called a cave plot, which resembles the index plot of Figure 1.7(a) to visualize a multivariate time series.

1.3.2 One-dimensional Curves

One-dimensional curves in the d-dimensional Euclidean space may be visualized naturally with parallel level plots by taking the parameter of the curve to be the level: if we have a curve $z(t) = (z_1(t), \ldots, z_d(t)) \in \mathbf{R}^d$, $t \in [0, \infty)$, we may visualize it with a parallel level plot by choosing a fine grid $0 \leq t_1 < \cdots < t_n < \infty$, defining $x_i = z(t_i)$, and $\text{level}(x_i) = t_i$, $i = 1, \ldots, n$. We will illustrate parallel level plots in visualizing spirals.

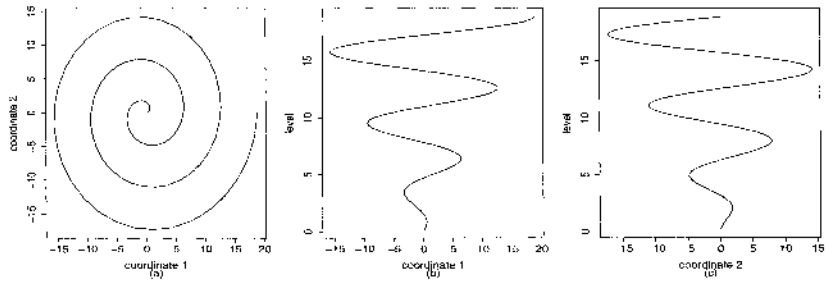

Figure 1.16 *(Parallel level plot, 2D.)* (a) An Archimedean spiral in the 2D Euclidean space; (b-c) a parallel level plot of the spiral.

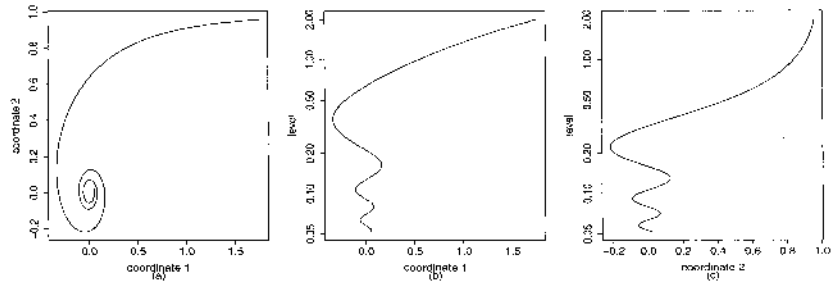

Figure 1.17 *(Parallel level plot, 2D.)* (a) A hyperbolic spiral in the 2D Euclidean space; (b-c) a parallel level plot of the spiral.

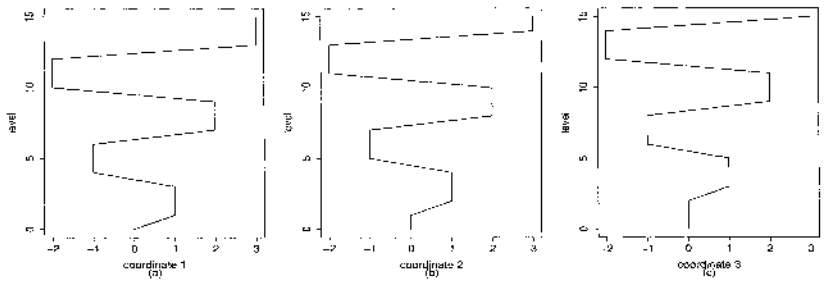

Figure 1.18 *(Parallel level plot, 3D.)* A parallel level plot of a 3D rectangular spiral.

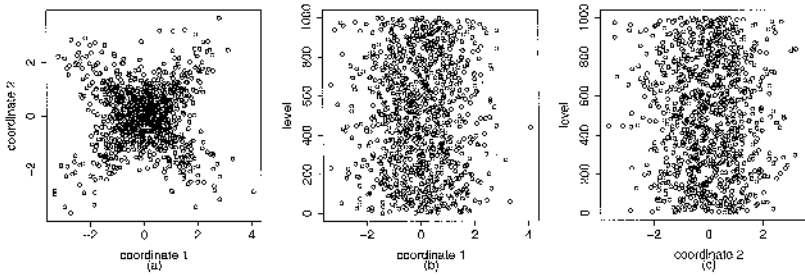

Figure 1.19 *(Parallel level plot, point cloud.)* (a) A scatter plot of size 1000 from a mixture of two Gaussian distributions. (b-c) A parallel level plot of the data in (a) when the level is the index of observations.

1. Figure 1.15(a) shows a 2D rectangular spiral and Figure 1.15(b)-(c) shows a parallel level plot of the spiral. We have taken the level equal to the length of the spiral, starting at the origin.

2. Figure 1.16(a) shows a 2D Archimedean spiral and Figure 1.16(b)-(c) shows a parallel level plot of the spiral. The Archimedean spiral is defined by $r \mapsto r(\cos r, \sin r)$, $r > 0$. We have taken the level equal to the radius r.

3. Figure 1.17(a) shows a 2D hyperbolic spiral and Figure 1.17(b)-(c) shows a parallel level plot of the spiral. The hyperbolic spiral is defined by $r \mapsto (r\cos(1/r), r\sin(1/r))$, $r > 0$. We have taken the level equal to the radius r, and $r \geq 0.05$.

4. Figure 1.18 shows a parallel level plot of a 3D rectangular spiral.

1.3.3 Point Clouds

When visualizing point clouds in general, the usefulness of a parallel level plot depends on the choice of level. A natural choice for the level is the distance from a center point, say, from the arithmetic mean. Define

$$\text{level}(z) = \|z - \mu\|, \qquad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.2}$$

An other choice for the level is the value of the first coordinate, say,

$$\text{level}(z) = z^1, \qquad z = (z^1, \ldots, z^d) \in \mathbf{R}^d. \tag{1.3}$$

We illustrate the effect of the choice of the level with four figures.

1. Figure 1.19 shows a 2D point cloud which is a sample of size 1000 from a mixture of two Gaussians. Panels (b)-(c) show the same data with a parallel

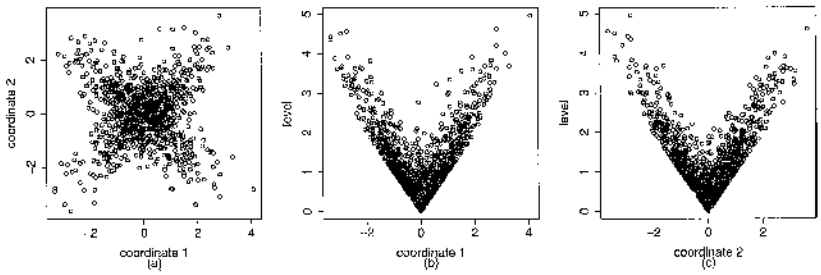

Figure 1.20 *(Parallel level plot, point cloud.)* (a) The same data as in Figure 1.19(a). (b-c) A parallel level plot when the level is the distance from the arithmetic mean.

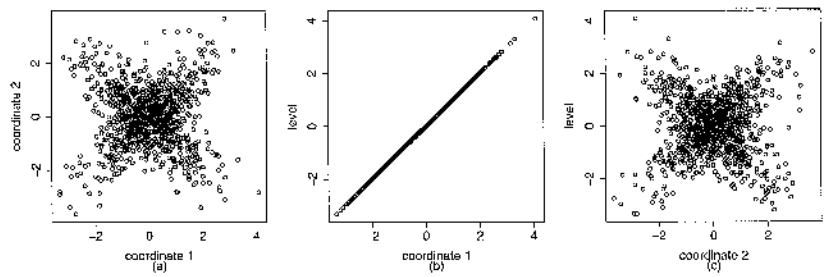

Figure 1.21 *(Parallel level plot, point cloud.)* (a) The same data as in Figure 1.19(a). (b-c) A parallel level plot when the level is the first coordinate.

