Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Estimation of weights and validation: a marginal likelihood approach

Andrey A. Lebedev,[a]† Ian J. Tickle,[b] Roman A. Laskowski[c] and David S. Moss[a]*

[a]School of Crystallography, Birkbeck College, London WC1E 7HX, England, [b]Astex Technology Ltd, 250 Cambridge Science Park, Milton Road, Cambridge CB4 0WE, England, and [c]EMBL Outstation Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England

† Current address: Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 5DD, England.

Correspondence e-mail:
d.moss@cryst.bbk.ac.uk

Received 7 January 2003
Accepted 26 June 2003

The estimation of weights is quite an important aspect of the restrained refinement of macromolecular structures and related procedures such as the estimation of coordinate errors and structure validation using geometrical criteria. In principle, the method of maximum likelihood can be used for estimation of both atomic and weighting parameters. However, the low observation-to-parameter ratio in macromolecular refinement makes this kind of estimate of weighting parameters seriously biased; thus, the weighting parameters have traditionally been estimated separately from atomic parameters using a special technique, such as minimizing the free *R* factor. However, the variance of the latter estimate is large, as only a small portion of all data is used. In this work, an estimator of weights is proposed which is based on an approximation of a marginal likelihood function of the weighting parameters and which uses all the X-ray data. There is a known true value for the overall scaling coefficient for distance variances in restrained refinement and therefore the (maximum-likelihood) estimator for this coefficient may be used as a validation statistic.

1. Introduction

The problem of validating protein structures solved by X-ray crystallography has been widely investigated during the last decade (Dodson, 1998; Kleywegt, 1999, 2000). This problem has become even more relevant with the rapidly growing number of structures in the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2002) and with the development of highly automated methods of structure determination (Perrakis *et al.*, 1999; Morris *et al.*, 2002), a constituent of high-throughput protein crystallography.

One of the validation criteria is the consistency of covalent-bond parameters with those obtained from previously solved structures. In the case of unrestrained refinement, the average weighted geometrical residual R_G is a statistic with known expected value $\mathcal{E}(R_G)$ and variance. However, in refinements with geometrical restraints, the equation $\mathcal{E}(R_G) = 1$ transforms into the inequality $\mathcal{E}(R_G) < 1$ (Tickle *et al.*, 1998*a*) and $\mathcal{E}(R_G)$ depends on the variable weighting parameters in the X-ray partition of the covariance matrix. Badger & Hendle (2002) have found that root-mean-square deviations from ideality for bond lengths and bond angles appear to be unrelated to both the estimate of the overall accuracy of a structure given by the diffraction-component precision index (DPI; Cruickshank, 1999) and the number of local errors that the molecular model contains.

Several applications make use of the correlation between $R_{\rm free}$ and the phase accuracy of a crystal structure. Brünger (1992*a*) has shown that the $R_{\rm free}$ method can be used to optimize the overall weighting between diffraction data and chemical restraints in crystallographic refinements. Examples of the implementation of this idea can be found in Brünger (1992*b*) and Murshudov *et al.* (1997). The overall weight of the restraints is a special case of the more general weights that apply to subsets of diffraction data or particular classes of restraints, such as bond lengths, bond angles and dihedral angles. Brünger (1993) optimized $R_{\rm free}$ as a function of all these individual weights using the penicillopepsin crystal structure at 1.8 Å resolution as a test case and demonstrated that the uncertainties in bond lengths and bond angles derived from small-molecule crystal structures (Engh & Huber, 1991) appear to be applicable to macromolecules. This conclusion follows from the fact that the optimized weights w_b and w_a for bond-length and bond-angle terms, respectively, are close to one. From the point of view of structure validation, it is a notable fact because it means that in principle the validation criterion $\mathcal{E}(R_G) = 1$ can be replaced in restrained refinement by the criterion $\mathcal{E}(w_{b,a}) = 1$, where $w_{b,a}$ is the weight of a joint term involving bond-length and bond-angle restraints. The optimization problem under discussion is computer-intensive, as each point in the multi-dimensional space of weighting coefficients requires a complete round of refinement to evaluate $R_{\rm free}$. However, the major problem is that the variances of the estimates of weights are unlikely to approach the Cramér–Rao lower bound (Cramér, 1946; Leonard & Hsu, 2001), as only a small part of experimental information is used in this estimation.

The free R factor is also used as a validation statistic itself. It is certainly useful for judging protocols and progress in refinement. It is particularly valuable as a tool for detecting overfitting (Dodson *et al.*, 1996; Brünger, 1997). However, a rigorous validation test requires knowledge of the distribution of the tested statistic. In a more relaxed but often acceptable approach, one has to know the value of the statistic calculated for a given data set together with estimates of the mean and variance of the statistic. Tickle *et al.* (1998*b*) derive the expected value of the free residual from which estimates of the expected values of both $R_{\rm free}$ and the ratio $R_{\rm free}/R$ are calculated. The work is taken further in Tickle *et al.* (2000), where the variation of the above ratio about its expected value is explored. In these papers, the estimates of the expected values and variances are derived on the assumption that the weights used in structure refinement correctly reflect the errors. Thus, the practical implementation of this approach requires reliable estimates of weights.

In contrast to the method of least squares, the method of maximum likelihood takes into account the probability distribution of observations and may perform better than least squares if this distribution is not normal. The normal distribution can be, in a first approximation, attributed to the distribution of structure factors in the complex plane (Luzzati, 1952; Srinivasan & Ramachandran, 1965). The marginal distribution density of structure amplitudes corresponding to this approximation is known in crystallography as the Sim (Sim, 1959) or Rice distribution (Bricogne, 1988). The latter is sufficiently simple to handle and is the basis for the implementation of the method of maximum likelihood in the macromolecular refinement. Bricogne & Irwin (1996), Pannu & Read (1996) and Murshudov *et al.* (1997) find that compared with least-squares refinement, maximum-likelihood refinement can achieve a considerable improvement in average phase error. The resulting electron-density maps are correspondingly clearer and suffer less from model bias.

