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

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

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

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

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

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

where $\sigma_{\rm geom}^2(l)$ 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

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

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

$$\sigma_{\rm res}^2(x_i) = (a+b)/[a(a+2b)] < \sigma_{\rm diff}^2(x_i). \qquad (18.5.3.12)$$

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 $a \ll b$, $\sigma_{\rm res}^2(x_i)$ is only half $\sigma_{\rm diff}^2(x_i)$.

The general formula for the variance of the length $l = x_2 - x_1$ is

$$\sigma^2(l) = \sigma^2(x_2) - 2\mathrm{cov}(x_2, x_1) + \sigma^2(x_1). \qquad (18.5.3.13)$$

For the diffraction data alone, this gives

$$\sigma^2_{\rm diff}(l) = 1/a + 0 + 1/a = 2/a = 2\sigma^2_{\rm diff}(x_i), \qquad (18.5.3.14)$$

as expected. For the diffraction data plus restraint,

$$\begin{aligned}\sigma^2_{\rm res}(l) &= [1/a(a+2b)][(a+b) - 2b + (a+b)] \\ &= 1/(a/2 + b) \\ &< \sigma^2_{\rm diff}(l).\end{aligned} \qquad (18.5.3.15)$$

For small a, $\sigma^2_{\rm res}(l) \rightarrow 1/b = \sigma^2_{\rm geom}(l)$, as expected. The variance of the restrained length, (18.5.3.15), can be re-expressed as

$$1/\sigma^2_{\rm res}(l) = 1/\sigma^2_{\rm diff}(l) + 1/\sigma^2_{\rm geom}(l). \qquad (18.5.3.16)$$

For the two-atom protein, it can be proved directly, as one would expect from (18.5.3.16), that *restrained refinement determines a length which is the weighted mean of the diffraction-only length and the geometric dictionary length*.

The centroid has coordinate $c = (x_1 + x_2)/2$. It is easily found that $\sigma^2_{\rm res}(c) = \sigma^2_{\rm diff}(c) = 1/2a$. Thus, as expected, the restraint says nothing about the position of the molecule in the cell.

For numerical illustrations of the s.u.'s in restrained refinement, suppose the stereochemical length restraint has $\sigma_{\rm geom}(l) = 0.02$ Å. Equation (18.5.3.16) gives the length s.u. $\sigma_{\rm res}(l)$ in restrained refinement. If the diffraction-only $\sigma_{\rm diff}(x_i) = 0.01$ Å, the restrained $\sigma_{\rm res}(l)$ is 0.012 Å. If $\sigma_{\rm diff}(x_i) = 0.05$ Å, $\sigma_{\rm res}(l)$ is 0.019 Å. However large $\sigma_{\rm diff}(x_i)$, $\sigma_{\rm res}(l)$ never exceeds 0.02 Å.

Equation (18.5.3.12) gives the position s.u. $\sigma_{\rm res}(x_i)$ in restrained refinement. If the diffraction-only $\sigma_{\rm diff}(x_i) = 0.01$ Å, the restrained $\sigma_{\rm res}(x_i)$ is 0.009 Å. If $\sigma_{\rm diff}(x_i) = 0.05$ Å, $\sigma_{\rm res}(x_i) = 0.037$ Å. For large $\sigma_{\rm diff}(x_i)$, $\sigma_{\rm res}(x_i)$ tends to $\sigma_{\rm diff}(x_i)/(2)^{1/2}$ as the strong restraint couples the two atoms together. For very small $\sigma_{\rm diff}(x_i)$, the relatively weak restraint has no effect.

18.5.3.3. *Relative weighting of diffraction and restraint terms*

When only relative diffraction weights are known, as in equation (18.5.2.13), it has been common (Rollett, 1970) to scale the geometric restraint terms against the diffraction terms by replacing the restraint weights $w_{\rm geom} = 1/\sigma^2_{\rm geom}$ by $w_{\rm geom} = S^2/\sigma^2_{\rm geom}$, where $S^2 = (\sum w_h \Delta^2_h)/(n_{\rm obs} - n_{\rm params})$. However, this scheme cannot be used for low-resolution structures if $n_{\rm obs} < n_{\rm params}$.

The treatment by Tickle *et al.* (1998*a*) shows that the reduction $n_{\rm params}$ in the number of degrees of freedom has to be distributed among all the data, both diffraction observations and restraints. Since the geometric restraint weights are on an absolute scale (Å^{-2}), they propose that the (absolute) scale of the diffraction weights should be determined by adjustment until the restrained residual R' (18.5.3.1) is equal to its expected value $(n_{\rm obs} + n_{\rm restraints} - n_{\rm params})$.

For a method of determining the scale of the diffraction weights based on $R_{\rm free}$, see Brünger (1993).

The geometric restraint weights were classified by the IUCr Subcommittee (Schwarzenbach *et al.*, 1995) as derived from observations supplementary to the diffraction data, with uncertainties of type B (Section 18.5.2.3).

