20 Introductory concepts of multivariate analysis

20.1 Introduction

So far, all the analyses discussed in this book have been for either univariate or bivariate data. Often, however, earth scientists need to analyze samples of **multivariate** data – where **more than two variables are measured on each sampling or experimental unit** – because univariate or bivariate data do not give enough detail to realistically describe the material or the environment being investigated.

For example, a large ore body may contain several different metals, and the concentrations of each of these may vary considerably within it. It would be useful to have a good estimate of this variation because some parts of the deposit may be particularly worth mining, others may not be worth mining at all, or certain parts may have to be mined and processed in different ways. Data for only one or two metals (e.g. copper and silver) are unlikely to be sufficient to estimate the full variation in composition and value within a deposit that also includes lead and zinc.

Samples on which multivariate data have been measured are often difficult to compare with one another because there are so many variables. In contrast, samples where only univariate data are available can easily be visualized and compared (e.g. by summary statistics such as the mean and standard error). Bivariate data can be displayed on a two-dimensional graph, with one axis for each variable. Even data for three variables can be displayed in a three-dimensional graph. But as soon as you have four or more variables, the visualization of these in a multidimensional space and comparison among samples becomes increasingly difficult. For example, Table 20.1 gives data for the concentrations of five metals at four sites. Although this is only a small data set, it is difficult to assess which sites are most similar or dissimilar. (Incidentally, you may be thinking this is a very poor sampling design, because data are only given for one sampling unit at each site. This is true, but here we are presenting a simplified data set for clarity.)

Table 20.1 The concentrations of five metals at four sites (A–D). From these raw data, it is difficult to evaluate which sites are most similar or dissimilar.

Metal	Site A	Site B	Site C	Site D
Copper	12	43	26	21
Silver	11	40	28	19
Lead	46	63	26	21
Gold	32	5	19	7
Zinc	6	40	21	38

Earth scientists need **ways of simplifying and summarizing multivariate data** to compare samples. Because univariate data are so easy to visualize, the comparison among the four sites in Table 20.1 would be greatly simplified if the data for the five metals could somehow be **reduced to a single statistic or measure**. Multivariate methods do this by reducing the complexity of the data sets while retaining as much information as possible about each sample. The following explanations are simplified and conceptual, but they do describe how these methods work.

20.2 Simplifying and summarizing multivariate data

The methods for simplifying and comparing samples of multivariate data can be divided into two groups.

(a) The first group of analyses works on the variables themselves. They **reduce the number of variables** by identifying **the ones that have the most influence upon the observed differences among sampling units** so that **relationships among the units** can be summarized and visualized more easily. These "variable-oriented" methods are often called ***R*- mode analyses.**

(b) The second group of analyses works on the sampling units. They often summarize the multivariate data by calculating **a single measure, or statistic**, that helps to **quantify differences among sampling units.** These "sample-oriented" methods are often called ***Q*-mode analyses.**

This chapter will describe an example of an *R*-mode analysis, followed by two *Q*-mode ones.

20.3 An *R*-mode analysis: principal components analysis

Principal components analysis (PCA) (which is called "principal component analysis" in some texts) is one of the oldest multivariate techniques. The mathematical procedure of PCA is complex and uses matrix algebra, but the concept of how PCA works is very easy to understand. The following explanation only assumes an understanding of the correlation between two variables (Chapter 15).

If you have a set of data where you have measured several variables on a set of sampling units (e.g. a number of sites or cores), which for PCA are often called **objects**, it is very difficult to compare them when you have data for more than three variables (e.g. the data in Table 20.1).

Quite often, however, a set of multivariate data shows a lot of **redundancy** – that is, two or more variables are **highly correlated** with each other. For example, if you look at the data in Table 20.1, it is apparent that the concentrations of copper, silver and zinc are positively correlated (when there are relatively high concentrations of copper there are also relatively high concentrations of silver and zinc and vice versa). Furthermore, the concentrations of copper, silver and zinc are also correlated with gold, but we have deliberately made these correlations negative (when there are relatively high concentrations of gold, there are relatively low concentrations of copper, silver and zinc and vice versa) because negative correlations are just as important as positive ones.

