3.8. Clustering and visualization of powder-diffraction data

C. J. Gilmore, G. Barr and W. Dong

3.8.1. Introduction

In high-throughput crystallography, crystallization experiments using robotics coupled with automatic sample changers and two-dimensional (2D) detectors can generate and measure over 1000 powder-diffraction patterns on a series of related compounds, often polymorphs or salts, in a day (Storey *et al.*, 2004). It is also possible to simultaneously measure spectroscopic data, especially Raman (Alvarez *et al.*, 2009). The analysis of these patterns poses a difficult statistical problem: a need to classify the data by putting the samples into clusters based on diffraction-pattern similarity so that unusual samples can be readily identified. At the same time, suitable visualization tools to help in the data-classification process are required; the techniques of classification and visualization go hand-in-hand. Interestingly, the techniques developed for large data sets with poor-quality data also have great value when looking at smaller data sets, and the visualization tools developed for high-throughput studies are especially useful when looking at phase transitions, mixtures *etc.*

In this chapter the methods for comparing whole patterns will be described. The mathematics of cluster analysis will then be explained, followed by a discussion of the associated visualization tools. Examples using small data sets from pharmaceuticals, inorganics and phase transitions will be given; the techniques used can be readily scaled up for handling large, high-throughput data sets. The same methods also work for spectroscopic data and the use of such information with and without powder X-ray diffraction (PXRD) data will be discussed. Finally, the use of visualization tools in quality control is demonstrated.

3.8.2. Comparing 1D diffraction patterns

Comparing 1D diffraction patterns or spectra cannot be done by simply using the peaks and their relative intensities for a number of reasons:

(1) The accurate determinations of the peak positions may be difficult, especially in cases where peak overlap occurs or there is significant peak asymmetry.
(2) The hardware and the way in which the sample is prepared can also affect the d-spacing (or 2θ value) that is recorded for the peak. Shoulders to main peaks and broad peaks can also be problematic.
(3) There is a subjective element to deciding how many peaks there are in the pattern, especially for weak peaks and noisy data.
(4) Weak peaks may be discarded. This can affect the quantitative analysis of mixtures if one component diffracts weakly or is present only in small amounts.
(5) Differences in sample preparation and instrumentation can lead to significant differences in the powder-diffraction patterns of near-identical samples.
(6) Preferred orientation may be present: this is a very difficult and common problem.
(7) The reduction of the pattern to point functions can also make it difficult to design effective algorithms.

In order to use the information contained within the full profile, algorithms are required that utilize each measured data point in the analysis. We use two correlation coefficients for the purpose of comparing PXRD patterns: the Pearson and the Spearman coefficients.

3.8.2.1. *Spearman's rank order coefficient*

Consider two diffraction patterns, i and j, each with n measured points $n((x_1, y_1), \ldots, (x_n, y_n))$. These are transformed to ranks $R(x_k)$ and $R(y_k)$. The Spearman test (Spearman, 1904) then gives a correlation coefficient (Press *et al.*, 2007),

$$R_{ij} = \frac{\sum_{k=1}^{n} R(x_k)R(y_k) - n\left(\frac{n+1}{2}\right)^2}{\left(\sum_{k=1}^{n} R(x_k)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{1/2}\left(\sum_{k=1}^{n} R(y_k)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{1/2}}, \tag{3.8.1}$$

where $-1 \leq R_{ij} \leq 1$.

3.8.2.2. *Pearson's r coefficient*

Pearson's r is a parametric linear correlation coefficient widely used in crystallography. It has a similar form to Spearman's test, except that the data values themselves, and not their ranks, are used:

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})}{\left[\sum_{k=1}^{n} (x_k - \overline{x})^2 \sum_{k=1}^{n} (y_k - \overline{y})^2\right]^{1/2}}, \tag{3.8.2}$$

where $\overline{x}$ and $\overline{y}$ are the means of intensities taken over the full diffraction pattern. Again, r can lie between -1.0 and $+1.0$.

Fig. 3.8.1 shows the use of the Pearson and Spearman correlation coefficients (Barr *et al.*, 2004*a*). In Fig. 3.8.1(*a*) $r = 0.93$ and $R = 0.68$. The high parametric coefficient arises from the perfect match of the two biggest peaks, but the much lower Spearman coefficient acts as a warning that there are unmatched regions in the two patterns. In Fig. 3.8.1(*b*) the situation is reversed: $r = 0.79$, whereas $R = 0.90$, and it can be seen that there is a strong measure of association with the two patterns, although there are some discrepancies in the region 15–35°. In Fig. 3.8.1(*c*) $r = 0.66$ and $R = 0.22$; in this case the Spearman test is again warning of missing match regions. Thus, the use of the two coefficients acts as a valuable balance of their respective properties when processing complete patterns. The Spearman coefficient is also robust in the statistical sense and useful in the case of preferred orientation.

