evaluated and saved in a file. To make room for new strong pixels as the spot search proceeds, all entries of strong pixels that are no longer needed are removed from the hash table and the remaining ones are rehashed. On termination, a list X_i', Y_i', Z_i' $(i = 1, \ldots, n)$ of the centroids of strong spots is available.

11.3.2.6. *Basis extraction*

Any reciprocal-lattice vector can be written in the form $\mathbf{p}_0^* = h\mathbf{b}_1^* + k\mathbf{b}_2^* + l\mathbf{b}_3^*$ where h, k, l are integer numbers and $\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*$ are basis vectors of the lattice. The basis vectors which describe the orientation, metric and symmetry of the crystal, as well as the reflection indices h, k, l, have to be determined from the list of strong diffraction spots X_i', Y_i', Z_i' $(i = 1, \ldots, n)$. Ideally, each spot corresponds to a reciprocal-lattice vector $\mathbf{p}_0^*$ which satisfies the Laue equations after a crystal rotation by φ. Substituting the observed value Z' for the unknown φ angle (see Section 11.3.2.4), $\mathbf{p}_0^*$ is found from the observed spot coordinates as

$$\mathbf{p}_0^* = D(\mathbf{m}_2, -Z')(\mathbf{S}' - \mathbf{S}_0)$$

$$\mathbf{S}' = [(X' - X_0)\mathbf{d}_1 + (Y' - Y_0)\mathbf{d}_2 + F\mathbf{d}_3] \times \left\{\lambda \cdot \left[(X' - X_0)^2 + (Y' - Y_0)^2 + F^2\right]^{1/2}\right\}^{-1}.$$

Unfortunately, the reciprocal-lattice vectors $\mathbf{p}_{0i}^*$ $(i = 1, \ldots, n)$ derived from the above list of strong diffraction spots often contain a number of 'aliens' (spots arising from fluctuations of the background, from ice, or from satellite crystals) and a robust method has to be used which is still capable of recognizing the dominant lattice. One approach, suggested by Bricogne (1986) and implemented in a number of variants (Otwinowski & Minor, 1997; Steller *et al.*, 1997), is to identify a lattice basis as the three shortest linear independent vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$, each at a maximum of the Fourier transform $\sum_{i=1}^{n} \cos(2\pi\mathbf{b} \cdot \mathbf{p}_{0i}^*)$. Alternatively, a reciprocal basis for the dominant lattice can be determined from short differences between the reciprocal-lattice vectors (Howard, 1986; Kabsch, 1988*a*). As implemented in *XDS*, a lattice basis is found by the following procedure.

The list of given reciprocal-lattice points $\mathbf{p}_{0i}^*$ $(i = 1, \ldots, n)$ is first reduced to a small number m of low-resolution difference-vector clusters $\mathbf{v}_\mu^*$ $(\mu = 1, \ldots, m)$. f_μ is the population of a difference-vector cluster $\mathbf{v}_\mu^*$, that is the number of times the difference between any two reciprocal-lattice vectors $\mathbf{p}_{0i}^* - \mathbf{p}_{0j}^*$ is approximately equal to $\mathbf{v}_\mu^*$. In a second step, three linear independent vectors $\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*$ are selected among all possible triplets of difference-vector clusters that maximize the function Q:

$$Q(\mathbf{b}_1^*, \mathbf{b}_2^*, \mathbf{b}_3^*) = \sum_{\mu=1}^{m} f_\mu q(\xi_1^\mu, \xi_2^\mu, \xi_3^\mu)$$

$$q(\xi_1^\mu, \xi_2^\mu, \xi_3^\mu) = \exp\left(-2\sum_{k=1}^{3}\left\{[\max(|\xi_k^\mu - h_k^\mu| - \varepsilon, 0)/\varepsilon]^2 + [\max(|h_k^\mu| - \delta, 0)]^2\right\}\right)$$

$$\xi_k^\mu = \mathbf{v}_\mu^* \cdot \mathbf{b}_k, \quad \mathbf{v}_\mu^* = \sum_{k=1}^{3} \xi_k^\mu \mathbf{b}_k^*, \quad \mathbf{b}_k \cdot \mathbf{b}_l^* = \begin{cases} 1 \text{ if } k = l; \\ 0 \text{ otherwise} \end{cases}$$