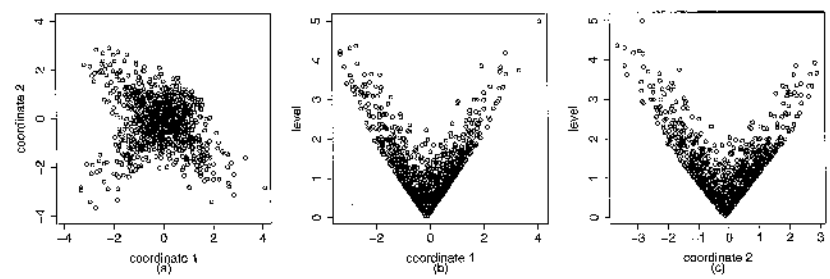

Figure 1.22 *(Parallel level plot, point cloud.)* (a) A scatter plot of the data in Figure 1.19 when one tail is removed. (b-c) A parallel level plot when the level is the distance from the arithmetic mean.

level plot where we have attached a unique random level in $\{1, \ldots, 1000\}$ to each observation (the level is in fact the index of the observations). The scatter plot and the parallel level plot contain the same information (for a computer program), but the scatter plot lets a human eye to detect the pattern of four tails in the data.

2. Figure 1.20(b)–(c) shows a parallel level plot of the data in panel (a) when we choose the level as the distance to the arithmetic mean, as in equation (1.2). The parallel level plot indicates that there are at least two tails in the data. We define in Section 6.2 a *tail tree plot*. A tail tree plot adds a sufficient structure to a parallel level plot making it possible for humans to detect the four tails in the data, as is shown in Figure 6.9.

3. Figure 1.21(b)–(c) shows a parallel level plot when the level is the first coordinate, as in (1.3). In this case the second window of the parallel level plot is a scatter plot where the coordinates have been switched.

4. Figure 1.22 shows the data in Figures 1.19 through 1.21 when one tail is cut away. Panel (a) shows a scatter plot and panels (b–c) show a parallel level plot when the level is the Euclidean distance from the arithmetic mean, as in equation (1.2). The parallel level plot of the data does not appear to differ from the parallel level plot in Figure 1.20(b)–(c), where the data has four tails. This demonstrates further the need for tail tree plots.

1.4 GRAPHICAL MATRICES

A graphical matrix is a $n \times d$ matrix of graphical elements; a $n \times d$ data matrix of real numbers is transformed by representing each real number by a graphical element. Graphical matrices were studied by Bertin (1967, 1981). There are many ways to code real numbers with graphical elements. When the data consists only of binary values, one may represent these values with white and black rectangles. Figure 1.23 shows a graphical matrix containing data about townships. The variables take binary values, and the values are coded with black and white. Figure 1.24 shows the same data when the columns and rows have been reordered, so that the patterns are revealed. Real values can be represented with bars or lines, whose length is proportional to the value. Bertin (1981) mentions seven visual variables that can be used to code numerical values: form, position, size, value, texture, color, and orientation. Further developments of a graphical matrix include the Data Image as in Minnotte and West (1999).

1.4.1 Bar Matrix

Figure 1.25(a) shows a scatter plot of five observations. Panel (b) shows a graphical matrix of these data, where the measurements are pictured as bars whose heights are proportional to the values of measurement (relative to the range of the variable). We

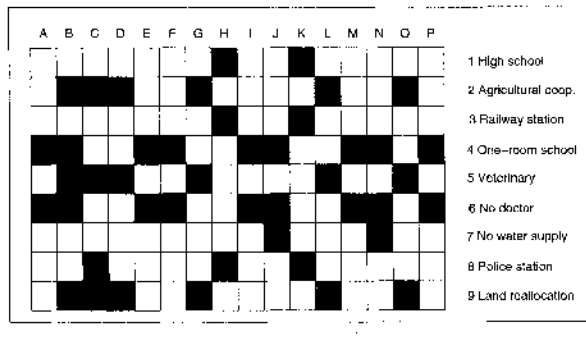

Figure 1.23 *(Unordered table.)* The unordered table of Bertin is reproduced from Bertin (1981).

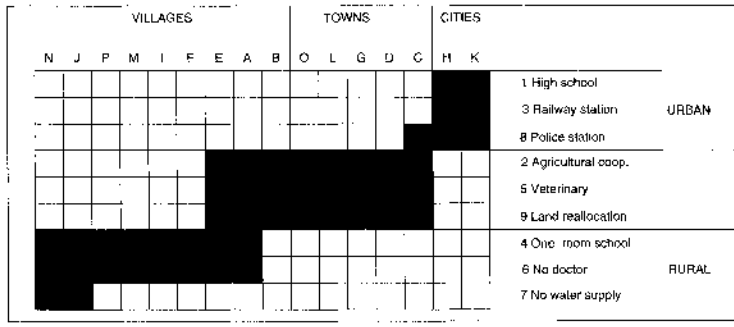

Figure 1.24 *(Ordered table.)* The ordered table of Bertin is reproduced from Bertin (1981).

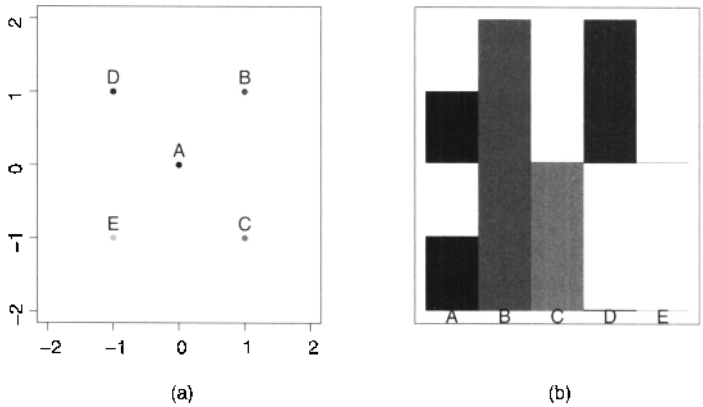

Figure 1.25 *(Bar matrix, definition.)* (a) A scatter plot of five points, and (b) a graphical matrix of the data with bars as graphical elements (bar matrix).

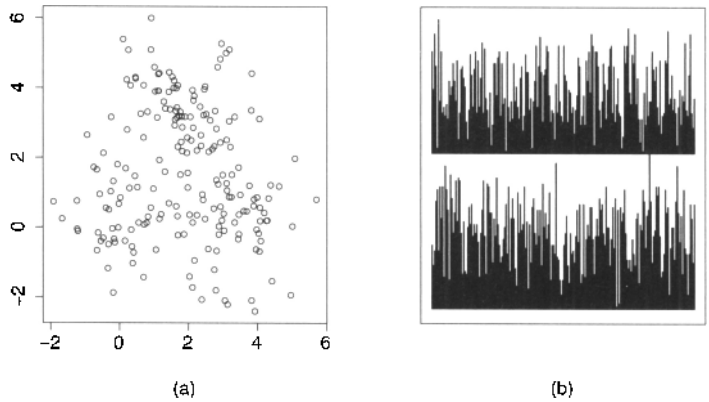

Figure 1.26 *(Bar matrix, illustration.)* (a) A scatter plot of 200 points and (b) a graphical matrix of the data with bars as graphical elements (bar matrix).

have labeled the bars to show the correspondence with the points in the scatter plot. This kind of a graphical matrix may be called a *bar matrix*.

Figure 1.26 shows an other example of a scatter plot and a bar matrix. Panel (a) shows a data of size 200 (the data are generated from the density shown in Figure 2.7), and panel (b) shows a graphical bar matrix with a random permutation.