These correlations are an example of **redundancy** within the data set – because four of the five variables are well-correlated, and knowing which correlations are negative and which are positive, **you really only need the data for one of these variables** to describe differences among the sites. Therefore, you could reduce the data for these four metals down to only one (copper, silver, gold or zinc) plus lead in Table 20.2 with little loss of information about the sites.

A principal components analysis uses such cases of redundancy to reduce the number of variables in a data set, although it does not exclude variables. Instead, PCA **identifies variables that are highly correlated** with each other and combines these to **construct a reduced set of new variables that still**

Table 20.2 Because the concentrations of copper, silver, gold and zinc are correlated, you only need data for one of these (e.g. silver), plus the concentration of lead, to describe the differences among the sites.

Metal	Site A	Site B	Site C	Site D
Silver	11	40	28	19
Lead	46	63	26	21

describes the differences among samples. These new variables are called **principal components** and are listed in decreasing order of importance (beginning with the one that explains the most variation among sampling units, followed by the next greatest, etc.). With a reduced number of variables, any differences among sampling units are likely to be easier to visualize.

20.4 How does a PCA combine two or more variables into one?

This is a straightforward example where data for two variables are combined into one new variable, and we are using a simplified version of the conceptual explanation presented by Davis (2002). Imagine you need to assess variation within a large ore body for which you have data for the concentration of silver and gold at ten sites. It would be helpful to know which sites were most similar (and dissimilar) and how the concentrations of silver and gold varied among them.

The data for the ten sites have been plotted in Figure 20.1, which shows a negative correlation between the concentrations of silver and gold. This strong **relationship between two variables** can be used to **construct a single, combined variable to help make comparisons among the ten sites**. Note that you are not interested in whether the variables are positively or negatively correlated – you only want to compare the sites.

The bivariate distribution of points for these two highly correlated variables could be enclosed by a **boundary**. This is analogous to the way a set of univariate data has a 95% confidence interval (Chapter 8). For this bivariate data set the boundary will be two dimensional, and because the variables are correlated it will be elliptical as shown in Figure 20.2.

An ellipse is symmetrical and its relative length and width can be described by the length of the longest line that can be drawn through it

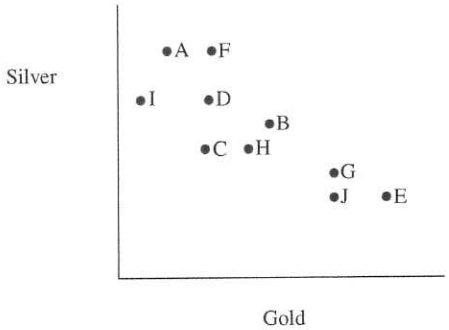


Figure 20.1 The concentration of silver versus the concentration of gold at ten sites.

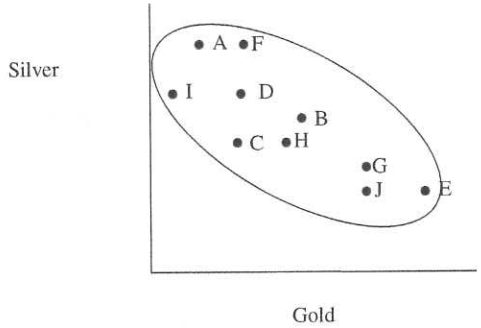


Figure 20.2 An ellipse drawn around the set of data for the concentration of silver versus the concentration of gold in ore at ten sites. The elliptical boundary can be thought of as analogous to the 95% confidence interval for this bivariate distribution.

(which is called the major axis), and the length of a line drawn halfway down and perpendicular to the major axis (which is called the minor axis) (Figure 20.3).

The relative lengths of the two axes describing the ellipse **will depend upon the strength of the correlation between the two variables**. Highly correlated data like those in Figure 20.3 will be enclosed by a long and narrow ellipse, but for weakly correlated data the ellipse will be far more circular.

