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

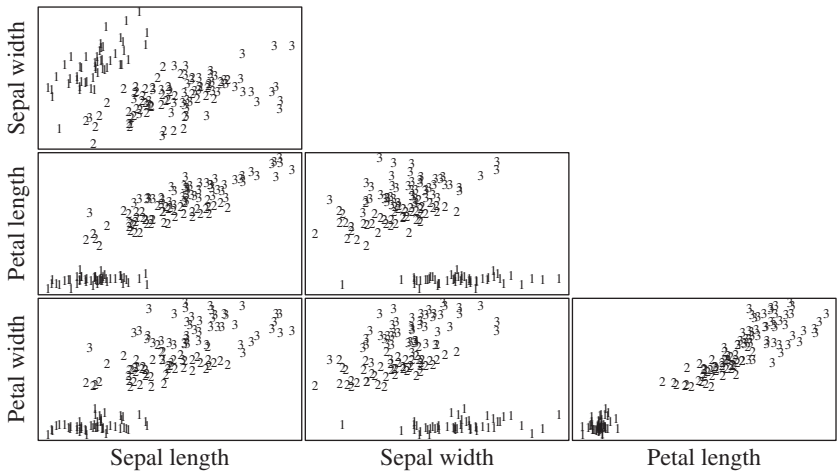

FIGURE 1.1 Pairwise scatter diagrams of the *Iris* data with the three species labeled. 1, setosa; 2, versicolor; 3, virginica.

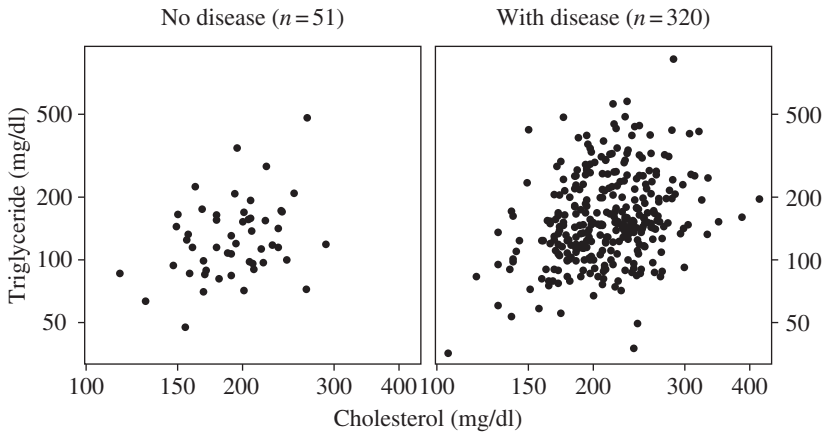

FIGURE 1.2 Scatter diagrams of blood lipid concentrations for 320 diseased and 51 nondiseased males.

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

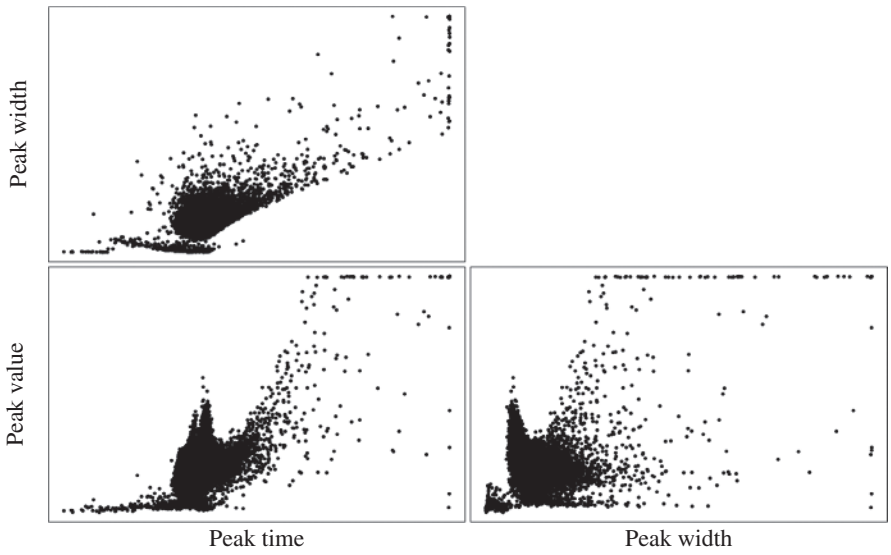

FIGURE 1.3 Pairwise scatter diagram of transformed Landsat data from 22,932 pixels over a 5 by 6 nautical mile region. The range on all the axes is (0, 255).

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.

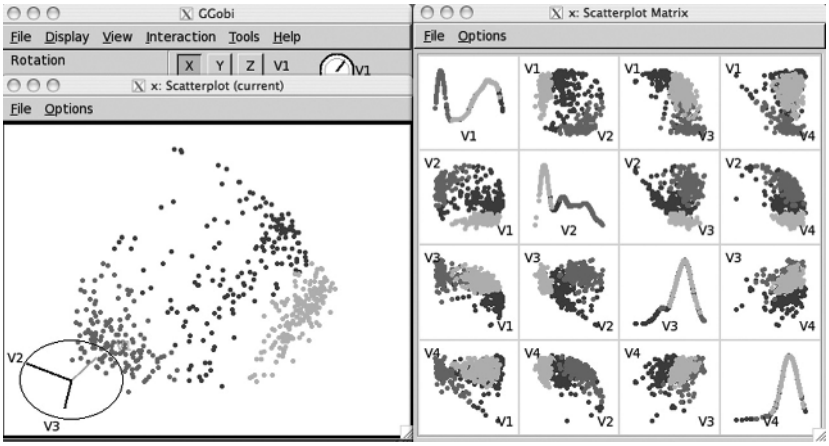

FIGURE 1.4 Pairwise scatterplots of the transformed PRIM4s data using the ggobi visualization system. Two clumps of points are highlighted by brushing.

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.

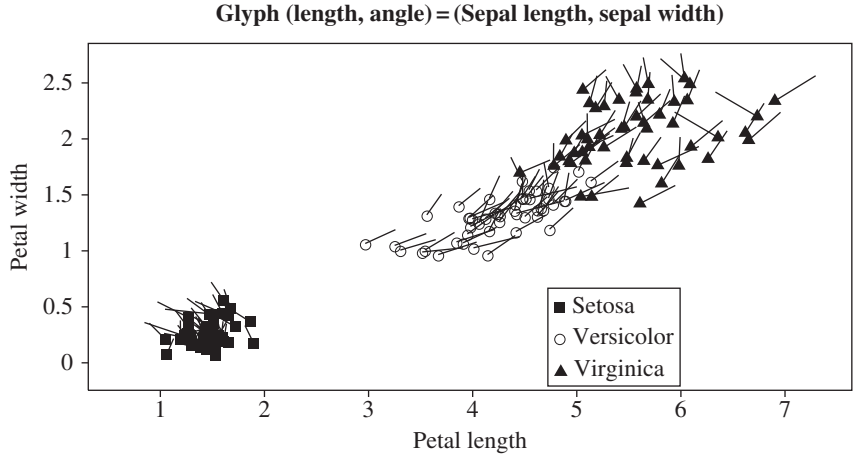

FIGURE 1.5 Glyph scatter diagram of the *Iris* data.

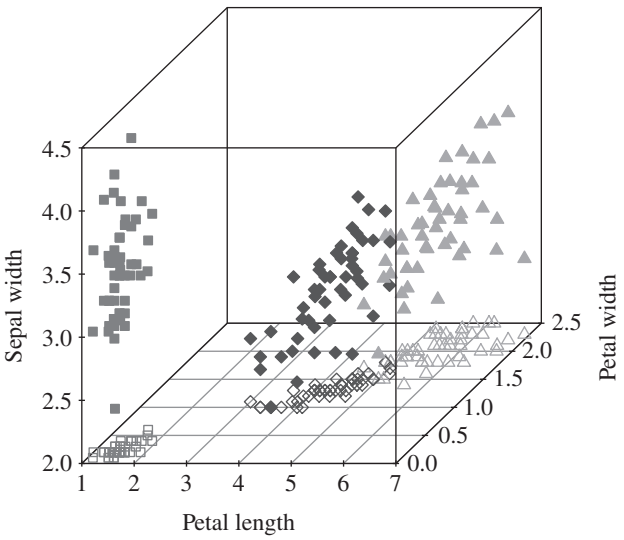

FIGURE 1.6 A three-dimensional scatter diagram of the Fisher–Anderson *Iris* data, omitting the sepal length variable. From left to right, the 50 points for each of the three varieties of *setosa*, *versicolor*, and *virginica* are distinguished by symbol type (square, diamond, triangle), respectively. The symbol is required to indicate the presence of three clusters rather than only two. The same basic picture results from any choice of three variables from the full set of four variables.

A second glyph representation shown in Figure 1.6 is a 3-D scatterplot omitting sepal length, one of the four variables. This figure clearly depicts the structure in these data. Plotting glyphs in 3-D scatter diagrams with stereography is a more powerful visual tool (Carr and Nicholson, 1988). The glyph technique does not treat variables "symmetrically" and all variable–glyph combinations could be considered. This complaint affects most multivariate procedures (with a few exceptions).

All of these techniques are an outgrowth of a powerful system devised to analyze data in up to nine dimensions called PRIM-9 (Fisherkeller et al., 1974; reprinted in Cleveland and McGill, 1988). The PRIM-9 system contained many of the capabilities of current systems. The letters are an acronym for "Picturing, Rotation, Isolation, and Masking." The latter two serve to identify and select subsets of the multivariate data. The "picturing" feature was implemented by pressing two buttons that cycled through all of the $\binom{9}{2}$ pairwise scatter diagrams in current coordinates. An IBM 360 mainframe was specially modified to drive the custom display system.

