Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Estimates of the twinning fraction for macromolecular crystals using statistical models accounting for experimental errors

Vladimir Y. Lunin,[a] Natalia L. Lunina[a] and Manfred W. Baumstark[b]*

[a]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino 142290, Russia, and [b]Medizinische Universitätsklinik Freiburg, Hugstetter Strasse 55, D-79106 Freiburg, Germany

Correspondence e-mail: maba@uni-freiburg.de

Received 28 June 2007
Accepted 11 September 2007

An advanced statistical model is suggested that is designed to estimate the twinning fraction in merohedrally (or pseudo-merohedrally) twinned crystals. The model takes experimental errors of the measured intensities into account and is adapted to the accuracy of a particular X-ray experiment through the standard deviations of the reflection intensities. The theoretical probability distributions for the improved model are calculated using a Monte Carlo-type simulation procedure. The use of different statistical criteria (including likelihood) to estimate the optimal twinning-fraction value is discussed. The improved model enables better agreement of theoretical and observed cumulative distribution functions to be obtained and produces twinning-fraction estimates that are closer to the refined values in comparison to the conventional model, which disregards experimental errors. The results of the two approaches converge when applied to selected subsets of measured intensities of high accuracy.

1. Introduction

In recent decades, twinning has been shown to be an important feature of macromolecular crystals (Yeates, 1997; Yeates & Fam, 1999; Dauter, 2003; Parsons, 2003; Lebedev *et al.*, 2006). Twinned crystals are composed of separate differently orientated crystal domains. If certain conditions for unit-cell parameters and orientation of the domains are met (merohedral twinning), the reciprocal-space lattices of different domains coincide and the measured intensity of a diffracted beam becomes the sum of two (or more) different 'true' intensities that come from different domains. In this case, the structure-factor magnitude can no longer be estimated as the square root of the corresponding measured intensity and special efforts must be applied to restore it. The possibility of merohedral twinning occurs when the symmetry of the crystal lattice is higher than the symmetry of the unit-cell content. It is sometimes the case that the lattice has additional approximate symmetry not conditioned by crystal syngony (for example, the angle $\beta \simeq 90^\circ$ in space group $P2$). Such cases are usually referred to as pseudo-merohedral twinning. We do not distinguish between merohedral and pseudo-merohedral twinning in this paper, assuming sufficiently small obliquity (such that the complete overlap of spots from different lattices occurs within the whole resolution range of the data set). For brevity, we restrict our consideration in this paper to one twofold twin operator (hemihedrally twinned specimens). More complicated cases may be considered in a similar way.

In the case of hemihedral twinning, every measured intensity is a linear combination of two 'true' intensities with coefficients depending on the relative volume of the smaller

Table 1
Crystals used to test the method.

The twinning is merohedral for WGA-div and 1l2h and pseudo-merohedral for other crystals.

Crystal	Reference	d_{min} (Å)	Space group	Unit-cell parameters (Å, °)	Twinning law
1c5e	Capsid-stabilizing protein of bacteriophage λ (Yang *et al.*, 2000)	1.1	$P2_1$	$a = 45.66, b = 69.03, c = 45.67, \alpha = \gamma = 90.0, \beta = 104.34$	$l\ -k\ h$
WGA-div	Wheat-germ agglutinin (Diederichs *et al.*, in preparation)	1.7	$R3$	$a = b = 101.3, c = 144.9, \alpha = \beta = 90.0, \gamma = 120.0$	$k\ h\ -l$
WGA-18	Wheat-germ agglutinin (Diederichs *et al.*, in preparation)	1.4	$P2_1$	$a = 44.4, b = 87.7, c = 44.5, \alpha = \gamma = 90.0, \beta = 111.7$	$l\ -k\ h$
LDL l463	Low-density lipoprotein (Ritter *et al.*, 1999; Baumstark *et al.*, work in progress)	27.0	$C2$	$a = 181.5, b = 425.5, c = 390.8, \alpha = \gamma = 90.0, \beta = 91.2$	$h\ -k\ -l$
1l2h	Interleukin-1β (Rudolph *et al.*, 2003)	1.5	$P4_3$	$a = b = 53.89, c = 77.36, \alpha = \beta = \gamma = 90.0$	$-h\ k\ -l$

twin mate. This relative volume is usually referred to as the twinning fraction or twinning ratio. If the twinning fraction α is known (and is not equal to 0.5), the true intensities can be restored. The key problem here is to define the value α. To estimate the twinning fraction, several methods have been suggested (Fisher & Sweet, 1980; Murray-Rust, 1973; Britton, 1972; Rees, 1980; Yeates, 1988; Redinbo & Yeates, 1993; Gomis-Rüth *et al.*, 1995; Lebedev *et al.*, 2006). In this paper, we base our estimation on the method suggested by Yeates (1988). This method is based on the study of the normalized difference H of two twinned intensities $J^{obs}(\mathbf{h})$ and $J^{obs}(\mathbf{Gh})$ (where $\mathbf{G}$ represents the twinning operator in reciprocal space and $\mathbf{h}$ varies). The set of H values calculated for different twinned pairs of observed intensities is studied with the use of a collection of theoretical distributions of H corresponding to different twinning-fraction values. The customary theoretical distributions are derived from a simple statistical model that considers the experimental errors to be negligible. An experimental set