$$h_k^\mu = \text{ nearest integer to } \xi_k^\mu.$$

The absolute maximum of Q is assumed if all difference vectors can be expressed as small integral multiples of the best triplet. Deviations from this ideal situation are quantified by the quality measure q. The value of q declines sharply if the expansion coefficients ξ_k^μ deviate by more than ε from their nearest integers h_k^μ or if the indices are absolutely larger than δ. The constraint on the allowed range of indices prevents the selection of a spurious triplet of very short difference vector clusters which might be present in the set. Excellent results have been obtained using $\varepsilon = 0.05$ and $\delta = 5$. The best vector triplet thus found is refined against the observed difference-vector clusters. Finally, a reduced cell is derived from the refined reciprocal-base vector triplet as defined in IT *A* (1995), p. 743.

11.3.2.7. *Indexing*

Once a basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ of the lattice is available, integral indices h_i, k_i, l_i must be assigned to each reciprocal-lattice vector $\mathbf{p}_{0i}^*$ $(i = 1, \ldots, n)$. Using the integers nearest to $\mathbf{p}_{0i}^* \cdot \mathbf{b}_k$ $(k = 1, 2, 3)$ as indices of the reciprocal-lattice vectors $\mathbf{p}_{0i}^*$ could easily lead to a misindexing of longer vectors because of inaccuracies in the basis vectors $\mathbf{b}_k$ and the initial values of the parameters describing the instrumental setup. A more robust solution of the indexing problem is provided by the *local indexing method* which assigns only small index differences $h_i - h_j, k_i - k_j, l_i - l_j$ between pairs of neighbouring reciprocal-lattice vectors (Kabsch, 1993).

The reciprocal-lattice points can be considered as the nodes of a tree. The tree connects the n points to each other with the connections as its branches. The length ℓ_{ij} of a possible branch between nodes i and j is defined here as

$$\ell_{ij} = 1 - \exp\left(-2\sum_{k=1}^{3}\left\{[\max(|\xi_k^{ij} - h_k^{ij}| - \varepsilon, 0)/\varepsilon]^2 + [\max(|h_k^{ij}| - \delta, 0)]^2\right\}\right)$$

$$\xi_k^{ij} = (\mathbf{p}_{0i}^* - \mathbf{p}_{0j}^*) \cdot \mathbf{b}_k, \quad h_k^{ij} = \text{ nearest integer of } \xi_k^{ij}, \quad k = 1, 2, 3.$$

Reliable index differences are indicated by short branches; in fact, ℓ_{ij} is 0 if none of the indices h_k^{ij} is absolutely larger than δ and the ξ_k^{ij} are integer values to within ε. Typical values of ε and δ are $\varepsilon = 0.05$ and $\delta = 5$. Defining the length of a tree as the sum of the lengths of its branches, a shortest tree among all n^{n-2} possible trees is determined by the elegant algorithm described by Dijkstra (1976). Starting with arbitrary indices 0, 0, 0 for the root node, the local indexing method then consists of traversing the shortest tree and thereby assigning each node the indices of its predecessor plus the small index differences between the two nodes.

During traversal of the tree, each node is also given a subtree number. Starting with subtree number 1 for the root node, each successor node is given the same subtree number as its predecessor if the length of the connecting branch is below a minimal length $\ell_{\min}$. Otherwise its subtree number is incremented by 1. Thus all nodes in the same subtree have internally consistent reflection indices. Defining the size of a subtree by the number of its nodes, aliens are usually found in small subtrees. Finally, a constant index offset is determined such that the centroids of the observed reciprocal-lattice points $\mathbf{p}_{0i}^*$ belonging to the largest subtree and their corresponding grid vectors $\sum_{k=1}^{3} h_k^i \mathbf{b}_k^*$ are as close as possible. This offset is added to the indices of each reciprocal-lattice point.