3.8.2.3. *Combining the correlation coefficients*

Correlation coefficients are not additive, so it is invalid to average them directly; they need to be transformed into the Fisher Z value to give

$$\rho_{ij} = \tanh\left[\left(\tanh^{-1}R_{ij} + \tanh^{-1}r_{ij}\right)/2\right]. \quad (3.8.3)$$

3.8.2.4. *Full-profile qualitative pattern matching*

Before performing pattern matching, some data pre-processing may be necessary. In order not to produce artefacts, this should be minimized. Typical pre-processing activities are:

(1) The data are normalized such that the maximum peak intensity is 1.0.
(2) The patterns need to be interpolated if necessary to have common increments in 2θ. High-order polynomials using Neville's algorithm can be used for this (Press *et al.*, 2007).
(3) If backgrounds are large they should be removed. High-throughput data are often very noisy because of low counting times and the sample itself. If this is the case, smoothing of the data can be carried out. The SURE (Stein's Unbiased Risk Estimate) thresholding procedure (Donoho & Johnstone, 1995; Ogden, 1997) employing wavelets is ideal for this task since it does not introduce potentially damaging artefacts, for example ringing around peaks (Barr *et al.,* 2004*a*; Smrčok *et al.*, 1999).

After pre-processing, which needs to be carried out in an identical way for each sample, the following steps are carried out:

(1) The intersecting 2θ range of the two data sets is calculated, and each of the pattern correlation coefficients is calculated using only this region.
(2) A minimum intensity is set, below which profile data are set to zero. This reduces the contribution of background noise to the matching process without reducing the discriminating power of the method. We usually set this to $0.1I_{\max}$ as a default, where $I_{\max}$ is the maximum measured intensity.
(3) The Pearson correlation coefficient is calculated.
(4) The Spearman R is computed in the same way.
(5) An overall ρ value is calculated using (3.8.3).
(6) A shift in 2θ values between patterns is often observed, arising from equipment settings and data-collection protocols. Three possible simple corrections are

$$\Delta(2\theta) = a_0 + a_1\cos\theta, \quad (3.8.4)$$

which corrects for the zero-point error *via* the a_0 term and, *via* the $a_1\cos\theta$ term, for varying sample heights in reflection mode, or

$$\Delta(2\theta) = a_0 + a_1\sin\theta, \quad (3.8.5)$$

which corrects for transparency errors, for example, and

$$\Delta(2\theta) = a_0 + a_1\sin 2\theta, \quad (3.8.6)$$

which provides transparency coupled with thick specimen error corrections, where a_0 and a_1 are constants that can be determined by shifting patterns to maximize their overlap as measured by ρ. It is difficult to obtain suitable expressions for the derivatives $\partial a_0/\partial\rho_{ij}$ and $\partial a_1/\partial\rho_{ij}$ for use in the optimization, so we use the downhill simplex method (Nelder & Mead, 1965), which does not require their calculation.

3.8.2.5. *Generation of the correlation and distance matrices*

Using equation (3.8.3), a correlation matrix is generated in which a set of n patterns is matched with every other to give a symmetric $(n \times n)$ correlation matrix $\boldsymbol{\rho}$ with unit diagonal. The matrix $\boldsymbol{\rho}$ can be converted to a Euclidean distance matrix, **d**, of the same dimensions *via*