Along with the atomic parameters, a number of parameters appear in the method of maximum likelihood that are linked to the probabilistic model used to describe the distribution of structure factors. These parameters generalize the weights of the X-ray terms in the method of least squares and are therefore further referred to as weighting parameters. Just as in the case of restrained least-squares refinement, estimation of weighting parameters in the method of maximum likelihood is a complicated problem. Lunin & Urzhumtsev (1984) observe that the accuracy of phases obtained from models refined using likelihood targets is overestimated. Read (1986) finds that the use of maximum-likelihood estimators for weighting parameters results in systematically overestimated figures of merit but only within the resolution limits used in the refinement of atomic coordinates. Lunin & Skovoroda (1995) and Brünger (1997) show that the use of the likelihood function calculated from the free set of reflections allows one to eliminate the bias. Skovoroda & Lunin (2000) suggest a method for reducing the statistical dispersion of the estimates based on a small number of free reflections.

The work described in this paper demonstrates that weighting parameters and, in particular, the weight of geometrical term $w_{b,a}$ can be estimated without significant bias using an approximation of a marginal likelihood function of all the observed data. As we discussed earlier, the weight of the geometrical term has a known expected value that equals one and it can be proposed as a potential validation statistic.

2. Model of experiment

The refinement of macromolecular structures is a statistical problem with a strong prior knowledge of stereochemical parameters. We are particularly interested in the parameters that have well determined variances, *i.e.* in the covalent-bond and covalent-angle distances. This is because we seek to determine whether the model structure proposed is consistent with these known variances.

The above prior knowledge is formally expressed in terms of target interatomic distances provided in a dictionary containing means and variances of the distances that are derived from a database of small molecules (Engh & Huber, 1991).

2.1. Sampling model

Let x, u and f^o be the vector of atomic parameters, the vector of weighting parameters and the vector of experimental data, including the stereochemical data, respectively.[1]

As long as we consider geometrical data as observations, the prior knowledge associated with them is already accounted for and therefore it is valid to assume that atomic parameters are sampled from a uniform prior density,

$$p(x|u) = p(x) = \text{constant}. \tag{1}$$

The prior probability distribution density (1) is an improper prior (Lindley, 1965), which is the limiting case of densities, uniform on an infinitely growing range of the parameters.

The experimental uncertainties and the features of electron density which are not accounted for by the atomic model are described by the conditional density $p(f^o|x, u)$. The marginal density function of f^o conditioned by the weighting parameters is given by

$$p(f^o|u) = \int p(f^o|x, u)\ p(x|u)\ \mathrm{d}x. \tag{2}$$

Formally, the integration in (2) is over all possible x, but the dominant contribution to the integral comes from the values of x that are close to the sharp maximum of the likelihood function $p(f^o|x, u)$.

The numerical experiments carried out in this work deal with the case where the relative values of true atomic temperature factors are either known (tests with simulated data) or are assumed to be known (tests with real data). Thus, the vector x includes atomic coordinates plus two overall parameters, namely a temperature factor and a structure-amplitude scale factor.

2.2. Model of mean and covariance

Let $\mathcal{F}$ be a parametric family of random vectors with the parameters x and u, such that any random vector $f^o \in \mathcal{F}$ of the family satisfies the following relations

$$\begin{aligned} &\mathcal{E}(f^o|x, u) = f(x) \\ &\mathcal{E}(f^o {f^o}^T|x, u) = f(x)f(x)^T + \Sigma(u), \end{aligned} \tag{3}$$

i.e. components of x are parameters of the mean vector $f(x)$ and components of u are parameters of the covariance matrix $\Sigma(u)$ of the random vector f^o.

The X-ray data are divided into P bins according to some reasonable criteria, for instance resolution and/or intensity, so that we can assume that the observed structure amplitudes from the same bin have the same variance. Also, we assume that all observations, including both X-ray and distance data, are independent and therefore the covariance matrix is diagonal. Thus,

$$\Sigma(u) = \begin{pmatrix} u_0\Sigma_0 & 0 & \cdots & 0 \\ 0 & u_1\Sigma_1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & u_P\Sigma_P \end{pmatrix}, \tag{4}$$

where Σ_0 is the diagonal distance covariance matrix determined from a dictionary of experimental small-molecule data and Σ_1, Σ_2, $\ldots$, Σ_P are the unit matrices of appropriate dimensions. The scalar u_0, the overall scaling coefficient for distance variances, and scalars u_1, $\ldots$, u_P, the variances of structure amplitudes from bins 1, 2, $\ldots$, P (reciprocal weights), constitute the vector of weighting parameters $u = (u_0, u_1, \ldots, u_P)^T$.

Accordingly, the random vector f^o and the mean vector $f(x)$ have $P + 1$ partitions:

$$f^o = \begin{pmatrix} f_0^o \\ f_1^o \\ \cdots \\ f_P^o \end{pmatrix} \tag{5}$$

and

$$f(x) = \begin{bmatrix} f_0(x) \\ f_1(x) \\ \cdots \\ f_P(x) \end{bmatrix}, \tag{6}$$

where $f_0(x)$ is the vector of calculated distances and $f_1(x)$, $\ldots$, $f_P(x)$ are vectors of calculated structure amplitudes from bins 1, $\ldots$, P.

We suppose that our data are sampled from a random vector that belongs to the family $\mathcal{F}$. If f^o is this random vector, then the components of x and u in (3) are the true values of atomic and weighting parameters, respectively. Given the observed value of f^o, we need to estimate these true values of x and u.

Note that the problem is formulated in such a way that u_0 is a known constant:

$$u_0 = 1. \tag{7}$$

However, if we let u_0 be an unknown parameter, then any unbiased or insignificantly biased estimator for u_0 is a validation statistic, provided that the variance of the estimator can also be estimated.

3. Maximum-likelihood estimators for weighting parameters

Our goal is to understand how good the method of maximum likelihood is for the determination of weights in macromolecular refinement and, particularly, for the validation of the structure using (7). It is essential for us that the variances of likelihood estimators can be estimated using the second derivatives at the maximum of the likelihood function.