At present the ten sites are described by two variables – the concentrations of silver and gold. But because these two variables are highly correlated, all the sites are quite close to the major axis of the ellipse, so most of the variation among them can be described by just that axis (Figure 20.3). Therefore, you can think of the major axis as a **new single variable** that is a good indication of

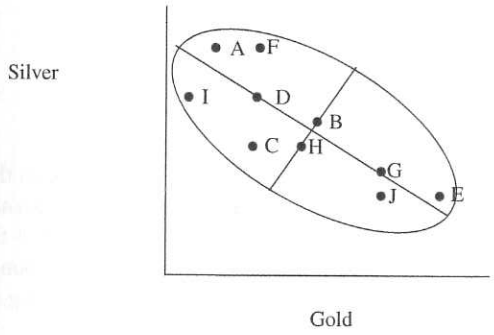


Figure 20.3 The long major axis and shorter minor axis give the dimensions of the ellipse that encloses the set of data.

most of the variation among sites. **So instead of using two variables to describe the ten sites, the information can be combined into just one.**

The two axes are called **eigenvectors** and the relative length of each that falls within the ellipse is its **eigenvalue.** Once the longest eigenvector of the ellipse has been drawn, it is rotated (in the case of Figure 20.3 this will simply be anticlockwise by about 45°) so that it becomes the new X axis (Figure 20.4). This new, **artificially constructed principal component** explains most of the variation among the ten sites. It has no name except principal component number 1 (PC1). It is important to remember that PC1 is a new variable – in this case it is a **combination** of the two variables "concentration of silver" and "concentration of gold." The plot of the points in relation to PC1 in Figure 20.4 only shows the sites in terms of this new variable – there is nothing about silver or gold in the graph.

The new X axis, PC1, is rescaled to assign the midpoint of the axis the value of zero. This makes the axis symmetrical about zero, so the objects will have both positive and negative coordinates for PC1 (Figure 20.5).

In this example, the points are all close to the major axis, so principal component 1 explains the majority of the variation among the sites, and can be used to easily assess similarities among them. From Figures 20.4 and 20.5 it is clear that sites A, I and F are more similar to each other than A is to E because the distance between the former three is much shorter.

Because there are two variables in the initial data set, principal components analysis also constructs a second component that is completely independent and uncorrelated with principal component 1. The second axis is called principal component 2 (PC2) and is simply the minor axis of the ellipse

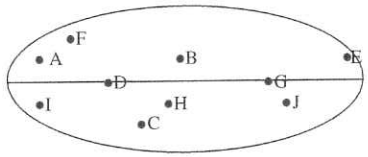


Figure 20.4 The long axis of the ellipse has been drawn through the set of highly correlated data for the concentration of silver and the concentration of gold (Figure 20.3), and then rotated to give a new X axis (which is the major axis of the ellipse) for the artificial variable called principal component number 1. This new variable explains most of the variation among sites.

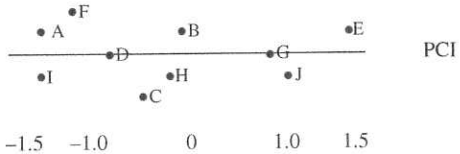


Figure 20.5 The values for PC1 are expressed in relation to the midpoint of the principal eigenvector, which is assigned the value of zero.

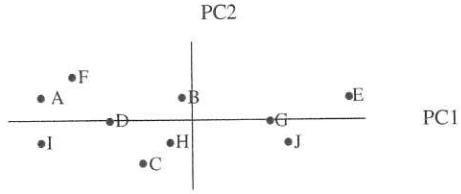
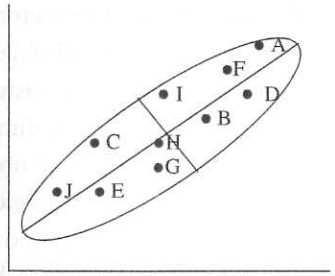


Figure 20.6 Principal component 2 is the short axis of the ellipse shown in Figure 20.5 and constructed by drawing a line perpendicular to the line showing PC1. Note that PC2 explains very little of the variation among sites.

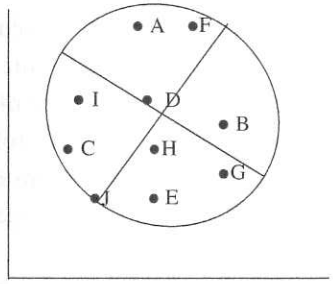
shown in Figure 20.3, which after the rotation described above will be a line perpendicular to PC1. Here too, the eigenvalue for PC2 corresponds to its relative length and its midpoint is given the value of zero. It is clear that PC2 does not explain very much of the variation among the sites – the objects are quite widely dispersed around it, so it is a relatively short eigenvector (Figure 20.6). Therefore, most of the variation is described by PC1, and the analysis has effectively reduced the number of variables from two to one.