$$\mathbf{d} = 0.5(1.0 - \boldsymbol{\rho}) \quad (3.8.7)$$

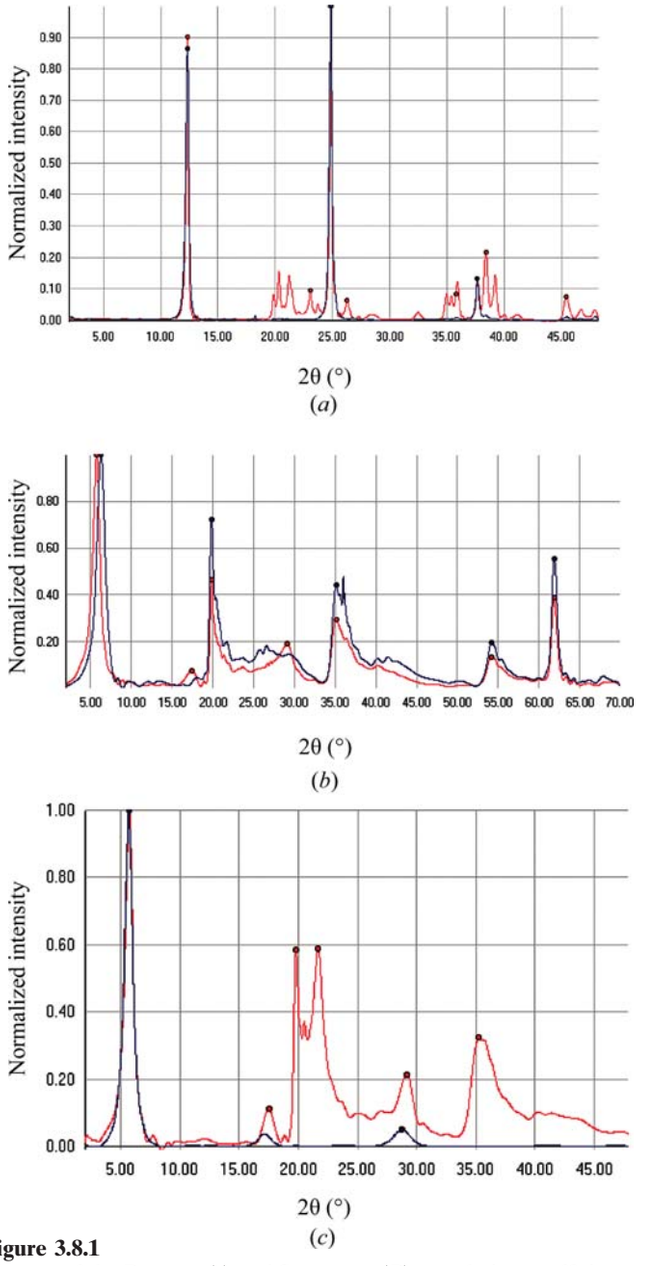

Figure 3.8.1
The use of the Pearson (r) and Spearman (R) correlation coefficients to quantitatively match powder patterns: (*a*) $r = 0.93$, $R = 0.68$; (*b*) $r = 0.79$, $R = 0.90$; (*c*) $r = 0.66$, $R = 0.22$.

or a distance-squared matrix,

$$\mathbf{D} = 0.25(1 - \boldsymbol{\rho})^2 \quad (3.8.8)$$

for each entry i, j in **d**, $0.0 \leq d_{ij} \leq 1.0$. A correlation coefficient of 1.0 translates to a distance of 0.0, a coefficient of -1.0 to 1.0, and zero to 0.5. There are other methods of generating a distance matrix from $\boldsymbol{\rho}$ (see, for example, Gordon, 1981, 1999), but we have found this to be both simple and as effective as any other.

For other purposes a dissimilarity matrix **s** is also needed, whose elements are defined *via*

$$s_{ij} = 1 - d_{ij}/d^{\max}, \quad (3.8.9)$$

where $d^{\max}$ is the maximum distance in matrix **d**. A dissimilarity matrix, $\boldsymbol{\delta}$, is also generated with elements

$$\delta_{ij} = d_{ij}/d_{ij}^{\max}. \quad (3.8.10)$$

Table 3.8.1
Six commonly used clustering methods

For each method, the coefficients α_i, β and γ in equation (3.8.11) are given.

Method	α_i	β	γ
Single link	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete link	$\frac{1}{2}$	0	$\frac{1}{2}$
Average link	$n_i/(n_i+n_j)$	0	0
Weighted-average link	$\frac{1}{2}$	0	0
Centroid	$n_i/(n_i+n_j)$	$-n_in_j/(n_i+n_j)^2$	0
Sum of squares	$(n_i+n_k)/(n_i+n_j+n_k)$	$-n_k/(n_i+n_j+n_k)$	0

In some cases it can be advantageous to use $I^{1/2}$ in the distance-matrix generation; this can enhance the sensitivity of the clustering to weak peaks (Butler *et al.*, 2019).

3.8.3. Cluster analysis

Cluster analysis uses **d** (or **s**, or $\boldsymbol{\delta}$) to partition the patterns into groups based on the similarity of their diffraction profiles. Associated with cluster are a number of important ancillary techniques all of which will be discussed here. A flowchart of these methods is shown in Fig. 3.8.4.

