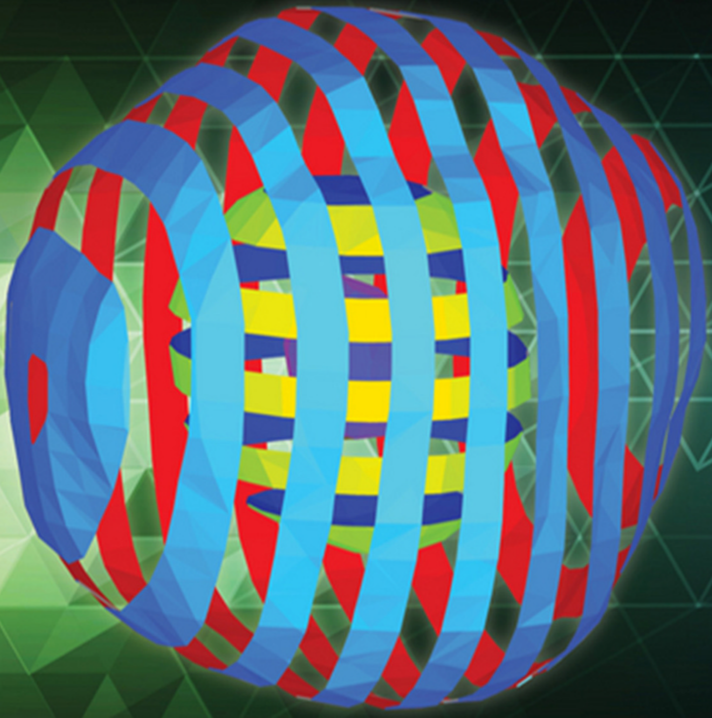
Wiley Series in Probability and Statistics

MULTIVARIATE DENSITY ESTIMATION

Theory, Practice, and Visualization

David W. Scott

SECOND EDITION



WILEY

MULTIVARIATE DENSITY ESTIMATION

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

MULTIVARIATE DENSITY ESTIMATION

Theory, Practice, and Visualization

Second Edition

DAVID W. SCOTT
Rice University
Houston, Texas

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Scott, David W., 1950–
Multivariate density estimation : theory, practice, and visualization / David W. Scott. – Second edition.
pages cm
Includes bibliographical references and index.
ISBN 978-0-471-69755-8 (cloth)
1. Estimation theory. 2. Multivariate analysis. I. Title.
QA276.8.S28 2014
519.5′35–dc23
2014043897

Set in 10/12pts Times Lt Std by SPi Publisher Services, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

1 2015

To Jean, Hilary,
Elizabeth, Warren,
and my parents, John
and Nancy Scott

CONTENTS

PREFACE TO SECOND EDITION

The past 25 years have seen confirmation of the importance of density estimation and nonparametric methods in modern data analysis, in this era of “big data.” This updated version retains its focus on fostering an intuitive understanding of the underlying methodology and supporting theory. I have sought to retain as much of the original material as possible and, in particular, the point of view of its development from the histogram. In every chapter, new material has been added to highlight challenges presented by massive datasets, or to clarify theoretical opportunities and new algorithms. However, no claim to comprehensive coverage is professed.

I have benefitted greatly from interactions with a number of gifted doctoral students who worked in this field—Lynette Factor, Donna Nezames, Rod Jee, Ferdie Wang, Michael Minnotte, Steve Sain, Keith Baggerly, John Salch, Will Wojciechowski, H.-G. Sung, Alena Oetting, Galen Papkov, Eric Chi, Jonathan Lane, Justin Silver, Jaime Ramos, and Yeshaya Adler—their work is represented here. In addition, contributions were made by many students taking my courses. I would also like to thank my colleagues and collaborators, especially my co-advisor Jim Thompson and my frequent co-authors George Terrell (VPI), Bill Szewczyk (DoD) and Masahiko Sagae (Kanazawa University). They have made the lifetime of learning, teaching, and discovery especially delightful and satisfying. I especially wish to acknowledge the able help of Robert Kosar in assembling the final versions of the color figures and reviewing new material.

Not a few mistakes have been corrected. For example, the constant in the expression for the asymptotic mean integrated squared error for the multivariate histogram in Theorem 3.5 is now correct. The content of Tables 3.6 and 3.7 has been modified accordingly, and the effect of dimension on sample size is seen to be even more dramatic in the corrected version. Any mistakes remain the responsibility of the author, who would appreciate hearing of such. All will be recorded in an appropriate repository.