20.5 What happens if the variables are not highly correlated?

As described above, if the two variables **are highly correlated** the ellipse enclosing the data will be very long and narrow. Therefore the first



(a)



(b)

Figure 20.7 (a) Highly correlated data. The long axis is a good indication of variation among sites. (b) Uncorrelated data. The major and minor axes of the ellipse surrounding the data points are both similar in length. Therefore neither axis is a good single summary of the variation among sites.

eigenvector will be relatively long with a large eigenvalue, and the second will be relatively short with a small eigenvalue. In this case, **by itself the new combined variable of the first eigenvector is a good indicator of the differences among sites.**

In contrast, if the two variables **are not correlated** the ellipse will be more circular and the first and second eigenvectors will both have similar eigenvalues (Figure 20.7). Therefore, **neither can be used by themselves as a good indication of the differences among sites.**

20.6 PCA for more than two variables

Principal components analysis becomes particularly useful when you have data for three or more variables.

If you have n variables a PCA will calculate n eigenvectors (with n eigenvalues) that give the dimensions of an n-dimensional object in an n-dimensional space. This may sound daunting but it is easy to visualize for only three variables, where the three eigenvectors will give the dimensions for a three-dimensional object in three-dimensional space. The object will be close to spherical for a data set with no correlations and therefore little redundancy, but a very elongated three-dimensional hyperellipsoid for a set of two or three highly correlated variables. The same applies to however many additional dimensions there are.

For three or more variables the PCA procedure is an extension of the explanation given for two variables in Section 20.4.

The longest axis of the object is found and rotated so that it becomes the X axis lying horizontally to the viewer on a two-dimensional plane with its flat surface facing the viewer (like the page you are reading at the moment). If there are many variables and therefore many dimensions, the rotation is likely to be complex – for example, an eigenvector in three dimensions may have to be rotated in both the transverse and the horizontal. The eigenvector for the longest axis then becomes principal component 1.

After this the other eigenvectors are drawn. For example, if you have measured three variables, then the three-dimensional boundary enclosing the data points will have three eigenvectors describing its length, breadth and depth, all at 90° to each other.

In many cases **several** variables may be highly correlated with each other, so the hyperellipsoid may be relatively simple and may even describe most of the variation among sites in just one or two dimensions.

Here is an example. An environmental geochemist sampled sediments along a 100 mile section of coastline, including five estuaries (A–E) that received storm water runoff from urban areas and five control estuaries (F–J) that did not. At each site, they obtained data for the concentration of copper, lead, chromium, nickel, cadmium, aluminum, mercury, zinc, total polycyclic aromatic hydrocarbons (ΣPAHs) and total polychlorinated biphenyls (ΣPCBs). These ten variables were subject to principal components analysis and re-expressed as ten principal components giving the shape of a ten-dimensional hyperellipsoid. Because several of the initial variables were highly correlated, the first principal component (PC1) **explained 70% of the variation among estuaries**. The second, PC2, explained 15% more of the variation and the third, PC3, only 5% of the

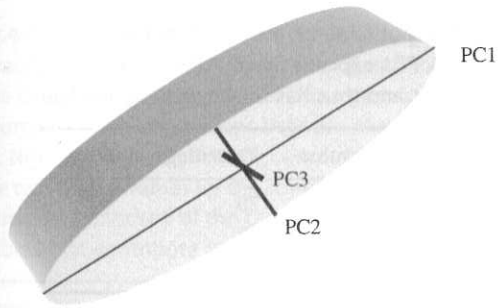


Figure 20.8 Because several variables are highly correlated they can be re-expressed as a hyperellipsoid with one very long axis (PC1), a shorter one (PC2) and a very short one (PC3). Most of the variation can be explained by PC1 and PC2. The third component, PC3, accounts for very little variation and could be ignored.

variation. Therefore, in this case 85% of the variation among site could be described by a two-dimensional ellipse with axes of PC1 and PC2, and 90% could be described by a three-dimensional ellipse with axes of PC1, PC2 and PC3. So the three-dimensional hyperellipsoid will approximate a very elongate, not very wide, and even less thick object suspended in three-dimensional space (Figure 20.8) and the remaining seven dimensions will make little contribution to its shape.

