International Tables for Crystallography (2018). Vol. H, Section 3.8.6, pp. 333–337.

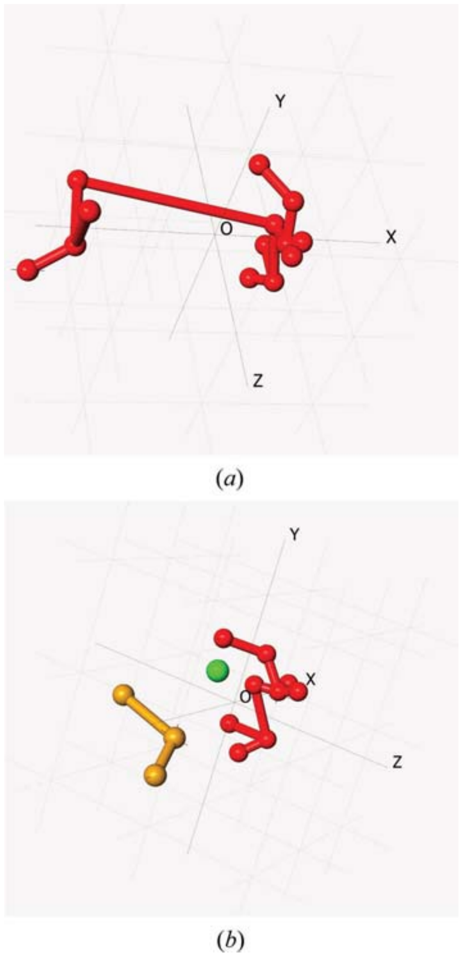


Figure 3.8.7
The use of minimum spanning trees (MSTs). (*a*) The MST with 12 links. (*b*) The MST with 10 links; three clusters are now present.

$$h_i = \frac{b_i - a_i}{\max(a_i,\ b_i)}. \qquad (3.8.21)$$

Clearly $-1 \leq h_i \leq 1.0$. It is not possible to define silhouettes for clusters with only one member (singleton clusters). Silhouettes are displayed such that each cluster is represented as a histogram of frequency plotted against silhouette values so that one can look for outliers or poorly connected plots.

From our experience with powder data collected in reflection mode on both organic and inorganic samples (Barr *et al.*, 2004*b*), we conclude that for any given pattern

(1) $h_i > 0.5$ implies that pattern i is probably correctly classified;
(2) $0.2 < h_i < 0.5$ implies that pattern i should be inspected, since it may belong to a different or new cluster;
(3) $h_i < 0.2$ implies that pattern i belongs to a different or new cluster.

The use of silhouettes in defining the details of the clustering is shown for the aspirin data in Fig. 3.8.8. The silhouettes for the red cluster corresponding to the dendrogram in Fig. 3.8.6(*a*) are shown in Fig. 3.8.8(*a*) and those for the corresponding orange cluster are shown in Fig. 3.8.8(*b*). Both sets of silhouettes have values < 0.5, which indicates that the clustering is not optimally defined. When the cut line is moved to give the dendrogram in Fig. 3.8.6(*c*), the silhouettes for the red cluster are shown in Fig. 3.8.8(*c*). The entry centred on a silhouette value of 0.15 is pattern 3. This implies that pattern 3 is only loosely connected to the cluster and this is demonstrated in Fig. 3.8.8(*d*) where pattern 3 and the most representative pattern for the cluster (No. 9) are superimposed. Although there is a general sense of similarity there are significant differences and the combined correlation coefficient is only 0.62. In Fig. 3.8.8(*e*), the silhouettes for the orange cluster are shown. They imply that this is a single cluster without outliers. The silhouettes for the green cluster corresponding to the dendrogram in Fig. 3.8.6(*c*) are shown in Fig. 3.8.8(*f*). The clustering is poorly defined here.