1.3.2 Chernoff Faces

Chernoff (1973) proposed a special glyph that associates variables to facial features, such as the size and shape of the eyes, nose, mouth, hair, ears, chin, and facial outline. Certainly, humans are able to discriminate among nearly identical faces very well. Chernoff has suggested that most other multivariate point methods "seem to be

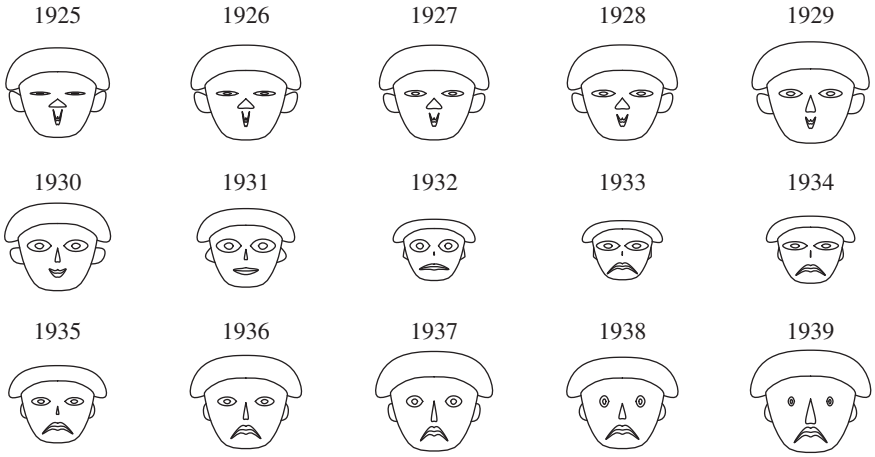

FIGURE 1.7 Chernoff faces of the economic dataset spanning 1925–1939.

less valuable in producing an emotional response" (Wang, 1978, p. 6).Whether an emotional response is desired is debatable. Chernoff faces for the time series dataset in Table B.1 are displayed in Figure 1.7. (The variable–feature associations are listed in the table.) By carefully studying an individual facial feature such as the smile over the sequence of all the faces, simple trends can be recognized. But it is the overall multivariate impression that makes Chernoff faces so powerful. Variables should be carefully assigned to features. For example, Chernoff faces of the colleges' data in Table B.2 might logically assign variables relating to the library to the eyes rather than to the mouth (see Problem 1.3). Such subjective judgments should not prejudice our use of this procedure.

One early application not in a statistics journal was constructed by Hiebert-Dodd (1982), who had examined the performance of several optimization algorithms on a suite of test problems. She reported that several referees felt this method of presentation was too frivolous. Comparing the endless tables in the paper as it appeared to the Chernoff faces displayed in the original technical report, one might easily conclude the referees were too cautious. On the other hand, when Rice University administrators were shown Chernoff faces of the colleges' dataset, they were quite open to its suggestions and enjoyed the exercise. The practical fact is that repetitious viewing of large tables of data is tedious and haphazard, and broad-brush displays such as faces can significantly improve data digestion. Several researchers have noted that Chernoff faces contain redundant information because of symmetry. Flury and Riedwyl (1981) have proposed using asymmetrical faces, as did Turner and Tidmore (1980), although Chernoff has stated he believes the additional gain does not justify such nonrealistic figures.

1.3.3 Andrews' Curves and Parallel Coordinate Curves

Three intriguing proposals display not the data points themselves but rather a unique curve determined by the data vector $\mathbf{x}$. Andrews (1972) proposed representing

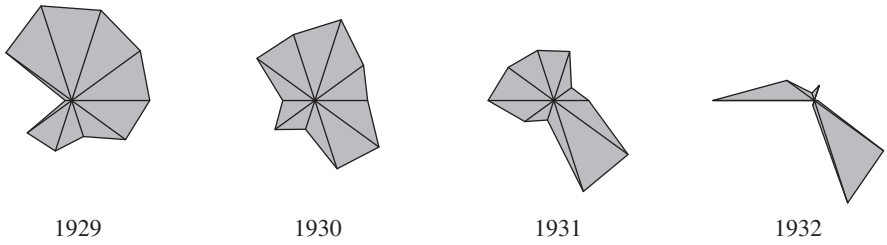

FIGURE 1.8 Star diagram for 4 years of the economic dataset shown in Figure 1.7.

high-dimensional data by replacing each point in $\Re^d$ with a curve $s(t)$ for $|t| < \pi$, where

$$s(t \mid x_1, \ldots, x_d) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots,$$

the so-called *Fourier series representation*. This mapping provides the first "complete" continuous view of high-dimensional points on the plane, because, in principle, the original multivariate data point can be recovered from this curve. Clearly, an Andrews' curve is dominated by the variables placed on the low-frequency terms, so care should be taken to put the most interesting variables early in the expansion (see Problem 1.4).

A simple graphical device that treats the d variables symmetrically is the star diagram, which is discussed by Fienberg (1979). The d axes are drawn as spokes on a wheel. The coordinate data values are plotted on those axes and connected as shown in Figure 1.8.

Another novel multivariate approach that treats variables in a symmetric fashion is the *parallel coordinates plot*, introduced by Inselberg (1985) in a mathematical setting and extended by Wegman (1990) to the analysis of stochastic data. Cartesian coordinates are abandoned in favor of d axes drawn parallel and equally spaced. Each multivariate point $\mathbf{x} \in \Re^d$ is plotted as a piecewise linear curve connecting the d points on the parallel axes. For reasons shown by Inselberg and Wegman, there are advantages to simply drawing piecewise linear line segments, rather than a smoother line such as a spline. The disadvantage of this choice is that points that have identical values in any coordinate dimension cannot be distinguished in parallel coordinates. However, with this choice a duality may be deduced between points and lines in Euclidean and parallel coordinates. In the left frame of Figure 1.9, six points that fall on a straight line with negative slope are plotted. The right frame shows those same points in parallel coordinates. Thus a scatter diagram of highly correlated normal points displays a nearly common point of intersection in parallel coordinates. However, if the correlation is positive, that point is not "between" the parallel axes (see Problem 1.6). The location of the point where the lines all intersect can be used to recover the equation of the line back in Euclidean coordinates (see Problem 1.8).

A variety of other properties with potential applications are explored by Inselberg and Wegman. One result is a graphical means of deciding if a point $\mathbf{x} \in \Re^d$ is on the

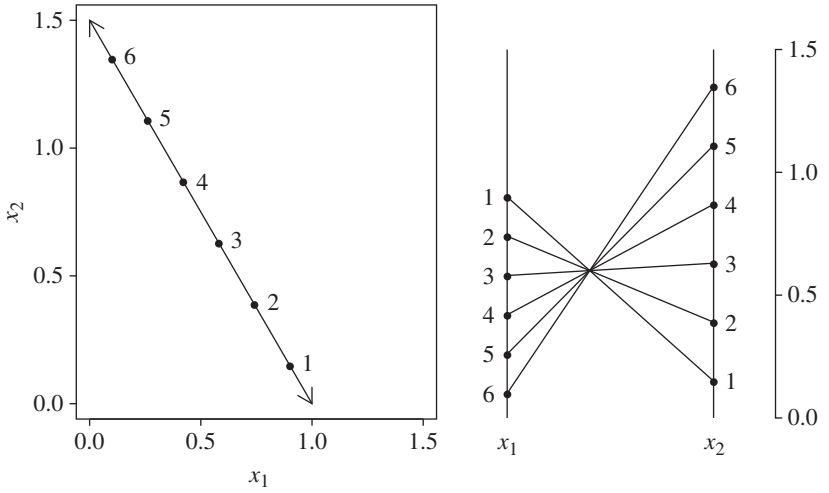

FIGURE 1.9 Example of duality of points and lines between Euclidean and parallel coordinates. The points are labeled 1 to 6 in both coordinate systems.

inside or the outside of a convex closed hypersurface. If all the points on the hypersurface are plotted in parallel coordinates, then a well-defined geometrical outline will appear on the plane. If a portion of the line segments defining the point **x** in parallel coordinates fall outside the outline, then **x** is not inside the hypersurface, and vice versa. One of the more fascinating extensions developed by Wegman is a grand tour of all variables displayed in parallel coordinates. The advantage of parallel coordinates is that all d of the rotating variables are visible simultaneously, whereas in the usual presentation, only two of the grand tour variables are visible in a bivariate scatterplot.

Figure 1.10 displays parallel coordinate plots of the *Iris* and earthquake data. The earthquake dataset represents the epicenters of 473 tremors beneath the Mount St. Helens volcano in the several months preceding its March 1982 eruption (Weaver et al., 1983). Clearly, the tremors are mostly small in magnitude, increasing in frequency over time, and clustered near the surface, although depth is clearly a bimodal variable. The longitude and latitude variables are least effective on this plot, because their natural spatial structure is lost.

1.3.4 Limitations

Tools such as Chernoff faces and scatter diagram glyphs tend to be most valuable with small datasets where individual points are "identifiable" or interesting. Such individualistic exploratory tools can easily generate "too much ink" (Tufte, 1983) and produce figures with black splotches, which convey little information. Parallel coordinates and Andrews' curves generate much ink. One obvious remedy is to plot

