Introduction

We will analyze data that are given as an $n \times d$ matrix of real numbers. The number in the ith row and in the jth column is the measurement of the jth property of the ith object. For example, the objects might be companies and the properties might be the stock price, debt, number of employees, earnings, or the objects might be persons and the measurements might be height, weight, age.

I.1 SMOOTHING

A fundamental idea is to *smooth* the data. Smoothing means that we interpret the data as n realizations of d-dimensional identically distributed random vectors and estimate the density function of the observations. A density function is a function $\mathbf{R}^d \to \mathbf{R}$ that describes the distribution of the probability mass in the d-dimensional Euclidean space.

The invention of the Cartesian coordinate system made it possible to visualize two-dimensional data with scatter plots. One may interpret the $n \times d$ data matrix as n points in the d-dimensional Euclidean space, and when $d = 2$ to plot the points in the Cartesian coordinate system. Scatter plots may be used, for example, to find the regions where most of the observations are concentrated. Finding regions where the observations are concentrated translates into the problem of finding regions where the density function takes large values, since a density function is a function that describes

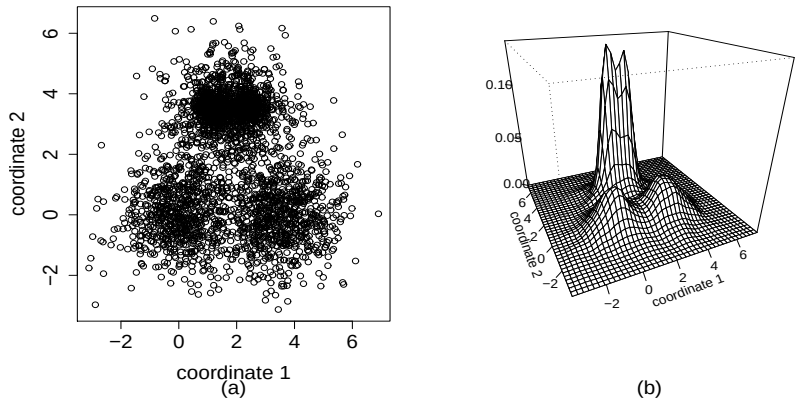

Figure I.1 Shown are (a) a scatter plot and (b) a kernel estimate of data of size $n = 3000$.

the denseness of the observations in the regions of the sample space. Figure I.1(a) shows a scatter plot of data of size $n = 3000$, and panel (b) shows a perspective plot of a kernel estimate. The figure illustrates the fact that the scatter plot makes it possible to identify individual points but the perspective plot of the density estimate visualizes the overall denseness of the observations.

I.2 VISUALIZATION

Functions $\mathbf{R}^d \to \mathbf{R}$ are much more complex objects than $n \times d$ data matrices. Thus it would seem that smoothing multivariate data is not useful in visualization. Is it possible to extend the success story of smoothing from the cases $d = 1$ and $d = 2$ to the cases $d \geq 3$? In our opinion only the very first steps have been made in finding visualization tools for multivariate functions, sets, and data.

The usual graphs seem simple to us, but the idea did not occur to the Greeks or Romans, or to Newton and Leibniz. Lambert (1779) used bivariate function graphs to analyze physical data, and Playfair (1786, 1801) invented the histogram, the pie chart, and the line graph. Still the progress in using these graphs in scientific reporting was slow, and even the scientifically trained readers had to learn how to cope with the new methods. (Spence and Lewandowsky 1990, pp. 13-14). Visualization makes the data visible, and seeing is one of the basic ways for humans to perceive reality. This does not mean that visualization is trivial and that new tools cannot be developed.

Humans can see only one-, two-, or three-dimensional objects. Thus visualization of multidimensional objects is possible only by transforming multidimensional objects to one, two, or three-dimensional objects. Furthermore science is communicated through paper and the computer screen, and this puts emphasis on the two-dimensional case. How to transform multidimensional objects to one, two, or

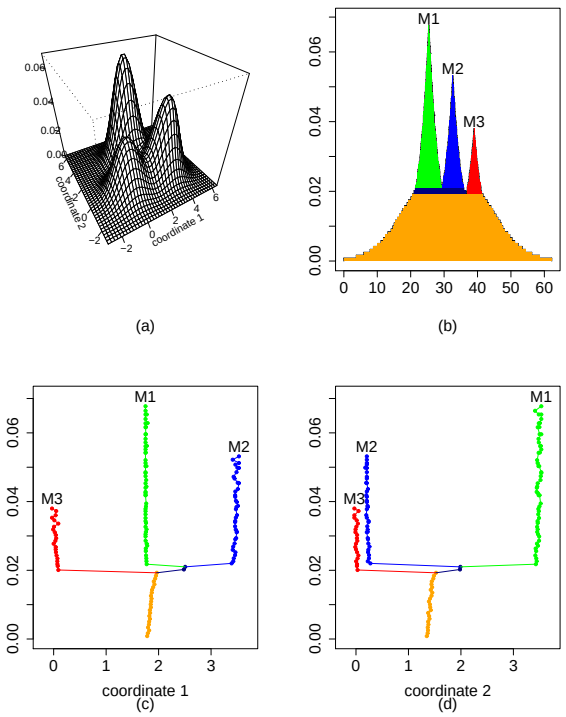

Figure I.2 Shown are (a) a perspective plot, (b) volume plot, and (c–d) barycenter plots of a three-modal density.

three-dimensional objects? A useful method is to apply projections and slices, but there are other possibilities.