The usefulness of a graphical matrix depends on the choice of an informative permutation of the rows and the columns of the matrix. Similarly to the parallel level plot in Figure 1.19(b)–(c), and similarly to the Bertin's table in Figure 1.23, a

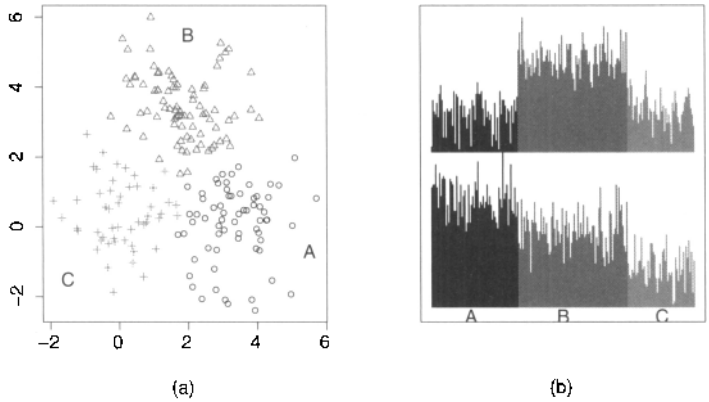

Figure 1.27 *(Bar matrix, ordered.)* (a) A scatter plot of 200 points when the k-means clustering algorithm has been applied to label the data points. (b) A graphical matrix of the data with bars as graphical elements (bar matrix). The observations belonging to a same cluster are grouped together.

random permutation leads typically to useless graphical matrices. A suggestion of Bertin (1981) is to order indexes so that the values of one variable are in a monotonic order. The permutation could also be done with the help of a clustering algorithm. The observations belonging to the same cluster are grouped together in the graphical matrix.

Figures 1.26 and 1.27 illustrate the usefulness of a permutation. Figure 1.26(b) shows an unordered graphical matrix and Figure 1.27(b) shows an ordered graphical matrix. The ordering has been done by a permutation where the observations belonging to a same cluster are grouped together. The clusters were obtained by the k-means algorithm with $k = 3$; see Section 8.2 for the definition of k-means clustering.

1.4.2 Index Plot Matrix

Figure 1.28(a) shows a scatter plot of five observations. Panel (b) shows a graphical matrix of these data, where each measurement is drawn as a circle that is connected with a line to the arithmetic mean of the coordinate. This graphical matrix is identical with a parallel level plot, where the level is the index of the observation, and we delete the lines and keep only the circles. (Another possibility would be to delete the circles and keep only the lines.) Note also that this graphical matrix generalizes the 1D index plot, as shown in Figure 1.7(a).

Figure 1.29 shows an other example of an index plot matrix. Panel (a) shows a scatter plot of data of size 200 that contains four tail clusters and panel (b) shows an index plot matrix with random ordering of the observations.

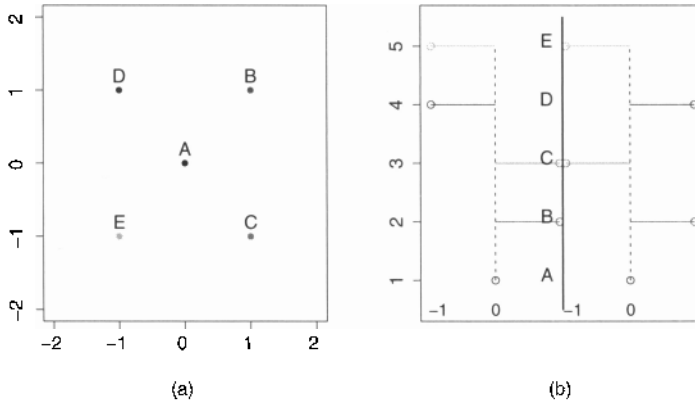

Figure 1.28 *(Index plot matrix, definition.)* (a) A scatter plot of five points, and (b) a graphical matrix of the data with lines as graphical elements (index plot matrix).

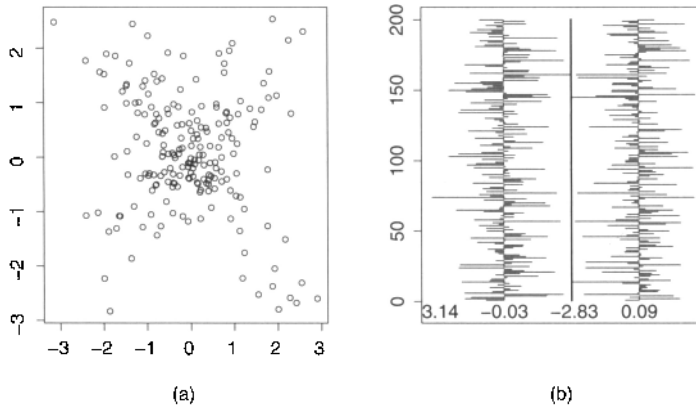

Figure 1.29 *(Index plot matrix, illustration.)* (a) A scatter plot of 200 points and (b) a graphical matrix of the data with lines as graphical elements (index plot matrix).

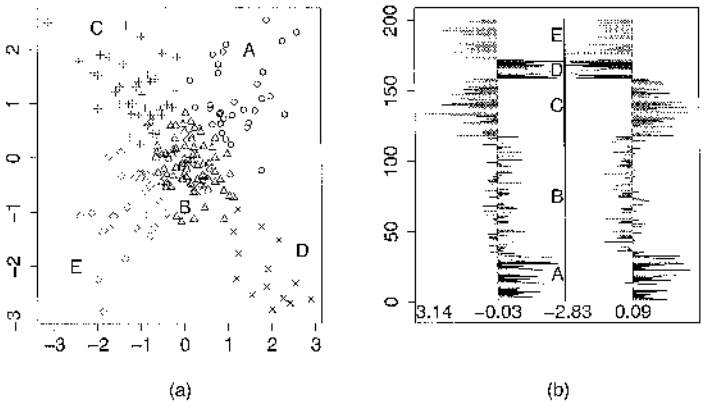

Figure 1.30 *(Index plot matrix, ordered.)* (a) A scatter plot of 200 points when the k-means clustering algorithm has been applied to label the data points. (b) A graphical matrix of the data with lines as graphical elements (index plot matrix). The observations belonging to a same cluster are grouped together.

Figures 1.29 and 1.30 illustrate the usefulness of a permutation based on a clustering. Figure 1.29(b) shows an unordered matrix, and Figure 1.30(b) shows an ordered matrix for groups that were obtained by the k-means algorithm with $k = 5$, The observations belonging to a same cluster are grouped together. Cluster analysis is discussed more in detail in Chapter 8. Cluster analysis may be used to visualize data with graphical matrices, or conversely graphical matrices may be used to visualize and validate the results of a cluster analysis.

1.5 OBSERVATIONS AS OBJECTS

Parallel coordinate plots and Andrew's curves represent observations as 1D curves. Chernoff faces represent observations with schematic human faces. We mention also other possibilities for visualizing data by representing the data points as objects. These methods work typically for small data sets. Also, the methods are usually sensitive for the ordering of the variables.

1.5.1 Parallel Coordinate Plots

Parallel coordinate plots may be used to visualize a multivariate point cloud by looking at the one-dimensional projections of the data to the coordinate axes, which are drawn parallel to each other. One needs to identify the observations across the d projections. Parallel coordinate plots identify the observations across the coordinate axes by joining them with straight lines.

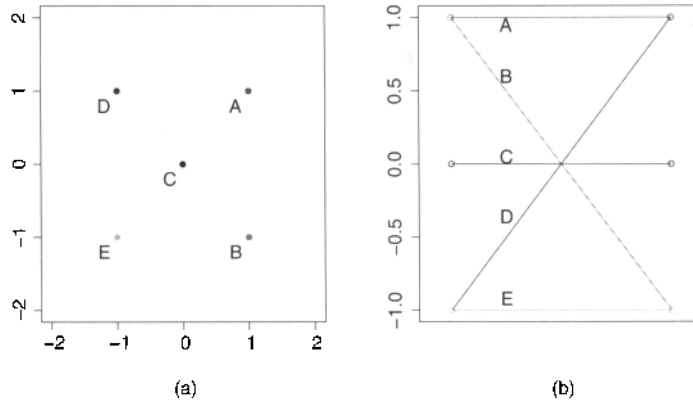

Figure 1.31 *(Parallel coordinate plot, definition.)* (a) A scatter plot of five points and (b) a parallel coordinate plot of these points.