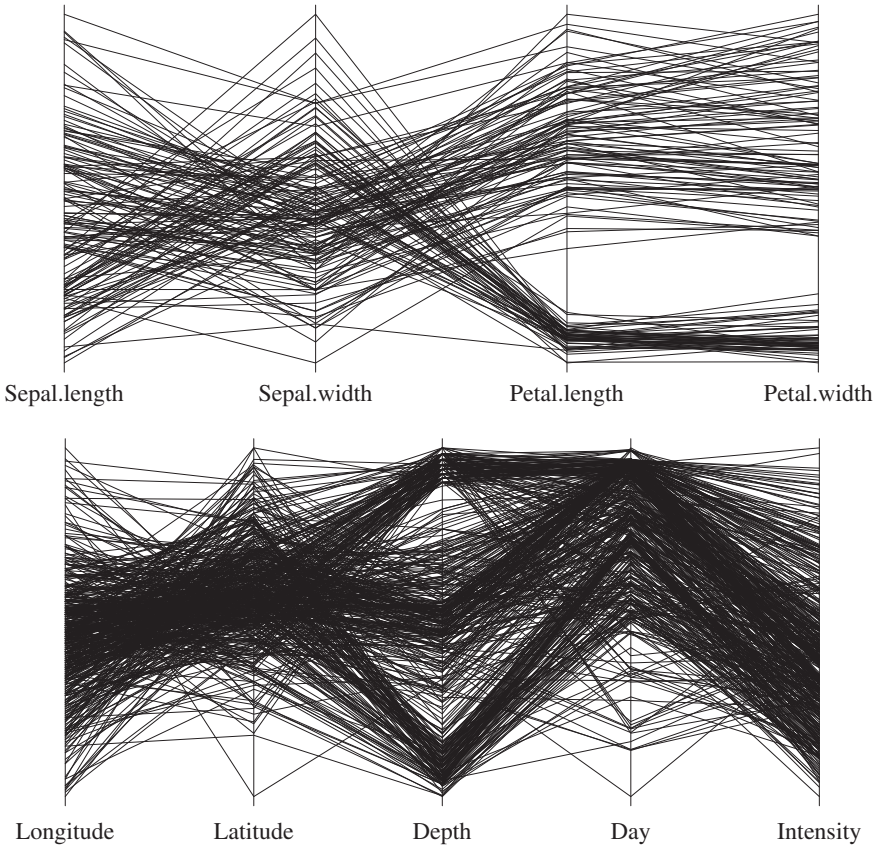

FIGURE 1.10 Parallel coordinate plot of the earthquake dataset.

only a subset of the data in a process known as "thinning." However, plotting random subsets no longer makes optimal use of all the data and does not result in precisely reproducible interpretations. Point-oriented methods typically have a range of sample sizes that is most appropriate: $n < 200$ for faces; $n < 2000$ for scatter diagrams.

Since none of these displays is truly d-dimensional, each has limitations. All pairwise scatterplots can detect distinct clusters and some two-dimensional structure (if perhaps in a rotated coordinate system). In the latter case, an interactive supplement such as brushing may be necessary to confirm the nature of the links among the scatterplots (not really providing any higher dimensional information). On the positive side, variables are treated symmetrically in the scatterplot matrix. But many different and highly dissimilar d-dimensional datasets can give rise to visually similar scatterplot matrix diagrams; hence the need for brushing. However, with increasing number of variables, individual scatterplots physically decrease in size and fill up with ink ever faster. Scatter diagrams provide a highly subjective view of data, with poor density perception and greatest emphasis on the tails of the data.

1.4 GRAPHICAL DISPLAY OF MULTIVARIATE FUNCTIONALS

1.4.1 Scatterplot Smoothing by Density Function

As graphical exploratory tools, each of the point-based procedures has significant value. However, each suffers from the problem of too much ink, as the number of objects (and hence the amount of ink) is linear in the sample size n. To mix metaphors, point-based graphs cannot provide a consistent picture of the data as $n \to \infty$. As Scott and Thompson (1983) wrote,

> the scatter diagram points to the bivariate density function.

In other words, the raw data points need to be smoothed if a consistent view is to be obtained.

A histogram is the simplest example of a *scatterplot smoother*. The amount of smoothness is controlled by the bin width. For univariate data, the histogram with bin width narrower than $\min |x_i - x_j|$ is precisely a univariate scatter diagram plotted with glyphs that are tall, thin rectangles. For bivariate data, the glyph is a beam with a square base. Increasing the bin width, the histogram represents a count per unit area, which is precisely the unit of a probability density. In Chapter 3, the histogram will be shown to provide a consistent estimate of the density function in any dimension.

Histograms can provide a wealth of information for large datasets, even well-known ones. For example, consider the 1979–1981 decennial life table published by the U.S. and Bureau of the Census (1987). Certain relevant summary statistics are well-known: life expectancy, infant mortality, and certain conditional life expectancies. But what additional information can be gleaned by examining the mortality histogram itself? In Figure 1.11, the histogram of age of death for individuals is depicted. Not surprisingly, the histogram is skewed with a short tail for older ages. Not as well-known perhaps is the observation that the most common age of death is 85! The absolute and relative magnitude of mortality in the first year of life is made strikingly clear.

Careful examination reveals two other general features of interest. The first feature is the small but prominent bump in the curve between the ages of 13 and 27 years. This "excess mortality" is due to an increase in a variety of risky activities, the most notable being obtaining a driver's license. In the right frame of Figure 1.11, comparison of the 1959–1961 (Gross and Clark, 1975) and 1979–1981 histograms shows an impressive reduction of death in all preadolescent years. Particularly striking is the 60% decline in mortality in the first year and the 3-year difference in the locations of the modes.

These facts are remarkable when placed in the context of the *mortality histogram* constructed by John Graunt from the Bills of Mortality during the plague years. Graunt (1662) estimated that 36% of individuals died before attaining their sixth birthday! Graunt was a contemporary of the better-known William Petty, to whom some credit for these ideas is variously ascribed, probably without cause. The circumstantial evidence that Graunt actually invented the histogram while looking at these mortality data seems quite strong, although there is reason to infer that Galileo had used

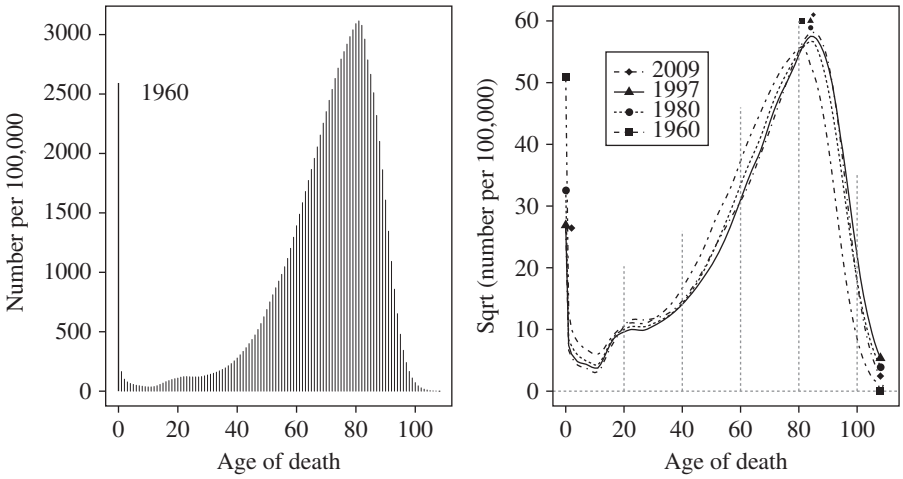

FIGURE 1.11 Histogram of the U.S. mortality data in 1960. Rootgrams (histograms plotted on a square-root scale) of the mortality data for 1960, 1980, and 1997.

histogram-like diagrams earlier. Hald (1990) recounts a portion of Galileo's *Dialogo*, published in 1632, in which Galileo summarized his observations on the star that appeared in 1572. According to Hald, Galileo noted the symmetry of the "observation errors" and the more frequent occurrence of small errors than large errors. Both points suggest Galileo had constructed a frequency diagram to draw those conclusions.

Many large datasets are in fact collected in binned or histogram form. For example, elementary particles in high-energy physics scattering experiments are manifested by small bumps in the frequency curve. Good and Gaskins (1980) considered such a large dataset $(n = 25{,}752)$ from the Lawrence Radiation Laboratory (LRL) (see Figure 1.12). The authors devised an ingenious algorithm for estimating the odds that a bump observed in the frequency curve was real. This topic is covered in Chapter 9.

Multivariate scatterplot smoothing of time series data is also easily accomplished with histograms. Consider a univariate time series and smooth both the raw data $\{x_t\}$ as well as the lagged data $\{x_t, x_{t+1}\}$. Any strong elliptical structure present in the smoothed lagged-data diagram provides a graphical version of the first-order autocorrelation coefficient. Consider the Old Faithful geyser dataset listed in Table B.6. These data are the durations in minutes of 107 eruptions of the Old Faithful geyser (Weisberg, 1985). As there was a gap in the recording of data between midnight and 6 A.M., there are only 99 pairs $\{x_t, x_{t+1}\}$ available. The univariate histogram in Figure 1.13 reveals a simple bimodal structure—short and long eruption durations. The most notable feature in the bivariate (smoothed) histogram is the missing fourth bump corresponding to the short-short duration sequence. Clearly, graphs of $\hat{f}(x_{t+1}|x_t)$ would be useful for improved prediction compared to a regression estimate.

For more than two dimensions, only slices are available for viewing with histogram surfaces. Consider the Landsat data again. Divide the (jittered) data into four pieces using quartiles of x_1, which is the time of peak greenness. Examining a series of

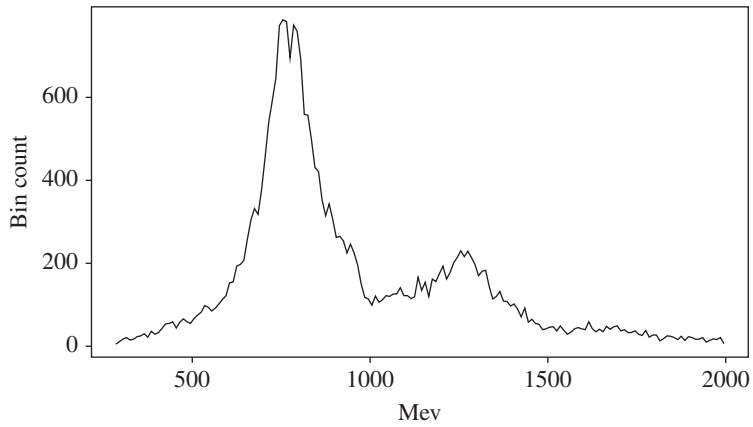

FIGURE 1.12 Histogram of LRL dataset.

