# 5

# Pattern Recognition

## 5.1 INTRODUCTION

One of the first and most publicized success stories in chemometrics is pattern recognition. Much chemistry involves using data to determine patterns. For example, can infrared (IR) spectra be used to classify compounds into ketones and esters? Is there a pattern in the spectra allowing physical information to be related to chemical knowledge? There have been many spectacular successes of chemical pattern recognition. Can a spectrum be used in forensic science, for example to determine the cause of a fire? Can a chromatogram be used to decide on the origin of a wine and, if so, what main features in the chromatogram distinguish different wines? And is it possible to determine the time of year the vine was grown? Is it possible to use measurements of heavy metals to discover the source of pollution in a river?

There are several groups of methods for chemical pattern recognition.

### 5.1.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) consists mainly of the techniques of Principal Components Analysis (PCA) and Factor Analysis (FA). The statistical origins are in biology and psychology. Psychometricians have for many years had the need to translate numbers such as answers to questions in tests into relationships between individuals. How can verbal ability, numeracy and the ability to think in three dimensions be predicted from a test? Can different people be grouped by these abilities? And does this grouping reflect the backgrounds of the people taking the test? Are there differences according to educational background, age, sex, or even linguistic group?

In chemistry, we, too, need to ask similar questions, but the raw data are normally chromatographic or spectroscopic. An example involves chemical communication between animals: animals recognize each other more by smell than by sight, and different animals often lay scent trails, sometimes in their urine. The chromatogram of a urine sample may contain several hundred compounds, and it is often not obvious to the untrained observer which are most significant. Sometimes the most potent compounds are only present in small quantities. Yet animals can often detect through scent marking whether there is an in-heat member of the opposite sex looking for a mate, or whether there is a dangerous intruder

*Applied Chemometrics for Scientists* R. G. Brereton

entering their territory. EDA of chromatograms of urine samples can highlight differences in chromatograms of different social groups or different sexes, and give a simple visual idea as to the main relationships between these samples.

### 5.1.2 Unsupervised Pattern Recognition

A more formal method of treating samples is unsupervised pattern recognition, often called cluster analysis. Many methods have their origins in numerical taxonomy. Biologists measure features in different organisms, for example various body length parameters. Using a couple of dozen features, it is possible to see which species are most similar and draw a picture of these similarities, such as a dendrogram, phylogram or cladogram, in which more closely related species are closer to each other. The main branches can represent bigger divisions, such as subspecies, species, genera and families.

These principles can be directly applied to chemistry. It is possible to determine similarities in amino acid sequences in myoglobin in a variety of species, for example. The more similar the species, the closer the relationship: chemical similarity mirrors biological similarity. Sometimes the amount of information is huge, for example in large genomic or crystallographic databases such that cluster analysis is the only practicable way of searching for similarities.

Unsupervised pattern recognition differs from exploratory data analysis in that the aim of the methods are to detect similarities, whereas using EDA there is no particular prejudice as to whether or how many groups will be found. This chapter will introduce these approaches which will be expanded in the context of biology in Chapter 11.

### 5.1.3 Supervised Pattern Recognition

There are many reasons for supervised pattern recognition, mostly aimed at classification. Multivariate statisticians have developed a large number of discriminant functions, many of direct interest to chemists. A classic example is the detection of forgery in banknotes. Can physical measurements such as width and height of a series of banknotes be used to identify forgeries? Often one measurement is not enough, so several parameters are required before an adequate mathematical model is available.

Equivalently in chemistry, similar problems occur. Consider using a chemical method such as IR spectroscopy to determine whether a sample of brain tissue is cancerous or not. A method can be set up in which the spectra of two groups, cancerous and noncancerous tissues, are recorded: then some form of mathematical model is set up and finally the diagnosis of an unknown sample can be predicted.

Supervised techniques require a training set of known groupings to be available in advance, and try to answer a precise question as to the class of an unknown sample. It is, of course, always first necessary to establish whether chemical measurements are actually good enough to fit into the predetermined groups. However, spectroscopic or chromatographic methods for diagnosis are often much cheaper than expensive medical tests, and provide a valuable first diagnosis. In many cases chemical pattern recognition can be performed as a form of screening, with doubtful samples being subjected to more sophisticated tests. In areas such as industrial process control, where batches of compounds might be produced at hourly intervals, a simple on-line spectroscopic test together with chemical data analysis is often an essential first step to determine the possible acceptability of a batch. The methods in this chapter are expanded in Chapter 10 in the context of biology and medicine, together with several additional techniques.

## 5.2 PRINCIPAL COMPONENTS ANALYSIS

### 5.2.1 Basic Ideas

PCA is probably the most widespread multivariate statistical technique used in chemometrics, and because of the importance of multivariate measurements in chemistry, it is regarded by many as the technique that most significantly changed the chemist's view of data analysis.

There are numerous claims to the first use of PCA in the literature. Probably the most famous early paper was by Pearson in 1901 [1]. However, the fundamental ideas are based on approaches well known to physicists and mathematicians for much longer, namely those of eigen-analysis. In fact, some school mathematics syllabuses teach ideas about matrices which are relevant to modern chemistry. An early description of the method in physics was by Cauchy in 1829 [2]. It has been claimed that the earliest nonspecific reference to PCA in the chemical literature was in 1878 [3], although the author of the paper almost certainly did not realize the potential, and was dealing mainly with a simple problem of linear calibration. It is generally accepted that the revolution in the use of multivariate methods took place in psychometrics in the 1930s and 1940s of which Hotelling's paper is regarded as a classic [4]. An excellent more recent review of the area with a historical perspective, available in the chemical literature has been published by the Emeritus Professor of Psychology from the University of Washington, Paul Horst [5].

Psychometrics is well understood to most students of psychology and one important area involves relating answers in tests to underlying factors, for example, verbal and numerical ability as illustrated in Figure 5.1. PCA relates a data matrix consisting of these answers to a number of psychological 'factors'. In certain areas of statistics, ideas of factor analysis and PCA are intertwined, but in chemistry both approaches have a different meaning.

Natural scientists of all disciplines, from biologists, geologists and chemists have caught on to these approaches over the past few decades. Within the chemical community the first major applications of PCA were reported in the 1970s, and form the foundation of many modern chemometric methods.

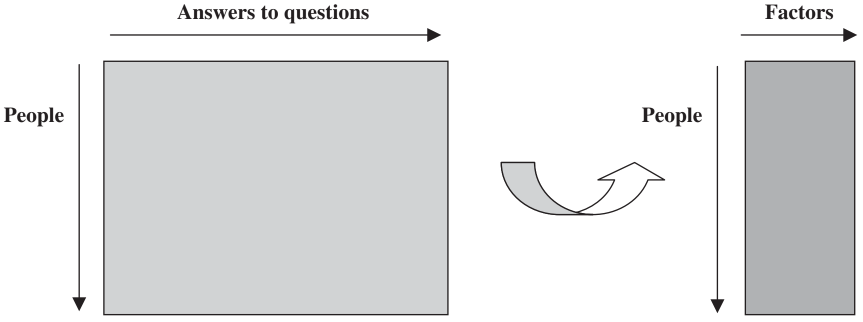


**Figure 5.1** Typical psychometric problems

A key idea is that most chemical measurements are inherently *multivariate*. This means that more than one measurement can be made on a single sample. An obvious example is spectroscopy: we can record a spectrum at hundreds of wavelengths on a single sample. Traditional approaches are *univariate* in which only one wavelength (or measurement) is used per sample, but this misses much information. Another common area is quantitative structure – property relationships, in which many physical measurements are available on a number of candidate compounds (bond lengths, dipole moments, bond angles, etc.); can we predict, *statistically*, the biological activity of a compound? Can this assist in pharmaceutical drug development? There are several pieces of information available. PCA is one of several multivariate methods that allows us to explore patterns in these data, similar to exploring patterns in psychometric data. Which compounds behave similarly? Which people belong to a similar group? How can this behaviour be predicted from available information?

As an example, Figure 5.2 represents a chromatogram in which a number of compounds are detected with different elution times, at the same time as their spectra [such as an ultraviolet (UV)/visible or mass spectrum] are recorded. Coupled chromatography, such as diode array high performance chromatography or liquid chromatography mass spectrometry, is increasingly common in modern laboratories, and represents a rich source of multivariate data. The chromatogram can be represented as a data matrix.

What do we want to find out about the data? How many compounds are in the chromatogram would be useful information. Partially overlapping peaks and minor impurities are the bugbears of modern chromatography. What are the spectra of these compounds? Figure 5.3 represents some coeluting peaks. Can we reliably determine their spectra? By looking at changes in spectral information across a coupled chromatogram, multivariate methods can be employed to resolve these peaks and so find their spectra. Finally, what are the quantities of each component? Some of this information could undoubtedly be obtained by better chromatography, but there is a limit, especially with modern trends to recording more and more data, more and more rapidly. In many cases the identities and amounts of unknowns may not be available in advance. PCA is one tool from multivariate statistics that can help sort out these data. Chapter 7 expands on some of these methods in the context of coupled chromatography, whereas the discussion in this chapter is restricted primarily to exploratory approaches.

The aims of PCA are to determine underlying information from multivariate raw data. There are two principal needs in chemistry.

The first is to interpret the Principal Components (PCs) often in a quantitative manner.

- The number of significant PCs. In the case of coupled chromatography this could relate to the number of compounds in a chromatogram, although there are many other requirements.

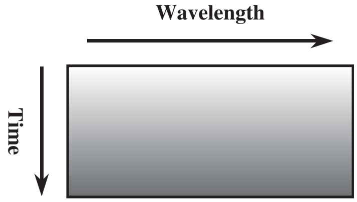


**Figure 5.2** Typical multivariate chromatographic information

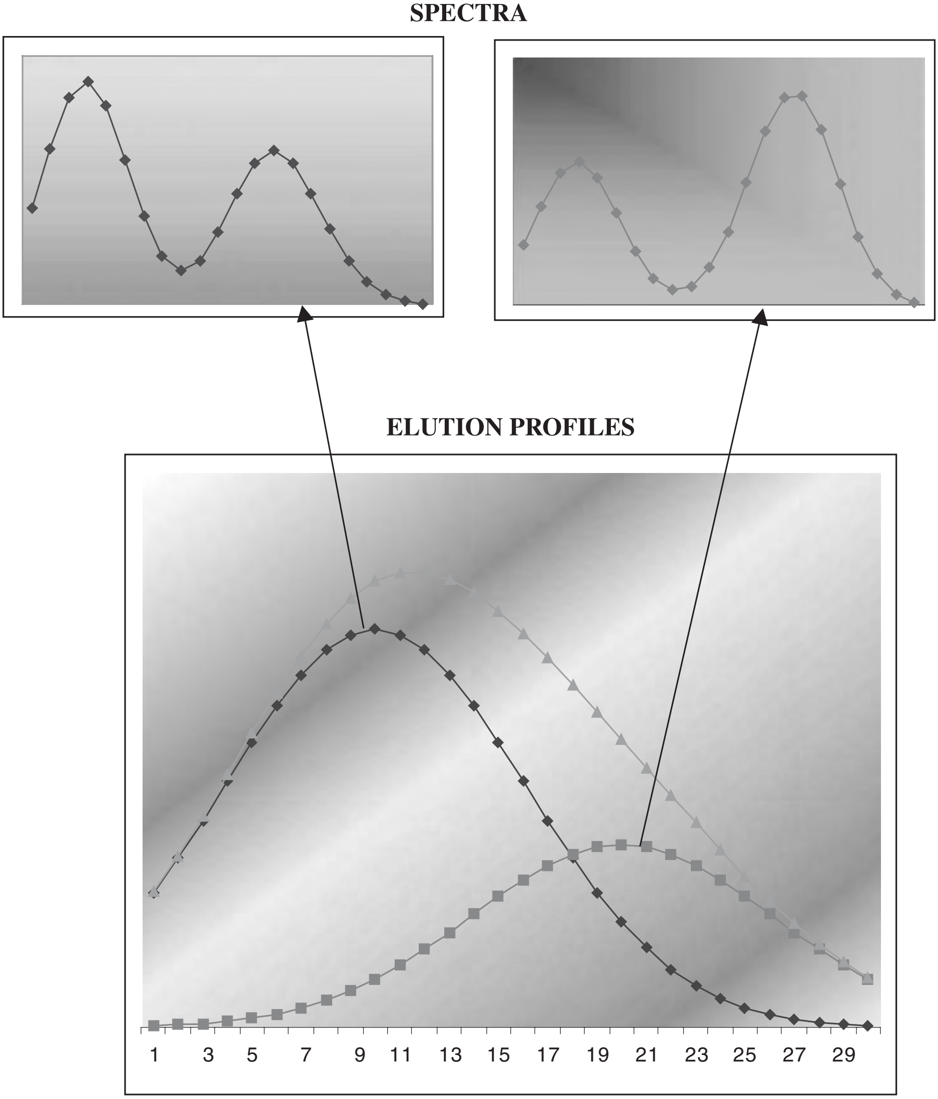


**Figure 5.3** Using a chromatogram of partially overlapping peaks to obtain spectra of individual compounds

- The characteristics of each PC, usually the *scores* relating to the objects or samples (in the example of coupled chromatography, the elution profiles) and the *loadings* relating to the variables or measurements (in coupled chromatography, the spectra).

In the next section we will look in more detail how this information is obtained. Often this information is then related to physically interpretable parameters of direct interest to the chemist, or is used to set up models for example to classify or group samples. The numerical information is interesting and can be used to make predictions often of the origin or nature of samples.

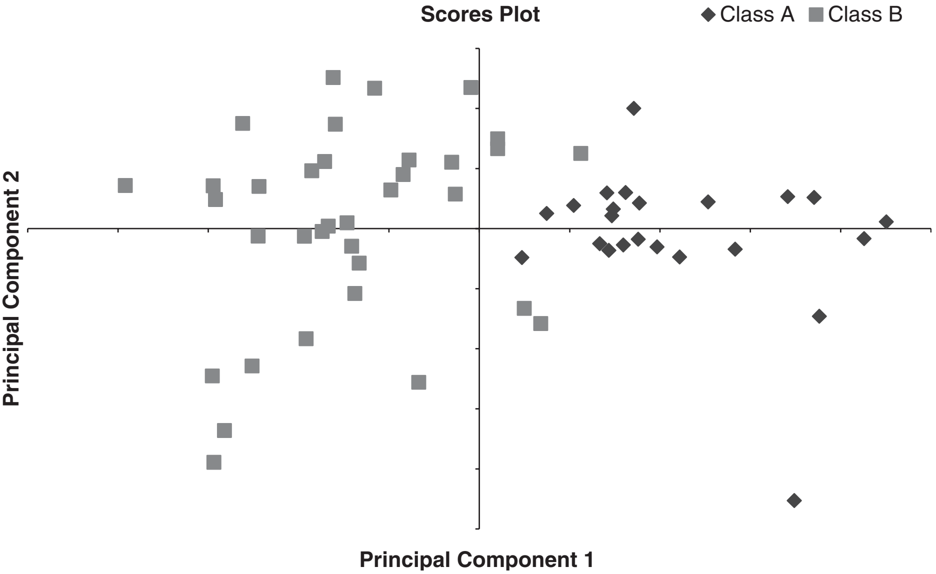


**Figure 5.4** Principal Component scores plot for elemental composition of pots in two different groups (A and B)

The second need is simply to obtain patterns. Figure 5.4 represents the scores plots (see Section 5.3) obtained after performing PCA on a standardized data matrix (see Section 5.5) whose rows (objects) correspond to archaeological finds of pottery and whose columns (variables) correspond to the amount of different elements found in these pots. The pots come from two different regions and the graph shows that these can be distinguished using their elemental composition. It also shows that there is a potential outlier (bottom right). The main aim is to simplify and explore the data rather than to make hard physical predictions, and graphical representation in itself is an important aim. Sometimes datasets are very large or difficult to interpret as tables of numbers and PC plots can simplify and show the main trends, and are easier to visualize than tables of numbers.

## 5.2.2 Method

In order to become familiar with the method it is important to appreciate the main ideas behind PCA. Although chemists have developed their own terminology, it is essential to recognize that similar principles occur throughout scientific data analysis, whether in physics, quantum mechanics or psychology.

As an illustration, we will use the case of coupled chromatography, such as diode array high performance liquid chromatography (HPLC). For a simple chromatogram, the underlying dataset can be described as a sum of responses for each significant compound in the data, which are characterized by (a) an elution profile and (b) a spectrum, plus noise or instrumental error. In matrix terms, this can be written as:

$$X = C.S + E \tag{5.1}$$

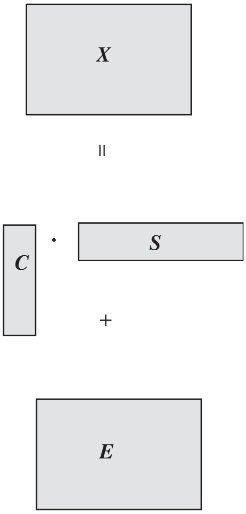


**Figure 5.5** Multivariate data such as occurs in diode array high performance liquid chromatography

where $\boldsymbol{X}$ is the original data matrix or coupled chromatogram, $\boldsymbol{C}$ is a matrix consisting of the elution profiles of each compound, $\boldsymbol{S}$ is a matrix consisting of the spectra of each compound and $\boldsymbol{E}$ is an error matrix.

