3 *Visualizing and Exploring Data*

3.1 Introduction

This chapter explores visual methods for finding structures in data. Visual methods have a special place in data exploration because of the power of the human eye/brain to detect structures—the product of aeons of evolution. Visual methods are used to display data in ways that capitalize upon the particular strengths of human pattern processing abilities. This approach lies at quite the opposite end of the spectrum from methods for formal model building and for testing to see whether observed data could have arisen from a hypothesized data generating structure. Visual methods are important in data mining because they are ideal for sifting through data to find unexpected relationships. On the other hand, they do have their limitations, particularly, as we illustrate below, with very large data sets.

Exploratory data analysis can be described as *data-driven hypothesis generation*. We examine the data, in search of structures that may indicate deeper relationships between cases or variables. This process stands in contrast to *hypothesis testing* (we use the phrase here in an informal and general sense; more formal methods are described in chapter 4) which begins with a proposed model or hypothesis and undertakes statistical manipulations to determine the likelihood that the data arose from such a model. The phrase *data based* in the above description indicates that it is the patterns in the data that give rise to the hypotheses—in contrast to situations in which hypotheses are generated from theoretical arguments about underlying mechanisms. This distinction has implications for the legitimacy of subsequent testing of the hypotheses. It is closely related to the issues of overfitting discussed in chapter 7 (and again in 10 and 11). A simple example will illustrate the problem.

If we take 10 random samples of size 20 from the same population, and measure the values of a single variable, the random samples will have different means (just by virtue of random variability). We could compare the means using formal tests. Suppose, however, we took only the two samples giving rise to the smallest and largest means, ignoring the others. A test of the difference between these means might well show significance. If we took 100 samples, instead of 10, then we would be even more likely to find a significant difference between the largest and the smallest means. By ignoring the fact that these are the largest and smallest in a set of 100, we are biasing the analysis toward detecting a difference—even though the samples were generated from the same population.

In general, when searching for patterns, we cannot test whether a discovered pattern is a real property of the underlying distribution (as opposed to a chance property of the sample) without taking into account the size of the search—the number of possible patterns we have examined. The informal nature of exploratory data analysis makes this very difficult—it is often impossible to say how many patterns have been examined. For this reason researchers often use a separate data set, obtained from the same source as the first, to conduct formal testing for the existence of any pattern. (Alternatively, they may use some kind of sophisticated method such as cross-validation and sample re-use, as described in chapter 7.)

This chapter examines informal graphical data exploration methods, which have been widely used in data analysis down through the ages. Early books on statistics contain many such methods. They were often more practical than lengthy, number crunching alternatives in the days before computers. However, something of a revolution has occurred in recent years, and now such methods are even more widely used. As with the bulk of the methods decribed in this book, the revolution has been driven by the computer: computers enable us to view data in many different ways, both quickly and easily, and have led to the development of extremely powerful data visualization tools.

We begin the discussion in section 3.2 with a description of simple summary statistics for data. Section 3.3 discusses visualization methods for exploring distributions of values of single variables. Such tools, at least for small data sets, have been around for centuries, but even here progress in computer technology has led to the development of novel approaches. Moreover, even when using univariate displays, we often want simultaneous univariate displays of many variables, so we need concise displays that readily convey the main features of distributions.

Section 3.4 moves on to methods for displaying the relationships between pairs of variables. Perhaps the most basic form is the scatterplot. Due to the sizes of the data sets often encountered in data mining applications, scatterplots are not always enlightening—the diagram may be swamped by the data. Of course, this qualification can also apply to other graphical displays.

Moving beyond variable pairs, section 3.5 describes some of the tools used to examine relationships between multiple variables. No method is perfect, of course: unless a very rare relationship holds in the data, the relationship between multiple variables cannot be completely displayed in two dimensions.

Principal components analysis is illustrated in section 3.6. This method can be regarded as a special (indeed, the most basic) form of multidimensional scaling analysis. These are methods that seek to represent the important structure of the data in a reduced number of dimensions. Section 3.7 discusses additional multidimensional scaling methods.

There are numerous books on data visualization (see section 3.8) and we could not hope to examine all of the possibilities thoroughly in a single chapter. There are also several software packages motivated by an awareness of the importance of data visualization that have very powerful and flexible graphics facilities.

3.2 Summarizing Data: Some Simple Examples

We mentioned in earlier chapters that the *mean* is a simple summary of the average of a collection of values. Suppose that $x(1), \ldots, x(n)$ comprise a set of n data values. The *sample mean* is defined as

$$\hat{\mu} = \sum_i x(i)/n. \tag{3.1}$$

(Note that we use μ to refer to the true mean of the population, and $\hat{\mu}$ to refer to a sample-based *estimate* of this mean). The sample mean has the property that it is the value that is "central" in the sense that it minimizes the sum of squared differences between it and the data values. Thus, if there are n data values, the mean is the value such that the sum of n copies of it equals the sum of the data values.

The mean is a measure of *location*. Another important measure of location is the *median*, which is the value that has an equal number of data points above and below it. (Easy if n is an odd number. When there is an even number it is usually defined as halfway between the two middle values.)

The most common value of the data is the *mode*. Sometimes distributions have more than one mode (for example, there may be 10 objects which take the value 3 on some variable, and another 10 which take the value 7, with all other values taken less often than 10 times) and are therefore called *multimodal*.

Other measures of location focus on different parts of the distribution of data values. The first *quartile* is the value that is greater than a quarter of the data points. The third quartile is greater than three quarters. (We leave it to you to discover why we have not mentioned the second quartile.) Likewise, *deciles* and *percentiles* are sometimes used.

Various measures of *dispersion* or *variability* are also common. These include the *standard deviation* and its square, the *variance*. The variance is defined as the average of the squared differences between the mean and the individual data values:

$$\hat{\sigma}^2 = \sum_i (x(i) - \mu)^2/n. \tag{3.2}$$

Note that since the mean minimizes the sum of these squared differences, there is a close link between the mean and the variance. If μ is unknown, as is often the case in practice, we can replace μ above with $\hat{\mu}$, our data based estimate. When μ is replaced with $\hat{\mu}$, to get an unbiased estimate (as discussed in chapter 4), the variance is estimated as

$$\sum_i (x(i) - \hat{\mu})^2/(n-1). \tag{3.3}$$

The standard deviation is the square root of the variance:

$$\hat{\sigma} = \sqrt{\sum_i (x(i) - \mu)^2/n}. \tag{3.4}$$

The *interquartile range*, common in some applications, is the difference between the third and first quartile. The *range* is the difference between the largest and smallest data point.

Skewness measures whether or not a distribution has a single long tail and is commonly defined as

$$\frac{\sum (x(i) - \hat{\mu})^3}{\left(\sum (x(i) - \hat{\mu})^2)\right)^{3/2}}. \tag{3.5}$$

For example, the distribution of peoples' incomes typically shows the vast majority of people earning small to moderate amounts, and just a few people earning large sums, tailing off to the very few who earn astronomically large sums—the Bill Gateses of the world. A distribution is said to be *right-skewed* if the long tail extends in the direction of increasing values and *left-skewed* otherwise. Right-skewed distributions are more common. Symmetric distributions have zero skewness.

3.3 Tools for Displaying Single Variables

One of the most basic displays for univariate data is the histogram, showing the number of values of the variable that lie in consecutive intervals. With small data sets, histograms can be misleading: random fluctuations in the values or alternative choices for the ends of the intervals can give rise to very different diagrams. Apparent multimodality can arise, and then vanish for different choices of the intervals or for a different small sample. As the size of the data set increases, however, these effects diminish. With large data sets, even subtle features of the histogram can represent real aspects of the distribution.

Figure 3.1 shows a histogram of the number of weeks during 1996 in which owners of a particular credit card used that card to make supermarket purchases (the label on the vertical axis has been removed to conceal commercially sensitive details). There is a large mode to the left of the diagram: most people did not use their card in a supermarket, or used it very rarely. The number of people who used the card a given number of times decreases rapidly with increases in the number of times. However, the relatively large number of people represented in this diagram allows us to detect another, much smaller mode toward the right hand end of the diagram. Apparently there is a tendency for people to make regular weekly trips to a supermarket, though this is reduced from 52 annual transactions, probably by interruptions such as holidays.

Example 3.1

Figure 3.2 shows a histogram of diastolic blood pressure for 768 females of Pima Indian heritage. This is one variable out of eight that were collected for the purpose of building classification models for forecasting the onset of diabetes. The documentation for this data set (available online at the UCI Machine Learning data archive) states that there are no missing values in the data. However, a cursory glance at the histogram reveals that about 35 subjects have a blood pressure value of zero, which is clearly impossible if these subjects were alive when the measurements were taken (presumably they were). A

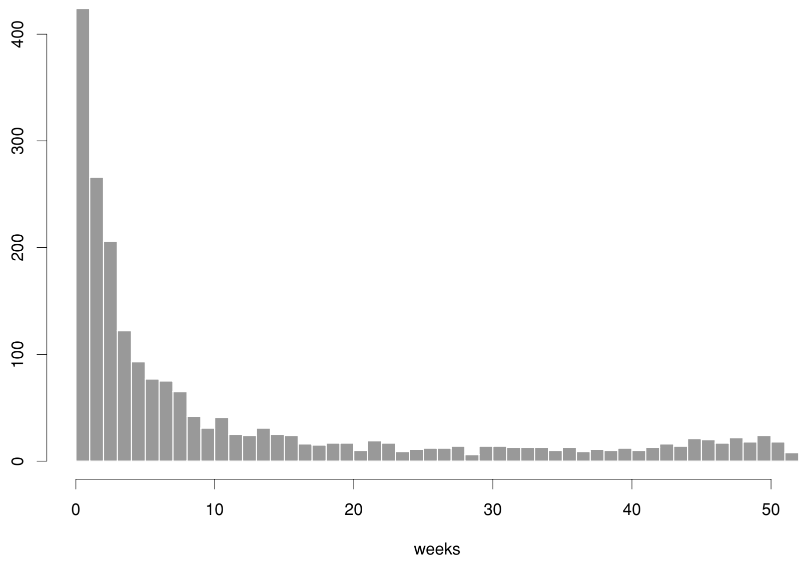

Figure 3.1 Histogram of the number of weeks of the year a particular brand of credit card was used.

plausible explanation is that the measurements for these 35 subjects are in fact missing, and that the value "0" was used in the collection of the data to code for "missing." This seems likely given that a number of the other variables (such as `triceps-fold-skin-thickness`) also have zero-values that are physically impossible.