Therefore, you could take only PC1 and PC2 and plot a two-dimensional ellipse from which you can easily visualize the relationships among the sites. The two principal components explain 85% of the variation, so the closeness of the objects in two dimensions will give a realistic indication of their similarities (Figure 20.9). The analysis shows two relatively distinct clusters corresponding to the five urban and five control estuaries, consistent with urban storm water runoff having a relatively consistent effect (although you need to bear in mind that this is only a mensurative experiment).

20.7 The contribution of each variable to the principal components

Although the analysis described above has reduced the ten variables to two principal components, it is often useful to know which specific variables contribute to each of these components. For example, most of the variation (i.e. PC1) might only be related to ΣPAHs and ΣPCBs; such an outcome might suggest ways of reducing the effects of urban development upon

Table 20.3 Typical output table for only the first three components of a PCA. PC1 explains most (70%) of the variation in the data set and thus has the largest eigenvalue.

Principal component	Eigenvalue	Percentage variation
1	3.54	70
2	1.32	15
3	0.64	5

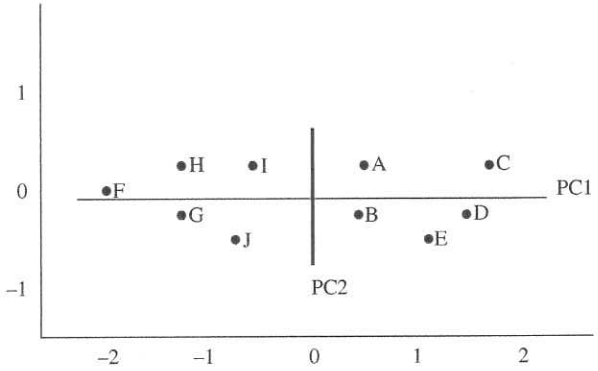


Figure 20.9 A plot of only PC1 and PC2 can still explain most of the variation among sites A–J. Note that the five urban estuaries are clustered to the right of the plot and the five control estuaries are clustered to the left.

estuaries. To address questions such as these, PCA also gives the relative contribution of each variable to each component.

The output from a PCA usually includes a plot such as Figure 20.9 and a table of **eigenvalues**. As described above, an eigenvalue gives the relative length of each eigenvector for the dimensions of the hyperellipsoid. As an example, a list of eigenvalues is given in Table 20.3, which also gives **the percentage of variation explained by each principal component**. Here too the hyperellipsoid is non-spherical, so you know the variables show redundancy and the PCA procedure has usefully reduced the number of variables.

Importantly, as well as reducing the number of variables to help visualize the relationships among objects, **PCA also gives the relative contribution of the original variables to each eigenvalue.** The output table from a PCA will contain a list of the original variables and their correlations with each of the principal components. Table 20.4 gives an example for the ten variables in the

Table 20.4 Typical output table from a PCA. The far left-hand column lists the original variables (in this case, variables 1–10) and the elements they represent. The next three columns represent the first three principal components and the values in these columns are the correlations between the new components and the original variables. Note that PC1 is primarily composed of the concentrations of variables 3 and 6 (the two largest values for the correlation coefficients and shown in bold) while PC2 is primarily composed of the concentrations of variables 1, 2 and 10 (also bold). The variables that contribute most to PC3 are 4 and 5.

Original variable	Component 1	Component 2	Component 3
1 Copper	0.01	**0.60**	0.22
2 Lead	0.24	**0.61**	0.37
3 Chromium	**0.91**	0.26	−0.06
4 Nickel	−0.18	0.32	0.57
5 Cadmium	0.15	0.05	0.52
6 Aluminum	**−0.87**	−0.22	0.44
7 Mercury	0.42	0.19	0.37
8 Zinc	0.30	−0.02	−0.22
9 ΣPAHs	−0.17	0.21	−0.06
10 ΣPCBs	0.05	**−0.71**	0.32

estuarine study described above. It is clear that principal component 1 is mainly composed of variables 3 and 6, which are chromium and aluminum (the two highest positive and negative correlations). In contrast, principal component 2 is largely composed of variables 1, 2 and 10, which are copper, lead and ΣPCBs. Which two variables make the major contribution to principal component 3? You need to look for the highest correlations, irrespective of their signs. (They are nickel and cadmium.)