3.8.5.2. *Fuzzy clustering*

In standard clustering methods a set of n diffraction patterns are partitioned into c disjoint clusters. Cluster membership is defined *via* a membership matrix $\mathbf{U}(n \times c)$, where individual coefficients, u_{ik}, represent the membership of pattern i of cluster k. The coefficients are equal to unity if i belongs to c and zero otherwise, *i.e.*

$$u_{ik} \in [0, 1] \quad (i = 1, \ldots, n;\, k = 1, \ldots, c). \qquad (3.8.22)$$

If these constraints are relaxed, such that

$$0 \leq u_{ik} \leq 1 \quad (i = 1, \ldots, n;\, k = 1, \ldots, c), \qquad (3.8.23)$$

$$0 < \sum_{i=1}^{n} u_{ik} < n \quad (k = 1, \ldots, c) \qquad (3.8.24)$$

and

$$\sum_{k=1}^{c} u_{ik} = 1, \qquad (3.8.25)$$

then fuzzy clusters are generated, in which there is the possibility that a pattern can belong to more than one cluster (see, for example, Everitt *et al.*, 2001; Sato *et al.*, 1966). Such a situation is quite feasible in the case of powder diffraction, for example, when mixtures can be involved. It is described in detail by Barr *et al.* (2004*b*).

3.8.5.3. *The PolySNAP program and DIFFRAC.EVA*

All these techniques have been incorporated into the *PolySNAP* computer program (Barr *et al.*, 2004*a*,*b*,*c*; Barr, Dong, Gilmore & Faber, 2004; Barr, Dong & Gilmore, 2009), which was developed from the *SNAP-D* software (Barr, Gilmore & Paisley, 2004). *PolySNAP* has subsequently been incorporated into the Bruker *DIFFRAC.EVA* program (Bruker, 2018), and the following sections are based on its use.

3.8.6. Examples

All the elements for clustering and visualization are now in place. Fig. 3.8.4 shows this as a flowchart. Hitherto we have looked at elements of the aspirin data to demonstrate how methods work; we now examine the aspirin data in detail as a single analysis.

3.8.6.1. *Aspirin data*

In this example we use 13 powder patterns from commercial aspirin samples collected in reflection mode on a Bruker D8 diffractometer. Since these samples include fillers, the active pharmaceutical ingredient (API) and other formulations, it is not surprising that peak widths are high: ~0.5° full width at half maximum (FWHM). The data-collection range was 10–43° in 2θ using Cu $K\alpha$ radiation. The 13 powder data sets are shown in Fig. 3.8.5 arranged into groups based on similarity. We have already described the methods of analysis and have shown typical results in Figs. 3.8.6 to 3.8.8, and now present detailed examples. The correlation matrix derived from equation (3.8.3) is shown in Fig.

