18.5.6. The diffraction-component precision index

18.5.6.1. *Statistical expectation of error dependence*

From general statistical theory, one would expect the s.u. of an atomic coordinate determined from the diffraction data alone to show dependence on four factors:

$$\sigma(x) \propto (\mathcal{R})\left[(n_{\rm atoms})/(n_{\rm obs} - n_{\rm params})\right]^{1/2}(1/s_{\rm rms}). \qquad (18.5.6.1)$$

Here, $\mathcal{R}$ is some measure of the precision of the data; $n_{\rm atoms}$ is the recognition that the information content of the data has to be shared out; $n_{\rm obs}$ is the number of independent data, but to achieve the correct number of degrees of freedom this must be reduced by $n_{\rm params}$, the number of parameters determined; and $1/s_{\rm rms}$ is a more specialized factor arising from the sensitivity $\partial|F|/\partial x$ of the data to the parameter x. Here $s_{\rm rms}$ is the r.m.s. reciprocal radius of the data. Any statistical error estimate must show some correspondence to these four factors.

18.5.6.2. *A simple error formula*

Cruickshank (1960) offered a simple order-of-magnitude formula for $\sigma(x)$ in small molecules. It was intended for use in experimental design: how many data of what precision are needed to achieve a given precision in the results? The formula, derived from a very rough estimate of a least-squares diagonal element in non-centrosymmetric space groups, was

$$\sigma(x_i) = (1/2)(N_i/p)^{1/2}[R/s_{\rm rms}] \qquad (18.5.6.2)$$

Here $p = n_{\rm obs} - n_{\rm params}$, R is the usual residual $\sum|\Delta F|/\sum|F|$ and N_i is the number of atoms of type i needed to give scattering power at $s_{\rm rms}$ equal to that of the asymmetric unit of the structure, *i.e.*, $\sum_j f_j^2 \equiv N_i f_i^2$. [The formula has also proved very useful in a systematic study of coordinate precision in the many thousands of small-molecule structure analyses recorded in the Cambridge Structural Database (Allen *et al.*, 1995*a*,*b*).]

For small molecules, the above definition of N_i allowed the treatment of different types of atom with not-too-different B's. However, it is not suitable for individual atoms in proteins where there is a very large range of B values and some atoms have B's so large as to possess negligible scattering power at $s_{\rm rms}$.

Often, as in isotropic refinement, $n_{\rm params} \simeq 4n_{\rm atoms}$, where $n_{\rm atoms}$ is the total number of atoms in the asymmetric unit. For fully anisotropic refinement, $n_{\rm params} \simeq 9n_{\rm atoms}$.

A first very rough extension of (18.5.6.2) for application in proteins to an atom with $B = B_i$ is

$$\sigma(x_i) = k(N_i/p)^{1/2}\left[g(B_i)/g(B_{\rm avg})\right]C^{-1/3}Rd_{\rm min}, \qquad (18.5.6.3)$$

where k is about 1.0, $N_i = \sum Z_j^2/Z_i^2$, $B_{\rm avg}$ is the average B for fully occupied sites and C is the fractional completeness of the data to $d_{\rm min}$. In deriving (18.5.6.3) from (18.5.6.2), $1/s_{\rm rms}$ has been replaced by $1.3d_{\rm min}$, and the factor $(1/2)(1.3) = 0.65$ has been increased to 1.0 as a measure of caution in the replacement of a full matrix by a diagonal approximation. $g(B) = 1 + a_1B + a_2B^2$ is an empirical function to allow for the dependence of $\sigma(x)$ on B. However, the results in Section 18.5.4.2 showed that the parameters a_1 and a_2 depend on the structure.