In a parallel coordinate plot the number of the coordinate is mapped onto the horizontal axis and the value is mapped onto the vertical axis, and then these points are linearly interpolated. Observation x_i, whose index is $i \in \{1, \ldots, n\}$, is represented as the curve that linearly interpolates the points

$$(1, x_i^1), (2, x_i^2), \ldots, (d, x_i^d).$$

Parallel coordinate plots were introduced by Inselberg (1985). See also Inselberg and Dimsdale (1990), Wegman (1990), and Inselberg (1997).

Parallel coordinate plots resemble parallel level plots: in both cases one visualizes 1D projections of the data at the coordinate axis, and uses a method to identify the observations among the d projections. Parallel coordinate plots provide a more compact visualization of the data than parallel level plots. Consequently parallel coordinate plots suffer more from overplotting. Indeed the lines connecting the observations introduce more ink to the plot (bad "signal-to-ink-ratio"). In contrast, parallel level plots introduce a new variable to the plot (level), and spread the observations with the help of the new variable, thus diminishing the problem of overplotting. Miller and Wegman (1991) introduces a parallel coordinate density plot which that may be used to solve the problem of black ink.

1. Figure 1.31 illustrates the definition of a parallel coordinate plot. Panel (a) shows a scatter plot of five points, and panel (b) shows a parallel coordinate plot of these five points.

2. Figure 1.32 shows a scatter plot and the corresponding parallel coordinate plot. We have generated 500 observations from an equal mixture of two Gaussian distributions. The distribution is the same as in Figure 1.19, but the sample size is only half of that in Figure 1.19.

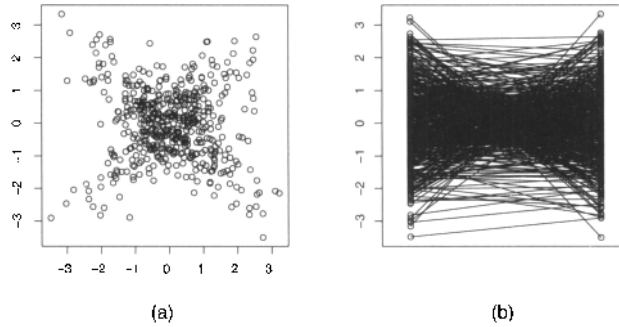

Figure 1.32 *(Parallel coordinate plot, illustration.)* (a) A scatter plot and (b) a corresponding parallel coordinate plot. The sample of size 500 is generated from a mixture of two Gaussian distributions. The distribution is the same as in Figure 1.19, where the sample size was 1000.

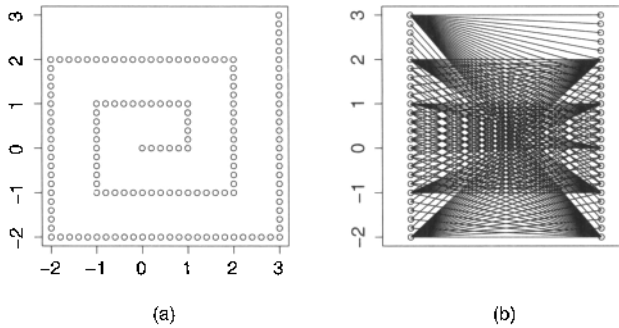

Figure 1.33 *(Parallel coordinate plot, rectangular spiral.)* (a) A scatter plot of points lying on a rectangular spiral and (b) a parallel coordinate plot of the points in panel (a).

3. Figure 1.33 visualizes a 2D rectangular spiral with a parallel coordinate plot.
4. Figure 1.34 visualizes a 2D Archimedean spiral with a parallel coordinate plot.
5. Figure 1.35 visualizes a 2D hyperbolic spiral with a parallel coordinate plot.
6. Figure 1.36 visualizes a 3D rectangular spiral with a parallel coordinate plot.

1.5.2 Multivariate Time Series

If we have n time series, and each time series is a vector of measurements at the same d time points, then the plot of these time series in a single window, joining the

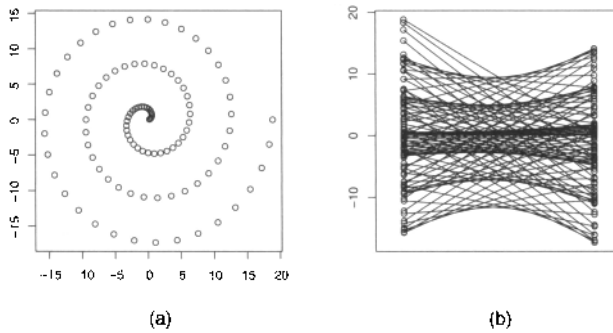

Figure 1.34 *(Parallel coordinate plot, Archimedean spiral.)* (a) A scatter plot and (b) a parallel coordinate plot.

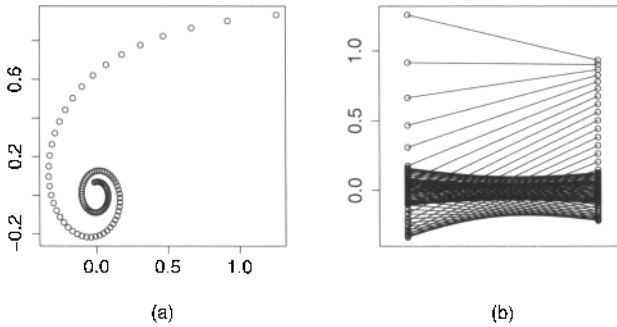

Figure 1.35 *(Parallel coordinate plot, hyperbolic spiral.)* (a) A scatter plot and (b) a parallel coordinate plot.

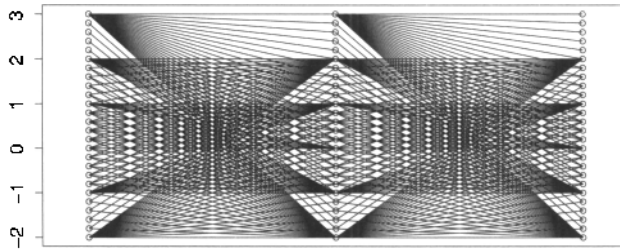

Figure 1.36 *(Parallel coordinate plot, 3D rectangular spiral.)* A parallel coordinate plot of points lying on a 3D rectangular spiral.

observations of a single time series with straight lines, is a parallel coordinate plot. Figure 1.7(c) shows a time series plot of a single time series.

1.5.3 Andrew's Curves

Andrew's curves, introduced in Andrews (1972), represent observations as one-dimensional curves. The curves are Fourier series whose coefficients are equal to the the observation values. The definition of the ith Andrew's curve is

$$f_i(t) = \begin{cases} 2^{-1/2}x_i^1 + \sum_{k=1}^{(d-1)/2} x_i^{2k}\sin(kt) + x_i^{2k+1}\cos(kt), \\ \qquad \text{when } d \text{ is odd,} \\ 2^{-1/2}x_i^1 + \sum_{k=1}^{d/2-1} \left[x_i^{2k}\sin(kt) + x_i^{2k+1}\cos(kt)\right] + x_i^d\sin(dt/2), \\ \qquad \text{when } d \text{ is even,} \end{cases}$$

for $i = 1, \ldots, n$, $d \geq 3$, where $t \in [-\pi, \pi]$. The ordering of the variables affects the visualization: the last variables will have only a small contribution to the visualization. Andrews (1972) suggests to use the ordering of the variables given by the principal component analysis; see Section 1.8.1 for the definition of the principal component analysis.

1.5.4 Faces

Faces are graphical representations of a data matrix where the size of face elements are assigned to variables (the columns of the data matrix). Chernoff (1973) introduced faces, and one often uses the term "Chernoff faces". Flury and Riedwyl (1981) and Flury and Riedwyl (1988) have further developed the face technique, and defined the following characteristics: right eye size, right pupil size, position of right pupil, right eye slant, horizontal position of right eye, curvature of right eyebrow, density of right eyebrow, horizontal position of right eyebrow, vertical position of right eyebrow,

right upper hair line, right lower hair line, right face line, darkness of right hair, right hair slant, right nose line, right size of mouth, right curvature of mouth, and the same for the left side of the face.