Steve Quigley of John Wiley & Sons was infinitely patient awaiting this second edition until his retirement, and Kathryn Sharples completed the project. Steve made a freshly minted LaTeX version available as a starting point. All figures in S-Plus have been re-engineered into R. Figures in color or using color have been transformed to gray scale for the printed version, but the original figures will also be available in the same repository. In the original edition, I also neglected to properly acknowledge the generous support of the ARO (DAAL-03-88-G-0074 through my colleague James Thompson) and the ONR (N00014-90-J-1176).

As with the original edition, this revision would not have been possible without the tireless and enthusiastic support of my wife, Jean, and family. Thanks for everything.

Houston, Texas
August, 2014

David W. Scott

PREFACE TO FIRST EDITION

With the revolution in computing in recent years, access to data of unprecedented complexity has become commonplace. More variables are being measured, and the sheer volume of data is growing. At the same time, advancements in the performance of graphical workstations have given new power to the data analyst. With these changes has come an increasing demand for tools that can detect and summarize the multivariate structure in difficult data. Density estimation is now recognized as a tool useful with univariate and bivariate data; my purpose is to demonstrate that it is also a powerful tool in higher dimensions, with particular emphasis on trivariate and quadrivariate data. I have written this book for the reader interested in the theoretical aspects of nonparametric estimation as well as for the reader interested in the application of these methods to multivariate data. It is my hope that the book can serve as an introductory textbook and also as a general reference.

I have chosen to introduce major ideas in the context of the classical histogram, which remains the most widely applied and most intuitive nonparametric estimator. I have found it instructive to develop the links between the histogram and more statistically efficient methods. This approach greatly simplifies the treatment of advanced estimators, as much of the novelty of the theoretical context has been moved to the familiar histogram setting.

The nonparametric world is more complex than its parametric counterpart. I have selected material that is representative of the broad spectrum of theoretical results available, with an eye on the potential user, based on my assessments of usefulness, prevalence, and tutorial value. Theory particularly relevant to application or understanding is covered, but a loose standard of rigor is adopted in order to emphasize the methodological and application topics. Rather than present a cookbook of techniques, I have adopted a hierarchical approach that emphasizes the similarities among the different estimators. I have tried to present new ideas and practical advice, together with numerous examples and problems, with a graphical emphasis.

Visualization is a key aspect of effective multivariate nonparametric analysis, and I have attempted to provide a wide array of graphic illustrations. All of the figures in this book were composed using S, S-PLUS, Exponent Graphics from IMSL, and Mathematica. The color plates were derived from S-based software. The color graphics with transparency were composed by displaying the S output using the MinneView program developed at the Minnesota Geometry Project and printed on hardware under development by the 3M Corporation. I have not included a great deal of computer code. A collection of software, primarily Fortran-based with interfaces to the S language, is available by electronic mail at scottdw@rice.edu. Comments and other feedback are welcomed.

I would like to thank many colleagues for their generous support over the past 20 years, particularly Jim Thompson, Richard Tapia, and Tony Gorry. I have especially drawn on my collaboration with George Terrell, and I gratefully acknowledge his major contributions and influence in this book. The initial support for the high-dimensional graphics came from Richard Heydorn of NASA. This work has been generously supported by the Office of Naval Research under grant N00014-90-J-1176 as well as the Army Research Office. Allan Wilks collaborated on the creation of many of the color figures while we were visiting the Geometry Project, directed by Al Marden and assisted by Charlie Gunn, at the Minnesota Supercomputer Center.

I have taught much of this material in graduate courses not only at Rice but also during a summer course in 1985 at Stanford and during an ASA short course in 1986 in Chicago with Bernard Silverman. Previous Rice students Lynette Factor, Donna Nezames, Rod Jee, and Ferdie Wang all made contributions through their theses. I am especially grateful for the able assistance given during the final phases of preparation by Tim Dunne and Keith Baggerly, as well as Steve Sain, Monnie McGee, and Michael Minnotte. Many colleagues have influenced this work, including Edward Wegman, Dan Carr, Grace Wahba, Wolfgang Härdle, Matthew Wand, Simon Sheather, Steve Marron, Peter Hall, Robert Launer, Yasuo Amemiya, Nils Hjort, Linda Davis, Bernhard Flury, Will Gersch, Charles Taylor, Imke Janssen, Steve Boswell, I.J. Good, Iain Johnstone, Ingram Olkin, Jerry Friedman, David Donoho, Leo Breiman, Naomi Altman, Mark Matthews, Tim Hesterberg, Hal Stern, Michael Trosset, Richard Byrd, John Bennett, Heinz-Peter Schmidt, Manny Parzen, and Michael Tarter. Finally, this book could not have been written without the patience and encouragement of my family.

