Acta Cryst. (1978). A**34**, 517–525

On the Treatment of Negative Intensity Observations

By Simon French* and Keith Wilson

Laboratory of Molecular Biophysics, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, England

(*Received* 21 *October* 1977; *accepted* 23 *January* 1978)

A method is described which produces sensible estimates of structure factor moduli from intensity observations, whether the latter are positive or negative. Preliminary applications of the method to data from the protein phosphorylase b are summarized.

Introduction

Reflections with small structure factor moduli have always caused problems. Their true intensities are, of course, non-negative, but because of counting errors their measured intensities may not be. What should one do with negative intensity measurements? If they are left untouched, a problem results upon encountering the first square rooting. If they are omitted from the data set, bias will result in the determined structure. If they are set to zero, and no adjustment made to their standard deviations, bias will again result, although on a smaller scale (Hirshfeld & Rabinovich, 1973).

To avoid these biases, Hirshfeld & Rabinovich (1973) recommend that all intensity measurements should be included in a crystallographic refinement, whether their observed values are positive or negative. Furthermore they suggest that least-squares refinement should proceed by minimization of the discrepancies between the observed and calculated squares of the structure factor moduli, rather than the moduli themselves. This allows the negative observations to be included and so prevents any bias entering the calculations. We are in complete agreement with these conclusions.

Nevertheless, in many crystallographic studies it is essential to obtain estimates of the structure factor moduli and their associated error, whatever the value of the observed intensity. Estimates of the moduli are necessary, for example, in the calculation of Fourier or, more importantly, difference Fourier syntheses. How should the problems associated with negative intensity observations be tackled in such circumstances?

Fortunately the problems are almost entirely due to poor statistical methodology. Instead of thanking the data for the information that certain structure factor moduli are small, we accuse them of assuming 'impossible' negative values. What we should do is combine our knowledge of the non-negativity of the true intensities with the information concerning their magnitude contained in the data. Bayesian statistics provide the techniques for doing precisely this.

The Bayesian approach to statistical inference (see, for example, Box & Taio, 1973; Lindley, 1965) differs fundamentally from the more conventional, frequentist approach (see, for example, Hamilton, 1964) and the reader is strongly urged before continuing to acquaint himself with the Bayesian theory of which the following is a brief description.

Probability distributions are taken to represent degrees of belief rather than relative frequencies. Prior to performing an experiment designed to gain information on a certain parameter θ, we have a certain distribution of belief in the possible values of θ. Assume that this distribution has density function $p_\theta(.)$. Note that we use subscripts to indicate the quantity about which we are expressing our beliefs. Note also that we must always have some belief about θ, otherwise we could never design an informative experiment. The experiment gives rise to an observation X which we believe is related to the parameter through the sampling distribution with density function $p_X(.|\theta)$. Note that this distribution is conditional on the unknown parameter, θ. It describes simply how we would expect our observations to be distributed if we knew θ. If the observation $X = x$ is actually made, then Bayes's theorem tells us that our belief in θ is modified by the data to

$$p_\theta(\theta|x) \underset{\theta}{\propto} p_X(x|\theta)\,p_\theta(\theta). \qquad (0.0)$$

The $\underset{\theta}{\propto}$ means 'is proportional to as a function of θ'. The constant of proportionality may be determined by remembering that a probability density function integrates to unity.

1. Intensity measurements

At a given reflection the parameter that concerns us is the true intensity. It is, perhaps, arguable that the true

* Present address: Department of Decision Theory, University of Manchester, Manchester M13 9PL, England.

structure factor modulus is the parameter of interest, but, as will become apparent, this would lead to precisely the same results. We shall denote the true intensity by J, *i.e.* the above $\theta = J$ in this particular case. If the observed intensity is denoted by I, *i.e.* $X = I$, then Bayes's theorem (0.0) becomes

$$p_J(J|I) \underset{J}{\propto} p_I(I|J)\, p_J(J), \qquad (1.0)$$

where $p_J(.)$ is our prior belief in the true intensity, $p_I(.|J)$ is our belief in the relation between the observed and true intensity.

This is an appropriate point at which to explain our interpretation of an 'intensity observation'. We assume that all the relevant data sets, collected by either diffractometer or photographic methods, have been corrected for Lorentz, polarization, absorption, extinction and radiation-damage effects, have been reduced to a common scale and have been merged over equivalents. I is this 'merged intensity' containing all the available observational information at the given unique reflection. All the operations needed to produce this merged intensity are assumed to have been carried out on the raw intensity measurements, be they positive or negative.