Each observation is represented by one face. We have defined together 36 characteristics of a face. Thus we may visualize data up to 36 variables, but we are not able to visualize many observations with faces. With faces we may find clusters from data by looking groups of similar looking faces.

1.5.5 Other Possibilities

- *Profiles.* Profiles represent each observation by d vertical bars. Each bar has a height proportional to the value of the corresponding variable. Sometimes profiles are drawn as polygonal lines. Bertin (1967) uses profiles to make graphical matrices, as discussed in Section 1.4. Bar matrices shown in Figures 1.25 through 1.27 are related to profiles: a transposed bar matrix would show n rows of profiles.

- *Stars.* Stars are similar to profiles. The values of the variables are represented by equally spaced rays emanating from a fixed point. The endpoints of the rays are usually connected in a polygon. This definition of stars can be found in Kleiner and Hartigan (1981).

- *Glyphs.* Glyphs are similar to profiles and stars. The values of the variables are represented by not equally spaced rays emanating from a boundary of a circle of a fixed positive radius. Anderson (1960) defines glyphs.

- *Boxes.* Boxes represent each variable as a length in one of the three dimensions of a box. If there are more than three variables, then there will be several segments within each dimension. Hartigan (1975*b*) defines boxes.

- *Trees.* To draw trees, one makes first a tree with d nodes; each node corresponds to a variable (a hierarchical clustering algorithm is applied to construct the tree). All the observations have the same tree structure, and the values of the variables are represented with the widths of the branches of the tree. Using a tree structure for representing points makes the method less sensitive to the problem of the ordering of the variables. Kleiner and Hartigan (1981) introduce trees.

Härdle and Simar (2003) discuss parallel coordinate plots, Andrews's curves, and Chernoff faces. Kleiner and Hartigan (1981) introduce trees and castles and discuss profiles, stars, glyphs, faces, and boxes.

1.6 LINKING ACROSS DIMENSIONS

We have shown many ways to visualize multivariate data by making d univariate visualizations and linking these visualizations together, by identifying observations

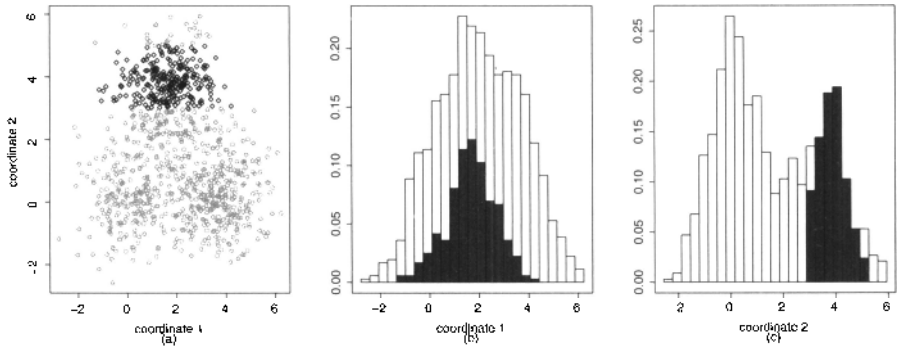

Figure 1.37 *(Linking across dimensions.)* Linking across dimensions illustrated with histograms.

across the dimensions. Parallel level plots identify observations across the dimensions by using an additional height-variable. Parallel coordinate plots identify observations across the dimensions by connecting them with straight lines. Graphical matrices make a layout of the data so that the attributes of a single observations are in the same row (or in a single column, depending on the arrangement). Tail tree plots, to be introduced in Chapter 6, enhance parallel level plots by coloring the observations with the help of a tail tree, and this helps identify observations across the 1D scatter plots.

When one makes several 2D projections and shows the scatter plots of the projected data, it is useful to apply brushing to identify observations across the scatter plots.

Brushing techniques can be used with all 1D visualization tools introduced in Section 1.2. For d one-dimensional visualizations brushing is used to connect observations across the dimensions. For example, one may select observations lying in an interval in one dimension and look how these observations are distributed in the other dimensions. It is useful to have a dynamic brushing tool to be able to try various subsets and refine the selections during the progress of exploration.

A particular example of linking across the dimensions is provided by the Attribute Explorer, introduced in Tweedie, Spence, Williams and Bhogal (1994) and Spence and Tweedie (1998), and whose use is illustrated in Spence (2001, Section 5.4). The Attribute Explorer makes d one-dimensional histograms, and each observation is identified with a rectangular region of a histogram bar. Figure 1.37 shows an example of using histograms in linking across dimensions. Panel (a) shows a data of size 1000. The data are generated from the same density as the data in Figure 1.1. Panel (b) shows a histogram of the first coordinate and panel (c) shows a histogram of the second coordinate, when the number of bins is 25. The observations whose second coordinate lies in the interval $[3, 5]$ are brushed.

1.7 DESCRIPTIVE STATISTICS

We will discuss descriptive statistics that characterize the distribution. The location of the data is characterized by a d dimensional vector that gives a center point for the data. We can make a very condensed visualization of the data by calculating a center point of the data. We can also characterize the dispersion of the data with various descriptive statistics. For example, box plots, shown in Figure 1.12, utilize descriptive statistics to make a visualization of data.

1.7.1 Location

Mean The expected value of random variable $X \in \mathbf{R}^d$ is written as EX. We use the notation EF for the mean of the distribution with distribution function

$$F(x) = P((-\infty, x]), \qquad x \in \mathbf{R}^d,$$

where $(-\infty, x] = (-\infty, x_1] \times \ldots \times (-\infty, x_d]$. When the distribution has density $f : \mathbf{R}^d \to \mathbf{R}$, we have

$$EF = \int_{\mathbf{R}^d} x f(x)\, dx,$$

assuming the integral is finite. A data version of the expected value may be taken to be the arithmetic mean:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Univariate Median A median of a univariate distribution F may be defined as such value $\text{med}(F) \in \mathbf{R}$ that

$$F(\text{med}(F)) = 0.5.$$

In terms of the density function f we have

$$\int_{-\infty}^{\text{med}(F)} f(y)\, dy = 0.5.$$

This equation may not uniquely define the median, since the distribution function F might not be genuinely increasing inside the support. The not genuinely increasing distribution function happens when the support is not an interval, or when the distribution is discrete. In general, we define

$$\text{med}(F) = \inf\{x \in \mathbf{R} : F(x) \geq 0.5\}.$$

A data version of median is a midpoint of the data, for example,

$$\text{med}_n = X_{(\lceil n/2 \rceil)},$$

where $X_{(1)} < \cdots < X_{(n)}$ is the ordered sample. Then $\text{med}_n \approx \text{med}(F_n)$, where F_n is the empirical distribution function. Unlike the mean, the median always exists.

Multivariate Median In the multivariate case it is not straightforward to define a median. We give several definitions for *data depthness*. A depth function $D : \mathbf{R}^d \to [0, \infty)$ corresponding to a distribution function $F : \mathbf{R}^d \to \mathbf{R}$, is such that when $D(x)$ is large, x is close to the center of the distribution (it is deep inside the distribution), and when $D(x)$ is small (close to 0), x is far away from the center. A median would be a value maximizing D. We may define a data version of the depth function by defining the depth function corresponding to the empirical distribution function.

1. *Mahalabonis depth.* Define the Mahalabonis distance, corresponding to the distribution function F, by

$$d_M^2(x) = (x - EF)^T \Sigma_F^{-1}(x - EF),$$

where EF is the mean vector of the distribution and Σ_F is the covariance matrix of the distribution, defined in (1.4). Define the Mahalabonis depth function as

$$D_M(x) = (1 + d_M(x))^{-1}.$$

The data version is obtained by applying the sample mean and covariance.