Houston, Texas
February, 1992

DAVID W. SCOTT

1

REPRESENTATION AND GEOMETRY OF MULTIVARIATE DATA

A complete analysis of multidimensional data requires the application of an array of statistical tools—parametric, nonparametric, and graphical. Parametric analysis is the most powerful. Nonparametric analysis is the most flexible. And graphical analysis provides the vehicle for discovering the unexpected.

This chapter introduces some graphical tools for visualizing structure in multidimensional data. One set of tools focuses on depicting the data points themselves, while another set of tools relies on displaying of functions estimated from those points. Visualization and contouring of functions in more than two dimensions is introduced. Some mathematical aspects of the geometry of higher dimensions are reviewed. These results have consequences for nonparametric data analysis.

1.1 INTRODUCTION

Classical linear multivariate statistical models rely primarily on analysis of the covariance matrix. So powerful are these techniques that analysis is almost routine for datasets with hundreds of variables. While the theoretical basis of parametric models lies with the multivariate normal density, these models are applied in practice to many kinds of data. Parametric studies provide neat inferential summaries and parsimonious representation of the data.

For many problems second-order information is inadequate. Advanced modeling or simple variable transformations may provide a solution. When no simple

Multivariate Density Estimation, First Edition. David W. Scott.

parametric model is forthcoming, many researchers have opted for fully "unparametric" methods that may be loosely collected under the heading of exploratory data analysis. Such analyses are highly graphical; but in a complex non-normal setting, a graph may provide a more concise representation than a parametric model, because a parametric model of adequate complexity may involve hundreds of parameters.

There are some significant differences between parametric and nonparametric modeling. The focus on optimality in parametric modeling does not translate well to the nonparametric world. For example, the histogram might be proved to be an inadmissible estimator, but that theoretical fact should not be taken to suggest histograms should not be used. Quite to the contrary, some methods that are theoretically superior are almost never used in practice. The reason is that the ordering of algorithms is not absolute, but is dependent not only on the unknown density but also on the sample size. Thus the histogram is generally superior for small samples regardless of its asymptotic properties. The exploratory school is at the other extreme, rejecting probabilistic models, whose existence provides the framework for defining optimality.

In this book, an intermediate point of view is adopted regarding statistical efficacy. No nonparametric estimate is considered wrong; only different components of the solution are emphasized. Much effort will be devoted to the data-based calibration problem, but nonparametric estimates can be reasonably calibrated in practice without too much difficulty. The "curse of optimality" might suggest that this is an illogical point of view. However, if the notion that optimality is all important is adopted, then the focus becomes matching the theoretical properties of an estimator to the assumed properties of the density function. Is it a gross inefficiency to use a procedure that requires only two continuous derivatives when the curve in fact has six continuous derivatives? This attitude may have some formal basis but should be discouraged as too heavy-handed for nonparametric thinking. A more relaxed attitude is required. Furthermore, many "optimal" nonparametric procedures are unstable in a manner that slightly inefficient procedures are not. In practice, when faced with the application of a procedure that requires six derivatives, or some other assumption that cannot be proved in practice, it is more important to be able to recognize the signs of estimator failure than to worry too much about assumptions. Detecting failure at the level of a discontinuous fourth derivative is a bit extreme, but certainly the effects of simple discontinuities should be well understood. Thus only for the purposes of illustration are the best assumptions given.

The notions of efficiency and admissibility are related to the choice of a criterion, which can only imperfectly measure the quality of a nonparametric estimate. Unlike optimal parametric estimates that are useful for many purposes, nonparametric estimates must be optimized for each application. The extra work is justified by the extra flexibility. As the choice of criterion is imperfect, so then is the notion of a single optimal estimator. This attitude reflects not sloppy thinking, but rather the imperfect relationship between the practical and theoretical aspects of our methods. Too rigid a point of view leads one to a minimax view of the world where nonparametric methods should be abandoned because there exist difficult problems.

Visualization is an important component of nonparametric data analysis. *Data visualization* is the focus of exploratory methods, ranging from simple scatterplots to sophisticated dynamic interactive displays. *Function visualization* is a significant component of nonparametric function estimation, and can draw on the relevant literature in the fields of scientific visualization and computer graphics. The focus of multivariate data analysis on points and scatterplots has meant that the full impact of scientific visualization has not yet been realized. With the new emphasis on smooth functions estimated nonparametrically, the fruits of visualization will be attained. Banchoff (1986) has been a pioneer in the visualization of higher dimensional mathematical surfaces. Curiously, the surfaces of interest to mathematicians contain singularities and discontinuities, all producing striking pictures when projected to the plane. In statistics, visualization of the smooth density surface in four, five, and six dimensions cannot rely on projection, as projections of smooth surfaces to the plane show nothing. Instead, the emphasis is on contouring in three dimensions and slicing of surfaces beyond. The focus on three and four dimensions is natural because one and two are so well understood. Beyond four dimensions, the ability to explore surfaces carefully decreases rapidly due to the curse of dimensionality. Fortunately, statistical data seldom display structure in more than five dimensions, so guided projection to those dimensions may be adequate. It is these threshold dimensions from three to five that are and deserve to be the focus of our visualization efforts.