We base our discussion on the comparison of two log-likelihood functions, both corresponding to the distribution density of the random vector f^o, the first being conditioned by the vector of atomic parameters x and the vector of weighting

[1] Associated with any of the parameters of any Bayesian model of an experiment are a random variable, a sampled value(s) and a formal variable of the probability distribution density. Dependent on the context, the sampled value is a true value or an observed value. Similar objects are associated with functions of the parameters, including estimators. To avoid the abuse of notation we use one-letter notations, the actual sense of the letter being obvious from the context. This is effectively the usual practice.

parameters u and the second being conditioned only by the vector of weighting parameters u. In this paper, we compare these two likelihood functions using a number of simulated data sets corresponding to a small synthetic crystal structure. The better of the two likelihood functions is then used to estimate the parameters of the geometry and structure-amplitude variances using three real polypeptide structures.

3.1. Definitions of likelihood functions

Let $L(x, u)$ be the log-likelihood of atomic parameters x and weighting parameters u given the vector of observations f^o, *i.e.*

$$L(x, u) = \ln l(x, u|f^o) = \ln p(f^o|x, u). \tag{8}$$

In the likelihood maximization we have assumed the density $p(f^o|x, u)$ to be normal, *i.e.*

$$\begin{aligned} L(x, u) = &-\tfrac{1}{2}[f^o - f(x)]^T \Sigma(u)^{-1}[f^o - f(x)] \\ &- \tfrac{1}{2}\ln[\det \Sigma(u)] \; + \; \text{constant}. \end{aligned} \tag{9}$$

This simplified likelihood is however sufficient to illustrate our ideas. Our tests show that there are some other factors that influence the behaviour of estimators for weighting parameters much more than the choice of likelihood approximation.

Let $M(u)$ be the following function of u,

$$M(u) = L[t(u), u] + \tfrac{1}{2}\ln[\det \sigma(u)], \tag{10}$$

where

$$\sigma(u) = \left[\frac{\partial f(x)}{\partial x}^T \Sigma(u)^{-1} \frac{\partial f(x)}{\partial x} \bigg|_{x=t(u)} \right]^{-1} \tag{11}$$

and where the vector-function $t(u)$ is such that for a given value of u it maximizes $L(x, u)$ with respect to x, thus

$$\frac{\partial L(x, u)}{\partial x}\bigg|_{x=t(u)} = 0. \tag{12}$$

Neglecting the term independent of u, the function $M(u)$ approximates the marginal log-likelihood of u given f^o:

$$M(u) \approx \ln l(u|f^o) = \ln p(f^o|u). \tag{13}$$

Using (1), (2), (8), (10), (11) and (12), one can check that approximation (13) becomes an exact equation if the conditional distribution of f^o given x and u is multivariate normal (9) and $f(x)$ is a linear function of x.

3.2. Definitions and comparison of estimators

Let $(\hat{x}, \hat{u})$ and $\tilde{u}$ maximize the log-likelihood functions $L(x, u)$ and $M(u)$, respectively. Consequently,

$$\begin{cases} \dfrac{\partial L(x, u)}{\partial x}\bigg|_{u=\hat{u}, x=\hat{x}} = 0 \\ \dfrac{\partial L(x, u)}{\partial u}\bigg|_{u=\hat{u}, x=\hat{x}} = 0 \end{cases} \tag{14}$$

and

$$\frac{\partial M(u)}{\partial u}\bigg|_{u=\tilde{u}} = 0. \tag{15}$$

(14) and (15) are likelihood equations and the random vectors $(\hat{x}, \hat{u})$ and $\tilde{u}$ are maximum-likelihood estimators corresponding to the log-likelihood functions $L(x, u)$ and $M(u)$, respectively.

Note that (12) and (14) imply that

$$\frac{\partial L[t(u), u]}{\partial u}\bigg|_{u=\hat{u}} = 0, \tag{16}$$

where the maximized function coincides with the first term on the right-hand side of (10). Therefore, it is the second term in (10) that is responsible for the difference between $\tilde{u}$ and $\hat{u}$.

Which is better for estimating u, the partition $\hat{u}$ of the maximum-likelihood estimator $(\hat{x}, \hat{u})$ or the maximum-likelihood estimator $\tilde{u}$? The observation-to-parameter ratios for the two cases give a clear, although heuristic, answer. In the first case ($\hat{u}$), the number of unknown parameters is equal to the number of atomic parameters (hundreds or many more) plus the number of weighting parameters (a few dozen or less). The ratio 10 would be good for macromolecular refinement. Only the weighting parameters are unknown in the second case ($\tilde{u}$) and therefore the above ratio is hundreds or thousands. Thus, the behaviour of the estimator $\tilde{u}$ is likely to approach the asymptotic behaviour of maximum-likelihood estimators, *i.e.* we would expect that $\tilde{u}$ is an almost normally distributed random variable and an almost unbiased estimator for u and that the matrix

$$Z = \left[-\frac{\partial^2 M(u)}{\partial u^T \partial u}\bigg|_{u=\tilde{u}} \right]^{-1}, \tag{17}$$

which is reciprocal to the likelihood information matrix (Leonard & Hsu, 2001), is a good estimator of the covariance matrix of $\tilde{u}$, *i.e.*

$$\begin{aligned} \mathcal{E}(\tilde{u}|u) &\approx u \\ \mathcal{E}(\tilde{u}\tilde{u}^T|u) &\approx uu^T + \mathcal{E}(Z|u). \end{aligned} \tag{18}$$

On the other hand, the computational cost of calculating $M(u)$ and its derivatives is huge compared with those for $L[t(u), u]$ owing to the term $\ln[(\det\sigma(u)]$ (see §6). It is clear that any practical application would use certain approximations for σ and its derivatives. Nevertheless, we first of all want to be sure that such efforts are necessary. Maybe the estimator $\hat{u}$ is sufficiently good for practical purposes? Also, in spite of the above rather intuitive reasoning, $\tilde{u}$ might have an unacceptable bias even if the exact value of σ is used in its calculation.

4. Numerical experiments

Even in the case of only two weighting parameters considered below, there is no analytical expression for both $\hat{u}$ and $\tilde{u}$ for the general case of the matrix $\partial f/\partial x$. (The case of one weighting parameter is discussed in Appendix *A*.) Therefore, we address the questions formulated in the previous section by means of tests with simulated data, where we know the true values of all estimated parameters and can estimate bias.

