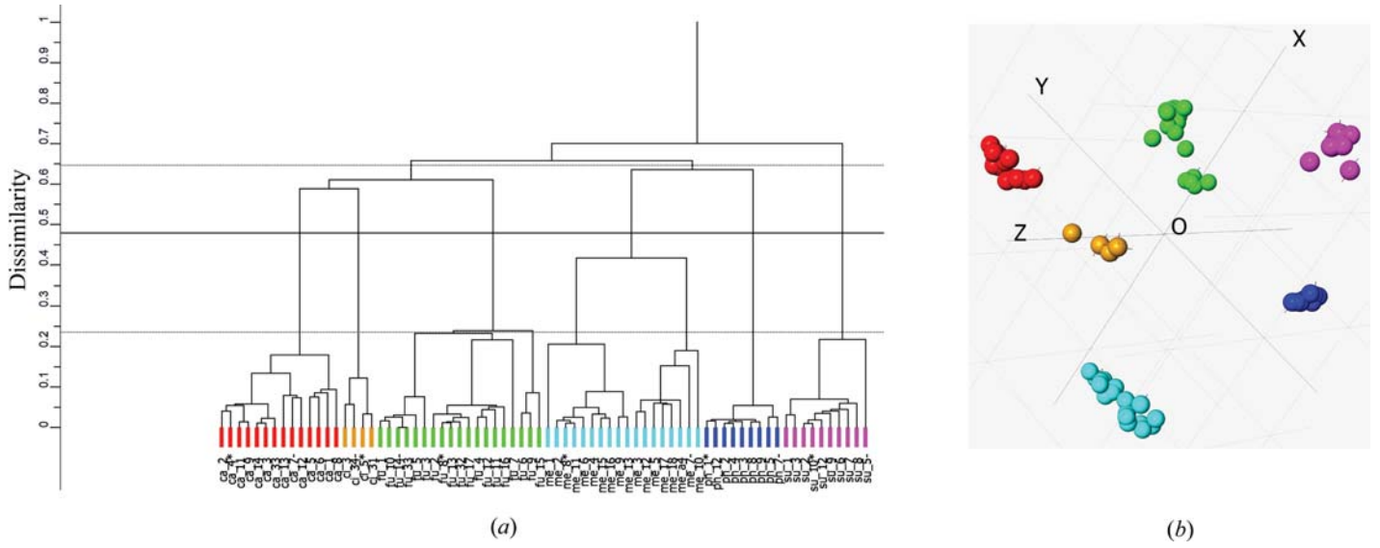

Figure 3.8.13
(*a*) The dendrogram generated from 74 Raman spectra without background corrections applied. Labelling from the left-hand side, the red samples are carbamazepine, the orange are cimetidene, the green are two forms of furosemide, the light blue is mefenamic acid, the dark blue is phenilbutazone and the purple at the right-hand side is sulfamerazine. (*b*) The MMDS plot. The sphere colours are taken from the dendrogram. This representation shows clearly discrete clusters in correspondence with those generated by the dendrogram.