Throughout we shall assume that $p_I(.|J)$ is a normal density, *viz*

$$I \sim N(J, \sigma^2). \qquad (1.1)$$

Note that we assume that the observation is unbiased on the true intensity J and has known variance σ^2. Further note that we take σ^2 to be particular to each individual reflection. Thus we have three assumptions – normality, unbiasedness and known variance – which require further discussion.

Firstly, whilst we accept the data are certainly not exactly normally distributed, we do contend that the normal distribution is an adequate approximation for our purposes. From a theoretical point of view we are encouraged in this belief, since the merged intensity I is made up of sums of differences of theoretically Poisson-distributed counts. Such operations on Poisson variables reduce them to normality quickly (Irwin, 1937; Skellam, 1946). Furthermore, it has been our observation that the actual distributions of I have not been noticeably skew, although they have been slightly sharper or flatter than the normal. We believe that such deviations have little effect on our main results.

Our second assumption of unbiasedness is more suspect. It is well known that some data reduction methods produce biased measurements on small intensities (Tickle, 1975; French, 1975). However, such bias is invariably introduced because the data reduction method 'forces' the observed intensities to be non-negative, *e.g.* ordinate analysis of diffractometer data (Watson, Shotton, Cox & Muirhead, 1970). Our methods are both inappropriate and unnecessary in such cases. For diffractometer data reduced by background–peak–background scanning or profile analysis (Diamond, 1969; French, 1975, 1978) or photographic data reduced by similar methods (Wilson & Yeates, 1978), unbiasedness is a fair assumption.

Finally, we have assumed σ^2 is known. This is, of course, completely untrue. We no more know the exact accuracy of our data than the true intensity. To deal with this problem in a Bayesian fashion, we should express all our prior beliefs about σ^2 in a probability distribution, update this in the light of the data and then expect the nuisance parameter σ^2 from our posterior distribution for J (see, for example, Box & Taio, 1973, pp. 70–72). However, it is adequate for our method to use a good estimate of σ^2 in the distribution (1.1). We emphasize that this estimate must be as good as possible. Not only must σ^2 allow for variations due to counting statistics but also for errors arising from instrument instability and poor correction factors. Procedures for producing such estimates are discussed in Dodson (1976), and McCandlish, Stout & Andrews (1975).

Thus, all we have to do is define our prior density $p_J(.)$ and we may use (1.0) to obtain our posterior density, $p_J(.|I)$. We shall discuss three possibilities for $p_J(.)$ shortly, but defer doing so in order to indicate how one should proceed upon obtaining $p_J(.|I)$.

Most, if not all, crystallographic structure-solution techniques do not use the posterior density for the intensity in its entirety, but use approximations based upon the mean and variance of it or its square root, *i.e.* the structure factor modulus. Least-squares refinement is an extremely common example of such a technique. So usually we do not require the density $p_J(.|I)$, but the posterior means and variances:

$$E_J(J|I) = \int_0^\infty J p_J(J|I)\, \mathrm{d}J, \qquad (1.2)$$

$$\mathrm{var}_J(J|I) = \int_0^\infty [J - E_J(J|I)]^2 p_J(J|I)\, \mathrm{d}J, \qquad (1.3)$$

or, letting $F = \sqrt{J}$,

$$E_J(F|I) = \int_0^\infty F pJ(J|I)\, \mathrm{d}J, \qquad (1.4)$$

$$\mathrm{var}_J(F|I) = \int_0^\infty [F - E_J(F|I)]^2 p_J(J|I)\, \mathrm{d}J. \qquad (1.5)$$

Note that (1.4) and (1.5) provide the solution to the problem mentioned above where the structure factor modulus is the parameter of interest.

It should be noted that equations (1.0), (1.2), (1.3), (1.4) and (1.5) apply equally well to positive and negative intensity measurements. Our procedure has no arbitrary cut-off points in it. All the data are treated in the same fashion.

In the next section, we discuss three possibilities for our prior density $p_J(.)$. Appendix A indicates how the mean and variances (1.2), (1.3), (1.4) and (1.5) may be calculated on a computer for the three different choices of $p_J(.)$.