This is illustrated in Figure 5.5. For those not expert in coupled chromatography, this example is of a data matrix each of whose dimensions correspond to variables related sequentially, one in time (chromatography) and one in frequency (spectroscopy), and each compound in the mixture has a characteristic time/frequency profile. The data matrix consists of a combination of signals from each constituent, mixed together, plus noise.

Consider a two way chromatogram recorded over 10 min at 1 s intervals (600 points in time), and over 200 nm at 2 nm intervals (100 spectroscopic points), containing three underlying compounds:

- $\boldsymbol{X}$ is a matrix of 600 rows and 100 columns;
- $\boldsymbol{C}$ is a matrix of 600 rows and 3 columns, each column corresponding to the elution profile of a single compound;
- $\boldsymbol{S}$ is a matrix of 3 rows and 100 columns, each row corresponding to the spectrum of a single column;
- $\boldsymbol{E}$ is a matrix of the same size as $\boldsymbol{X}$.

For more on matrices see Section 2.4.

If we observe $\boldsymbol{X}$, can we then predict $\boldsymbol{C}$ and $\boldsymbol{S}$? Many chemometricians use a ‘hat’ notation to indicate a *prediction* so it is also possible to write Equation (5.1) as:

$$\boldsymbol{X} \approx \hat{\boldsymbol{C}}.\hat{\boldsymbol{S}}$$

Ideally the predicted spectra and chromatographic elution profiles are close to the true ones, but it is important to realize that we can *never directly or perfectly* observe the underlying data. There will always be measurement error even in practical spectroscopy.

Chromatographic peaks may be partially overlapping or even embedded meaning that chemometric methods will help resolve the chromatogram into individual components.

One aim of chemometrics is to obtain these predictions after first treating the chromatogram as a multivariate data matrix, and then performing PCA. Each compound in the mixture can be considered a 'chemical' factor with its associated spectra and elution profile, which can be related to PCs, or 'abstract' factors, by a mathematical transformation.

A fundamental first step is to determine the number of significant factors or components in a matrix. In a series of mixture spectra or portion of a chromatogram, this should, ideally, correspond to the number of compounds under observation.

The *rank* of a matrix relates to the number of significant components in the data, in chemical terms to the number of compounds in a mixture. For example, if there are six components in a chromatogram the rank of the data matrix from the chromatogram should ideally equal 6. However, life is never so simple. What happens is that noise distorts this ideal picture, so even though there may be only six compounds, it may sometimes appear that the rank is 10 or more.

Normally the data matrix is first transformed into a number of PCs and the *size* of each component is measured. This is often called an *eigenvalue*: the earlier (or more significant) the components, the larger their size. It is possible to express eigenvalues as a percentage of the entire data matrix, by a simple technique.

- Determine the sum of squares of the entire data, $S_{total}$.
- For each PC determine its own sum of squares (which is usually equal to the sum of squares of the scores vector as discussed below), $S_k$ for the $k$th component. This is a common definition the *eigenvalue* although there is other terminology in the literature.
- Determine the percentage contribution of each PC to the data matrix $(100S_k/S_{total})$. Sometimes the *cumulative* contribution is calculated.

Note that there are several definitions of eigenvalues, and many chemometricians have adopted a rather loose definition, that of the sum of squares of the scores, this differs from the original formal definitions in the mathematical literature. However, there is no universally agreed set of chemometrics definitions, every group or school of thought has their own views.

One simple way of determining the number of significant components is simply by the looking at the size of each successive eigenvalue. Table 5.1 illustrates this. The total sum of squares for the entire dataset happens to be 670, so since the first three PCs account for around 95 % of the data (or 639/670), so it is a fair bet that there are only three components in the data. There are, of course, more elaborate approaches to estimating the number of significant components, to be discussed in more detail in Section 5.10.

The number of nonzero components will never be more than the smaller of the number of rows and columns in the original data matrix $\boldsymbol{X}$. Hence if this matrix consists of 600 rows

**Table 5.1** Illustration of size of eigenvalues in Principal Component Analysis

| Total | | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|---|
| 670 | Eigenvalue | 300 | 230 | 109 | 20 | 8 |
| | % | 44.78 | 34.34 | 16.27 | 2.99 | 1.19 |
| | Cumulative % | 44.78 | 79.11 | 95.37 | 98.36 | 99.55 |

(e.g. chromatographic elution times) and 100 columns (e.g. spectral wavelengths), there will never be more than 100 nonzero eigenvalues, but, hopefully, the true answer will be very much smaller, reflecting the number of compounds in the chromatogram.

PCA results in an abstract mathematical transformation of the original data matrix, which, for the case of a coupled chromatogram, may take the form:

$$\mathbf{X} \approx \hat{C}.\hat{S} = \boldsymbol{T}.\boldsymbol{P}$$

where $\boldsymbol{T}$ are called the scores, and ideally have the same dimensions as $\boldsymbol{C}$, and $\boldsymbol{P}$ the loadings ideally having the same dimensions as $\boldsymbol{S}$. A big interest is how to relate the abstract factors (scores and loadings) to the chemical factors, and Sections 7.8, 7.9, 8.1, 8.3 and 8.4 will introduce a number of techniques in various applications. Note that the product and number of abstract factors should ideally equal the product and number of chemical factors. Purely numerical techniques can be use to obtain the abstract factors.

Each scores matrix consists of a series of column vectors, and each loadings matrix a series of row vectors, the number of such vectors equalling the rank of the original data matrix, so if the rank of the original data matrix is 8 and spectra are recorded at 100 wavelengths, the loadings matrix consists of 8 row vectors 100 data points in length. Many authors denote these vectors by $\boldsymbol{t}_a$ and $\boldsymbol{p}_a$ where $a$ is the number of the PC (1, 2, 3 up to the matrix rank). The scores matrices $\boldsymbol{T}$ and $\boldsymbol{P}$ are composed of several such vectors, one for each PC. If we want to interpret these vectors, these can be related to the true spectra and elution profiles by the transformations discussed in Chapter 7 in greater detail. In many areas of chemical pattern recognition, of course, the scores and loadings are an end in themselves and no further transformation to physical factors is required.

Scores and loadings have important properties, the main one being called *orthogonality* (introduced also in Sections 2.6, 2.8 and 2.9). This is often expressed in a number of ways:

- The product between any two loadings or scores vectors is 0.
- The correlation coefficient between any two loadings or scores vectors is 0 providing they are centred.

The original variables (e.g. 100 wavelengths) are reduced to a number of significant PCs (e.g. 3 or 4) each of which is orthogonal to each other. In practice, PCA has acted as a form

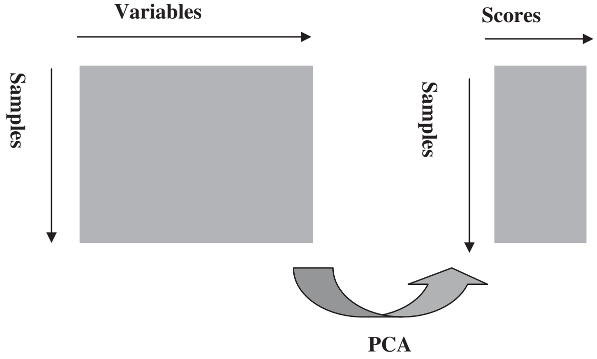


**Figure 5.6** Overview of data simplification by PCA in chemistry

of variable reduction, reducing the large original dataset (e.g. recorded at 100 wavelengths) to a much smaller more manageable dataset (e.g. consisting of three PCs) which can be interpreted more easily, as illustrated in Figure 5.6. The loadings represent the means to this end.

The loadings vectors for each component are also generally *normalized*, meaning that their sum of squares equals one, whereas the sum of squares of the scores vectors are often equal to the corresponding eigenvalue. There are of course several different PC algorithms and not everyone uses the same scaling methods, however orthogonality is always obeyed.

## 5.3 GRAPHICAL REPRESENTATION OF SCORES AND LOADINGS

Many revolutions in chemistry relate to the graphical presentation of information. For example, fundamental to the modern chemist's way of thinking is the ability to draw structures on paper in a convenient and meaningful manner. Years of debate preceded the general acceptance of the Kekulé structure for benzene: today's organic chemist can write down and understand complex structures of natural products without the need to plough through pages of numbers of orbital densities and bond lengths. Yet, underlying these representations are quantum mechanical probabilities, so the ability to convert from numbers to a simple diagram has allowed a large community to think clearly about chemical reactions.

So with statistical data, and modern computers, it is easy to convert from numbers to graphs. Many modern multivariate statisticians think geometrically as much as numerically, and concepts such as PCs are often treated as much as objects in an imaginary space than mathematical entities. The algebra of multidimensional space is the same as that of multivariate statistics. Older texts, of course, were written before the days of modern computing, so the ability to produce graphs was more limited. However, now it is possible to obtain a large number of graphs rapidly using simple software. There are many ways of visualizing PCs and this section will illustrate some of the most common.

We will introduce two case studies.

### 5.3.1 Case Study 1

The first relates to the resolution of two compounds (I=2-hydroxypyridine and II=3-hydroxypyridine) by diode array HPLC. The chromatogram (summed over all wavelengths) is illustrated in Figure 5.7. More details are given in Dunkerley *et al.* [6]. The aim is to try to obtain the individual profiles of each compound in the chromatogram, and also their spectra. Remember that a second, spectroscopic, dimension has been recorded also. The raw data are a matrix whose *columns* relate to wavelengths and whose *rows* relate to elution time. Further discussions of data of this nature are included in Chapter 7.

### 5.3.2 Case Study 2

In this case five physical constants are measured for 27 different elements, namely melting point, boiling point, density, oxidation number and electronegativity, to form a $27 \times 5$ matrix, whose *rows* correspond to elements and whose *columns* to constants. The data are presented in Table 5.2. The aims are to see which elements group together and also which

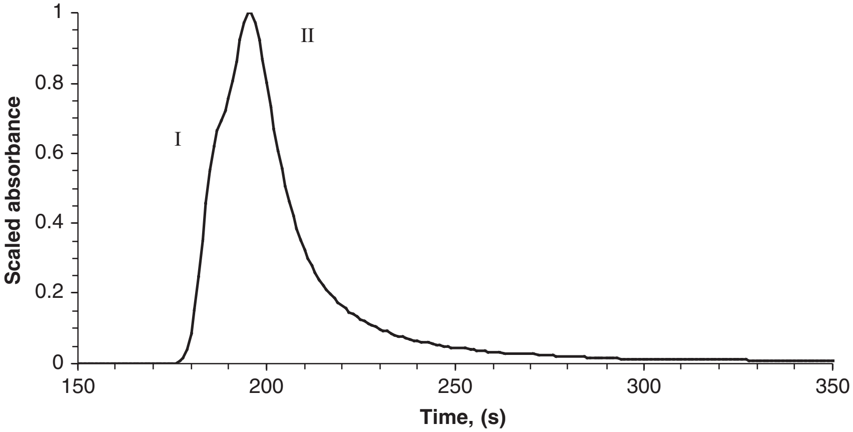


**Figure 5.7** Chromatographic profile for case study 1

**Table 5.2** Case study 2

| Element | Group | Melting point (K) | Boiling) point K | Density (mg/cm$^3$) | Oxidation number | Electronegativity |
|---|---|---|---|---|---|---|
| Li | 1 | 453.69 | 1615 | 534 | 1 | 0.98 |
| Na | 1 | 371 | 1156 | 970 | 1 | 0.93 |
| K | 1 | 336.5 | 1032 | 860 | 1 | 0.82 |
| Rb | 1 | 312.5 | 961 | 1530 | 1 | 0.82 |
| Cs | 1 | 301.6 | 944 | 1870 | 1 | 0.79 |
| Be | 2 | 1550 | 3243 | 1800 | 2 | 1.57 |
| Mg | 2 | 924 | 1380 | 1741 | 2 | 1.31 |
| Ca | 2 | 1120 | 1760 | 1540 | 2 | 1 |
| Sr | 2 | 1042 | 1657 | 2600 | 2 | 0.95 |
| F | 3 | 53.5 | 85 | 1.7 | −1 | 3.98 |
| Cl | 3 | 172.1 | 238.5 | 3.2 | −1 | 3.16 |
| Br | 3 | 265.9 | 331.9 | 3100 | −1 | 2.96 |
| I | 3 | 386.6 | 457.4 | 4940 | −1 | 2.66 |
| He | 4 | 0.9 | 4.2 | 0.2 | 0 | 0 |
| Ne | 4 | 24.5 | 27.2 | 0.8 | 0 | 0 |
| Ar | 4 | 83.7 | 87.4 | 1.7 | 0 | 0 |
| Kr | 4 | 116.5 | 120.8 | 3.5 | 0 | 0 |
| Xe | 4 | 161.2 | 166 | 5.5 | 0 | 0 |
| Zn | 5 | 692.6 | 1180 | 7140 | 2 | 1.6 |
| Co | 5 | 1765 | 3170 | 8900 | 3 | 1.8 |
| Cu | 5 | 1356 | 2868 | 8930 | 2 | 1.9 |
| Fe | 5 | 1808 | 3300 | 7870 | 2 | 1.8 |
| Mn | 5 | 1517 | 2370 | 7440 | 2 | 1.5 |
| Ni | 5 | 1726 | 3005 | 8900 | 2 | 1.8 |
| Bi | 6 | 544.4 | 1837 | 9780 | 3 | 2.02 |
| Pb | 6 | 600.61 | 2022 | 11340 | 2 | 1.8 |
| Tl | 6 | 577 | 1746 | 11850 | 3 | 1.62 |

physical constants are responsible for this grouping. Because all the physical constants are on different scales, it is first necessary to standardize (Section 5.5) the data prior to performing PCA.

### 5.3.3 Scores Plots

One of the simplest plots is that of the scores (Section 5.2.2) of one PC against the other. Below we will look only at the first two PCs, for simplicity.

Figure 5.8 illustrates the PC plot for case study 1. The horizontal axis is the scores for the first PC, and the vertical axis for the second PC. This 'picture' can be interpreted as follows:

- The linear regions of the graph represent regions of the chromatogram where there are pure compounds, I and II.
- The curve portion represents a region of coelution.
- The closer to the origin, the lower the intensity.

Hence the PC plot suggests that the region between 187 and 198 s (approximately) is one of coelution. The reason why this method works is that the spectrum over the chromatogram changes with elution time. During coelution the spectral appearance changes most, and PCA uses this information.

How can these graphs help?

- The pure regions can inform us of the spectra of the pure compounds.
- The shape of the PC plot informs us of the amount of overlap and quality of chromatography.
- The number of bends in a PC plot can provide information about the number of different compounds in a complex multipeak cluster.

Figure 5.9 illustrates the scores plot for case study 2. We are not in this case trying to determine specific factors or pull out spectra, but rather to determine where the main

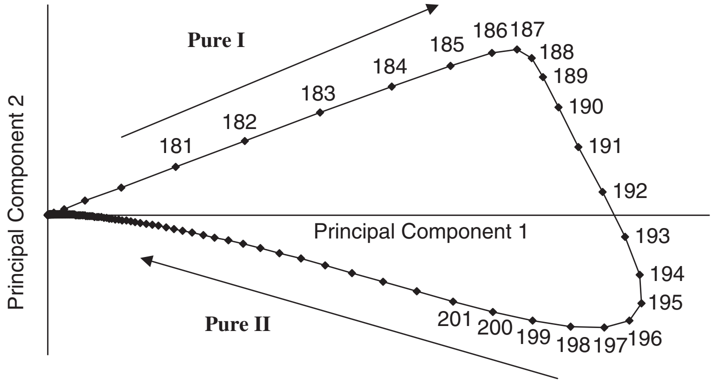


**Figure 5.8** Scores plot for the chromatographic data of case study 1: the numbers refer to elution times (in s)

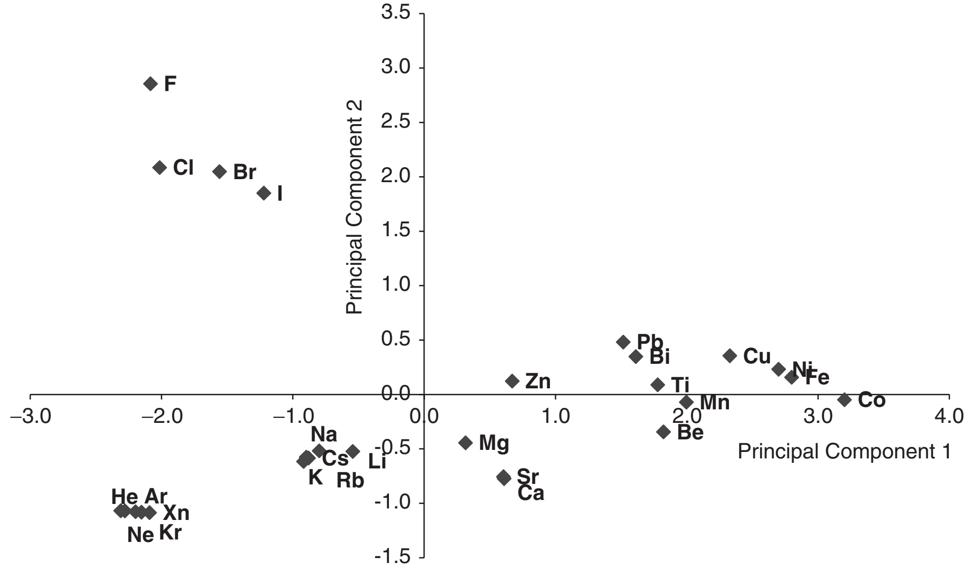


**Figure 5.9** Scores plot of the first two PCs for case study 2

groupings are. We can see that the halides cluster together at the top left, and the inert gases in the bottom left. The metals are primarily clustered according to their groups in the periodic table. This suggests that there are definitive patterns in the data which can be summarized graphically using PCs. Many more statistically based chemometricians often do not particularly like these sort of graphical representations which cannot very easily be related to physical factors, but they are nevertheless an extremely common way of summarizing complex data, which we will use in several contexts later in this book.