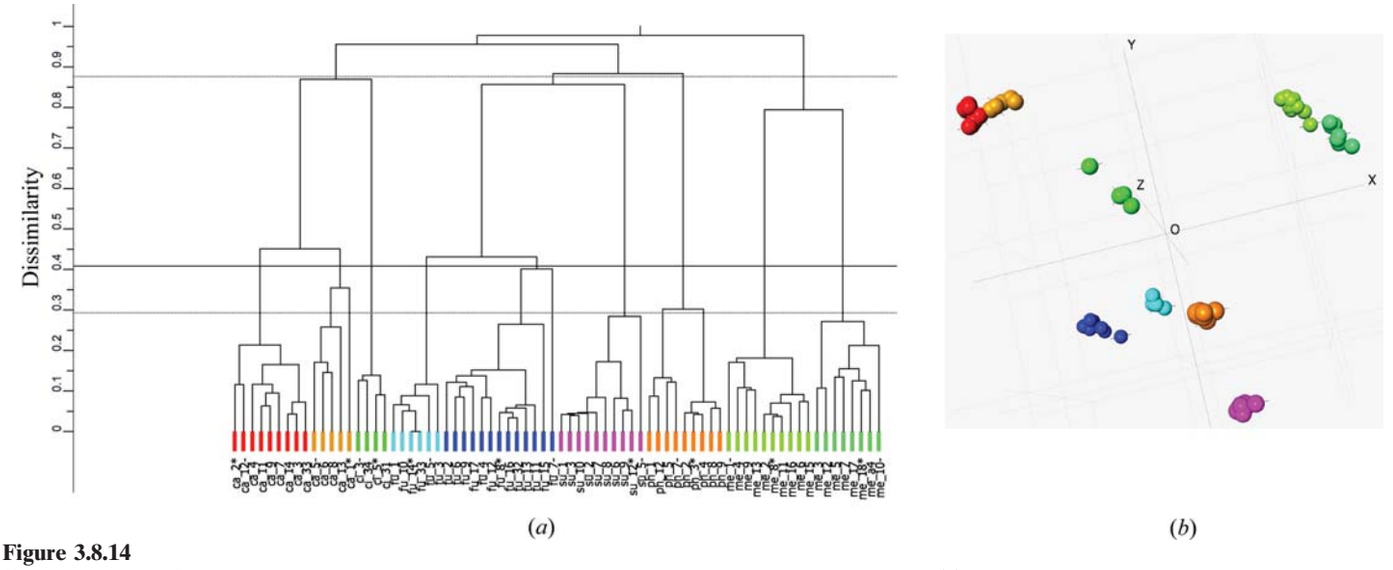

Figure 3.8.14
Clustering the 74 Raman spectra without background corrections applied using first-derivative data. (*a*) The dendrogram. Labelling from the left-hand side, the red and orange entries are carbamazepine; the green are cimetidene; the light blue and dark blue are two forms of furosemide; the purple are sulfamerazine; the brown are phenilbutazone and the right-hand light and dark green are two forms of mefenamic acid. (*b*) The MMDS plot. The clusters are well defined but the orange and red (both carbamazepine) are very close to each other.

For further details of this method with organic samples, see Dong *et al.* (2008).

3.8.8. Using spectroscopic data

There is no reason why the methodology described in this chapter cannot be used for other 1D data sets, *e.g.* Raman, IR, NMR and near-IR spectroscopies, although different data pre-processing is usually required. Raman spectroscopy is well suited to high-throughput screening: good-quality spectra can be collected in a few minutes, and sample preparation is straightforward and flexible, although the resulting spectra are not always as distinct as the PXRD equivalents (Mehrens *et al.*, 2005; Boccaleri *et al.*, 2007).

As an example we show the results of cluster analysis carried out on samples of carbamazepine, cimetidene, furosemide, mefenamic acid, phenilbutazone and sulfamerazine using Raman spectroscopy. A total of 74 samples were measured on a LabRam HR-800/HTS-Multiwell spectrometer at room temperature, equipped with a backscattering light path system of a light-emitting diode laser (785 nm, 300 mW) as an excitation source and an air-cooled charge-coupled device detector. A 20-fold superlong working distance objective lens was used to collect the backscattered light. The spectra were acquired with 5.84 cm^{-1} spectral width and at least 30 s exposure (Kojima *et al.*, 2006). The spectra had backgrounds subtracted but no other corrections were carried out.

The initial clustering is shown in Fig. 3.8.13(*a*) with the default cut level in the dendrogram. There are six clusters: labelling from the left-hand side, the red are three polymorphs of carbamazepine; the orange are cimetidene; the green cluster contains three polymorphs of furosemide; the light blue contains three poly-