Figure I.2, Figure I.3, and Figure I.4, show visualizations of objects of three different types: a function, a set, and data. These three objects have something in common: they are all three-modal objects. The visualizations in the figures reveal the modality of the objects by way of shape isomorphic transforms. These visualizations are one of the main subjects of the book.

In topology one says that two sets are topologically equivalent if they are diffeomorphic or homeomorphic. For example, a donut and a coffee cup may be said to be topologically equivalent. The definition of topological equivalence in terms of diffeomorphisms or homeomorphisms applies to objects of same dimension, but we are interested in the similarity of objects of different dimensions; a multivariate function may be visualized by a one- or two-dimensional function if these functions are similar in some sense.

"Visual geometry is like an experienced doctor's savvy in reading a patient's complexion, charts, and X-rays. Precise analysis is like the medical test results–the

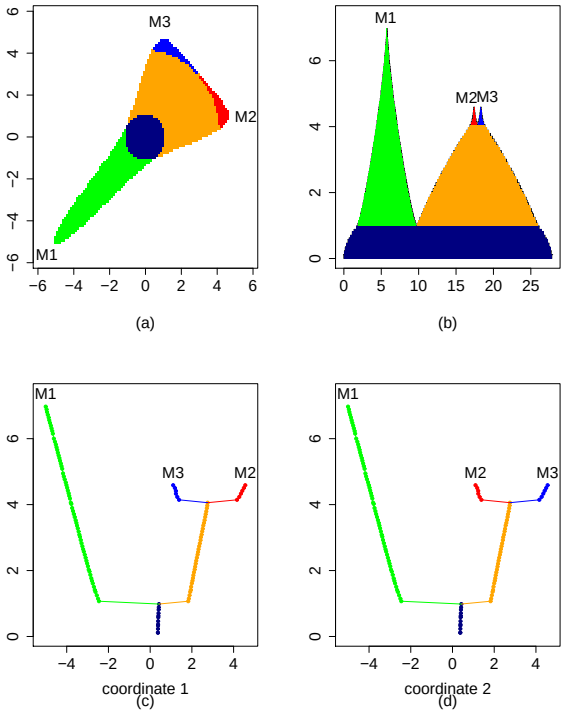

Figure I.3 Shown are (a) a standard plot, (b) radius plot, and (c–d) location plots of a level set of a density with Clayton copula, and Student marginals.

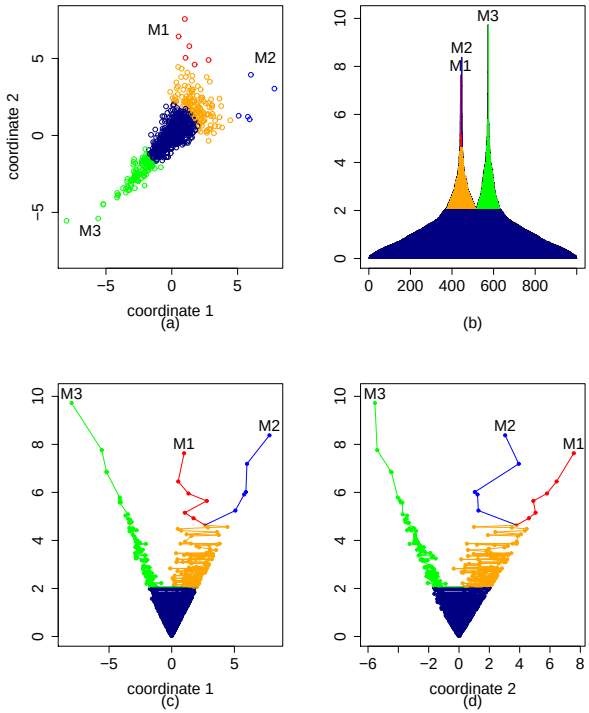

Figure I.4 Shown are (a) a scatter plot, (b) tail frequency plot, and (c–d) tail tree plots of a sample of size 1000 from a density with Clayton copula and Student marginals.

raw numbers of blood pressure and chemistry." (Mandelbrot and Hudson 2004, the prelude). Visualization cannot replace probabilistic inference, but probabilistic inference cannot alone be sufficient for scientific inference, without the support of visualization. Sometimes graphical methods are the only tool we need. For example, when the sample size is very large, we do not need to worry about the random fluctuation part of the experiment.