18.5.4. Two examples of full-matrix inversion

18.5.4.1. *Unrestrained and restrained inversions for concanavalin A*

G. M. Sheldrick extended his *SHELXL*96 program (Sheldrick & Schneider, 1997) to provide extra information about protein precision through the inversion of least-squares full matrices. His programs have been used by Deacon *et al.* (1997) for the high-resolution refinement of native concanavalin A with 237 residues, using data at 110 K to 0.94 Å refined anisotropically. After the convergence and completion of full-matrix restrained refinement for the structure, the unrestrained full matrix (coordinates only) was computed and then inverted in a massive calculation. This led to s.u's $\sigma(x)$, $\sigma(y)$, $\sigma(z)$ and $\sigma(r)$ for all atoms, and to $\sigma(l)$ and $\sigma(\theta)$ for all bond lengths and angles. $\sigma(r)$ is defined as $[\sigma^2(x) + \sigma^2(y) + \sigma^2(z)]^{1/2}$. For concanavalin A the restrained full matrix was also inverted, thus allowing the comparison of restrained and unrestrained s.u.'s.

The results for concanavalin A from the inversion of the coordinate matrices of order 6402 $(= 2134 \times 3)$ are plotted in Figs. 18.5.4.1 and 18.5.4.2. Fig. 18.5.4.1 shows $\sigma(r)$ *versus* $B_{\rm eq}$ for the fully occupied atoms of the protein (a few atoms with $B >$ 60 Å^2 are off-scale). The points are colour-coded black for carbon, blue for nitrogen and red for oxygen. Fig. 18.5.4.1(*a*) shows the restrained results, and Fig. 18.5.4.1(*b*) shows the unrestrained diffraction-data-only results. Superposed on both sets of data points are least-squares quadratic fits determined with weights $1/B^2$. At high B, the unrestrained $\sigma_{\rm diff}(r)$ can be at least double the restrained $\sigma_{\rm res}(r)$, *e.g.*, for carbon at $B = 50$ Å^2, the unrestrained $\sigma_{\rm diff}(r)$ is about 0.25 Å, whereas the restrained $\sigma_{\rm res}(r)$ is about 0.11 Å. For $B < 10$ Å^2, both $\sigma(r)$'s fall below 0.02 Å and are around 0.01 Å at $B = 6$ Å^2.

For $B < 10$ Å^2, the better precision of oxygen as compared with nitrogen, and of nitrogen as compared with carbon, can be clearly seen. At the lowest B, the unrestrained $\sigma_{\rm diff}(r)$ in Fig. 18.5.4.1(*b*) are almost as small as the restrained $\sigma_{\rm res}(r)$ in Fig. 18.5.4.1(*a*). [The quadratic fits of the restrained results in Fig. 18.5.4.1(*a*) are evidently slightly imperfect in making $\sigma_{\rm res}(r)$ tend almost to 0 as B tends to 0.]

Fig. 18.5.4.2 shows $\sigma(l)$ *versus* $B_{\rm eq}$ for the bond lengths in the protein. The points are colour-coded black for C—C, blue for C—N and red for C—O. The restrained and unrestrained distributions are very different for high B. The restrained distribution in Fig. 18.5.4.2(*a*) tends to about 0.02 Å, which is the standard uncertainty of the applied restraint for 1–2 bond lengths, whereas the unrestrained distribution in Fig. 18.5.4.2(*b*) goes off the scale of the diagram. But for $B < 10$ Å^2, both distributions fall to around 0.01 Å.

The differences between the restrained and unrestrained $\sigma(r)$ and $\sigma(l)$ can be understood through the two-atom model for restrained refinement described in Section 18.5.3. For that model, the equation

$$1/\sigma^2_{\rm res}(l) = 1/\sigma^2_{\rm diff}(l) + 1/\sigma^2_{\rm geom}(l) \qquad (18.5.3.16)$$

relates the bond-length s.u. in the restrained refinement, $\sigma_{\rm res}(l)$, to the $\sigma_{\rm diff}(l)$ of the unrestrained refinement and the s.u. $\sigma_{\rm geom}(l)$ assigned to the length in the stereochemical dictionary. In the refinements, $\sigma_{\rm geom}(l)$ was 0.02 Å for all bond lengths. When this is combined in (18.5.3.16) with the unrestrained $\sigma_{\rm diff}(l)$ of any bond, the predicted restrained $\sigma_{\rm res}(l)$ is close to that found in the restrained full matrix.

It can be seen from Fig. 18.5.4.2(*b*) that many bond lengths with average $B < 10$ Å^2 have $\sigma_{\rm diff}(l) < 0.014$ Å. For these bonds the diffraction data have greater weight than the stereochemical dictionary. Some bonds have $\sigma_{\rm diff}(l)$ as low as 0.0080 Å, with

references