NSXTool Internal Documentation of Implemented Theory

The NSXTool Collaboration

internal draft, work in progress, do not circulate, April 22, 2020

1 Overview

We mention a few predecessors, for the sake of listing references: RETREAT [1], XDS [2], HKL2000 [3, 4, 5].

The NSXTool project was initially started as a replacement for RETREAT but many features were added in order to add support for BioDiff.

The basic workflow of NSXTool is as follows

1. Load data, apply masks
2. Peak search 2.1
3. Determine lattice 2.2 2.3
4. Determine spacegroup 2.5
5. Predict peaks 3
6. Integrate peaks 4
7. Computed merged statistics 5

2 Lattice determination

We use [6] as one of our primary references.

2.1 Peak search

The initial peak search is essentially a pure image processing step, with no crystallographic input. The technique is roughly as follows

1. Apply an image filter to subtract local background
2. Apply a threshold to the resulting image
3. Find connected components ("blobs") of the resulting thresholded image
4. Merge blobs that overlap, according to some cutoff

In the first step, we apply a filter which consists of a central circular region with positive weight, and an outer annular region with negative weight. The weights are chosen so that the convolution computes the local average of the circular region subtracted by the average of the annular region, effectively giving a local background subtraction. The radii of the circle and annulus may be specified by the user.

To find connected components, we use a standard blob-search algorithm, as described e.g. on the wikipedia page (do we have a better reference for this?) In the last step, we compute inertia ellipsoids for each blob, and merge those blobs whose ellipsoids overlap, after a user-defined scaling factor has been applied. The merging process is repeated until there are no longer any overlapping ellipsoids.

The collision detection problem for ellipsoids is sped up by storing them in an octree. The ellipsoid overlap detection is implemented using the criterion described in [**?**].

2.2 Autoindexing

We use an autoindexing method based on 1d Fourier-transform, see [7, 8].

The algorithm works as follows. We are given some set of $\boldsymbol{q}$ vectors which lie approximately on a lattice, yet to be determined. To find candidate lattice directions, we take a random sample of directions. For each direction, we perform the orthogonal projection of each $\boldsymbol{q}$ vector to the infinite line specified by the direction. We then take a finite number of bins along this line (the way the binning is performed can be controlled by user-defined parameters), and then take FFT of the resulting histogram. The histogram will be strongly periodic when the direction corresponds to a lattice direction, so we identify lattice vectors by taking the strongest Fourier modes of the histograms.

The FFT method produces a finite set of potential lattice vectors. To find a basis, we enumerate over triples of these basis vectors and rank them according to

1. The percentage of peaks that can be indexed (with integer indices)
2. The volume of the resulting unit cell

This provides a ranked list of candidate unit cells, from which the user may choose.

We remark that when indexing fails, try one of the following:

- Reduce the number of input peaks, e.g. by taking just the very strong peaks
- Change the parameters of the FFT method so that the bin sizes change

In all cases that we have encountered, we were able to correctly index the peaks, but sometimes trial and error is required to determine the best FFT parameters.

2.3 Computing the reduced unit cell

Once we have a lattice basis, we compute the Niggli cell following [9, 10] and then transform the Niggli cell to a reduced standard form following [11]. The cell is given an index from 1 to 44 corresponding to the 44 cases in the International Tables [**?**].

2.4 Computation of the first Brillouin zone

Let B be matrix whose rows form a basis for the reciprocal space lattice. Thus the reciprocal lattice consists of the set

$$\{(hk\ell)B \mid h, k, \ell \in \mathbf{Z}\}. \tag{1}$$

The Bragg plane $B_{\boldsymbol{q}}$ corresponding to $\boldsymbol{q}$ is the plane which bisects $\boldsymbol{q}$ at right angles, i.e. the plane

$$\{\boldsymbol{q}' \mid \boldsymbol{q} \cdot \boldsymbol{q}' = -\frac{1}{2}\boldsymbol{q}^2\} \tag{2}$$

Note that the Bragg plane separates the reciprocal lattice into two subsets, we denote by $H_{\boldsymbol{q}}$ that subset which contains the origin, i.e.

$$H_{\boldsymbol{q}} = \{\boldsymbol{q}' \mid \boldsymbol{q} \cdot \boldsymbol{q}' \leq \frac{1}{2}\boldsymbol{q}^2\}. \tag{3}$$

For convenience, let us define a *zone* to be an convex intersubsection of the half-spaces $H_{\boldsymbol{q}}$. The *Brillouin zone* is the smallest zone, which equivalently is the intersubsection

$$Z_B := \bigcap_{\boldsymbol{q} \in \Lambda^*} H_{\boldsymbol{q}}. \tag{4}$$

The goal of this subsection is to describe an algorithm to compute Z_B, without any assumptions on B.

2.4.1 Step 1: Compute an initial zone.

Let $\boldsymbol{b}_1, \boldsymbol{b}_2, \boldsymbol{b}_3$ denote the rows of B, which form a basis of the dual lattice. We let the initial zone Z_0 be the parallelpiped centered at the origin bounded by the planes $H_{\pm \boldsymbol{q}_i}$.