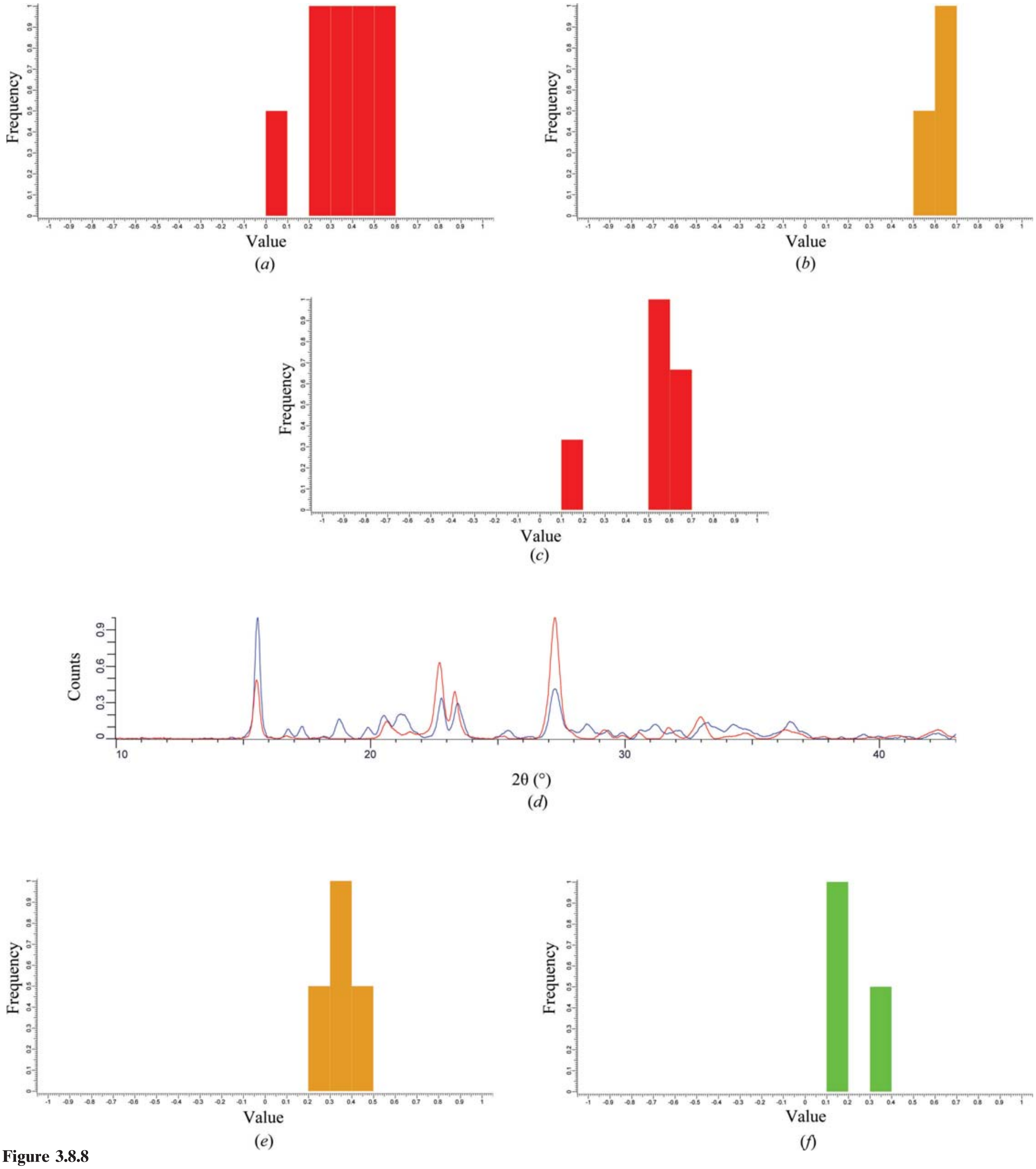


Figure 3.8.8
The use of silhouettes in defining the details of the clustering. (*a*) The silhouettes for the red cluster in the dendrogram from Fig. 3.8.6(*a*). (*b*) The corresponding orange cluster. Both sets of silhouettes have values that are less than 0.5, which indicates that the clustering is not well defined. (*c*) The silhouettes for the red cluster corresponding to the dendrogram in Fig. 3.8.6(*c*). The entry centred on a silhouette value of 0.15 is pattern 3. This implies that pattern 3 is only loosely connected to the cluster and this is demonstrated in part (*d*), where pattern 3 and the most representative pattern for the cluster (No. 9) are superimposed. Although there is a general sense of similarity there are significant differences and the combined correlation coefficient is only 0.62. (*e*) The silhouettes for the orange cluster corresponding to the dendrogram in Fig. 3.8.6(*c*). The silhouettes imply that this is a single cluster without outliers. (*f*) The silhouettes for the green cluster corresponding to the dendrogram in Fig. 3.8.6(*c*). The clustering is poorly defined here.