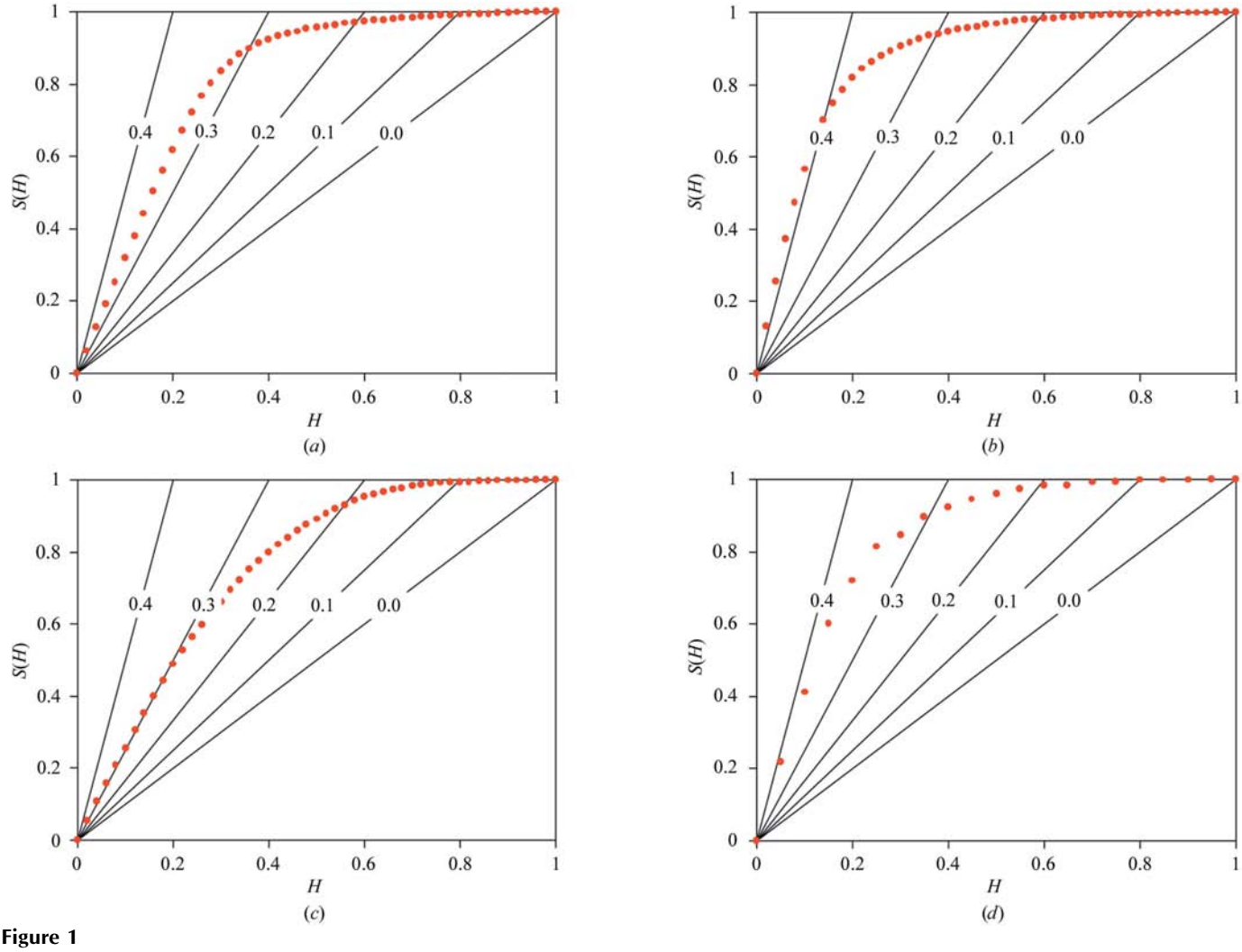

Figure 1
Cumulative distributions of H for four twinned crystals: (*a*) 1c5e, (*b*) WGA-div, (*c*) WGA-18 and (*d*) LDL l463. The observed distributions are represented by red dotted lines. Thin black lines represent theoretical cumulative distributions derived for different twinning fractions using the commonly used model that does not account for experimental errors. The values of the twinning fractions are shown over the corresponding lines.

of H values fits one of the theoretical distributions with reasonable quality if the measurement errors are small and the resolution of the data set is high enough, but the correspondence becomes poor in other cases (Fig. 1). To improve the correspondence, a more sophisticated statistical model is suggested below that takes experimental errors into account. This model does not allow the derivation of a set of theoretical distributions in a closed analytical form, but this problem is overcome with the use of a Monte Carlo-type simulation procedure. The error-accounting theoretical distributions reveal a better correspondence to the empirical distributions (Fig. 2) and allow a more accurate estimation of the twinning fraction.

To illustrate this approach, several experimental data sets of different quality were used. Table 1 summarizes their main parameters. This paper was inspired by problems we encountered when detecting twinning in very low resolution data sets from low-density lipoprotein particles. Nevertheless, the approach developed might also be useful when working with medium-resolution data sets.

2. Statistical modelling of *H*-ratio value: model disregarding experimental errors

Let us suppose that the twinning law is specified and so for every reflection $\mathbf{h}_1$ its twinning mate $\mathbf{h}_2 = \mathbf{G}\mathbf{h}_1$ is known. In the following, we distinguish three types of intensities.

(i) 'True' intensities $I^{\rm true}(\mathbf{h})$ that correspond to a uniquely orientated domain; they can be calculated as squares of the structure-factor magnitudes corresponding to the unit-cell content.

(ii) 'Theoretical' (error-free) twinned intensities corresponding to a hemihedrally twinned sample (*i.e.* $\mathbf{G}\mathbf{h}_2 = \mathbf{G}\mathbf{G}\mathbf{h}_1 = \mathbf{h}_1$); they are calculated as

$$J^{\rm theor}(\mathbf{h}_1) = (1-\alpha)I^{\rm true}(\mathbf{h}_1) + \alpha I^{\rm true}(\mathbf{h}_2)$$
$$J^{\rm theor}(\mathbf{h}_2) = (1-\alpha)I^{\rm true}(\mathbf{h}_2) + \alpha I^{\rm true}(\mathbf{h}_1), \qquad (1)$$

where α represents the twinning fraction (twinning ratio) and is equal to the relative volume of the smaller twin mate.