2.4.2 Step 2: Compute a bounding sphere.

The initial zone Z_0 contains the Brillouin zone. To find a bounding radius, compute the 6 vertices $\boldsymbol{q}_i$ of Z_0, and take $R = \max_i |\boldsymbol{q}_i|$. Since the Brillouin zone is contained in Z_0, and Z_0 is contained in the ball of radius R, we know that the Brillouin zone is contained in the ball of radius R. Therefore we have reduced the problem to checking Bragg planes for $\boldsymbol{q}$ satisfying $|\boldsymbol{q}| \leq R$, of which there are finitely many.

To simplify the problem further, take $\boldsymbol{q} = (h, k, \ell)^t B$. Then

$$|\boldsymbol{q}|^2| = |(h, k, \ell)^t B|^2 = (h, k, \ell)^t B B^t (h, k, \ell) \geq \lambda(h^2 + k^2 + \ell^2), \tag{5}$$

where λ is the smallest eigenvalue of the symmetric matrix BB^t. Therefore

$$|(h,k,\ell)| \leq \frac{R}{\sqrt{\lambda}}, \tag{6}$$

so we obtain an effective bound on the (finitely-many) values of h, k, ℓ which must be checked.

2.4.3 Step 3: Find the finitely-many bounding planes of Z_B.

Now we iterate.

1. Set $i = 0$, and let Z_0 be as above.
2. For each (h,k,ℓ) satisfying $|(h,k,\ell)| < R/\sqrt{\lambda}$
   (a) Set $\boldsymbol{q} = (h,k,\ell)^t B$.
   (b) If $\boldsymbol{q}/2 \in Z_i$, then
      i. Set $Z_{i+1} = Z_i \cap H_{\boldsymbol{q}}$.
      ii. Set $i := i + 1$.
      iii. (Optional) remove extraneous bounding planes from Z_{i+1}.
3. Stop.

2.5 Space Groups

We fix a real-space lattice Λ which is spanned by real-space vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \boldsymbol{a}_3$, i.e.

$$\Lambda = \{n_1\boldsymbol{a}_1 + n_2\boldsymbol{a}_2 + n_3\boldsymbol{a}_3 \mid n_1, n_2, n_3 \in \mathbf{Z}\} \tag{7}$$

we denote by Λ^* the dual lattice, defined by

$$\Lambda^* = \{q \in \mathbf{R}^3 \mid \forall x \in \mathbf{Z}, q \cdot x \in \mathbf{Z}\} \tag{8}$$

Let Aff denote the group of affine translations of Euclidean space. As a set it is equal to $O(3) \times \mathbf{R}^3$, which acts on $\mathbf{R}^3$ via $(A,b) \cdot x = Ax + b$. From this definition, we see that the group law is

$$(A_1, b_1) \cdot (A_2, b_2) = (A_1A_2, A_1b_2 + b_1). \tag{9}$$

A *space group* is a discrete cocompact subgroup of Aff. It is a classic theorem that there are exactly 230 space groups. The *translational subgroup* of a spacegroup G is the subgroup $1 \times \mathbf{R}^3 \cap G$, i.e. the subgroup consisting of affine transformations with trivial rotational part.

If $f(x)$ is a function which is invariant under G, then it is in particular invariant under the translational subgroup, and therefore we can express it as a Fourier series

$$f(x) = \sum_q f_q e^{2\pi i q \cdot x}. \tag{10}$$

Now consider the action of $g = (A, b)$:

$$f(x) = f(g \cdot x) \tag{11}$$
$$= f(Ax + b) \tag{12}$$
$$= \sum_q f_q e^{2\pi i q \cdot (Ax+b)} \tag{13}$$
$$= \sum_q f_q e^{2\pi i ((A^t q) \cdot x + q \cdot b)} \tag{14}$$
$$= \sum_q f_q e^{2\pi i q \cdot b} e^{2\pi i (A^t q) \cdot x} \tag{15}$$

which shows that $f_{A^t q} = f_q e^{2\pi i q \cdot b}$ whenever $(A, b) \in G$.

For space group determination see e.g. [12, 13].

3 Shape Prediction

We make a simplifying assumption, that for a *perfect plane wave* $\boldsymbol{k}_\mathrm{i}$, the observed scattering function has the form

$$\sum_{hkl} I_{hkl} f(\boldsymbol{q} - \boldsymbol{q}_{hkl}), \tag{16}$$

i.e. that the peak shape is independent of its intensity and Miller index, specified by a single function $f(\boldsymbol{q})$.

Now suppose that the incoming plane wave actually has momentum $\boldsymbol{k}_\mathrm{i} + \delta\boldsymbol{k}_\mathrm{i}$, with $\delta\boldsymbol{k}_\mathrm{i}$ sampled from a probability distribution $p(\delta\boldsymbol{k}_\mathrm{i})$. Let $\boldsymbol{u}$ be the unit vector pointing from the sample origin to a given detector pixel. As we only consider elastic scattering, we can write the wavenumber as $K \coloneqq k_\mathrm{i} = k_\mathrm{f}$. Then the outgoing momentum associated with this pixel is

