non-zero eigenvalues a set of coordinates can be defined *via* the matrix $\mathbf{X}(n \times p)$,

$$\mathbf{X} = \mathbf{V}\Lambda^{1/2}, \tag{3.8.16}$$

where $\mathbf{\Lambda}$ is the vector of eigenvalues.

If $p = 3$, then we are working in three dimensions, and the $\mathbf{X}(n \times 3)$ matrix can be used to plot each pattern as a single point in a 3D graph. This assumes that the dimensionality of the problem can be reduced in this way while still retaining the essential features of the data. As a check, a distance matrix $\mathbf{d}^{\rm calc}$ can be calculated from $\mathbf{X}(n \times 3)$ and correlated with the observed matrix $\mathbf{d}$ using both the Pearson and Spearman correlation coefficients. In general the MMDS method works well, and correlation coefficients greater than 0.95 are common. For large data sets this can reduce to ∼0.7, which is still sufficiently high to suggest the viability of the procedure. Parallel coordinates based on the MMDS analysis can also be used, and this is discussed in Sections 3.8.4.2.1 and 3.8.4.2.2.

There are occasions in which the underlying dimensionality of the data is 1 or 2, and in these circumstances the data project onto a plane or a line in an obvious way without any problems.

An example of an MMDS plot is shown in Fig. 3.8.6(*b*), which is linked to the dendrogram in Fig. 3.8.6(*a*).

3.8.3.4. *Principal-component analysis*

It is also possible to carry out principal-component analysis (PCA) on the correlation matrix. The eigenvalues of the correlation matrix can be used to estimate the number of clusters present *via* a scree plot, as shown in Fig. 3.8.2(*a*), and the eigenvectors can be used to generate a score plot, which is an $\mathbf{X}(n \times 3)$ matrix and can be used as a visualization tool in exactly the same way as the MMDS method to indicate which patterns belong to which class. Score plots traditionally use two components with the data thus projected on to a plane; we use 3D plots in which three components are represented. In general, we find that the MMDS representation of the data is nearly always superior to the PCA analysis for powder and spectroscopic data.

3.8.3.5. *Choice of clustering method*

It is possible to use the MMDS plot (or, alternatively, PCA score plots) to assist in the choice of clustering method, since the two methods operate semi-independently. The philosophy here is to choose a technique that results in the tightest, most isolated clusters as follows:

(1) The MMDS formalism is used to derive a set of three-dimensional coordinates stored in matrix $\mathbf{X}(n \times 3)$.
(2) The number of clusters, c, is estimated as described in Section 3.8.3.2.
(3) Each of six dendrogram methods (see Table 3.8.1) is employed in turn, stopping when c clusters have been generated. Each entry in $\mathbf{X}$ can now be assigned to a cluster.
(4) A sphere is drawn around each point in $\mathbf{X}$ and the average between-cluster overlap of the spheres is calculated for each of the N clusters C_1 to C_N. If the total number of overlaps is m, this can be written as

$$S = \sum_{i=1}^{n} \sum_{\substack{j=1,n \\ j\neq i}}^{n} \left(\int_V s_{i\in C_i} s_{j\in C_j} \, {\rm d}s \right) \Big/ m. \tag{3.8.17}$$

If the clusters are well defined then S should be a minimum. Conversely, poorly defined clusters will tend to have large values of S. In the algorithm used in *PolySNAP* (Barr, Dong & Gilmore, 2009) and *DIFFRAC.EVA* (Bruker, 2018), the sphere size depends on the number of diffraction patterns.
(5) The tightness of each cluster is also estimated by computing the mean within-cluster distance. This should also be a minimum for well defined, tight clusters.
(6) The mean within-cluster distance from the centroid of the cluster can also be computed, which should also be a minimum.
(7) Steps (4)–(6) are repeated using coordinates derived from PCA 3D score plots.
(8) Tests (4)–(7) are combined in a weighted, suitably scaled mean to give an overall figure of merit (FOM); the minimum is used to select which dendrogram method to use (Barr *et al.*, 2004*b*).