### 5.3.4 Loadings Plots

It is not, however, only the scores that are of interest but sometimes the loadings. Exactly the same principles apply in that the value of the loadings at one PC can be plotted against that at the other PC. The result for case study 1 is shown in Figure 5.10. This figure looks quite complicated, this is because both spectra overlap and absorb at similar wavelengths, and should be compared with the scores plot of Figure 5.8, the pure compounds lie in the same directions. The pure spectra are presented in Figure 5.11. Now we can understand these graphs a little more:

- High wavelengths, above 325 nm belong mainly to compound I and are so along the direction of pure I.
- 246 nm is a wavelength where the ratio of absorbance of compound I to II is a maximum, whereas for 301 nm, the reverse is true.

More interpretation is possible, but it can easily be seen that the loadings plots provide detailed information about which wavelengths are most associated with which compound.

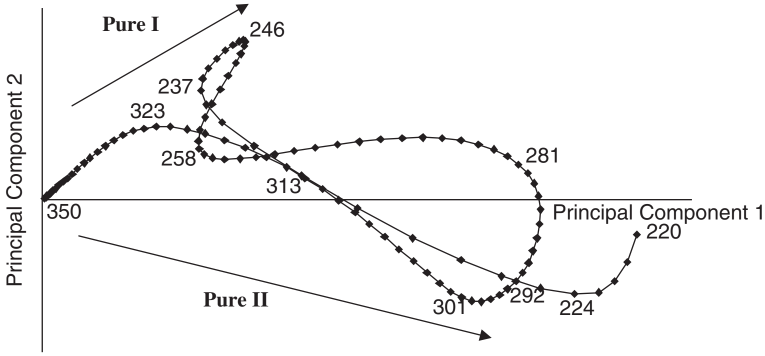


**Figure 5.10** Loadings plot for case study 1 (compare with Figure 5.8): wavelengths (in nm) are indicated

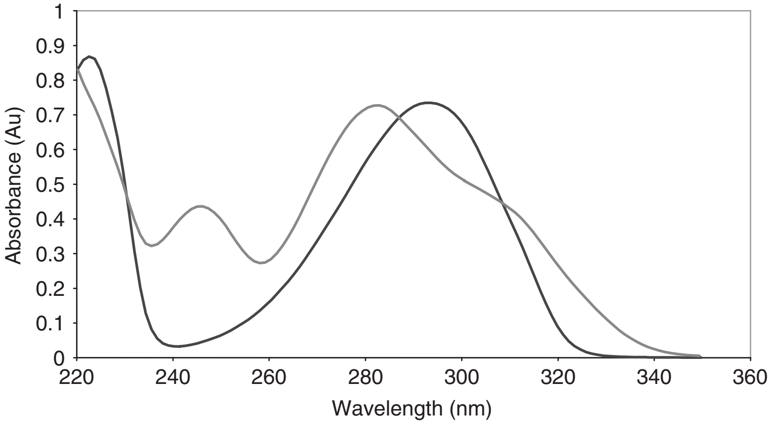


**Figure 5.11** Spectra of the two pure compounds in case study 1

The loadings plot for case study 2 is illustrated in Figure 5.12. We can see several features. The first is that melting point, boiling point and density seem closely clustered. This suggests that these three parameters measure something very similar, which is unsurprising, as the higher the melting point, the higher the boiling point in most cases. The density (at room temperature) should have some relationship to melting/boiling point also particularly whether an element is in gaseous, liquid or solid state. We can see that electronegativity is in quite a different place, almost at right angles to the density/boiling/melting point axis, and this suggests it follows entirely different trends.

We can see also that there are relationships between scores and loadings in case study 2. The more dense, high melting point, elements are on the right in the scores plot, and the more electronegative elements at the top end, so we can look at which variable influences which object by looking at both plots together, as discussed.

Loadings plots can be used to answer a lot of questions about the data, and are a very flexible facility available in almost all chemometrics software.

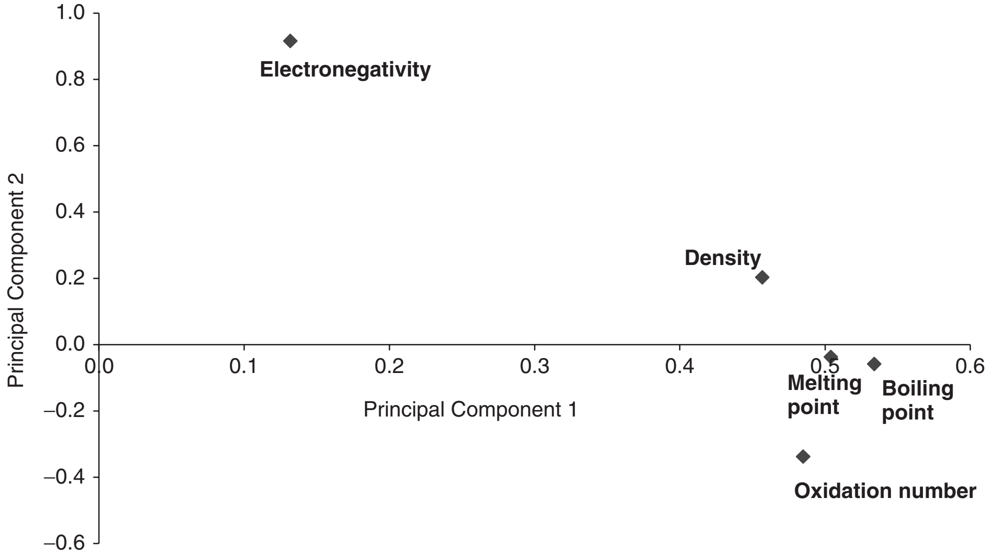


**Figure 5.12** Loadings plot for case study 2

### 5.3.5 Extensions

In many cases, more than two significant PCs are necessary to adequately characterize the data, but the same concepts can be employed, except there are many more possible graphs. For example, if four significant components are calculated, we can produce six possible graphs, of each possible combination of PCs, for example, PC 4 versus 2, or PC 1 versus 3 and so on. Each graph could reveal interesting trends. It is also possible to produce three-dimensional PC plots, whose axes consist of the scores or loadings of three PCs (normally the first three) and so visualize relationships between and clusters of variables in three-dimensional space.

## 5.4 COMPARING MULTIVARIATE PATTERNS

PC plots are often introduced only by reference to the independent loadings or scores plot of a single dataset. Yet there are common patterns within different graphs. Consider taking measurements of the concentration of a mineral in a geochemical deposit. This information could be presented as a table of sampling sites and observed concentrations. A much more informative approach would be to produce a picture in which physical location and mineral concentration are superimposed, such as a coloured map, each different colour corresponding to a concentration range of the mineral. Two pieces of information are connected, namely geography and concentration. So in many areas of multivariate analysis, one aim may be to connect the samples (e.g. geographical location/sampling site) represented by scores, to the variables (e.g. chemical measurements), represented by loadings. Graphically this requires the superimposition of two types of information.

A biplot involves superimposition of a scores and a loadings plot, with the variables and samples represented on the same diagram. It is not necessary to restrict biplots to two PCs, but, of course, when more than three are used, graphical representation becomes difficult, and numerical measures of fit between the scores and loadings are often employed, using statistical software, to determine which variables are best associated with which samples.

A different need is to be able to compare different types of measurements using procrustes analysis. Procrustes was a Greek god who kept a house by the side of the road where he offered hospitality to passing strangers, who were invited in for a meal and a night's rest in his very special bed which Procrustes described as having the unique property that its length exactly matched whomsoever lay down upon it. What he did not say was the method by which this 'one-size-fits-all' was achieved: as soon as the guest lay down Procrustes went to work upon them, stretching them if they were too short for the bed or chopping off their legs if they were too long.

Similarly, procrustes analysis in chemistry involves comparing two diagrams, such as two PC scores plots originating from different types of measurement. One such plot is the reference and the other is manipulated to resemble the reference plot as closely as possible. This manipulation is done mathematically, involving rotating, stretching and sometimes translating the second scores plot, until the two graphs are as similar as possible.

It is not necessary to restrict data from each type of measurement technique to two PCs, indeed in many practical cases four or five PCs are employed. Computer software is available to compare scores plots and provide a numeric indicator of the closeness of the fit. Procrustes analysis can be used to answer quite sophisticated questions. For example, in sensory research, are the results of a taste panel comparable with chemical measurements? If so, can the rather expensive and time-consuming taste panel be replaced by chromatography? A second use of procrustes analysis is to reduce the number of tests: an example being of clinical trials. Sometimes 50 or more bacteriological tests are performed but can these be reduced to 10 or less? A way to check this is by performing PCA on the results of all 50 tests, and compare the scores plot when using a subset of 10 tests. If the two scores plots provide comparable information, the 10 selected tests are just as good as the full set of tests. This can be of significant economic benefit. A final and important application is when several analytical techniques are employed to study a process, an example being the study of a reaction by IR, UV/visible and Raman spectroscopy, does each type of spectrum give similar answers? A consensus can be obtained using procrustes analysis.

## 5.5 PREPROCESSING

Many users of chemometric software simply accept without much insight the results of PCA: yet interpretation depends critically on how the original data have been handled. Data preprocessing or scaling can have a significant influence on the outcome, and also relate to the chemical or physical aim of the analysis. In fact in many modern areas such as metabolomics (Section 10.10), it is primarily the method for preprocessing that is difficult and influences the end result.

As an example, consider a data matrix consisting of 10 rows (labelled from 1 to 10) and eight columns (labelled from A to H), illustrated in Table 5.3(a). This could represent a portion of a two way diode array HPLC data matrix, whose elution profile in given in Figure 5.13, but similar principles apply to other multivariate data matrices, although the chromatographic example is especially useful for illustrative purposes as both dimensions

**Table 5.3** Simple example for Section 5.5. (a) Raw data; (b) column mean centred data;(c) Column standardized data (d) row scaled

(a)

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.318 | 0.413 | 0.335 | 0.196 | 0.161 | 0.237 | 0.290 | 0.226 |
| 2 | 0.527 | 0.689 | 0.569 | 0.346 | 0.283 | 0.400 | 0.485 | 0.379 |
| 3 | 0.718 | 0.951 | 0.811 | 0.521 | 0.426 | 0.566 | 0.671 | 0.526 |
| 4 | 0.805 | 1.091 | 0.982 | 0.687 | 0.559 | 0.676 | 0.775 | 0.611 |
| 5 | 0.747 | 1.054 | 1.030 | 0.804 | 0.652 | 0.695 | 0.756 | 0.601 |
| 6 | 0.579 | 0.871 | 0.954 | 0.841 | 0.680 | 0.627 | 0.633 | 0.511 |
| 7 | 0.380 | 0.628 | 0.789 | 0.782 | 0.631 | 0.505 | 0.465 | 0.383 |
| 8 | 0.214 | 0.402 | 0.583 | 0.635 | 0.510 | 0.363 | 0.305 | 0.256 |
| 9 | 0.106 | 0.230 | 0.378 | 0.440 | 0.354 | 0.231 | 0.178 | 0.153 |
| 10 | 0.047 | 0.117 | 0.212 | 0.257 | 0.206 | 0.128 | 0.092 | 0.080 |

(b)

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | −0.126 | −0.231 | −0.330 | −0.355 | −0.285 | −0.206 | −0.175 | −0.146 |
| 2 | 0.083 | 0.045 | −0.095 | −0.205 | −0.163 | −0.042 | 0.020 | 0.006 |
| 3 | 0.273 | 0.306 | 0.146 | −0.030 | −0.020 | 0.123 | 0.206 | 0.153 |
| 4 | 0.360 | 0.446 | 0.318 | 0.136 | 0.113 | 0.233 | 0.310 | 0.238 |
| 5 | 0.303 | 0.409 | 0.366 | 0.253 | 0.206 | 0.252 | 0.291 | 0.229 |
| 6 | 0.135 | 0.226 | 0.290 | 0.291 | 0.234 | 0.185 | 0.168 | 0.139 |
| 7 | −0.064 | −0.017 | 0.125 | 0.231 | 0.184 | 0.062 | 0.000 | 0.010 |
| 8 | −0.230 | −0.243 | −0.081 | 0.084 | 0.064 | −0.079 | −0.161 | −0.117 |
| 9 | −0.338 | −0.414 | −0.286 | −0.111 | −0.093 | −0.212 | −0.287 | −0.220 |
| 10 | −0.397 | −0.528 | −0.452 | −0.294 | −0.240 | −0.315 | −0.373 | −0.292 |

(c)

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | −0.487 | −0.705 | −1.191 | −1.595 | −1.589 | −1.078 | −0.760 | −0.818 |
| 2 | 0.322 | 0.136 | −0.344 | −0.923 | −0.909 | −0.222 | 0.087 | 0.035 |
| 3 | 1.059 | 0.933 | 0.529 | −0.133 | −0.113 | 0.642 | 0.896 | 0.856 |
| 4 | 1.396 | 1.361 | 1.147 | 0.611 | 0.629 | 1.218 | 1.347 | 1.330 |
| 5 | 1.174 | 1.248 | 1.321 | 1.136 | 1.146 | 1.318 | 1.263 | 1.277 |
| 6 | 0.524 | 0.690 | 1.046 | 1.306 | 1.303 | 0.966 | 0.731 | 0.774 |
| 7 | −0.249 | −0.051 | 0.452 | 1.040 | 1.026 | 0.326 | 0.001 | 0.057 |
| 8 | −0.890 | −0.740 | −0.294 | 0.376 | 0.357 | −0.415 | −0.698 | −0.652 |
| 9 | −1.309 | −1.263 | −1.033 | −0.497 | −0.516 | −1.107 | −1.247 | −1.228 |
| 10 | −1.539 | −1.608 | −1.635 | −1.321 | −1.335 | −1.649 | −1.620 | −1.631 |

(d)

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.146 | 0.190 | 0.154 | 0.090 | 0.074 | 0.109 | 0.133 | 0.104 |
| 2 | 0.143 | 0.187 | 0.155 | 0.094 | 0.077 | 0.109 | 0.132 | 0.103 |
| 3 | 0.138 | 0.183 | 0.156 | 0.100 | 0.082 | 0.109 | 0.129 | 0.101 |
| 4 | 0.130 | 0.176 | 0.159 | 0.111 | 0.090 | 0.109 | 0.125 | 0.099 |
| 5 | 0.118 | 0.166 | 0.162 | 0.127 | 0.103 | 0.110 | 0.119 | 0.095 |
| 6 | 0.102 | 0.153 | 0.167 | 0.148 | 0.119 | 0.110 | 0.111 | 0.090 |
| 7 | 0.083 | 0.138 | 0.173 | 0.171 | 0.138 | 0.111 | 0.102 | 0.084 |
| 8 | 0.066 | 0.123 | 0.178 | 0.194 | 0.156 | 0.111 | 0.093 | 0.078 |
| 9 | 0.051 | 0.111 | 0.183 | 0.213 | 0.171 | 0.112 | 0.086 | 0.074 |
| 10 | 0.041 | 0.103 | 0.186 | 0.226 | 0.181 | 0.112 | 0.081 | 0.071 |

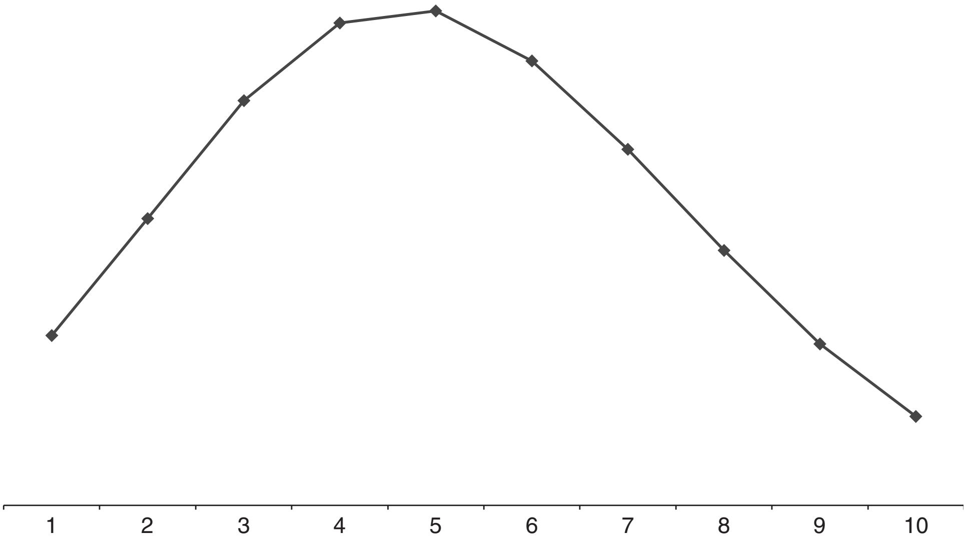


**Figure 5.13** Summed profile formed from data in Table 5.3

have an interpretable sequential meaning (which is not necessarily so in most other types of data analysis) and provides a situation that illustrates several different consequences of data preprocessing.