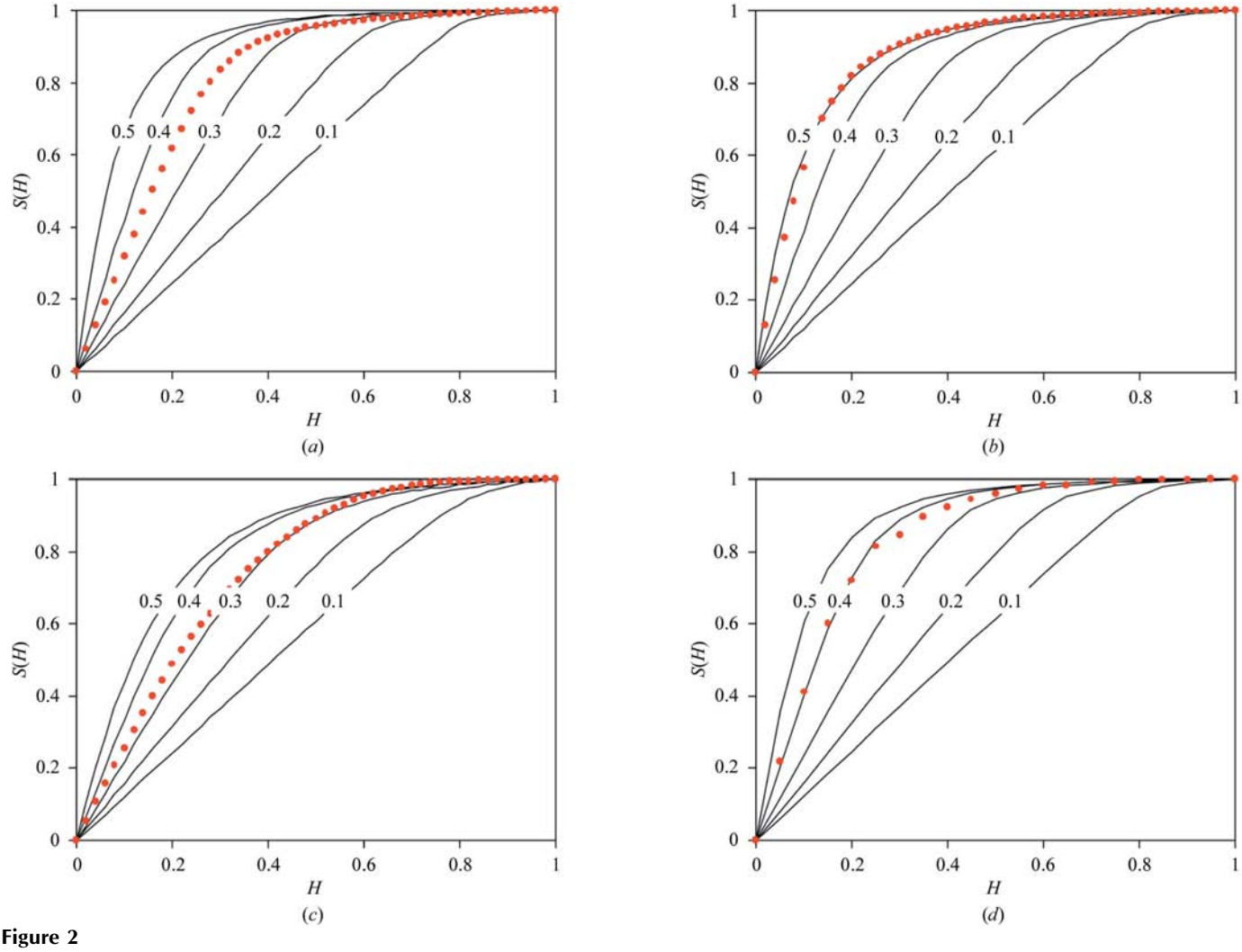

Figure 2
Corrected cumulative distributions of H for four twinned crystals: (*a*) 1c5e, (*b*) WGA-div, (*c*) WGA-18 and (*d*) LDL l463. The observed distributions are represented by red dotted lines. Thin black lines represent theoretical cumulative distributions derived for different twinning fractions with the use of a statistical model that takes experimental errors into account. The values of the twinning fractions are shown over the corresponding lines.

Table 2
Twinning fraction estimated from different statistical criteria.

Initial estimates of the twinning fraction were obtained using the statistical model disregarding experimental errors (13). Refined estimates (§6) are shown as reported by the authors of the structures. The estimates accounting for experimental errors correspond to the four criteria of consistency of theoretical distributions with observed H values as described in §§5.1–5.4.

			Estimates accounting for experimental errors			
Crystal	Initial estimate	Refined estimate	Mean	Likelihood	χ^2	D_{KS}
1c5e	0.31	0.36	0.34	0.36	0.36	0.35
WGA-div	0.37	0.44	0.47	0.46	0.46	0.46
WGA-18	0.25	0.33	0.30	0.31	0.32	0.32
LDL l463	0.34	—	0.39	0.41	0.41	0.40
1l2h	0.30	0.37	0.34	0.34	0.34	0.34

(iii) 'Observed' intensities $J^{obs}(\mathbf{h})$ that were obtained in an X-ray experiment; they differ from theoretical twinned intensities $J^{theor}(\mathbf{h})$ by experimental errors $\delta(\mathbf{h})$,

$$J^{obs}(\mathbf{h}_1) = J^{theor}(\mathbf{h}_1) + \delta(\mathbf{h}_1)$$
$$J^{obs}(\mathbf{h}_2) = J^{theor}(\mathbf{h}_2) + \delta(\mathbf{h}_2). \quad (2)$$

The goal of the following study is to find a realistic estimate of the twinning fraction α in (1) starting from the set of experimentally observed intensities $\{J^{obs}(\mathbf{h})\}$. In this paper, we discuss an approach (Yeates, 1988) based on the statistics of the discrepancy between the intensities of twinned reflections. For every pair of twinned reflections $\mathbf{h}_1$, $\mathbf{h}_2$, we define the observed discrepancy as

$$H^{obs} = H^{obs}(\mathbf{h}_1, \mathbf{h}_2) = \frac{|J^{obs}(\mathbf{h}_1) - J^{obs}(\mathbf{h}_2)|}{J^{obs}(\mathbf{h}_1) + J^{obs}(\mathbf{h}_2)}. \quad (3)$$

The set of experimentally observed intensities provides the set of observed H values $\{H_1^{obs}, H_2^{obs}, \ldots, H_M^{obs}\}$ calculated as in (3), where M is the number of twinned pairs in the respective set of intensities. [In this paper, we restrict our consideration to the case of all $J^{obs}(\mathbf{h})$ being positive and so $0 \leq H < 1$.]

If the experimental errors $\delta(\mathbf{h})$ are negligible, then (3) is reduced to

$$H^Y = \frac{(1-2\alpha)|I^{true}(\mathbf{h}_1) - I^{true}(\mathbf{h}_2)|}{I^{true}(\mathbf{h}_1) + I^{true}(\mathbf{h}_2)} = \frac{(1-2\alpha)|z_1 - z_2|}{z_1 + z_2}, \quad (4)$$

where

$$z_1 = \frac{I^{true}(\mathbf{h}_1)}{\langle I \rangle}, \quad z_2 = \frac{I^{true}(\mathbf{h}_2)}{\langle I \rangle} \quad (5)$$

are normalized intensities and the normalizing factor $\langle I \rangle$ is assumed to be the same for both $\mathbf{h}_1$ and $\mathbf{h}_2$. It is worthy of note that in (4) the value H is defined through the initially unknown quantities $I^{true}(\mathbf{h}_1)$, $I^{true}(\mathbf{h}_2)$. This makes direct calculation of H^{obs} with the use of (4) impossible, but allows theoretical study.

It is easy to see that for α close to 0.5, the values of H for many twinned pairs will be close to zero. In the opposite case, when α is close to zero many H values will be significantly different from zero, reflecting the difference of the unrelated intensities z_1 and z_2. This demonstrates that the distribution of H values through the set of twinned pairs of reflections is sensitive to the twinning fraction α and may in principle be used to estimate α. A statistical approach provides a tool to determine this estimator.

The basis of the use of the H statistics is the hypothesis that after a proper normalization (5) (the normalizing value $\langle I \rangle$ is generally different for different reflections) the normalized intensities for noncentric reflections obey Wilson statistics (Wilson, 1949), *i.e.* the value of the normalized intensity z for a randomly chosen reflection $\mathbf{h}$ may be considered as a random variable of exponential distribution,

$$P(z) = \exp(-z). \quad (6)$$

(The term 'Wilson distribution' is customarily used in crystallography when applied to intensity distribution.) On the basis of the definition of H in (4) and on the hypothesis that z_1, z_2 in (4) are independent random variables distributed by (6) it becomes possible (Yeates, 1988) to consider H^Y as a random variable and to calculate the theoretical probability distribution $P_{H,\alpha}^{Yeates}(H)$ corresponding to a particular twinning fraction value α. The cumulative functions for these distributions for several α values are shown in Fig. 1 and may be presented as

$$S_{H,\alpha}^{Yeates}(t) = \begin{cases} 0 & \text{for } t < 0 \\ \dfrac{1}{(1-2\alpha)}t & \text{for } 0 \leq t \leq 1-2\alpha \\ 1 & \text{for } t > 1-2\alpha \end{cases}. \quad (7)$$

The calculation of $\{H_1^{obs}, H_2^{obs}, \ldots, H_M^{obs}\}$ values by means of (3) from the experimental data allows an observed cumulative function to be obtained,

$$S_H^{obs}(t) = \frac{\{\text{No. of twinned pairs } \mathbf{h}_1, \mathbf{h}_2 \text{ resulting in } H^{obs}(\mathbf{h}_1, \mathbf{h}_2) \leq t\}}{\{\text{total no. of twinned pairs considered}\}}. \quad (8)$$

The comparison of the plot of this observed cumulative function with the set of theoretical plots (7) allows a decision on the presence of merohedral twinning to be made and an estimation of the twinning-fraction value α as the value α^{opt} that provides the best fit of the corresponding theoretical distribution $S_{H,\alpha}^{Yeates}$ to the observed distribution S_H^{obs}.