The point here is that even though the histogram has limitations it is nonetheless often quite valuable to plot data before proceeding with more detailed modeling. In the case of the Pima Indians data, the histogram clearly reveals some suspicious values in the data that are incompatible with the physical interpretations of the variables being measured. Performing such simple checks on the data is always advisable before proceeding to use a data mining algorithm. Once we apply an algorithm it is unlikely that we will notice such data quality problems, and these problems may distort our analysis in an unpredictable manner.

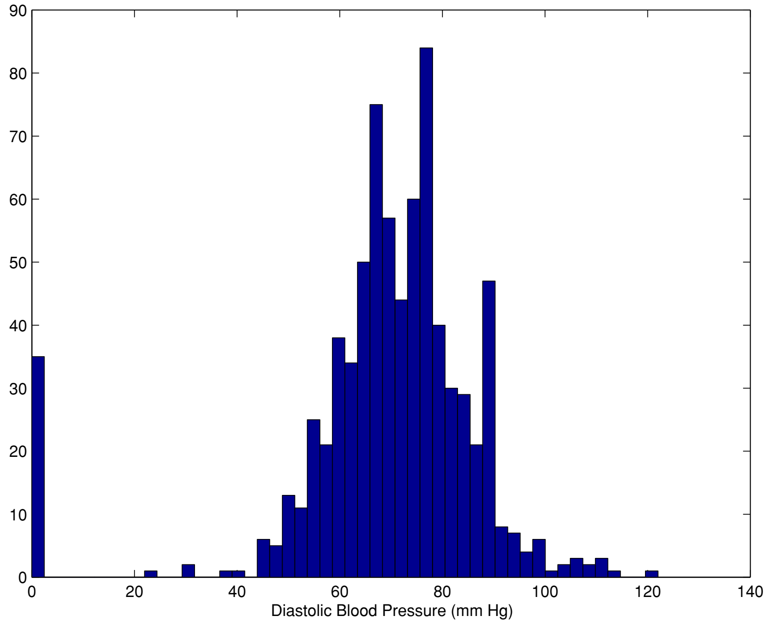

Figure 3.2 Histogram of diastolic blood pressure for 768 females of Pima Indian descent.

The disadvantages of histograms have also been tackled by smoothing estimates. One of the most widely used types is the kernel estimate.

Kernel estimates smooth out the contribution of each observed data point over a local neighborhood of that point (we will revisit the kernel method again in chapter 9). Consider a single variable X for which we have measured values $\{x(1), \ldots, x(n)\}$. The contribution of data point $x(i)$ to the estimate at some point x^* depends on how far apart $x(i)$ and x^* are. The extent of this contribution is dependent upon on the shape of the *kernel function* adopted and the width accorded to it. Denoting the kernel function by K and its width (or bandwidth) by h, the estimated density at any point x is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x(i)}{h}\right), \qquad (3.6)$$

where $\int K(t)dt = 1$ to ensure that the estimate $f(x)$ itself integrates to 1 (i.e., is a proper density) and where the kernel function K is usually chosen to be a smooth unimodal function with a peak at 0. The quality of a kernel estimate depends less on the shape of K than on the value of h.

A common form for K is the Normal (Gaussian) curve, with h as its spread parameter (standard deviation), i.e.,

$$K(t, h) = Ce^{-\frac{1}{2}(\frac{t}{h})^2} \tag{3.7}$$

where C is a normalization constant and $t = x - x(i)$ is the distance of the query point x to data point $x(i)$. The bandwidth h is equivalent to σ, the standard deviation (or width) of the Gaussian kernel function.

There are formal methods for optimizing the fit of these estimates to the unknown distribution that generated the data, but here our interest is in graphical procedures. For our purposes the attraction of such estimates is that by varying h, we can search for peculiarities in the shape of the sample distribution. Small values of h lead to very spiky estimates (not much smoothing at all), while large values lead to oversmoothing. The limits at each extreme of h are the empirical distribution of the data points (i.e., "delta functions" on each data point $x(i)$) as $h \to 0$, and a uniform flat distribution as $h \to \infty$. These limits correspond to the extremes of total commitment to the data (with no mass anywhere except at the observed data points), versus completely ignoring the observed data.

Figure 3.3 shows a kernel estimate of the density of the weights of 856 elderly women who took part in a study of osteoporosis. The distribution is clearly right skewed and there is a hint of multimodality. Certainly the assumption often made in classical statistical work that distributions are normal does not apply in this case. (This is not to say that statistical techniques nominally based on that assumption might not still be valid. Often the arguments are asymptotic—based on normality arising from the central limit theorem. In this case, the assumption that the sample *mean* of 856 subjects would vary from sample to sample according to a normal distribution would be reasonable for practical purposes.)

Figure 3.4 shows what happens when a larger value is used for the smoothing parameter h. Which of the two kernel estimates is "better" is a difficult question to answer. Figure 3.4 is more conservative in that less credence is given to local (potentially random) fluctuations in the observed data values.

Although this section focuses on displaying single variables, it is often desirable to display different groups of scores on a single variable separately, so that the groups may be compared. (Of course, we can think of this as a two-variable situation, in which one of the variables is the grouping factor.)

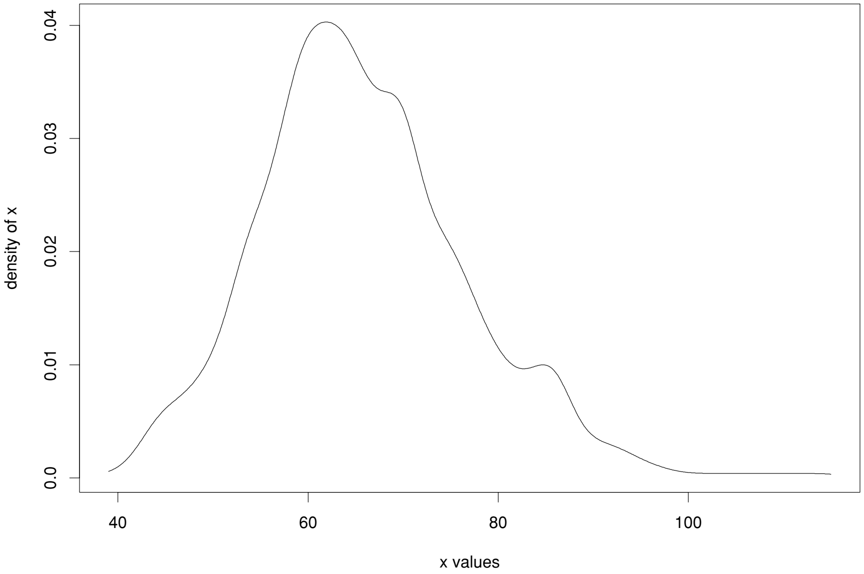

Figure 3.3 Kernel estimate of the weights (in kg) of 856 elderly women.

Histograms, kernel plots, and other unidimensional displays can be used separately for each group. However, this can become unwieldy if there are more than two or three groups. In such cases a useful alternative display is the box and whisker plot.

Although various versions of box and whisker plots exist, the essential ideas are the same. A *box* containing which the bulk of the data is defined—for example, the interval between the first and third quartiles. A line across this box indicates some measure of location—often the median of the data. Whiskers project from the ends of the box to indicate the spread of the tails of the empirical distribution.

We illustrate the boxplot using a subset of the diabetes data set from figure 3.2. Figure 3.5 shows four panels of box plots, each containing a separate boxplot for each of the two classes in the data, *healthy* (1) and *diabetic* (2).The diagrams show clearly how mean, dispersion, and skewness vary with val-

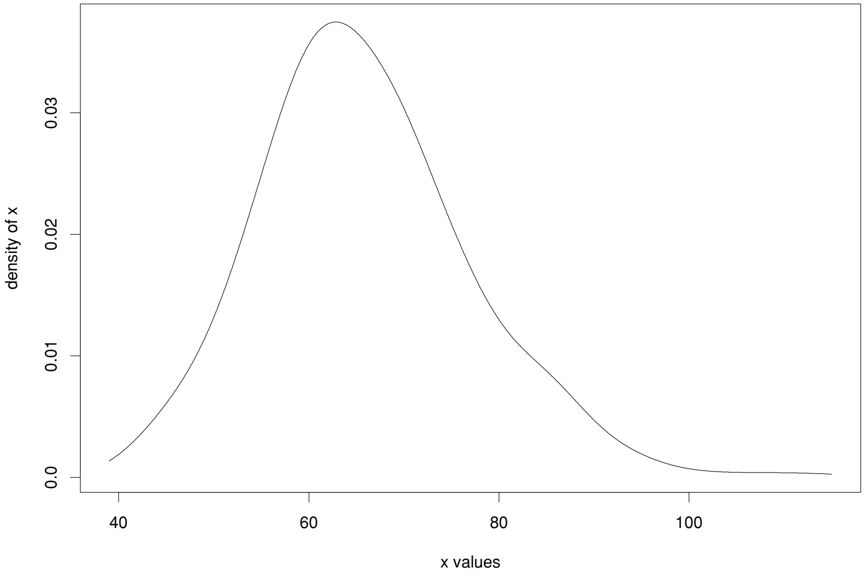

Figure 3.4 As figure 3.3, but with more smoothing.

ues of the grouping variable.

3.4 Tools for Displaying Relationships between Two Variables

The scatterplot is a standard tool for displaying two variables at a time. Figure 3.6 shows the relationship between two variables describing credit card repayment patterns (the details are confidential). It is clear from this diagram that the variables are strongly correlated—when one value has a high (low) value, the other variable is likely to have a high (low) value. However, a significant number of people depart from this pattern; showing high values on one of the variables and low values on the other. It might be worth investigating these individuals to find out why they are unusual.