3.8.3.6. *The most representative sample*

Similar techniques can be used to identify the most representative sample in a cluster. This is defined as the sample that has the minimum mean distance from every other sample in the clusters, *i.e.* for cluster J containing m patterns, the most representative sample, i, is defined as that which gives

$$\min \left(\sum_{\substack{j=1 \\ i,j\in J}}^{m} d(i,j)/m \right). \tag{3.8.18}$$

The most representative sample is useful in visualization and can, with care, be used to create a database of known phases (Barr *et al.*, 2004*b*).

3.8.3.7. *Amorphous samples*

Amorphous samples are an inevitable consequence of high-throughput experiments, and need to be handled correctly if they are not to lead to erroneous indications of clustering. To identify amorphous samples the total background for each pattern is estimated and its intensity integrated; the integrated intensity of the non-background signal is then calculated. If the ratio falls below a preset limit (usually 5%, but this may vary with the type of samples under study) the sample is treated as amorphous. The distance matrix is then modified so that each amorphous sample is given a distance and dissimilarity of 1.0 from every other sample, and a correlation coefficient of zero. This automatically excludes the samples from the clustering until the last amalgamation steps, and also limits their effect on the estimation of the number of clusters (Barr *et al.*, 2004*b*). Of course, the question of amorphous samples is not a binary (yes/no) one: there are usually varying degrees of amorphous content, which further complicates matters.

3.8.4. Data visualization

3.8.4.1. *Primary data visualization*

It is important when dealing with large data sets to have suitable visualization tools. These tools are also a valuable resource for exploring smaller data sets. This methodology provides four primary aids:

(1) A pie chart is produced for each sample, corresponding to the sample wells used in the data-collection process, in which each well is given a colour as defined by the dendrogram. If mixtures of known phases are detected, the pie charts give the relative proportions of the pure samples as estimated by quantitative analysis (see Section 3.8.7).

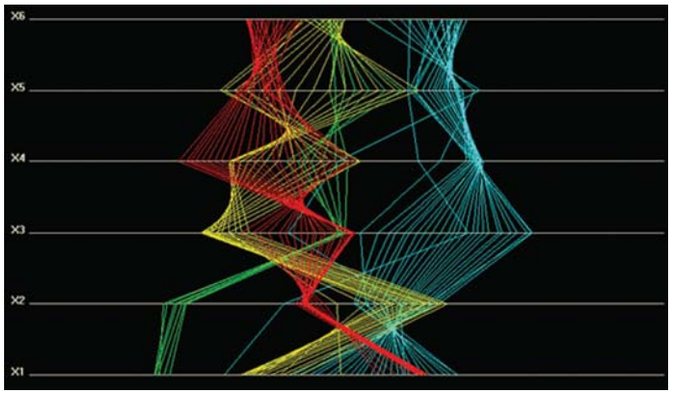

Figure 3.8.3
Example of a parallel-coordinates plot in six dimensions, with axes labeled X1, X2, ..., X6, for a set of 80 organic PXRD samples partitioned into four clusters. The plot shows that the clustering looks realistic and that it is maintained when the data are examined in six dimensions.

(2) The dendrogram gives the clusters, the degree of association within the clusters and the differential between a given cluster and its neighbours. Different colours are used to distinguish each cluster. The cut line is also drawn along with the associated confidence levels. The dendrogram is the primary visualization tool.
(3) The MMDS method reproduces the data as a 3D plot in which each point represents a single powder pattern. The colour for each point is taken from the dendrogram. The most representative sample for each cluster is marked with a cross.
(4) Similarly, the eigenvalues from principal-component analysis can be used to generate a 3D score plot in which each point also represents a powder pattern. Just as in the MMDS formalism, the colour for each point is taken from the dendrogram, and the most representative sample is marked with a cross.