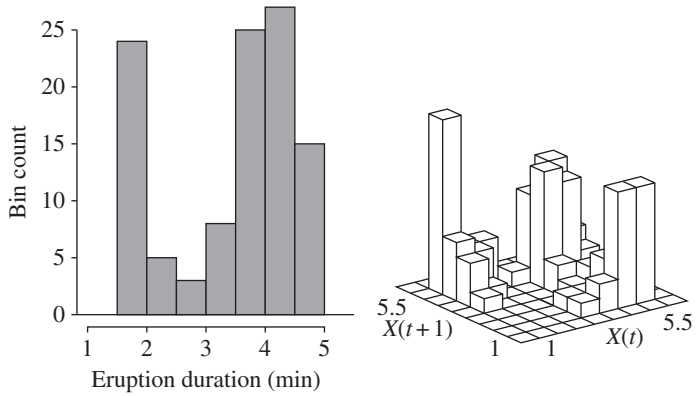

FIGURE 1.13 Histogram of $\{x_t\}$ for the Old Faithful geyser dataset, and a bivariate histogram of the lagged data (x_t, x_{t+1}).

bivariate pictures of (x_2, x_3) for each quartile slice provides a crude approximation of the four-dimensional surface $\hat{f}(x_1, x_2, x_3)$ (see Figure 1.14). The histograms are all constructed on the subinterval $[-5, 100] \times [-5, 100]$. Compare this representation of the Landsat data to that in Figure 1.3. From Figure 1.3, it is clear that most of the outliers are in the last quartile of x_1. How well can the relative density levels be determined from the scatter diagrams? Visualization of a smoothed histogram of these data will be considered in Section 1.4.3.

1.4.2 Scatterplot Smoothing by Regression Function

The term *scatterplot smoother* is most often applied to regression data. For bivariate data, either a nonparametric regression line can be superimposed upon the data, or the points themselves can be moved toward the regression line. Tukey (1977) presents

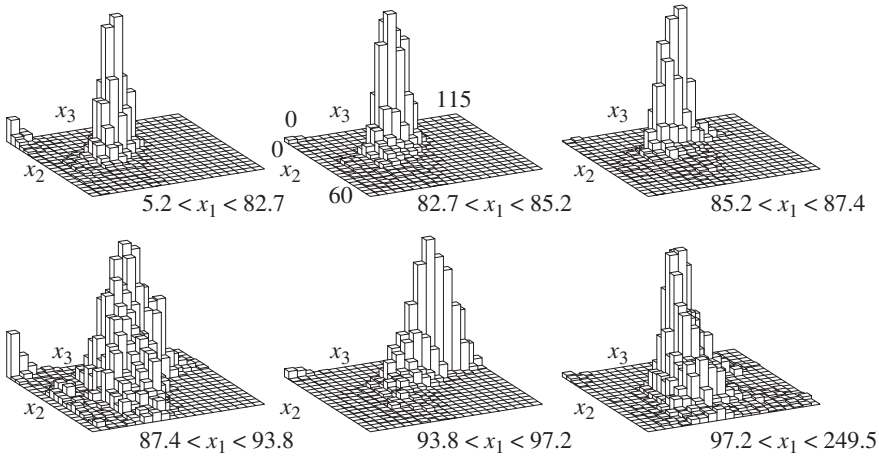

FIGURE 1.14 Bivariate histogram slices of the trivariate Landsat data. Slicing was performed at the quartiles of variable x_1.

the "3R" smoother as an example of the latter. Suppose that the n data points, $\{x_t\}$, are measured on a fixed time scale. The 3R smoothing algorithm replaces each point $\{x_t\}$ with the median of the three points $\{x_{t-1}, x_t, x_{t+1}\}$ recursively until no changes occur. This algorithm is a powerful filter that removes isolated outliers effectively. The 3R smoother may be applied to unequally spaced data or repeated data. Tukey also proposes applying a Hanning filter, by which $\tilde{x}_t \leftarrow 0.25 \times (x_{t-1} + 2x_t + x_{t+1})$. This filter may be applied several times as necessary. In Figure 1.15, the Tukey smoother (S function *smooth*) is applied to the gas flow dataset given in the Table B.5. Observe how the single potential outlier at $x = 187$ is totally ignored. The least-squares fit is shown for reference.

The simplest nonparametric regression estimator is the *regressogram*. The x-axis is binned and the sample averages of the responses are computed and plotted over the intervals. The regressogram for the gas flow dataset is also shown in Figure 1.15. The Hanning filter and regressogram are special cases of nonparametric kernel regression, which is discussed in Chapter 8.

The gas flow dataset is part of a larger collection taken at seven different pressures. A stick-pin plot of the complete dataset is shown in Figure 1.16 (the 74.6 psia data are second from the right). Clearly, the accuracy is affected by the flow rate, while the effect of psia seems small. These data will be revisited in Chapter 8.

1.4.3 Visualization of Multivariate Functions

Visualization of functions of more than two variables has not been common in statistics. The Landsat example in Figure 1.14 hints at the potential that visualization of 4-D surfaces would bring to the data analyst. In this section, effective visualization of surfaces in more than three dimensions is introduced.

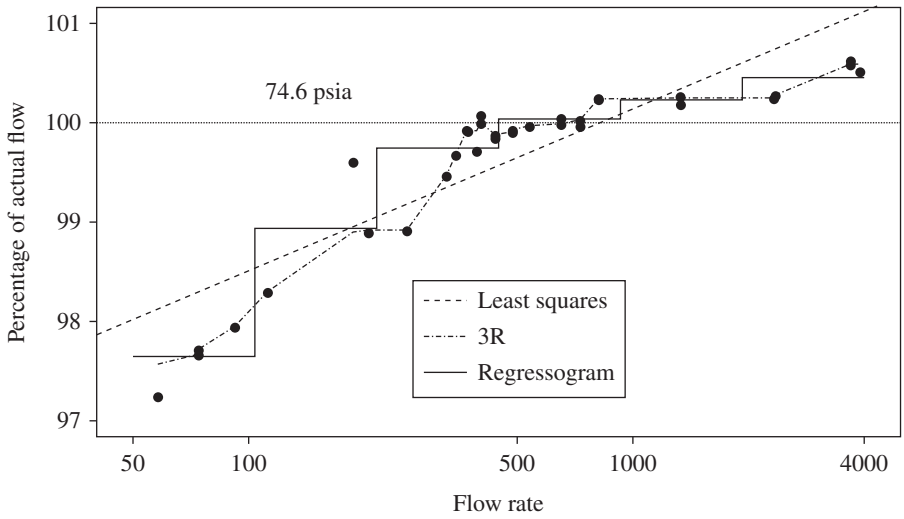

FIGURE 1.15 Accuracy of a natural gas meter as a function of the flow rate through the valve at 74.6 psia. The raw data ($n = 33$) are shown by the filled points. The three smooths (least squares, Tukey's 3R, and Tukey's regressogram) are superimposed.

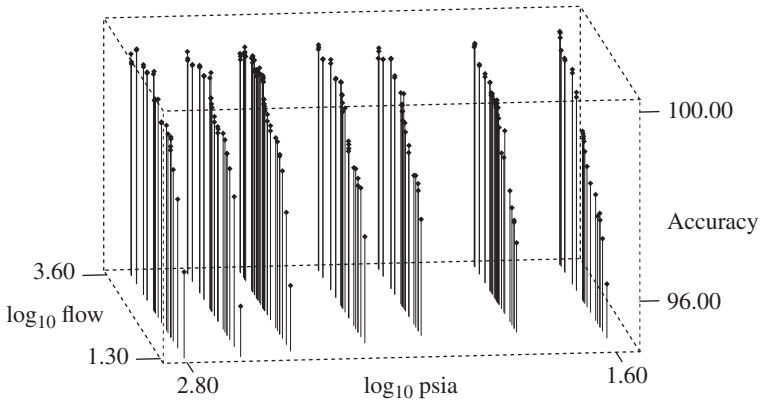

FIGURE 1.16 Complete 3-D view of the gas flow dataset.

Displaying a three-dimensional perspective plot of the surface $f(x, y)$ of a bivariate function requires one more dimension than the corresponding bivariate contour representation (see Figure 1.17). There are trade-offs. The contour representation lacks the exact detail and visual impact available in a perspective plot; however, perspective plots usually have portions obscured by peaks and present less precise height information. One way of expressing the difference is to say that a contour plot displays, loosely speaking, about 2.6–2.9 dimensions of the entire 3-D surface (more, as more contour lines are drawn). Some authors claim that one or the other representation is superior, but it seems clear that both can be useful for complicated surfaces.

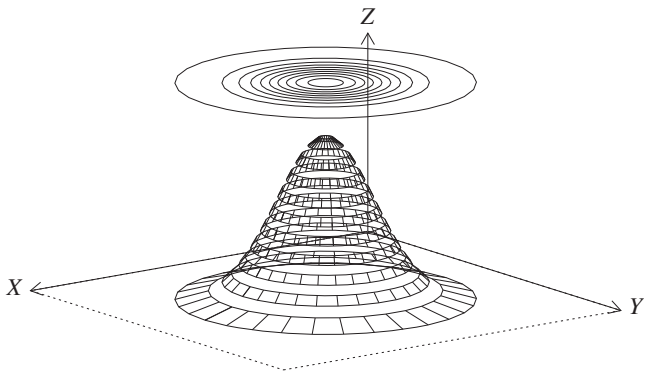

FIGURE 1.17 Perspective plot of bivariate normal density with a "floating" representation of the corresponding contours.

The visualization advantage afforded by a contour representation is that it lives in the *same dimension* as the data, whereas a perspective plot requires an additional dimension. Hence with trivariate data, the third dimension can be used to present a 3-D contour. In the case of a density function, the corresponding 3-D contour plot comprises one or more α-*level contour surfaces*, which are defined for $\mathbf{x} \in \Re^d$ by

$$\boxed{\alpha\text{-Contour}: \qquad S_\alpha = \{\mathbf{x} : f(\mathbf{x}) = \alpha f_{\max}\}, \qquad 0 \le \alpha \le 1\,,}$$

where $f_{\max}$ is the maximum or modal value of the density function.

For normal data, the general contour surfaces are hyper-ellipses defined by the easily verified equation (see Problem 1.14):

$$(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) = -2\log\alpha. \tag{1.1}$$

A trivariate contour plot of $f(x_1, x_2, x_3)$ would generally contain several "nested" surfaces, $\{S_{0.1}, S_{0.3}, S_{0.5}, S_{0.7}, S_{0.9}\}$, for example. For the independent standard normal density, the contours would be nested hyperspheres centered on the mode. In Figure 1.18, three contours of the trivariate standard normal density are shown in stereo. Many if not most readers, will have difficulty crossing their eyes to obtain the stereo effect. But even without the stereo effect, the three spherical contours are well-represented.

