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

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

$$\Delta(2\theta) = a_0 + a_1 \cos\theta, \tag{3.8.4}$$

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

$$\Delta(2\theta) = a_0 + a_1 \sin\theta, \tag{3.8.5}$$

which corrects for transparency errors, for example, and

$$\Delta(2\theta) = a_0 + a_1 \sin 2\theta, \tag{3.8.6}$$

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, $\mathbf{d}$, of the same dimensions *via*

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

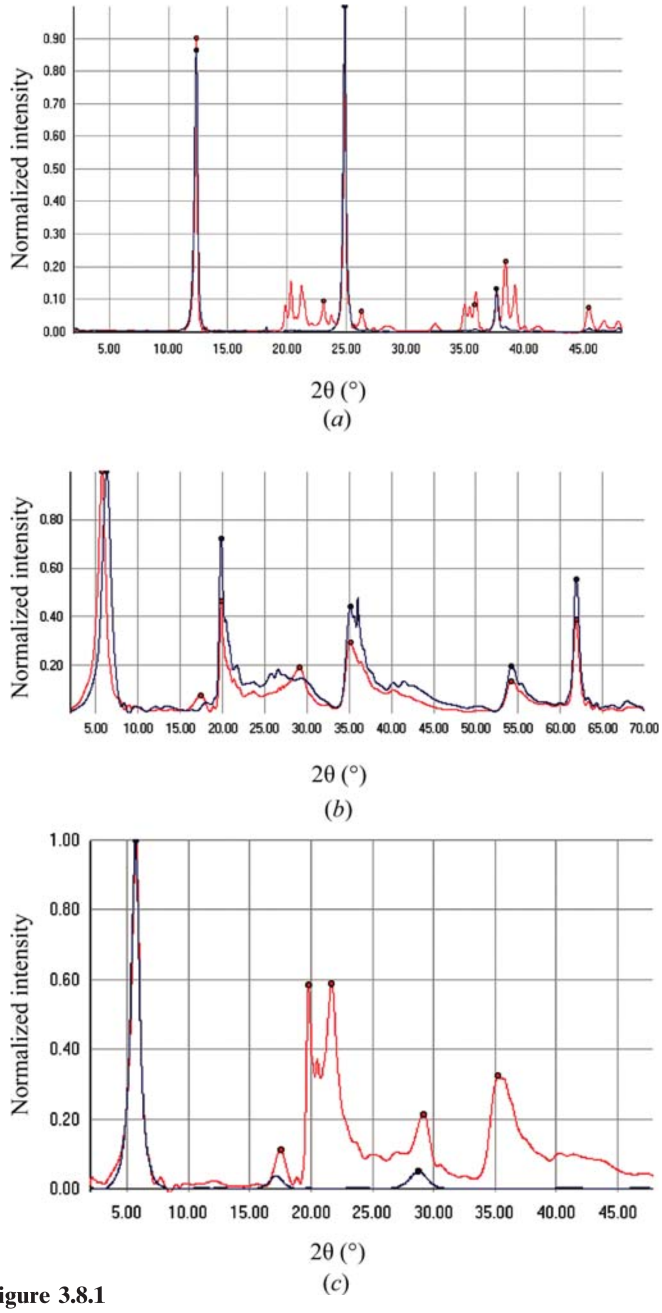

Figure 3.8.1
The use of the Pearson (r) and Spearman (R) correlation coefficients to quantitatively match powder patterns: (a) $r = 0.93$, $R = 0.68$; (b) $r = 0.79$, $R = 0.90$; (c) $r = 0.66$, $R = 0.22$.

or a distance-squared matrix,

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

for each entry i, j in $\mathbf{d}$, $0.0 \le d_{ij} \le 1.0$. 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 $\mathbf{s}$ is also needed, whose elements are defined *via*

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

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

$$\delta_{ij} = d_{ij}/d_{ij}^{\max}. \tag{3.8.10}$$