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

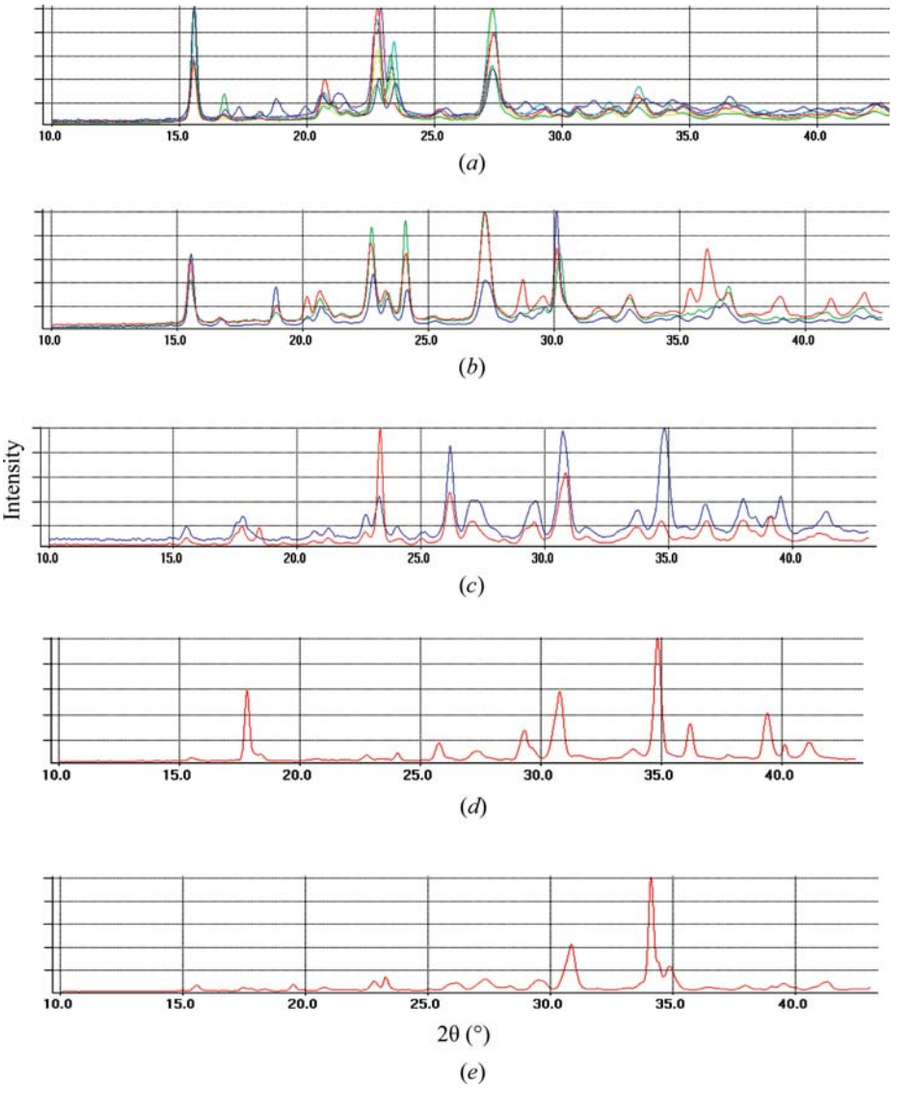

Figure 3.8.5
Powder patterns for 13 commercial aspirin samples partitioned into five sets. The patterns are in highly correlated sets: (*a*) comprises patterns 1, 3, 5, 6, 9 and 12; (*b*) comprises patterns 10, 11 and 13; (*c*) contains patterns 2 and 4; (*d*) contains pattern 7 and (*e*) contains pattern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

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.

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-