The determination of the twinning factor on the basis of theoretical cumulative distributions (7) has been successfully applied for many years using high-resolution data of reasonable accuracy. Nevertheless, the choice of α^{opt} by means of comparing observed and theoretical distributions (7) becomes less obvious when the data accuracy is relatively low or the data are restricted to low resolution (Figs. 1*c* and 1*d*). A possible explanation of the disagreement of observed with theoretical curves may be the presence of experimental errors in the observed intensities that are disregarded in the model

(4). These errors are sometimes considered to be insignificant, but it will be shown below that they can substantially change the shape of theoretical distributions.

3. Statistical modelling of the *H* ratio accounting for experimental errors

If experimental errors are assumed to be essential, then (4) is replaced by

$$H = \left| \frac{(1-2\alpha)(z_1 - z_2) + \dfrac{\delta_1}{\langle I \rangle} - \dfrac{\delta_2}{\langle I \rangle}}{z_1 + z_2 + \dfrac{\delta_1}{\langle I \rangle} + \dfrac{\delta_2}{\langle I \rangle}} \right|, \qquad (9)$$

where $\delta_1 = \delta(\mathbf{h}_1)$, $\delta_2 = \delta(\mathbf{h}_2)$ represent the experimental errors when measuring $J^{\rm theor}(\mathbf{h}_1)$, $J^{\rm theor}(\mathbf{h}_2)$. The theoretical study of such an extended model requires decisions on the following.

(i) A statistical model for experimental errors.

(ii) Theoretical cumulative functions $S^{\rm theor}_{H,\alpha}$ for different α values.

(iii) 'Expected' intensities $\langle I \rangle(\mathbf{h})$ for every pair of reflections. It is worth mentioning that we did not need to know the particular values while working with (4); it was sufficient to assume that both reflections from a twinned pair have the same $\langle I \rangle$ value.

In our study, we assume δ_1, δ_2 to be normally distributed independent random variables with zero mean and a variance derived from the experimental estimate $\sigma^{\rm obs}(\mathbf{h})$ of the accuracy of $J^{\rm obs}(\mathbf{h})$. In the simplest case we can model the standard deviation of $\delta(\mathbf{h})$ simply as $\sigma^{\rm obs}(\mathbf{h})$, but more complicated models may also be considered (see Appendix *A*). It is noteworthy that the input of the procedure is now both $\{J^{\rm obs}(\mathbf{h})\}$ and $\{\sigma^{\rm obs}(\mathbf{h})\}$, so that the derived theoretical cumulative functions for H are no longer universal but relate to a particular X-ray experiment.

The development of analytical expressions for the distribution of random value (9) is a difficult mathematical task and these expressions are hardly ever obtained in the closed form in a general case. To overcome this difficulty, a Monte Carlo-type procedure can be used to obtain these distributions by means of computer simulation. In this approach, to calculate theoretical distributions $S^{\rm theor}_{H,\alpha}(t)$, a number of trials is performed that consist of generating random variables z_1, z_2, δ_1, δ_2 followed by calculation of H in accordance with (9). The cumulative functions obtained in such a computer simulation can be considered as an approximation of the theoretical functions. This approach is close to that used previously to calculate probability distributions in the case of absent analytical expression (Lunin *et al.*, 1998; Petrova *et al.*, 2000; Zwart, 2005).

The expected intensities $\langle I \rangle(\mathbf{h})$ in our study were assumed to be constant in thin shells in reciprocal space and were estimated from the observed intensities $J^{\rm obs}(\mathbf{h})$ in corresponding shells. Three approaches were evaluated to estimate $\langle I \rangle(\mathbf{h})$: the mean, the mean weighted by $\sigma^{\rm obs}(\mathbf{h})$ and the median value of the intensity in the shell (see §6).

4. Visual analysis of test cases

To check to what extent the complicated model improves the correspondence of theoretical and experimental distributions, several experimental data sets of different quality were used. Table 1 summarizes their main parameters. It is worthy of note that in our approach both measured intensities $\{J^{\rm obs}(\mathbf{h})\}$ and estimates of their accuracy $\{\sigma^{\rm obs}(\mathbf{h})\}$ form the input of the procedure, so that accurate estimates of $\{\sigma^{\rm obs}(\mathbf{h})\}$ become important for the success of the procedure. The deposited $\{\sigma^{\rm obs}(\mathbf{h})\}$ values were used in our tests for these purposes. They had been calculated by *HKL* (Otwinowski & Minor, 1997) in the case of 1c5e and 1l2h and by *XDS* (Kabsch, 1993) for the other three test cases.

Fig. 1 presents observed cumulative functions for four test cases superposed with customary theoretical cumulative functions (7) that disregard experimental errors. Fig. 2 presents the same observed cumulative functions, but now superposed with the simulated theoretical cumulative functions that take experimental errors into account. Comparison of these two sets of figures shows that the correction for measurement errors essentially changes the shape of the theoretical cumulative function, making it much closer to the observed cumulative function.

A more precise analysis of test cases is presented in Table 2 and is discussed in the following sections.

5. The choice of a theoretical model

The choice of a theoretical distribution that is most consistent with the observed values is a standard problem in the theory of probability and there are many approaches to solve it. Four of them used in our study are discussed below.

5.1. Method of moments

A traditional approach to define the twinning fraction α on the basis of theoretical distributions (7) is the simplest type of method of moments. It suggests the optimal value $\alpha^{\rm opt}$ as one that results in the theoretically expected value

$$\langle H \rangle_\alpha = \int H P^{\rm theor}_{H,\alpha}(H)\, {\rm d}H \qquad (10)$$

equal to the experimental mean

$$\langle H \rangle_{\rm obs} = \frac{1}{M} \sum_{j=1}^{M} H_j^{\rm obs}. \qquad (11)$$

For the statistical model disregarding experimental errors, direct calculation with the use of (7) gives

$$\langle H \rangle_\alpha = \frac{1}{2} - \alpha, \qquad (12)$$

which results in the estimate

$$\alpha^{\rm opt} = \frac{1}{2} - \langle H \rangle_{\rm obs} = \frac{1}{2} - \frac{1}{M} \sum_{j=1}^{M} H_j^{\rm obs} \qquad (13)$$

used in many computer programs, such as the *CCP*4 program suite (Collaborative Computational Project, Number 4, 1994) programs *DETWIN* and *SFCHECK* (Vaguine *et al.*, 1999).

Fig. 3 shows theoretical distributions disregarding experimental errors corresponding to $\alpha^{\rm opt}$ as defined in (13).

In the case of error-sensitive models, the expected values $\langle H\rangle_\alpha$ for different α may be calculated by Monte Carlo simulation (Fig. 4). The corresponding values $\alpha^{\rm opt}$ derived from the condition $\langle H\rangle_\alpha = \langle H\rangle_{\rm obs}$ are generally higher than those calculated as in (13), but this increment is only essential for large twinning fractions.