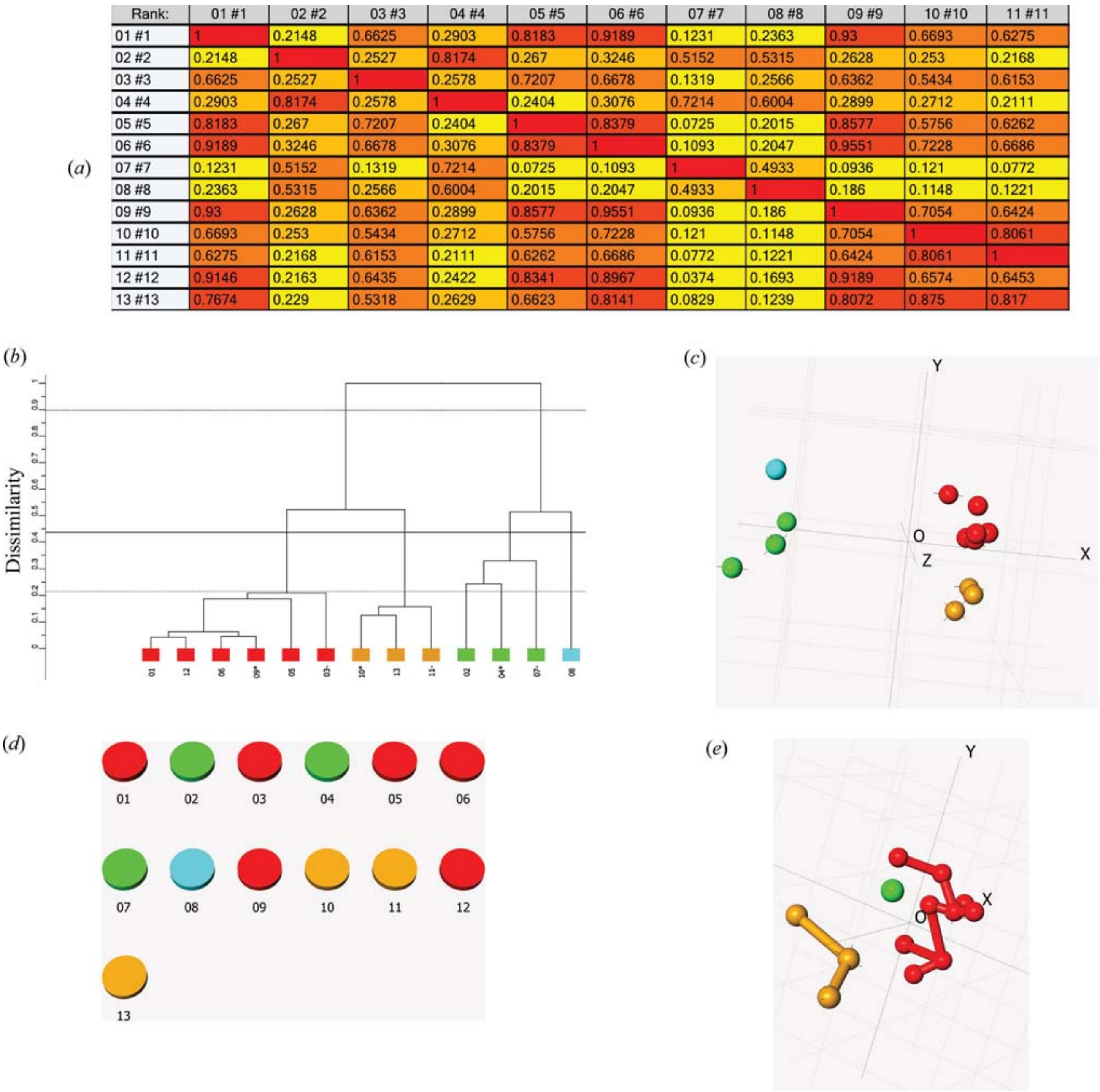
3.8.9(*a*), colour coded to reflect the values of the coefficients; the darker the shade, the higher the correlation. The resulting dendrogram and MMDS plot are shown in Figs. 3.8.9(*b*) and (*c*), respectively. Four clusters are identified in the dendrogram and these have been appropriately coloured. Other visualization tools are now shown. In Fig. 3.8.9(*d*) the pie chart is displayed; the number of rows can be adjusted to reflect the arrangement of the samples in a multiple sample holder. Fig. 3.8.9(*e*) shows the default minimum spanning tree with 12 links. In Fig. 3.8.9(*f*) the scree plot indicates that three clusters will account for more than 95% of the data variability. The steep initial slope is a clear indication of good cluster estimation. The silhouettes are shown in Fig. 3.8.9(*g*–*i*). These were discussed in Section 3.8.5.1. In Fig. 3.8.9(*j*) the default parallel-coordinates plot for the same data is shown, and in Fig. 3.8.9(*k*) there is another view taken from the grand tour. These two plots validate the clustering and also indicate that there is no significant error introduced into the MMDS plot by truncating it into three dimensions.