11.3.2.8. *Refinement*

For a fixed detector, the diffraction pattern depends on the parameters $\mathbf{S}_0, \mathbf{m}_2, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, X_0, Y_0$ and F. Starting values for the parameters can be obtained by the procedures described above that do not rely on prior knowledge of the crystal orientation, space-group symmetry or unit-cell metric. Better estimates of the parameter values, as required for the subsequent integration step, can be obtained by the method of least squares from the list of n observed indexed reflection centroids h_i, k_i, l_i, X_i', Y_i', Z_i' $(i = 1, \ldots, n)$. In this method, the parameters are chosen to minimize a weighted sum of squares of the residuals

$$E = w_X \sum_{i=1}^{n} (\Delta_X^i)^2 + w_Y \sum_{i=1}^{n} (\Delta_Y^i)^2 + w_Z \sum_{i=1}^{n} (\Delta_Z^i)^2.$$

The residuals between the calculated (X_i, Y_i, Z_i) and observed spot centroids are

$$\Delta_X^i = X_i - X_i' = X_0 + F\mathbf{S}_i \cdot \mathbf{d}_1 / \mathbf{S}_i \cdot \mathbf{d}_3 - X_i'$$
$$\Delta_Y^i = Y_i - Y_i' = Y_0 + F\mathbf{S}_i \cdot \mathbf{d}_2 / \mathbf{S}_i \cdot \mathbf{d}_3 - Y_i'$$
$$\Delta_Z^i = Z_i - Z_i' = \varphi_0 + \Delta_\varphi \sum_{j=-\infty}^{\infty} (j - 1/2) R_j^i - Z_i'.$$

Let s_μ $(\mu = 1, \ldots, k)$ denote the k independent parameters for which initial estimates are available. Expanding the residuals to first order in the parameter changes δs_μ gives

$$\Delta(s_\mu + \delta s_\mu) \approx \Delta(s_\mu) + \sum_{\mu=1}^{k} \frac{\partial \Delta}{\partial s_\mu} \delta s_\mu.$$

The parameters should be changed in such a way as to minimize $E(\delta s_\mu)$, which implies $\partial E / \partial \delta s_\mu = 0$ for $\mu = 1, \ldots, k$. The δs_μ are found as the solution of the k normal equations

$$\sum_{\mu'=1}^{k} \left(w_X \sum_{i=1}^{n} \frac{\partial \Delta_X^i}{\partial s_\mu} \frac{\partial \Delta_X^i}{\partial s_{\mu'}} + w_Y \sum_{i=1}^{n} \frac{\partial \Delta_Y^i}{\partial s_\mu} \frac{\partial \Delta_Y^i}{\partial s_{\mu'}} + w_Z \sum_{i=1}^{n} \frac{\partial \Delta_Z^i}{\partial s_\mu} \frac{\partial \Delta_Z^i}{\partial s_{\mu'}} \right) \delta s_{\mu'}$$
$$= - \left(w_X \sum_{i=1}^{n} \Delta_X^i \frac{\partial \Delta_X^i}{\partial s_\mu} + w_Y \sum_{i=1}^{n} \Delta_Y^i \frac{\partial \Delta_Y^i}{\partial s_\mu} + w_Z \sum_{i=1}^{n} \Delta_Z^i \frac{\partial \Delta_Z^i}{\partial s_\mu} \right).$$

The parameters are corrected by δs_μ and a new cycle of refinement is started until a minimum of E is reached. The weights

$$w_X = 1 / \sum_{i=1}^{n} (\Delta_X^i)^2, \quad w_Y = 1 / \sum_{i=1}^{n} (\Delta_Y^i)^2, \quad w_Z = 1 / \sum_{i=1}^{n} (\Delta_Z^i)^2$$

are calculated with the current guess for s_μ at the beginning of each cycle.

The derivatives appearing in the normal equations can be worked out from the definitions given in Sections 11.3.2.2 and 11.3.2.4, and only the form of the gradient of the Z residuals is shown. Assuming $\sigma_i = \sigma_M / |\zeta_i|$ $(i = 1, \ldots, n)$ is constant for each reflection, the gradients of the Z residuals are obtained from the chain rule and the relation $\mathrm{d}\,\mathrm{erf}(z)/\mathrm{d}z = [2/(\pi)^{1/2}] \exp(-z^2)$.