4.1. The case of two unknown weighting parameters

The synthetic model is a crystal structure of a fragment of an α-helix with the sequence Ser-Val-Val-Ser-Gln in space group $P2_1$, with unit-cell parameters a = 16.2, b = 12.5, c = 11.1 Å, β = 96°. It is assumed that this structure is a true structure, *i.e.* its atomic coordinates are the components of the true value of vector x.

The dimension of the vector x and the dimension n_0 of the vector $f_0(x)$ are determined by the model and are 109 and 82, respectively. The same set of reflections in the resolution range 16.1–1.3 Å have been used in all simulations. The number of reflections, *i.e.* the dimension n_1 of the vector $f_1(x)$, is equal to 901. If we treat restraints as observations, then the observation-to-parameter ratio equals 9.0, which is a rather large value for macromolecular crystallography.

We generated a set of size 130 of $(n_0 + n_1)$-dimensional vectors that can be considered as a sample of size 130 drawn from a population of vectors ζ, the components of ζ being independent random variables with a truncated normal distribution, with zero mean and unit variance, *i.e.*

$$\mathcal{E}(\zeta) = 0 \tag{19}$$

and

$$\mathcal{E}(\zeta\,\zeta^T) = E, \tag{20}$$

where E is the unit matrix.

Using the above sample, we generated a sample of size 130 drawn from one-parameter family of random vectors $f^o = f^o(u_1)$ given by

$$\begin{pmatrix} f_0^o \\ f_1^o \end{pmatrix} = \begin{bmatrix} f_0(x) \\ f_1(x) \end{bmatrix} + \begin{pmatrix} u_0\Sigma_0 & 0 \\ 0 & u_1\Sigma_1 \end{pmatrix}^{1/2} \begin{pmatrix} \zeta_0 \\ \zeta_1 \end{pmatrix}, \tag{21}$$

where ζ_0 and ζ_1 are appropriate partitions of the random vector ζ and thus (3) holds, where $u_0 = 1$ and thus (7) holds and where u_1 is the variable parameter.

In the description of the results, we use the parameter R_e instead of the parameter u_1, the two parameters being related by

$$u_1 = \frac{\pi}{2}\left[\frac{|f_1(x)|}{\mathrm{tr}(\Sigma_1^{1/2})} R_e\right]^2, \tag{22}$$

where $|f|$ denotes the sum of absolute values of components of a vector f. Therefore,

$$\mathcal{E}\left[\frac{|f_1^o - f_1(x)|}{|f_1(x)|}\middle| x, u\right] \approx R_e, \tag{23}$$

where exact equality would hold if the components of ζ had a normal distribution. Thus, the parameter R_e approximates the expected value of the R factor between X-ray data f_1^o and the true structure amplitudes $f_1(x)$. The description of the simulated experiment in terms of R_e is preferable compared with that in terms of u_1, as the value of R_e has a similar meaning and the same scale as the usual crystallographic R factor.

We solved equations (15) and (16) numerically with respect to unknown $u = (u_0, u_1)^T$ for the above 130 one-parameter data-set families at 91 values of R_e in the range 2–16%. The variation of R_e models the variation of the quality of experimental data.

Table 1
Tests with simulated data partitioned into one restraint and nine X-ray bins (the case of ten unknown covariance parameters); s.u.s multiplied by 10^3 are given in parentheses.

		Estimated magnitudes		
R_e (%)	p	$\mathcal{E}(\tilde{u}_p\|u)/u_p$	$[\mathcal{D}(\tilde{u}_p\|u)]^{1/2}/u_p$	$\mathcal{E}(Z_{pp}^{1/2}\|u)/u_p$
4	0	1.007 (6)	0.176 (4)	0.177 (1)
	1	0.999 (6)	0.176 (4)	0.173 (1)
			...	
	9	1.000 (5)	0.150 (3)	0.147 (1)
8	0	1.007 (7)	0.214 (5)	0.211 (1)
	1	0.998 (5)	0.172 (4)	0.168 (1)
			...	
	9	0.999 (5)	0.148 (3)	0.146 (1)
12	0	1.008 (8)	0.241 (5)	0.241 (2)
	1	1.000 (5)	0.165 (4)	0.165 (1)
			...	
	9	0.999 (5)	0.149 (3)	0.146 (1)
16	0	1.018 (9)	0.273 (6)	0.271 (2)
	1	0.994 (5)	0.162 (4)	0.162 (1)
			...	
	9	1.000 (5)	0.144 (3)	0.145 (1)

We observe a continuous dependence of $\tilde{u}$ on R_e in all data-set families (Figs. 1*a*, 1*b*, 1*e* and 1*f*). Continuous behaviour of $\hat{u}$ takes place for 125 of 130 data set families (Figs. 1*a* and 1*b*). In the remaining five data-set families the values of $\hat{u}_1$ jump and the values of $\hat{u}_0$ drop to almost zero (10^{-4}–10^{-2}) at a certain value of R_e. The interatomic distances in these cases almost coincide with target distances (the zero value of u_0 corresponds to a refinement with constraints). The first discontinuity occurs at $R_e \simeq 12\%$ and the frequency of such events then grows with the increase in R_e (see Figs. 1*e* and *f*).

Figs. 1(*c*) and 1(*d*) represent the totals for data-set families with continuous behaviour of both $\tilde{u}$ and $\hat{u}$. On each of these plots there are five red and five blue lines corresponding to $\tilde{u}$ and $\hat{u}$, respectively. The middle line of a particular colour is the sample mean. The two lines adjacent to it are the sample mean ± the standard uncertainty (s.u.; Schwarzenbach *et al.*, 1995) of the sample mean. The uppermost and lowermost lines are the sample mean ± the s.u. These figures clearly demonstrate that $\tilde{u}$, in contrast to $\hat{u}$, is a practically unbiased estimator of the unknown weighting parameters.

The changes of $\hat{u}_0$ and $\hat{u}_1$ with R_e indicate that bias cannot be removed by simple multiplication of the covariance by a constant coefficient, as is the case where there is only one unknown weighting parameter (Appendix *A*).

4.2. The case of ten unknown weighting parameters

The simulation of data was carried out in the same way as in the two-parameter case, but for only four values of R_e. At the same time, the sample size was increased to 1024, giving rise to a total of 4 × 1024 data sets.