3.8.6.1.1. *Aspirin data with amorphous samples included*

As a demonstration of the handling of data from amorphous samples, five patterns for amorphous samples were included in the aspirin data and the clustering calculation was repeated. The results are shown in Fig. 3.8.10. Fig. 3.8.10(*a*) shows the dendrogram. It can be seen that the amorphous samples are positioned as isolated clusters on the right-hand end. They also appear as an isolated cluster in the MMDS plot and the parallel-coordinates plots, as shown in Figs. 3.8.10(*b*) and (*c*). It could be argued that these samples should be treated as a single, five-membered cluster rather than five individuals, but we have found that this confuses the clustering algorithms, and it is clearer to the user if the data from amorphous samples are presented as separate classes.

3.8.6.2. *Phase transitions in ammonium nitrate*

Ammonium nitrate exhibits temperature-induced phase transformations. Between 256 and 305 K it crystallizes in the orthorhombic space group *Pmmm* with $a = 5.745$, $b = 5.438$, $c = 4.942$ Å and $Z = 2$; from 305 to 357 K it crystallizes in *Pbnm* with



Rank:	01 #1	02 #2	03 #3	04 #4	05 #5	06 #6	07 #7	08 #8	09 #9	10 #10	11 #11
01 #1	1	0.2148	0.6625	0.2903	0.8183	0.9189	0.1231	0.2363	0.93	0.6693	0.6275
02 #2	0.2148	1	0.2527	0.8174	0.267	0.3246	0.5152	0.5315	0.2628	0.253	0.2168
03 #3	0.6625	0.2527	1	0.2578	0.7207	0.6678	0.1319	0.2566	0.6362	0.5434	0.6153
04 #4	0.2903	0.8174	0.2578	1	0.2404	0.3076	0.7214	0.6004	0.2899	0.2712	0.2111
05 #5	0.8183	0.267	0.7207	0.2404	1	0.8379	0.0725	0.2015	0.8577	0.5756	0.6262
06 #6	0.9189	0.3246	0.6678	0.3076	0.8379	1	0.1093	0.2047	0.9551	0.7228	0.6686
07 #7	0.1231	0.5152	0.1319	0.7214	0.0725	0.1093	1	0.4933	0.0936	0.121	0.0772
08 #8	0.2363	0.5315	0.2566	0.6004	0.2015	0.2047	0.4933	1	0.186	0.1148	0.1221
09 #9	0.93	0.2628	0.6362	0.2899	0.8577	0.9551	0.0936	0.186	1	0.7054	0.6424
10 #10	0.6693	0.253	0.5434	0.2712	0.5756	0.7228	0.121	0.1148	0.7054	1	0.8061
11 #11	0.6275	0.2168	0.6153	0.2111	0.6262	0.6686	0.0772	0.1221	0.6424	0.8061	1
12 #12	0.9146	0.2163	0.6435	0.2422	0.8341	0.8967	0.0374	0.1693	0.9189	0.6574	0.6453
13 #13	0.7674	0.229	0.5318	0.2629	0.6623	0.8141	0.0829	0.1239	0.8072	0.875	0.817

Figure 3.8.9
The complete cluster analysis for the aspirin samples. (*a*) The correlation matrix, which is the source of all the clustering results. The entries are colour coded: the darker the shade, the higher the correlation. (*b*) The dendrogram. The colours assigned to the samples are used in all the visualization tools. (*c*) The corresponding MMDS plot. The clustering defined by the dendrogram is well defined. (*d*) The pie-chart view. (*e*) The minimum spanning tree.

