Merging of equivalent reflections

Luc J. Bourhis
(Bruker-AXS SAS, France)

September 5, 2012

Two steps are required to merge equivalent reflections with the `cctbx`. Given a `miller.array` `m`,

1. `m1 = m.map_to_asu()` projects each Miller index into the asymmetric unit, i.e. for each group of equivalent reflection, each index of that group is replaced by the same Miller index;

2. `merging = m1.merge_equivalents()` finds the group of identical Miller indices, gathers the data and sigma's for each group in turn, computes an average datum and an associated sigma; `merging.array()` is then the `miller.array` containing those unique indices associated to those averaged data and sigma.

The first step is only about space-group algebra whereas the second step is only about statistics and this division is therefore optimally orthogonal in a sense. We will now expound each step, starting from the second one.

1 Averaging of equivalent reflections

Given n data $y_1, \ldots, y_n$ and the associated estimated standard deviations $\sigma_1, \ldots, \sigma_n$, either the amplitudes or the intensities for a group of symmetry equivalent reflections, we sought to combine those data and sigma's into a single datum and an associated standard deviation.

That merged amplitude or intensity $\bar{y}$ is computed as a weighted average of the $\{y_i\}_{i=1,\ldots,n}$,

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}. \tag{1}$$

There are two ways to handle this from a statistical point of view.

1.1 External variance

The first one gives a mathematical meaning to the loose assertion that all y_i should be equal within the uncertainties quantified by the σ_i (the exact equality is required by those being equivalent reflections but this is spoiled by all sources of errors in measurement and data processing up to this point). Each y_i is then seen as an outcome of a random variable $\hat{y}_i$ which is an unbiased estimator for the value y_{eq} that all equivalent reflections should ideally share, i.e. mathematically

$$E(\hat{y}_i) = y_{\text{eq}},\ \forall i = 1, \ldots, n \qquad (2)$$
$$V(\hat{y}_i) = \sigma_i^2.$$

Then the average $\bar{y}$ is the outcome of the random variable

$$\hat{y} = \frac{\sum_{i=1}^n w_i \hat{y}_i}{\sum_{i=1}^n w_i}. \qquad (3)$$

which is obviously an unbiased estimator of y_{eq} (i.e. $E(\hat{y}) = y_{\text{eq}}$). If we postulate that the measurement and data reduction lead to uncorrelated $\hat{y}_i$, then

$$V(\hat{y}) = \sum_{i=1}^n \omega_i^2 V(\hat{y}_i) \qquad (4)$$

where

$$\omega_i = \frac{w_i}{\sum_{i=1}^n w_i}. \qquad (5)$$

This is often called the "external" variance. Its lowest possible value is obtained for the weights

$$\tilde{w}_i = \frac{1}{V(\hat{y}_i)} = \frac{1}{\sigma_i^2}, \qquad (6)$$

as well as for any weights differing from those by a common proportionality factor, as demonstrated in appendix A and this minimum is equal to

$$V(\hat{y}) = \frac{1}{\sum_{i=1}^n \tilde{w}_i} = \frac{1}{n \langle \tilde{w}_i \rangle}. \qquad (7)$$

Those are the weights and the external variance used by the `cctbx`.

This is not the only popular choice. Indeed ShelXL [?] uses instead

$$w_i = \begin{cases} \frac{y_i}{\sigma_i^2} & \text{if } \frac{y_i}{\sigma_i} > 3, \\ \frac{3}{\sigma_i} & \text{otherwise.} \end{cases} \qquad (8)$$