At least three categories of research on visualization have been classified: statistical visualization or data visualization, scientific visualization, and information visualization. Data visualization studies the direct visualization of data matrices, including the visualization of categorical data. Scientific visualization has concentrated on the visualization of 3D objects, functions, and processes, addressing the issues of industrial design and the medical, chemical, and meteorological visualization. Information visualization has addressed the visualization of various kinds of abstract data structures, like networks and text corpuses. Our main emphasis is on the visualization of multivariate functions. The research on the visualization of functions can be seen as a part of the discipline of information visualization. Our basic setting is to analyze statistical data and thus our research could also be seen as belonging to statistical visualization.

I.3 DENSITY ESTIMATION

Multivariate density estimation is difficult. Parametric Gaussian models fail because they have $2d + d(d-1)/2$ parameters (d parameters for the mean, d parameters for the diagonal of the covariance matrix, and $d(d-1)/2$ parameters for the off-diagonals of the symmetric covariance matrix). Nonparametric estimators that use local averaging fail because local neighborhoods are almost empty of observations in high-dimensional Euclidean spaces. Although some classical methods fail, this does not mean that some other methods could not work. There exists a rich and growing population of density estimators that add to the toolbox of fully parametric and fully nonparametric methods.

The additions to the toolbox could include structured nonparametric methods that utilize structural restrictions in the underlying function. Consider, for example, estimating a multivariate density with a product density, or estimating a regression function with an additive function. Consider imposing shape restrictions like unimodality or imposing structural restrictions on the level sets of the density. Recent additions to the toolbox of density estimators include the estimators based on semiparametric models and mixture models. Infinite mixture models are convex hulls of a base class of densities. This leads to the use of ensemble methods, like bootstrap aggregation, boosting, and stagewise minimization estimators.

Density estimation is a high-precision tool for statistical inference. It can give detailed knowledge about the distribution. Functions defined in moderate-dimensional Euclidean spaces, say four- or five-dimensional spaces, can be extremely complex, and it can be almost impossible to detect all features of a joint distribution of four random variables. Sometimes the data contain hundreds of variables, and there is

no hope to reach detailed knowledge about the full joint distribution with a finite amount of measurements. In this case it may be useful to apply dimension reduction techniques before continuing the analysis. Just like a sculptor starts with a hammer and a chisel to create the first contours of the sculpture, and then proceeds with high precision instruments to create the final details, a scientist could start with dimension reduction techniques and then proceed with density estimation. Statistics needs different kinds of tools to be used for different purposes. The right tools are chosen taking into account the available material and taking into account the aims of the work.

I.4 PLAN OF THE BOOK

Part I of the book covers visualization of multivariate functions, sets, data, and scales of multivariate density estimates. Part II gives basic mathematical tools to analyze asymptotically the behavior of multivariate density estimators and describes algorithms that are needed in visualization and in estimation of multivariate densities. Part III presents a toolbox of multivariate density estimators.

I.5 WEB PAGE AND THE CODE

Our hope is that the book satisfies the requirements of reproducible research. We provide software packages to reproduce the main figures and experiments of the book. The R-packages "denpro"and "delt" may be downloaded from the Web page http://www.denstruct.net or from the Web page http://r-project.org. The Web page of the book contains instructions for applying the packages. The Web page of the book contains also the colored figures of the book and the code for reproducing the figures.

I.6 BIBLIOGRAPHIC NOTES

The classic introductions to multivariate density estimation are those by Tapia and Thompson (1978), who discuss penalized likelihood density estimation, Silverman (1986), who considers applications of kernel density estimation, and Scott (1992), who addresses issues of visualization. Kernel estimation is studied in Wand and Jones (1995). A mathematical exposition with the L_1 view is given by Devroye and Györfi (1985) and Devroye (1987). An applied view is given by Simonoff (1996). Efromovich (1999) covers curve estimation with an emphasis on series methods. Tsybakov (2004) covers asymptotic minimax theory of density estimation.

A semiological study of graphics is given by Bertin (1967, 1981). Tukey (1977) gives a foundation for exploratory data analysis. Visualization of information is considered in Tufte (1983, 1990, 1997). Cleveland (1993*b*, 1994) considers principles of graph construction and strategies for data analysis, treating curve fitting as a visualization tool. The topological concepts of scientific visualization are presented in Fomenko and Kunii (1997). Information visualization, as visualization of graphs,

trees, knowledge domains, and virtual environments, is discussed in Chen (2004). Spence (2001) treats general information visualization and includes also classical statistical visualization from an information visualization viewpoint.