$$|\boldsymbol{k}_\mathrm{i} + \delta\boldsymbol{k}_\mathrm{i}|\boldsymbol{u} = \boldsymbol{u}\sqrt{\boldsymbol{k}_\mathrm{i}^2 + 2\boldsymbol{k}_\mathrm{i} \cdot \delta\boldsymbol{k}_\mathrm{i} + (\delta\boldsymbol{k}_\mathrm{i})^2} \tag{17}$$
$$\approx \boldsymbol{u}\sqrt{\boldsymbol{k}_\mathrm{i}^2 + 2\boldsymbol{k}_\mathrm{i} \cdot \delta\boldsymbol{k}_\mathrm{i}} \tag{18}$$
$$= \boldsymbol{u}K\sqrt{1 + 2\frac{\boldsymbol{k}_\mathrm{i} \cdot \delta\boldsymbol{k}_\mathrm{i}}{\boldsymbol{k}_\mathrm{i}^2}} \tag{19}$$
$$\approx \boldsymbol{u}K\left(1 + \frac{\boldsymbol{k}_\mathrm{i} \cdot \delta\boldsymbol{k}_\mathrm{i}}{\boldsymbol{k}_\mathrm{i}^2}\right) \tag{20}$$
$$= \boldsymbol{k}_\mathrm{f} + \delta\boldsymbol{k}_\mathrm{f}, \tag{21}$$

where $\boldsymbol{k}_\mathrm{f} = \boldsymbol{u}K$ and $\delta\boldsymbol{k}_\mathrm{f} = \boldsymbol{u}(\boldsymbol{k}_\mathrm{i} \cdot \delta\boldsymbol{k}_\mathrm{i})/K$. Therefore, we have

$$\delta\boldsymbol{q} = \delta\boldsymbol{k}_\mathrm{f} - \delta\boldsymbol{k}_\mathrm{i} = \boldsymbol{u}(\boldsymbol{k}_\mathrm{i} \cdot \delta\boldsymbol{k}_\mathrm{i})/K - \delta\boldsymbol{k}_\mathrm{i} =: A\delta\boldsymbol{k}_\mathrm{i}, \tag{22}$$

where A is the matrix $A = K^{-1}\boldsymbol{u}\boldsymbol{k}_\mathrm{i}^t - \mathbf{1} = K^{-2}\boldsymbol{k}_\mathrm{f}\boldsymbol{k}_\mathrm{i}^t - \mathbf{1}$. Note that $A\boldsymbol{k}_\mathrm{i} = \boldsymbol{q}$ and therefore $\boldsymbol{q} + \delta\boldsymbol{q} = A(\boldsymbol{k}_\mathrm{i} + \delta\boldsymbol{k}_\mathrm{i})$.

Therefore, the observed intensity at detector position (x, y) should be proportional to

$$= \int f(\boldsymbol{q} - \boldsymbol{q}_{hkl} + \delta\boldsymbol{q}) p(\delta\boldsymbol{k}_\mathrm{i}) d(\delta\boldsymbol{k}_\mathrm{i}) \tag{23}$$
$$= \int f(\boldsymbol{q} - \boldsymbol{q}_{hkl} + RA\delta\boldsymbol{k}_\mathrm{i}) p(\delta\boldsymbol{k}_\mathrm{i}) d(\delta\boldsymbol{k}_\mathrm{i}) \tag{24}$$

where R is the rotation matrix taking lab coordinates to sample-fixed coordinates. **Assuming that A is invertible (why? $\det A = -\frac{1}{2}\boldsymbol{q}^2$)**, we have $\delta\boldsymbol{k}_{\mathrm{i}} = A^{-1}\delta\boldsymbol{q}$ and $d(\delta\boldsymbol{k}_{\mathrm{i}}) = |\det A|^{-1}d(\delta\boldsymbol{q})$. Let Σ_M denote the variance-covariance matrix of the profile shape f and let Σ_D denote the variance-covariance matrix of the beam divergence $\delta\boldsymbol{k}_{\mathrm{i}}$. Then from the above formula we see that the *observed* profile shape will have (in sample-fixed q-space) a variance-covariance matrix given by

$$\Sigma_M + RA\Sigma_D A^t R^t, \tag{25}$$

where R is the rotation matrix from lab space to sample space and $A = K^{-2}\boldsymbol{k}_{\mathrm{f}}\boldsymbol{k}_{\mathrm{i}}^t - \mathbf{1}$. Note that the matrix A depends only on $\boldsymbol{k}_{\mathrm{f}}$, i.e. on the detector pixel location, and the matrix R depends on the sample orientation, i.e. the frame number.

Now make a simplifying assumption, $\Sigma_M = \sigma_M^2\mathbf{1}$ and $\Sigma_D = \sigma_D^2\mathbf{1}$. For a set of observations (Σ_i, R_i, A_i), we form the loss function

$$L(\sigma_M^2, \sigma_D^2) = \sum_i |\sigma_M^2 + \sigma_D^2(R_iA_i)(R_iA_i)^t - \Sigma_i|^2 \tag{26}$$

Setting $\nabla L = 0$ we obtain the 2x2 system of linear equations...

$$\begin{bmatrix} 3N & \sum_i \mathrm{tr}((R_iA_i)(R_iA_i)^t) \\ \sum_i \mathrm{tr}((R_iA_i)^t(R_iA_i)) & \sum_i \mathrm{tr}(((R_iA_i)^t(R_iA_i))^2) \end{bmatrix} \begin{bmatrix} \sigma_M^2 \\ \sigma_D^2 \end{bmatrix} = \begin{bmatrix} \sum_i \mathrm{tr}(\Sigma_i) \\ \sum_i \mathrm{tr}((R_iA_i)^t\Sigma_i(R_iA_i)) \end{bmatrix} \tag{27}$$

which is easily solved. One can also solve for the the full covariance matrices Σ_M, Σ_D via gradient descent, since the gradient is easily computed analytically.. **Note: tested this out in Python, and it seems to work pretty well. So the assumptions may be justified!**