The resultant PC scores and loadings plots are given in Figure 5.14 for the first two PCs. Several deductions are possible, for example:

- There are probably two main compounds, one which has a region of purity between points 1 and 3, and the other between points 8 and 10.
- Measurements (e.g. spectral wavelengths) A, B, G and H correspond mainly to the first (e.g. fastest eluting) chemical component, whereas measurements D and E to the second chemical component.

PCA has been performed directly on the raw data, something statisticians in other disciplines very rarely do. It is important to be very careful when using packages that have been designed primarily by statisticians, on chemical data. Traditionally, what is mainly interesting to statisticians is deviation around a mean, for example, how do the mean characteristics of a forged banknote vary? What is an ‘average’ banknote? In chemistry we are often (but by no means exclusively) interested in deviation above a baseline, such as in spectroscopy.

It is, though, possible to mean centre the columns. The result of this is presented in Table 5.3(b). Notice now that the sum of each column is now 0. Almost all traditional statistical packages perform this operation prior to PCA whether desired or not. The PC plots are presented in Figure 5.15. The most obvious difference is that the scores plot is now centred around the origin. However, the relative positions of the points in both graphs change slightly, the biggest effect being on the loadings. In practice, mean centring can have quite a large influence in some cases, for example if there are baseline problems or only a small region of the data is recorded.

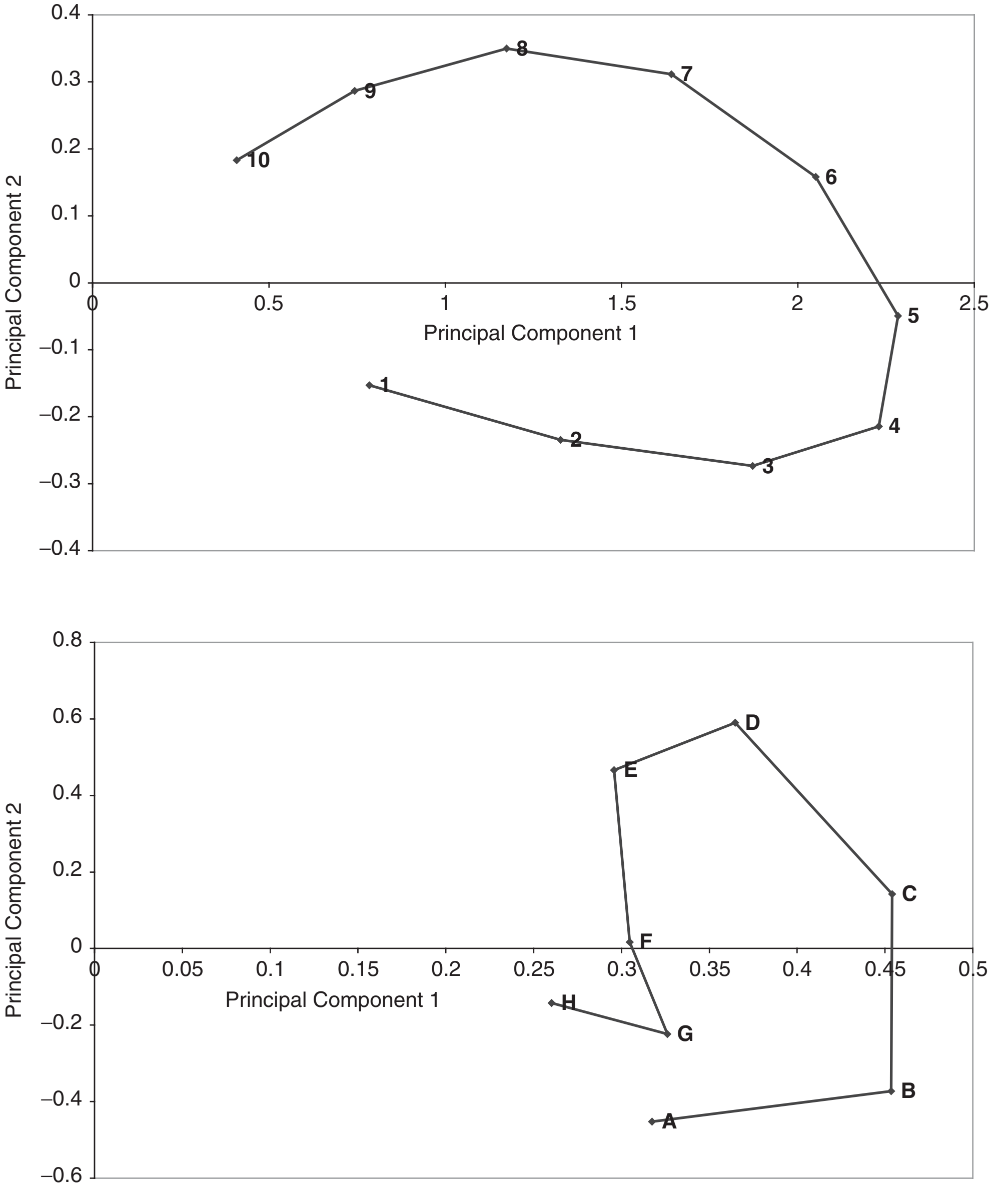


**Figure 5.14** Scores and loadings plots from the raw data in Table 5.3

Note that it is also possible to mean centre the rows, and also double mean centre data both simultaneously down the columns and along the rows (see Section 7.2.1 for more details), however, this is rarely done in chemometrics.

Standardization is another common method for data scaling and first requires mean centring: in addition, each variable is also divided by its standard deviation, Table 5.3(c) for our example. This procedure has been discussed in the context of the normal distribution in Section 3.4. Note an interesting feature that the sum of squares of each column equals

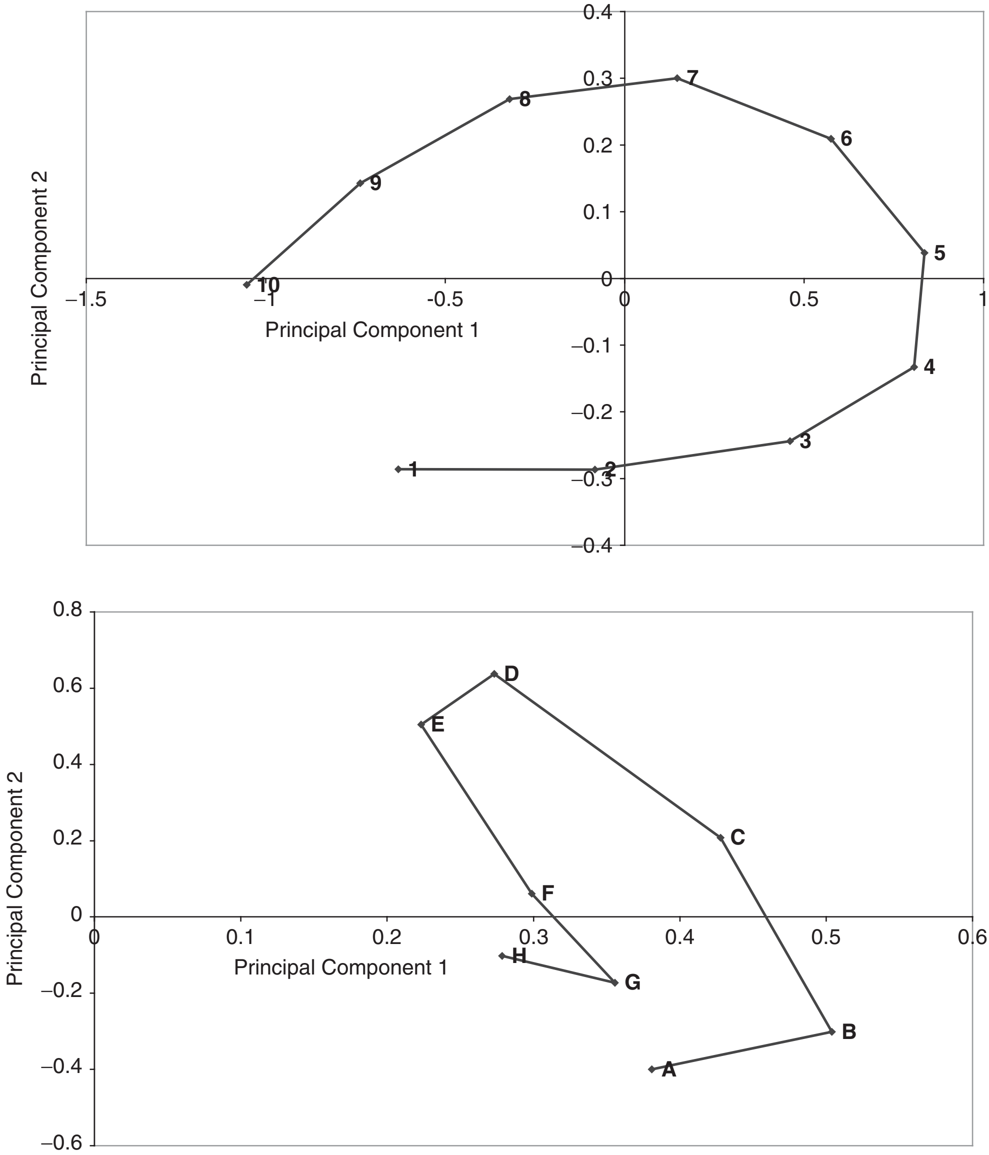
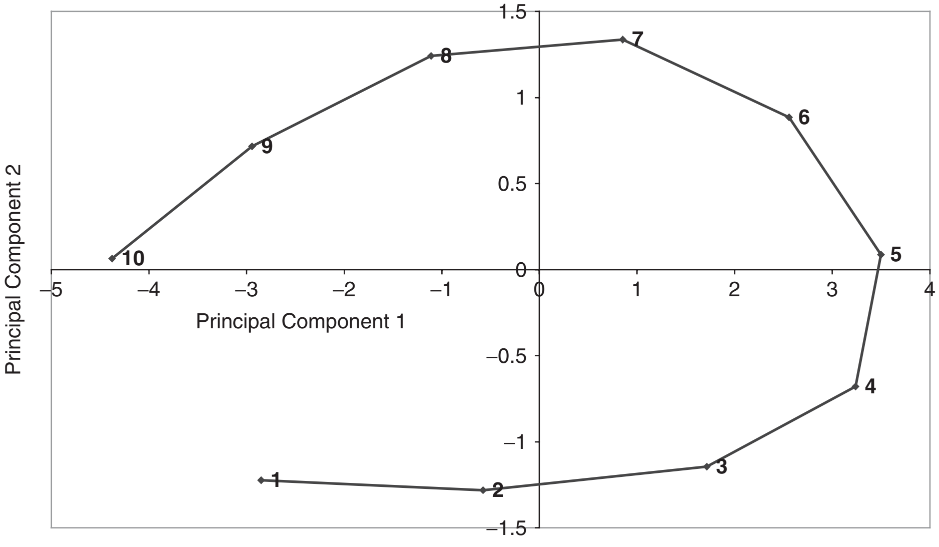


**Figure 5.15** Scores and loadings plots corresponding to Figure 5.14 but for column mean centred data

10 in this example (which is the number of objects in the dataset): the population standard deviation (Section 3.3.1) is usually employed as the aim is data scaling and not parameter estimation. Figure 5.16 represents the new PC plots. Whereas the scores plot hardly changes in appearance, there is a dramatic difference in the appearance of the loadings. The reason is that standardization puts all the variables on approximately the same scale. Hence variables (such as wavelengths) of low intensity assume equal significance to those of high intensity,



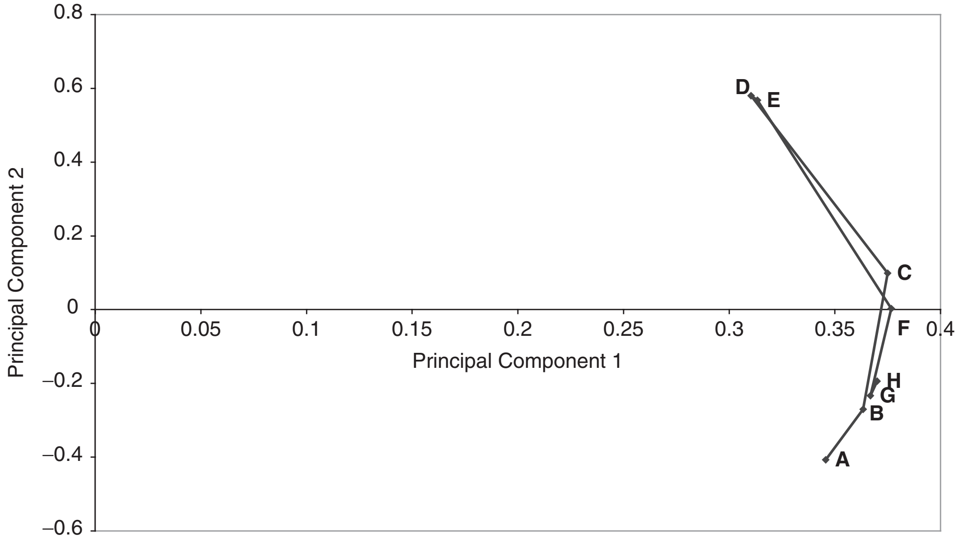


**Figure 5.16** Scores and loadings plots corresponding to Figure 5.14 but for column standardized data

and, in this case all variables are roughly the same distance away from the origin, on an approximate circle (this looks distorted simply because the horizontal axis is longer than the vertical one in the graph).

Standardization can be important in many real situations. Consider, for example, a case where the concentrations of 30 metabolites are monitored in a series of organisms. Some metabolites might be abundant in all samples, but their variation is not very significant. The change in concentration of the minor compounds might have a significant relationship to the underlying biology. If standardization is not performed, PCA will be dominated by the most intense compounds. In some cases standardization (or closely similar types of scaling) is essential. In the case of using physical properties to look at relationships between elements as discussed in Section 5.3, each raw variable is measured on radically different scales and standardization is required so that each variable has an equal influence. Standardization is useful in areas such as quantitative structure – property relationships, where many different pieces of information are measured on very different scales, such as bond lengths and dipoles.

Row scaling involves scaling the rows to a constant total, usually 1 or 100 (this is sometimes called normalization but there is a lot of confusion and conflicting terminology in the literature: usually normalization involves the sum of squares rather than the sum equalling 1, as we use for the loadings – see Section 5.2.2). This is useful if the absolute concentrations of samples cannot easily be controlled. An example might be biological extracts: the precise amount of material might vary unpredictably, but the *relative* proportions of each chemical can be measured. Row scaling introduces a constraint which is often called *closure*. The numbers in the multivariate data matrix are proportions and some of the properties have analogy to those of mixtures (Sections 2.12 and 9.5).

The result of row scaling is presented in Table 5.3(d) and the PC plots are given in Figure 5.17. The scores plot appears very different from those of previous figures. The data points now lie on a straight line (this is a consequence of there being exactly two components in this particular dataset and does not always happen). The 'mixed' points are in the centre of the straight line, with the pure regions at the extreme ends. Note that sometimes if extreme points are primarily influenced by noise, the PC plot can be quite distorted, and it can be important to select carefully an appropriate region of the data.

There are a very large battery of other methods for data preprocessing, although the ones described above are the most common.

- It is possible to combine approaches, for example, first to row scale and then standardize a dataset.
- Weighting of each variable according to any external criterion of importance is sometimes employed.
- Logarithmic scaling of measurements might be useful if there are large variations in intensities, although there can be problems if there are missing or zero intensity measurements.
- Selective row scaling over part of the variables can sometimes be used. It is even possible to divide the measurements into blocks and perform row scaling separately on each block. This could be useful if there were several types of measurement, for example, a couple of spectra and one chromatogram, each constituting a single block, and each of equal importance, but recorded on different physical scales.

Undoubtedly, however, the appearance and interpretation not only of PC plots but the result of almost all chemometric techniques, depends on data preprocessing. The influence of preprocessing can be quite dramatic, so it is essential for the user of chemometric software to understand and question how and why the data has been scaled prior to interpreting the result from a package.