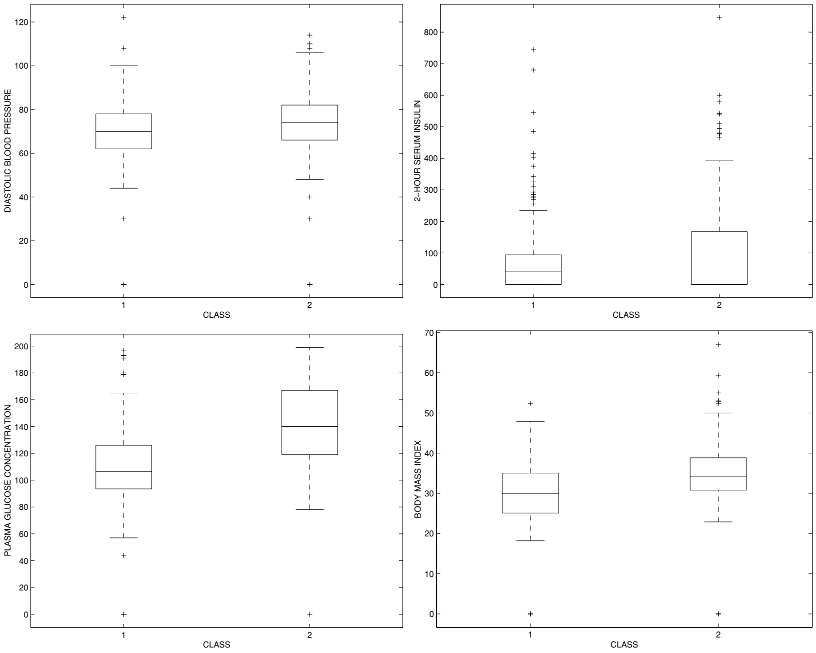

Figure 3.5 Boxplots on four different variables from the Pima Indians diabetes data set. For each variable, a separate boxplot is produced for the healthy subjects (labeled 1) and the diabetic subjects (labeled 2). The upper and lower boundaries of each box represent the upper and lower quartiles of the data respectively. The horizontal line within each box represents the median of the data. The whiskers extend 1.5 times the interquartile range from the end of each box. All data points outside the whiskers are plotted individually (although some overplotting is present, e.g., for values of 0).

Unfortunately, in data mining, scatterplots are not always so useful. If there are too many data points we will find ourselves looking at a purely black rectangle. Figure 3.7 illustrates this sort of problem. This shows a scatterplot of 96,000 points from a study of bank loans. Little obvious structure is discernible, although it might appear that later applicants in general are older. On the other hand, the apparent greater vertical dispersion toward the right end of the diagram could equally be caused by a greater number of samples on the right side. In fact, the linear regression fit to these data has a very small but highly significant *downward* slope.

Even when the situation is not quite so extreme, scatterplots with large

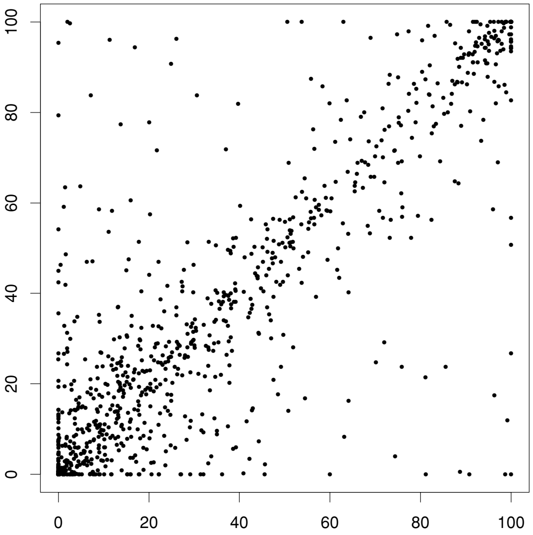

Figure 3.6 A standard scatterplot for two banking variables.

numbers of points can conceal more than they reveal. Figure 3.8 plots the number of weeks a particular credit card was used to buy petrol (gasoline) in a given year against the number of weeks the card was used in a supermarket (each data point represents an individual credit card). There is clearly some correlation, but the actual correlation 0.482 is much higher than it appears here. The diagram is deceptive because it conceals a great deal of overprinting in the bottom left corner—there are 10,000 customers represented here altogether. The bimodality shown in figure 3.1 can also be discerned in this figure, though not as easily as in figure 3.1.

Another curious phenomenon is also apparent in figure 3.8. The distribution of the number of weeks the card was used in a petrol station is skewed for low values of the supermarket variable, but fairly uniform for high values. What could explain this? (Of course, bearing in mind the point above, this apparent phenomenon needs to be checked for overprinting.)

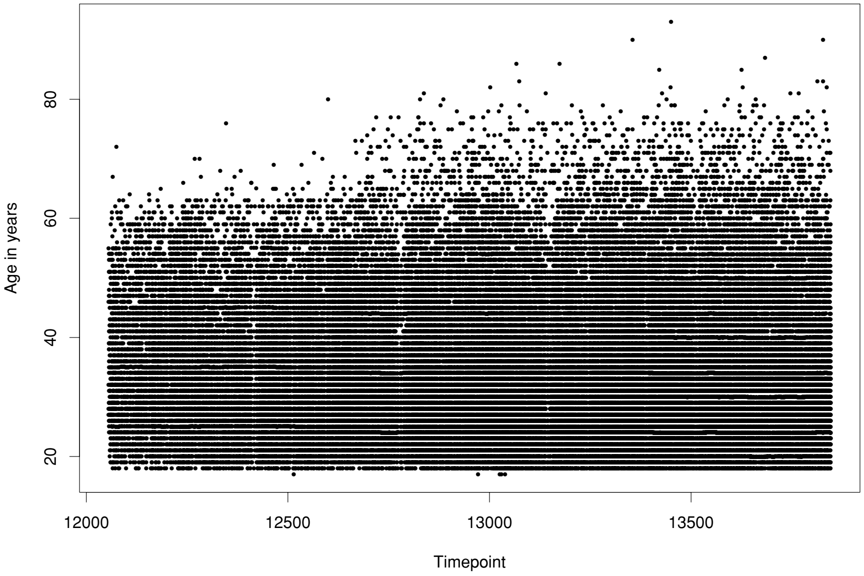

Figure 3.7 A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.

Contour plots can help overcome some of these problems. Note that creating a contour plot in two dimensions effectively requires us to construct a two-dimensional density estimate, using something like a two-dimensional generalization of the kernel method of equation 3.6, again raising the issue of bandwidth selection but now in a two-dimensional context. A contour plot of the 96,000 points shown in figure 3.7 is given in figure 3.9. Certain trends are clear from this display that cannot be discerned in figure 3.7. For instance the density of points increases toward the right side of the diagram; the apparent increasing dispersion of the vertical axis is due to there being a greater concentration of points in that area. The vertical skewness of the data is also very evident in this diagram. The unimodality of the data, and

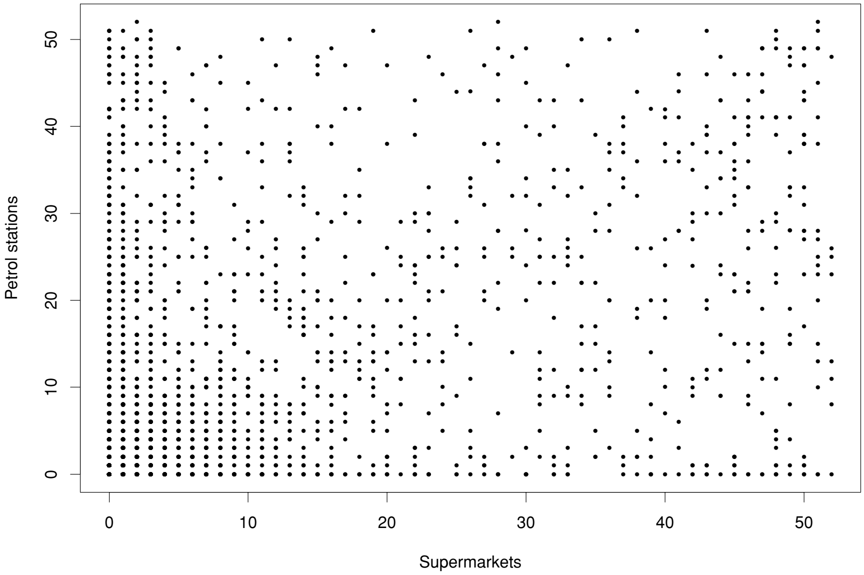

Figure 3.8 Overprinting conceals the actual strength of the correlation.

the position of the single mode cannot be seen at all in figure 3.7 but is quite clear in figure 3.9. Note that since the horizontal axis in these plots is time, an alternative way to display the data is to plot contours of constant conditional probability density, as time progresses.

Other standard forms of display can be used when one of the two variables is time, to show the value of the other variable as time progresses. This can be a very effective way of detecting trends and departures from expected or standard behaviour. Figure 3.10 shows a plot of the number of credit cards issued in the United Kingdom from 1985 to 1993 inclusive. A smooth curve has been fitted to the data to place emphasis on the main features of the relationship. It is clear that around 1990 something caused a break in a growth pattern that had been linear up to that point. In fact, what happened was that in 1990 and 1991 annual fees were introduced for credit cards, and many users reduced their holding to a single card.

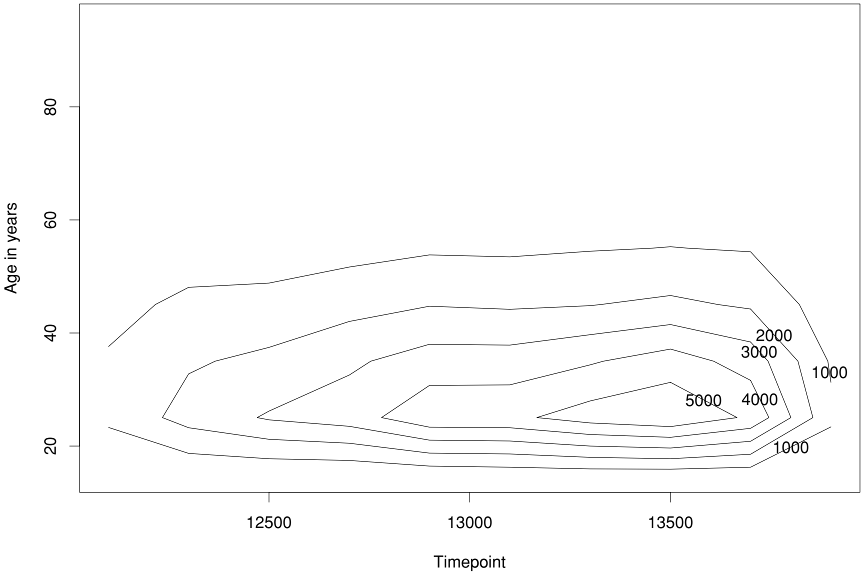

Figure 3.9 A contour plot of the data from figure 3.7.

Figure 3.11 shows a plot of the number of miles flown by UK airlines, during each month from January 1963 to December 1970. There are several patterns immediately apparent from this display that conform with what one might expect to observe, such as the gradually increasing trend and the periodicity (with large peaks in the summer and small peaks around the new year). The plot also reveals an interesting bifurcation of the summer peak, suggesting a tendency for travelers to favor the early and late summer over the middle period.