There is a natural flow among the parametric, exploratory, and nonparametric procedures that represents a rational approach to statistical data analysis. Begin with a fully exploratory point of view in order to obtain an overview of the data. If a probabilistic structure is present, estimate that structure nonparametrically and explore it visually. Finally, if a linear model appears adequate, adopt a fully parametric approach. Each step conceptually represents a willingness to more strongly *smooth* the raw data, finally reducing the dimension of the solution to a handful of interesting parameters. With the assumption of normality, the mind's eye can easily imagine the d-dimensional egg-shaped elliptical data clusters. Some statisticians may prefer to work in the reverse order, progressing to exploratory methodology as a diagnostic tool for evaluating the adequacy of a parametric model fit.

There are many excellent references that complement and expand on this subject. In exploratory data analysis, references include Tukey (1977), Tukey and Tukey (1981), Cleveland and McGill (1988), and Wang (1978).

In density estimation, the classic texts of Tapia and Thompson (1978), Wertz (1978), and Thompson and Tapia (1990) first indicated the power of the nonparametric approach for univariate and bivariate data. Silverman (1986) has provided a further look at applications in this setting. Prakasa Rao (1983) has provided a theoretical survey with a lengthy bibliography. Other texts are more specialized, some focusing on regression (Müller, 1988; Härdle, 1990), some on a specific error criterion (Devroye and Györfi, 1985; Devroye, 1987), and some on particular solution classes such as splines (Eubank, 1988; Wahba, 1990). A discussion of additive models may be found in Hastie and Tibshirani (1990).

1.2 HISTORICAL PERSPECTIVE

One of the roots of modern statistical thought can be traced to the empirical discovery of correlation by Galton in 1886 (Stigler, 1986). Galton's ideas quickly reached Karl Pearson. Although best remembered for his methodological contributions such as goodness-of-fit tests, frequency curves, and biometry, Pearson was a strong proponent of the geometrical representation of statistics. In a series of lectures a century ago in November 1891 at Gresham College in London, Pearson spoke on a wide-ranging set of topics (Pearson, 1938). He discussed the foundations of the science of pure statistics and its many divisions. He discussed the collection of observations. He described the classification and representation of data using both numerical and geometrical descriptors. Finally, he emphasized statistical methodology and discovery of statistical laws. The syllabus for his lecture of November 11, 1891, includes this cryptic note:

> Erroneous opinion that Geometry is only a means of popular representation: *it is a fundamental method of investigating and analysing statistical material.* (his italics)

In that lecture Pearson described 10 methods of geometrical data representation. The most familiar is a representation "by columns," which he called the "histogram." (Pearson is usually given credit for coining the word "histogram" later in a 1894 paper.) Other familiar-sounding names include "diagrams," "chartograms," "topograms," and "stereograms." Unfamiliar names include "stigmograms," "euthygrams," "epipedograms," "radiograms," and "hormograms."

Beginning 21 years later, Fisher advanced the numerically descriptive portion of statistics with the method of maximum likelihood, from which he progressed on to the analysis of variance and other contributions that focused on the optimal use of data in parametric modeling and inference. In *Statistical Methods for Research Workers*, Fisher (1932) devotes a chapter titled "Diagrams" to graphical tools. He begins the chapter with this statement:

> The preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them.

An emphasis on optimization and the efficiency of statistical procedures has been a hallmark of mathematical statistics ever since. Ironically, Fisher was criticized by mathematical statisticians for relying too heavily upon geometrical arguments in proofs of his results.

Modern statistics has experienced a strong resurgence of geometrical and graphical statistics in the form of exploratory data analysis (Tukey, 1977). Given the parametric emphasis on optimization, the more relaxed philosophy of exploratory data analysis has been refreshing. The revolution has been fueled by the low cost of graphical workstations and microcomputers. These machines have enabled current work on *statistics in motion* (Scott, 1990), that is, the use of animation and kinematic display for visualization of data structure, statistical analysis, and algorithm performance. No longer are static displays sufficient for comprehensive analysis.