2. Choice of prior distribution

What do we know about the true intensity *a priori*? First and foremost, we know it is positive. A distribution that would express precisely this condition is

$$p_J(J)\begin{cases} = 1 & \text{if } J \geq 0 \\ = 0 & \text{if } J < 0. \end{cases} \qquad (2.0)$$

Admittedly this density has the embarrassing property of integrating to infinity rather than unity, but such an 'improper' prior distribution is permissible under certain circumstances, which, fortunately, apply here. (2.0) may be thought of as an approximation to the proper prior distribution given by the density

$$p_J(J)\begin{cases} = 1/k & \text{if } 0 \leq J \leq k \\ = 0 & \text{if } J < 0 \text{ or } J > k, \end{cases} \qquad (2.1)$$

i.e. to the distribution which demands that the intensity be a non-negative number not greater than some upper limit k. The fact that $1/k \neq 1$ needs no concern, since this difference would be absorbed into the constant of proportionality in (1.0). As $k \to \infty$, we obtain (2.0). Of course, we could use (2.1) in its own right and set k to some unreachable intensity, *e.g.* that of the main beam. If we did so, the results we obtained for such k would be indistinguishable from those resulting from (2.0).

Usually, arguably always, we have considerably more knowledge of the intensity than expressed by (2.0). We know that, taken as a whole, any moderate or large data set obeys Wilson's (1949) statistics. So, for an acentric distribution of reflections:

$$p_J(J)\begin{cases} = (\Sigma)^{-1}\exp(-J/\Sigma) & \text{if } J \geq 0 \\ = 0 & \text{if } J < 0, \end{cases} \qquad (2.2)$$

while for a centric distribution of reflections:

$$p_J(J)\begin{cases} = (2\pi\Sigma J)^{-1/2}\exp(-J/2\Sigma) & \text{if } J \geq 0 \\ = 0 & \text{if } J < 0. \end{cases} \qquad (2.3)$$

In both cases:

Σ is the mean intensity in the appropriate shell of reciprocal space. (2.4)

Now, of course, Σ is unknown and, as for σ^2, we should express our beliefs about Σ in a probability distribution *a priori*, analyse all the data to learn about Σ, and then expect the nuisance parameter Σ from the posterior distribution for J. Again, however, since the data sets that we usually collect contain a fair number of reflections, estimating Σ in shells of reciprocal space and then using these estimates in (2.2) and (2.3) does not cause noticeable bias. This is precisely what we do: estimate Σ in shells of reciprocal space. In doing so we include all the data, positive and negative. The appropriate value of Σ then serves in (2.2) or (2.3) to calculate the values (1.2), (1.3), (1.4) and (1.5) needed in the later stages of the structure determination.

It would, of course, be possible to use prior densities specific to particular space groups, but we have not done this.

3. Illustration of the effect of the algorithm

In Fig. 1 we illustrate the application of the algorithm. The three curves in (1.0) are shown: the prior, $p_J(J)$, is dotted; the likelihood, $p_I(I|J)$, is dashed; and the posterior, $p_J(J|I)$, is continuous. The observed intensity and the posterior mean intensity are indicated.

For the most part the effect of the algorithm depends on the ratio of I to σ (*i.e.* on the significance of the observation) with secondary dependence on the ratio of σ^2 to Σ (see Appendix *A*). In order to illustrate the main effect we have applied the algorithm to a sequence of hypothetical intensity observations ranging from -3 to 50 standard deviations, *i.e.* σ^2 has been set to unity. For this illustration Σ was chosen to be 20·0. We should point out that this has the implication that we are considering observations on a sequence of different crystals. Table 1 presents our results.

The effect of the procedure is maximal on the smallest observations and is negligible for an observation greater than about three standard deviations in the acentric case and about six standard deviations in the centric. For the lower values of the observed intensity, we note that the posterior mean is lower in the centric than in the acentric case, whilst for the higher values it is slightly higher. This reflects the relative shapes of Wilson's distributions. We also note that an observation of minus three standard deviations leads to a posterior standard deviation lower than that resulting from an observation of zero.This confirms our intuition