Figure 3.12 provides a third example of the power of plots in which time is one of the two variables. From February to June 1930, an experiment was carried out in Lanarkshire, Scotland to investigate whether adding milk to children's diets had an effect on "physique, general health and increasing mental alertness" (Leighton and McKinlay, 1930). In this study 20,000 children were allocated to one of three groups; 5000 of the children received

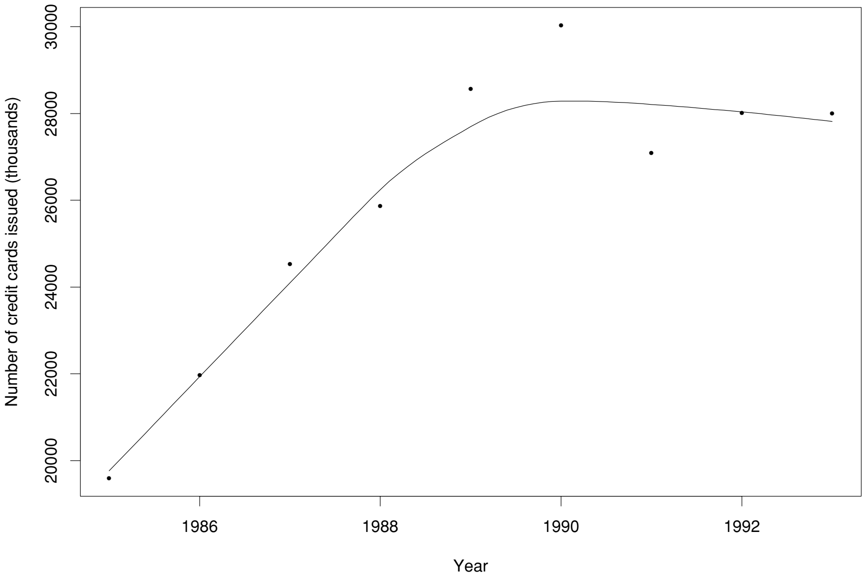

Figure 3.10 A plot of the number of credit cards in circulation in the United Kingdom, by year.

three-quarters of a pint of raw milk per day, 5000 received three-quarters of a pint of pasteurized milk per day, and 10,000 formed a control group receiving no dietary milk supplement. The children were weighed at the start of the experiment and again four months later. Interest lay in whether there was differential growth between the three groups.

Figure 3.12 plots the mean weight of the control group of girls against the mean age of the group they are in. The first point corresponds to the youngest age group (mean age 5.5 years) at the start of the experiment, and the second point corresponds to this group four months later. The third and fourth points correspond to the second age group, and so on. The points are connected by lines to make the shape easier to discern. Similar shapes are apparent for all groups in the experiment.

The plot immediately reveals an unexpected pattern that cannot be seen

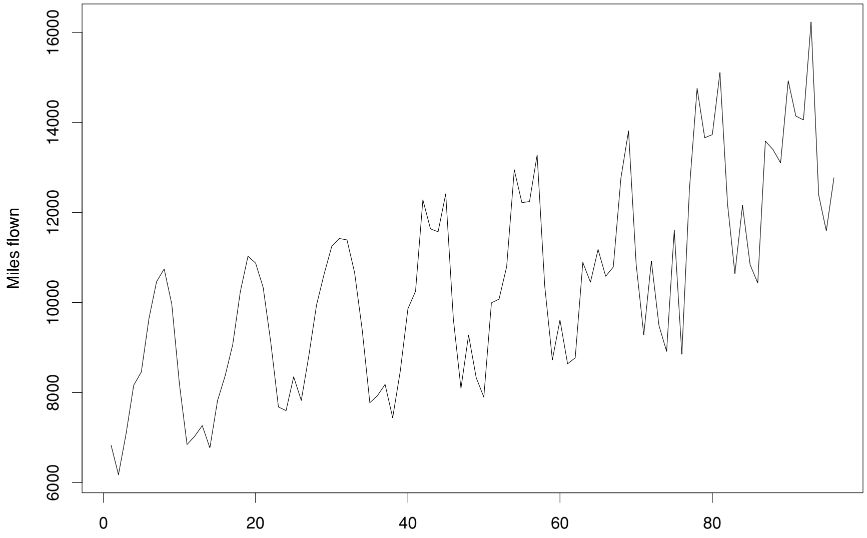

Figure 3.11 Patterns of change over time in the number of miles flown by UK airlines in the 1960s.

from a table of the data. We would expect a smooth plot, but there are clear steps evident here. It seems that each age group does not gain as much weight as expected. There are various possible explanations for this shape. Perhaps children grow less during the early months of the year than during the later ones. However, similar plots of heights show no such intermittent growth, so we need a more elaborate explanation in which height increases uniformly but weight increases in spurts. Another possible explanation arises from the fact that the children were weighed in their clothes. The report does say, "All of the children were weighed without their boots or shoes and wearing only their ordinary outdoor clothing. The boys were made to turn out the miscellaneous collection of articles that is normally found in their pockets, and overcoats, mufflers, etc., were also discarded. Where a child was found to be wearing three or four jerseys—a not uncom-

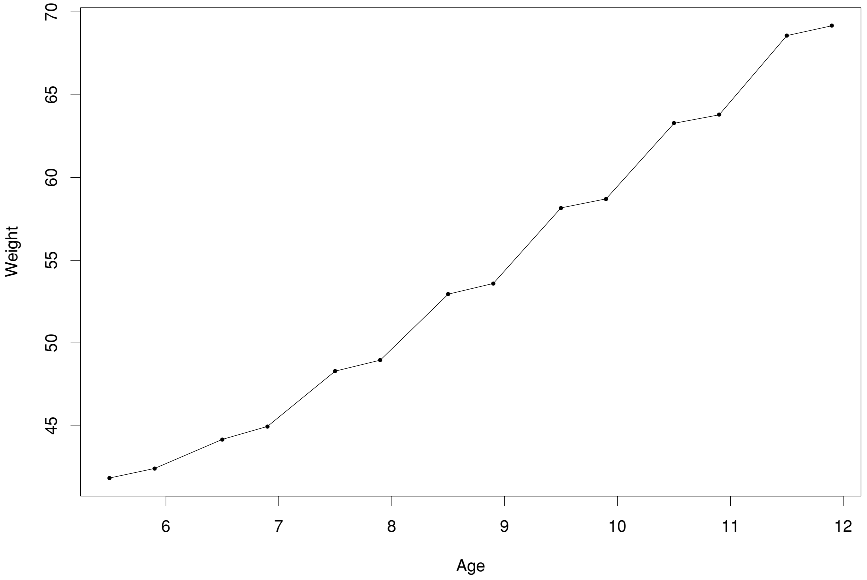

Figure 3.12 Weight changes over time in a group of 10,000 school children in the 1930s. The steplike pattern in the data highlights a problem with the measurement process.

mon experience—all in excess of one were removed." It still seems likely, however, that the summer garb was lighter than the winter garb. This example illustrates that the patterns discovered by data mining may not shed much light on the phenomena under investigation, but finding data anomalies and shortcomings may be just as valuable.

3.5 Tools for Displaying More Than Two Variables

Since sheets of paper and computer screens are flat, they are readily suited for displaying two-dimensional data, but are not effective for displaying higher dimensional data. We need some kind of projection, from the higher dimensional data to a two dimensional plane, with modifications to show (aspects of) the other dimensions. The most obvious approach along these lines is to examine the relationships between all pairs of variables, extending the basic scatterplot described in section 3.3 to a *scatterplot matrix*.

Figure 3.13 illustrates a scatterplot matrix for characteristics, performance measures, and relative performance measures of 209 computer CPUs dating from over 10 years ago. The variables are cycle time, minimum memory (kb), maximum memory (kb), cache size (kb), minimum channels, maximum channels, relative performance, and estimated relative performance (relative to an IBM 370/158-3). While some pairs of variables appear to be unrelated, others are strongly related. *Brushing* allows us to highlight points in a scatterplot matrix in such a way that the points corresponding to the same objects in each scatterplot are highlighted. This is particularly useful in interactive exploration of data.

Of course, scatterplot matrices are not really multivariate solutions: they are multiple bivariate solutions, in which the multivariate data are projected into multiple two-dimensional plots (and in each two-dimensional plot all other variables are ignored). Such projections necessarily sacrifice information. Picture a cube formed from eight smaller cubes. If data points are uniformly distributed in alternate subcubes, with the others being empty, all three one-dimensional and all three two-dimensional projections show uniform distributions. (This "exclusive-or" structure caused great difficulty with perceptrons—the precursors of today's neural networks which we will discuss in chapters 5 and 11.)

Interactive graphics come into their own when more than two variables are involved, since then we can rotate ("spin") the direction of projection in a search for structure. Some systems even let the software follow random rotations, while we watch and wait for interesting structures to become apparent. While this is a good idea in principle, the excitement of watching a cloud of points shift relative position as the direction of viewing changes can quickly pall, and more structured methods are desirable. Projection pursuit, described in chapter 11, is one such method.

Trellis plotting also utilizes multiple bivariate plots. Here, however, rather than displaying a scatterplot for each pair of variables, they fix a particular pair of variables that is to be displayed and produce a series of scatterplots conditioned on levels of one or more other variables.

Figure 3.14 shows a trellis plot for data on epileptic seizures. The horizontal axis of each plot gives the number of seizures that 58 patients experienced over a certain two week period, and the vertical axis gives the number of seizures experienced over a later two week period. The two left hand graphs

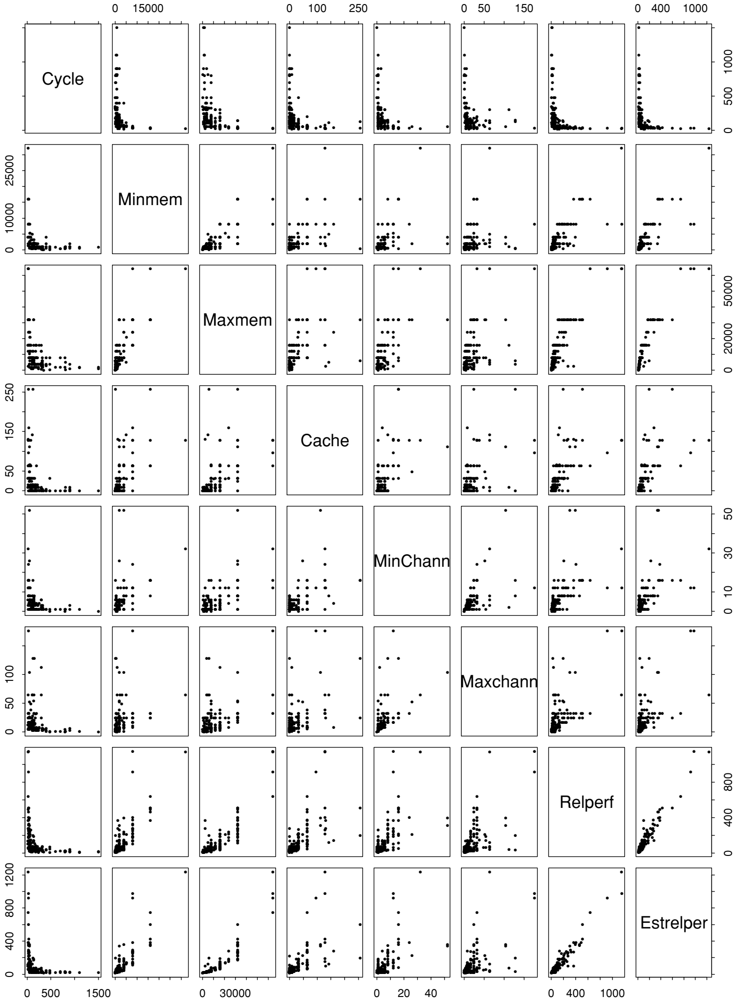

Figure 3.13 A scatterplot matrix for the computer CPU data.