3.8.3.1. *Dendrograms*

Using **d** and **s**, agglomerative, hierarchical cluster analysis is now carried out, in which the patterns are put into clusters as defined by their distances from each other. [Gordon (1981, 1999) and Everitt *et al.* (2001) provide excellent and detailed introductions to the subject. Note that the two editions of Gordon's monograph are quite distinct and complementary.] The method begins with a situation in which each pattern is considered to be in a separate cluster. It then searches for the two patterns with the shortest distance between then, and joins them into a single cluster. This continues in a stepwise fashion until all the patterns form a single cluster. When two clusters (C_i and C_j) are merged, there is the problem of defining the distance between the newly formed cluster $C_i \cup C_j$ and any other cluster C_k. There are a number of different ways of doing this, and each one gives rise to a different clustering of the patterns, although often the difference can be quite small. A general algorithm has been proposed by Lance & Williams (1967), and is summarized in a simplified form by Gordon (1981). The distance from the new cluster formed by merging C_i and C_j to any other cluster C_k is given by

$$d(C_i \cup C_j,\ C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma\left|d(C_i, C_k) - d(C_j, C_k)\right|. \quad (3.8.11)$$

There are many possible clustering methods. Table 3.8.1 defines six commonly used clustering methods, defined in terms of the parameters α, β and γ. All these methods can be used with powder data; in general, the group-average-link or single-link formalism is the most effective, although differences between the methods are often slight.

The results of cluster analysis are usually displayed as a dendrogram, a typical example of which is shown in Fig. 3.8.6(*a*), where a set of 13 powder patterns is analysed using the centroid method. Each pattern begins at the bottom of the plot as a separate cluster, and these amalgamate in stepwise fashion linked by horizontal tie bars. The height of the tie bar represents a similarity measure as measured by the relevant distance. As an indication of the differences that can be expected in the various algorithms used for dendrogram generation, Fig. 3.8.6(*e*) shows the same data analysed using the single-link method: the resulting clustering is slightly different: the similarity measures are larger, and, in consequence, the tie bars are higher on the graph. [For further examples see Barr *et al.* (2004*b*,*c*) and Barr, Dong, Gilmore & Faber (2004).]

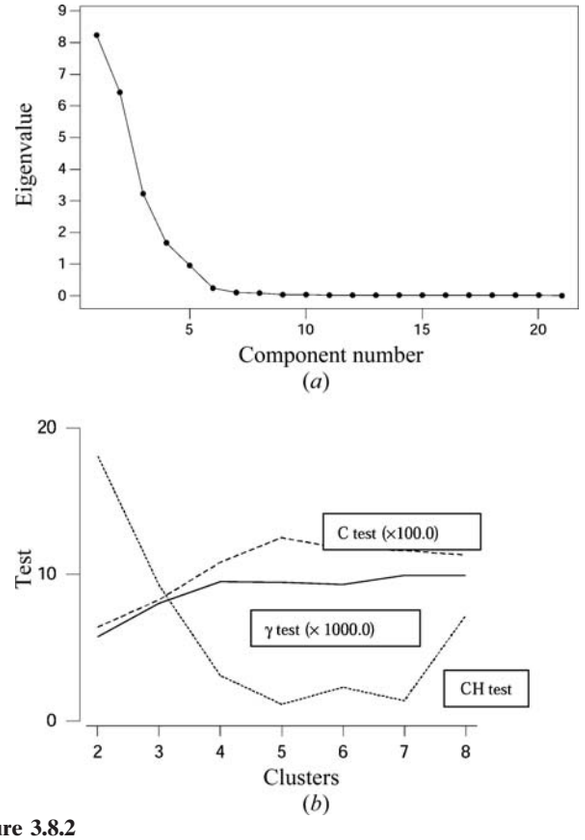

Figure 3.8.2
Four different methods of estimating the number of clusters present in a set of 23 powder patterns for the drug doxazosin. A total of five polymorphs are present, as well as two mixtures of these polymorphs. (*a*) A scree plot from the eigenvalue analysis of the correlation matrix; (*b*) the use of the C test (the coefficients have been multiplied by 100.0), which gives an estimate of five clusters using its local maximum. The γ test estimates that there are seven clusters and the CH test has a local maximum at seven clusters. Numerical details are given in Table 3.8.2.

3.8.3.2. *Estimating the number of clusters*

An estimate of the number of clusters present in the data set is needed. In terms of the dendrogram, this is equivalent to 'cutting the dendrogram' *i.e.* the placement of a horizontal line across it such that all the clusters as defined by tie lines above this line remain independent and unlinked. The estimation of the number of clusters is an unsolved problem in classification methods. It is easy to see why: the problem depends on how similar the patterns need to be in order to be classed as the same, and how much variability is allowed within a cluster. We use two approaches: (*a*) eigenvalue analysis of matrices $\boldsymbol{\rho}$ and **A**, and (*b*) those based on cluster analysis.

Eigenvalue analysis is a well used technique: the eigenvalues of the relevant matrix are sorted in descending order and when a fixed percentage (typically 95%) of the data variability has been accounted for, the number of eigenvalues is selected. This is shown graphically *via* a scree plot, an example of which is shown in Fig. 3.8.2.