At the refinement stage, the simulated X-ray data were partitioned into three resolution shells containing equal numbers of reflections and each resolution shell was partitioned into three bins according to the value of $f(x)$. Equation (15) has been solved with respect to unknown $u = (u_0, \ldots, u_9)^T$ for the above 4 × 1024 data sets.

As a result, we have samples of size 1024 drawn from the random vector $\tilde{u}$ and the random matrix Z for four values of R_e. The expected value $\mathcal{E}(\tilde{u}_p|u)$ and variance $\mathcal{D}(\tilde{u}_p|u)$ of the pth component of the random vector $\tilde{u}$ and the expected value $\mathcal{E}(Z_{pp}^{1/2}|u)$ of the square root of the pth diagonal component of the matrix Z have been estimated for four values of R_e and for $p = 0, 1, \ldots, 9$ by appropriate sample means and sample variances. Table 1 represents these estimates and their s.u.s (the s.u. in the second column have been calculated assuming normal distri-

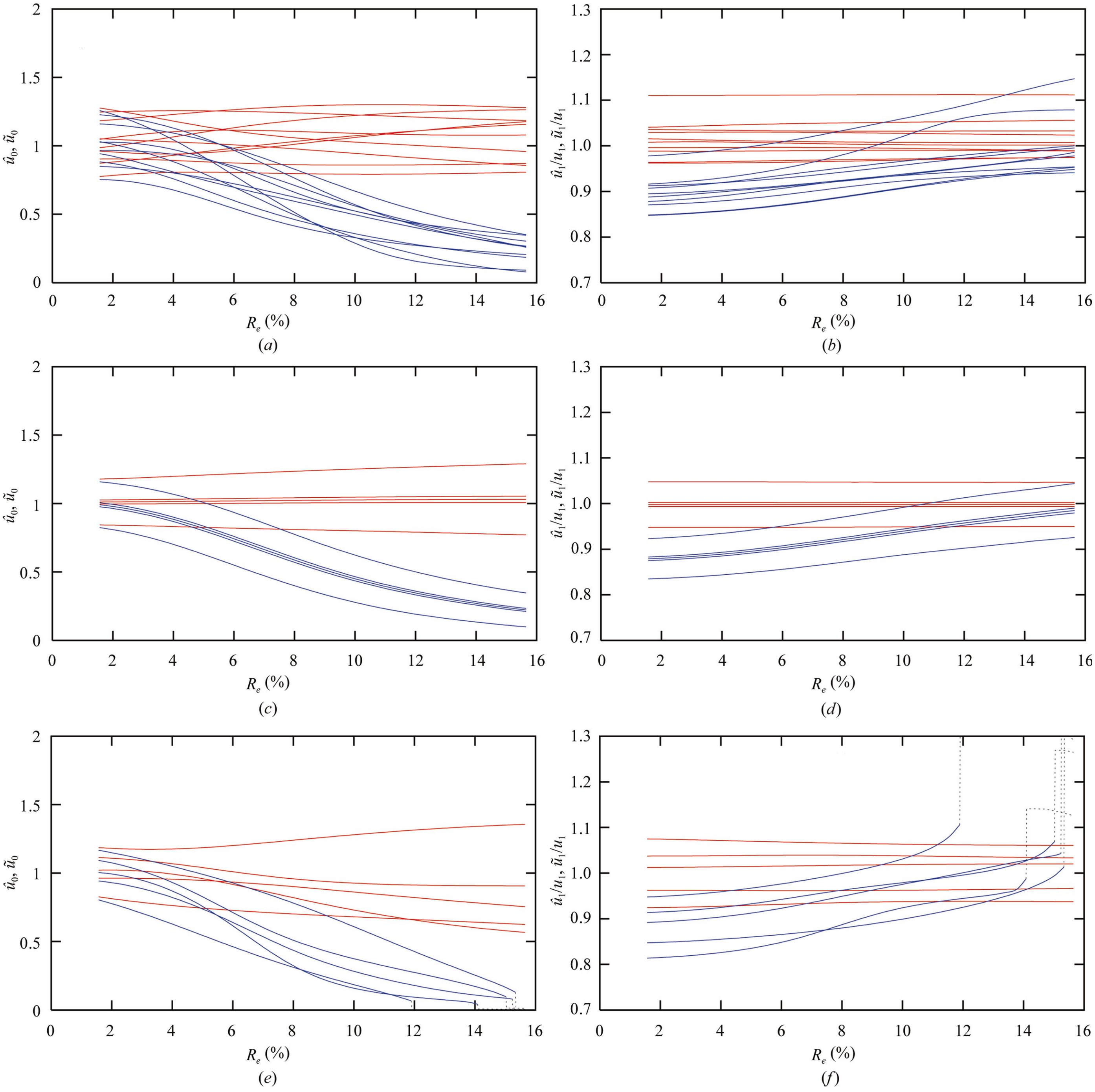

Figure 1
Dependence of estimators $\hat{u}$ (blue lines) and $\tilde{u}$ (red lines) on R_e in the case of two unknown covariance parameters: (*a*) $\hat{u}_0$, $\tilde{u}_0$ and (*b*) $\hat{u}_1/u_1$, $\tilde{u}_1/u_1$ for ten of 125 data-set families where both $\hat{u}$ and $\tilde{u}$ are continuous functions of R_e; (*c*) the sample mean, the band of the s.u. of the sample mean and the s.u. band for $\hat{u}_0$, $\tilde{u}_0$ and (*d*) for $\hat{u}_1/u_1$, $\tilde{u}_1/u_1$ corresponding to the above 125 data-set families; (*e*) $\hat{u}_0$, $\tilde{u}_0$ and (*f*) $\hat{u}_1/u_1$, $\tilde{u}_1/u_1$ for five data series with discontinuous behaviour of $\hat{u}$.

bution of $\tilde{u}$). Sections of Table 1 correspond to different values of R_e.

The approximate equation (18) holds within 1–2 s.u.s, confirming that $\tilde{u}_p$ and Z_{pp} are practically unbiased estimators for u_p and $\mathcal{D}(\tilde{u}_p|u)$, respectively. This is exactly what we would expect for an almost asymptotic behaviour of a maximum-likelihood estimator.