All of these events were anticipated by Pearsonand his visionary statistical computing laboratory. In his lecture of April 14, 1891, titled “The Geometry of Motion,” he spoke of the “ultimate elements of sensations we represent as motions in space and time.” In 1918, after his many efforts during World War I, he reminisced about the excitement created by wartime work of his statistical laboratory:

> The work has been so urgent and of such value that the Ministry of Munitions has placed eight to ten computers and draughtsmen at my disposal ... (Pearson, 1938, p. 165).

These workers produced hundreds of statistical graphs, ranging from detailed maps of worker availability across England (chartograms) to figures for sighting antiaircraft guns (diagrams). The use of stereograms allowed for representation of data with three variables. His “computers,” of course, were not electronic but human. Later, Fisher would be frustrated because Pearson would not agree to allocate his “computers” to the task of tabulating percentiles of the *t*-distribution. But Pearson’s capabilities for producing high-quality graphics were far superior to those of most modern statisticians prior to 1980. Given Pearson’s joint interests in graphics and kinematics, it is tantalizing to speculate on how he would have utilized modern computers.

1.3 GRAPHICAL DISPLAY OF MULTIVARIATE DATA POINTS

The modern challenge in data analysis is to be able to cope with whatever complexities may be intrinsic to the data. The data may, for example, be strongly non-normal, fall onto a nonlinear subspace, exhibit multiple modes, or be asymmetric. Dealing with these features becomes exponentially more difficult as the dimensionality of the data increases, a phenomenon known as the *curse of dimensionality*. In fact, datasets with hundreds of variables and millions of observations are routinely compiled that exhibit all of these features. Examples abound in such diverse fields as remote sensing, the US Census, geological exploration, speech recognition, and medical research. The expense of collecting and managing these large datasets is often so great that no funds are left for serious data analysis. The role of statistics is clear, but too often no statisticians are involved in large projects and no creative statistical thinking is applied. The goal of statistical data analysis is to extract the maximum information from the data, and to present a product that is as accurate and as useful as possible.

1.3.1 Multivariate Scatter Diagrams

The presentation of multivariate data is often accomplished in tabular form, particularly for small datasets with named or labeled objects. For example, Table B.1 contains economic data spanning the depression years of the 1930s, and Table B.2 contains information on a selected sample of American universities. It is easy enough to scan an individual column in these tables, to make comparisons of library size, for example, and to draw conclusions *one variable at a time* (see Tufte (1983) and Wang (1978)). However, variable-by-variable examination of multivariate data can be overwhelming and tiring, and cannot reveal any relationships among the variables. Looking at all pairwise scatterplots provides an improvement (Chambers et al., 1983). Data on four variables of three species of *Iris* are displayed in Figure 1.1. (A listing of the Fisher–Anderson *Iris* data, one of the few familiar four-dimensional datasets, may be found in several references and is provided with the S package (Becker et al., 1988)). What multivariate structure is apparent from this figure? The *setosa* variety does not overlap the other two varieties. The *versicolor* and *virginica* varieties are not as well separated, although a close examination reveals that they are almost nonoverlapping. If the 150 observations were unlabeled and plotted with the same symbol, it is likely that only two clusters would be observed. Even if it were known *a priori* that there were three clusters, it would still be unlikely that all three clusters would be properly identified. These alternative presentations reflect the two related problems of discrimination and clustering, respectively.

If the observations from different categories overlap substantially or have different sample sizes, scatter diagrams become much more difficult to interpret properly. The data in Figure 1.2 come from a study of 371 males suffering from chest pain (Scott et al., 1978): 320 had demonstrated coronary artery disease (occlusion or narrowing of the heart's own arteries) while 51 had none (see Table B.3). The blood fat concentrations of plasma cholesterol and triglyceride are predictive of heart disease, although the correlation is low. It is difficult to estimate the predictive power of these variables in this setting solely from the scatter diagram. A nonparametric analysis will reveal some interesting nonlinear interactions (see Chapters 5 and 9).

An easily overlooked practical aspect of scatter diagrams is illustrated by these data, which are integer valued. To avoid problems of overplotting, the data have been *jittered* or *blurred* (Chambers et al., 1983); that is, uniform $U(-0.5, 0.5)$ noise is

