Journal of
Applied Crystallography

ISSN 0021-8898

Received 22 November 2009
Accepted 19 May 2010

Using the *t*-distribution to improve the absolute structure assignment with likelihood calculations

Rob W. W. Hooft,[a]* Leo H. Straver[b] and Anthony L. Spek[c]

[a]Maasdijk 93, 2691 PC's-Gravenzande, The Netherlands, [b]Chopinrode 8, 2717 BK Zoetermeer, The Netherlands, and [c]Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. Correspondence e-mail: rob@hooft.net

The previously described method for absolute structure determination [Hooft, Straver & Spek (2008). *J. Appl. Cryst.* **41**, 96–103] assumes a Gaussian error distribution. The method is now extended to make it robust against poor data with large systematic errors with the introduction of the Student *t*-distribution. It is shown that this modification makes very little difference for good data but dramatically improves results for data with a non-Gaussian error distribution.

1. Introduction

In our paper on the determination of the absolute structure using likelihood analysis on Bijvoet differences (Hooft *et al.*, 2008), we introduced a method that is sensitive to very small Bijvoet differences. The implicit assumption was a carefully collected data set. After publication of this method we have received a number of example data sets that appeared to misbehave in the analysis.

A careful analysis of these data sets showed that the deviations were due to (1) individual observations that deviated very far from the expected values and (2) systematic deviations of the estimated uncertainties in the reflection intensities. Obviously, data sets do not always follow a Gaussian error distribution, and the assumption of underlying Gaussian statistics in our analysis can bias the results. We therefore set out to extend our method such that it would become less sensitive to these effects.

A possible approach is to recognize the obvious outliers and either downweight them or eliminate them altogether from the data analysis. However, such a procedure would lead potentially to an arbitrary cut-off and it only partially addresses the non-Gaussian behaviour of the data. The method we propose is to describe the data with a more appropriate error model. This robust method renders rejection of outliers superfluous and properly describes non-Gaussian behaviour not only of the tails of the distribution but also of the bulk of the data.

2. Theoretical background

The following sections will explain how incorrectly determined standard uncertainties and outliers influence the determination of absolute structure by statistical analysis of Bijvoet differences.

2.1. Standard uncertainties

The basis of our statistical analysis is the equation that describes the difference between measured and observed Bijvoet differences for reflections h and $-h$:

$$z_h = \frac{\Delta F_c^2(h) - \Delta F_o^2(h)}{\sigma_{\Delta F_o^2(h)}}. \quad (1)$$

In this equation, the standard uncertainty in the observation is in the denominator. It will be obvious that the calculation relies not only on accurate intensity measurements but also on accurate estimates of their accuracy. The 'error model' that is used to determine the standard uncertainty is therefore very important.

The various software packages that are used for data reduction mostly apply the same technique to obtain the variance of a reflection intensity: they add up a term accounting for Poisson statistics and a term accounting for inaccuracies in the measurement. The last term is proportional to the square of the measured intensity; the proportionality constant is sometimes named the 'instability factor' (we will abbreviate this as g) (McCandlish *et al.*, 1975).

For a measurement with a point-detector system in which each reflection is measured once, g can be determined from a few reflections that are measured repeatedly during the whole experiment (so-called 'intensity control reflections'). This value can be used to calculate standard uncertainties for all measured reflections.

When a system with an area detector is used to collect the data, many reflection intensities are obtained multiple times through independent observations. In such a case the calculated standard uncertainty values – the *internal* standard uncertainty – can be verified from the multiplicity in the measurements: equivalent reflections should be consistent. The variance of different observations can be used to calculate an *external* standard uncertainty. Advanced data reduction software such as *SADABS* (Sheldrick, 1996) will use the multiplicity to calculate different values of g for different segments of the data, and apply these different values appropriately to obtain the best possible correspondence between all calculated internal and external standard uncertainties.

We can conclude that there is an accepted practice for data reduction that will result in reasonable estimated uncertainties

in the reflection intensities and rejection of obvious outliers. However, when Bijvoet differences are calculated after structure refinement, there is no guarantee that this has been done correctly. Systematic deviations in the standard uncertainties are therefore possible.