How effective is this in practice? Consider a smoothed histogram $\hat{f}(x, y, z)$ of 1000 trivariate normal points with $\Sigma = I_3$. Figure 1.19 shows surfaces of nine equally spaced bivariate slices of the trivariate estimate. Each slice is approximately bivariate normal but without rescaling. Of course, the surfaces are not precisely bivariate normal, due to the finite size of the sample.

A natural question to pose is: Why not plot the corresponding sequence of *conditional densities*, $\hat{f}(x, y|z = z_0)$, rather than the *slices*, $\hat{f}(x, y, z_0)$? If this were done, all the surfaces in Figure 1.19 would be nearly identical. (Theoretically, the condition

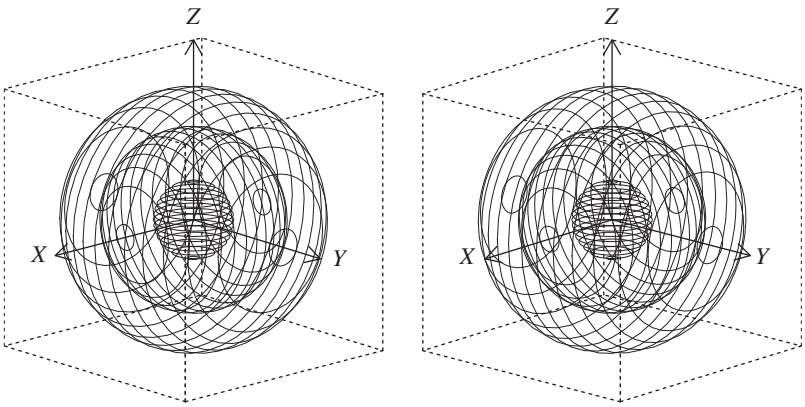

FIGURE 1.18 Stereo representation of three α-contours of a trivariate normal density. Gently crossing your eyes should allow the two frames to fuse in the middle.

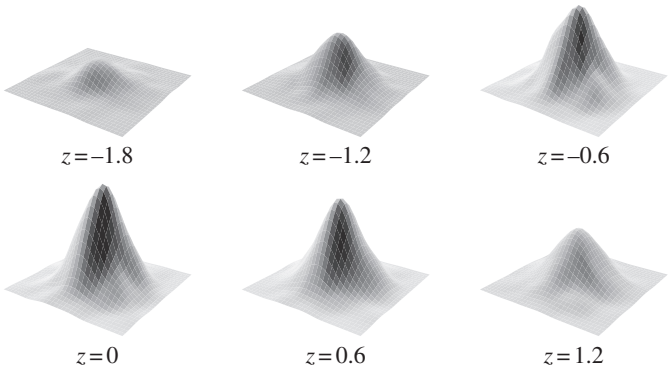

FIGURE 1.19 Sequence of bivariate slices of a trivariate smoothed histogram.

densities are all exactly $N(\mathbf{0}_2, I_2)$.) If the goal is to understand the 4-D density surface, then the sequence of conditional densities overemphasizes the (visual) importance of the tails and obscures information about the location of the "center" of the data. Furthermore, as nonparametric estimates in the tail will be relatively noisy, the estimates will be especially rough upon normalization (see Figure 1.20). For these reasons, it seems best to look at slices and to reserve normalization for looking at conditional densities that are particularly interesting.

Several trivariate contour surfaces of the same estimated density are displayed in Figure 1.21. Clearly, the trivariate contours give an improved "big picture"—just as a rotating trivariate scatter diagram improves on three static bivariate scatter diagrams. The complete density estimate is a 4-D surface, and the trivariate contour view in the final frame of Figure 1.21 may present only 3.5 dimensions, while the series of bivariate slices may yield a bit more, perhaps 3.75 dimensions, but without the visual impact. Examine the 3-D contour view for the Landsat data in the first frame of Figure 7.8 in comparison to Figures 1.3 and 1.14. The structure is quite complex.

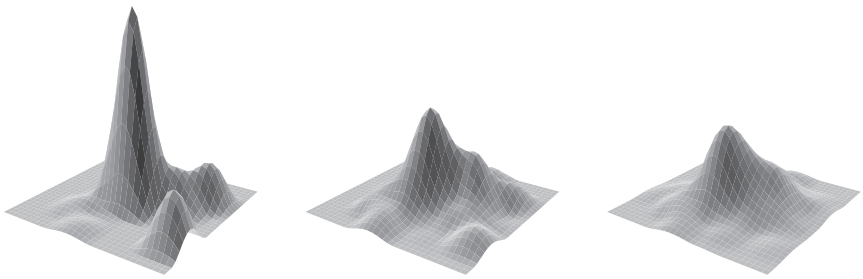

FIGURE 1.20 Normalized slices in the left tail of the smoothed histogram.

The presentation of clusters is stunning and shows multiple modes and multiple clusters. This detailed structure is not apparent in the scatterplot in Figure 1.3.

Depending on the nature of the variables, slicing can be attempted with four-, five-, or six-dimensional data. Of special importance is the 5-D surface generated by 4-D data, for example, space–time variables such as the Mount St. Helens data in Figure 1.10. These higher dimensional estimates can be animated in a fashion similar to Figure 1.19 (see Scott and Wilks (1990)).

In the 4-D case, the α-level contours of interest are based on the slices:

$$S_{\alpha,t} = \{(x,y,z) : f(x,y,z,t) = \alpha f_{\max}\},$$

where $f_{\max}$ is the global maximum over the 5-D surface. For a fixed choice of α, as the slice value t changes continuously, the contour shells will expand or contract smoothly, finally vanishing for extreme values of t. For example, a single theoretical contour of the $N(0, I_4)$ density would vanish outside a symmetric interval around the origin, but within that interval, the contour shell would be a sphere centered on the origin with greatest diameter when $t = 0$. With several α-shells displayed simultaneously, the contours would be nested spheres of different radii, appearing at different values of t, but of greatest diameter when $t = 0$.

One particularly interesting slice of the smoothed 5-D histogram estimate of the entire *Iris* dataset is shown in Figure 1.22. The $\alpha = 4\%$ contour surface reveals two well-separated clusters. However, the $\alpha = 10\%$ contour surface is trimodal, revealing the true structure in this dataset even with only 150 points. the *virginica* and *versicolor* data may not be separated in the point cloud but apparently can be separated in the density cloud.

The 3-D contour slices in Figure 1.22 were assembled from a 2-D contouring algorithm, then projected into the plane. The sequence of 2-D contour slices is shown in Figure 1.23. Study these two diagrams and think about the possibilities for exploring the entire five-dimensional surface.

To emphasize the potential value of additional variables, we conclude this vignette, we examine the *Iris* data excluding the sepal width variable. Figure 1.24 displays a 3-D scatterplot, as well as contours of the smoothed histogram at levels $\alpha = 0.17$ and $\alpha = 0.44$. A litle study supports the speculation that the data might contain a hybrid

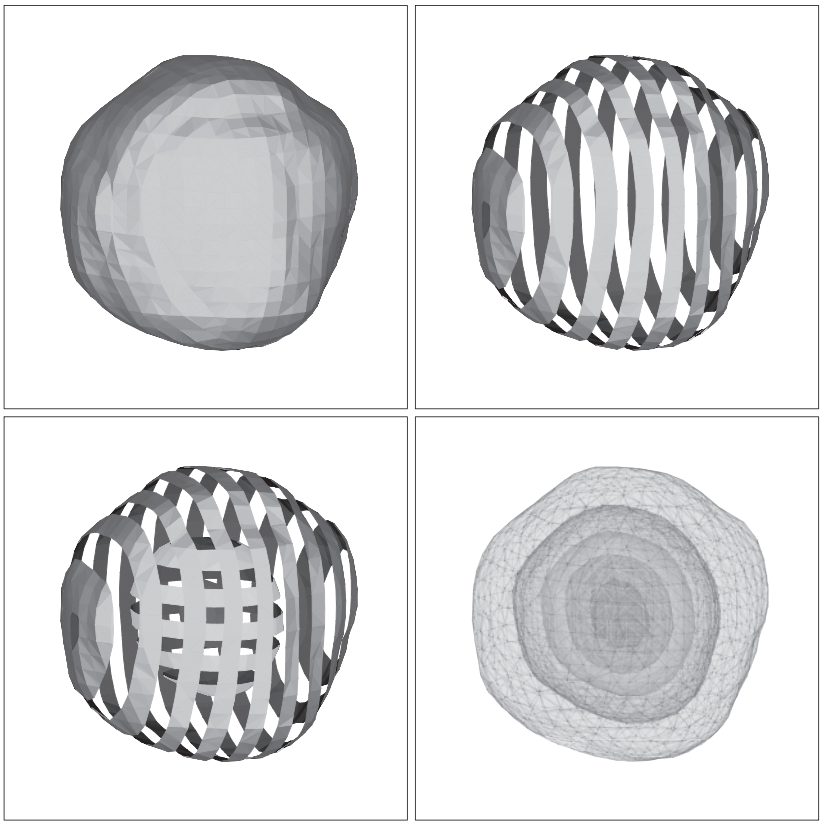

FIGURE 1.21 Trivariate normal examples.

species of the *versicolor* and *virginica* species. With such a small sample, that may be an embellishment.

With more than four variables, the most appropriate sequence of slicing is not clear. With five variables, bivariate contours of (x_4, x_5) may be drawn; then a sequence of trivariate slices may be examined tracing along one of these bivariate contours. With more than five or six variables, deciding where to slice at all is a difficult problem because the number of possibilities grows exponentially. That is why projection-based methods are so important (see Chapter 7).