5.2. Likelihood-based estimation of the twinning fraction α

Let $H_1^{\rm obs}, H_2^{\rm obs}, \ldots, H_M^{\rm obs}$ be the values calculated as in (3) based on the input list of twinned pairs and observed intensities. If it is assumed that these values were obtained in the process of generating H values randomly with a $P_{H,\alpha}^{\rm theor}(H)$ distribution, then the statistical likelihood

$$L = L(\alpha) = \prod_{j=1}^{M} P_{H,\alpha}^{\rm theor}(H_j^{\rm obs}) \qquad (14)$$

might be used as a measure of the consistency of this hypothesis and the observed values. This likelihood value might be interpreted as the probability of reproducing the observed values $\{H_j^{\rm obs}\}$ in the framework of the statistical hypothesis tested. The likelihood-based estimate of the twinning fraction is the $\alpha^{\rm opt}$ that maximizes likelihood. Obviously, the logarithm of the likelihood is a more convenient target function for practical purposes,

$$-{\rm LLG}(\alpha) = -\log L(\alpha) = -\sum_{j=1}^{M} \log P_{H,\alpha}^{\rm theor}(H_j^{\rm obs}) \Rightarrow \min . \qquad (15)$$

Plots of $-{\rm LLG}(\alpha)$ for test cases are shown in Fig. 5.

5.3. χ^2 criterion

One of the most popular criteria when fitting theoretical curves to observed data is the χ^2 criterion. Depending on the circumstances, it may have a slightly different form. In the case of simulated theoretical distributions, it might be presented as

$$\chi^2 = \sum_{k=1}^{K_{\rm bin}} \frac{\left[\left(\frac{N_{\rm total}^{\rm theor}}{N_{\rm total}^{\rm obs}}\right)^{1/2} n_k^{\rm obs} - \left(\frac{N_{\rm total}^{\rm obs}}{N_{\rm total}^{\rm theor}}\right)^{1/2} n_k^{\rm theor}\right]^2}{n_k^{\rm obs} + n_k^{\rm theor}}. \qquad (16)$$

It is assumed here that the interval [0, 1] of H values was disjointed into $K_{\rm bin}$ bins; $n_k^{\rm theor}$ is the number of H values obtained in the simulation that fell in the kth bin and $N_{\rm total}^{\rm theor}$ is the total number of simulated H values. $n_k^{\rm obs}$ and $N_{\rm total}^{\rm obs} = M$ are the same for observed $\{H_j^{\rm obs}\}$ values. The plots of $\chi^2(\alpha)$ for test cases are shown in Fig. 5.

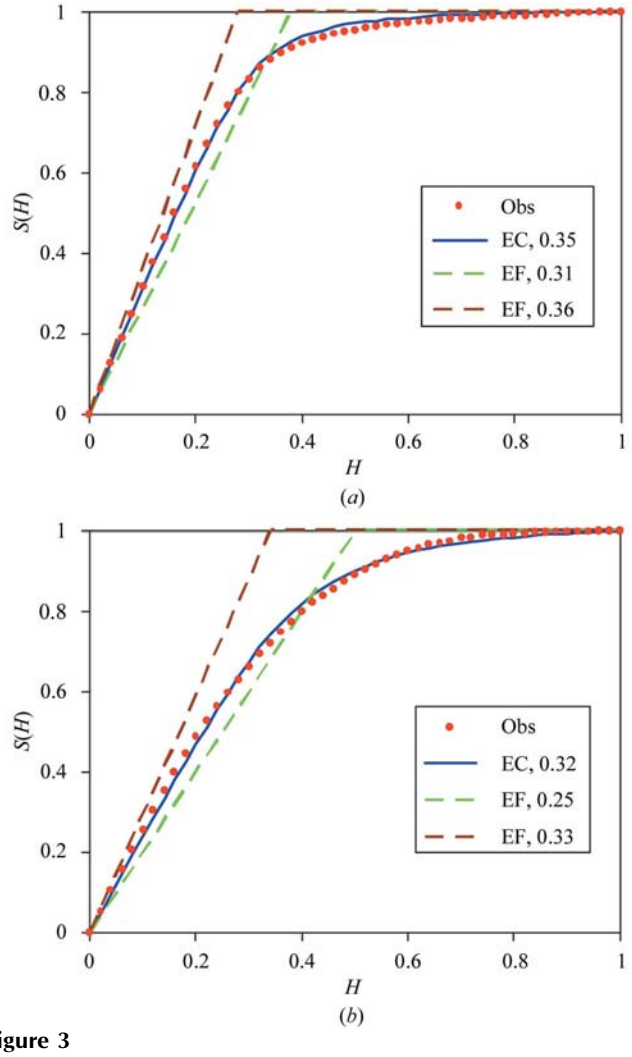

Figure 3
Cumulative distributions of H for (*a*) 1c5e and (*b*) WGA-18. The observed distributions are represented by red dotted lines. The broken lines represent theoretical cumulative distributions not accounting for experimental errors (EF). The two chosen twinning-fraction values correspond to the initial values estimated as in (13) (green) and the reported refined values (brown). The solid lines represent theoretical cumulative distributions derived on the basis of the statistical model that takes into account experimental errors (EC). The values of the twinning fraction chosen correspond to the optimal values derived using the Kolmogorov–Smirnov criterion (see §5.4).

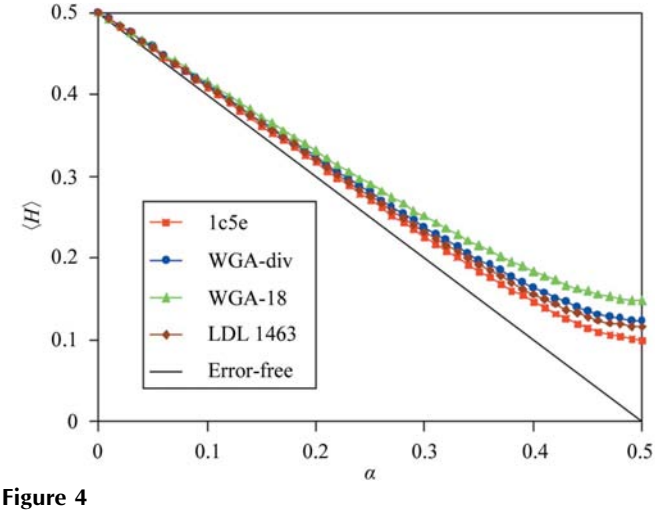

Figure 4
The expected value of H *versus* twinning fraction α for statistical models that take into account experimental errors (the four marked lines) and for the conventional approach (thin diagonal line) that neglects experimental errors.

5.4. Kolmogorov–Smirnov test

One further test used in our studies is based on the direct comparison of theoretical and empirical cumulative functions. The measure of consistency in this case is defined as

$$D_{\rm KS}(\alpha) = \sup_{0\leq H\leq 1} |S^{\rm theor}_{H,\alpha}(H) - S^{\rm obs}(H)|. \qquad (17)$$

The plots of $D_{\rm KS}(\alpha)$ for test cases are shown in Fig. 5.