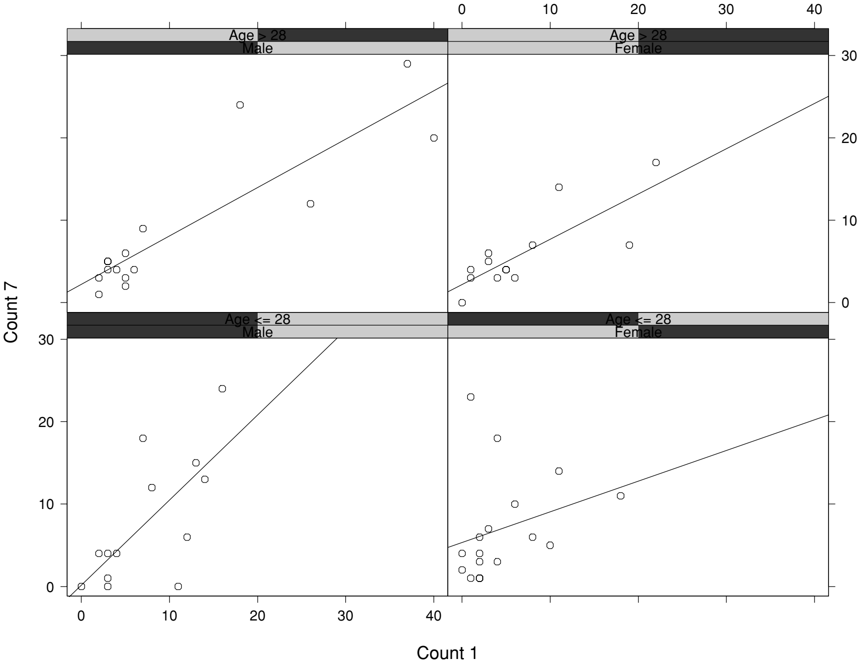

Figure 3.14 A trellis plot for the epileptic seizures data.

show the figures for males, and the two right hand graphs the figures for females. The two upper graphs show ages 29 to 42 while the two lower graphs show ages 18 to 28. (The original data set included the record of another subject who had much higher counts. We have removed this subject here so that we can more clearly see the relationships between the scores of the other subjects.) From these plots, we can see that the younger group show lower average counts than the older group. The figures also hint at some possible differences between the slopes of the estimated best fitting lines relating the y and x axes, though we would need to carry out formal tests to be confident that these differences were real.

Trellis plots can be produced with any kind of component graph. Instead of scatterplots in each cell, we could have histograms, time series plots, contour plots, or any other types of plots.

An entirely different way to display multivariate data is through the use of *icons*, small diagrams in which the sizes of different features are determined by the values of particular variables. Star icons are among the most popular. In these, different directions from the origin correspond to different variables, and the lengths of radii projecting in these directions correspond to the magnitudes of the variables. Figure 3.15 shows an example. The data displayed here come from 12 chemical properties that were measured on 53 mineral samples equally spaced along a long drill into the Earth's surface.

Another type of icon plot, Chernoff's faces, is discussed frequently in introductory texts on the subject. In these plots, the sizes of features in cartoon faces (length of nose, degree of smile, shape of eyes, etc.) represent the values of the variables. The method is based on the principle that the human eye is particularly adept at recognizing and distinguishing between faces. Although they are entertaining, plots of this type are seldom used in serious data analysis since the idea does not work very well in practice with more than a handful of cartoon faces. In general, iconic representations are effective only for relatively small numbers of cases since they require the eye to scan each case separately.

Parallel coordinates plots show variables as parallel axes, representing each case as a piecewise linear plot connecting the measured values for that case. Figure 3.16 shows such a plot for four repeated measurements of the number of epileptic seizures experienced by 58 patients during successive two week periods. The data are clearly skewed and might be modeled by a Poisson distribution (see Appendix). Since the data set is not too large, we can follow the trajectories of individual patients.

Another way of representing dimensions is through the use of color. Line styles, as in the parallel coordinates plot above, can serve the same purpose.

No single method of representing multivariate data is a universal solution. Which method is most useful in a given situation will depend on the data and on the structures being sought.

3.6 Principal Components Analysis

Scatterplots project multivariate data into a two-dimensional space defined by just two of the variables. This allows us to examine pairwise relationships between variables, but such simple projections might conceal more complicated relationships. To detect these relationships we can use projec-

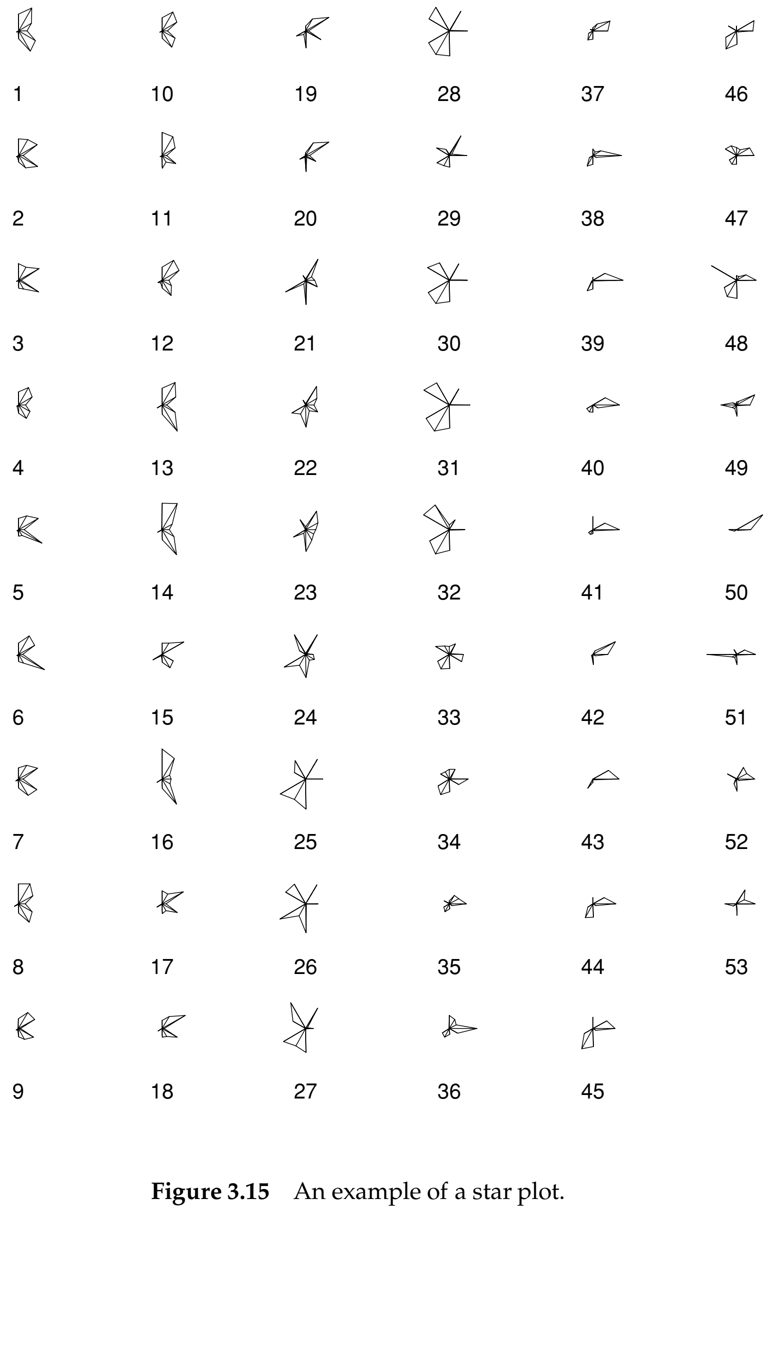

Figure 3.15 An example of a star plot.

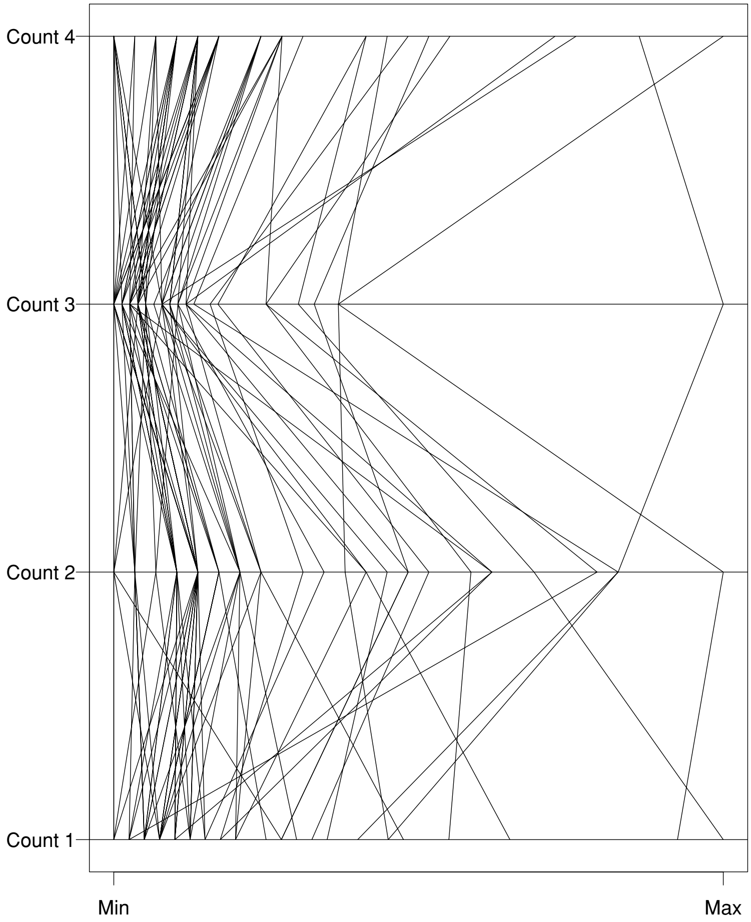

Figure 3.16 A parallel coordinates plot for the epileptic seizure data.

tions along different directions, defined by any weighted linear combination of variables (e.g., along the direction defined by $2x_1 + 3x_2 + x_3$).