Now, if we work in lab-based q-space, under the simplifying assumptions above, we find a covariance matrix

$$\Sigma = \sigma_M^2\mathbf{1} + \sigma_D^2 A_i A_i^t \tag{28}$$

Now let

$$\boldsymbol{e}_1 = (\boldsymbol{k}_{\mathrm{f}} \times \boldsymbol{k}_{\mathrm{i}})/|\boldsymbol{k}_{\mathrm{f}} \times \boldsymbol{k}_{\mathrm{i}}| \tag{29}$$
$$\boldsymbol{e}_2 = (\boldsymbol{k}_{\mathrm{f}} \times e_1)/|\boldsymbol{k}_{\mathrm{f}} \times e_1| \tag{30}$$
$$\boldsymbol{e}_3 = (\boldsymbol{k}_{\mathrm{f}} + \boldsymbol{k}_{\mathrm{i}})/|\boldsymbol{k}_{\mathrm{f}} + \boldsymbol{k}_{\mathrm{i}}| \tag{31}$$

3.1 Kabsch's Coordinate System

In [14] Kabsch introduced a per-peak coordinate system intented to undo effects from detector geometry. See also [13] for an updated description of the coordinates and integration technique. The basis introduced by Kabsch is the following:

$$\boldsymbol{e}_1 = (\boldsymbol{q} \times \boldsymbol{k}_{\mathrm{i}})/|\boldsymbol{q} \times \boldsymbol{k}_{\mathrm{i}}| \tag{32}$$
$$\boldsymbol{e}_2 = (\boldsymbol{q} \times \boldsymbol{e}_1)/|\boldsymbol{q} \times \boldsymbol{e}_1| \tag{33}$$
$$\boldsymbol{e}_3 = (\boldsymbol{k}_{\mathrm{f}} + \boldsymbol{k}_{\mathrm{i}})/|\boldsymbol{k}_{\mathrm{f}} + \boldsymbol{k}_{\mathrm{i}}| \tag{34}$$

with corresponding coordinates

$$\epsilon_1 = \boldsymbol{e}_1 \cdot (\boldsymbol{k}'_{\mathrm{f}} - \boldsymbol{k}_{\mathrm{f}})/|\boldsymbol{k}_{\mathrm{f}}| \tag{35}$$

$$\epsilon_2 = \boldsymbol{e}_2 \cdot (\boldsymbol{k}'_{\mathrm{f}} - \boldsymbol{k}_{\mathrm{f}})/|\boldsymbol{k}_{\mathrm{f}}| \tag{36}$$

$$\epsilon_3 = \boldsymbol{e}_3 \cdot (R_{\phi'-\phi}\boldsymbol{q} - \boldsymbol{q})/|\boldsymbol{q}| \tag{37}$$

The coordinates ϵ_1, ϵ_2 correspond to the angular distribution (in radians) of the peak, as if it were measured on the Ewald sphere. Hence this corresponds to beam divergence and we may model the intensity distribution as $\exp(-(\epsilon_1^2 + \epsilon_2^2)/2\sigma_D^2)$.

To understand the last coordinate, consider the following. Take a peak with center $\boldsymbol{q}$ and consider a nearby point $\boldsymbol{q}'$. We project $\boldsymbol{q}'$ back to the Ewald sphere by rotating along the axis $\boldsymbol{e}_1$ (which is the normal of the plane containing $\boldsymbol{k}_{\mathrm{f}}$ and $\boldsymbol{k}_{\mathrm{i}}$). The velocity of q when it crosses the Ewald sphere by rotating along this axis is $\boldsymbol{e}_1 \times \boldsymbol{q}$. It is easy to verify that

$$\boldsymbol{e}_1 \times \boldsymbol{q} = |q|\boldsymbol{e}_3 \tag{38}$$

and therefore the coordinate ϵ_3 may be interpreted as (approximately) and angular distance from the Ewald sphere.

To better understand $\boldsymbol{e}_3$, consider the following: we want to find the axis $\boldsymbol{a}$ such that $\boldsymbol{q}$ passes through the Ewald sphere as fast as possible. Hence, we want to maximize $(\boldsymbol{a} \times \boldsymbol{q}) \cdot \boldsymbol{k}_{\mathrm{f}}$ subject to the constraint $\boldsymbol{a} \cdot \boldsymbol{a} = 1$. Now $(\boldsymbol{a} \times \boldsymbol{q}) \cdots \boldsymbol{k}_{\mathrm{f}}) = \boldsymbol{a} \cdot (\boldsymbol{a} \times \boldsymbol{k}_{\mathrm{f}}) = \boldsymbol{a} \cdot (\boldsymbol{k}_{\mathrm{f}} \times \boldsymbol{k}_{\mathrm{i}})$, so by the method of Langrange multipliers we must solve $\boldsymbol{k}_{\mathrm{f}} \times \boldsymbol{k}_{\mathrm{i}} = \lambda \boldsymbol{a}$, which tells us immediately that the axis is in the direction of $\boldsymbol{e}_1$.