1.4.3.1 Visualizing Multivariate Regression Functions The same graphical representation can be applied to regression surfaces. However, the interpretation can be more difficult. For example, if the regression surface is monotone, the α-level contours of the surface will not be "closed" and will appear to "float" in space. If the regression surface is a simple linear function such as $ax + by + cz$, then a set of trivariate α-contours will simply be a set of parallel planes. Practical questions arise that do not appear for density surfaces. In particular, what is the natural extent of the regression surface; that is, for what region in the design space should the surface be

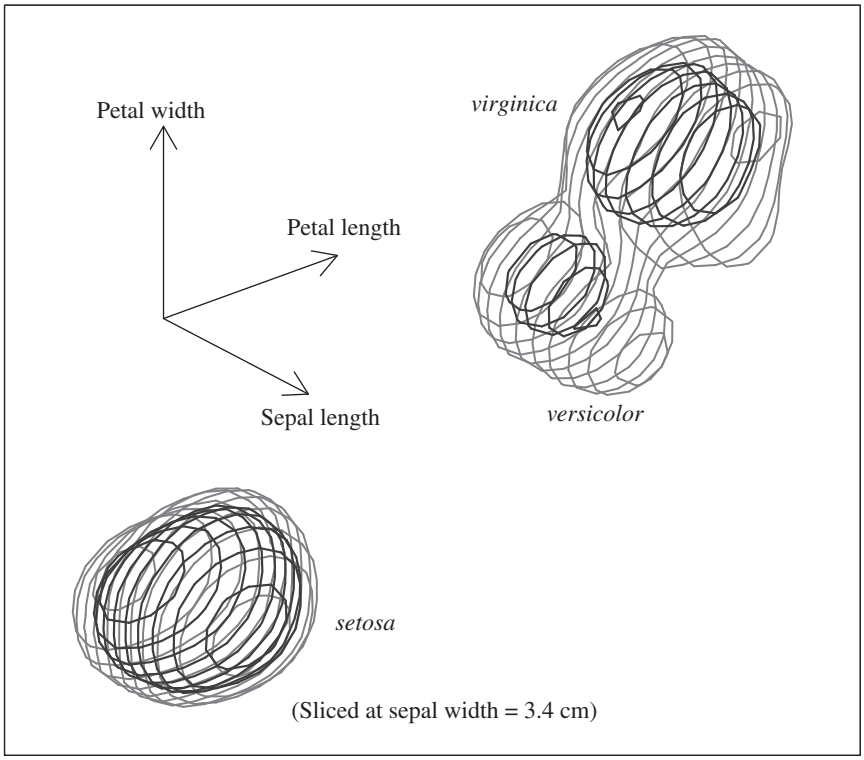

FIGURE 1.22 Two α-level contour surfaces from a slice of a five-dimensional averaged shifted histogram estimate, based on all 150 *Iris* data points. The displayed variables x, y, and z are sepal length, petal length and width, respectively, with the sepal width variable sliced at $t = 3.4$ cm. The (outer) darker $\alpha = 4\%$ contour reveals only two clusters, while the (inner) lighter $\alpha = 10\%$ contour reveals the three clusters.

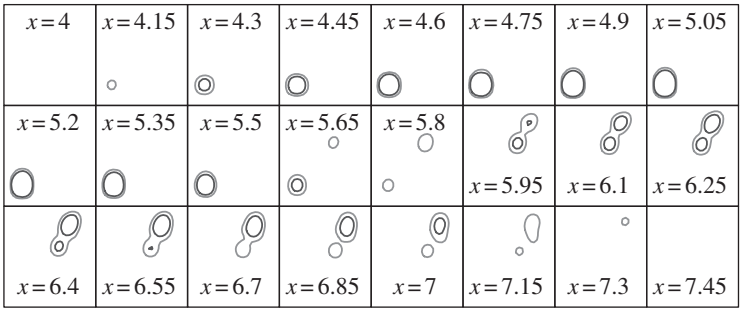

FIGURE 1.23 A detailed breakdown of the 3-D contours shown in Figure 1.22 taken from the ASH estimate $\hat{f}(x, y, z, t = 3.4)$ as the sepal length, x, ranges from 4.00 to 7.45 cm.

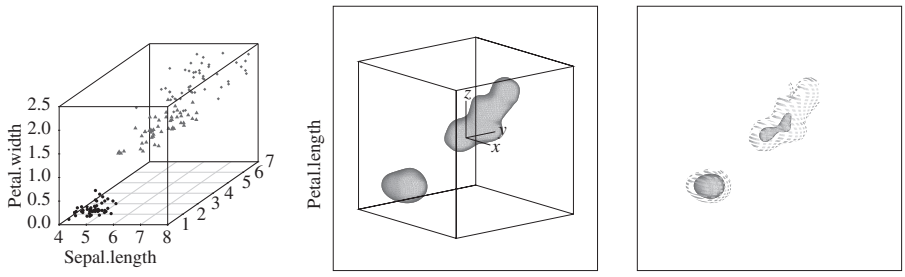

FIGURE 1.24 Analysis of three of the four *Iris* variables, omitting sepal width entirely, which should be compared to the slice shown in Figure 1.22. The middle contour ($\alpha = 0.17$) is superimposed upon the contour ($\alpha = 0.44$) in the right frame to help locate the shells.

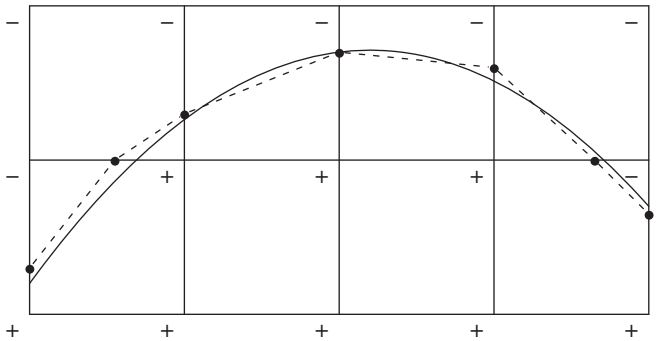

FIGURE 1.25 A portion of a bivariate contour at the $\alpha = 0$ level of a smooth function measured on a regular grid and using linear interpolation (dotted lines).

plotted? Perhaps one answer is to limit the plot to regions where there is sufficient data, that is, where the density of design points is above a certain threshold.

1.4.4 Overview of Contouring and Surface Display

Suppose that a general bivariate function $f(x,y)$ (taking on positive and negative values) is sampled on a regular grid, and the $\alpha = 0$ contour S_0 is desired; that is, $S_0 = \{(x,y) : f(x,y) = 0\}$. Label the values of the grid as $+$, 0, or $-$ depending on whether $f > 0$, $f = 0$, or $f < 0$, respectively. Then the desired contour is shown in Figure 1.25. The piecewise linear approximation and the true contour do not match along the bin boundaries since the interpolation is not exact.

However, bivariate contouring is not as simple a task as one might imagine. Usually, the function is sampled on a rectangular mesh, with no gradient information or possibility for further refinement of the mesh. If too coarse a mesh is chosen, then small local bumps or dips may be missed, or two distinct contours at the same level may be inadvertently joined. For speed and simplicity, one wants to avoid having to do any global analysis before drawing contours. A local contouring algorithm avoids multiple passes over the data. In any case, global analysis is based on certain

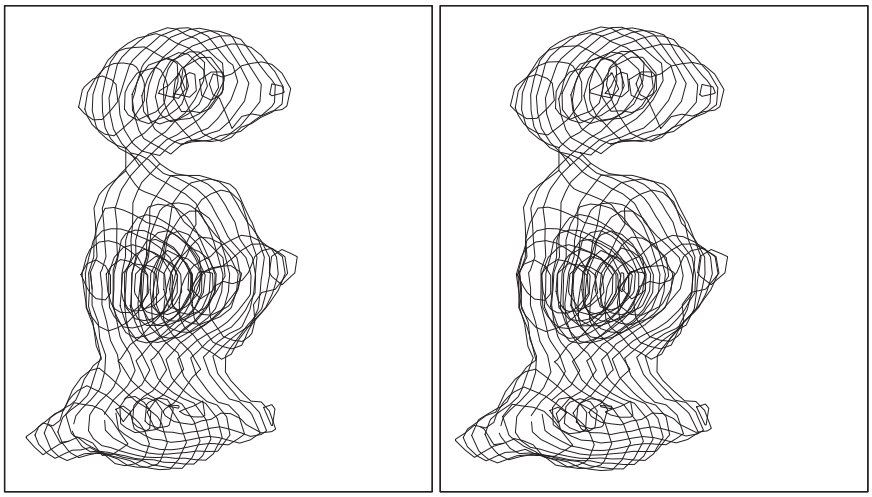

FIGURE 1.26 Simple stereo representation of four 3-D nested shells of the earthquake data.

smoothness assumptions and may fail. The difficulties and details of contouring are described more fully in Section A.1.

There are several varieties of 3-D contouring algorithms. It is assumed that the function has been sampled on a lattice, which can be taken to be cubical without loss of generality. One simple trick is to display a set of 2-D contour slices that result from intersecting the 3-D contour shell with a set of parallel planes along the lattice of the data, as was done in Figures 1.18 and 1.22. In this representation, a single spherical shell becomes a set of circular contours (Figure 1.26). This approach has the advantage of providing a shell representation that is "transparent" so that multiple α-level contour levels may be visualized. Different colors can be used for different contour levels (see Scott (1983, 1984, 1991a), Scott and Thompson (1983), Härdle and Scott (1988), and Scott and Hall (1989)).

