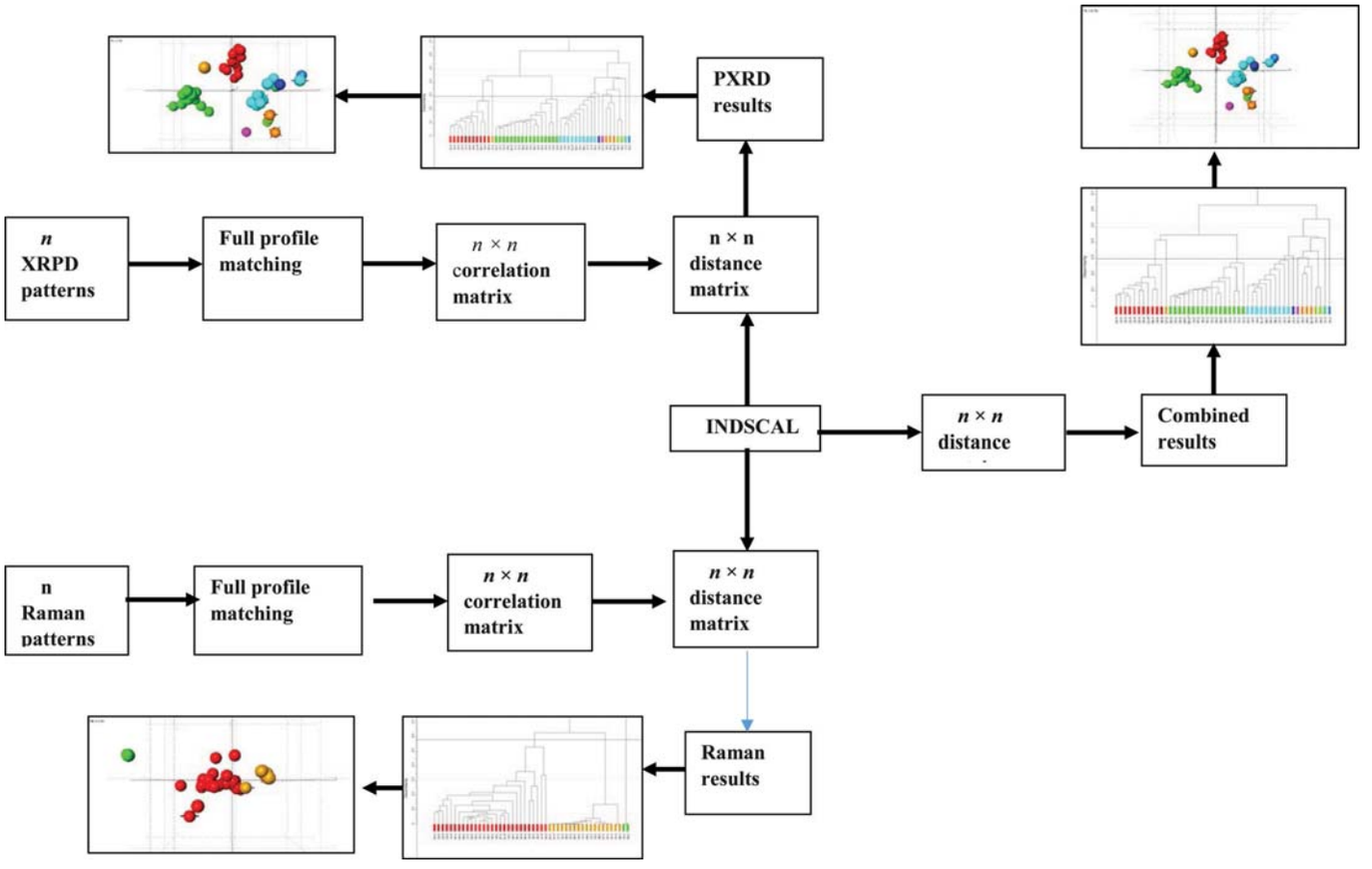

Figure 3.8.15
A flowchart for the INDSCAL method using Raman and PXRD data. Note that any combination of any 1D data can be used here.

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