5. Effect of model imperfections

We studied the following cases: (i) an erroneous atomic model, (ii) a case in which the distribution of observed structure amplitudes strongly deviate from normal, although the mean and covariance are modelled correctly, (ii) a case with unaccounted random atoms with low occupancy (the models of X-ray partitions of both mean and covariance are wrong) and (iv) a case with strong correlations between interatomic distances (a wrong model of the geometrical partition of covariance). The goal was to understand the impact of different imperfections of the model on the behaviour of the statistic $\tilde{u}_0$.

5.1. Wrong atomic model

The true structure is the same as described in §4 and the data sets (observed values of both structure amplitudes and distances) were generated in a similar way except for non-essential variations in the sample size and the number of tested values of R_e. Four models have been refined against these data sets: the model with both valines substituted with threonines and three models with different combinations of missing protein atoms. The results of these refinements are presented in Fig. 2. The mean crystallographic R factor (see caption of Fig. 2) is larger than R_e, as R_e represents only experimental errors.

The effect of wrong target distances (*i.e.* the effect of the wrong residues in the refined model) on $\tilde{u}_0$ is larger compared with the effect of missing atoms. Had our model of structure-amplitude errors included the errors arising from the missing atoms, then the curve associated with missing atoms might have been even closer to $\mathcal{E}(\tilde{u}_0) = 1$.

The errors in the atomic models caused a mixture of effects, *e.g.* systematic discrepancies between the mean refined values and the true values of both distances and structure amplitudes. In the remaining tests we deal with simpler cases.

5.2. Influence of distribution of observed random vector

As long as $\Sigma(\cdot)$ and $f(\cdot)$ are correct functions, the expected value of the first derivative of M over u at $u = u_{\rm true}$ equals zero, independent of the distribution of the components of the random vector f^o. As a result, $\tilde{u}$ remains an almost unbiased estimate of u even if the actual distribution of f^o strongly deviates from normal. This is the case in this simulation, where the observed structure amplitudes f^o are equal to $f_{\rm true} + \sigma_{\rm true}$ or $f_{\rm true} - \sigma_{\rm true}$ with equal probability. The results are indicated by a red line in Fig. 3. This test justifies the simplified likelihood function (9), which corresponds to the normal distribution of f^o.

5.3. Unaccounted random atoms

In the conventional crystallographic likelihood function, the coordinates of unknown atoms are integrated out, assuming that these atoms are distributed uniformly over the whole unit cell. In the next simulation this assumption is satisfied exactly: 64 O atoms with low occupancies were added to the true crystal structure in each simulation at random positions to generate observed structure amplitudes and in the refined model these atoms were missing. In this test, the X-ray measurements were assumed to be precise (R_e = 0), but occupancies of the missing atoms were varied to generate a plot of $\mathcal{E}(\tilde{u})$ against $\mathcal{E}(R_{\rm cryst})$ shown by the blue lines in Fig. 3.

With this test we may judge the improvement that could be achieved had the model of structure-amplitude errors

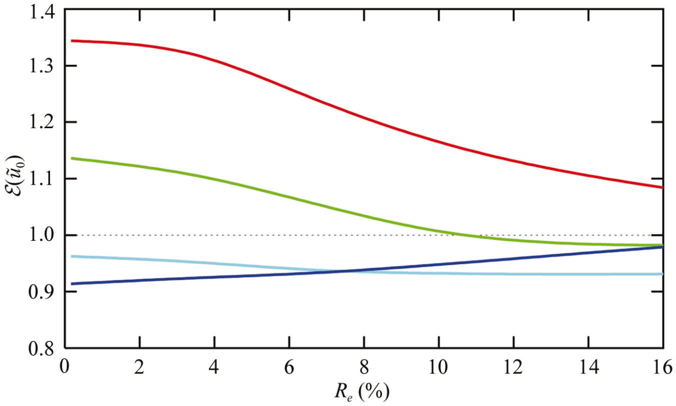

Figure 2
(*a*) The plots of $\mathcal{E}(\tilde{u}_0)$ estimated by the sample means *versus* R_e for a model with Val2 and Val3 substituted with Thr (red) and for models with the following missing atoms: CG1 of Val3 (blue), CG1 and CG2 of Val2 and Val3 (light blue), CD, OE1, OE2 and OXT of Gln5 (green). In these four tests $\mathcal{E}(R_{\rm cryst})$ (%) estimated by the sample means grows from 7 to 16, from 9 to 17, from 20 to 25 and from 24 to 28, respectively.

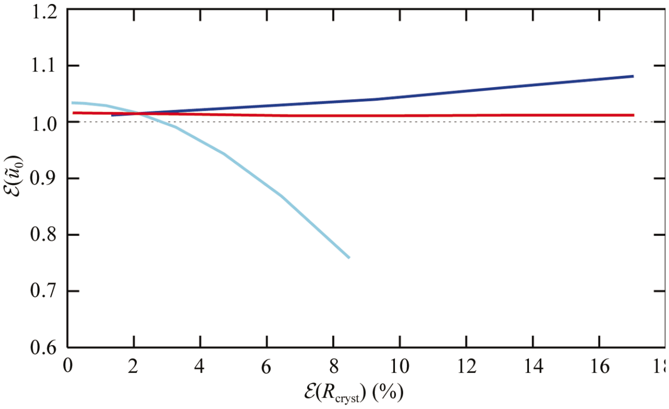

Figure 3
The plots of $\mathcal{E}(\tilde{u}_0)$ *versus* $\mathcal{E}(R_{\rm cryst})$ estimated by the sample means in the cases where the following factors are present in simulations, but ignored in the refinements: (red) strongly non-normal distribution of the observed structure amplitudes, (blue) unaccounted random atoms with low occupancy and (light blue) strong correlations between interatomic distances.

included the errors arising from the missing atoms. Such a correction of the model would remove the bias of $\tilde{u}$ observed in this test, which is significant as it is comparable with the s.u. of $\tilde{u}$ (Fig. 1).

5.4. Neglect of correlation of geometrical errors

In reality, the variations in interatomic distances are correlated. The following exaggerated simulation demonstrates the possible impact of these correlations on $\tilde{u}_0$ if they are not accounted for (with off-diagonal elements of Σ_0).