The signs of the correlations are also useful. For example, for principal component 1 (Table 20.4), the correlation coefficient for variable 3 (chromium) is positive, and the one for variable 6 (aluminum) is negative. This means that as PC1 increases, chromium concentration also increases, but aluminum decreases.

In summary, a PCA has the potential to express multivariate data in a form that we can more easily understand, by reducing the number of dimensions so the data can be plotted on a two- or three-dimensional graph. It also gives a good indication of which variables contribute most to the differences among sampling units.

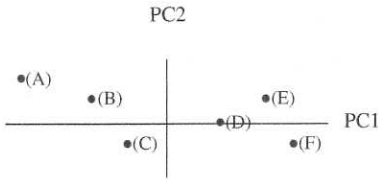


Figure 20.10 A plot of PC1 and PC2 for six sites increasingly distant (site A = closest, site F = most distant) from a petrochemical plant. The analysis shows a clear gradation through sites A to F.

20.8 An example of the practical use of principal components analysis

A marine geochemist was interested in comparing the hydrocarbons in sediments at six sampling sites, each one mile apart, running south along the shore and increasingly distant from a petrochemical plant to (a) see if there were differences in hydrocarbon levels among the sites, and (b) if so, to find out which compounds might be the best indicators of pollution.

The geochemist sampled ten hydrocarbons at each of six sites (A–F). A principal components analysis showed that only two hydrocarbons, 1 and 6 (combined as PC1), contributed to most of the variation among sites and were negatively correlated with PC1, followed by 5 and 9 (combined as PC2). When plotted on a graph of PC1 and PC2 there was a clear pattern (Figure 20.10) in that the rank order of the sites, running from left to right, corresponded to their distance from the petrochemical plant. Thus they concluded that the concentrations of only two hydrocarbons can explain most of the variation among sites.

20.9 How many principal components should you plot?

There are several ways of deciding upon how many components to use in a plot. If you are lucky, you might be in the situation where only one or two are needed, but this will only occur if they account for almost all the percentage variation among sampling units. **Generally, however, you should not use components with eigenvalues of 1.0 or less,** because this is the level of variation that you would expect by chance when there are no strong correlations among variables and therefore all original variables contribute equally to a component.

20.10 How much variation must a PCA explain before it is useful?

Very generally, if the first two or three components describe more than 70% of the variation among sampling units, then the analysis will produce a plot in two or three dimensions that is reasonably realistic.

Sometimes, however, it may be useful to know that **none** of the variables within a multivariate data set can explain very much of the variation among sampling units. For example, PCA of a multivariate data set for indicators of air pollution (nitrogen dioxide, sulfur dioxide, ozone, ammonia and the concentration of fine particles per cubic meter of air) at sites throughout a city, including the center and the fringes of the outer suburbs, showed no component with an eigenvalue greater than 0.9; none explained more than 16% of the variation among sites. The two-dimensional plot of the data was almost circular, and the three-dimensional plot was spheroidal. It was concluded that there was no obvious difference in air quality (in relation to these five indicators) across the city.

20.11 Summary and some cautions and restrictions on use of PCA

PCA is a way of reducing the complexity of a multivariate data set, but it can only do this if some variables are highly correlated. Any highly correlated variables are combined to form principal components, which may allow sampling units on which multivariate data have been measured to be plotted in two or three dimensions. The contribution of each original variable to the principal components is also given.

PCA is best suited to data where there are few zero values (e.g. grain size or concentration). It is not well suited for data such as counts, where many cells in the table of sites versus variables have a count of zero (e.g. the number of diamonds in each of several 1 m^3 sampling units of kimberlite). This restriction can be thought of in terms of the PCA constructing new axes from highly correlated variables. If the data contain a lot of zero values for each variable with only some larger numbers, the PCA is likely to overestimate redundancy, just as a group of points close to zero and a few points within a bivariate plot are likely to overestimate the strength of a correlation.