These aids give graphical views of the data that are semi-independent and thus can be used to check for consistency and discrepancies in the clustering. They are also interactive. No one method is optimal, and a combination of mathematical and visualization techniques is required, techniques that often need tuning for each individual application (Barr, Cunningham *et al.*, 2009; Barr, Dong & Gilmore, 2009).

3.8.4.2. *Secondary visualization using parallel coordinates, the grand tour and minimum spanning trees*

In the MMDS and PCA methods $p = 3$ [equation (3.8.16)] to work in three dimensions; the **X** matrix can then be used to plot each pattern as a single point in a 3D graph. However, this has reduced the dimensionality of the data to three, and the question arises as to the validity of this: are three dimensions sufficient? The use of parallel-coordinates plots coupled with the grand tour can assist here as well as giving us an alternative view of the data.

3.8.4.2.1. *Parallel-coordinates plots*

A parallel-coordinates plot is a graphical data-analysis technique for plotting multivariate data. Usually orthogonal axes are used when doing this, but in parallel-coordinates plots orthogonality is abandoned and replaced with a set of N equidistant parallel axes, one for each variable and labelled X1, X2, X3, ..., XN (Inselberg, 1985, 2009; Wegman, 1990). Each data point is plotted on each axis and the points are joined *via* a line connecting each data point. The data now become a set of lines. The lines are given the colours of the cluster to which they belong as defined by the current dendrogram. A parallel-coordinates display can be interpreted as a generalization of a two-dimensional scatterplot, and it allows the display of an arbitrary number of dimensions. The method can also be used to validate the clustering itself without using dendrograms. Using this technique it is possible to determine whether the clustering shown by the MMDS (or PCA) plot in three dimensions continues in higher dimensions.

Fig. 3.8.3 shows a typical example for a set of 80 organic samples partitioned into four clusters (Barr, Dong & Gilmore, 2009). The plot shows that the clustering looks realistic when