2. *Half-space depth.* The half-space depth function is defined as

$$D_h(x) = \inf\{P_F(H) : H \text{ is a closed half-space in } \mathbf{R}^d \text{ such that } x \in H\}.$$

The data version is obtained by applying the empirical distribution function. We may also define the sample version in the following way: The depth of a point $x \in \mathbf{R}$, with respect to a one-dimensional data set, is the minimum of the number of data points in one side of x. The depth of a point $x \in \mathbf{R}^d$, with respect to a high-dimensional data set, is the smallest depth in any one-dimensional projection of the data set. That is, the depth of a point $x \in \mathbf{R}^d$, with respect to a data set $\{X_1, \ldots, X_n\} \subset \mathbf{R}^d$, is

$$D_{h,n}(x) = n^{-1} \min\{\#\{i : u^T X_i \le u^T x\} : \|u\| = 1\}.$$

The half-space depth was defined by Hodges (1955) and Tukey (1975).

3. *Data depths based on simplexes.* Let

$$d_O(x) = E_F \text{volume}[S(x, X_1, \ldots, X_d)],$$

where $S(x, X_1, \ldots, X_d) \subset \mathbf{R}^d$ is the closed simplex with vertices x and $X_1, \ldots, X_d$, and $X_1, \ldots, X_d$ are i.i.d. with distribution F. We define the Oja depth by

$$D_O(x) = (1 + d_O(x))^{-1},$$

as in Oja (1983). We define the simplicial depth by

$$D_s(x) = E_F I_{S(X_1, \ldots, X_{d+1})}(x),$$

where $S(X_1, \ldots, X_{d+1}) \subset \mathbf{R}^d$ is the closed simplex with $X_1, \ldots, X_{d+1}$ as vertices, and $X_1, \ldots, X_{d+1}$ are i.i.d. with distribution F. Simplicial depth was defined by Liu (1990).

Data based versions may be calculated with U-statistics. Define $d_n(x)$ as the average of $h(x, x_{i_1}, \ldots, x_{i_m})$ taken over all subsets $\{i_1, \ldots, i_m\}$ of size m from $\{1, \ldots, n\}$. That is,

$$d_n(x) = \binom{n}{m}^{-1} \sum_{*} h(x, x_{i_1}, \ldots, x_{i_m}),$$

where the summation is taken over all m-lets $(i_1, \ldots, i_m)$ such that $1 \le i_1 \le \cdots \le i_m \le n$. For the Oja depth we choose $m = d$ and h as the volume of the simplex formed by x and d data points: $h(x) = \text{volume}[S(x, X_{i_1}, \ldots, X_{i_d})]$, and $D_{O,n}(x) = (1 + d_n(x))^{-1}$. For the simplicial depth we choose $m = d+1$ and h as the indicator of a simplex formed by $d + 1$ data points: $h(x) = I_{S(X_{i_1}, \ldots, X_{i_{d+1}})}(x)$, and $D_{s,n}(x) = d_n(x)$. In the case of the simplicial depth one computes the fraction of the sample random simplexes containing the point x.

4. *Data depths based on the L_1 norm.* Define

$$D_{L_1}(x) = (1 + d_{L_1}(x))^{-1},$$

where we may consider the following possibilities:

$$d_{L_1}(x) = \begin{cases} E\|x - X\| \\ E\|x - X\|_{\Sigma^{-1}}, \end{cases}$$

where Σ is the covariance matrix and $\|x\|^2_{\Sigma^{-1}} = x^T \Sigma^{-1} x$. These depth functions were studied, for example, by Zuo and Serfling (2000). Note that after defining med(F) with an L_1 depth function, the Mahalabonis distance may be modified to $d_{M,2}(x) = \|x - \text{med}(F)\|_{\tilde{\Sigma}^{-1}}$, where $\tilde{\Sigma}$ is a suitable covariance matrix. This distance was proposed in Serfling (2002*a*).

5. *Convex hull peeling depth.* The convex hull peeling depth has only the data version, and this depth is defined only at the data points $X_1, \ldots, X_n$, but not at arbitrary points $x \in \mathbf{R}^d$. Let the data points at the perimeter of the convex hull of the data have the smallest depth 1. Remove these data points and form the convex hull of the remaining data points. The points at the perimeter have the one step larger depth 2. Continue until a nested sequence of convex hulls is formed and all data points have been reached. Convex hull peeling depth was defined by Barnett (1976).

Liu, Parelius and Singh (1999) give a review of applications of data depthness. Many of the notions of a data depth are related to level sets or to estimates of the level sets; consider the depth regions

$$\{x \in \mathbf{R}^d : D(x) \ge \lambda\}.$$

Depth regions are often close to level sets of the density f of a unimodal distribution. Indeed one way to define a depth function is to take the depth equal to the density: $D(x) = f(x)$. We may call this the density depth, or the likelihood depth.

When the density is elliptical, as defined in Section 10.1.3, then the depth regions of Mahalabonis depth are equal to the level sets of the density. The simplicial depths measure the local concentration of the probability mass and are thus related to the density depth. The depth regions of convex peeling depth estimate the level sets of densities whose level sets are convex.

Mode A mode may be defined to be a point at which the density has a local maximum. In the univariate case the unimodality is relatively straightforward to define.

Definition 1.3 (Unimodal 1D distribution.) *A distribution is* unimodal *about a mode $\mu \in \mathbf{R}$, if the distribution function $F(t) = P((-\infty, t])$ is convex on $(-\infty, \mu)$ and concave on (μ, ∞).*

In the multivariate case there are many alternative definitions of unimodality. A very general definition is the following:

Definition 1.4 (Star unimodality.) *A distribution with density f is* star unimodal *about $\mu \in \mathbf{R}^d$, if for every $\lambda > 0$ the level set $\{x : f(x) \geq \lambda\}$ is star-shaped about μ.*

Set $A \subset \mathbf{R}^d$ is called star shaped when there exists a reference point $\mu \in A$ so that for every $x \in A$ the line segment joining x and μ belongs to A: $\{\mu + p(x - \mu) : p \in [0, 1]\} \subset A$; see also Definition 5.2. The set of reference points about which A is star shaped is called the kernel of A. The star unimodality is equivalent to the following two conditions:

1. A distribution with density f is star unimodal about $\mu \in \mathbf{R}^d$ if $f(\mu + ux) \leq f(\mu + tx)$ for every $0 < t < u < \infty$ and $x \neq 0$. That is, the density is decreasing on rays emanating from μ.

2. A distribution is star unimodal about $\mu \in \mathbf{R}^d$ if it belongs to the closed convex hull of the set of all uniform distributions on sets that are star shaped about μ. Here the topology on the set of probability measures is the topology of weak convergence. This criterion does not require that the density exists.

The equivalence of these two conditions with the star unimodality is proved in Dharmadhikari and Joag-dev (1988, p. 38). Figure 1.38(a) shows a density that is not star unimodal. We may define the unimodality so that also this density is unimodal. (Figure 1.38(b) is discussed in the connection of Lemma 2.1.)

Definition 1.5 (Level set unimodality.) *A distribution with density f is* level set unimodal, *if for every $\lambda > 0$ the level set $\{x : f(x) \geq \lambda\}$ is connected.*

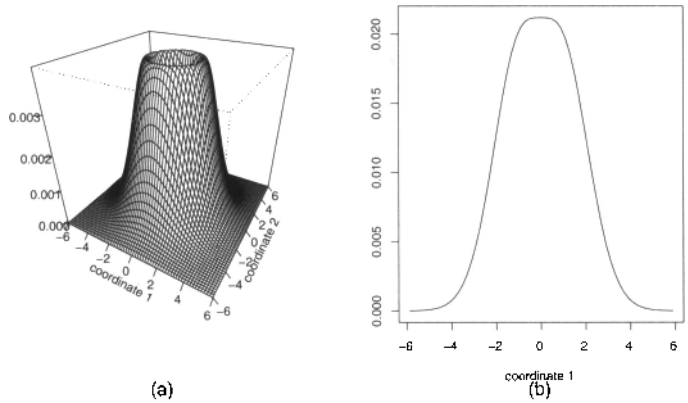

Figure 1.38 *(Star unimodality and hat function.)* (a) A density which is not star unimodal. (b) A marginal density of the density in (a). The density is defined in (2.5).