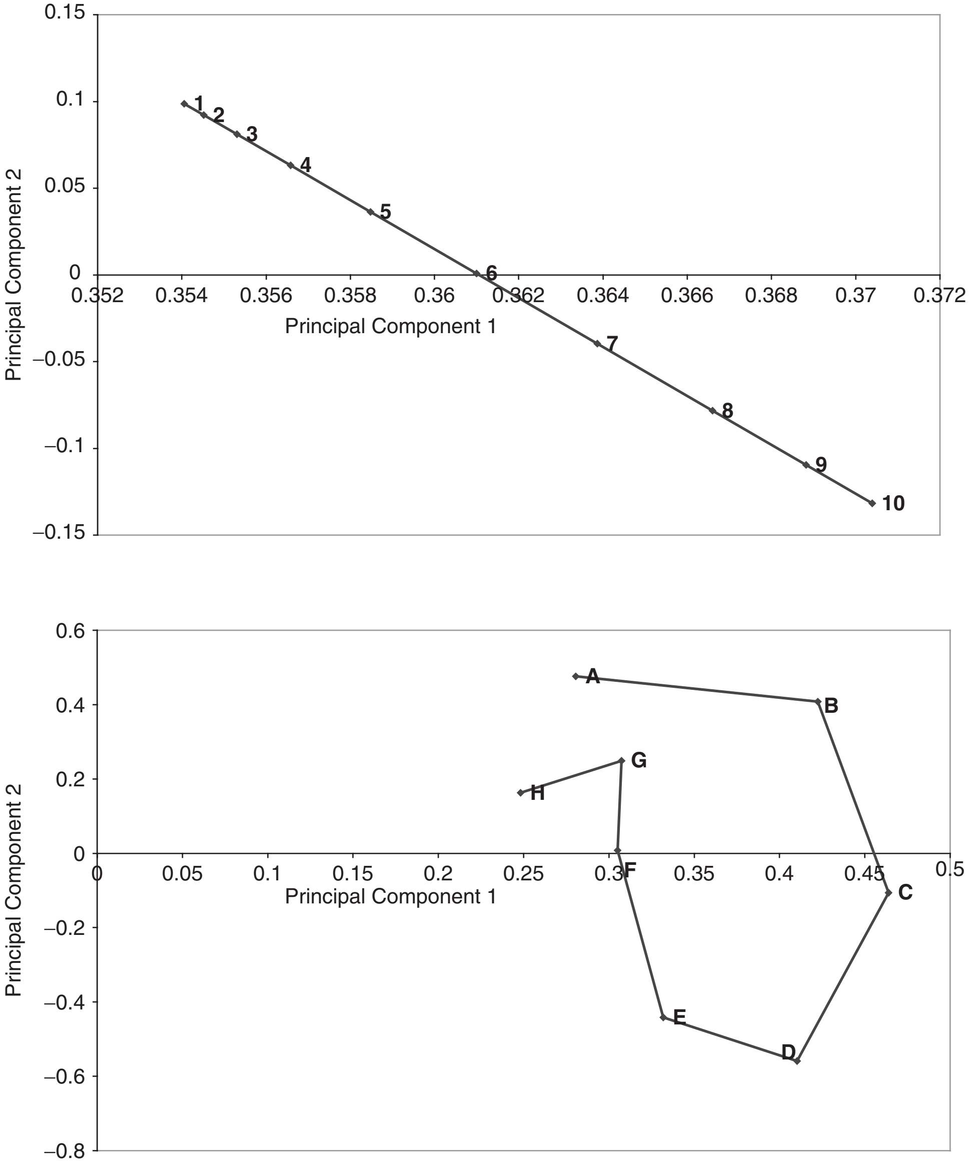


**Figure 5.17** Scores and loadings plots corresponding to Figure 5.14 but for row scaled data

## 5.6 UNSUPERVISED PATTERN RECOGNITION: CLUSTER ANALYSIS

Exploratory data analysis such as PCA is used primarily to determine general relationships between data. Sometimes more complex questions need to be answered such as, do the samples fall into groups? Cluster analysis is a well established approach that was developed primarily by biologists to determine similarities between organisms. Numerical taxonomy emerged from a desire to determine relationships between different species, for example genera, families and phyla. Many textbooks in biology show how organisms are related using family trees. In Chapter 11 we will expand on how cluster analysis can be employed by biological chemists.

However, the chemist also wishes to relate samples in a similar manner. Can the chemical fingerprint of wines be related and does this tell us about the origins and taste of a particular wine? Unsupervised pattern recognition employs a number of methods, primarily cluster analysis, to group different samples (or objects) using chemical measurements.

The first step is to determine the similarity between objects. Table 5.4 represents six objects, (1–6) and seven measurements (A–G). What are the similarities between the objects? Each object has a relationship to the remaining five objects.

A number of common numerical measures of similarity are available.

1. *Correlation coefficient* between samples (see Section 3.3.3 for the definition). A correlation coefficient of 1 implies that samples have identical characteristics.
2. *Euclidean distance*. The distance between samples $k$ and $l$ is defined by:

$$d_{kl} = \sqrt{\sum_{j=1}^{J} (x_{kj} - x_{lj})^2}$$

where there are $j$ measurements, and $x_{ij}$ is the $j$th measurement on sample $i$, for example, $x_{23}$ is the third measurement on the second sample, equalling 0.6 in Table 5.4. The smaller this value, the more similar the samples, so this distance measure works in an opposite manner to the correlation coefficient.

3. *Manhattan distance*. This is defined slightly differently to the Euclidean distance and is given by:

$$d_{kl} = \sum_{j=1}^{J} |x_{kj} - x_{lj}|$$

Once a distance measure has been chosen, a similarity (or dissimilarity) matrix can be drawn up. Using the correlation coefficients (measure 1 above), the matrix is presented in Table 5.5, for our dataset. Notice that the correlation of any object with itself is always

**Table 5.4** Simple example for cluster analysis; six objects (1–6) and seven variables (A–G)

| Objects | Variables | | | | | | |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | G |
| 1 | 0.9 | 0.5 | 0.2 | 1.6 | 1.5 | 0.4 | 1.5 |
| 2 | 0.3 | 0.2 | 0.6 | 0.7 | 0.1 | 0.9 | 0.3 |
| 3 | 0.7 | 0.2 | 0.1 | 0.9 | 0.1 | 0.7 | 0.3 |
| 4 | 0.5 | 0.4 | 1.1 | 1.3 | 0.2 | 1.8 | 0.6 |
| 5 | 1.0 | 0.7 | 2.0 | 2.2 | 0.4 | 3.7 | 1.1 |
| 6 | 0.3 | 0.1 | 0.3 | 0.5 | 0.1 | 0.4 | 0.2 |

**Table 5.5** Correlation matrix for the six objects in Table 5.4

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | | | | | |
| 2 | −0.338 | 1 | | | | |
| 3 | 0.206 | 0.587 | 1 | | | |
| 4 | −0.340 | 0.996 | 0.564 | 1 | | |
| 5 | −0.387 | 0.979 | 0.542 | 0.990 | 1 | |
| 6 | −0.003 | 0.867 | 0.829 | 0.832 | 0.779 | 1 |

1, and that only half the matrix is required, because the correlation of any two objects is always identical no matter which way round the coefficient is calculated. This matrix gives an indication of relationships: for example, object 5 appears very similar to both objects 2 and 4, as indicated by the high correlation coefficient. Object 1 does not appear to have a particularly high correlation with any of the others.

The next step is to link the objects. The most common approach is called *agglomerative* clustering whereby single objects are gradually connected to each other in groups.

- From the raw data, find the two most similar objects, in our case the objects with the highest correlation coefficient (or smallest distance). According to Table 5.5, these are objects 2 and 4, as their correlation coefficient is 0.996.
- Next form a ‘group’ consisting of the two most similar objects. The original six objects are now reduced to five groups, namely objects 1, 3, 5 and 6 on their own and a group consisting of objects 2 and 4 together.
- The tricky bit is to decide how to represent this new grouping. There are quite a few approaches, but it is common to change the data matrix from one consisting of six rows to a new one of five rows, four corresponding to original objects and one to the new group. The numerical similarity values between this new group and the remaining objects have to be recalculated. There are three principal ways of doing this:
  - *Nearest neighbour*. The similarity of the new group from all other groups is given by the *highest* similarity of either of the original objects to each other object. For example, object 6 has a correlation coefficient of 0.867 with object 2, and 0.837 with object 4. Hence the correlation coefficient with the new combined group consisting of objects 2 and 4 is 0.867.
  - *Farthest neighbour*. This is the opposite to nearest neighbour, and the *lowest* similarity is used, 0.837 in our case.
  - *Average linkage*. The average similarity is used, 0.852 in our case. There are, in fact, two different ways of doing this, according to the size of each group being joined together. Where they are of equal size (e.g. each group consists of one object), both methods are equivalent. The two different ways are as follows. *Unweighted* linkage involves taking the each group size into account when calculating the new similarity coefficient, the more the objects the more significant the similarity measure is whereas *weighted* linkage ignores the group size. The terminology indicates that for the unweighted method, the new similarity measure takes into consideration the number of objects in a group, the conventional terminology possibly being the opposite to what is expected. For the first link, each method provides identical results.

**Table 5.6** First step of clustering of data from Table 5.5, with the new correlation coefficients indicated as shaded cells, using nearest neighbour linkage

| | 1 | 2 and 4 | 3 | 5 | 6 |
|---|---|---|---|---|---|
| 1 | 1 | | | | |
| 2 and 4 | −0.338 | 1 | | | |
| 3 | 0.206 | 0.587 | 1 | | |
| 5 | −0.387 | 0.990 | 0.542 | 1 | |
| 6 | −0.003 | 0.867 | 0.829 | 0.779 | 1 |

As an illustration, the new data matrix using nearest neighbour clustering is presented in Table 5.6, with the new values shaded. Remember that there are many similarity measures and methods for linking, so this table is only one possible way for handling the information.

The next steps consist of continuing to group the data just as above, until only one group, consisting of all the original objects, remains. Since there are six original objects, there will be five steps before the data are reduced to a single group.

It is normal to then determine at what similarity measure each object joined a larger group, and so which objects resemble each other most.

Often the result of hierarchical clustering is presented in a graphical form called a dendrogram: note that many biologists call this a phylogram and it differs from a cladogram where the size of the branches are the same (see Section 11.4). The objects are organized in a row, according to their similarities: the vertical axis represents the similarity measure at which each successive object joins a group. Using nearest neighbour linkage and correlation coefficients for similarities, the dendrogram is presented in Figure 5.18. It can be seen that object 1 is very different from the others. In this case all the other objects appear to form a single group, but other clustering methods may give slightly different results. A good approach is

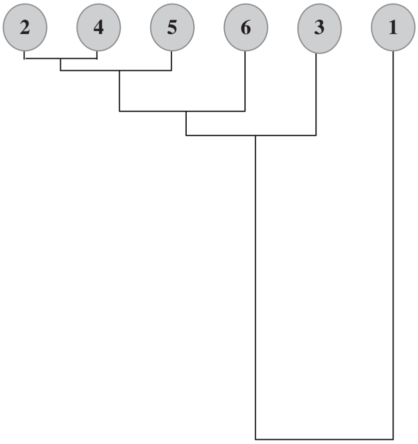


**Figure 5.18** Dendrogram for data in Table 5.4, using correlation coefficients as similarity measures and nearest neighbour clustering

to perform several different methods of cluster analysis and compare the results. If similar groupings remain, no matter which method is employed, we can rely on the results.

There are a large number of books on clustering but a well recognized text, for chemists, is written by Massart and Kaufman [7]. Although quite an early vintage chemometrics text, it has survived the passage of time, and there are many clear explanations in this book. More aspects of cluster analysis are discussed in Chapter 11.

## 5.7 SUPERVISED PATTERN RECOGNITION

Classification (often called supervised pattern recognition) is at the heart of chemistry. Mendeleev's periodic table, the grouping of organic compounds by functionality and listing different reaction types all involve classification. Much of traditional chemistry involves grouping chemical behaviour. Most traditional texts in organic and inorganic chemistry are systematically divided into chapters according to the behaviour or structure of the underlying compounds or elements.

So the modern chemist also has a significant need for classification. Can a spectrum be used to determine whether a compound is a ketone or an ester? Can the chromatogram of a tissue sample be used to determine whether a patient is cancerous or not? Can we record the spectrum of an orange juice and decide its origin? Is it possible to monitor a manufacturing process and decide whether the product is acceptable or not? Supervised pattern recognition is used to assign samples to a number of groups (or classes). It differs from unsupervised pattern recognition (Section 5.6) where, although the relationship between samples is important, there are no predefined groups.

Although there are numerous algorithms in the literature, chemists have developed a common strategy for classification.

### 5.7.1 Modelling the Training Set

The first step is normally to produce a mathematical model between some measurements (e.g. spectra) on a series of objects and their known groups. These objects are called a *training set*. For example, a training set might consist of the near infrared (NIR) spectra of 30 orange juices, 10 known to be from Spain, 10 known to be from Brazil and 10 known to be adulterated. Can we produce a mathematical equation that predicts which class an orange juice belongs to from its spectrum?

Once this is done it is usual to determine how well the model predicts the groups. Table 5.7 illustrates a possible scenario. Of the 30 spectra, 24 are correctly classified. Some classes are modelled better than others, for example, nine out of 10 of the Spanish orange juices are correctly classified, but only seven of the Brazilian orange juices. A parameter %CC (percentage correctly classified) can be calculated and is 80 % overall. There appears some risk of making a mistake, but the aim of a spectroscopic technique might be to perform screening, and there is a high chance that suspect orange juices (e.g. those adulterated) would be detected, which could then be subject to further detailed analysis. Chemometrics combined with spectroscopy acts like a 'sniffer dog' in a customs checkpoint trying to detect drugs: the dog may miss some cases, and may even get excited when there are no drugs, but there will be a good chance the dog is correct. Proof, however, only comes when the suitcase is opened.

Sometimes the number of false positives or false negatives can be computed as an alternative measure of the quality of a classification technique. This, however, can only be

**Table 5.7** Classification ability using a training set

| | | Predicted | | | |
|---|---|---|---|---|---|
| Known | Spain | Brazil | Adulterated | Correct | %CC |
| Spain | 9 | 0 | 1 | 9 | 90 |
| Brazil | 1 | 7 | 2 | 7 | 70 |
| Adulterated | 0 | 2 | 8 | 8 | 80 |
| Overall | | | | 24 | 80 |

done for what is called a 'one class classifier', i.e. one class against the rest. We may however be interested in whether an orange juice is adulterated or not. The data of Table 5.7 suggest that there are three false positives (situations where the orange juice is not adulterated but the test suggests it is adulterated) and two false negatives. This number can often be changed by making the classification technique more or less 'liberal'. A liberal technique lets more samples into the class, so would have the effect of increasing the number of false positives but decreasing the number of false negatives. Ultimately we would hope that a method can be found for which there are no false negatives at the cost of several more false positives. Whether this is useful or not depends a little on the application. If we are, for example, screening people for cancer it is better that we reduce the number of false negatives, so all suspect cases are then examined further. If, however, we are deciding whether to cut a person's leg off due to possible disease it is preferable to err on the side of false negatives so we are very sure when we cut the leg off that it is really necessary. More discussion of this approach is provided in Section 10.5.

### 5.7.2 Test Sets, Cross-validation and the Bootstrap

It is normal that the training set results in good predictions, but this does not necessarily mean that the method can safely be used to predict unknowns. A recommended second step is to test the quality of predictions often using a *test* set. This is a series of samples that has been left out of the original calculations, and is a bit like a 'blind test'. These samples are assumed to be unknowns at first. Table 5.8 is of the predictions from a test set (which does not necessarily need to be the same size as the training set), and we see that now only 50 % are correctly classified so the model is not particularly good.

Using a test set to determine the quality of predictions is a form of *validation*. The test set could be obtained, experimentally, in a variety of ways, for example, 60 orange juices might be analysed in the first place, and then randomly divided into 30 for the training set and 30 for the test set. Alternatively, the test set could have been produced in an independent laboratory.

An alternative approach is *cross-validation*. Only a single training set is required, but what happens is that one (or a group) of objects is removed at a time, and a model determined on the remaining samples. Then the prediction on the object (or set of objects) left out is tested. The most common approach is Leave One Out (LOO) cross-validation where one sample is left out at a time. This procedure is repeated until all objects have been left out in turn.

**Table 5.8** Classification ability using a test set

| | Predicted | | | | |
|---|---|---|---|---|---|
| Known | Spain | Brazil | Adulterated | Correct | %CC |
| Spain | 5 | 3 | 2 | 5 | 50 |
| Brazil | 1 | 6 | 3 | 6 | 60 |
| Adulterated | 4 | 2 | 4 | 4 | 40 |
| Overall | | | | 15 | 50 |

For example, it would be possible to produce a class model using 29 out of 30 orange juices. Is the 30th orange juice correctly classified? If so this counts towards the percentage correctly classified. Then, instead of removing the 30th orange juice, we decide to remove the 29th and see what happens. This is repeated 30 times, which leads to a value of %CC for cross-validation. Normally the cross-validated %CC is lower (worse) than the %CC for the training set.

Finally, mention should be made of a third alternative called the *bootstrap* [8]. This is a half way house between cross-validation and having a single independent test set, and involves iteratively producing several internal test sets, not just removing samples once as in cross-validation, but not just having a single test set. A set of samples may be removed for example 50 times, each time including a different combination of the original samples (although the same samples will usually be part of several of these test sets). The prediction ability each time is calculated, and the overall predictive ability is the average of each iteration.