Suppose that atoms are distributed normally around their position in an 'average crystal', that the s.u.s of all atomic coordinates are equal to 0.04 Å (roughly the s.u. of a distance between neighbouring atoms) and that no correlation between the coordinates occurs. Let the relevant distances in the 'average crystal' (the target distances in the refinement) and their variances be known (the distance variance is the doubled variance of the atomic coordinate in this case). In such circumstances the covariances of neighbouring distances are large, but let them be unknown and ignored in the refinement. In contrast, let our model of the means and covariances of the X-ray measurements used in the refinement be correct. For a sample of such random crystals we simulated X-ray data of different quality (with different R_e) and refined the atomic and weighting parameters. The light-blue plot in Fig. 3 represents the dependence of $\mathcal{E}(\tilde{u}_0)$ on $\mathcal{E}(R_{\rm cryst})$ estimated using the sample means.

In real cases the correlations of the distance variances will be much smaller, as will their effect on $\tilde{u}_0$. However, this effect may be more pronounced in cases with large R factors or with low-resolution X-ray data.

6. Tests with real data

We have chosen three small protein structures from the PDB, making sure that the X-ray data vary in quality in terms of observation-to-parameter ratio. We excluded low-resolution reflections because we had no bulk-water correction. Table 2 contains some characteristics of the crystallographic models and data (which differ slightly from the PDB data owing to the resolution cutoff that we applied) and some results of our refinements, including the values of $\tilde{u}_0$ and the standard uncertainties of $\tilde{u}_0$ estimated, according to the results of the previous subsection, by $Z_{00}^{1/2}$, the square root of the 0th diagonal component of the matrix Z.

The values of $\tilde{u}$ have been obtained by solving the likelihood equation (15), alternating the refinement of atomic parameters with the refinement of weighting parameters. The latter refinement is a time-consuming procedure, as we do not use any approximation for $\sigma(u)$, but compute it exactly according to (11). For the whole refinement to converge, it takes a total of 20–30 cycles of the refinement of weights, each cycle costing about 10 (1gk7, 1be7) or 30 (1jo8) min of CPU time on a DEC Alpha Station.

Table 2 shows that u is significantly less than one and thus the validation equation (7) does not hold. Such an effect is greatest when the observation-to-parameter ratio is lowest, as is the case for data set 1be7. This could be because of the temperature factors, which were held at their PDB values. The tests with simulated data show that lack of off-diagonal terms in the geometrical partition of the covariance matrix could also contribute to depression of the $\tilde{u}_0$ value. Much more experience with real data sets will be needed to learn more about the effects of different error sources on $\tilde{u}_0$.

Table 2
Refinements of three small protein crystal structures.

X-ray data below 5 Å resolution are disregarded and the remainder are partitioned into 25 bins (five resolution bins, each containing five intensity bins) each with an approximately equal number of reflections. The s.u.s multiplied by 10^2 are given in parentheses. The results of these refinements and control refinements with 9 × 9 bins differ insignificantly.

PDB code	1gk7	1jo8	1be7
No. of atoms (total)	375	649	459
No. of atoms (anisotropic)	375	649	0
No. of refined atomic parameters†	1127	1949	1378
No. of reflections (total)	11140	13436	5833
No. of reflections (working set)	10690	12766	5569
No. of reflections (free set)	450	670	264
No. of restraints	796	1219	1021
No. of observations	11486	13985	6590
High-resolution limit (Å)	1.40	1.30	1.65
Observation-to-parameter ratio	10.1	7.1	4.8
Model from PDB			
$R_{\rm cryst}$ (%)	21.5	15.7	17.4
$R_{\rm free}$ (%)	24.1	19.9	19.9
R.m.s. deviations from ideal values			
Bond distances (Å)	0.021	0.011	0.011
Angle distances (Å)	0.036	0.027	0.024
All restrained distances (Å)	0.030	0.022	0.020
Average weighted geometrical residual	0.88	0.40	0.40
Model refined with optimized weights			
$R_{\rm cryst}$ (%)	21.0	14.9	17.3
$R_{\rm free}$ (%)	24.8	20.2	19.9
R.m.s. deviations from ideal values			
Bond distances (Å)	0.012	0.007	0.004
Angle distances (Å)	0.024	0.019	0.009
All restrained distances (Å)	0.020	0.015	0.008
Average weighted geometrical residual	0.34	0.18	0.05
$\tilde{u}_0$ and ($Z_{00}^{1/2} \times 10^2$)	0.73 (8)	0.41 (3)	0.23 (4)

† Number of atomic coordinates plus two overall scaling parameters.

In the case of poorest observation-to-parameter ratio (1be7), the r.m.s. deviation of restrained distances from dictionary values becomes very small. This is the boundary case, where the experimental information about the bond lengths is weak compared with the prior knowledge. In this case the automated (marginal likelihood) weighting suggests the relative weights of X-ray and geometry to be such that we are rather dealing with constrained refinement.

7. Discussion

We have compared two estimators for weighting parameters. In the linear, normal case the estimator $\hat{u}$ is a partition of the maximum-likelihood estimator $(\hat{x}, \hat{u})$, the likelihood function corresponding to the probability distribution of the random vector f^o given the vector of atomic parameters x and the vector of weighting parameters u. The estimator $\tilde{u}$ is a maximum-likelihood estimator itself, the likelihood function corresponding to the probability distribution of the random vector f^o given only the vector of weighting parameters u. A comparison of the observation-to-parameter ratio of the two estimators suggests that, in contrast to $\hat{u}$, the behaviour of $\tilde{u}$ is likely to approach the asymptotic behaviour of maximum-likelihood estimators.

The difference between $\hat{u}$ and $\tilde{u}$ arises from the second term in (10). The use of this term (or, preferably, its approximation) may be required to reduce the bias to an acceptable level if the method of maximum likelihood is used for estimating both atomic and weighting parameters. The estimator $\hat{u}$ (Fig. 1) gives a clear illustration of this point, exhibiting the case when this term is totally disregarded.