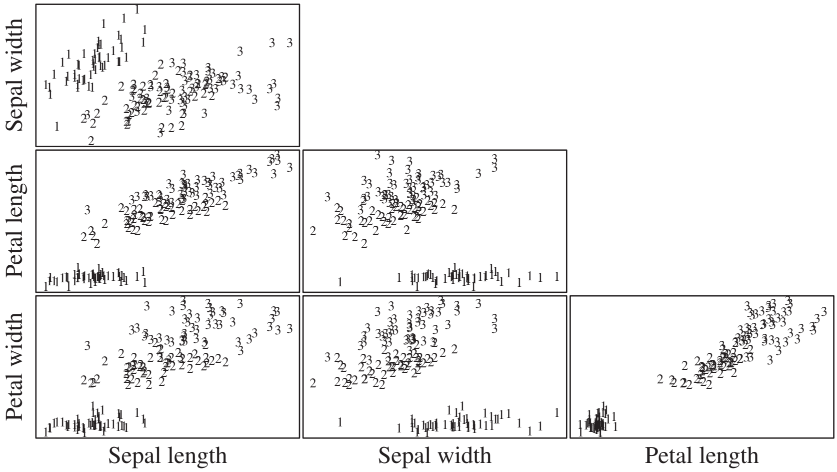


FIGURE 1.1 Pairwise scatter diagrams of the *Iris* data with the three species labeled. 1, setosa; 2, versicolor; 3, virginica.

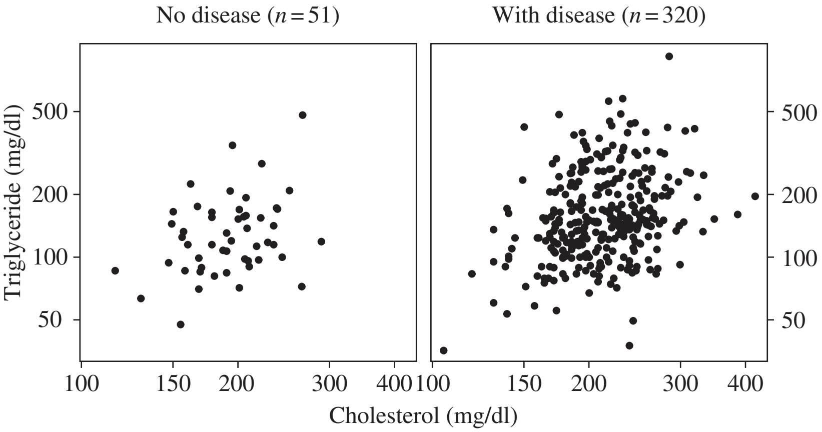


FIGURE 1.2 Scatter diagrams of blood lipid concentrations for 320 diseased and 51 nondiseased males.

added to each element of the original data. This trick should be regularly employed for data recorded with three or fewer significant digits (with an appropriate range on the added uniform noise). Jittering reduces visual miscues that result from the vertical and horizontal synchronization of regularly spaced data.

The visual perception system can easily be overwhelmed if the number of points is more than several thousand. Figure 1.3 displays three pairwise scatterplots derived from measurements taken in 1977 by the Landsat remote sensing system over a 5 mile by 6 mile agricultural region in North Dakota with $n = 22{,}932 = 117 \times 196$ *pixels* or picture elements, each corresponding to an area approximately 1.1 acres in size (Scott and Thompson, 1983; Scott and Jee, 1984). The Landsat instrument measures the intensity of light in four spectral bands reflected from the surface of the earth. A principal components transformation gives two variables that are commonly referred to as the "brightness" and "greenness" of each pixel. Every pixel is measured at regular intervals of approximately 3 weeks. During the summer of 1977, six useful replications were obtained, giving 24 measurements on each pixel. Using an agronometric growth model for crops, Badhwar et al. (1982) nonlinearly transformed this 24-dimensional data to three dimensions. Badhwar described these synthetic variables, (x_1, x_2, x_3), as (1) the calendar time at which peak greenness is observed, (2) the length of crop ripening, and (3) the peak greenness value, respectively. The scatter diagrams in Figure 1.3 have also been enhanced by jittering, as the raw data are integers between $(0, 255)$. The use of integers allows compression to eight bits of computer memory. Only structure in the boundary and tails is readily seen. The overplotting problem is apparent and the blackened areas include over 95% of the data. Other techniques to enhance scatter diagrams are needed to see structure in the bulk of the data cloud, such as plotting random subsets (see Tukey and Tukey (1981)).

Pairwise scatter diagrams lack one important property necessary for identifying more than two-dimensional features—strong interplot linkage among the plots. In