4 Integration

4.1 Pixel sum method

Consider a peak centered at q_0 in (sampled-fixed) reciprocal space. The background-subtracted intensity distribution in a neighborhood of q_0 has a covariance matrix Σ, which is either computed (for strong peaks) or predicted (for weak peaks). The integration routine takes three parameters r_1, r_2, r_2 and produces two sets $\mathcal{B}$ and $\mathcal{P}$ of background and peak events, respectively, as follows:

$$\mathcal{P} = \{i \mid (q_i - q_0)^t \Sigma^{-1} (q_i - q_0) < r_1^2\} \tag{39}$$

$$\mathcal{B} = \{i \mid r_2^2 < (q_i - q_0)^t \Sigma^{-1} (q_i - q_0) < r_3^2\} \tag{40}$$

The local background is estimated as

$$\mu_B = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} M_i \tag{41}$$

$$\sigma_B^2 = \frac{1}{|\mathcal{B}| - 1} \sum_{i \in \mathcal{B}} (M_i - \mu_B)^2 \tag{42}$$

Note that in the case of Poisson statistics, we should have $\mu_B \approx \sigma_B^2$. By recording the pairs (μ_B, σ_B^2) from strong peaks, we can estimate the deviation from Poisson statistics and use this to improve error estimates (more later).

The integrated peak intensity is then estimated as

$$I = \sum_{i \in \mathcal{P}} (M_i - \mu_B) \tag{43}$$

$$\sigma_I^2 = I + \frac{|\mathcal{P}|^2}{|\mathcal{B}|} \overline{B} \tag{44}$$

It should be noted that for weak peaks, the error is dominated by the error in the background estimate, and therefore we will obtain unsatisfactorily low values of I/σ_I.

4.2 Fitted Intensity

As shown in [15], the integration error for weak peaks is dominated by background subtraction and it is typically better to find the integrated intensity by fitting to a profile learned from strong peaks.

3D profile fitting is used by XDS [2] and is described in some detail in [14, 13].

As in the previous subsection, using a covariance matrix and a parameters $r_1 < r_2 < r_3$ we produce sets $\mathcal{P}$ and $\mathcal{B}$ of peak and background points. Assume that we know the resolution function R_i, normalized as

$$\sum_i R_i = 1. \tag{45}$$

We model the observed intensities M_i as

$$M_i \simeq B + IR_i, \tag{46}$$

where B, I are the mean background and integrated intensity, yet to be fit. To find optimal values of B, I we minimize the chi-squared loss

$$\chi^2 = \sum_{i \in \mathcal{P}} \frac{(B + IR_i - M_i)^2}{\sigma_i^2}. \tag{47}$$

For a fixed set of variances, minimizing χ^2 reduces to the 2x2 linear system below:

$$\begin{bmatrix} \sum 1/\sigma_i^2 & \sum R_i/\sigma_i^2 \\ \sum R_i/\sigma_i^2 & \sum R_i^2/\sigma_i^2 \end{bmatrix} \begin{bmatrix} B \\ I \end{bmatrix} = \begin{bmatrix} \sum M_i/\sigma_i^2 \\ \sum M_i R_i/\sigma_i^2 \end{bmatrix} \tag{48}$$

Write this equation as $Ax = b$. It is easy to compute that the covariance matrix of b is exactly the coefficient matrix A, and therefore the variance-covariance matrix of the solution vector $x = (B, I)$ is given by A^{-1}.

The solution given above depends on the pixel uncertainties σ_i^2. As suggested by Kabsch 2010, we solve this iteratively. To begin, we set all σ_i^2 equal to some fixed value, say 1. This allows us to solve for B and I. We then put the solved values into the error model

$$\sigma_i^2 = B + IR_i \tag{49}$$

and iterate until either I becomes negative, or (B, I) do not change within some given convergence criterion.

Bayesian approach [JWu apr19]: Determine expectation values or most probable values of B, I from the conditional probability

$$p(B, I|M) \propto p(M|B, I)p(B)p(I). \tag{50}$$

Count statistics of the single pixels are independent of each other, hence

$$p(M|B, I) = \prod_i c(M_i|B, I; R_i) \tag{51}$$

with the single-pixel count probability distribution given by Poisson statistics,

$$c(m|B, I; r) = \frac{\lambda^m \mathrm{e}^{-\lambda}}{m!} \tag{52}$$

with $\lambda = B + Ir$.

Rewrite (50) as

$$\ln p(B, I|M) = \sum_i \Big\{ M_i \ln(B + IR_i) - (B + IR_i) \Big\} + \ln p(B) + \ln p(I) + \text{const.} \tag{53}$$

Let N pixels contribute to the sum. Use the normalization (45). Then

$$\ln p(B, I|M) = \sum_i \Big\{ M_i \ln(B + IR_i) \Big\} - (NB + I) + \ln p(B) + \ln p(I) + \text{const.} \tag{54}$$

We now must specify the a priori distributions of B and I. This cannot be done without some arbitrariness. For instance, assume equal probability per decade within given limits. Then

$$p(I) = \frac{1}{\ln(I_+/I_-)} \frac{1}{I}, \tag{55}$$

and similarly for B, lest we breed a better idea.

Now, compute the most probable parameter values from

$$\begin{aligned} \partial \ln p(B, I|M)/\partial B = 0, \\ \partial \ln p(B, I|M)/\partial I = 0. \end{aligned} \tag{56}$$

4.3 I/σ Integration

This is the integration technique used by RETREAT [1]. The method is described in detail in [16]. In the article [17] there is a detailed comparison between this method and profile fitting.