More visually pleasing surfaces can be drawn using the *marching cubes* algorithm (Lorensen and Cline, 1987). The overall contour surface is represented by a large number of connected triangular planar sections, which are computed for each cubical bin and then displayed. Depending on the pattern of signs on the eight vertices of each cube in the data lattice, up to six triangular patches are drawn within each cube (see Figure 1.27). In general, there are 2^8 cases (each corner of the cube being either above or below the contour level). Taking into consideration certain symmetries reduces this number. By scanning through all the cubes in the data lattice, a collection of triangles is found that defines the contour shell. Each triangle has an inner and outer surface, depending on the gradient of the density function. The inner and outer surfaces may be distinguished by color shading. A convenient choice is various shades of red for surfaces pointing toward regions of higher (hotter) density, and shades of blue toward regions of lower (cooler) density; see the cover jacket of this book for an example. Each contour is a patchwork of several thousand triangles. Smoother surfaces may be

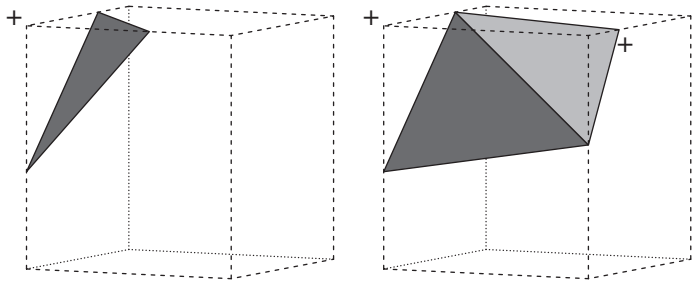

FIGURE 1.27 Examples of marching cube contouring algorithm. The corners with values above the contour level are labeled with a + symbol.

obtained by using higher-order splines, but the underlying bin structure information would be lost.

In summary, visualizing trivariate functions directly is a powerful adjunct to data analysis. The gain of an additional dimension of visible structure without resort to slices greatly improves the ability of a data analyst to perceive structure. The same visualization applies to slices of density function with more than three variables. A demonstration tape that displays 4-D animation of $S_{\alpha,t}$ contours as α and t vary is available (Scott and Wilks, 1990).

1.5 GEOMETRY OF HIGHER DIMENSIONS

The geometry of higher dimensions provides a few surprises. In this section, a few standard figures are considered. This material is available in scattered references (see Kendall (1961), for example).

1.5.1 Polar Coordinates in *d* Dimensions

In d dimensions, a point $\mathbf{x}$ can be expressed in spherical polar coordinates by a radius r, a base angle θ_{d-1} ranging over $(0, 2\pi)$, and $d-2$ angles $\theta_1, \ldots, \theta_{d-2}$ each ranging over $(-\pi/2, \pi/2)$ (see Figure 1.28). Let $s_k = \sin\theta_k$ and $c_k = \cos\theta_k$. Then the transformation back to Euclidean coordinates is given by

$$
\begin{aligned}
x_1 &= r\,c_1\,c_2 \cdots c_{d-3}\,c_{d-2}\,c_{d-1} \\
x_2 &= r\,c_1\,c_2 \cdots c_{d-3}\,c_{d-2}\,s_{d-1} \\
x_3 &= r\,c_1\,c_2 \cdots c_{d-3}\,s_{d-2} \\
&\vdots \\
x_j &= r\,c_1 \cdots c_{d-j} s_{d-j+1} \\
&\vdots \\
x_d &= r\,s_1\,.
\end{aligned}
$$

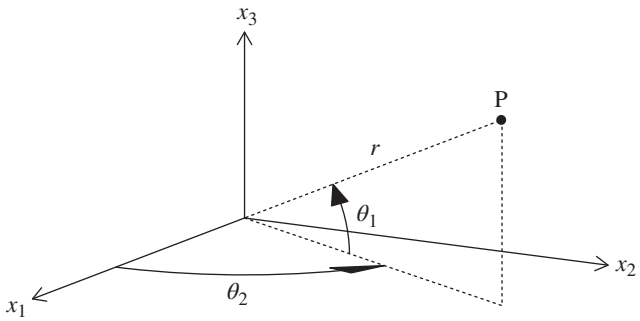

FIGURE 1.28 Polar coordinates (r, θ_1, θ_2) of a point P in $\Re^3$.

After some work (see Problem 1.11), the Jacobian of this transformation may be shown to be

$$J = r^{d-1} c_1^{d-2} c_2^{d-3} \cdots c_{d-2}\,. \tag{1.2}$$

1.5.2 Content of Hypersphere

The volume of the d-dimensional hypersphere $\{\mathbf{x} : \sum_{i=1}^{d} x_i^2 \le a^2\}$ is given by

$$\begin{aligned} V_d(a) &= \int\limits_{\sum_{i=1}^{d} x_i^2 \le a^2} 1\ \mathbf{dx} \\ &= \int_0^a dr \int_{-\pi/2}^{\pi/2} d\theta_1 \int_{-\pi/2}^{\pi/2} d\theta_2 \cdots \int_0^{2\pi} d\theta_{d-1} r^{d-1} c_1^{d-2} c_2^{d-3} \cdots c_{d-2}\,. \end{aligned}$$

This can be simplified using the identity

$$\int_{-\pi/2}^{\pi/2} \cos^k\theta\, d\theta = 2\int_0^{\pi/2} \cos^k\theta\, d\theta = 2\int_0^{\pi/2} \cos^k\theta \frac{d(\cos^2\theta)}{-2\cos\theta\sin\theta}\,,$$

which, using the change of variables $u = \cos^2\theta$,

$$= \int_0^1 u^{k/2} \frac{du}{u^{1/2}(1-u)^{1/2}} = \mathrm{B}\left(\tfrac{1}{2}, \tfrac{k+1}{2}\right) = \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k+2}{2}\right)}\,.$$

As $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$,

$$V_d(a) = 2\pi \frac{a^d}{d} \cdot \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \cdot \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d-2}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)} \cdots \frac{\Gamma\left(\frac{1}{2}\right)\Gamma(1)}{\Gamma\left(\frac{3}{2}\right)}$$
$$= \frac{a^d \pi^{d/2}}{\frac{d}{2}\Gamma\left(\frac{d}{2}\right)} = \frac{a^d \pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}. \tag{1.3}$$

1.5.3 Some Interesting Consequences

1.5.3.1 Sphere Inscribed in Hypercube Consider the hypercube $[-a,a]^d$ and an inscribed hypersphere with radius $r = a$. Then using (1.3), the fraction of the volume of the cube contained in the hypersphere is given by

$$f_d = \frac{\text{Volume sphere}}{\text{Volume cube}} = \frac{a^d \pi^{d/2} / \Gamma\left(\frac{d}{2}+1\right)}{(2a)^d} = \frac{\pi^{d/2}}{2^d \Gamma\left(\frac{d}{2}+1\right)}.$$

For lower dimensions, the fraction f_d is as shown in Table 1.1. It is clear that the center of the cube becomes less important. As the dimension increases, the volume of the hypercube concentrates in its corners. This distortion of space (at least to our three-dimensional way of thinking) has many potential consequences for data analysis.

1.5.3.2 Hypervolume of a Thin Shell Wegman (1990) demonstrates the distortion of space in another setting. Consider two spheres centered on the origin, one with radius r and the other with slightly smaller radius $r - \epsilon$. Consider the fraction of the volume of the larger sphere in between the spheres. By Equation (1.3),

$$\frac{V_d(r) - V_d(r-\epsilon)}{V_d(r)} = \frac{r^d - (r-\epsilon)^d}{r^d} = 1 - \left(1 - \frac{\epsilon}{r}\right)^d \xrightarrow[d\to\infty]{} 1.$$

Hence, virtually all of the content of a hypersphere is concentrated close to its surface, which is only a $(d-1)$-dimensional manifold. Thus for data distributed uniformly over both the hypersphere and the hypercube, most of the data fall near the boundary and edges of the volume. Most statistical techniques exhibit peculiar behavior if the data fall in a lower dimensional subspace. This example illustrates one important aspect of the *curse of dimensionality*, which is discussed in Chapter 7.

TABLE 1.1 Fraction of the Volume of a Hypercube Lying in the Inscribed Hypersphere

Dimension (d)	1	2	3	4	5	6	7
Fraction volume (f_d)	1	0.785	0.524	0.308	0.164	0.081	0.037

1.5.3.3 Tail Probabilities of Multivariate Normal The preceding examples make it clear that if we are trying to view uniform data over the hypercube in $\Re^{10}$, most (spherical) neighborhoods will be empty!

Let us examine what happens if the data follow the standard d-dimensional normal distribution:

$$f_d(\mathbf{x}) = (2\pi)^{-d/2} e^{-\mathbf{x}^T\mathbf{x}/2}.$$

Clearly, the origin (mode) is the most likely point and the equiprobable contours are spheres. Consider the spherical contour, $S_{0.01}(\mathbf{x})$, where the density value is only 1% of the value at the mode. Now

$$\frac{f(\mathbf{x})}{f(\mathbf{0})} = e^{-\mathbf{x}^T\mathbf{x}/2} \quad \text{and} \quad -2\log\frac{f(\mathbf{x})}{f(\mathbf{0})} = \sum_{i=1}^{d} x_i^2 \sim \chi^2(d);$$

therefore, the probability that a point is *within* the 1% spherical contour may be computed as

$$\Pr\left(\frac{f(\mathbf{x})}{f(\mathbf{0})} \geq \frac{1}{100}\right) = \Pr\left(\chi^2(d) \leq -2\log\frac{1}{100}\right). \tag{1.4}$$

Equation (1.4) gives the probability a random point will not fall in the "tails" or, in other words, will fall in the medium- to high-density region. In Table 1.2, these probabilities are tabulated for several dimensions. Around five or six dimensions, the probability mass of a multivariate normal begins a rapid migration into the extreme tails. In fact, more than half of the probability mass is in a very low-density region for 10-dimensional data. Silverman (1986) has dramatized this in 10 dimensions by noting that $\text{Prob}(\|\mathbf{x}\| \geq 1.6) = 0.99$. In very high dimensions, virtually the entire sample will be in the tails in a sense consistent with low-dimensional intuition. Table 1.2 is also applicable to normal data with a general full-rank covariance matrix, except that the contour is a hyper-ellipsoid.