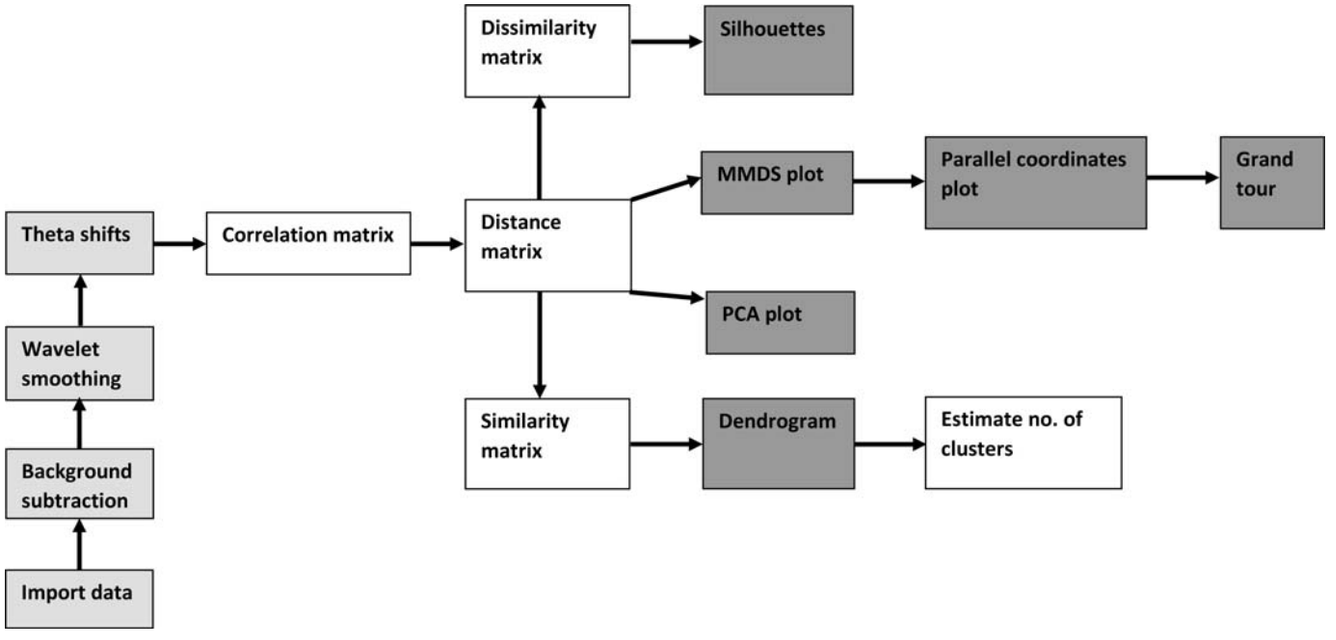

Figure 3.8.4
Flowchart for the cluster-analysis and data-visualization procedure described in this chapter. The light grey boxes denote data-visualization elements and the dark grey objects are optional data pre-processing operations.

viewed in this way and that it is maintained when the data are examined in six dimensions.

3.8.4.2.2. *The grand tour*

The grand tour is a method of animating the parallel-coordinates plot to examine it from all possible viewpoints. Consider a 3D data plot using orthogonal axes: a grand tour takes 2D sections through these data and displays them in parallel-coordinates plots in a way that explores the entire space in a continuous way. The former is important, because the data can be seen from all points of view, and the latter allows the user the follow the data without abrupt discontinuities. This concept was devised by Asimov (1985) and further developed by Wegman (1990). In more than three dimensions it becomes a generalized rotation of all the coordinate axes. A d-dimensional tour is a continuous geometric transformation of a d-dimensional coordinate system such that all possible orientations of the coordinate axes are eventually achieved. The algorithm for generating a smooth and complete view of the data is described by Asimov (1985).

To do this, the restriction of $p = 3$ in the MMDS calculation is relaxed to 6, so that there is now a 6D data set with six orthogonal axes. The choice of six is somewhat arbitrary – more can be used, but six is sufficient to see whether the clustering is maintained without generating unduly complex plots and requiring extensive computing resources. The data are plotted as a parallel-coordinates plot. The grand-tour method is then applied by a continuous geometric transformation of the 6D coordinate system such that all possible orientations of the axes are achieved. Each orientation is reproduced as a parallel-coordinates plot using six axes.