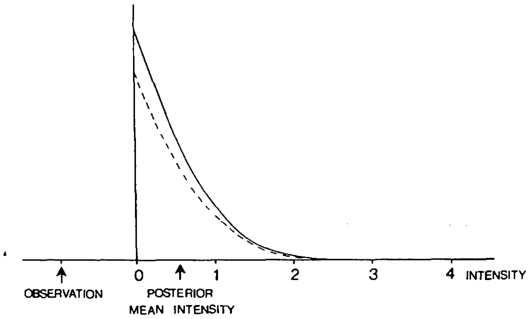

Fig. 1. Probability density functions for a hypothetical acentric reflection with $\Sigma = 20{\cdot}0$, $I = -1{\cdot}0$ and $\sigma = 1{\cdot}0$. Dotted line: prior density $p_J(.)$, dashed line: likelihood, $p_I(I|J)$, solid line: posterior density, $p_J(J|I)$. Posterior mean intensity = 0·515, posterior standard deviation = 0·440. *N.b.* The curves have not been drawn on the same vertical scale.

Table 1. *Posterior means and standard deviations for both the intensities and the structure factor moduli derived from a sequence of hypothetical observations with unit standard deviation*

The prior distributions are Wilson's for both the centric and acentric cases. Thoughout $\Sigma = 20{\cdot}0$.

Observation		Posterior moments		
I	$E(J\|I)$	$\sigma(J\|I)$	$E(F\|I)$	$\sigma(F\|I)$
(*a*) Acentric series				
−3·0	0·280	0·264	0·472	0·238
−2·0	0·368	0·335	0·545	0·268
−1·0	0·515	0·440	0·650	0·305
0·0	0·780	0·595	0·812	0·347
1·0	1·257	0·786	1·056	0·376
2·0	2·011	0·938	1·372	0·360
3·0	2·955	0·995	1·691	0·307
4·0	3·950	1·000	1·987	0·252
5·0	4·950	1·000	2·225	0·225
6·0	5·950	1·000	2·439	0·205
10·0	9·950	1·000	3·154	0·159
20·0	19·950	1·000	4·467	0·112
50·0	49·950	1·000	7·068	0·071
(*b*) Centric series				
−3·0	0·144	0·196	0·304	0·226
−2·0	0·194	0·255	0·355	0·260
−1·0	0·284	0·352	0·435	0·308
0·0	0·469	0·516	0·574	0·373
1·0	0·876	0·766	0·824	0·444
2·0	1·678	1·000	1·216	0·447
3·0	2·757	1·052	1·623	0·352
4·0	3·910	1·028	1·938	0·274
5·0	4·868	1·020	2·194	0·233
6·0	5·888	1·015	2·417	0·210
10·0	9·924	1·006	3·146	0·160
20·0	19·950	1·001	4·465	0·112
50·0	49·965	1·000	7·068	0·071

that negative observations actually carry more information than small positive ones.

The behaviour of the posterior standard deviation for the true intensity in the centric case requires further comment. As can be seen, for observed intensities of moderate size the posterior standard deviation takes values greater than unity, *i.e.* greater than the standard deviation of the observation. At first sight this is surely wrong. However, consider the situation more carefully. The density of Wilson's distribution for a centric reflection, see (2.3), has an infinite spike at the origin. It is emphatic that we should expect small observations. The increase in the posterior standard deviation over unity reflects the conflict between our prior expectation of a small intensity and our observation of a moderate or large one. Our first intuition that the observation standard deviation is an upper bound for the posterior standard deviation of the intensity is fallacious. It derives from the common misunderstanding in frequentist statistics that the distribution of an estimate of a parameter is identically the same as the distribution of the true parameter. In fact, the second distribution is an ill-defined concept in frequentist statistics (Barnett, 1973; French, 1977).

We have checked our algorithms (see Appendix *A*) for calculating the posterior moments against full-scale numerical integration. The errors induced by our approximations are at most 3% and in the vast majority of cases much less than 1%. The errors are, in fact, confined to the region in which we linearly interpolate values from a table generated by full-scale numerical integration and also the region of smaller intensity observation in which we first use our series approximation. For the higher values of the observed intensities the errors are negligible, as might be expected since the error is at most of order $1/I^3$

4. Implementation

The algorithms which evaluate the posterior means and variances are described in Appendix *A*, and have been programmed in Fortran.* In this section, we describe the overall organization of the program in which these routines are implemented. The four principal steps are as follows.

(1) The maximum and minimum values of $4(\sin^2\theta)/\lambda^2$ are found and the data divided into a suitable number of ranges of this parameter. The maximum number of such ranges is defined to be 50. If the total number of unique reflections is small, the number of ranges is restricted so that there is a minimum of 20 reflections in each. This ensures a reasonable estimate of Σ in each range.

