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