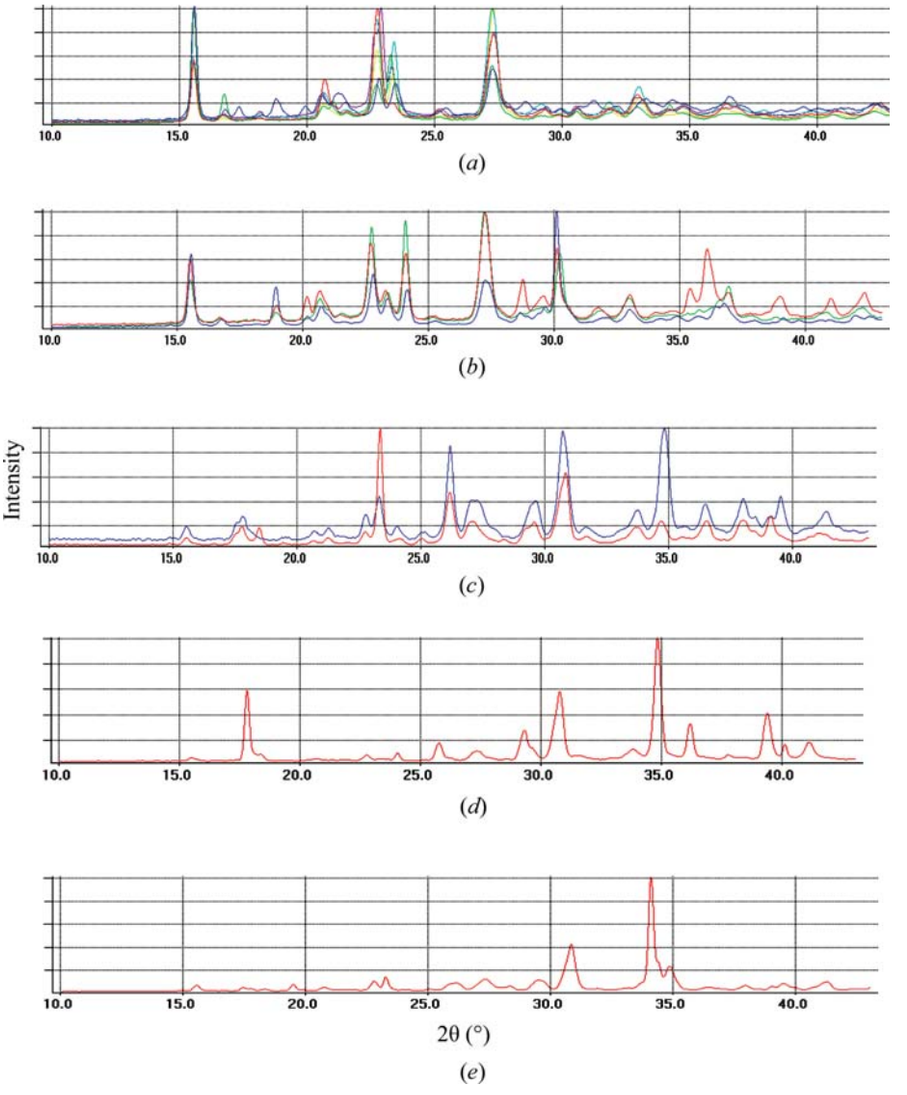

Figure 3.8.5
Powder patterns for 13 commercial aspirin samples partitioned into five sets. The patterns are in highly correlated sets: (*a*) comprises patterns 1, 3, 5, 6, 9 and 12; (*b*) comprises patterns 10, 11 and 13; (*c*) contains patterns 2 and 4; (*d*) contains pattern 7 and (*e*) contains pattern 8.

Figs. 3.8.9(*j*) and (*k*) show an example from the clustering of the 13 aspirin samples using PXRD data. Fig. 3.8.9(*j*) shows the default parallel-coordinates plot. Fig. 3.8.9(*k*) shows alternative views of the data taken from the grand tour. In Fig. 3.8.9(*j*) there appears to be considerable overlap between clusters in the 4th, 5th and 6th dimensions (X4, X5 and X6), but the alternative view given in Fig. 3.8.9(*k*) show that the clustering is actually well defined in all six dimensions (Barr, Dong & Gilmore, 2009).

3.8.4.2.3. *Powder data as a tree: the minimum spanning trees*

The minimum spanning tree (MST) displays the MMDS plot as a tree whose points are the data from the MMDS calculation (in three dimensions) and whose weights are the distances between these points. The minimum-spanning-tree problem is that of joining the points with a minimum total edge weight. (As an example, airlines use minimum spanning trees to work out their basic route systems: the best set of routes taking into account airport hubs, passenger numbers, fuel costs *etc.* is the minimum spanning tree.) Because a tree is used, each point is only allowed a maximum of three connections to other points.

To do this Kruskal's (1956) algorithm can be used, in which the lowest weight edge is always added to see if it builds a spanning tree; if so, it is added or otherwise discarded. This process continues until the tree is constructed. An example is shown in Figs. 3.8.7 for the 13-sample aspirin data. A complete tree for this data set using three dimensions and the MMDS-derived coordinates is shown in Fig. 3.8.7(*a*). This has 12 links between the 13 data points. Reducing the number of links to 10 gives Fig. 3.8.7(*b*).

3.8.5. Further validating and visualizing clusters: silhouettes and fuzzy clustering

Other techniques exist to validate the clusters, and these are discussed here.

3.8.5.1. *Silhouettes*

Silhouettes (Rousseeuw, 1987; Kaufman & Rousseeuw, 1990) are a property of every member of a cluster and define a coef-