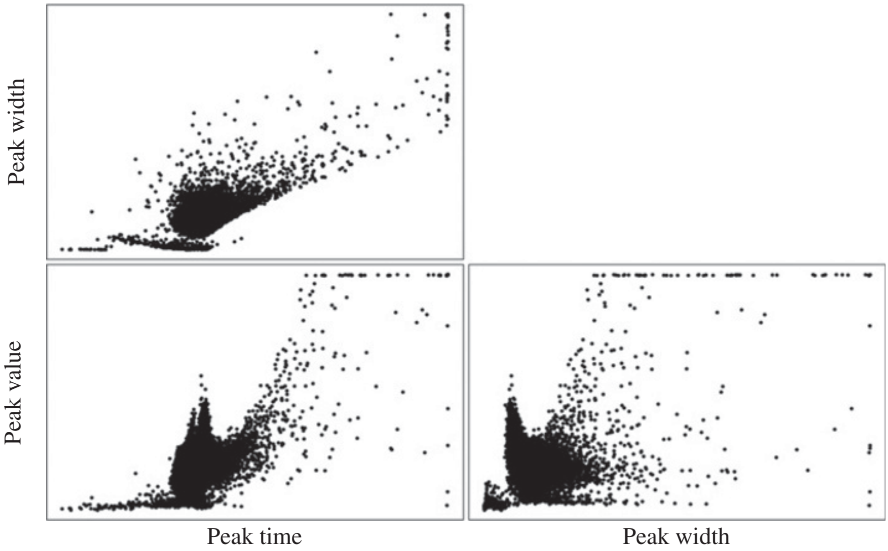


FIGURE 1.3 Pairwise scatter diagram of transformed Landsat data from 22,932 pixels over a 5 by 6 nautical mile region. The range on all the axes is (0, 255).

principle, it should be possible to locate the same point in each figure, assuming the data are free of ties. But it is not practical to do so for samples of any size. For quadrivariate data, Diaconis and Friedman (1983) proposed drawing lines between corresponding points in the scatterplots of (x_1,x_2) and (x_3,x_4) (see Problem 1.2). But a more powerful dynamic technique that takes full advantage of computer graphics has been developed by several research groups (McDonald, 1982; Becker and Cleveland, 1987; see the many references in Cleveland and McGill, 1988). The method is called *brushing* or *painting* a scatterplot matrix. Using a pointing device such as a mouse, a subset of the points in one scatter diagram is selected and the corresponding points are simultaneously highlighted in the other scatter diagrams. Conceptually, a subset of points in $\Re^d$ is tagged, for example, by painting the points red or making the points blink synchronously, and that characteristic is inherited by the linked points in all the "linked" graphs, including not only scatterplots but also histograms and regression plots as well. The *Iris* example in Figure 1.1 illustrates the flavor of brushing with three tags. Usually the color of points is changed rather than the symbol type. Brushing is an excellent tool for identifying outliers and following well-defined clusters. It is well-suited for conditioning on some variable, for example, $1 < x_3 < 3$.

These ideas are illustrated in Figure 1.4 for the PRIM4 dataset (Friedman and Tukey, 1974; the data summarize 500 high-energy particle physics scattering experiments) provided in the S language. Using the brushing tool in S-PLUS (1990), the left cluster in the 1–2 scatterplot was brushed, and then the left cluster in the 2–4 scatterplot was brushed with a different symbol. Try to imagine linking the clusters throughout the scatterplot matrix without any highlighting.

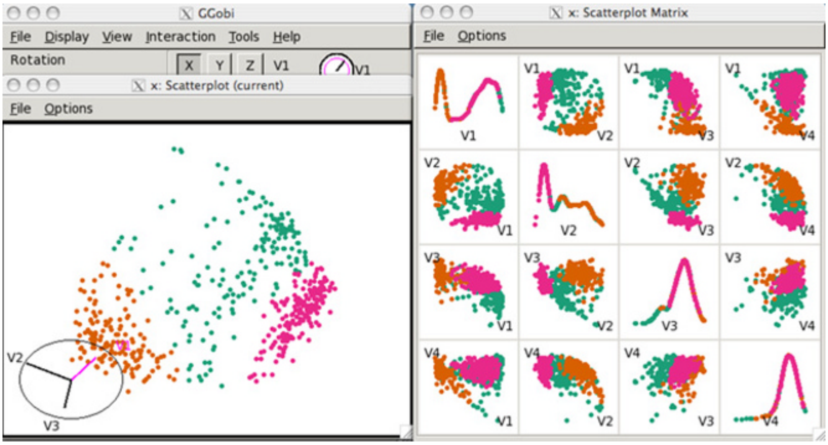


FIGURE 1.4 Pairwise scatterplots of the transformed PRIM4s data using the ggobi visualization system. Two clumps of points are highlighted by brushing.

There are limitations to the brushing technique. The number of pairwise scatterplots is $\binom{d}{2}$, so viewing more than 5 or 10 variables at once is impractical. Furthermore, the physical size of each scatter diagram is reduced as more variables are added, so that fewer distinct data points can be plotted. If there are more than a few variables, the eye cannot follow many of the dynamic changes in the pattern of points during brushing, except with the simplest of structure. It is, however, an open question as to the number of dimensions of structure that can be perceived by this method of linkage. Brushing remains an important and well-used tool that has proven successful in real data analysis.