(2) The mean intensity Σ is computed within each range and a table set up to allow linear interpolation of Σ for any particular reflection in the data.

(3) The observed value of each intensity (possibly negative) is read, together with its standard deviation. The value of Σ appropriate to this reflection is linearly interpolated from the table calculated in step 2. This value of Σ is multiplied by the multiplicity of the reflection (Iwasaki & Izo, 1977). The value of Σ thus obtained is our estimate of the mean intensity parameter of either of Wilson's distributions (2.2) or (2.3). Using this and the observed intensity and standard deviation, the posterior mean and standard deviation of the structure factor moduli are calculated by our Fortran routines. All observations are treated in the same manner, be they positive or negative. There is no discontinuity in the manner in which reflections are considered.

* The routines have been deposited with the British Library Lending Division as Supplementary Publication No. SUP 33352 (8 pp.). Copies may be obtained through The Executive Secretary, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

(4) In the same loop as step 3, the cumulative distribution of the ratio of observed to mean intensity is calculated in each of the ranges of $4(\sin^2\theta)/\lambda^2$. These distributions are calculated from the 'raw' intensity observations before they have been processed by our routines. They allow the validation of our assumption of Wilson's distributions as our prior beliefs.

5. On the validity of Wilson's distributions

A question which must be answered before the method can be applied is whether the distributions described by Wilson (1949) are valid prior beliefs. This question is especially important in the study of protein structures, where the atoms are certainly not randomly distributed throughout the crystal, as is assumed in deriving the distributions. This is evident from Fig. 2, where the fall-off in intensity with scattering angle for data from phosphorylase b shows a characteristic minimum (for a protein) at about 6 Å resolution and a maximum at 4·5 Å.

However, the behaviour of the ratio of the observed intensity (I) to the mean intensity (Σ) does conform to that predicted by Wilson. In Fig. 3 we show the experimental cumulative distribution of I/Σ for data from phosphorylase b extending to a resolution of 3 Å. The agreement with the theoretical distributions derived by Wilson is apparent for both the centric and acentric terms. The value of Σ was calculated for each reflection by the procedure described in the previous section.

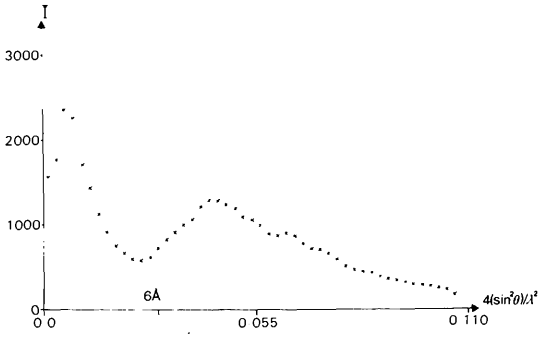

Fig. 2. Fall-off in mean observed intensity (I) with $4(\sin^2\theta)/\lambda^2$ for the native phosphorylase b data described in Table 2.

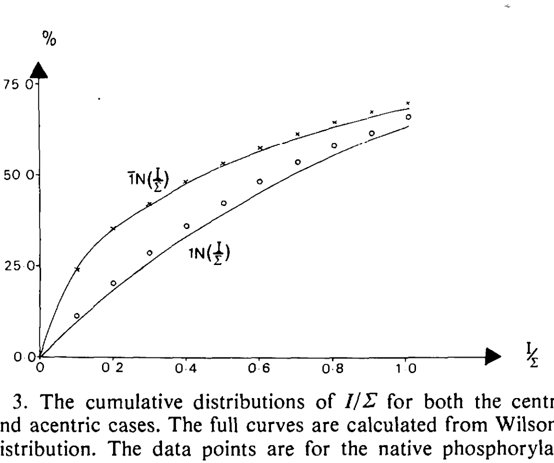

Fig. 3. The cumulative distributions of I/Σ for both the centric and acentric cases. The full curves are calculated from Wilson's distribution. The data points are for the native phosphorylase data described in Table 2 (before the intensities had been processed through the Bayesian estimation routines).

