The square root S of the expression S^2, (18.5.2.12) above, is called the *goodness of fit* when the weights are the reciprocals of the absolute variances of the observations.

One recommendation in the second report does call for comment here. While agreeing that formulae like (18.5.2.13) lead to conservative estimates of parameter variances, the report suggests that this practice is based on the questionable assumption that the variances of the observations by which the weights are assigned are relatively correct but uniformly underestimated. When the goodness of fit $S > 1$, then either the weights or the model or both are suspect.

Comment is needed. The account in Section 18.5.2.2 describes two distinct ways of estimating parameter variances, covering two ranges of error. The kind of weights envisaged in the reports (based on variances of type A and/or of type B) are of a class described for method (1). They are not the weights to be used in method (2) (though they may be a component in such weights). Method (2) implicitly assumes from the outset that there are experimental errors, some covered and others not covered by method (1), and that there are imperfections in the calculated model (as is obviously true for proteins). Method (2) avoids exploring the relative proportions and details of these error sources and aims to provide a realistic estimate of parameter uncertainties which can be used in external comparisons. It can be formally objected that method (2) does not conform to the criteria of random-variable theory, since clearly the Δ's are partially correlated through the remaining model errors and some systematic experimental errors. But it is a useful procedure. Method (1) on its own would present an optimistic view of the reliability of the overall investigation, the degree of optimism being indicated by the inverse of the goodness of fit (18.5.2.12). In method (2), if the weights are on an arbitrary scale, then S^2 can have an arbitrary value.

For an advanced-level treatment of many aspects of the refinement of structural parameters, see Part 8 of *International Tables for Crystallography*, Volume C (1999). The detection and treatment of systematic error are discussed in Chapter 8.5 therein.

18.5.3. Restrained refinement

18.5.3.1. *Residual function*

Protein structures are often refined by a restrained refinement program such as *PROLSQ* (Hendrickson & Konnert, 1980). Here, a function of the type

$$R' = \textstyle\sum w_h(\Delta F)^2 + \sum w_{\rm geom}(\Delta Q)^2 \qquad (18.5.3.1)$$

is minimized, where Q denotes a geometrical restraint such as a bond length. Formally, all one is doing is extending the list of observations. One is adding to the protein diffraction data geometrical data from a stereochemical dictionary such as that of Engh & Huber (1991). A chain C—N bond length may be known from the dictionary with much greater precision $1/w_{\rm geom}^{1/2}$, say 0.02 Å, than from an unrestrained diffraction-data-only protein refinement.

In a high-resolution unrestrained refinement of a small molecule, the standard uncertainty (s.u.) of a bond length A—B is often well approximated by

$$\sigma(l) = (\sigma_A^2 + \sigma_B^2)^{1/2}. \qquad (18.5.3.2)$$

However, in a protein determination $\sigma(l)$ is often much smaller than either σ_A or σ_B because of the excellent information from the stereochemical dictionary, which correlates the positions of A and B.

Laying aside computational size and complexity, the protein precision problem is straightforward in principle. When a restrained refinement has converged to an acceptable structure and the shifts in successive rounds have become negligible, invert the full matrix. The inverse matrix immediately yields estimates of the variances and covariances of all parameters.

The dimensions of the matrix are the same whether or not the refinement is restrained. The full matrix will be rather sparse, but not nearly as sparse as in a small-molecule refinement. For the purposes of Section 18.5.3, it is irrelevant whether the residual for the diffraction data is based on $|F|$ or $|F|^2$. On the relative weighting of the diffraction and restraint terms, see Section 18.5.3.3.

18.5.3.2. *A very simple protein model*

Some aspects of restrained refinement are easily understood by considering a *one-dimensional protein consisting of two like atoms* in the asymmetric unit, with coordinates x_1 and x_2 relative to a fixed origin and bond length $l = x_2 - x_1$. In the refinement, the normal equations are of the type $\mathbf{N}\Delta\mathbf{x} = \mathbf{e}$. For two non-overlapping like atoms, the *diffraction data* will yield a normal matrix

$$\mathbf{N} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \qquad (18.5.3.3)$$

with inverse

$$\begin{pmatrix} 1/a & 0 \\ 0 & 1/a \end{pmatrix}, \qquad (18.5.3.4)$$

where

$$a = \textstyle\sum w_h(\partial|F_n|/\partial x_i)^2. \qquad (18.5.3.5)$$

A *geometric restraint* on the length will yield a normal matrix

$$\begin{pmatrix} b & -b \\ -b & b \end{pmatrix} \qquad (18.5.3.6)$$

with no inverse, since its determinant is zero, where

$$b = w_{\rm geom}(\partial l/\partial x_i)^2. \qquad (18.5.3.7)$$

Note $\partial l/\partial x_2 = -\partial l/\partial x_i = 1$, so that

$$b = w_{\rm geom} = 1/\sigma^2_{\rm geom}(l), \qquad (18.5.3.8)$$

where $\sigma^2_{\rm geom}(l)$ is the variance assigned to the length in the stereochemical dictionary.

Combining the diffraction data and the restraint, the normal matrix becomes

$$\begin{pmatrix} a+b & -b \\ -b & a+b \end{pmatrix}, \qquad (18.5.3.9)$$

with inverse

$$\{1/[a(a+2b)]\}\begin{pmatrix} a+b & b \\ b & a+b \end{pmatrix}. \qquad (18.5.3.10)$$

For the diffraction data alone, the variance of x_i is

$$\sigma^2_{\rm diff}(x_i) = 1/a. \qquad (18.5.3.11)$$

For the diffraction data plus restraint, the variance of x_i is

$$\begin{aligned}\sigma^2_{\rm res}(x_i) &= (a+b)/[a(a+2b)] \qquad (18.5.3.12)\\ &< \sigma^2_{\rm diff}(x_i).\end{aligned}$$

Note that though the restraint says nothing about the position of x_i, the variance of x_i has been reduced because of the coupling to the position of the other atom. In the limit when $a \ll b$, $\sigma^2_{\rm res}(x_i)$ is only half $\sigma^2_{\rm diff}(x_i)$.

references