With only a few variables, it might be feasible to search for such interesting spaces manually, rotating the distribution of the data. With more than a few variables, however, it is best to let the computer loose to search by itself. To do this, we need to define what an "interesting" projection might look like, so that the computer knows when it has found one. *Projection pursuit methods* are based on this general principle of allowing the computer to search for interesting directions. (Such techniques, however, are computationally quite intensive: we will return to projection pursuit in chapter 11 when we discuss regression.)

However, in one special case—for one specific definition of what constitutes an "interesting" direction—a computationally efficient explicit solution can be found. This is when we seek the projection onto the two-dimensional plane for which the sum of squared differences between the data points and their projections onto this plane is smaller than when any other plane is used. (We use two-dimensional projections here for convenience, but in general we can use any k-dimensional projection, $1 \leq k \leq p-1$). This two-dimensional plane can be shown to be spanned by (1) the linear combination of the variables that has maximum sample variance and (2) the linear combination that has maximum variance subject to being uncorrelated with the first linear combination. Thus "interesting" here is defined in terms of the *maximum variability* in the data.

Of course, we can take this process further, seeking additional linear combinations that maximize the variance subject to being uncorrelated with all those already selected. In general, if we are lucky, we find a set of just a few such linear combinations ("components") that describes the data fairly accurately. The mathematics of this process is described below. Our aim here is to capture the intrinsic variability in the data. This is a useful way of reducing the dimensionality of a data set, either to ease interpretation or as a way to avoid overfitting and to prepare for subsequent analysis.

Suppose that $\mathbf{X}$ is an $n \times p$ data matrix in which the rows represent the cases (each row is a data vector $\mathbf{x}\,(i)$) and the columns represent the variables. Strictly speaking, the ith row of this matrix is actually the transpose $\mathbf{x}^T$ of the ith data vector $\mathbf{x}(i)$, since the convention is to consider data vectors as being $p \times 1$ column vectors rather than $1 \times p$ row vectors. In addition, assume that $\mathbf{X}$ is mean-centered so that the value of each variable is relative to the sample mean for that variable (i.e., the estimated mean has been subtracted from each column).

Let $\mathbf{a}$ be the $p \times 1$ column vector of projection weights (unknown at this point) that result in the largest variance when the data $\mathbf{X}$ are projected along $\mathbf{a}$. The projection of any particular data vector $\mathbf{x}$ is the linear combination $\mathbf{a}^T\mathbf{x} = \sum_{j=1}^{p} a_j x_j$. Note that we can express the projected values onto $\mathbf{a}$ of all data vectors in $\mathbf{X}$ as $\mathbf{Xa}$ ($n \times p$ by $p \times 1$, yielding an $n \times 1$ column vector of projected values). Furthermore, we can define the *variance* along $\mathbf{a}$ as

$$\begin{aligned} \sigma_{\mathbf{a}}^2 &= \left(\mathbf{Xa}\right)^T\left(\mathbf{Xa}\right) \\ &= \mathbf{a}^T\mathbf{X}^T\mathbf{Xa} \\ &= \mathbf{a}^T V\mathbf{a}, \end{aligned} \tag{3.8}$$

where $V = \mathbf{X}^T\mathbf{X}$ is the $p \times p$ covariance matrix of the data (since $\mathbf{X}$ has zero mean), as defined in chapter 2. Thus, we can express $\sigma_{\mathbf{a}}^2$ (the variance of the projected data (a scalar) that we wish to maximize) as a function of both $\mathbf{a}$ and the covariance matrix of the data $\mathbf{V}$.

Of course, maximizing $\sigma_{\mathbf{a}}^2$ directly is not well-defined, since we can increase $\sigma_{\mathbf{a}}^2$ without limit simply by increasing the size of the components of $\mathbf{a}$. Some kind of constraint must be imposed, so we impose a normalization constraint on the $\mathbf{a}$ vectors such that $\mathbf{a}^T\mathbf{a} = 1$.

With this normalization constraint we can rewrite our optimization problem as that of maximizing the quantity

$$u = \mathbf{a}^T\mathbf{Va} - \lambda(\mathbf{a}^T\mathbf{a} - 1), \tag{3.9}$$

where λ is a Lagrange multiplier. Differentiating with respect to $\mathbf{a}$ yields

$$\frac{\partial u}{\partial \mathbf{a}} = 2\mathbf{Va} - 2\lambda\mathbf{a} = 0, \tag{3.10}$$

which reduces to the familiar eigenvalue form of

$$(\mathbf{V} - \lambda\mathbf{I})\mathbf{a} = 0. \tag{3.11}$$

Thus, the first principal component $\mathbf{a}$ is the eigenvector associated with the largest eigenvalue of the covariance matrix $\mathbf{V}$. Furthermore, the second principal component (the direction orthogonal to the first component that has the largest projected variance) is the eigenvector corresponding to the second largest eigenvalue of $\mathbf{V}$, and so on (the eigenvector for the kth largest eigenvalue corresponds to the kth principal component direction).

In practice of course we may be interested in projecting to more than two-dimensions. A basic property of this projection scheme is that if the data are

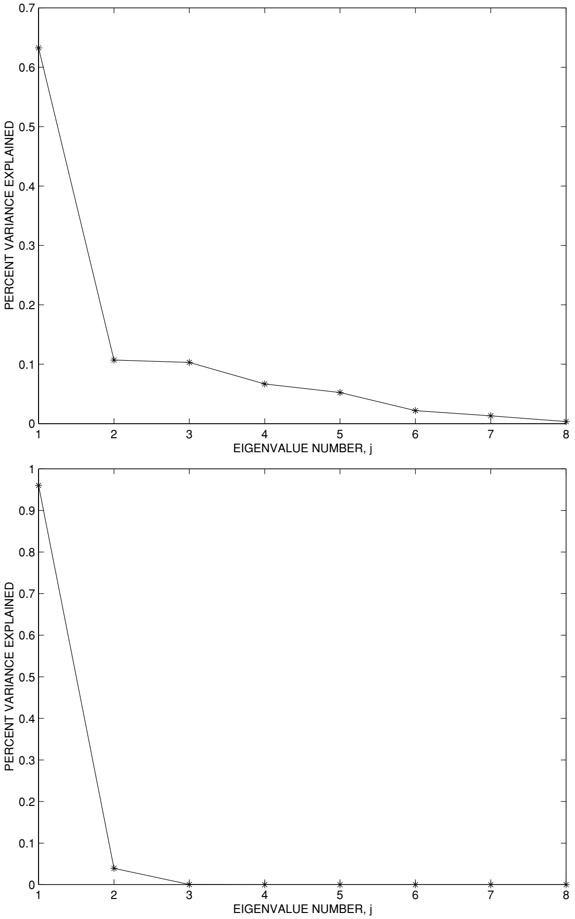

Figure 3.17 Scree plots for the computer CPU data set. The upper plot displays the eigenvalues from the correlation matrix, and the lower plot is for the covariance matrix.

projected into the first k eigenvectors, the variance of the projected data can be expressed as $\sum_{j=1}^{k} \lambda_j$, where λ_j is the jth eigenvalue. Equivalently, the squared error in terms of approximating the true data matrix X using only the first k eigenvectors can be expressed as

$$\frac{\sum_{j=k+1}^{p} \lambda_j}{\sum_{l=1}^{p} \lambda_l}. \tag{3.12}$$

Thus, in choosing an appropriate number k of principal components, one approach is to increase k until the squared error quantity above is smaller than some acceptable degree of squared error. For high-dimensional data sets, in which the variables are often relatively well-correlated, it is not uncommon for a relatively small number of principal components (say, 5 or 10) to capture 90% or more of the variance in the data.

A useful visual aid in this context is the *scree plot*—which shows the amount of variance explained by each consecutive eigenvalue. This is necessarily nonincreasing with the number of the component, and the hope is that it demonstrates a sudden dramatic fall toward zero. A principal components analysis of the correlation matrix of the computer CPU data described earlier gives rise to eigenvalues proportional to 63.26, 10.70, 10.30, 6.68, 5.23, 2.18, 1.31, and 0.34 (see figure 3.17). The fall from the first to the second eigenvalue is dramatic, but after that the decline is gradual. (The weights that the first component puts on the eight variables are (0.199, -0.365, -0.399, -0.336, -0.331, - 0.298, -0.421, -0.423). Note that, it gives them all roughly similar weights, but gives the first variable (cycle time) a weight opposite in sign to those of the other variables.) If, instead of the correlation matrix, we analyzed the covariance matrix, the variables with larger ranges of values would tend to dominate. In the case of these data, the values given for memory are much larger than those for the other variables. (This is because they are given in kilobytes. Had they been given in megabytes, this would not be the case—an example of the arbitrariness of the scaling of noncommensurate variables (see chapter 2)). Principal components analysis of the covariance matrix gives proportions of variation attributable to the different components as 96.02, 3.93, 0.04, 0.01, 0.00, 0.00, 0.00, and 0.00 (see figure 3.17). Here the fall from the first component is very striking—the variability in the data can, indeed, be explained almost entirely by the differences in memory capacity. Often, however, there is no obvious fall such as this—no point at which the remaining variance in the data can be attributed to random variation. Then the choice of how many components to extract is fairly arbitrary. The proportion of the total variance that we regard as providing an adequate simplified description of the data depends on the field of application. In some cases it might be sufficient for the first few components to describe 60% of the variance, but in other fields one might hope for 95% or more.