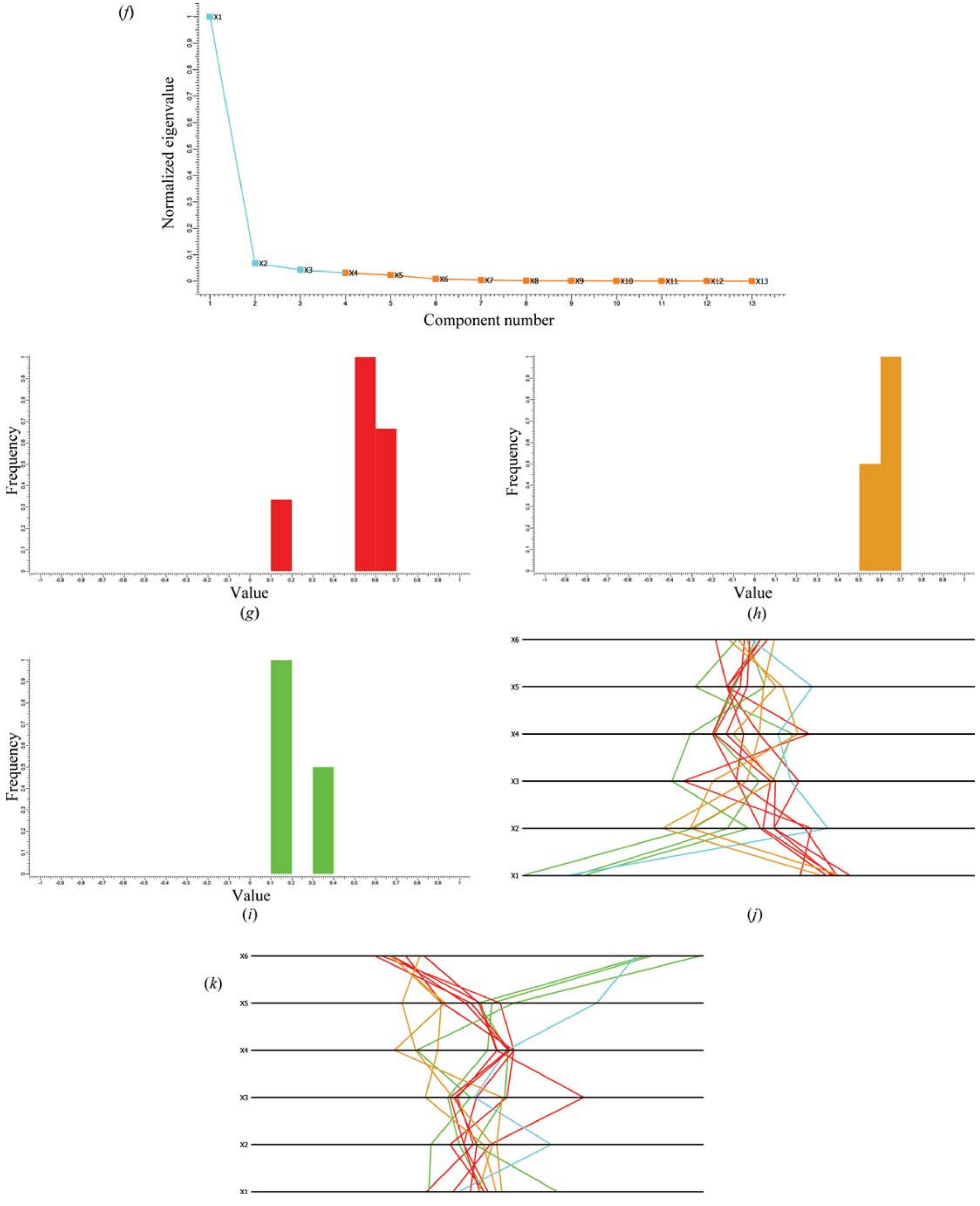


Figure 3.8.9 (continued)
The complete cluster analysis for the aspirin samples (continued). (*f*) The scree plot. It indicates that three clusters explain 95% of the variance of the distance matrix derived from (*a*). (*g–i*) The silhouettes for the red, the orange and the green clusters, respectively. These are discussed in detail in the caption to Fig. 3.8.8. (*j*) The default parallel-coordinates plot. The clusters are well maintained into the 4th, 5th and 6th dimensions. (*k*) Another view of the parallel coordinates using the grand tour. The clustering remains well maintained in higher dimensions.