For a given peak with mean background μ_b, center x_0, and covariance matrix Σ, define

$$X_s = \{x \mid (x - x_0)^t \Sigma^{-1} (x - x_0) \le s^2\} \tag{57}$$

$$I_s = \sum_{X_\sigma} I_x \tag{58}$$

Then the error of I_σ can be estimated (assuming Poisson statistics) as

$$\sigma^2(I_s) = I_s + n_s(1 + \frac{n_s}{n_b})\overline{B} \tag{59}$$

where $n_s = |X_s|$ is the number of points contributing to I_σ and n_b is the number of points used for background estimation.

Important Remark: The function I_σ is, to a good approximation, *independent of the coordinate system x*. It is an *intrinsic* property of the intensity distribution, independent of the coordinates used to express the distribution. We therefore do not have to worry about changes of coordinates, as in Kabsch's paper.

Now, suppose that we take some value t to be the cutoff for strong peak integration. We can define the integrated peak profile

$$p_s := I_s/I_t \tag{60}$$

The uncertainty in p_s:

$$\sigma^2(p_s) = \frac{\sigma^2(I_s)}{I_t^2} - 2\frac{I_s}{I_t^3}\mathrm{cov}(I_s, I_t) + \frac{I_s^2}{I_t^4}\sigma^2(I_t) \tag{61}$$

Assuming $s <$, we have

$$\mathrm{cov}(I_s, I_t) = I_s + n_s(1 + n_t/n_b)\overline{B} \tag{62}$$

and therefore we have everything we need to estimate p_s and $\sigma^2(p_s)$. Finally, if we have N independent strong peaks with measured profiles $p_s^i, \sigma^2(p_s^i)$, then (assuming the peaks are non-overlapping) we can estimate the true profile as

$$\hat{\mathbf{p}}_s = N^{-1}\sum_i p_s^i \tag{63}$$

$$\sigma^2(\hat{\mathbf{p}}_s) = N^{-2}\sum_i \sigma^2(p_s^i) \tag{64}$$

Assumptions: We now assume that the intensity distributions for all peaks are approximately equal, or at least slowly varying as a function of detector position and sample orientation. Therefore, we model the function I_σ as

$$I_\sigma = I_0 p(\sigma), \tag{65}$$

where I_0 is the "true" integrated intensity and $p(\sigma)$ is a function independent of the particular peak. Given a collection of N strong peaks, we can estimate $p(\sigma)$ as

$$p_\sigma = \frac{1}{N}\sum_i \frac{I_\sigma^i}{I_0^i} \tag{66}$$

$$\sigma^2(p_\sigma) = \frac{1}{N^2}\sum_i \sigma^2\left(\frac{I_\sigma^i}{I_0^i}\right) \tag{67}$$

Remark When calculating $\sigma^2(I_\sigma/I_0)$ be very careful, because I_σ and I_0 are definitely correlated!! Assuming $s < t$, and the sets of peak points and backgruond points are disjoint, *and Poisson statistics*, we have

$$\mathrm{cov}(I_s, I_t) = I_s + n_s(1 + n_t/n_b)\bar{B} \tag{68}$$

Now, suppose that we estimate the true intensity as $I = I_t$ for some t. Then for $s < t$ we have

$$\sigma^2(p_s) = \frac{\sigma^2(I_s)}{I_t^2} + \frac{I_s^2}{I_t^4}\sigma^2(I_t) - 2\frac{I_s}{I_t^3}\mathrm{cov}(I_s, I_t) \tag{69}$$

Integration Method: Now suppose we have a good estimate of $p_\sigma, \sigma^2(p_\sigma)$ and we have computed I_σ for some weak peak (note: this assumes we can accurately predict the covariance matrix; see below). From the model intensity distribution, we have $I_\sigma \approx I_o p_\sigma$, and therefore $I_0 \approx I_\sigma/p_\sigma$. We have

$$\sigma^2(I_\sigma/p_\sigma) \approx \frac{\sigma^2(I_\sigma)}{p_\sigma^2} + \frac{I_\sigma^2}{p_\sigma^4}\sigma^2(p_\sigma) \tag{70}$$

Therefore, the relative error $\sigma^2(I_\sigma/p_\sigma)/(I_\sigma/p_\sigma)^2$ is

$$\frac{\sigma^2(I_\sigma/p_\sigma)}{(I_\sigma/p_\sigma)^2} \approx \frac{\sigma^2(I_\sigma)}{I_\sigma^2} + \frac{\sigma^2(p_\sigma)}{p_\sigma^2} \tag{71}$$

The fitted intensity is then defined to be

$$I_{\mathrm{fit}} = I_{s'}/p_{s'} \tag{72}$$

$$s' = \underset{s}{\mathrm{argmin}} \left(\frac{\sigma^2(I_s)}{I_s^2} + \frac{\sigma^2(p_s)}{p_s^2} \right) \tag{73}$$

4.4 Combining estimates

In the previous subsections, we have described two methods of integration, producing estimates I_{fit} and I_{sum}, and also how to estimate their errors. Provided our error estimate is reasonable, we can combine these estimates into an even stronger estimate I_{wt}, which is a weighted linear combination of I_{fit} and I_{sum}:

$$I_{\mathrm{wt}} = aI_{\mathrm{fit}} + bI_{\mathrm{sum}}, \tag{74}$$

subject to $a + b = 1$. We wish to minimize $\sigma^2_{I_{\mathrm{wt}}}$, which is equal to

$$\sigma^2_{\mathrm{wt}} = a^2\sigma^2_{\mathrm{fit}} + b^2\sigma^2_{\mathrm{sum}} + 2ab\mathrm{cov}(I_{\mathrm{fit}}, I_{\mathrm{sum}}). \tag{75}$$

This is easily minimized using the method of Lagrange multipliers. The solution is $(a, b) = x/(x_1 + x_2)$, where $x = \Sigma^{-1}\begin{bmatrix}1\\1\end{bmatrix}$ and Σ is the variance-covariance matrix of I_{fit} and I_{sum}.