When conducting principal components analysis prior to further analyses, it is risky to choose a small number of components that fail to explain the variability in the data very well. Information is lost, and there is no guarantee that the sacrificed information is not relevant to the aims of further analyses. (Indeed, this is true even if the retained components do explain the variability well, short of 100%.) For example, we might perform principal components analysis prior to classifying our data. Since the aims of dimension reduction and classification are somewhat different, it is possible that the reduction to a few spanning components may lose valuable information about the differences between the classes—we will see an example of this at the end of chapter 9. Likewise, for many multivariate data sets in which the points fall into two (or more) classes, a prior principal components analysis may completely obliterate the differences between the distributions of the classes. On the other hand, in regression problems (chapter 11) with many explanatory variables, unless the data set is large, there may be problems of instability of the estimated coefficients. A principal components analysis is sometimes performed to reduce the large number of explanatory variables to a few linear combinations prior to carrying out the regression analysis.

Despite the risks of failing to extract relevant information, principal components analysis is a powerful and valuable tool. Because it is based on linear projections and minimizing the variance (or sum of squared errors), numerical manipulations can be carried out explicitly, without any iterative searches. Computing the principal component solutions directly from the eigenvector equations will scale roughly as $O(np^2 + p^3)$ (np^2 to calculate $\mathbf{V}$ and p^3 to solve the eigenvalue equations for the $p \times p$ matrix). This means that it can be applied to data sets with large numbers of records n (but does not scale so well as a function of dimensionality p). As illustrated above when we applied principal components analysis to both correlation and covariance matrices, the method is not invariant under rescalings of the original variables. The appropriate steps to take will depend on the objectives of the analysis. Typically we rescale the data if different variables measure different attributes (e.g., height, weight, and lung capacity) since otherwise the results of a direct principal components analysis depend on the arbitrary choice of units used for each attribute.

To illustrate the simple graphical use of principal components analysis, figure 3.18 shows the projections (indicated by the numbers) of 17 pills onto the space spanned by the first two principal components. The six measurements

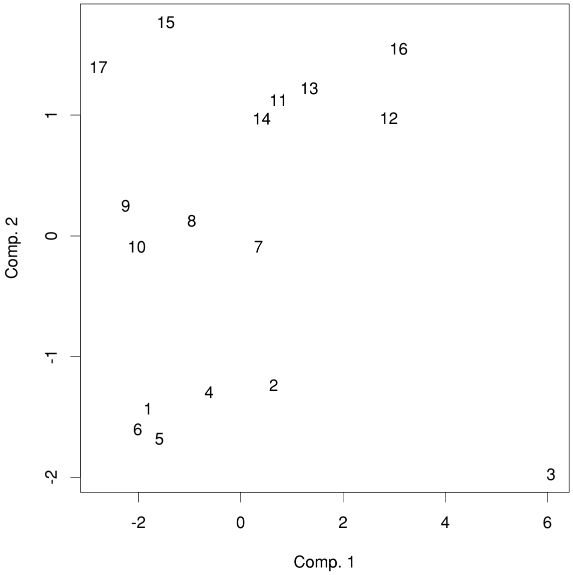

Figure 3.18 Projection onto the first two principal components.

on each pill are the times at which a specified proportion (10%, 30%, 50%, 70%, 75%, and 90%) of the pill has dissolved. It is clear from this diagram that one of the pills is very different from the others, lying in the bottom right corner, far from the other points.

Sometimes we can gain insights from the pattern of weights (or loadings, as they are sometimes called) defining the components of a principal components analysis. Huba et al. (1981) collected data on 1684 students in Los Angeles showing consumption of each of thirteen legal and illegal psychoactive substances: cigarettes, beer, wine, spirits, cocaine, tranquilizers, drug store medications used to get high, heroin and other opiates, marijuana, hashish, inhalants (such as glue), hallucinogenics, and amphetamines. They scored each as 1 (never tried), 2 (tried only once), 3 (tried a few times), 4 (tried many times), 5 (tried regularly). Taking these variables in order, the weights of the first component from a principal components analysis were (0.278, 0.286, 0.265, 0.318, 0.208, 0.293, 0.176, 0.202, 0.339, 0.329, 0.276, 0.248, 0.329). This component assigns roughly equal weights to each of the variables and can be regarded as a general measure of how often students use such substances. Thus, the biggest difference between the students is in terms of how often they use psychoactive substances, regardless of which substances they use.

The second component had weights (0.280, 0.396, 0.392, 0.325, -0.288, -0.259, -0.189, -0.315, 0.163, -0.050, -0.169, -0.329, -0.232). This is interesting because it gives positive weights to the legal substances and negative weights to the illegal ones: therefore, once we have controlled for overall substance use, the major difference between the students lies in their use of legal versus illegal substances. This is just the sort of relationship one would hope to discover from a data mining exercise.

Another statistical technique, *factor analysis,* is often confused with principal components analysis, but the two have very different aims. As described above, principal components analysis is a transformation of the data to new variables. We can then select just some of these as providing an adequate description of the data. Factor analysis, on the other hand, is a *model* for data, based on the notion that we can define the measured variables $X_1, \ldots, X_p$ as linear combinations of a smaller number m ($m < p$) of "latent" (unobserved) factors—variables that cannot be measured explicitly. The objective of factor analysis is to unearth information about these latent variables.

We can define $\mathbf{F} = (F_1, ..., F_m)^T$ as the $m \times 1$ column vector of unknown latent variables, taking values $\mathbf{f} = (f_1, ..., f_m)$. Then a measured data vector $\mathbf{x} = (x_1, ..., x_p)^T$ (defined here as a $p \times 1$ column vector) is regarded as a linear function of $\mathbf{f}$ defined by

$$\mathbf{x} = \mathbf{\Lambda f} + \mathrm{e}. \tag{3.13}$$

Here $\mathbf{\Lambda}$ is a $p \times m$ matrix of *factor loadings* giving the weights with which each factor contributes to each manifest variable. The components of the $p \times 1$ vector e are uncorrelated random variables, sometimes termed *specific factors* since they contribute only to single manifest (observed) variables, $X_j, 1 \leq j \leq p$. Factor analysis is a special case of structural linear relational models described in chapter 9, so we will not dwell on estimation procedures here. However, since factor analysis was the earliest model structure of this form to be developed, it has a special place, not only because of its history, but also because it continues to be among the most widely used of such models.

Factor analysis has not had an entirely uncontroversial history, partly because its solutions are not invariant to various transformations. It is easy to see that new factors can be defined from equation 3.13 via $m \times m$ orthogonal matrices $\mathbf{M}$, such that $\mathbf{x} = (\mathbf{\Lambda M})(\mathbf{Mf}) + \mathrm{e}$. This corresponds to rotating the factors in the space they span. Thus, the extracted factors are essentially nonunique, unless extra constraints are imposed. There are various constraints in general use, including methods that seek to extract factors for which the weights are as close to 0 or 1 as possible, defining the variables as clearly as possible in terms of a subset of the factors.

3.7 Multidimensional Scaling

In the preceding section we described how to use principal components analysis to project a multivariate data set onto the plane in which the data has maximum dispersion. This allows us to examine the data visually, while sacrificing the minimum amount of information. Such a method is effective only to the extent that the data lie in a two-dimensional linear subspace of the area spanned by the measured variables. But what if the data forms a set that is intrinsically two-dimensional, but instead of being "flat," is curved or otherwise distorted in the space spanned by the original variables? (Imagine a crumpled piece of paper, intrinsically two-dimensional, but occupying three dimensions.) In this event it is quite possible that principal components analysis might fail to detect the underlying two-dimensional structure. In such cases, multidimensional scaling can be helpful. Multidimensional scaling methods seek to represent data points in a lower dimensional space while preserving, as far as is possible, the distances between the data points. Since, we are mostly concerned with two-dimensional representations, we shall restrict most of our discussion to such cases. The extension to higher dimensional representations is immediate.

Many multidimensional scaling methods exist, differing in how they define the distances that are being preserved, the distances they map to, and how the calculations are performed. Principal components analysis may be regarded as a basic form. In this approach the distances between the data points are taken as Euclidean (or Pythagorean), and they are mapped to distances in a reduced space that are also measured using the Euclidean metric. The sum of squared distances between the original data points and their projections provides a measure of quality of the representation. Other methods of multidimensional scaling also have associated measures of the quality of the representation.

Since multidimensional scaling methods seek to preserve interpoint distances, such distances can serve as the starting point for an analysis. That is, we do not need to know any measured values of variables for the objects being analyzed, only how similar the objects are, in terms of some distance measure. For example, the data may have been collected by asking respondents to rate the similarity between pairs of objects. (A classic example of this is a matrix showing the number of times the Morse codes for different letters are confused. There are no "variables" here, simply a matrix of "similarities" measuring how often is letter was mistaken for another.) The end point of the process is the same—a configuration of data points in a two–dimensional space. In a sense, the objects and the raters are used to determine on what dimensions "similarity" is to be measured. Multidimensional scaling methods are widely used in areas such as psychometrics and market research, in attempts to understand perceptions of relationships and similarities between objects.

From an $n \times p$ data matrix $\mathbf{X}$ we can compute an $n \times n$ matrix $\mathbf{B} = \mathbf{X}\mathbf{X}^T$. (Since this scales as $O(n^2)$ in both time and memory, it is clear that this approach is not practical for very large numbers of objects n). It is straightforward to see from this that the Euclidean distance between the ith and jth objects is given by

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}. \tag{3.14}$$

If we could invert this relationship, then, given a matrix of distances $\mathbf{D}$ (derived from original data points by computing Euclidean distances or obtained by other means), we could compute the elements of $\mathbf{B}$. $\mathbf{B}$ could then be factorized to yield the coordinates of the points. One factorization of $\mathbf{B}$ would be in terms of the eigenvectors. If we chose those associated with the two largest eigenvalues, we would have a two-dimensional representation that preserved the structure of the data as well as possible.

The feasibility of this procedure hinges upon our ability to invert equation 3.14. Unfortunately, this is not possible without imposing some extra constraints. Because shifting the mean and rotating a configuration of points does not affect the interpoint distances, for any given a set of distances there is an infinite number of possible solutions, differing in the location and orientation of the point configuration.

A sufficient constraint to impose is the assumption that the means of all the variables are 0. That is, we assume $\sum_i x_{ik} = 0$ for all $k = 1, ..., p$. This means that $\sum_i b_{ij} = \sum_j b_{ij} = 0$. Now, by summing equation 3.14 first over i, then over j, and finally over both i and j, we obtain

$$\sum_i d_{ij}^2 = tr(\mathbf{B}) + nb_{jj}$$

$$\begin{aligned}\sum_j d_{ij}^2 &= tr(\mathbf{B}) + nb_{ii} \\ \sum_{ij} d_{ij}^2 &= 2ntr(\mathbf{B})\end{aligned} \tag{3.15}$$

where $tr(\mathbf{B})$ is the trace of the matrix $\mathbf{B}$. The third equation expresses $tr(\mathbf{B})$ in terms of the d_{ij}^2, the first and second express b_{jj} and b_{ii} in terms of d_{ij}^2 and $tr(\mathbf{B})$, and hence in terms of d_{ij}^2 alone. Plugging these into equation 3.14 expresses b_{ij} as a function of d_{ij}^2, yielding the required inversion.