A density that is unimodal according to Definition 1.4 or Definition 1.5 may have many maxima, and we may define the "mode" to be the set of maxima. For a uniform distribution with a star-shaped or connected support the whole interior of the support could be defined as the mode. It is possible to generalize Definition 1.4 to the definition of *α-unimodality*.

Definition 1.6 (α-unimodality.) *A distribution P is* α-unimodal *about $\mu \in \mathbf{R}^d$, for the index $\alpha \geq 0$ if, for $X \sim P$, $t^\alpha E[g(t(X-\mu))]$ is nondecreasing in $t \in (0, \infty)$, for every bounded, nonnegative, Borel measurable function $g : \mathbf{R}^d \to \mathbf{R}$.*

If a distribution is α-unimodal, then the distribution is β-unimodal for $\beta > \alpha$. A distribution with density f is α-unimodal about $\mu \in \mathbf{R}^d$, if $t^{d-\alpha} f(\mu + tx)$ is non-increasing on $t \in (0, \infty)$, for every $x \neq 0$. This is proved in Dharmadhikari and Joag-dev (1988, Thm. 3.6). A star unimodal distribution is α-unimodal with $\alpha = d$. An interest in α-unimodality comes from the fact that the marginals of α-unimodal distributions are α-unimodal. This does not hold for the other two concepts of multivariate unimodality which we have defined; see Lemma 2.1.

We have followed Dharmadhikari and Joag-dev (1988) in the exposition of unimodality, except for Definition 1.5.

Estimation of the modes (the location of the modes) may be done with the help of density estimators: one estimates the density and finds the modes of the estimate. This was discussed in Parzen (1962). The estimation of the location of the mode is as difficult as the estimation of the gradient (first derivative) of the density: for s-smooth density the rate of convergence is $n^{(s-1)/(2s+d)}$.

1.7.2 Dispersion

Dispersion Matrix A dispersion matrix or a scatter matrix is defined as a functional that maps d-dimensional random vectors to $(d \times d)$-dimensional matrices. A dispersion matrix $\Sigma(X)$ of a random vector $X = (X_1, \ldots, X_d)$ is a symmetric and positive semi-definite matrix that is affine invariant. Affine invariance means that

$$\Sigma(a + BX) = B\Sigma(X)B^T,$$

for every d-vector a and invertible $d \times d$-matrix B.

In order to express the shape of an elliptical distributions with a dispersion matrix it is useful to write the matrix in a canonical form. A positive semi-definite symmetric matrix Σ can be written as

$$\Sigma = \lambda UCU^T,$$

where $\lambda > 0$, U is an orthogonal matrix of eigenvectors, and C is a diagonal matrix of standardized eigenvalues, so that $|\det C| = 1$. We may call λ the *scale*, λ^2 the *generalized variance*, C the *shape*, and U the *orientation*.

Below are some examples of dispersion matrices.

- The *covariance matrix* of distribution F, or the random vector $X \sim F$, is defined by
$$\Sigma = \mathrm{Cov}(X) = E\left[(X - EF)(X - EF)^T\right]. \tag{1.4}$$
The covariance matrix can be estimated by
$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)(X_i - \bar{X}_n)^T, \tag{1.5}$$
where $\bar{X}_n$ is the arithmetic mean.

- The positive semi-definite symmetric matrix Σ in the definition of an elliptical distribution in (10.4) is a dispersion matrix.

- *Modal dispersion* is defined for a random vector X with density f_X by
$$\Sigma(X) = -\left(D^2 f_X(\mathrm{mode}(f_X))\right)^{-1},$$
where $D^2(f)$ is the Hessian matrix of the second derivatives of f and $\mathrm{mode}(f)$ is the mode of f; see Meucci (2005), page 368. The modal dispersion may be motivated by the observation: the larger the absolute value of the second derivative, the thinner is the density function around its mode and thus the less disperse is the distribution. Clearly, modal dispersion is a local concept (unlike the covariance matrix). Also, in order for modal dispersion to be defined, we need to assume that the density is unimodal.

Note that we define various measures for dependency in Section 9.3.2. If we have a measure $\rho_{i,j}$ for the dependency of $X_i, X_j \in \mathbf{R}$, then we may construct the $d \times d$ matrix R with the elements $\rho_{i,j}$, and a similar matrix may be constructed for data versions. These matrices R are not always dispersion matrices.

Dispersion Vector We may use the generalized correlation, defined in (9.5), to define a dispersion vector. The generalized correlation is the correlation between a real valued random variable and a vector. Denote with Z_i the vector $X = (X_1, \ldots, X_d)$ without the ith component: $Z_i = (X_1, \ldots X_{i-1}, X_{i+1}, \ldots X_d)$. Define similarly to (9.5),

$$\rho_{G,i} = \max_{\phi} \rho(X_i, \phi^T Z_i),$$

where ρ is the linear correlation defined in (9.3), and the maximum is over k-vectors. Then we may use the vector $(\rho_{G,1}, \ldots, \rho_{G,d})$ to describe the spread of $(X_1, \ldots, X_d)$. However, a typical characterization of the spread is the covariance matrix, or more generally a dispersion matrix.

1.7.3 Higher Order Moments

We may use higher moments to characterize the distribution. Let $i_1, \ldots, i_k \in \{1, \ldots, d\}$. The kth central moment is defined as

$$E\left[(X_{i_1} - EX_{i_1}) \cdots (X_{i_k} - EX_{i_k})\right].$$

For a univariate distribution we may define the kth standardized moment as

$$\frac{\mu_k}{\sigma^k},$$

where $\mu_k = \int_{-\infty}^{\infty} (x - EF)^k f(x)\, dx$ is the kth moment about the mean and σ is the standard deviation, $\sigma^2 = \int_{\mathbf{R}^d} (x - EF)^2 f(x)\, dx$, and where f is the density and EF is the expectation. Value μ_3/σ^3 is called the skewness and μ_4/σ^4 is called the kurtosis. The kurtosis measures the heaviness of the tails of the density. Often the kurtosis is defined as $\mu_4/\sigma^4 - 3$, to make the kurtosis of the normal distribution equal to zero. For a definition of skewness and kurtosis in terms of cumulants, see (1.9).

1.8 DIMENSION REDUCTION OF DATA

1.8.1 Principal Components

Principal component analysis tries to find linear combinations of the variables that will contain as much information as possible. We reduce the dimension by choosing only the most important linear combinations.

Definition of the Principal Component Transformation In the theoretical version of the principal component transformation we use the spectral representation of the covariance matrix Σ of random vector $X \in \mathbf{R}^d$, as defined in (1.4). In the empirical version the covariance matrix Σ is replaced by the empirical covariance matrix Σ_n, defined in (1.5).

1. The *spectral representation*, or the *eigen decomposition*, of Σ is given by

$$\Sigma = A\Lambda A^T, \tag{1.6}$$

where Λ is the $d \times d$ diagonal matrix of the *eigenvalues* of Σ and A is the orthogonal $d \times d$ matrix of the *eigenvectors*, or the *principal components*, of Σ (the columns of A are the eigenvectors); see Appendix B.7. If Σ has rank d, then all λ_i are positive. Principal component transformation is defined by

$$Y = A^T(X - EX) \in \mathbf{R}^d. \tag{1.7}$$

The covariance matrix of Y is the matrix Λ of eigenvalues: $E(YY^T) = \Lambda$.

2. The empirical spectral representation can be calculated by finding the *singular value decomposition* of the centering of the $n \times d$ data matrix $\mathbb{X}_n$:

$$\mathbb{X}_n - 1_{n\times 1}\bar{X}_n^T = UDV^T,$$

where U is a $n \times d$ orthogonal matrix of left singular vectors, V is a $d \times d$ orthogonal matrix of right singular vectors, and D is a $d \times d$ diagonal matrix of singular values; see Appendix B.7. Here we denote with $1_{n\times 1}$ the $n \times 1$ column vector of ones and $\bar{X}_n$ is the $d \times 1$ column vector of arithmetic means. Then

$$\Sigma_n = \frac{1}{n}\left(UDV^T\right)^T \cdot UDV^T = \frac{1}{n}VD^2V.$$

Thus $A = V$ and $\Lambda = n^{-1}D^2$.