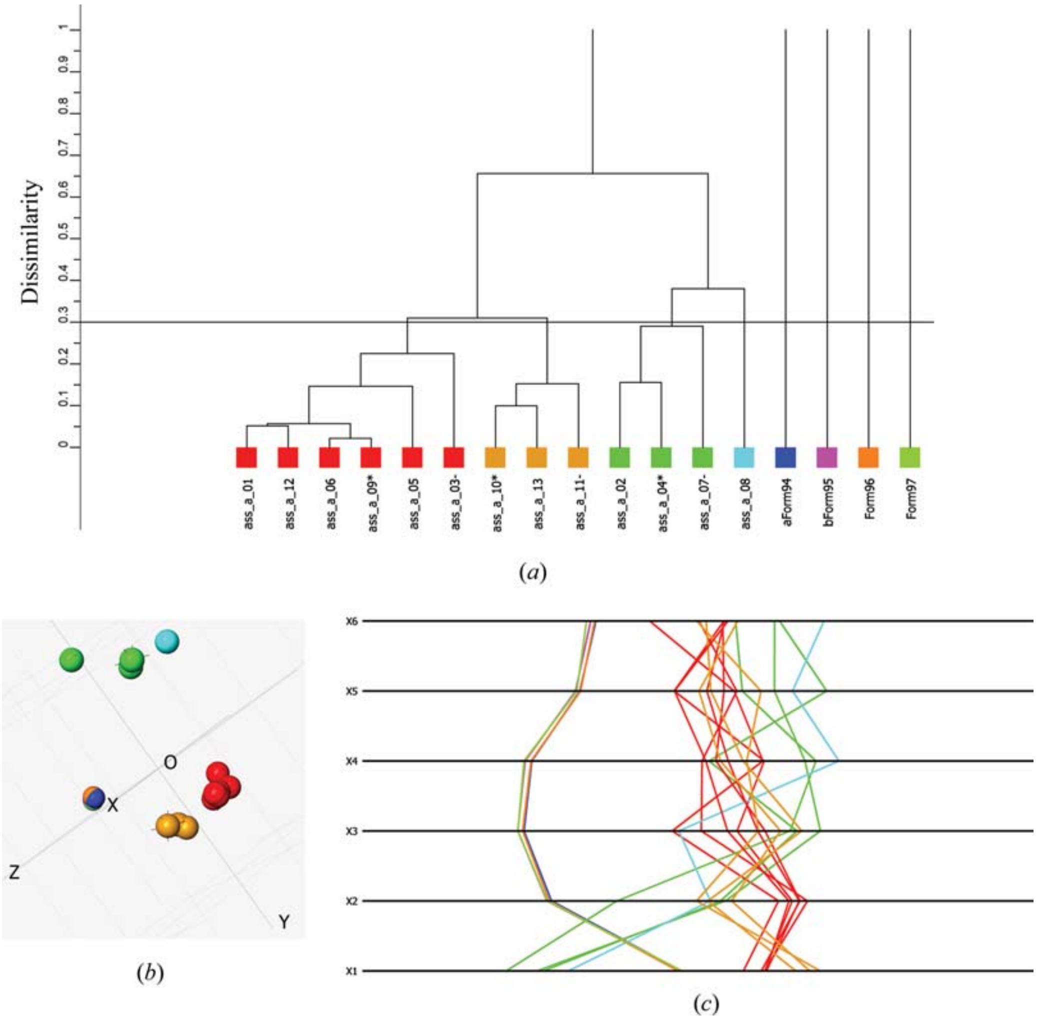


Figure 3.8.10

The aspirin data including data from five amorphous samples. (*a*) The resulting dendrogram and (*b*) the corresponding MMDS plot. (*c*) The parallel-coordinates plot.

a = 7.14, *b* = 7.65, *c* = 5.83 Å with *Z* = 4; between 357 and 398 K it crystallizes in the tetragonal space group $P\bar{4}2_1m$ with *a* = 5.719, *c* = 4.932 Å, *Z* = 2, and above 398 K it transforms to the cubic space group $Pm\bar{3}m$ with *a* = 4.40 Å and *Z* = 1. PXRD data containing 75 powder patterns taken at intervals of 3 K starting at 203 K using a D5000 Siemens diffractometer and Cu $K\alpha$ radiation with a 2θ range of 10–100° were used (Herrmann & Engel, 1997). Fig. 3.8.11(*a*) shows the data in the 2θ range 17–45°.

The visualization of these data following cluster analysis is shown in Fig. 3.8.11(*b*) using an MMDS plot on which has been superimposed a line showing the route followed by the temperature increments. The purple line follows the transition from a mixture of forms IV and V at low temperature (red) through form IV (yellow), form II (blue) and finally form I at high temperature (green). This is an elegant and concise representation of the data in a single diagram.

3.8.7. Quantitative analysis with high-throughput PXRD data without Rietveld refinement