This process is known as the *principal coordinates* method. It can be shown that the scores on the components calculated from a principal components analysis of a data matrix $\mathbf{X}$ (and hence a factorization of the matrix $\mathbf{X}^T$) are the same as the coordinates of the above scaling analysis.

Of course, if the matrix $\mathbf{B}$ does not arise not as a product $\mathbf{XX}^T$, but by some other route (such as simple subjective differences between pairs of objects), then there is no guarantee that all the eigenvalues will be non-negative. If the negative eigenvalues are small in absolute value, they can be ignored.

Classical multidimensional scaling into two dimensions finds the projection into two dimensions that is most accurate in the sense that it minimizes

$$\sum_i \sum_j (\delta_{ij} - d_{ij})^2, \tag{3.16}$$

where δ_{ij} is the observed distance between points i and j in the p-dimensional space and d_{ij} is the distance between the points representing these objects in the two-dimensional space. Expressed this way the process permits ready generalization. Given distances or dissimilarities, derived in one way or another, we can seek a distribution of points in a two-dimensional space that minimizes the sum of squared differences $\sum_i \sum_j (\delta_{ij} - d_{ij})^2$. Thus, we relax the restriction that the configuration must be found by projection. With this relaxation an exact algebraic solution will generally not be possible, so numerical methods must be used: we simply have a function of $2n$ parameters (the coordinates of the points in the two-dimensional space) that is to be minimized.

The score function $\sum_i \sum_j (\delta_{ij} - d_{ij})^2$, measuring how well the interpoint distances in the derived configuration match those originally provided, is invariant with respect to rotations and translations. However, it is not invariant to rescalings: if the δ_{ij} were multiplied by a constant, we would end up with the same solution, but a different value of $\sum_i \sum_j (\delta_{ij} - d_{ij})^2$. To permit different situations to be properly compared we divide $\sum_i \sum_j (\delta_{ij} - d_{ij})^2$

by $\sum_{i,j} d_{ij}^2$, yielding the standardized residual sum of squares. A common score function is the square root of this quantity, the *stress*. A variant on the stress is the *sstress*, defined as

$$\sqrt{\sum_i \sum_j (\delta_{ij}^2 - d_{ij}^2)^2 / \sum_i \sum_j d_{ij}^4}. \tag{3.17}$$

These measures effectively assume that the differences between the original dissimilarities and the distances in the two-dimensional configuration are due to random discrepancies and arbitrary distortions—that is, that $d_{ij} = \delta_{ij} + \epsilon_{ij}$. More sophisticated models can also be built. For example, we might assume that $d_{ij} = a + b\delta_{ij} + \epsilon_{ij}$. Now a two-stage procedure is necessary. Beginning with a proposed configuration, we regress the distances d_{ij} in the two-dimensional space on the given dissimilarities, yielding estimates for a and b. We then find new values of the d_{ij} that minimize the stress

$$\sqrt{\sum_i \sum_j (d_{ij} - a - b\delta_{ij})^2 / \sum_i \sum_j d_{ij}^2}, \tag{3.18}$$

and repeat this process until we achieve satisfactory convergence.

Multidimensional scaling methods such as the above, which attempt to model the dissimilarities as given, are called metric methods. Sometimes, however, a more general approach is required. For example, we may not be given the precise similarities, only their rank order (objects A and B are more similar than B and C, and so on); or we may not be prepared to assume that the relationship between d_{ij} and δ_{ij} has a particular form, just that some monotonic relationship exists. This requires a two-stage approach similar to that described in the preceding paragraph, but with a technique known as monotonic regression replacing simple linear regression, yielding non-metric multidimensional scaling. The term non-metric here indicates that the method seeks to preserve only ordinal relationships.

Multidimensional scaling is a powerful method for displaying data to reveal structure. However, as with the other graphical methods described in this chapter, if there are too many data points the structure becomes obscured. Moreover, since multidimensional scaling involves applying highly sophisticated transformations to the data (more so than a simple scatterplot or principal components analysis) there is a possibility that artifacts may be introduced. In particular, in some situations the dissimilarities between objects can be determined more accurately when the objects are similar than

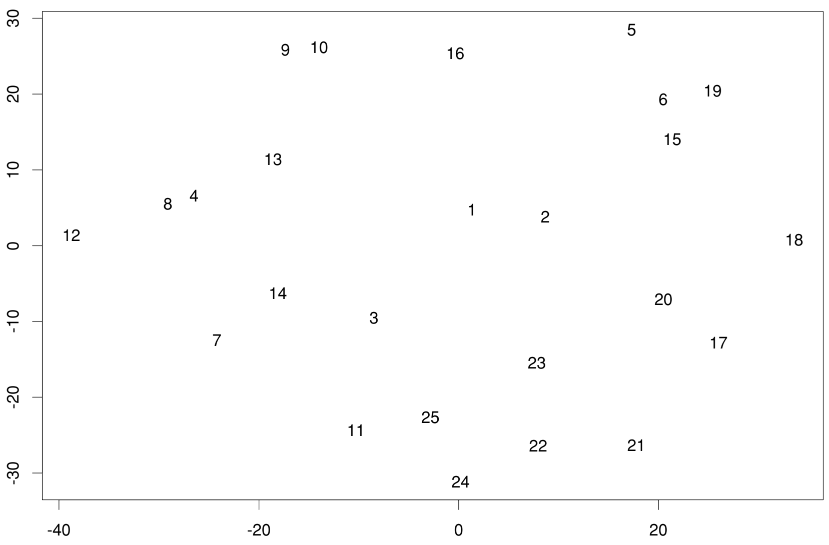

Figure 3.19 A multidimensional scaling plot of the village dialect similarities data.

when they are quite different. Consider the evolution of the style of a manufactured object. Those objects that are produced within a short time of each other will probably have much in common, while those separated by a greater time gap may have very little in common. The consequence will be an induced curvature in the multidimensional scaling plot, where we might have hoped to achieve a more or less straight line. This phenomenon is known as the *horseshoe effect*.

Figure 3.19 shows a plot produced using nonmetric scaling to minimize the sstress score function of equation 3.17. The data arose from a study of English dialects. Each pair of a group of 25 villages was rated according to the percentages of 60 items for which the villages used different words. The villages, and the counties in which they are located, are listed in table 3.1. The figure shows that villages from the same county (and hence that are relatively close geographically) tend to use the same words.

1	North Wheatley	Nottinghamshire
2	South Clifton	Nottinghamshire
3	Oxton	Nottinghamshire
4	Eastoft	Lincolnshire
5	Keelby	Lincolnshire
6	Wiloughton	Lincolnshire
7	Wragby	Lincolnshire
8	Old Bolingbroke	Lincolnshire
9	Fulbeck	Lincolnshire
10	Sutterton	Lincolnshire
11	Swinstead	Lincolnshire
12	Crowland	Lincolnshire
13	Harby	Leicestershire
14	Packington	Leicestershire
15	Goadby	Leicestershire
16	Ullesthorpe	Leicestershire
17	Empingham	Rutland
18	Warmington	Northamptonshire
19	Little Harrowden	Northamptonshire
20	Kislingbury	Northamptonshire
21	Sulgrave	Northamptonshire
22	Warboys	Huntingdonshire
23	Little Downham	Cambridgeshire
24	Tingewick	Buckinghamshire
25	Turvey	Bedfordshire

Table 3.1 Numerical codes, names, and counties for the 25 villages with dialect similarities displayed in figure 3.19.

Multidimensional scaling methods typically display the data points in a two-dimensional space. If the variables are also described in this space (provided the data are in vector form) the relationships between data points and variables may be clearly seen. Given the complicated nonlinear relationship between the space defined by the original variables and the space used to display the data, representing the original variables is a non-trivial task. Plots that display both data points and variables are known as *biplots*. The "bi" here signifies that there are two modes being displayed—the points and the variables—not that the display is two-dimensional. Indeed, three-dimensional biplots have also been developed. Forms of multidimensional scaling that involve nonlinear transformations produce nonlinear biplots. Biplots have even been produced for categorical data, and in this case the levels of the variables are represented by regions in the plot. Effective interpretation of multidimensional and biplot displays requires practice and experience.

3.8 Further Reading

Exploratory data analysis achieved an identity and respectability with the publication of John Tukey's book *Exploratory Data Analysis* (Tukey, 1977). Since then, as progress in computer technology facilitated rapid and straightforward production of accurate graphical displays, such methods have blossomed. Modern data visualization techniques can be very powerful ways of discovering structure. Books on graphical methods include those of Tufte (1983), Chambers et al. (1983), and Jacoby (1997). Wilkinson (1999) is a particularly interesting recent addition to the visualization literature, introducing a novel and general purpose language for analyzing and synthesizing a wide variety of data visualization techniques.

Interactive dynamic methods are emphasized by Asimov (1985), Becker, Cleveland, and Wilks (1987), Cleveland and McGill (1988), and Buja, Cook, and Swayne (1996). Books that describe smoothing approaches to displaying univariate distributions, as well as multivariate extensions, include those of Silverman (1986), Scott (1992), and Wand and Jones (1995). Carr et al. (1987) discuss scatterplot techniques for large data sets. Wegman (1990) discusses parallel coordinates. Categorical data is somewhat more difficult to visualize than quantitative real-valued data, and for this reason, visualization techniques for categorical data are not as widely developed or used. Still, Blasius and Greenacre (1998) provide a useful and broad review of recent developments in the visualization and exploratory data analysis of categorical data. Cook and Weisberg (1994) describe the use of graphical techniques for the task of regression modeling.

Card, MacKinlay, and Shneiderman (1999) contains a collection of papers on a variety of topics entitled "information visualization" and describe a number of techniques for displaying complex heterogeneous data sets in a useful manner. Keim and Kriegel (1994) describe a system specifically designed for database exploration.

Multidimensional scaling has become a large field in its own right. Books on this include those by Davidson (1983) and Cox and Cox (1994). Biplots are discussed in detail by Gower and Hand (1996).

The CPU data is from Ein-Dor and Feldmesser (1987), and is reproduced in Hand et al. (1994), dataset 325. The data on English dialects is from Morgan (1981) and is reproduced in Hand et al. (1994), dataset 145. The data on epileptic seizures is given in Thall and Vail (1990) and also in Hand et al. (1994). The mineral core data shown in the icon plot is described in Chernoff (1973).