The plot provided by a PCA is also sensitive to the scale on which each variable is measured. For example, data for the concentrations of ten metals

might include rare ones measured in ng/g of sediment and more abundant ones in g/kg of sediment. This will affect the shape of the hyperellipsoid, and if the data are rescaled (e.g. all expressed as ng/g) the PCA plot will stretch or shrink to reflect this. One solution, which is often automatically applied by many PCA programs, is to **normalize** the data. This is done by converting each datum to a standard Z score, as described in Chapter 8. For each variable, every datum is subtracted from the mean and the difference divided by the standard deviation. This always gives a distribution with a mean of zero and a standard deviation of 1.0, which provides a way of standardizing the data, in just the same way that a data set was standardized for a correlation analysis in Chapter 15, Equation (15.2).

20.12 *Q*-mode analyses: multidimensional scaling

Q-mode analyses are similar to *R*-mode ones in that they also reduce the effective number of variables in a data set, but they do it in a different way.

The previous sections describe how PCA combines highly correlated **variables** in order to create fewer new ones. **In contrast, multidimensional scaling (MDS) examines the similarities among sampling units.** For example, you might have data for ten variables (e.g. the concentrations of ten different hydrocarbons) measured at each of three polluted and three unpolluted sites. As discussed in relation to principal components analysis, if you were to graph all ten variables, you would need a ten-dimensional graph that would be impossibly difficult to interpret.

Multidimensional scaling is another way of condensing multivariate information so that samples can usually be displayed on a graph with fewer dimensions than the number of variables in the original data set. This method takes the data for the original set of samples and calculates a single measure of the **dissimilarity between each of the possible pairs of these.** These dissimilarity data, which are univariate, are then used to draw a plot of the samples in two- (or three-) dimensional space. Here is a very straightforward example.

Imagine that you are interested in the spatial relationships among pegmatites within a specific magmatic system. If you were to take four different pegmatites (for now we will call them A, B, C and D) within a few adjacent counties or quadrangles and measure the distances between every possible pair of these (A–B, A–C, A–D, B–C, B–D, C–D), then you could construct the matrix shown in Table 20.5. These data indicate the **dissimilarity** between

Table 20.5 The dissimilarities, expressed as distance apart in kilometers, for four pegmatites. Those close together will have a low dissimilarity score, while for those further apart the score will be higher. Note that each pegmatite is no distance from itself. The values are duplicated (i.e. the distance between Newry and Phillips is the same as that between Phillips and Newry) and the matrix is symmetrical: you only need the similarities either above or below the diagonal showing values of zero.

	Streaked Mtn.	Mount Mica	Newry	Phillips
Streaked Mtn.	0	7	58	84
Mount Mica	7	0	50	80
Newry	58	50	0	76
Phillips	84	80	76	0

pegmatites in terms of their distance apart: pegmatites that are very close together have a low score, while those further apart have a higher one.

Knowing the dissimilarity values from the matrix you could draw at least one map showing the position of the pegmatites in two dimensions. Not all of the maps would match the actual position of the pegmatites on a real geologic map, but they would be a convenient way of visualizing the relationships among the pegmatites. Two examples are shown in Figure 20.11.

This is what multidimensional scaling does. The example using pegmatites is very simple, but if you have a matrix of dissimilarities among sampling units you can use these univariate data to position the units in only two dimensions and easily visualize how closely they are related. Those close to each other will be more similar than those further apart.

20.13 How is a univariate measure of dissimilarity among sampling units extracted from multivariate data?

Univariate measures such as the **Euclidian distance** can be used to indicate dissimilarity between sampling units for which multivariate data are available. The Euclidian distance is just the distance between any two sampling units in two-, three-, four- or higher-dimensional space.

Here is an example for only two dimensions. The length of the hypotenuse of a triangle is the square root of the sum of the squared lengths of the two other sides of the triangle (Figure 20.12). For example, for two points (A and B) in two-dimensional space, with axes of Y_1 and Y_2 and coordinates for point A of ($Y_1 = 6$, $Y_2 = 11$) and for point B of ($Y_1 = 9$, $Y_2 = 13$) the