6. Discussion

The visual analysis shows that theoretical cumulative functions calculated considering experimental errors fit the curves calculated from observed intensities much better than theoretical cumulative functions disregarding experimental errors. The accuracy of the values of the twinning fraction α estimated by different means can be checked retrospectively after the final refinement of the atomic model taking twinning into account and considering the twinning fraction as an adjustable parameter. Such refinement can be performed, for example, with the use of *SHELXL* (Sheldrick & Schneider, 1997; Herbst-Irmer & Sheldrick, 1998), *XTAL* (Hall *et al.*, 2000) and *Phenix.refine* (Adams *et al.*, 2002; Afonine *et al.*, 2005). Table 2 gives the results of comparison of twinning-fraction estimates with the final refined value for several test structures. It can be concluded from this comparison that error-sensitive models allow more accurate estimates of the twinning fraction. It can be seen in Table 2 that the simplest estimates obtained from the simulated mean values of H are close to those obtained using more sophisticated methods. Nevertheless, analysis of plots of different criteria allows estimation of the reliability of the calculated value of the twinning fraction. Our tests have shown that the Kolmogorov–Smirnov criterion seems to be the most sensitive to the particular value of the twinning fraction α (Fig. 5), but it is also sensitive to the quality of the input set of $\{\sigma^{\rm obs}(\mathbf{h})\}$ values. The likelihood-based criterion, in contrast, has the broadest minimum but appears to be a more robust criterion. Overall, the calculation of all discussed criteria might represent the most reasonable strategy when analysing twinning.

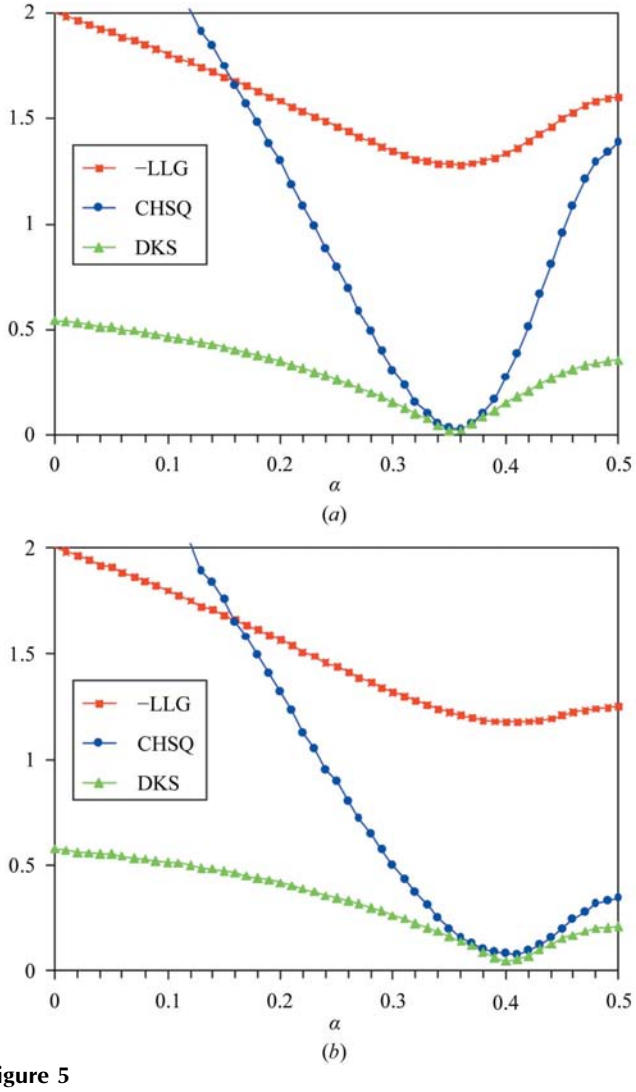

Figure 5
Three statistical criteria *versus* twinning fraction α for statistical models that take into account experimental errors. The values of the Kolmogorov–Smirnov criterion (DKS) are shown on a true scale. The values of the minus log likelihood (−LLG) and χ^2 (CHSQ) criteria are scaled for presentation purposes (the calculated χ^2 values were divided by 15 000 for 1c5e and for 200 for LDL; the corresponding numbers of bins were 50 and 20). (*a*) 1c5e, (*b*) LDL l463.

It is worthwhile noting that refinement of the twinning fraction is not yet common practice in macromolecular crystallography and requires some caution. As an example, Schneider *et al.* (2000) reported that in their work the value of the twinning fraction refined to a relatively large value (0.34) in the very first round of refinement and then gradually decreased towards its final value (0.24). A further problem mentioned by one of the referees is that within one crystalline specimen the twinning fraction can change throughout the whole volume of the sample. Consequently, when the crystal rotates during data collection in some cases (*e.g.* if the crystal is larger than the beam cross-section) different parts of the data may correspond to different twinning fractions, confusing the twinning estimators. A similar note could be made in the case when the crystal is shifted with respect to the beam during the course of an X-ray experiment (for example, when collecting different wavelengths). Data sets for different wavelengths may have different twinning fractions in this case.

It must be pointed out that some caution is necessary when forming the list of twinned mates for the calculation of different statistics (and in particular the mean value of H).

(i) The reflections that are on the twin axes should be excluded, as they result in $\mathbf{h}_1 = \mathbf{h}_2$ and $H = 0$.

(ii) The pairs of reflections $\mathbf{h}_1$, $\mathbf{h}_2$ should be excluded if both reflections belong to the same orbit of the Laue group (or of the crystal symmetry group if Friedel pairs are treated separately); in this case $I(\mathbf{h}_1) = I(\mathbf{h}_2)$ and $H = 0$.

(iii) The reflection $\mathbf{h}_1$, $\mathbf{h}_2$ should not belong to the centric zone, as the Wilson distribution for centric reflections should be used in this case instead of (6).

As pointed out in the original papers by Yeates, the $S(H)$ test does not require that all measured reflections must be used. Therefore, both models may be tried out with specially

selected subsets of measured intensities. This idea does not encounter any difficulty if the intensities are chosen randomly from the input list. As an example, an $R_{\rm free}$ type of approach can be used: *i.e.* one randomly chosen half of the reflections may be used to derive theoretical distributions, while the other half is used to calculate the observed cumulative function (8). Our tests have not revealed a visible difference from the case in which all reflections are used simultaneously, but such an approach may indicate problems in more difficult cases.

Some theoretical difficulties can be encountered if the selected subset includes the most accurately measured reflections only, *i.e.* those with

$$\frac{I^{\rm obs}(\mathbf{h})}{\sigma^{\rm obs}(\mathbf{h})} > \kappa, \tag{18}$$

where κ is some chosen cutoff. On the one hand, in this case the convergence of theoretical cumulative functions corresponding to models not accounting and accounting for experimental errors can be expected, as the influence of errors in (9),

$$\frac{\delta}{\langle I\rangle} = \frac{1}{\kappa}\cdot\frac{\delta}{\sigma^{\rm obs}}\cdot\frac{I^{\rm obs}}{\langle I\rangle}, \tag{19}$$

becomes small if the accuracy κ is sufficiently large. On the other hand, when selecting reflections on a special basis one can no longer be sure that the randomly chosen intensities obey Wilson's distribution. It is easy to overcome this difficulty in the Monte Carlo simulating procedure, but it produces problems regarding theoretical considerations. Nevertheless, our tests revealed that the estimate of α calculated using (13) converges to the refined twinning-fraction value if the accuracy κ of the intensities selected for analysis is sufficiently high (Fig. 6). It is worthwhile noting that the estimates of the twinning fraction obtained with the model accounting for experimental errors change only slightly with κ. Other selection strategies may also be incorporated in the Monte Carlo-type simulating procedure.