2.2. Outlier statistics

To illustrate the impact of strongly deviating observations, it is instructive to check what can happen with a single outlier observation when applying our absolute structure determination technique. Let us consider a Bijvoet difference measured as 250 ± 50 counts s^{-1}, and an inverted structure model m_1 and a correct structure model m_2, predicting differences of 250 and 300 counts s^{-1}, respectively. The likelihoods as calculated in our procedure are

$$z_1 = (300 - 250)/50 = 1.0, \quad p(m_1) \simeq 0.24, \tag{2}$$

$$z_2 = (250 - 250)/50 = 0.0, \quad p(m_2) \simeq 0.40, \tag{3}$$

$$p(m_2)/p(m_1) \simeq 1.6. \tag{4}$$

Based on this reflection alone, model m_2 is 1.6 times more likely than model m_1.

We will now study how moderate deviations and experimental errors affect the statistical analysis of this same example.

Imagine that this Bijvoet difference was measured as 150 ± 50 counts s^{-1} because of unfortunate fluctuations. The likelihood calculation now proceeds as

$$z_1 = (300 - 150)/50 = 3.0, \quad p(m_1) \simeq 4.4 \times 10^{-3}, \tag{5}$$

$$z_2 = (250 - 150)/50 = 2.0, \quad p(m_2) \simeq 5.4 \times 10^{-2}, \tag{6}$$

$$p(m_2)/p(m_1) \simeq 12. \tag{7}$$

This deviating observation shows a strong preference for model m_2. Deviations of a few times the standard uncertainty like this occur several times in a normal data set. Positive and negative deviations are equally likely and will statistically balance each other: this measurement of 150 ± 50 is equally likely for this data point as a measurement of 350 ± 50 which would lead to $p(m_1)/p(m_2) \simeq 4.5$. The effects of these fluctuations cancel each other in the composite probability; in this example a combination of the two observations leads to a preference of $p(m_2)/p(m_1) = 12/4.5 \simeq 2.7 \simeq 1.6^2$ in perfect correspondence to two ideal measurements (see above). This fact follows trivially from the mathematical form of the Gaussian distribution. In summary, the ensemble of all statistically deviating measurements forms an essential contribution to the absolute structure determination; this holds even when the individual model intensities differ by less than the standard uncertainty in the measured Bijvoet differences.

Now imagine that this same Bijvoet difference was measured as 800 ± 50 counts s^{-1} because of an experimental error. The likelihood calculation proceeds as

$$z_1 = (300 - 800)/50 = -10.0, \quad p(m_1) \simeq 7.7 \times 10^{-23}, \tag{8}$$

$$z_2 = (250 - 800)/50 = -11.0, \quad p(m_2) \simeq 2.1 \times 10^{-27}, \tag{9}$$

$$p(m_1)/p(m_2) \simeq 4 \times 10^4. \tag{10}$$

The erroneous observation leads to a huge preference for model m_1. This preference is incorrect in practice: deviations of this magnitude caused by statistical variations only are very unlikely; therefore these data points more likely indicate experimental problems. Once an experimental error occurs, the exact magnitude of the error is a pure chance value and the difference in likelihood between the two models should not contribute to the absolute structure assignment. Taking these data points that are due to experimental errors into account 'as-is' significantly impacts on the reliability of the resulting joint probabilities.

To make a reliable analysis without rejection of data points, a more robust error model is needed that will be near-Gaussian in behaviour for the ideal and statistically fluctuating measurements but will give a dramatic reduction of the strong influences of experimental errors.

3. Method

In his 1908 paper, William Sealy Gosset (Student, 1908) describes the derivation of a family of distributions to replace the normal (Gaussian) sampling distribution when the standard deviation of the underlying parent population is unknown and must be approximated by the square root of the variance of the measured data. The exact shape of each distribution depends on a variable $\nu > 0$ which expresses the number of degrees of freedom in the data set. The limiting case $\nu = \infty$ is identical to the normal distribution. Long after Student published his paper, this same family of distributions has been used in robust data analysis (*e.g.* Lange *et al.*, 1989).