$$\frac{\partial \Delta_Z^i}{\partial s_\mu} = \frac{\partial \Delta_Z^i}{\partial \varphi_i} \frac{\partial \varphi_i}{\partial s_\mu}$$
$$\frac{\partial \Delta_Z^i}{\partial \varphi_i} = \frac{\Delta_\varphi}{(2\pi)^{1/2} \sigma_i} \sum_{j=-\infty}^{\infty} \exp[-(\varphi_0 + j\Delta_\varphi - \varphi_i)^2 / 2\sigma_i^2]$$
$$\frac{\partial \varphi_i}{\partial s_\mu} = \cos\varphi_i \frac{\partial \sin\varphi_i}{\partial s_\mu} - \sin\varphi_i \frac{\partial \cos\varphi_i}{\partial s_\mu}.$$

Obviously, $\partial \Delta_Z^i / \partial s_\mu$ is small for a fully recorded reflection because of the small values of all exponentials appearing in $\partial \Delta_Z^i / \partial \varphi_i$. In contrast, the gradient for a partial reflection, equally recorded on two adjacent images, is most sensitive to parameter variations because one of the exponentials assumes its maximum value. In the limiting case of infinitely fine-sliced data, it can be shown that $\lim_{\Delta_\varphi \to 0} \partial \Delta_Z^i / \partial \varphi_i = 1$. Thus, the refinement scheme based on observed Z centroids, as described here and implemented in *XDS*, is applicable to fine-sliced data – and to data recorded with a large oscillation range as well.

11.3.3. Integration

A fundamental requirement for a general integration method is that it should distinguish carefully between signal and background points within its integration domain. For weak reflections, this distinction cannot be made reliably because of the errors superimposed on the signal. The problem can be solved, however, provided that both weak and strong reflections share the same profile shape – an assumption that has been adopted by most data-processing packages.

The intensity distribution of a reflection can be modelled analytically or derived from the observed profiles of neighbouring strong spots. For the rotation method, the profile shape depends strongly on the specific path of the reflection through the Ewald sphere and on variations in the angle of incidence of the diffracted beam on a flat detector. These geometrical distortions can be eliminated by mapping the reflections onto the coordinate system defined in Section 11.3.2.3, which simplifies the task of modelling the expected intensity distribution as all reflection profiles become similar.

11.3.3.1. *Spot extraction*

The region around a spot is defined by the two parameters δ_D and δ_M, which represent spot diameter and reflecting range, respectively. It is assumed that the coordinates of all image pixels contributing to the intensity of a spot satisfy $|\varepsilon_1| \leq \delta_D/2, |\varepsilon_2| \leq \delta_D/2$ and $|\varepsilon_3| \leq \delta_M/2$ when mapped to the profile coordinate system $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ defined in Section 11.3.2.3. Regions of neighbouring reflections may overlap. As implemented in *XDS*, potential overlap is dealt with by a simple strategy: pixels within the overlap region are assigned to the nearest spot. This is carried out in two steps. First, reflections predicted to occur on a given rotation image are found by generating and testing all possible indices h, k, l up to the highest resolution recorded by the detector. Reflection indices, coordinates of the diffracted beam wave vector and the expected fraction of spot intensity recorded on the image are saved in a table. In the second step, each reflection boundary is traced in the image and corrected to exclude pixels belonging to overlapping reflections, which are rapidly located in the table by the hash technique. The image scaling factor obtained from the mean image background and the neighbourhood pixel values belonging to the reflections recorded in the image are saved on a scratch file dedicated to the currently processed data image.

At regular intervals, these files are merged such that all pixel values belonging to a spot found in the contributing images follow each other. Reflections for which contributing pixels are expected further ahead in data processing are just copied to a scratch output file. The other reflections are mapped to the Ewald sphere, as described below, and their three-dimensional profiles and accompanying information are routed to the main output file of the spot-extraction step. After the file-merging procedure, spot extraction continues.

11.3.3.2. *Background*

The region around a spot is assumed to have been chosen to be large enough to include a sufficient number of pixels which can be used for determination of the background. Background determination, as implemented in *XDS*, begins by sorting all pixels belonging to a reflection by increasing intensity. For weak or absent reflections, these values should represent a random sample drawn from a normal distribution. If this is not the case, the pixel with the largest intensity is removed until the sampling distribution of the remaining smaller items satisfies the expected distribution. This method will also exclude pixels with unexpectedly high values, such as ice reflections. The background, determined as the mean value of the accepted pixels, is systematically overestimated for strong spots because of some residual intensity extending into the accepted background pixels. This residual intensity is estimated

references