As also mentioned in Section 18.5.4.2, Sheldrick has found that the Z_i in N_i is better replaced by $Z_i^{\#}$, the scattering factor at $\sin\theta/\lambda = 0.3$ Å^{-1}. Hence, N_i may be taken as

$$N_i = (\textstyle\sum Z_j^{\#2}/Z_i^{\#2}). \qquad (18.5.6.4)$$

A useful comparison of the relative precision of different structures may be obtained by comparing atoms with the respective $B = B_{\rm avg}$ in the different structures. (18.5.6.3) then reduces to

$$\sigma(x, B_{\rm avg}) = 1.0(N_i/p)^{1/2}C^{-1/3}Rd_{\rm min}. \qquad (18.5.6.5)$$

The smaller the $d_{\rm min}$ and the R, the better the precision of the structure. If the difference between oxygen, nitrogen and carbon atoms is ignored, N_i may be taken simply as the number of fully occupied sites. For heavy atoms, (18.5.6.4) must be used for N_i.

Equation (18.5.6.5) is not to be regarded as having absolute validity. It is a quick and rough guide for the diffraction-data-only error component for an atom with Debye B equal to the $B_{\rm avg}$ for the structure. It is named the *diffraction-component precision index*, or DPI. It contains none of the restraint data.

18.5.6.3. *Extension for low-resolution structures and use of* $R_{\rm free}$

For low-resolution structures, the number of parameters may exceed the number of diffraction data. In (18.5.6.3) and (18.5.6.5), $p = n_{\rm obs} - n_{\rm params}$ is then negative, so that $\sigma(x)$ is imaginary. This difficulty can be circumvented *empirically* by replacing p with $n_{\rm obs}$ and R with $R_{\rm free}$ (Brünger, 1992). The counterpart of the DPI (18.5.6.5) is then

$$\sigma(x, B_{\rm avg}) = 1.0(N_i/n_{\rm obs})^{1/2}C^{-1/3}R_{\rm free}d_{\rm min}. \qquad (18.5.6.6)$$

Here $n_{\rm obs}$ is the number of reflections included in the refinement, not the number in the $R_{\rm free}$ set.

It may be asked: how can there be any estimate for the precision of a coordinate from the diffraction data only when there are insufficient diffraction data to determine the structure? By following the line of argument of Cruickshank's (1960) analysis, (18.5.6.6) is a rough estimate of the square root of the reciprocal of one diagonal element of the diffraction-only least-squares matrix. All the other parameters can be regarded as having been determined from a diffraction-plus-restraints matrix.

Clearly, (18.5.6.6) can also be used as a general alternative to (18.5.6.5) as a DPI, irrespective of whether the number of degrees of freedom $p = n_{\rm obs} - n_{\rm params}$ is positive or negative.

Comment. When p is positive, (18.5.6.6) would be exactly equivalent to (18.5.6.5) only if $R_{\rm free} = R[n_{\rm obs}/(n_{\rm obs} - n_{\rm params})]^{1/2}$. Tickle *et al.* (1998*b*) have shown that the expected relationship in a restrained refinement is actually

$$R_{\rm free} = R\{[n_{\rm obs} + (n_{\rm params} - h)]/[n_{\rm obs} - (n_{\rm params} - h)]\}^{1/2}, \qquad (18.5.6.7)$$

where $h = n_{\rm restraints} - \sum w_{\rm geom}(\Delta Q)^2$, the latter term, as in (18.5.3.1), being the weighted sum of the squares of the restraint residuals.

18.5.6.4. *Position error*

Often an estimate of a position error $|\Delta\mathbf{r}|$, rather than a coordinate error $|\Delta x|$, is required. In the isotropic approximation,

$$\sigma(r, B_{\rm avg}) = 3^{1/2}\sigma(x, B_{\rm avg}). \qquad (18.5.6.8)$$

Consequently, the DPI formulae for the position errors are

$$\sigma(r, B_{\rm avg}) = 3^{1/2}(N_i/p)^{1/2}C^{-1/3}Rd_{\rm min} \qquad (18.5.6.9)$$

with R and

$$\sigma(r, B_{\rm avg}) = 3^{1/2}(N_i/n_{\rm obs})^{1/2}C^{-1/3}R_{\rm free}d_{\rm min} \qquad (18.5.6.10)$$

with $R_{\rm free}$.

references