We propose to replace the Gaussian distribution used by Hooft *et al.* (2008) by a Student *t*-distribution. This requires two steps: (1) a suitable value of ν must be determined and (2) the Gaussian distribution in the likelihood calculations must be replaced by the *t*-distribution.

3.1. Determination of a suitable value of ν

Hooft *et al.* (2008) briefly mentioned normal probability plots as a way to assess the reliability of the error model. The probability plot is made by determining G, and plotting the deviations between the observed and calculated Bijvoet differences at $\gamma = G$ in a normal probability plot. Our 2008 paper mentioned that the error model is good if the correlation coefficient of the normal probability plot is at least 0.999, but it did not describe a way of dealing with the case where the correlation is lower.

Table 1
Statistical analysis of ideal and deviating measurements using a Gaussian distribution and a *t*-distribution with different numbers of degrees of freedom ν.

Given is the likelihood ratio of each measurement relative to models that predict values of 250 and 300.

	p ratio			
Measurement	Gaussian	$\nu = 100$	$\nu = 10$	$\nu = 3$
250 ± 50	1.65	1.65	1.69	1.78
150 ± 50	12.1	10.7	5.36	2.94
350 ± 50	4.48	4.38	3.77	3.06
800 ± 50	36300	154	2.61	1.45

Hooft *et al.* (2009) described how to make probability plots based on a *t*-distribution. We will use this technique here. We calculate the observed deviations as $x(\gamma)$ from the Bijvoet differences following equation (17) of our 2008 paper. As γ we use the best estimate available: initially this is the optimized value G calculated using Gaussian statistics; if the final result of the analysis using the *t*-distribution deviates significantly from the Gaussian estimate, we can redo the calculation using that new value.

Since all Bijvoet differences can be calculated either as $F_h - F_{-h}$ or equivalently as $F_{-h} - F_h$, the data in the probability plot must go through the origin. The easiest way to enforce this also for the fit is to use all pairs of data points in the calculation. We have noticed that without explicit duplication of all differences the intercept of the fit unexpectedly deviates significantly from zero; we have not investigated where this deviation comes from.

Using the method described by Hooft *et al.* (2009) the value of ν is calculated that optimizes the linearity of the probability plot. We postulate that the *t*-distribution with the determined value of ν and the associated slope (a) of the fit line can be used to establish an error model for the data if the linear correlation that is obtained in the optimization process is sufficient (*e.g.* 0.999).

3.2. Using the *t*-distribution in the absolute structure determination

Once a proper value of ν has been calculated for the data, the remaining calculations must be performed as in our 2008 paper, replacing the Gaussian probability calculations with the *t*-distribution.

Instead of equation (5) from that paper, we will use

$$p(z, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{(\nu\pi)^{1/2}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{z^2}{\nu}\right)^{-(\nu+1)/2}. \qquad (11)$$

Similarly, equations (12) and (18) of that paper should be adapted to include a *t*-distribution with the determined value of ν.

The slope (a) of the probability plot that is obtained from the fit should also be used. If this slope is < 1, as is often observed, this means that the standard uncertainties in the Bijvoet differences are over-estimated. To account for this, the estimated standard deviations in the Bijvoet differences should be scaled by a. This can be done in the denominator of equation (17):

$$x(\gamma) = \frac{\gamma\Delta F_c^2 - \Delta F_o^2}{a\sigma_{\Delta F_o^2}}. \qquad (12)$$

This modification has practically no effect on the calculated value of G; it only leads to a proportional scaling of σ_G.

Performing the analysis with these modifications will result in new values for G and σ_G. If the value for G deviates significantly from the original estimate using Gaussian statistics, the new value should be used to make a new probability plot analysis. This whole procedure should be repeated until the results no longer change. In practice, for our relatively weak anomalous signals we have seen only very small changes after the first iteration; for large anomalous signals no significant change is expected at all.

4. Results and discussion

Table 1 shows how the outlier statistics as described in §2.2 change when *t*-distributions with different values for ν are applied. It can be concluded that with the replaced error model normal statistical fluctuations still lead to a similar probability ratio as using a Gaussian distribution. The new model satisfies our target of suppressing the erroneous bias caused by experimental errors. At very low ν even the relevance of deviations of three times the standard uncertainty is scaled down as outliers.