Since mixtures are so common in high-throughput experiments, and indeed in many situations with multiple data sets, it is useful to have a method of automatic quantitative analysis. The quality of data that results from high-throughput crystallography makes it unlikely that an accuracy better than 5–10% can be achieved but, nonetheless, the identification of mixtures can be carried out by whole-profile matching. First a database of *N* pure phases is created, or, if that is not possible, then the most representative patterns with appropriate safeguards can be used. Assume that there is a sample pattern, *S*, which is considered to be a mixture of up to *N* components. *S* comprises *m* data points, $S_1, S_2, \ldots, S_m$. The *N* patterns can be considered to make up fractions $p_1, p_2, p_3, \ldots, p_N$ of the sample pattern. The best possible combination of the database patterns to fit the sample pattern is required. A system of linear equations can be constructed in which x_{11} is measurement point 1 of pattern 1 *etc.*:

$$\begin{aligned}
x_{11}p_1 + x_{12}p_2 + x_{13}p_3 + \ldots + x_{1N}p_N &= S_1,\\
x_{21}p_1 + x_{22}p_2 + x_{23}p_3 + \ldots + x_{2N}p_N &= S_2,\\
&\vdots\\
x_{m1}p_1 + x_{m2}p_2 + x_{m3}p_3 + \ldots + x_{mN}p_N &= S_m. \qquad (3.8.26)
\end{aligned}$$

Writing these in matrix form, we get

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1N}\\ x_{21} & x_{22} & x_{23} & \ldots & x_{2N}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ x_{m1} & x_{m2} & x_{m3} & \ldots & x_{mN} \end{bmatrix} \begin{bmatrix} p_1\\ p_2\\ \vdots\\ p_N \end{bmatrix} = \begin{bmatrix} S_1\\ S_2\\ \vdots\\ S_N \end{bmatrix} \qquad (3.8.27)$$

or

$$\mathbf{xp} = \mathbf{S}. \qquad (3.8.28)$$

A solution for *S* that minimizes