In the considerations above, we assumed that the twinning law is fixed in advance. Nevertheless, the same procedure might be used for detecting possible twin axes in situations when the twinning law is not yet established. In this regard, the possibility of distinguishing perfect twinning ($\alpha = 0.5$) from a crystallographic dyad or a screw axis may be of special interest. In the model disregarding experimental errors both cases result in zero values of H and cannot be distinguished using H statistics. These two cases may only be differentiated on the basis of the statistics of intensities, such as the Wilson ratio, the cumulative intensity $N(z)$ test or the Padilla–Yeates L-test (Padilla & Yeates, 2003). In the error-sensitive model, the values H are calculated differently,

$$H_{0.5} = \left|\frac{\xi_1 - \xi_2}{z_1 + z_2 + \xi_1 + \xi_2}\right|, \quad H_{\rm sym} = \left|\frac{\xi_1 - \xi_2}{2z_1 + \xi_1 + \xi_2}\right| \tag{20}$$

(where ξ represents $\delta/\langle I\rangle$). Accordingly, there is a chance of distinguishing these two cases on the basis of H statistics. Unfortunately (see Fig. 7), although a difference between these distributions exists, it is fairly small, resulting in a low validity of such a tool.

In our studies, we assumed the 'expected intensity' $\langle I\rangle(\mathbf{h})$ to be the same for all reflections in a thin shell in reciprocal space. Obviously, this is not the case when high anisotropy is present in the data. Statistics based on local intensity differences (Padilla & Yeates, 2003) may be a more sensitive tool in this case. We evaluated three strategies to estimate expected intensities $\langle I\rangle(\mathbf{h})$ in shells in reciprocal space: the mean, the mean weighted by $\sigma^{\rm obs}(\mathbf{h})$ and the median value of the intensity in the shell. These approaches gave comparable results, with minor differences. Formally, the method of defining $\langle I\rangle(\mathbf{h})$ may be considered to be a parameter of a statistical model. The approach resulting in the best values of the criterion might be selected as the most appropriate for a particular case.

In this paper, we discuss the simplest statistical model, which contains only one adjustable parameter, namely the

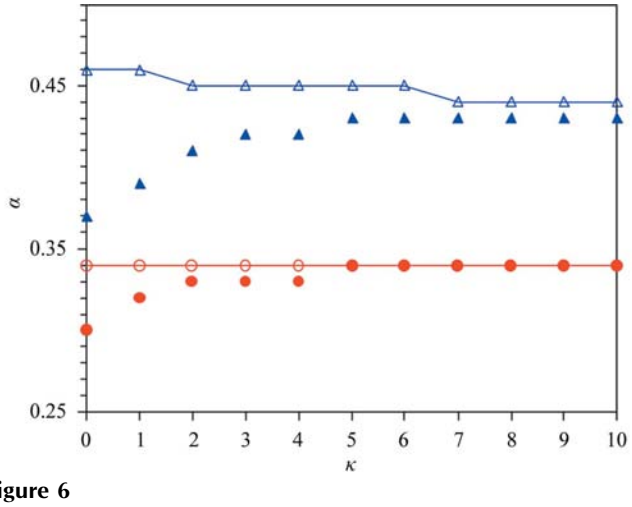

Figure 6
Estimates of twinning fraction α obtained on the basis of intensities with $I^{\rm obs}/\sigma^{\rm obs} > \kappa$ using the model disregarding experimental errors (13) (markers only) and by minimization of χ^2 (solid lines with markers) for 1c5e (circles) and WGA-div (triangles).

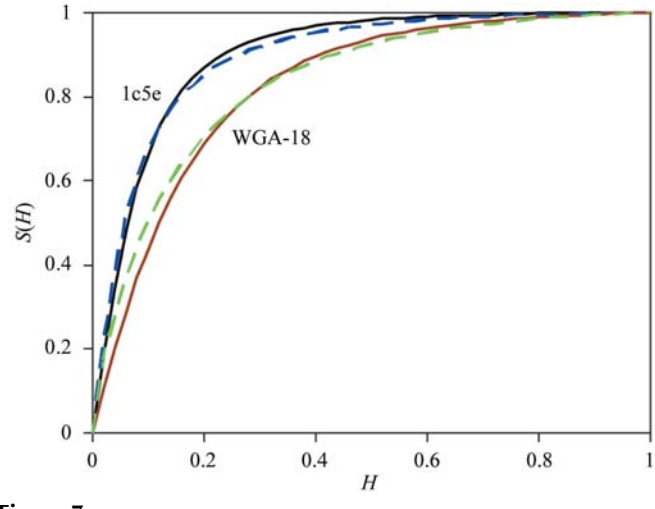

Figure 7
The error-sensitive theoretical cumulative distributions for H for the cases of perfect twinning (solid lines) and crystallographic symmetry (broken lines) for 1c5e and WGA-18 crystals.

twinning fraction α. Formally speaking, the model can easily be extended to more complicated cases with several adjustable parameters. For example, several twin operations (with different twinning fractions) may be studied simultaneously or the expected intensity values in resolution shells $\langle I\rangle(\mathbf{h})$ may be assumed to be adjustable parameters. At the same time, such an approach requires multidimensional minimization and coupling it with indirect (Monte Carlo) calculation of the target function may present serious computational difficulties. Nevertheless, such efforts may be worthwhile in some cases. The reflections related by the twinning operator are often simultaneously related by the NCS operator (Lebedev *et al.*, 2006). As suggested by one of the referees, the developed approach can be applied in this case with the addition of only one more parameter describing the dependency of the correlation coefficient between z_1 and z_2 on the resolution. To some extent, this also applies to the case of pseudo-merohedral twinning. Existing twinning tests as well as the new one assume sufficiently small obliquity (such that the complete overlap of spots from different lattices occurs within the whole resolution range of the data set). Some correction for a larger obliquity could also be tried in the framework of Monte Carlo modelling.

It is necessary to note that the apparent simplicity of the Monte Carlo approach may mask mathematical problems. For example, it was pointed out by one of the referees that in the case of zero errors both $\{H_i, i = 1, \ldots\}$ and $\langle H\rangle$ are sufficient statistics relative to the parameter α. This is not so for the case with nonzero errors, but it is reasonable to assume that $\{H_i, i = 1, \ldots\}$ remain 'good' statistics in the sense that the loss of information about α is insignificant. The sophisticated mathematical analysis is outside the scope of this paper. Nevertheless, it is worthwhile emphasizing that the suggested approach is by no means a panacea and does not remove the need for analytical mathematical study. The method proposed may be assumed as a temporary remedy as a more convincing tool has not yet been developed.

APPENDIX *A*
Modelling of errors

In the error-sensitive model, we assumed that the input of the procedure is the list of records

$$\mathbf{h}_1, J^{\rm obs}(\mathbf{h}_1), \sigma^{\rm obs}(\mathbf{h}_1), \mathbf{h}_2, J^{\rm obs}(\mathbf{h}_2), \sigma^{\rm obs}(\mathbf{h}_2), \qquad (21)$$

where $\mathbf{h}_1$, $\mathbf{h}_2$ are twinned reflections, $J^{\rm obs}(\mathbf{h})$ are measured intensities and $\sigma^{\rm obs}(\mathbf{h})$ are estimated standard deviations for the measured intensities. In the simplest case, we can generate measurement errors in the H-simulation procedure as normally distributed random values with zero mean and standard deviations $\sigma^{\rm obs}(\mathbf{h})$. This procedure is quite straightforward, but it is too sensitive to the input values $\sigma^{\rm obs}(\mathbf{h})$. The latter depend on a number of experimental parameters that are difficult to control and may therefore not always be reliable. We briefly describe below a more sophisticated yet more robust procedure used in our tests that refers to some 'average accuracy' of the data set rather than to the accuracy of a particular measured intensity. In this approach, the standard deviation of the measurement error for a given reflection is modelled as a random variable based on an empirical distribution derived from the input list of reflections.