However, if the %CC obtained when samples are left out is similar to the %CC on the training set (sometimes called the autopredictive model), the model is quite probably a good one. Where investigation is necessary is if the %CC is high for the training set but significantly lower when using one of the methods for validation. It is recommended that all classification methods are validated.

Naturally it is also possible to calculate the false positive or false negative rate as well using these, and which criterion is employed to judge whether a method is suitable or not depends very much on the perspective of the scientist.

If the model is not very satisfactory there are a number of ways to improve it. The first is to use a different computational algorithm. The second is to modify the existing method – a common approach might involve wavelength selection in spectroscopy, for example, instead of using an entire spectrum, many wavelengths which are not very meaningful, can we select the most diagnostic parts of the spectrum? Finally, if all else fails, change the analytical technique.

One important final consideration to remember that some people do not always watch out for is that there are two separate reasons for using the techniques described in this section. The first is to optimize a computational model. This means that different models can be checked and the one that gives the best prediction rate is retained. In this way the samples left out are actually used to improve the model. The second is as an independent test of how well the model performs on unknowns. This is a subtly different reason and sometimes both motivations are mixed up, which can lead to over-optimistic predictions of the quality of a model on unknowns, unless care is taken. This can be overcome by dividing the data into a training and test set, but then performing cross-validation or the bootstrap on the training set, to find the best model for the training set and testing its quality on the test set. Using iterative methods, this can be done several times, each time producing a different test set, and the predictive ability averaged.

### 5.7.3 Applying the Model

Once a satisfactory model is available, it can then be applied to unknown samples, using analytical data such as spectra or chromatograms, to make predictions. Usually by this stage, special software is required that is tailor made for a specific application, and measurement technique. The software will also have to determine whether a new sample really fits into the training set or not. One major difficulty is the detection of outliers that belong to none of the previously studied groups, for example if a Cypriot orange juice sample was measured when the training set consists just of Spanish and Brazilian orange juices. In areas such as clinical or forensic science outlier detection can be quite important, indeed an incorrect conviction or inaccurate medical diagnosis could be obtained otherwise. Multivariate outlier detection is discussed in Section 3.12.3.

Another problem is to ensure stability of the method over time, for example, instruments tend to perform slightly differently every day. Sometimes this can have a serious influence on the classification ability of chemometrics algorithms. One way around this is to perform a small test of the instrument on a regular basis and only accept data if the performance of this test falls within certain limits. However, in some cases such as chromatography this can be quite difficult because columns and instruments do change their performance with time, and this can be an irreversible process that means that there will never be absolutely identical results over a period of several months. In the case of spectroscopy such changes are often not so severe and methods called calibration transfer can be employed to overcome these problems often with great success.

There have been some significant real world successes of using classification techniques, a major area being in industrial process control using NIR spectroscopy. A manufacturing plant may produce samples on a continuous basis, but there are a large number of factors that could result in an unacceptable product. The implications of producing substandard batches may be economical, legal and environmental, so continuous testing using a quick and easy method such as on-line spectroscopy is valuable for rapid detection whether a process is going wrong. Chemometrics can be used to classify the spectra into acceptable or otherwise, and so allow the operator to close down a manufacturing plant in real time if it looks as if a batch can no longer be assigned to the group of acceptable samples.

## 5.8 STATISTICAL CLASSIFICATION TECHNIQUES

The majority of statistically based software packages contain substantial numbers of procedures, called by various names such as discriminant analysis and canonical variates analysis. It is important to emphasize that good practice requires methods for validation and optimization of the model as described in Section 5.7, together with various classification algorithms as discussed below.

### 5.8.1 Univariate Classification

The simplest form of classification is *univariate* where one measurement or variable is used to divide objects into groups. An example may be a blood alcohol reading. If a reading on a meter in a police station is above a certain level, then the suspect will be prosecuted for drink driving, otherwise not. Even in such a simple situation, there can be ambiguities, for example measurement errors and metabolic differences between people.

### 5.8.2 Bivariate and Multivariate Discriminant Models

More often, several measurements are required to determine the group a sample belongs to. Consider performing two measurements, and producing a graph of the values of these measurements for two groups, as in Figure 5.19. The objects denoted by squares are clearly distinct from the objects denoted by circles, but neither of the two measurements, alone, can discriminate between these groups, therefore both are essential for classification. It is, however, possible to draw a line between the two groups. If above the line, an object belongs to the group denoted by circles (class A), otherwise to the group denoted by squares (class B).

Graphically this can be represented by *projecting* the objects onto a line at right angles to the discriminating line as demonstrated in Figure 5.20. The projection can now be converted to a position along a single line (line 2). Often these numbers are converted to a *class distance* which is the distance of each object to the centre of the classes. If the distance to the centre of class A is greater than that to class B, the object is placed in class A and vice versa.

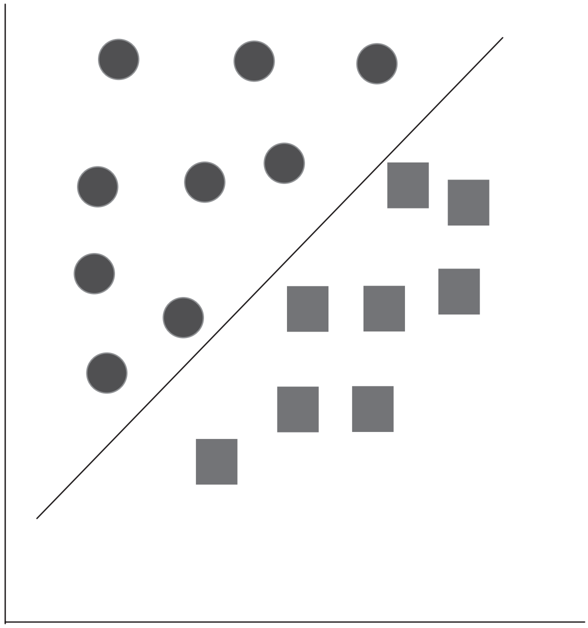


**Figure 5.19** Bivariate classification where no measurement alone can distinguish groups

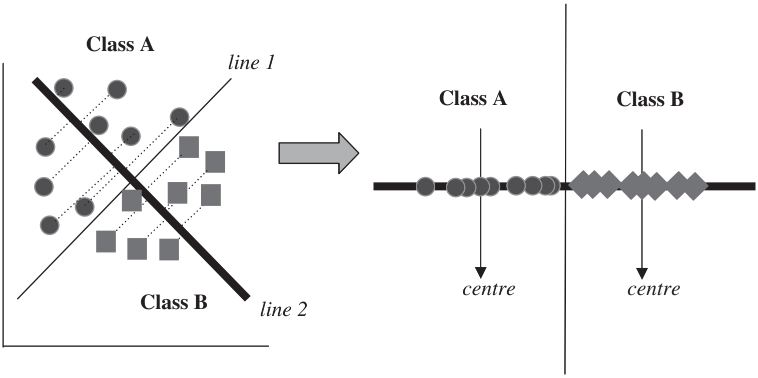


**Figure 5.20** Projections

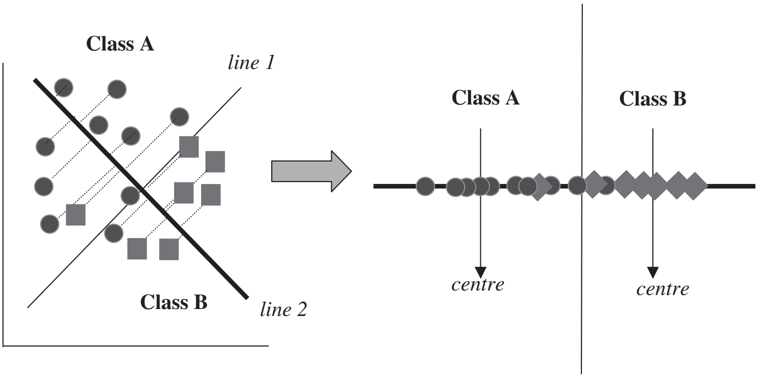


**Figure 5.21** Projections where it is not possible to unambiguously classify objects

Sometimes the projected line is called a canonical variate, although in statistics this can have quite a formal meaning.

It is not always possible to exactly divide the classes into two groups by this method (see Figure 5.21) but the mis-classified samples should be far from the centre of both classes, with two class distances that are approximately equal. The data can be presented in the form of a class distance plot where the distance of each sample to the two class centres are visualized, which can be divided into regions as shown in Figure 5.22. The top right-hand region is one in which classification is ambiguous.

Figure 5.22 is rather simple, and probably does not tell us much that cannot be shown from Figure 5.21. However, the raw data actually consist of more than one measurement, so it is possible to calculate the class distance using the raw two-dimensional information, as shown in Figure 5.23. The points no longer fall onto straight lines, but the graph can still be divided into four regions.

- Top left: almost certainly class A.
- Bottom left: unambiguous membership.
- Bottom right: almost certainly class B.
- Top right: unlikely to be a member of either class, sometimes called an outlier.

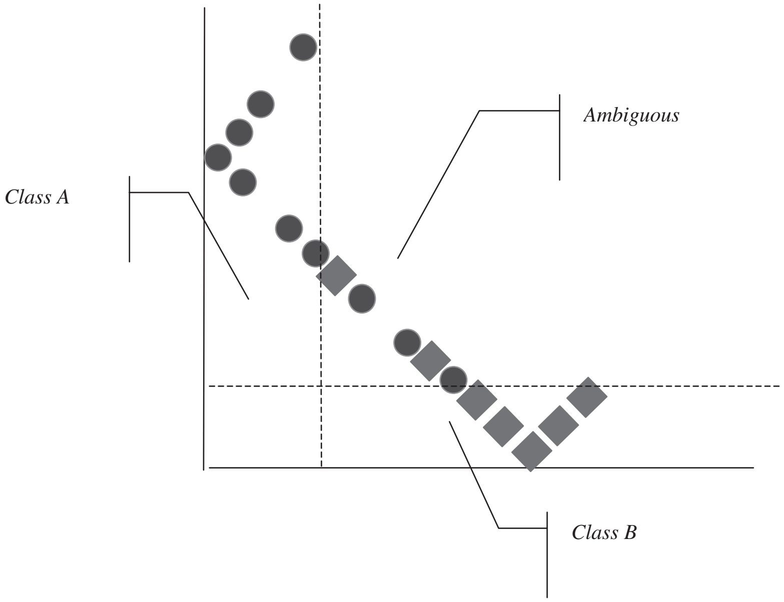


**Figure 5.22** A simple class distance plot corresponding to the projection in Figure 5.21: horizontal axis = distance from class A, vertical axis = distance from class B

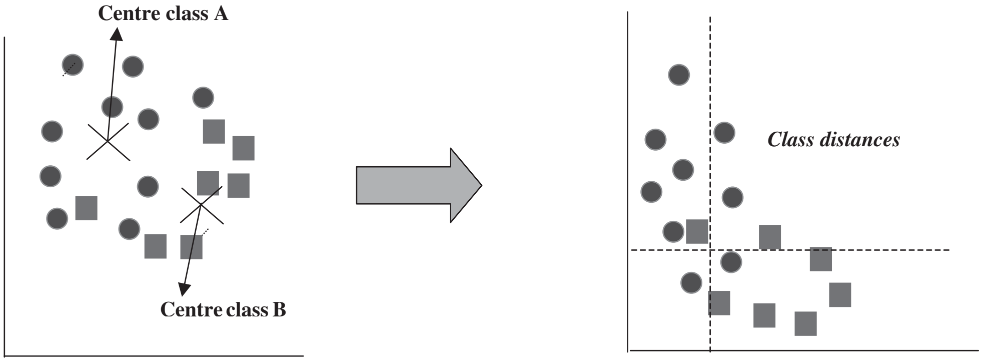


**Figure 5.23** Class distance plot using two-dimensional information about centroids

In chemistry, these four divisions are perfectly reasonable. For example, if we try to use spectra to classify compounds into ketones and esters, there may be some compounds that are both or neither. If, on the other hand, there are only two possible classifications, for example whether a manufacturing sample is acceptable or not, a conclusion about objects in the bottom left or top right is that the analytical data are insufficiently good to allow us to conclusively assign a sample to a group. This is a valuable conclusion, for example it is helpful to tell a laboratory that their clinical diagnosis or forensic test is inconclusive and that if they want better evidence they should perform more experiments or analyses.

It is easy to extend the methods above to multivariate situations where instead of two variables many (which can run to several hundreds in chromatography and spectroscopy) are used to form the raw data.

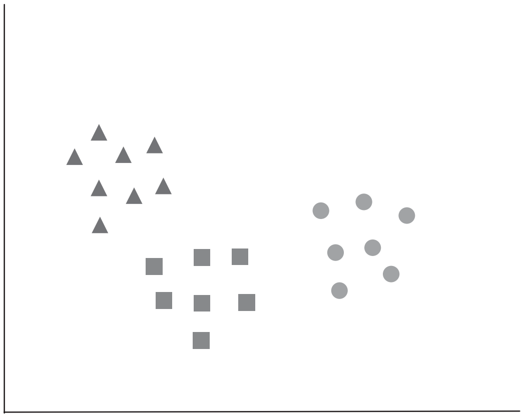


**Figure 5.24** Three classes

Most methods for discriminant analysis can contain a number of extensions. The most common is to scale the distances from the centre of a class by the variance (or spread) in the measurements for a particular class. The greater this variance, the less significant a large distance is. Hence in Figure 5.23, Class A is more dispersed compared with Class B, and so a large distance from the centre is indicative of a poor fit to the model. The class distance plot can be adjusted to take this into account. The most common distance measure that takes this into account is the *Mahalanobis distance* which is contrasted to the *Euclidean distance* above; the principles are described in greater detail in the context of biological pattern recognition in Section 10.4 but are generally applicable to all classification procedures.

In most practical cases, more than two variables are recorded, indeed in spectroscopy there may be several hundred measurements, and the aim of discriminant analysis is to obtain projections of the data starting with much more complex information. The number of 'canonical variates' equals the number of classes minus one, so, in Figure 5.24 there are three classes and two canonical variates.

## 5.8.3 SIMCA

The SIMCA method, first advocated by the Swedish organic chemist Svante Wold in the early 1970s, is regarded by many as a form of soft modelling used in chemical pattern recognition. Although there are some differences with discriminant analysis as employed in traditional statistics, the distinction is not as radical as many would believe. However, SIMCA has an important role in the history of chemometrics so it is important to understand the main steps of the method.

The acronym stands for *Soft Independent Modelling of Class Analogy* (as well as the name of a French car). The idea of soft modelling is illustrated in Figure 5.25. Two classes can overlap (hence are 'soft'), and there is no problem with an object belonging to both (or neither) class simultaneously: hence there is a region where both classes overlap. When we perform hard modelling we insist that an object belongs to a discrete class. For example, a biologist trying to sex an animal from circumstantial evidence (e.g. urine samples), knows that the animal cannot simultaneously belong to two sexes at the same time, and a forensic scientist trying to determine whether a banknote is forged or not, knows that there can

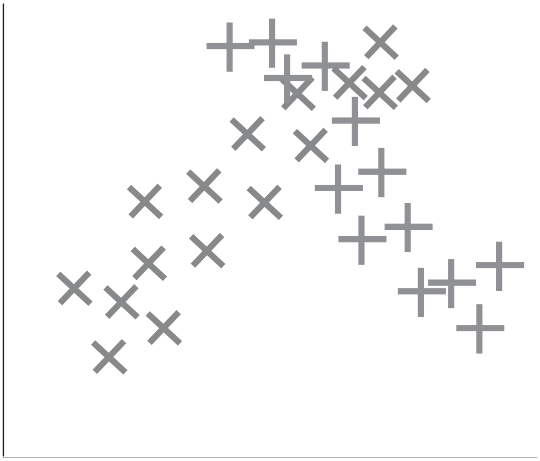


**Figure 5.25** Overlapping classes

be only one answer: if this appears not to be so, the problem lies with the quality of the evidence. The original philosophy of soft modelling was that, in many situations in chemistry, it is entirely legitimate for an object to fit into more than one class simultaneously, for example a compound may have an ester and an alkene group, so will exhibit spectroscopic characteristics of both functionalities, hence a method that assumes the answer must be either a ketone or an alkene is unrealistic. In practice, there is not such a large distinction between hard (traditional discriminant analysis) and soft models and it is possible to have a class distance derived from hard models that is close to two or more groups.

Independent modelling of classes, however, is a more useful feature. After making a number of measurements on ketones and alkenes, we may decide to include amides in the model. Figure 5.26 represents a third class (triangles). This new class can be added independently to the existing model without any changes. This contrasts to some other methods of classification in which the entire modelling procedure must be repeated if different numbers of groups are employed.

The main steps of SIMCA are as follows.