For practical tests of the method we used a data set we received for a monoclinic structure containing a single sulfur atom in the molecule, $Z' = 2$, space group $P2_1$, 7517 reflections, 2824 Bijvoet pairs. The calculation using Gaussian statistics gave $y = 0.43$ with $\sigma_y = 0.15$, seemingly indicating a racemic twin. However, the normal probability plot for the

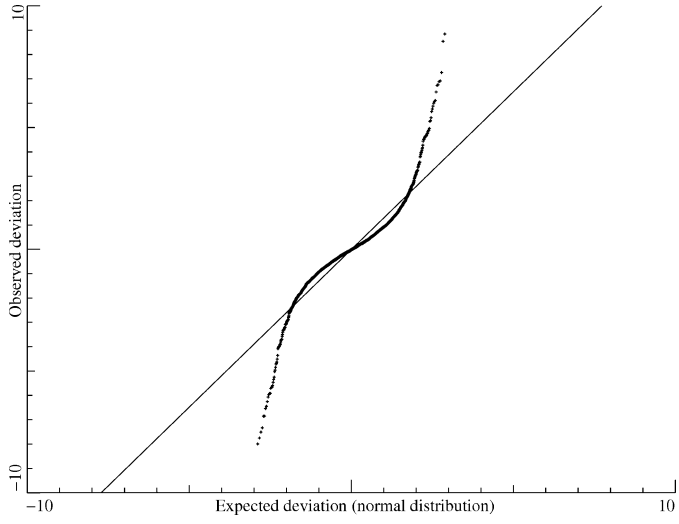

Figure 1
Normal probability plot of Bijvoet differences of a small-molecule crystal structure data set, showing curves due to non-normal behaviour of the errors. The diagonal straight line represents a least-squares fit: its slope is 1.32, the correlation coefficient 0.93.

Bijvoet differences showed a correlation coefficient of 0.93 and a slope of 1.32. Both a correlation coefficient much lower than 0.999 and a slope larger than 1.0 are clear indications of problems with the standard uncertainties in the data set. The normal probability plot for this data shows an inverted 'S' curve (Fig. 1).

Optimization of a *t*-probability plot resulted in $\nu = 2.2$ and a slope of $a = 0.76$ at a correlation coefficient of 0.998 (Fig. 2). The results of the absolute structure determination are $y = 0.10$ with $\sigma_y = 0.18$; no racemic twin is necessary to explain this value.

To verify on a somewhat larger scale how the augmented procedure behaves, a total of 11 data sets from diverse instruments and produced using several different data reduction software packages were analysed using a Gaussian distribution and a Student *t*-distribution. Three gave low correlation coefficients on the normal probability plot; these are labelled 'Test' and numbered 1 through 3. The other data sets are labelled 'Control' and numbered 1 through 8. The result is given in Table 2. As the result of the analysis, this table only gives y and σ_y; analogous observations can be made for *P*2 and *P*3 values, but these would not add to the discussion. All correlation coefficients, of both the test and the control data sets, improved significantly by using a *t*-distribution. The clearest difference between the two groups is the optimized value of ν: the highest value for the test data sets is below 6, whereas the lowest value for a control data set was above 12. Furthermore, for the control group none of the values of y changed; the only changes are in the standard deviation. For two out of the three data sets in the test group, y changed significantly.

From these results we can see that the *t*-distribution is quite suitable to describe the kinds of aberrations that can be observed in the three different test data sets. Apparently, the curves that are visible in the normal probability plot for these three can be very well described with the single parameter ν from the Student *t*-distribution. From a statistical point of view, this is also sufficient: the only condition for the distribution used as error model is that it can properly describe the probabilities for each deviation.

This solution to the robustness problem is quite appealing. There is no need to choose any parameters for the algorithm, only one parameter that is optimized. There are no arbitrary cut-offs. All data points are taken into account with appropriate weighting.