1.5.3.4 Diagonals in Hyperspace Pairwise scatter diagrams essentially project the multivariate data onto all the two-dimensional faces. Consider the hypercube $[-1,1]^d$ and let any of the diagonal vectors from the center to a corner be denoted by $\mathbf{v}$. Then $\mathbf{v}$ is any of the 2^d vectors of the form $(\pm 1, \pm 1, \ldots, \pm 1)^T$. The angle between

TABLE 1.2 Probability Mass *Not* in the "Tail" of a Multivariate Normal Density

d	1	2	3	4	5	6	7	8	9	10	15	20
$1000p$	998	990	973	944	899	834	762	675	582	488	134	20

a diagonal vector $\mathbf{v}$ and a Euclidean coordinate axis $\mathbf{e}_j = (0,\ldots,0,1,0,\ldots,0)^T$ is given by

$$\cos\theta_d = \frac{\langle \mathbf{v},\, \mathbf{e}_j\rangle}{\sqrt{\langle \mathbf{v},\, \mathbf{v}\rangle\langle \mathbf{e}_j, \mathbf{e}_j\rangle}} = \frac{\pm 1}{\sqrt{d}} \xrightarrow[d\to\infty]{} 0,$$

where $\langle \mathbf{u},\, \mathbf{v}\rangle = \mathbf{u}^T\mathbf{v}$, so that $\theta_d \to \pi/2$ as $d \to \infty$. Thus the diagonals are nearly orthogonal to all coordinate axes for large d. Hence, any data cluster lying near a diagonal in hyperspace will be mapped into the origin in every paired scatterplot, while a cluster along a coordinate axis should be visible in some plot.

Thus the choice of coordinate system in high dimensions is critical in data analysis and intuition is highly dependent on a good choice. Real data structures may be missed due to overstriking. The general conclusion is that one- to two-dimensional intuition is valuable but not infallible when continuing on to higher dimensions.

1.5.3.5 Data Aggregate Around Shell Returning to independent multivariate normal data, the mode is at the origin $\mathbf{x} = \mathbf{0}_d$. How far is a random point $\mathbf{X}$ from the origin for moderate to large dimensions d? Let $Z \sim N(0,1)$; then

$$\begin{aligned}\sqrt{\mathbf{X}^t\mathbf{X}} &= \sqrt{\sum_{j=1}^{d}\mathbf{X}_j^2} = \sqrt{\chi^2(d)} \approx \sqrt{d + Z\sqrt{2d}} = \sqrt{d}\sqrt{1 + Z\sqrt{2/d}} \\ &\approx \sqrt{d}\left(1 + \frac{1}{2}Z\sqrt{2/d}\right) = \sqrt{d} + \frac{1}{\sqrt{2}}Z \sim N\left(\sqrt{d}, \frac{1}{2}\right).\end{aligned}$$

Thus while the highest density region is near the origin, virtually all (99.7%) of the data lie within a distance $\pm 3/\sqrt{2} = \pm 2.12$ of the hypersphere with radius $\sqrt{d}$. This is the data version of the volume result in Section 1.5.3.2.

1.5.3.6 Nearest Neighbor Distances The derivation in Section 1.5.3.5 also addresses the question of the distribution of the closest pair of points in a random sample. As a conservative estimate, imagine one data point at the mode $\mathbf{x} = \mathbf{0}_d$ and compute the distribution of the distance from the origin to the closest of n sample points. Let D_i denote the distance the sample $\mathbf{X}_i$ is from the origin, and let D denote the minimum of $\{D_i\}$. Then

$$\begin{aligned}\Pr(D \le c) &= 1 - \Pr(D > c) = 1 - \Pr(D_1 > c, D_2 > c, \ldots, D_n > c) \\ &= 1 - \Pr(D_1 > c)^n = 1 - \Pr(D_1^2 > c^2)^n = 1 - \Pr(\chi_d^2 > c^2)^n \\ &= 1 - \left(1 - \Pr(\chi_d^2 \le c^2)\right)^n.\end{aligned}$$

Thus applying Leibniz's rule,

$$f_D(c) = \frac{d}{dc}\Pr(D \le c) = n\left(1 - \Pr(\chi_d^2 \le c^2)\right)^{n-1} \times 2c f_{\chi_d^2}(c^2). \tag{1.5}$$

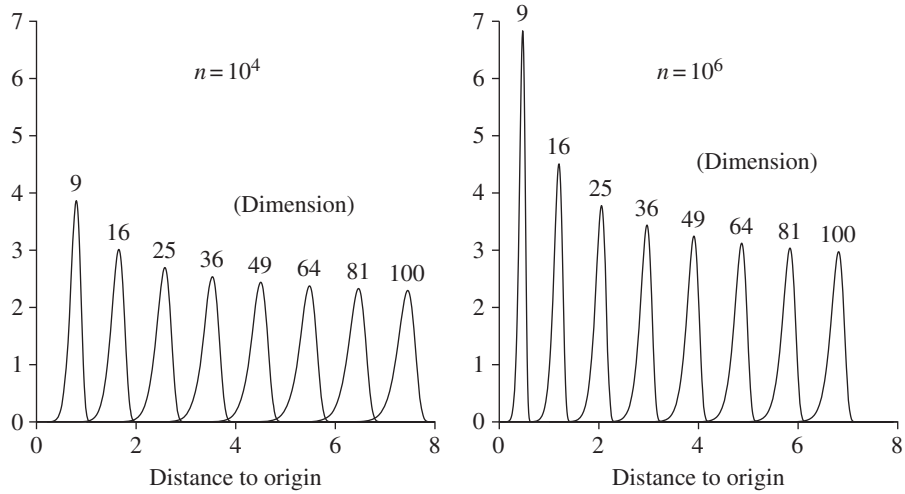

FIGURE 1.29 Densities of distance of closest point to the origin for sample sizes $n = 10^4$ and 10^6, for various dimensions $9 \leq d \leq 100$.

In Figure 1.29, this density is displayed for a sample of size $n = 10^4$ and several values of d. When $d > 25$, the closest pair of points is never closer than two units. Thus data in high dimensions are very sparse. A histogram bin will almost always be empty, or contain just one point. Increasing the sample size to 10^6 does not change the distribution very much. The sparseness of data in high dimensions is known as the *curse of dimensionality* (Bellman, 1961). This phenomenon will influence our thinking when analyzing data in more than five or six dimensions.

As a side note, there is a benefit to the curse of dimensionality in the field of biometrics. The uniqueness of fingerprints and other physical measurements (iris scans, for example) in a very large population is of much interest. What this analysis suggests is that if a feature space can be transformed into a reasonable number of independent measurements, and individuals may be viewed as independent samples from $N(\mathbf{0}_d, I_d)$, then unique identification is feasible with sufficiently accurate measurements of the features (see Kent and Millett (2002).)

PROBLEMS

1.1 A class of challenging data problems have a "hole" in them.

(a) Devise a simple way of creating radially symmetric trivariate data with a "hole" in it; that is, a region where the probability data points lie goes smoothly to zero at the center. *Hint*: Invent a rejection rule based on the distance a trivariate normal point is from the origin.

(b) Study a pairwise scatter diagram of 5000 trivariate data with either a "large" or a "small" hole in the middle. When does the hole become

difficult to discern? Use "o" and "." as plotting symbols. Plot a histogram of $(x_1^2 + x_2^2 + x_3^2)^{1/2}$ and see if the hole is apparent.

1.2 Try the Diaconis–Friedman idea of linked bivariate scatter diagrams using the *Iris* data. Draw the scatterplots side-by-side and try connecting all points or random subsets. Evaluate your findings.

1.3 Use Chernoff faces on the college data in Table B.2. Try to assign variables to facial features in a memorable way. Compare your subjective choices of variables with those of others. You will notice that if one variable is near the extreme value of the data, it may distort that facial feature to such a degree that it is impossible to recognize the levels of other variables controlling different aspects of that feature. How should that influence your choice of variables for the mouth and eyes?

1.4 Display Andrews, curves for the economic dataset in Table B.1 for several permutations of the variables. How do these curves reflect the onset of the Depression after 1929?

1.5 Research problem: Generalize Andrews' representation so that the representation of a multidimensional point is a trajectory in the three-dimensional rectangle $[-\pi, \pi]^2 \times [0, 1]$.

1.6 Plot in parallel coordinates random samples of bivariate normal data with correlations ranging from -1 to 1. When the correlation $\rho = +1$, where does the point of intersection fall? Can you guess how trivariate correlated normal data will appear in parallel coordinates? Try it.

1.7 Investigate the appearance in parallel coordinates of data with clusters. For example, generate bivariate data with clusters centered at (0, 0) and (3, 3). Try centers at (0, 0) and (3, 0). Try centers of three clusters at (0, 0), (1, 0), and (2, 0), where the data in each cluster have $\rho = -0.9$. The last example shows the duality between clusters and holes.

1.8 Prove that points falling on a straight line in Euclidean coordinates intersect in a point in parallel coordinates. What is the one exception? Superimposing Euclidean coordinates upon the parallel axes as shown in the right frame of Figure 1.9, find the (Euclidean) coordinates of the intersection point.

1.9 Investigate the literature for other ideas of data representation, including the star diagram, linear profile, weathervane, polygon star, and Kleiner–Hartigan faces.

1.10 What are the possible types of intersection of two planes (2-D) in four-space? *Hint*: Consider the two planes determined by pairs of coordinate axes (see Wegman (1990)).

1.11 Show that the Jacobian equals what is claimed in Equation (1.2). *Hint*: See Anderson (2003, p. 285). Interestingly, the signs of the determinant do not alternate as the dimension d increases, but go in pairs.

1.12 Verify Equation (1.3) for the well-known cases of a circle and a sphere.

1.13 Think of another way to represent high-dimensional data. Try using some other set of orthogonal functions for Andrews' curves (step functions, Legendre polynomials, or others). How sensitive is your method to permutations of the coordinate axes?

1.14 Show that the α-level contours of a multi-normal density are given by Equation (1.1). Use some of the techniques in Appendix A to display some contours when $d = 3$ with correlated and uncorrelated random variables.

1.15 (Problem 1.10 continued) What are the possible types of intersections of a k_1-dimensional subspace and a k_2-dimensional subspace in d dimensions? Think about the intersection of other types of hypersurfaces.

1.16 What fraction of a d-dimensional hypersphere lies in the inscribed d-dimensional hypercube? Find numerical values for dimensions up to 10.

1.17 Examine parallel coordinate plots of commonly observed bivariate and trivariate structure, including correlation and clustering. Summarize your findings.

1.18 Verify Equation (1.5) by simulation. What does Figure 1.29 look like for the world population of $n = 7.2 \times 10^9$? How many features would be required for a reliable biometrics system for the entire globe?