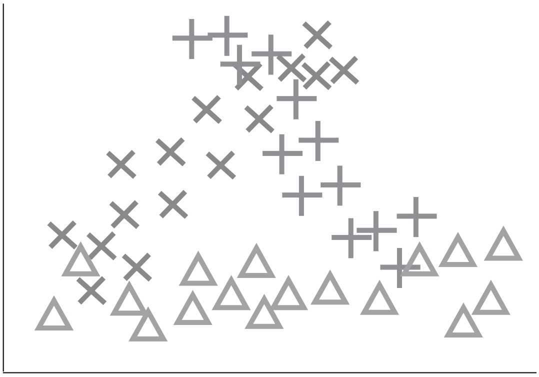


**Figure 5.26** Three classes

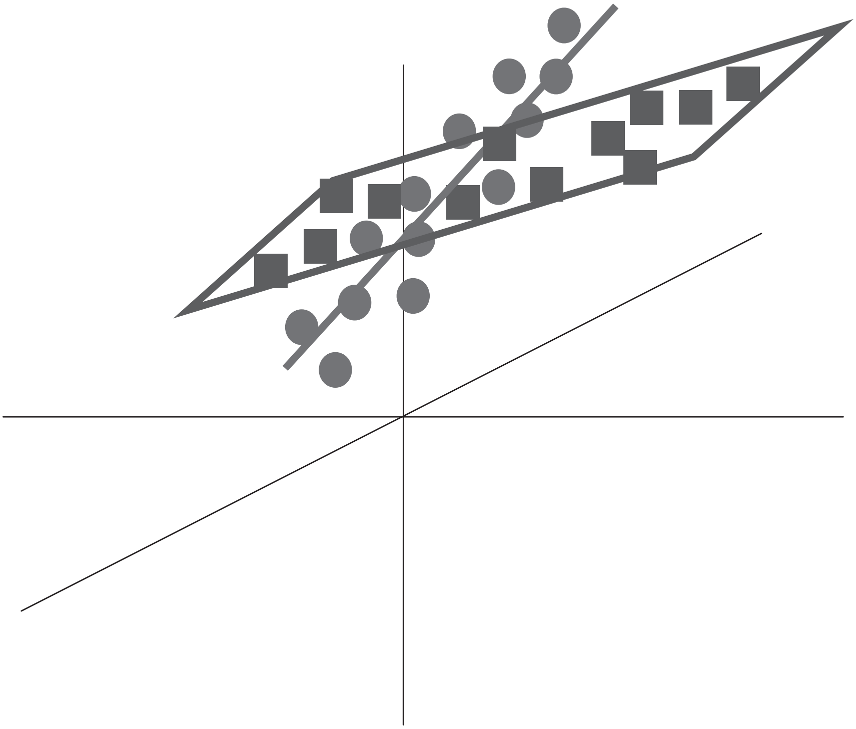


**Figure 5.27** Two groups, one modelled by one PC and one by two PCs

Each group is independently modelled using PCA. Note that each group could be described by a different number of PCs. Figure 5.27 represents two groups each characterized by three raw measurements, which may, for example, be chromatographic peak heights or physical properties. However, one group falls mainly on a straight line, which is defined as the first PC of the group. The second group falls roughly on a plane: the axes of this plane are the first two PCs of this group. This way of looking at PCs (axes that best fit the data) are sometimes used by chemists, and are complementary to the definitions introduced previously (Section 5.2.2). It is important to note that there are a number of proposed methods for determining how many PCs are most suited to describe a class, of which the original advocates of SIMCA preferred cross-validation (Section 5.10).

The class distance can be calculated as the geometric distance from the PC models (see Figure 5.28). The unknown is much closer to the plane formed from the group represented by squares than the line formed by the group represented by circles, and so is tentatively assigned to this class. A rather more elaborate approach is in fact usually employed in which each group is bounded by a region of space, which represents 95 % confidence that a particular object belongs to a class. Hence geometric class distances can be converted to statistical probabilities.

Sometimes it is interesting to see which variables are useful for discrimination. There are often good reasons, for example in gas chromatography-mass spectrometry we may have hundreds of peaks in a chromatogram and be primarily interested in a very small number

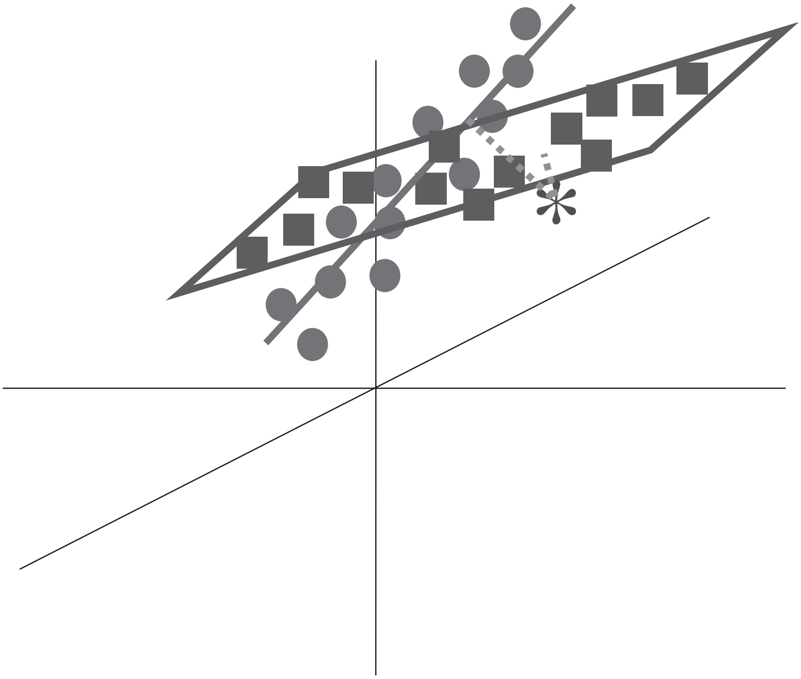


**Figure 5.28** Distance of an unknown sample (asterisk) to two known classes

of, for example, biomarkers that are used to distinguish two groups, so this interpretation can have a chemical basis.

The *modelling power* of each variable in each class is defined by:

$$M_j = 1 - s_{jresid}/s_{jsraw}$$

where $s_{jraw}$ is the standard deviation of the variable in the raw data, and $s_{jresid}$ the standard deviation of the variable in the residuals given by:

$$\boldsymbol{E} = \boldsymbol{X} - \boldsymbol{T}.\boldsymbol{P}$$

which is the difference between the observed data and the PC model as described earlier. The modelling power varies between 1 (excellent) and 0 (no discrimination). Variables with $M$ below 0.5 are of little use.

Another second measure is how well a variable discriminates between two classes. This is distinct from modelling power – being able to model one class well does not necessarily imply being able to discriminate two groups effectively. In order to determine this, it is necessary to fit each sample to both class models. For example, fit sample 1 to the PC model of both class A and class B. The residual matrices are then calculated, just as for discriminatory power, but there are now four such matrices:

1. Samples in class A fitted to the model of class A.
2. Samples in class A fitted to the model of class B.

3. Samples in class B fitted to the model of class B.
4. Samples in class B fitted to the model of class A.

We would expect matrices 2 and 4 to be a worse fit than matrices 1 and 3. The standard deviations are then calculated for these matrices to give:

$$D_j = \sqrt{\frac{{}^{\text{class A model B}}s_{jresid}^2 + {}^{\text{class B model A}}s_{jresid}^2}{{}^{\text{class A model A}}s_{jresid}^2 + {}^{\text{class B model B}}s_{jresid}^2}}$$

The bigger the value the higher the discriminatory power. This could be useful information, for example if clinical or forensic measurements are expensive, so allowing the experimenter to choose only the most effective measurements.

The original papers of SIMCA have been published by Wold and coworkers [9,10]. It is important not to get confused between the method for supervised pattern recognition and the SIMCA software package which, in fact, is much more broadly based. An alternative method proposed in the literature for soft modelling is UNEQ developed by Massart and coworkers [11].

### 5.8.4 Statistical Output

Software packages produce output in a variety of forms, some of which are listed below:

- The distances for each object from each class, suitably scaled as above.
- The most appropriate classification, and so per cent correctly classified (see Section 5.7).
- Probability of class membership, which relates to class distance. This probability can be high for more than one class simultaneously, for example if a compound exhibits properties both of a ketone or ester.
- Which variables are most useful for classification (e.g. which wavelengths or physical measurements), important information for future analyses.
- Variance within a class: how spread out a group is. For example, in the case of forgeries, the class of nonforged materials is likely to be much more homogeneous than the forged materials.

Information is not restricted to the training set, but can also be used in an independent test set or via cross-validation, as discussed above.

## 5.9 K NEAREST NEIGHBOUR METHOD

The methods of SIMCA (Section 5.8.3) and discriminant analysis (Section 5.8.2) discussed above involve producing statistical models, such as PCs and canonical variates. Nearest neighbour methods are conceptually much simpler, and do not require elaborate statistical computations.

The K Nearest Neighbour (KNN) method has been with chemists for over 30 years. The algorithm starts with a number of objects assigned to each class. Figure 5.29 represents five objects belonging to two classes, class A (diamonds) and class B (squares), recorded using two measurements which may, for example, be chromatographic peak areas or absorption intensities at two wavelengths.

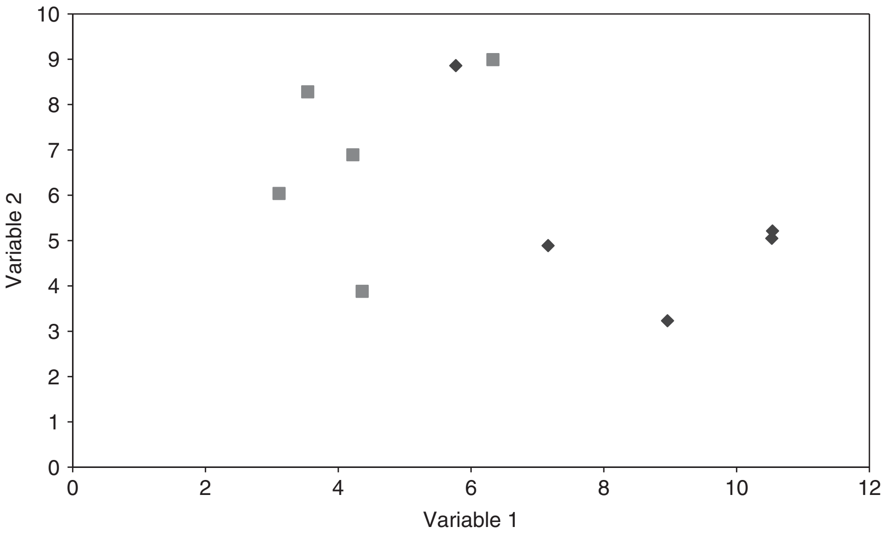


**Figure 5.29** Objects in two classes

**Table 5.9** Example for KNN calculations

| Class | Measurement 1 | Measurement 2 | Distance to unknown | Rank |
|---|---|---|---|---|
| A | 5.77 | 8.86 | 3.86 | 6 |
| A | 10.54 | 5.21 | 5.76 | 10 |
| A | 7.16 | 4.89 | 2.39 | 4 |
| A | 10.53 | 5.05 | 5.75 | 9 |
| A | 8.96 | 3.23 | 4.60 | 8 |
| B | 3.11 | 6.04 | 1.91 | 3 |
| B | 4.22 | 6.89 | 1.84 | 2 |
| B | 6.33 | 8.99 | 4.16 | 7 |
| B | 4.36 | 3.88 | 1.32 | 1 |
| B | 3.54 | 8.28 | 3.39 | 5 |
| unknown | 4.78 | 5.13 | | |

The method is implemented as follows:

1. Assign a training set to known classes.
2. Calculate the distance of an unknown to all members of the training set (see Table 5.9). Usually the simple geometric or Euclidean distance is computed.
3. Rank these in order (1 = smallest distance and so on).
4. Pick the $K$ smallest distances and see what classes the unknown in closest to. The case where $K = 3$ is illustrated in Figure 5.30. All objects belong to class B.
5. Take the ‘majority vote’ and use this for classification. Note that if $K = 5$, one of the five closest objects belongs to class A.
6. Sometimes it is useful to perform KNN analysis for a number of different values of $K$, e.g. 3, 5 and 7, and see if the classification changes. This can be used to spot anomalies.

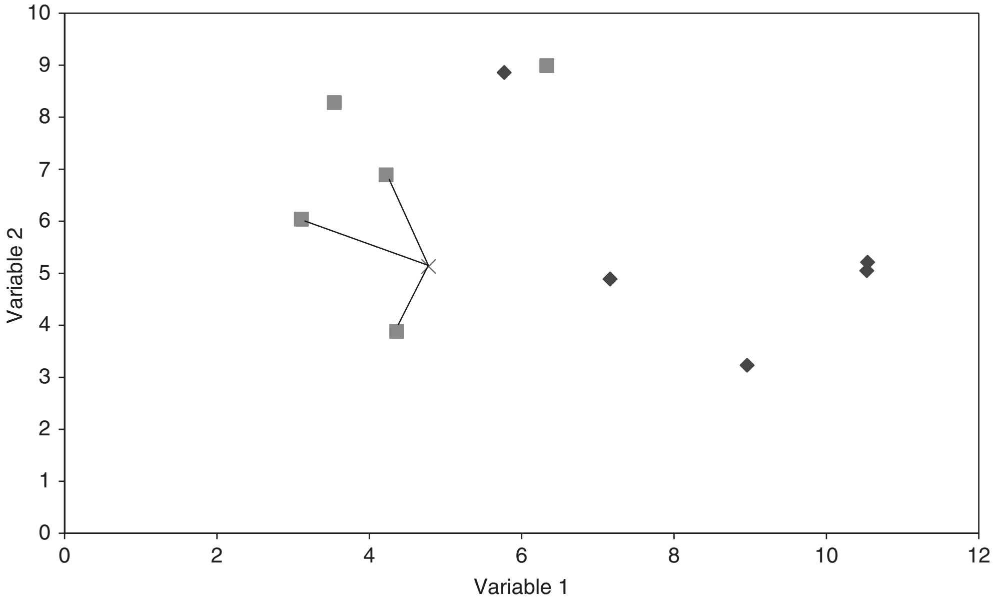


**Figure 5.30** Classifying an unknown using KNN (with $K = 3$)

If, as is usual in chemistry, there are many more than two measurements, it is simply necessary to extend the concept of distance to one in multidimensional space, each axis representing a variable. Although we cannot visualize more than three dimensions, computers can handle geometry in an indefinite number of dimensions, and the idea of distance is easy to generalize. In the case of Figure 5.30 it is not really necessary to perform an elaborate computation to classify the unknown, but when a large number of measurements have been made, it is often hard to determine the class of an unknown by simple graphical approaches.

This conceptually simple approach works well in many situations, but it is important to understand the limitations.

The first is that the numbers in each class of the training set should be approximately equal, otherwise the 'votes' will be biased towards the class with most representatives. The second is that for the simplest implementations, each variable is of equal significance. In spectroscopy, we may record hundreds of wavelengths, and some will either not be diagnostic or else be correlated. A way of getting round this is either to select the variables or else to use another distance measure. The third problem is that ambiguous or outlying samples in the training set can cause major problems in the resultant classification. Fourth, the methods take no account of the spread or variance in a class. For example, if we were trying to determine whether a forensic sample is a forgery, it is likely that the class of forgeries has a much higher variance to the class of nonforged samples.

It is, of course, possible to follow procedures of validation (Section 5.7) just as in all other methods for supervised pattern recognition. There are quite a number of diagnostics that can be obtained using these methods.

However, KNN is a very simple approach that can be easily understood and programmed. Many chemists like these approaches, whilst statisticians often prefer the more elaborate methods involving modelling the data. KNN makes very few assumptions, whereas methods based on modelling often inherently make assumptions such as normality of noise distributions that are not always experimentally justified, especially when statistical tests are employed to provide probabilities of class membership. In practice, a good strategy is to use several different methods for classification and see if similar results are obtained. Often the differences in performance of different approaches are not due to the algorithm itself but in data scaling, distance measures, variable selection, validation method and so on. In this chemometrics probably differs from many other areas of data analysis where there is much less emphasis on data preparation and much more on algorithm development.

## 5.10 HOW MANY COMPONENTS CHARACTERIZE A DATASET?

One of the most controversial and active areas in chemometrics, and indeed multivariate statistics, is the determination of how many PCs are needed to adequately model a dataset. These components may correspond to compounds, for example, if we measure a series of spectra of extracts of seawater, how many significant compounds are there? In other cases these components are simply abstract entities and do not have physical meaning.

Ideally when PCA is performed, the dataset is decomposed into two parts, namely, meaningful information and error (or noise). The transformation is often mathematically described as follows:

$$\boldsymbol{X} = \boldsymbol{T}.\boldsymbol{P} + \boldsymbol{E} = \hat{\boldsymbol{X}} + \boldsymbol{E}$$

where $\hat{\boldsymbol{X}}$ is the 'estimate' of $\boldsymbol{X}$ using the PC model. Further details have been described previously (Section 5.2.2).

There are certain important features of the PC model. The first is that the number of columns in the scores matrix and the number of rows in the loadings matrix should equal the number of significant components in a dataset. Second the error matrix $\boldsymbol{E}$, ideally, should approximate to measurement errors. Some chemists interpret these matrices physically, for example, one of the dimensions of $\boldsymbol{T}$ and $\boldsymbol{P}$ equals the number of compounds in a series of mixtures, and the error matrix provides information on instrumental noise distribution, however, these matrices are not really physical entities. Even if there are 20 compounds in a series of spectra, there may be only four or five significant components, because there are similarities and correlations between the signals from the individual compounds.