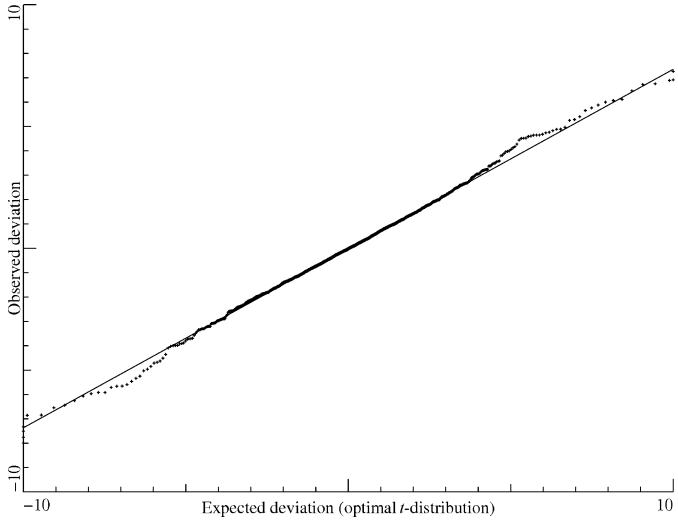

Figure 2
t-Probability plot on the same data represented in Fig. 1. A *t*-distribution with $\nu = 2.2$ was found to be optimal to model this data set. The least-squares line shows a slope of 0.76. Some points with expected deviations of larger than 10σ have been left out of the plot.

Table 2
Table of 11 data sets handled with both a Gaussian distribution and a Student *t*-distribution.

r_G is the correlation coefficient of the normal probability plot, ν is the optimized number of degrees of freedom for a *t*-distribution, and r_t and a are the correlation coefficient and the slope of the *t*-distribution probability plot, respectively. y is the absolute structure parameter.

Structure	y (Gaussian)	r_G	ν	r_t	a	y (t)
Test 1	0.0010 (16)	0.9951	5.63	0.9996	1.15	0.001 (2)
Test 2	0.43 (15)	0.9262	2.35	0.9982	0.76	0.10 (18)
Test 3	0.73 (3)	0.9478	2.67	0.9914	1.46	−0.09 (7)
Control 1	−0.04 (16)	0.9991	12.5	0.9998	0.76	−0.05 (13)
Control 2	0.0037 (19)	0.9992	25	0.9996	1.29	0.0041 (26)
Control 3	0.2 (2)	0.9998	33	0.9999	0.81	0.19 (17)
Control 4	0.06 (17)	0.9999	201	0.9999	0.95	0.05 (16)
Control 5	−0.10 (8)	0.9996	21	0.9999	0.92	−0.10 (7)
Control 6	0.28 (11)	0.9993	19	0.9995	0.89	0.29 (10)
Control 7	−0.1 (2)	0.9992	15	0.9996	0.89	−0.0 (2)
Control 8	0.022 (5)	0.9996	21	0.9999	0.62	0.022 (3)

5. Conclusions

We have seen cases of poor data quality where a Gaussian data analysis showed problems in the normal probability plot (NPP). Our original 2008 procedure could not cover such cases. With the new approach cases where the NPP is not linear can now be handled as well. For cases that work well with a Gaussian analysis, the application of a *t*-distribution gives very similar results. Therefore, the new analysis can be applied for all cases. For those test data sets that show a significant change in the value of y, the standard deviation of y increases slightly. The advantage is that, judging by the improved fit of the probability plot, the resulting $y(t)$ has a more realistic value.

References

Hooft, R. W. W., Straver, L. H. & Spek, A. L. (2008). *J. Appl. Cryst.* **41**, 96–103.
Hooft, R. W. W., Straver, L. H. & Spek, A. L. (2009). *Acta Cryst.* A**65**, 319–321.
Lange, K. L., Little, R. J. A. & Taylor, J. M. G. (1989). *J. Am. Stat. Assoc.* **84**, 881–896.
McCandlish, L. E., Stout, G. H. & Andrews, L. C. (1975). *Acta Cryst.* A**31**, 245–249.
Sheldrick, G. M. (1996). *SADABS*. University of Göttingen, Germany.
Student (1908). *Biometrika*, **6**, 1–25.