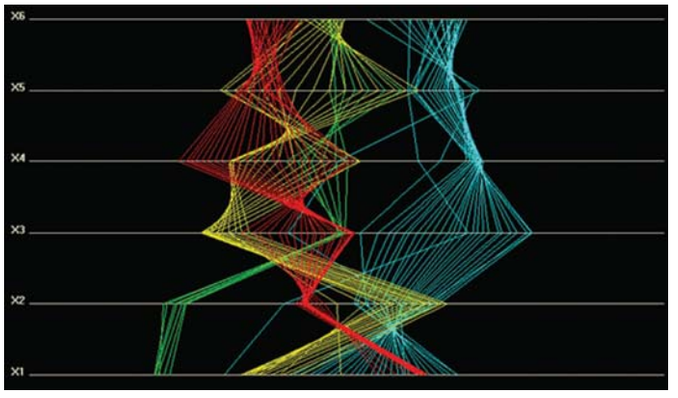

Figure 3.8.3
Example of a parallel-coordinates plot in six dimensions, with axes labeled X1, X2, ..., X6, for a set of 80 organic PXRD samples partitioned into four clusters. The plot shows that the clustering looks realistic and that it is maintained when the data are examined in six dimensions.

(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

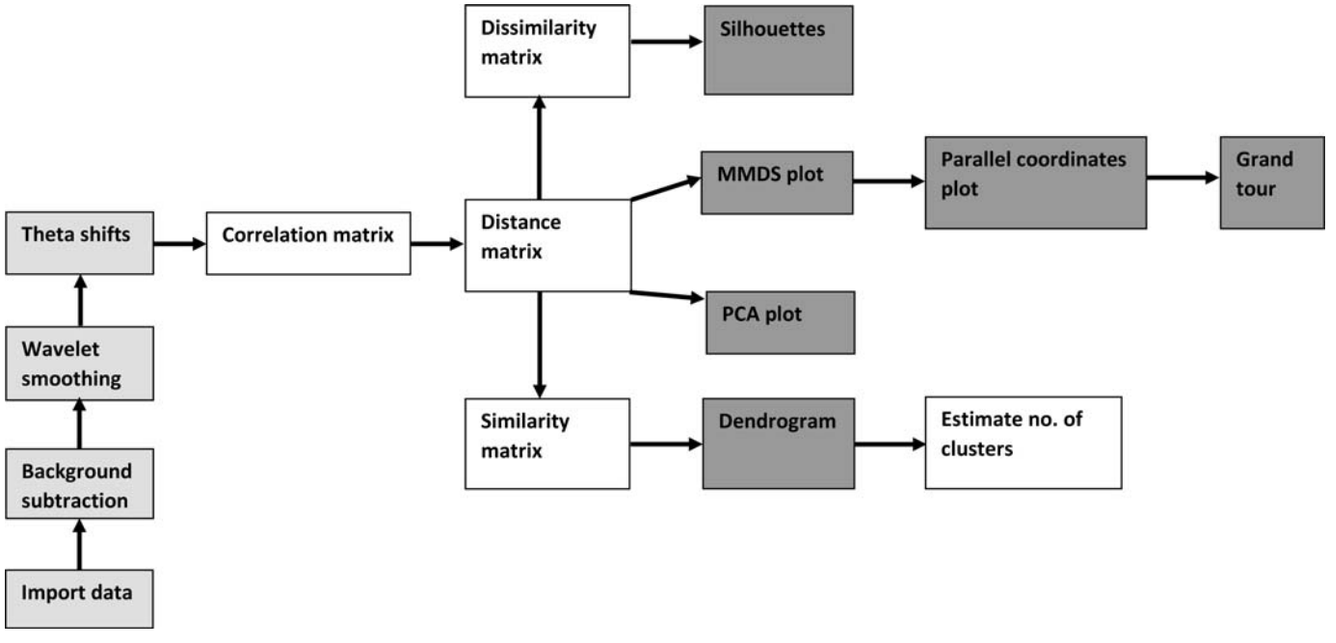

Figure 3.8.4
Flowchart for the cluster-analysis and data-visualization procedure described in this chapter. The light grey boxes denote data-visualization elements and the dark grey objects are optional data pre-processing operations.