1.2 Internal variance

The second way to handle the average (1) is to consider it as a mere sample mean, but a weighted one, ignoring the special property of the y_i. Those data are considered as the outcome of a sample $(Y_1, \ldots, Y_n)$ of a random variable Y, and $\bar{y}$ is then the outcome of the unbiased estimator of $E(Y)$,

$$\bar{Y} = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}. \qquad (9)$$

It is then natural to also compute a weighted sample variance

$$S^2 = \frac{\sum_{i=1}^n w_i (Y_i - \bar{Y})^2}{\sum_{i=1}^n w_i}. \tag{10}$$

However, it is a biased estimator of $V(Y)$, as it is well-known in the unweighted case, i.e. all weights w_i equal. The unbiased estimator

$$S_{n-1}^2 = \frac{S^2}{1 - \sum_{i=1}^n \omega_i^2} \tag{11}$$

$$= \frac{\sum_{i=1}^n w_i}{\left(\sum_{i=1}^n w_i\right)^2 - \sum_{i=1}^n w_i^2} \sum_{i=1}^n w_i (Y_i - \bar{Y})^2 \tag{12}$$

is therefore preferred. Those variances are called "internal" as opposed to the variance we have previously discussed. The `cctbx` computes it by using an instance of `scitbx::mean_and_variance` and calling its member function `gsl_stats_wvariance` whose implementation and naming follows the function with the same name in the GNU Scientific Library [**?**]. Since this formula is not that easily found in textbooks, we demonstrate it in appendix B.

Finally, it is customary to estimate the variance associated with $\bar{y}$ by taking the greatest of the internal and external variance. That is what the `cctbx` does as well as ShelXL.

Appendix A Minimum variance weights

We will demonstrate eqn (6). We seek the solution of the constrained minimisation problem

$$\min V(\hat{y}), \tag{13}$$

$$V(\hat{y}) = \sum_{i=1}^n \omega_i^2 V(\hat{y}_i), \tag{14}$$

$$\sum_{i=1}^n \omega_i = 1. \tag{15}$$

We can solve it by minimising the Lagrangian

$$L = V(\hat{y}) - \lambda \sum_{i=1}^n \omega_i, \tag{16}$$

$$= \sum_{i=1}^n \left[V(\hat{y}_i) \left(\omega_i - \frac{\lambda}{2V(\hat{y}_i)} \right)^2 - \frac{\lambda^2}{4V(\hat{y}_i)} \right] \tag{17}$$

Thus L reaches its minimum at

$$\omega_i = \frac{\lambda}{2V(\hat{y}_i)} \tag{18}$$

and using eqn (15), it comes

$$\frac{\lambda}{2} = \frac{1}{\sum_{i=1}^{n} \frac{1}{V(\hat{y}_i)}} \tag{19}$$

and therefore the minimum is reached at

$$\omega_i = \frac{\frac{1}{V(\hat{y}_i)}}{\sum_{j=1}^{n} \frac{1}{V(\hat{y}_i)}}. \tag{20}$$

That demonstrates eqn (6) and since weights differing by a common proportionality factor yield the same ω_i, QED.

Appendix B Weighted sample variance

First let us remember that, by definition of a sample,

$$E(Y_i) = E(Y),\ \forall i = 1, \dots, n \tag{21}$$
$$V(Y_i) = V(Y) \tag{22}$$

Therefore,

$$\begin{aligned} V(Y) &= \sum_{i=1}^{n} \omega_i V(Y_i) \\ &= E\left[\sum_{i=1}^{n} \omega_i (Y_i - E(Y))^2\right] \\ &= E\left[\sum_{i=1}^{n} \omega_i (Y_i - \bar{Y})^2\right] + 2E\left[\sum_{i=1}^{n} \omega_i (Y_i - \bar{Y})(\bar{Y} - E(Y))\right] + \sum_{i=1}^{n} \omega_i E\left[(\bar{Y} - E(Y))^2\right] \end{aligned}$$

Then,

- since $E(\bar{Y}) = E(Y)$, the last term is $V(\bar{Y})$;
- by definition of $\bar{Y}$, $\sum_{i=1}^{n} \omega_i (Y_i - \bar{Y}) = 0$ and the second term is therefore 0.

Thus

$$V(Y) = E(S^2) + V(\bar{Y}). \tag{23}$$

But

$$V(\bar{Y}) = \sum_{i=1}^{n} \omega_i^2 V(Y) \tag{24}$$

and therefore

$$V(Y) = \frac{E(S^2)}{1 - \sum_{i=1}^{n} \omega_i^2} \tag{25}$$