One aim of PCA is to determine a sensible number of columns in the scores and loadings matrices. Too few and some significant information will be missed out, too many and noise will be modelled or as many people say, the data are over-fitted. The number of significant components will never be more than the *smaller* of the number of variables (columns) or objects (rows) in the raw data. So if 20 spectra are recorded at 200 wavelengths, there will never be more than 20 nonzero components. Preprocessing (Section 5.5) may reduce the number of possible components still further.

In matrix terms the number of significant components is often denoted the 'rank' of a matrix. If a 15 × 300 $\boldsymbol{X}$ matrix (which may correspond to 15 UV/visible spectra recorded at 1 nm intervals between 201 nm and 500 nm) has a rank of 6, the scores matrix $\boldsymbol{T}$ has six columns, and the loadings matrix $\boldsymbol{P}$ has six rows.

Many approaches for determining the number of significant components relate to the size of successive eigenvalues. The larger an eigenvalue, the more significant the component. If each eigenvalue is defined as the sum of squares of the scores of the corresponding PC, then the sum of all the nonzero eigenvalues equals the overall sum of squares of the original data (after any preprocessing). Table 5.1 illustrates this. The eigenvalues can be converted to percentages of the overall sum of squares of the data, and as more components are calculated, the total approaches 100 %. Statisticians often preprocess their data by centring the columns, and usually define an eigenvalue by a variance, so many softwares quote a percentage variance which is a similar concept, although it is important not to get confused by different notation.

A simple rule might be to retain PCs until the cumulative eigenvalues account for a certain percentage (e.g. 95 %) of the data, in the case of Table 5.1, this means that the first three components are significant.

More elaborate information can be obtained by looking at the size of the error matrix. The sum of squares of the matrix $\boldsymbol{E}$ is simply the difference between the sum of squares of the matrices $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$. In Table 5.1, after three components are calculated the sum of squares of $\hat{\boldsymbol{X}}$ equals 639 (or the sum of the first 3 eigenvalues). However, the sum of square of the original data $\boldsymbol{X}$ equals 670. Therefore, the sum of squares of the error matrix $\boldsymbol{E}$ equals $670 - 639$ or 31.

This is sometimes interpreted physically. For example:

- if the dataset of Table 5.1 arose from six spectra recorded at 20 wavelengths:
- the error matrix is of size $6 \times 20$, consisting of 120 elements;
- so the root mean square error is equal to $(31/120)^{1/2} = 0.508$.

Is this a physically sensible number? This depends on the original units of measurement and what the instrumental noise characteristics are. If it is known that the root mean square noise is about 0.5 units, then it seems sensible. If the noise level, however, is around 5 units, far too many PCs have been calculated, as the error is way below the noise level and so the data have been over-fitted.

These considerations can be extended, and in spectroscopy, a large number of so-called ‘indicator’ functions have been proposed, many by Malinowski, whose text on factor analysis [12] is a classic in this area. Most functions involve producing graphs of functions of eigenvalues, and predicting the number of significant components using various criteria. Over the past decade several new functions have been proposed, some based on distributions such as the $F$-test. For more statistical applications, such as quantitative structure – activity relationships, these indicator functions are not so applicable, but in spectroscopy and certain forms of chromatography where there are normally a physically defined number of factors and well understood error (or noise) distributions, such approaches are valuable.

A complementary series of methods are based on cross-validation which has been introduced previously (Section 5.7.2) in a different context of classification. When performing PCA, as an increasing number of components is calculated, for prediction of the training set (often called ‘autoprediction’) the error reduces continuously, that is the difference between the $\boldsymbol{X}$ matrix predicted by PCA and the observed matrix reduces the more the components employed. However, if the later components correspond to error, they will not predict effectively an ‘unknown’ that is left out of the original training set. Cross-validation involves predicting a portion of the dataset using information from the remainder of the samples. The residual error using cross-validation should be a minimum as the correct number of components are employed, and unlike autoprediction will increase again afterwards, because later PCs correspond to noise and will not predict the data that is left out well.

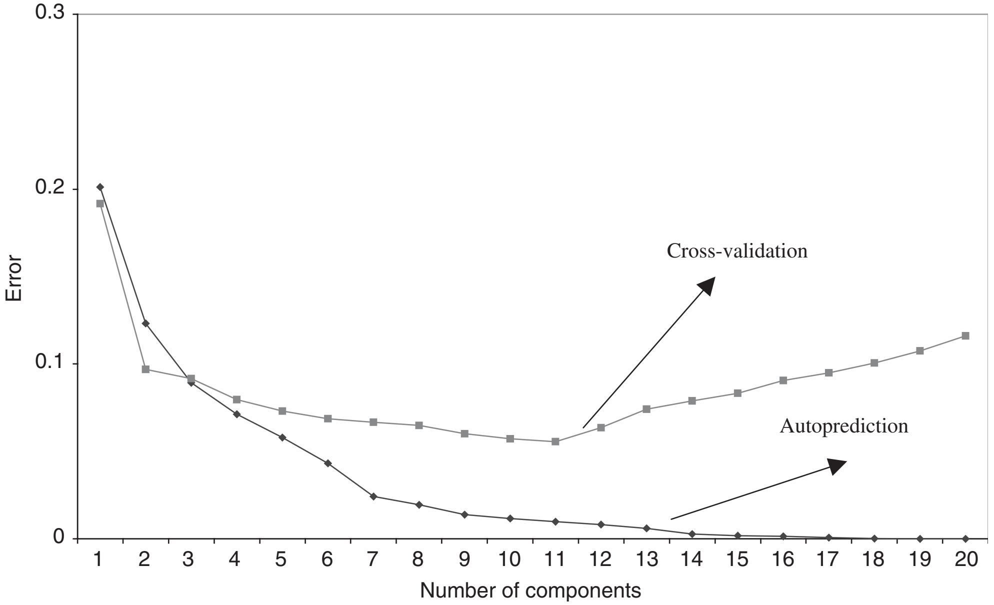


**Figure 5.31** Cross-validation and autoprediction errors

Figure 5.31 shows the autopredictive (on the full training set) and cross-validated errors for a typical dataset as an increasing number of components is calculated. Whereas the autopredictive error reduces all the time, the cross-validated error is a minimum at 11 components, suggesting that later PCs model mainly noise. Cross-validation is a good indicator of the quality of modelling whereas autoprediction often forces an unrealistically optimistic answer on a system. The cross-validated graphs are not always as straightforward to interpret. Of course are many different methods of cross-validation but the simplest (LOO) ‘leave one out’ at a time approach is normally adequate in most chemical situations. The bootstrap as discussed in section 5.7.2 in the context of PCA is an alternative but less common approach for determining the number of significant components.

Validation is very important in chemometrics and is also discussed in the context of classification in Section 5.7 and calibration in Section 6.7. It is always important to recognize that there are different motivations for validation, one being to optimize a model and the other to determine how well a model performs on an independent set of samples, and sometimes a clear head is required not to mix up these two reasons.

## 5.11 MULTIWAY PATTERN RECOGNITION

Most traditional chemometrics is concerned with two-way data, often represented by matrices. Yet over the past decade there has grown a large interest in what is often called three-way chemical data. Instead of organizing the information as a two-dimensional array [Figure 5.32(a)], it falls into a three-dimensional ‘tensor’ or box [Figure 5.32(b)]. Such datasets are surprisingly common.

Consider, for example, an environmental chemical experiment in which the concentrations of six elements are measured at 20 sampling sites on 24 days in a year. There will be

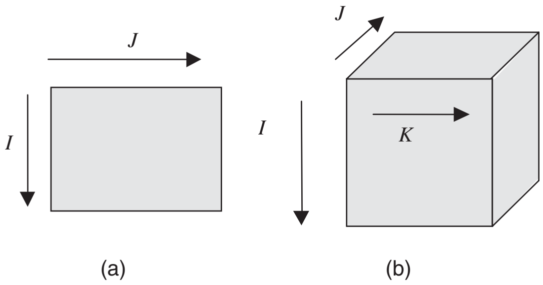


**Figure 5.32** Multiway data

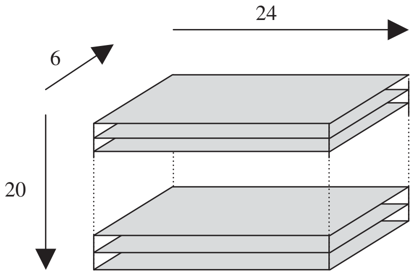


**Figure 5.33** Example of three-way data from environmental chemistry. Dimensions are elements, sampling sites and sampling days

20 × 24 × 6 or 2880 measurements, however, these can be organized as a 'box' with 20 planes each corresponding to a sampling site, and of dimensions 24 × 6 (Figure 5.33). Such datasets have been available for many years to psychologists and in sensory research. A typical example might involve a taste panel assessing 20 food products. Each food could involve the use of 10 judges who score eight attributes, resulting in a 20 × 10 × 8 box. In psychology, we might be following the reactions of 15 individuals to five different tests on 10 different days, possibly each day under slightly different conditions, so have a 15 × 5 × 10 box. These problems involve finding the main factors that influence the taste of a food or the source of pollutant or the reactions of an individual, and are a form of pattern recognition.

Three-dimensional analogies to PCs are required. The analogies to scores and loadings in PCA are not completely straightforward, so the components in each of the three dimensions are often called 'weights'.

There are a number of methods available to tackle this problem.

### 5.11.1 Tucker3 Models

These models involve calculating weight matrices corresponding to each of the three dimensions (e.g. sampling site, date and element), together with a 'core' box or array, which provides a measure of magnitude. The three weight matrices do not necessarily have the same dimensions, so the number of components for sampling sites may be different to those for date, unlike normal PCA where one of the dimensions of both the scores and

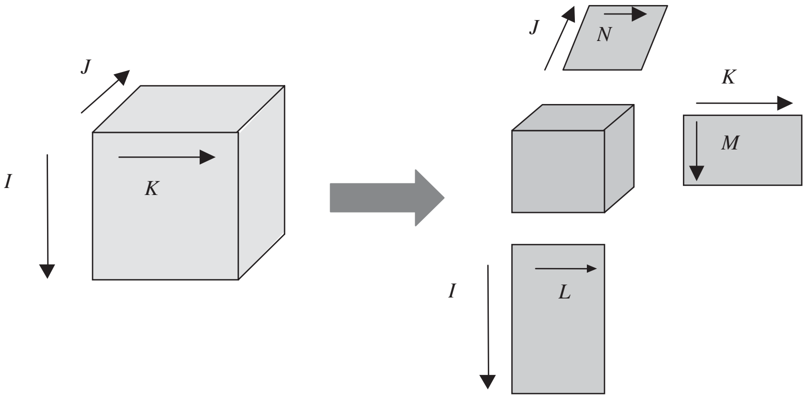


**Figure 5.34** Tucker3 models

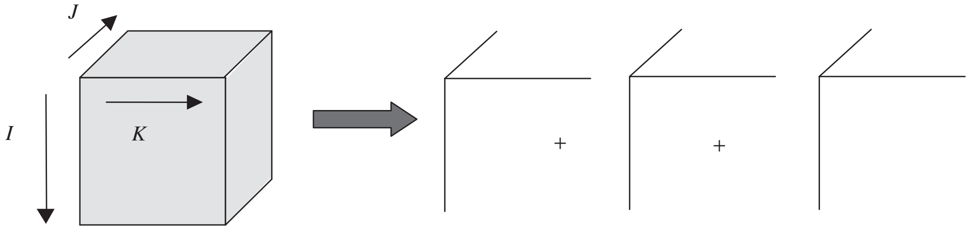


**Figure 5.35** PARAFAC models

loadings matrices must be identical. This model is represented in Figure 5.34. The easiest mathematical approach is by expressing the model as a summation:

$$x_{ijk} \approx \sum_{l=1}^{L} \sum_{m=1}^{M} \sum_{n=1}^{N} a_{il} b_{jm} c_{kn} z_{lmn}$$

where $z$ represents the core array. Some authors use the concept of 'tensor multiplication' being a three-dimensional analogy to 'matrix multiplication' in two dimensions, however, the details are confusing and it is conceptually probably best to stick to summations, which is what computer programs do well.

## 5.11.2 PARAFAC

Parallel Factor Analysis (PARAFAC) differs from Tucker3 models in that each of the three dimensions contains the same number of components. Hence, the model can be represented as the sum of contributions due to $g$ components, just as in normal PCA, as illustrated in Figure 5.35 and represented algebraically by:

$$x_{ijk} \approx \sum_{g=1}^{G} a_{ig} b_{jg} c_{kg}$$

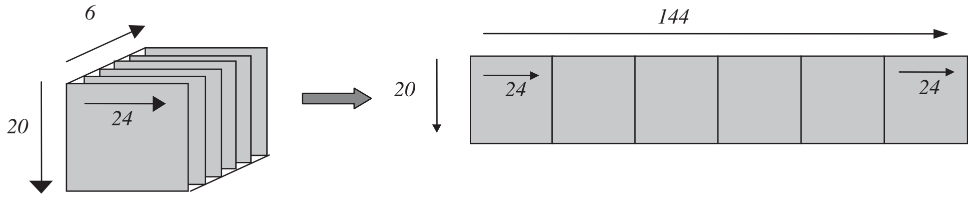


**Figure 5.36** Unfolding

Each component can be characterized by a vector that is analogous to a scores vector and two vectors that are analogous to loadings, but some keep to the notation of 'weights' in three dimensions. Components can, in favourable circumstances, be assigned a physical meaning. A simple example might involve following a reaction by recording a diode array HPLC chromatogram at different reaction times. A box whose dimensions are *reactiontime* × *elutiontime* × *wavelength* is obtained. If there are three factors in the data, this would imply three significant compounds in a cluster in the HPLC (or three significant reactants), and the weights should correspond to the reaction profile, the chromatogram and the spectrum of each compound.

PARAFAC, however, is quite difficult to use and, although the results are easy to interpret, is conceptually more complex than PCA. It can, however, lead to results that are directly relevant to physical factors, whereas the factors in PCA have a purely abstract meaning. Note that there are many complex approaches to scaling the data matrix prior to performing PARAFAC, which must be taken into account when using this approach.

## 5.11.3 Unfolding

Another approach is simply to 'unfold' the 'box' to give a long matrix. In the environmental chemistry example, instead of each sample being represented by a 24 × 6 matrix, it could be represented by a vector of length 144, each element consisting of the measurement of one element on one date, e.g. the measurement of Cd concentration on 15 July. Then a matrix of dimensions 20 (sampling sites) × 144 (variables) is produced (see Figure 5.36) and subjected to normal PCA. Note that a box can be divided into planes in three different ways (compare Figure 5.33 and Figure 5.36).

This comparatively simple approach is sometimes sufficient but the PCA calculation neglects to take into account the relationships between the variables. For example, the relationship between concentration of Cd on 15 July and that on 16 July, would be considered to be no stronger than the relationship between Cd concentration on 15 July and Hg on November 1 during the calculation of the components. However, after the calculations are performed it is still possible to regroup the loadings and sometimes an easily understood method such as unfolded PCA can be of value.

For more details, a tutorial review by Smilde is an excellent starting point in the literature [13].

## REFERENCES

1. K. Pearson, On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 (6) (1901), 559–572.

2. A.L. Cauchy, *Oeuvres, IX* (2) (1829), 172–175
3. R.J. Adcock, A problem in least squares, *Analyst*, 5 (1878), 53–54
4. H. Hotelling, Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 (1933), 417–441 and 498–520
5. P. Horst, Sixty years with latent variables and still more to come, *Chemometrics and Intelligent Laboratory Systems*, 14 (1992), 5–21
6. S. Dunkerley, J. Crosby, R.G. Brereton, K.D. Zissis and R.E.A. Escott, Chemometric analysis of high performance liquid chromatography – diode array detector - electrospray mass spectrometry of 2- and 3-hydroxypyridine, *Chemometrics and Intelligent Laboratory Systems*, 43 (1998), 89–105
7. D.L. Massart and L. Kaufman, *The Interpretation of Analytical Chemical Data by the use of Cluster Analysis*, John Wiley & Sons, Inc., New York, 1983.
8. B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993
9. S. Wold, Pattern Recognition by means of disjoint Principal Components models, *Pattern Recognition*, 8 (1976), 127–139
10. C. Albano, W.J. Dunn III, U. Edland, E. Johansson, B. Norden, M. Sjöström and S. Wold, 4 levels of Pattern Recognition, *Analytica Chimica Acta*, 103 (1978), 429–433
11. M.P. Derde and D.L. Massart, UNEQ – a disjoint modeling technique for Pattern Recognition based on normal-distribution, *Analytica Chimica Acta*, 184 (1986), 33–51
12. E.R. Malinowski, *Factor Analysis in Chemistry*, 3rd Edn, John Wiley & Sons, Inc., New York, 2002
13. A.K. Smilde, 3-way analyses – problems and prospects, *Chemometrics and Intelligent Laboratory Systems*, 15 (1992), 143–157