The counting statistics for diffracted beams may be assumed to follow the Poisson distribution, so that the standard deviation is equal to the square root of the number of measured quanta (Borek *et al.*, 2003). When considering a large number of registered quanta and calculating an integral intensity for a single reflection over many pixels, the corresponding probability distribution approaches the normal distribution, so that the error in a measured intensity $J^{\rm obs}$ may be considered as a normally distributed random value with zero mean and standard deviation

$$\sigma^{\rm theor}(\mathbf{h}) = a[J^{\rm theor}(\mathbf{h})]^{1/2}. \qquad (22)$$

Here, the coefficient a depends on the scale chosen for the intensities and might be estimated through the comparison of theoretical values (1.8) with the observed $\sigma^{\rm obs}(\mathbf{h})$ values obtained from multiple measured (or symmetric) reflections. In fact, the accuracy of the measured intensities depends on many additional factors. As an outcome, an attempt to link observed values by (22) results in the coefficient a depending on a particular reflection

$$\sigma^{\rm obs}(\mathbf{h}) = a(\mathbf{h})[J^{\rm obs}(\mathbf{h})]^{1/2}. \qquad (23)$$

It is convenient for our purposes to rewrite the last equation in the form

$$\sigma^{\rm obs}(\mathbf{h}) = \eta(\mathbf{h})[\langle I\rangle(h)]^{1/2}[J^{\rm obs}(\mathbf{h})]^{1/2} = \eta(\mathbf{h})\langle I\rangle[z^{\rm obs}(\mathbf{h})]^{1/2}, \quad (24)$$

where the mean $\langle I\rangle$ is the same as in (9). For every particular reflection the value $\eta(\mathbf{h})$ may be calculated through the input $J^{\rm obs}(\mathbf{h})$, $\sigma^{\rm obs}(\mathbf{h})$ values, provided that the mean values $\langle I\rangle(\mathbf{h})$ are assigned. The set of all calculated $\eta(\mathbf{h})$ values is then converted to an empirical probability distribution $P_\eta(\eta)$ of the coefficient η in the studied data set. When simulating distributions $P^{\rm theor}_{H,\alpha}$, the standard deviations of the measurement errors $\delta(\mathbf{h})$ are considered to be random variables,

$$\delta_1 = \eta_1\langle I\rangle[(1-\alpha)z_1 + \alpha z_2]^{1/2}, \quad \delta_1 = \eta_2\langle I\rangle[\alpha z_1 + (1-\alpha)z_2]^{1/2}, \qquad (25)$$

where η obeys the derived $P_\eta(\eta)$ distribution and z_1 and z_2 were generated with (6).

As a result, the statistical model for the H value may be described as follows. The value is calculated as

$$H = \left|\frac{(1-2\alpha)(z_1 - z_2) + \xi_1 - \xi_2}{z_1 + z_2 + \xi_1 + \xi_2}\right|,$$

where z_1 and z_2 are independent random variables distributed with the exponential distribution $\exp(-z)$, ξ_1 and ξ_2 are normally distributed random variables ($\xi = \delta/\langle I\rangle$) with zero mean and standard deviations

$$\sigma(\xi_1) = \eta_1[(1-\alpha)z_1 + \alpha z_2]^{1/2}, \sigma(\xi_2) = \eta_2[\alpha z_1 + (1-\alpha)z_2]^{1/2}$$

and η_1 and η_2 are independent random variables distributed with known $P_\eta(\eta)$ distribution.

The results obtained with the two procedures for modelling of measurement errors discussed were close enough in our tests, but the sophisticated procedure seems to be more stable.

The work was supported by DFG visiting grant 436 Rus 17/25/07 and RFBR grants 07-04-00137 and 05-01-22002. The authors thank K. Diederichs, P. Zwart and the referees of the paper for valuable discussions and comments.

References

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.
Afonine, P. V., Grosse-Kunstleve, R. W. & Adams, P. D. (2005). *CCP4 Newsl.* **42**, contribution 8.
Borek, D., Minor, W. & Otwinowski, Z. (2003). *Acta Cryst.* D**59**, 2031–2038.
Britton, D. (1972). *Acta Cryst.* A**28**, 296–297.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dauter, Z. (2003). *Acta Cryst.* D**59**, 2004–2016.
Fisher, R. G. & Sweet, R. M. (1980). *Acta Cryst.* A**36**, 755–760.
Gomis-Rüth, F. X., Fita, I., Kiefersauer, R., Huber, R., Avilés, F. X. & Navaza, J. (1995). *Acta Cryst.* D**51**, 819–823.
Hall, S. R., du Boulay, D. J. & Olthof-Hazekamp, R. (2000). Editors. *Xtal3.7 System.* Perth: University of Western Australia.
Herbst-Irmer, R. & Sheldrick, G. M. (1998). *Acta Cryst.* B**54**, 443–449.
Kabsch, W. (1993). *J. Appl. Cryst.* **26**, 795–800.
Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2006). *Acta Cryst.* D**62**, 83–95.
Lunin, V. Y., Lunina, N. L., Petrova, T. E., Urzhumtsev, A. G. & Podjarny, A. D. (1998). *Acta Cryst.* D**54**, 726–734.
Murray-Rust, P. (1973). *Acta Cryst.* B**29**, 2559–2566.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Padilla, J. E. & Yeates, T. O. (2003). *Acta Cryst.* D**59**, 1124–1130.
Parsons, S. (2003). *Acta Cryst.* D**59**, 1995–2003.
Petrova, T. E., Lunin, V. Y. & Podjarny, A. D. (2000). *Acta Cryst.* D**56**, 1245–1252.
Redinbo, M. R. & Yeates, T. O. (1993). *Acta Cryst.* D**49**, 375–380.
Rees, D. C. (1980). *Acta Cryst.* A**36**, 578–581.
Ritter, S., Diederichs, K., Frey, I., Berg, A., Keul, J. & Baumstark, M. W. (1999). *J. Cryst. Growth*, **196**, 344–349.
Rudolph, M. G., Kelker, M. S., Schneider, T. R., Yeates, T. O., Oseroff, V., Heidary, D. K., Jennings, P. A. & Wilson, I. A. (2003). *Acta Cryst.* D**59**, 290–298.
Schneider, T. R., Kärcher, J., Pohl, E., Lubini, P. & Sheldrick, G. M. (2000). *Acta Cryst.* D**56**, 705–713.
Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
Yang, F., Dauter, Z. & Wlodawer, A. (2000). *Acta Cryst.* D**56**, 959–964.
Yeates, T. O. (1988). *Acta Cryst.* A**44**, 142–144.
Yeates, T. O. (1997). *Methods Enzymol.* **276**, 344–358.
Yeates, T. O. & Fam, B. C. (1999). *Structure*, **7**, R25–R29.
Zwart, P. H. (2005). *Acta Cryst.* D**61**, 1437–1448.