Now, from the previous subsections, both I_{fit} and I_{sum} have an explicit, linear dependence on the measured intensities M_i. Assuming that M_i and M_j are independent for $i \neq j$, and that $\mathcal{P}$ and $\mathcal{B}$ are disjoint sets, we can now easily and explicitly compute the covariance $\mathrm{cov}(I_{\mathrm{fit}}, I_{\mathrm{sum}})$.

4.5 Peak Shape Model

Motivated by the Monte-Carlo technique of EVAL-14 and EVAL-15, we consider a probabilistic approach to estimate the covariance matrix of the intensity distribution for a given peak.

For a given detector event recorded at (x, y, θ), we have a corresponding reciprocal vector

$$q(x, y, \theta) = R_\theta(\boldsymbol{k}_{\mathrm{f}}(x, y) - \boldsymbol{k}_{\mathrm{i}}). \tag{76}$$

Note that by mapping to sample-fixed momentum space, we automatically undo distortion effects due to detector geometry. Expressed in these coordinates, the distribution of q values is a function of beam divergence (and wavelength uncertainty), mosaicity, and the shape of the crystal. Since the recorded peak is essentially a convolution of these effects, we can express the covariance matrix of the intensity distribution in sample-fixed momentum space as

$$\Sigma = P\Sigma_M P^t + |k|^2 J_k \Sigma_D J_k^t + J_p \Sigma_S J_p^t \tag{77}$$

where

- P is the projection matrix $|q|^2 1_{3\times 3} - qq^t$
- $J_k = \partial q/\partial k$ is the derivative of q with respect to incoming momentum k
- $J_p = \partial q/\partial p$ is the derivative of q with respect to sample position p
- Σ_M is a 3x3 covariance matrix representing the mosaicity of the crystal
- Σ_D is a 3x3 covariance matrix representing beam divergence and wavelength distribution
- Σ_S is a 3x3 covariance matrix representing the shape of the crystal in the beam

Remarks

- This model assumes that there is no covariance between beam divergence, scattering point, and mosaicity. This seems reasonable for typical experiments but could be violated by some extreme cases.
- P_q, J_k, J_p are all determined by experiment setup and geometry; can be computed analytically.
- For isotropic mosaicity, can take $\Sigma_M = \sigma_M^2 1_{3\times 3}$
- Similarly, can set off-diagonal components of Σ_D to zero if it is known there is no correlation.
- Σ_S represents the shape *in the beam for a fixed orientation of the sample*, and therefore will typically need to be updated every few frames (unless e.g. the crystal is small and approximately spherical).
- Without imposing any constraints, the model has 3x6=18 free parameters, and each strong peak covariance matrix gives a 6-component observation (due to symmetry), so even with just a few peaks the problem is massively overdetermined, which *should* result in very stable fits.
- Because this model is *physical*, once we have determined the parameters it makes sense to *extrapolate*, not just *interpolate*. So it should be helpful for predicting weak peaks at high q.

5 Measures of Data Quality

5.1 R factors

For the definition of R_{merge} see [18]. For R factors in general there is the paper [12].

5.2 The correlation coefficients $CC_{1/2}$ and $CC*$

The statistic $CC_{1/2}$ as introduced in [18]. The statistic $CC_{1/2}$ is defined as follows. Randomly divide the unmerged dataset into two subsets. For each symmetry-equivalence class $[hkl]$, we have a merged intensity x_{hkl} from the first set and y_{hkl} from the second set. $CC_{1/2}$ is defined as the Pearson correlation coefficient of the joint measurements (x_{hkl}, y_{hkl}).

$$CC_{1/2} := \rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y}, \tag{78}$$

where ρ denotes the Pearson correlation coefficient. Note that this depends on the choice of division of the unmerged datasets into two subsets, so it is itself a random variable. (However, under some assumptions, one can check that its variance should be small.)

Let J_{hkl} denote the true intensity (we use J instead of I to distinguish this from our measured and/or merged intensities). Then define random variables $\xi := x - J$ and $\eta := y - J$. We make the following assumption: ξ and η are independent with mean zero, that $\sigma_\xi = \sigma_\eta$, and that ξ, η are uncorrelated with J.