Data sphering is the principal component transformation followed by the normalization of the marginal variances to unity. Examples of data sphering are given in Section 9.2.1.

Optimality of the Principal Component Transformation Let Y_1 and Y_2 be the first two components of Y in (1.7). There is no other linear combination with larger variance than Y_1, and Y_2 is the linear combination, uncorrelated with Y_1, with the second largest variance. More precisely, let

$$a_1 = \text{argmax}_{a:\|a\|=1}\text{Var}(a^T X). \tag{1.8}$$

Then a_1 is the column of A corresponding to the largest eigenvalue. Let

$$a_2 = \text{argmax}_{a:\|a\|=1, a\perp a_1}\text{Var}(a^T X).$$

Then a_2 is the column of A corresponding to the second largest eigenvalue.

Dimension Reduction with the Principal Component Transformation We can define the dimension reduction for the theoretical version or for the empirical version.

1. The theoretical dimension reduction proceeds as follows: choose random vector $(Y_1, \ldots, Y_{d'})^T$, where $1 < d' \leq d$ and Y_i are the elements of vector Y defined in (1.7). Now d' could be chosen so that the ratio

$$\frac{\sum_{k=1}^{d'} \lambda_k}{\sum_{k=1}^{d} \lambda_k}$$

is sufficiently close to 1, for example, larger than 0.95, where λ_i, $i = 1, \ldots, d$, are the eigenvalues (the diagonal elements of Λ); we assume that the eigenvalues are in order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$.

2. The empirical dimension reduction finds new data matrix $\mathbb{Y}_n$ of dimension $n \times d'$,

$$\mathbb{Y}_n = \left(\mathbb{X}_n - 1_{n\times 1} X_n^T\right) \times (a_1 \cdots a_{d'}),$$

where $\mathbb{X}_n$ is the original $n \times d$ data matrix, and $(a_1 \cdots a_{d'})$ is the $d \times d'$ matrix of the d' first eigenvectors of Σ_n. When principal component analysis is used to reduce dimension, we obtain new random variables that are linear combinations of the original random variables. Thus it may be difficult to interpret the new random variables.

1.8.2 Projection Pursuit

Projection pursuit tries to find interesting low-dimensional projections of high-dimensional data by maximizing a projection index. In the theoretical version of projection pursuit we find a projection matrix A that maximizes $Q(AX)$, where $X \in \mathbf{R}^d$ is a random vector and Q is the projection index. Huber (1985) mentions the following projection indexes, which are defined for the case of 1D projections. In all cases one can show that $Q(X) \geq 0$ and $Q(X) = 0$ if X is normal.

1. *(Cumulant.)* Let

$$Q(X) = |c_m(X)|/c_2(X)^{m/2}, \qquad m > 2,$$

where c_m is the mth cumulant:

$$c_m = \frac{1}{i^m} \frac{\partial^m}{\partial t^m} \log E \exp\{itX\}\bigg|_{t=0}, \qquad m \geq 2, \tag{1.9}$$

where i is the imaginary unit. For $m = 3$ we obtain the absolute skewness and for $m = 4$ the absolute kurtosis.

2. *(Fisher information.)* Let

$$Q(X) = \mathrm{Var}(X) I(X) - 1,$$

where $I(X)$ is the Fisher information

$$I(X) = \frac{\int_{-\infty}^{\infty} (f_X')^2}{f_X},$$

where f_X is the density of X.

3. *(Shannon entropy.)* Let

$$Q(X) = -S(x) + \log\left(\text{std}(X)\sqrt{2\pi e}\right),$$

where $S(X)$ is the Shannon entropy

$$S(x) = -\int_{-\infty}^{\infty} f_X \log f_X,$$

where f_X is the density of X.

Note that principal component analysis can be seen as projection pursuit, where the projection index is the variance, see (1.8).

Huber (1985) gives a review of projection pursuit. The term "projection pursuit" was coined by Friedman and Tukey (1974). Further studies of projection pursuit include Cook, Buja and Cabrera (1993), Cook, Buja, Cabrera and Hurley (1995).

1.8.3 Self-organizing Maps

Self organizing map (SOM) is a discretized scatter plot of a two-dimensional data that is obtained from the original data by a nonlinear mapping. SOM finds k cluster centers, or prototypes, and each prototype is associated with a two-dimensional index in $I = \{i_1, \ldots, i_N\} \times \{j_1, \ldots, j_M\}$. Each observation in $\{X_1, \ldots, X_n\}$ is mapped to the closest cluster center, and thus each observation is associated to an index in I. The plot of those indexes is a self-organizing map of the data. Kohonen (2001) gives an exposition of SOM. We present the batch version of SOM following Hastie, Tibshirani and Friedman (2001).

1. Choose k cluster centers $\mathcal{M} = \{m^1, \ldots, m^k\} \subset \mathbf{R}^d$. Each cluster center is associated with an index $l(m) \in I$.

2. Iterate the following steps until convergence.

   (a) Go through the cluster centers $m \in \mathcal{M}$.

   (b) Consider cluster center m. Find the set of cluster centers in the neighborhood of m: $\mathcal{M}_m = \{m' : \delta(l(m'), l(m)) < \rho\}$, where δ is, for example, the Euclidean distance.

   (c) Find the observations $\mathbb{X}_m$ that belong to the Voronoi neighborhoods of the cluster centers in $\mathcal{M}_m$:

$$\mathbb{X}_m = \bigcup\{\mathcal{X}_{m'} : m' \subset \mathcal{M}_m\},$$

   where $\mathcal{X}_{m'}$ is the Voronoi neighborhood of m' (the set of those observations that are closer to m' than to any other cluster center):

$$\mathcal{X}_{m'} = \{X_i : \|X_i - m'\| = \min\{\|(X_i - m''\| : m'' \in \mathcal{M}\}\}.$$

(d) Update m to be the weighted average of the observations in $\mathbb{X}_m$:

$$m = \frac{\sum_{X \in \mathbb{X}_m} w_X X}{\sum_{X \in \mathbb{X}_m} w_X},$$

where $w_X > 0$ are weights.

Define $\mu : \{X_1, \ldots, X_n\} \to \mathcal{M}$ to be the mapping of observations to the closest cluster center:

$$\|\mu(X_i) - X_i\| = \min\{\|m' - X_i\| : m' \in \mathcal{M}\}.$$

(Then we may write $\mathcal{X}_m = \{X_i : m = \mu(X_i)\}$, where $\mathcal{X}_m$ is the Voronoi neighborhood of m.)

Definition 1.7 (Self-organizing map.) *) A* Self-organizing map *of data* $\{X_1, \ldots, X_n\}$ *is the scatter plot of* $\{l(\mu(X_i)) : i = 1, \ldots, n\}$.

The application of the algorithm requires a choice for the radius $\rho > 0$ and the weights w_X. When ρ is so small that $\mathcal{M}_m = \{m\}$, then the cluster centers are equal to the cluster centers of the k-means algorithm, defined in Chapter 8.2.

1.8.4 Multidimensional Scaling

In multidimensional scaling one tries to find such mapping of the d-dimensional data to a two-dimensional Euclidean space that the distances of the observations remain as close to the original distances as possible; that is, one tries to find such a mapping $Q : \{X_1, \ldots, X_n\} \to \mathbf{R}^2$, that the stress functional

$$\sum_{1 \le i < j \le n} (\|X_i - X_j\| - \|Q(X_i) - Q(X_j)\|)^2$$

is minimized. Sammon's mapping uses the stress functional

$$\sum_{1 \le i < j \le n} \frac{(\|X_i - X_j\| - \|Q(X_i) - Q(X_j)\|)^2}{\|X_i - X_j\|}.$$

This stress functional emphasizes small distances. For example, the Newton–Raphson algorithm can be used for the minimization of the stress functional.