If a 2-D array of bivariate scatter diagrams is useful, then why not construct a 3-D array of *trivariate* scatter diagrams? Navigating the collection of $\binom{d}{3}$ trivariate scatterplots is difficult even with modest values of d. But a single 3-D scatterplot can easily be rotated in real time with significant perceptual gain compared to three bivariate diagrams in the scatterplot matrix. Many statistical packages now provide this capability. The program MacSpin (Donoho et al., 1988) was the first widely used software of this type. The top middle panel in Figure 1.4 displays a particular orientation of a rotating 3-D scatterplot. The kinds of structure available in 3-D data are more complex (and hence more interesting) than in 2-D data. Furthermore, the overplotting problem is reduced as more data points can be resolved in a rotating 3-D scatterplot than in a static 2-D view (although this is resolution dependent—a 2-D view printed by a laser device can display significantly more points than is possible on a computer monitor). Density information is still relatively difficult to perceive, however, and the sample size definitely influences perception.

Beyond three dimensions, many novel ideas are being pursued (see Tukey and Tukey (1981)). Six-dimensional data could be viewed with two rotating 3-D scatter diagrams linked by brushing. Carr and Nicholson (1988) have actively pursued using stereography as an alternative and adjunct to rotation. Some workers report that stereo viewing of static data can be more precise than viewing dynamic rotation alone. Unfortunately, many individuals suffer from color blindness and various depth perception limitations, rendering some techniques useless. Nevertheless, it is clear that there is no limit to the possible combinations of ideas one might consider implementing. Such efforts can easily take many months to program without any fancy interface. This state of affairs would be discouraging but for the fact that a LISP-based system for easily prototyping such ideas is now available using object-oriented concepts (see Tierney (1990)). RStudio has made the *shiny* app available for this purpose as well: see http://shiny.rstudio.com. A collection of articles is devoted to the general topic of animation (Cleveland and McGill, 1988).

The idea of displaying 2- or 3-D arrays of 2- or 3-D scatter diagrams is perhaps too closely tied to the Euclidean coordinate system. It might be better to examine many 2- or 3-D projections of the data. An orderly way to do approximately just that is the "grand tour" discussed by Asimov (1985). Let P be a $d \times 2$ projection matrix, which takes the d-dimensional data down to a plane. The author proposed examining a sequence of scatterplots obtained by a smoothly changing sequence of projection matrices. The resulting kinematic display shows the n data points moving in a continuous (and sometimes seemingly random) fashion. It may be hoped that most interesting projections will be displayed at some point during the first several minutes of the grand tour, although for even 10 variables several hours may be required (Huber, 1985).

Special attention should be drawn to representing multivariate data in the bivariate scatter diagram with points replaced by *glyphs*, which are special symbols whose shapes are determined by the remaining data variables $(x_3, \ldots, x_d)$. Figure 1.5 displays the *Iris* data in such a form following Carr et al. (1986). The length and angle of the glyph are determined by the sepal length and width, respectively. Careful examination of the glyphs shows that there is no gap in 4-D between the *versicolor* and *virginica* species, as the angles and lengths of the glyphs are similar near the boundary.

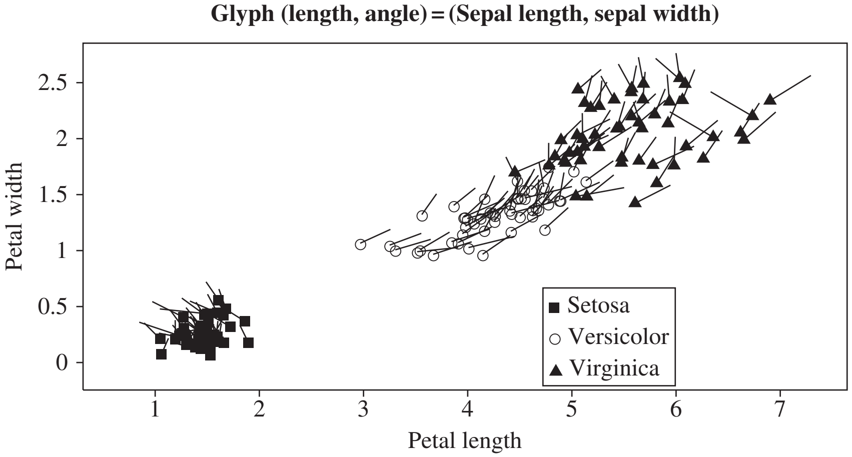


FIGURE 1.5 Glyph scatter diagram of the *Iris* data.