Since ξ, η are uncorrelated with J,

$$\sigma_x^2 = \sigma_J^2 + \sigma_\xi^2 \tag{79}$$

$$\sigma_y^2 = \sigma_J^2 + \sigma_\eta^2 = \sigma_J^2 + \sigma_\xi^2 \tag{80}$$

Then

$$\rho(x, y) = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} \tag{81}$$

$$= \frac{\operatorname{cov}(J + \xi, J + \eta)}{\sigma_x \sigma_y} \tag{82}$$

$$= \frac{\sigma_J^2 + \operatorname{cov}(\xi, J) + \operatorname{cov}(\eta, J) + \operatorname{cov}(\xi, \eta)}{\sigma_x \sigma_y} \tag{83}$$

$$= \frac{\sigma_J^2}{\sigma_J^2 + \sigma_\xi^2} \tag{84}$$

Thus we have

$$CC_{1/2} = \sigma_J^2 / \left(\sigma_J^2 + \sigma_\xi^2\right) \tag{85}$$

This expression will be useful in the following subsection.

5.3 CC_{true}

Let x, y, ξ, η, J be as in the previous subsection. Define

$$I = \frac{x + y}{2} \tag{86}$$

denote the merged intensities of the entire dataset. Then CC_{true} is defined to be the Pearson correlation coefficient of I and the true intensities J:

$$CC_{\text{true}} = \rho(I, J) = \frac{\text{cov}(I, J)}{\sigma_I \sigma_J} \tag{87}$$

Since in most cases we do not know the true intensities, this definition is not directly useful.

Making the same assumptions about measurement error as in the previous subsection, we have

$$\sigma_z^2 = \frac{1}{4}\sigma_x^2 + \frac{1}{4}\sigma_y^2 + \frac{1}{2}\text{cov}(x, y) \tag{88}$$

$$= \sigma_J^2 + \frac{1}{2}\sigma_\xi^2 \tag{89}$$

and furthermore,

$$\text{cov}(I, J) = \text{cov}(J + \frac{\xi + \eta}{2}, J) = \sigma_J^2. \tag{90}$$

Therefore,

$$CC_{\text{true}} = \frac{\sigma_J}{\sqrt{\sigma_J^2 + \frac{1}{2}\sigma_\epsilon^2}}. \tag{91}$$

From equation 85, we can express σ_ξ^2 as $\sigma_J^2(1/CC_{1/2} - 1)$. Putting this into the above expression for CC_{true}, we have

$$CC_{\text{true}} = \frac{\sigma_J}{\sqrt{\sigma_J^2 + \frac{1}{2}\sigma_\xi^2}} = \frac{\sigma_J}{\sqrt{\sigma_J^2 + \frac{1}{2}\sigma_J^2(1/CC_{1/2} - 1)}} \tag{92}$$

$$= \frac{1}{\sqrt{\frac{1}{2} - \frac{1}{2CC_{1/2}}}} = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}, \tag{93}$$

which amazingly is a function of $CC_{1/2}$ only. We therefore define

$$CC* := \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}, \tag{94}$$

to be an estimate of CC_{true}, which can be calculated directly from the data. The statistic was introduced in [18].

6 Other topics

6.1 Absorption correction

TODO

6.2 Monte-Carlo profile prediction

See papers [19, 20]. We have implemented a preliminary version of this algorithm, however it is definitely in need of testing and improvement.

References

[1] R. F. Stansfield and G. J. McIntyre, *RETREAT-PEAKINT: Program Manual*, Institut Laue-Langevin (2013).

[2] W. Kabsch, Acta Cryst. D **66**, 125 (2010).

[3] Z. Otwinowski and W. Minor, Methods in Enzymology **276**, 307 (1997).

[4] Z. Otwinowski and W. Minor, in *Macromolecular Crystallography*volume F of *International Tables for Crystallography* (International Tables for Crystallography F), Springer (2006).

[5] D. Gewirth, *The HKL Manual*.

[6] H. Burzlaff, H. Grimmer, B. Gruber, P. M. de Wolff and H. Zimmermann, in *Advanced topics on space-group symmetry*volume A of *International Tables for Crystallography* (International Tables for Crystallography A), Springer (2016).

[7] I. Steller, R. Bolotovsky and M. G. Rossmann, J. Appl. Cryst. **30**, 1036 (1997).

[8] N. K. Sauter, R. W. Grosse-Kunstleve and P. D. Adams, J. Appl. Cryst. **37**, 399 (2004).

[9] B. Gruber, Acta Cryst. A **29**, 433 (1973).

[10] I. Krivy and B. Gruber, Acta Cryst. A **32**, 297 (1976).

[11] R. W. Grosse-Kunstleve, N. K. Sauter and P. D. Adams, Acta Cryst. A **60**, 1 (2004).

[12] P. R. Evans, Acta Cryst. D **67**, 282 (2011).

[13] W. Kabsch, Acta Cryst. D **66**, 133 (2010).

[14] W. Kabsch, J. Appl. Cryst. **21**, 916 (1988).

[15] R. Diamond, Acta Cryst. A **25**, 43 (1969).

[16] C. Wilkinson, H. W. Khamis, F. D. Stansfield and G. J. McIntyre, J. Appl. Cryst. **21**, 471 (1988).

[17] E. Prince, C. Wilkinson and G. J. McIntyre, J. Appl. Cryst. **30**, 133 (1997).

[18] P. A. Karplus and K. Diederichs, Science **336**, 1030 (2012).

[19] A. J. M. Duisenberg, L. M. J. Kroon-Batenberg and A. M. M. Schreurs, J. Appl. Cryst. **36**, 220 (2003).

[20] A. M. M. Schreurs, X. Xian and L. M. J. Kroon-Batenberg, J. Appl. Cryst. **43**, 70 (2010).