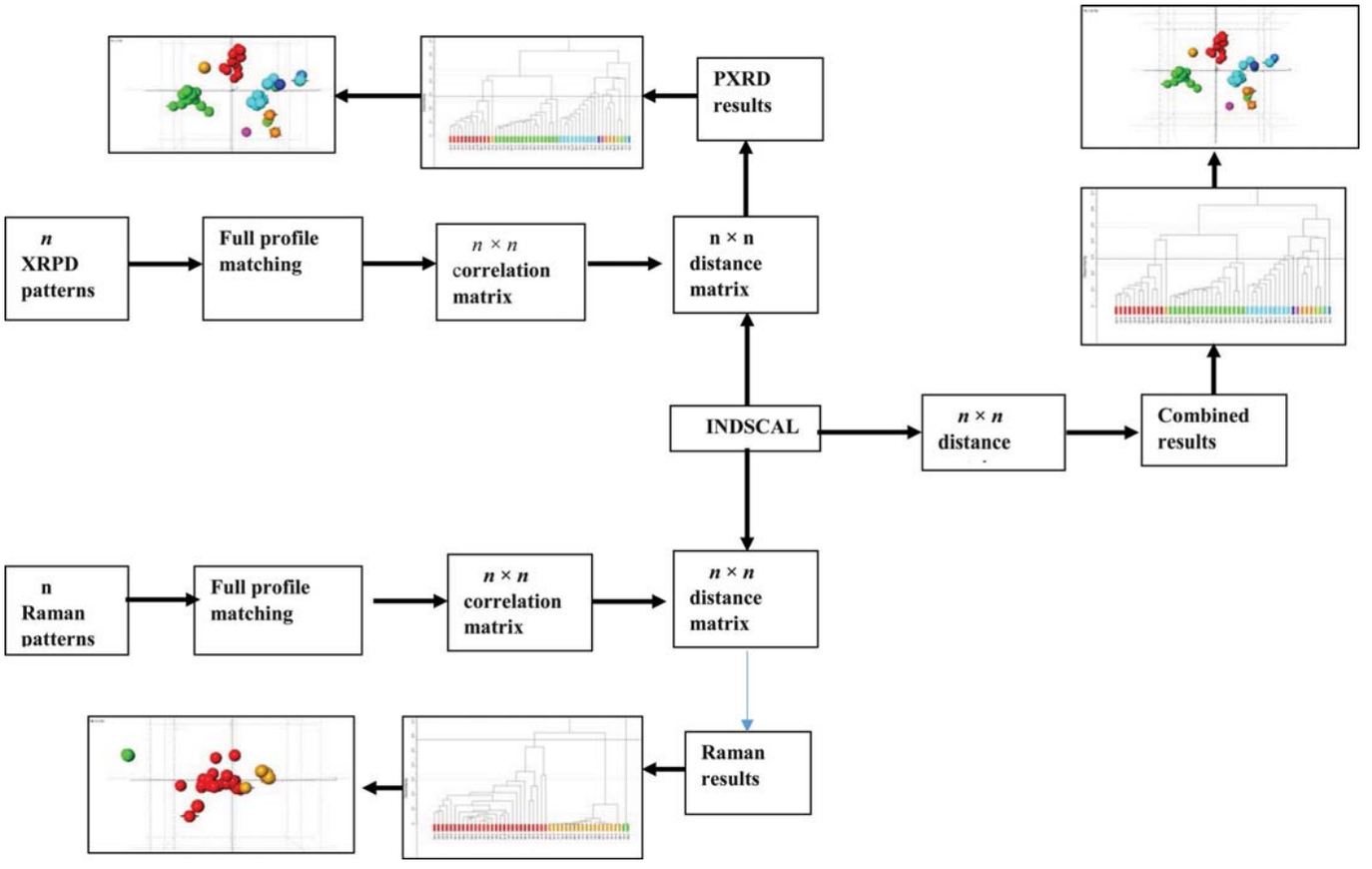

Figure 3.8.15
A flowchart for the INDSCAL method using Raman and PXRD data. Note that any combination of any 1D data can be used here.

morphs of mefenamic acid; the dark blue contains phenilbutazone; and finally the purple cluster contains sulfamerazine. The MMDS plot gives a complementary visualization of the data that supports the clustering.

It is also possible to use derivative data in place of the original spectra for clustering. The results of this for the 74 Raman spectra without initial background subtraction followed by the generation of first-derivative data are shown in Fig. 3.8.14. The clusters are well defined but now the carbamazepine data have split into two clusters. These correspond to forms I and III of carbamazepine, although the differences in the Raman spectra for these three species are small (O'Brien *et al.*, 2004). At the same time, both furosemide and mefenamic acid are each split into two groups. This is probably the best description of the data in terms of clustering and cluster membership corresponding to the chemical differences in the samples. The dendrogram also has the feature that the tie bars between samples are higher, *i.e.* the similarities are lower, reflecting the fact that the use of first derivatives accentuates small differences in the data.

It is interesting to note that, in general, PXRD works less well with derivative data. The reason for this is not clear, but possibly the presence of partial overlapping peaks and the associated issues of peak shape are partly responsible.

3.8.9. Combining data types: the INDSCAL method

It is now common to collect more than one data type, and some instruments now exist for collecting spectroscopic and PXRD data on the same samples, for example the Bruker D8 Screenlab, which combines PXRD and Raman measurement for high-throughput screening (Boccaleri *et al.*, 2007).

A technique for combining the results of more than one data type is needed. One method would be to take individual distance matrices from each data type and generate an average distance matrix using equation (3.8.3), but this leaves open the question of how best to define the associated weights in an optimal, objective way. Should, for example, PXRD be given a higher weight than Raman data? The individual differences scaling method (INDSCAL) of Carroll & Chang (1970) provides an unbiased solution to this problem by, as the name suggests, scaling the differences between individual distance matrices.

In this method, let $\mathbf{D}_k$ be the squared distance matrix of dimension $(n \times n)$ for data type k with a total of K data types. For example, if we have PXRD, Raman and differential scanning calorimetry (DSC) data for each of n samples, then $K = 3$. A group-average matrix $\mathbf{G}$ (which we will specify in two dimensions) is required that best represents the combination of the K data types. To do this, the $\mathbf{D}$ matrices are first put into inner-product form by the double-centring operation to give

$$\mathbf{B}_k = -\tfrac{1}{2}(\mathbf{I} - \mathbf{N})\mathbf{D}_k(\mathbf{I} - \mathbf{N}), \tag{3.8.33}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{N}$ is the centring matrix $I - \mathbf{1}\mathbf{1}'/N$; $\mathbf{1}$ is a column vector of ones. The inner-product matrices thus generated are matched to the weighted form of the group average, $\mathbf{G}$, which is unknown. To do this the function

$$S = \sum_{1}^{K} \left\| \mathbf{B}_k - \mathbf{G}\mathbf{W}_k^2\mathbf{G}' \right\| \tag{3.8.34}$$

is minimized. The weight matrices, $\mathbf{W}_k$, are scaled such that