The numerical tests presented in this paper correspond to a non-linear and non-normal case, resembling the major features of the usual model of a protein crystal structure. Thus, the estimator $\tilde{u}$ is not exactly a likelihood estimator. Nevertheless, the properties of this estimator (see Fig. 1 and Table 1) are almost those of a maximum-likelihood estimator, suggesting that the assumptions of linearity and normality are acceptable. Additional direct tests with the linearized problem give results that are almost indistinguishable from those listed in Table 1.

In the restrained macromolecular refinement it is assumed that the dictionary contains correct values of interatomic distances and their variances and that correlations between distances is negligible. In this case, we expect $\tilde{u}_0 \simeq 1$ within its uncertainty, provided that the errors in the X-ray term are modelled correctly and that all refined atomic parameters and all restraints are present in $\partial f/\partial x$ in the second term in (10). The latter condition is not satisfied in our tests with real data, where we disregard the refinement of B factors. Therefore, we were not surprised to observe $\tilde{u}_0 < 1$ in these tests. This effect may partially be a consequence of the neglected off-diagonal terms in the geometrical partition of the covariance matrix, as shown by tests with simulated data.

If the models of errors in both geometry and X-ray terms are correct, then the marginal likelihood technique presented in this paper enables the refinement of weighting parameters and the assessment of the quality of the refined structure from the scale factor of the geometrical variances, $\tilde{u}_0$, whose expected value is unity.

8. Further work

We have two immediate aims on the path toward a practical implementation of the maximum-likelihood method to determination of weights for macromolecular refinement. Firstly, we need to incorporate the derivatives of structure amplitudes with respect to atomic temperature factors into the matrix $\partial f/\partial x$ and add extra partitions to f^o, $f(x)$ and Σ in order to describe restraints on temperature factors of neighbouring atoms. Secondly, we need to test different approximations of the second term in (10) in order to effect a compromise between accuracy and efficiency in the refinement algorithm. There could be many approaches to the above tasks and the validation equation $\mathcal{E}(\tilde{u}_0|u) \simeq 1$ could help us to detect better solutions.

The estimator $\tilde{u}$ can be further improved by using a likelihood function corresponding to a more realistic distribution law of the observed random vector. In this case, the quadratic approximation of $L(x, u)$ with respect to x could be used in the integration in (2).

APPENDIX *A*

A1. The case of one unknown weighting parameter

In the special case of the covariance problem in (3) and (4), where $f^o = (x_1, x_2, \ldots, x_n)^T$, $f(x) = (x, x, \ldots, x)^T$, $u = u_0$ and Σ_0 is the unit matrix, (9) becomes

$$L(x, u) = -\frac{1}{2u}\sum_{i=1}^{n}(x_i - x)^2 - \frac{n}{2}\ln(u) \tag{24}$$

and (3) results in

$$\mathcal{E}(s|u) = [(n-1)/n]u, \tag{25}$$

where

$$s = \frac{1}{n}\sum_{i=1}^{n}(x_i)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2. \tag{26}$$

In this case $(x_1, x_2, \ldots, x_N)^T$ is a sample of size n drawn from a random variable with the mean x and with variance u.

In this case,

$$L[t(u), u] = -\frac{ns}{2u} - \frac{n\ln(u)}{2} \tag{27}$$

and

$$M(u) = -\frac{ns}{2u} - \frac{n\ln(u)}{2} + \frac{\ln(u)}{2} - \frac{\ln(n)}{2} \tag{28}$$

and (16) and (15) have analytical solutions given by

$$\hat{u} = s \tag{29}$$

and

$$\tilde{u} = [n/(n-1)]s, \tag{30}$$

respectively. Therefore,

$$\mathcal{E}(\hat{u}|u) = [(n-1)/n]\ u \tag{31}$$

and

$$\mathcal{E}(\tilde{u}|u) = u. \tag{32}$$

In other words, $\hat{u}$ and $\tilde{u}$ are non-corrected and corrected sample variances and therefore are biased and unbiased estimates of u, respectively.

The third and the fourth terms in (28) arise from $\ln[\det\sigma(u)]$ in (10) and it is the third term that performs the bias removal.

This situation also appertains to the case of one unknown weighting parameter, general linear vector-function $f(x)$ and general matrix Σ_0.

We thank Professor Randy Read for helpful discussions. The authors gratefully acknowledge BBSRC funding (grant reference B10670) that made this work possible.

References

Badger, J. & Hendle, J. (2002). *Acta Cryst.* D**58**, 284–291.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst.* D**58**, 899–907.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bricogne, G. (1988). *Acta Cryst.* A**44**, 517–545.

Bricogne, G. & Irwin, J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–474.

Brünger, A. T. (1992*b*). *X-PLOR Manual, Version* 3.1. Yale University, New Haven.

Brünger, A. T. (1993). *Acta Cryst.* D**49**, 24–36.

Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.

Cramér, H. (1946). *Mathematical Methods of Statistics.* Princeton: Princeton University Press.

Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.

Dodson, E. (1998). *Acta Cryst.* D**54**, 1109–1118.

Dodson, E. J., Kleywegt, G. J. & Wilson, K. (1996). *Acta Cryst.* D**52**, 228–234.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Kleywegt, G. J. (1999). *Acta Cryst.* D**55**, 1878–1884.

Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.

Leonard, T. & Hsu, J. S. J. (2001). *Bayesian Methods.* Cambridge University Press.

Lindley, D. V. (1965). *Introduction to Probability and Statistics, Part 2: Inference*. Cambridge University Press.

Lunin, V. Y. & Skovoroda, T. P. (1995). *Acta Cryst.* A**51**, 880–887.

Lunin, V. Y. & Urzhumtsev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968–975.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.

Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

Schwarzenbach, D., Abrahams, S. C., Flack, H. D., Prince, E. & Wilson, A. J. C. (1995). *Acta Cryst.* A**51**, 565–569.

Sim, G. A. (1959). *Acta Cryst.* **12**, 813–815.

Skovoroda, T. P. & Lunin, V. Y. (2000). *Crystallogr. Rep.* **45**, 195–198.

Srinivasan, R. & Ramachandran, G. N. (1965). *Acta Cryst.* **19**, 1008–1014.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998*a*). *Acta Cryst.* D**54**, 243–252.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998*b*). *Acta Cryst.* D**54**, 547–557.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (2000). *Acta Cryst.* D**56**, 442–450.