This agreement is not surprising from a theoretical point of view. The derivation of Wilson's distributions is based on the application of a central limit theorem, which assumes the atoms are uniformly and independently distributed about the unit cell. If this assumption breaks down, alternative central limit theorems may be applied (Dienanda, 1953, 1954, 1955). These show that the forms of Wilson's distributions remain, but the theoretical value of Σ is not that predicted in his 1949 paper (French, 1975).

The agreement between the theoretical and experimental curves shown in Fig. 3 and the remarks above provide confidence in the use of Wilson's distributions as valid prior beliefs for these phosphorylase data. This result concurs with those reported for crystals of small molecules by Howells, Phillips & Rogers (1950).

It is nevertheless important to validate this agreement for each set of data independently, as the presence of atoms in special positions or the existence of non-crystallographic elements of symmetry (or pseudo-symmetry) may abrogate the application of these prior beliefs for some crystal structures.

6. The standard deviation of an observation

As stated in § 1, it is necessary to have a meaningful estimate for the variance of each intensity observation. For data measured with a diffractometer, this can be obtained from counting statistics, empirically adjusted to allow for effects, *inter alia*, of instrument instability (McCandlish, Stout & Andrews, 1975; Dodson, 1976). The results described in this article are for data measured with an Arndt–Wonnacott oscillation camera (Arndt, Champness, Phizackerly & Wonnacott, 1973). An empirical approach has been used to estimate the standard deviation of the observations.

First, within a photograph the variance for each one of a set of equivalent terms is estimated from the expression:

$$\sigma^2 = \frac{\sum\limits_{i=1}^{N} (I_i - \bar{I})}{(N-1)}, \qquad (6.0)$$

where I_i is the intensity observation at each of N equivalent reflections, $\bar{I}$ is the unweighted mean intensity for these N observations. The standard deviation calculated from this variance for *all* sets of equivalent terms was then fitted to an expression of the form

$$\sigma = A + B\bar{I}, \qquad (6.1)$$

where A and B are evaluated by the method of least squares. These constants are then used to calculate the standard deviation of each observation from the value of the mean intensity for that set of equivalent terms. Wilson (1977) has pointed out that the variance rather than the standard deviation should be estimated from an expression of the form of (6.1) and we have therefore modified the program to use the expression:

$$\sigma^2 = A + BI. \tag{6.1a}$$

The results reported in this paper refer to the use of (6.1).

These initial estimates of the error standard deviation σ are modified when the complete set of photographs is merged to produce the unique set of data (Dodson, 1976). For each member of each set of equivalent reflections we compute the expression:

$$\frac{I_i - \bar{I}}{\sigma_i}, \tag{6.2}$$

where, now, I_i is the intensity of the ith observation, $\bar{I}$ is the weighted mean intensity for the set of equivalent terms over all the photographs, σ_i is the standard deviation of the ith observation as calculated from (5.1).

The data are divided into 20 equal ranges of $\bar{I}$ and the distribution of the expression (6.2) is evaluated for each range. The mean and standard deviation for each distribution are calculated. Provided that the σ_i are good estimates of the error variation in the data, these means should be 0·0 and the standard deviations 1·0. Furthermore, these values should be invariant with the range of intensity. Because of the manner in which the data have been processed up to this point, the means of the distributions do correspond well to zero. However, the standard deviations may depart from unity, because of error variation not yet allowed for in the σ_i.

The original σ_i are adjusted iteratively according to the expression

$$\sigma_i = C(\sigma_i^2 + D\bar{I}^2)^{1/2}, \tag{6.3}$$

where C and D are constants over the whole data set chosen at each iteration in such a way as to bring the standard deviation of expression (6.2) nearer to its desired value of unity. At present this procedure is not automated, but is a 'manual' one allowing the crystallographer to use his judgement. We believe that this procedure provides an adequate estimate of the standard errors. In a typical set of phosphorylase b data used in the example below, there are 70 000 observed reflections which merge to give 20 000 unique reflections (see Table 2). Thus there are a significant number of equivalent terms from which to estimate the errors. Overdetermination is likely to occur in all applications of the oscillation camera to X-ray intensity measurement.

Table 2. *Summary of the data sets used in the calculations*

The space group is $P4_32_12$ with cell dimensions $a = b = 128{\cdot}5$, $c = 115{\cdot}9$ Å. There are 800 000 daltons of protein in the unit cell.

Data set	Native protein	Native + maltotriose
Number of observations	82 439	66 287
Number of unique reflections	19 480	18 271
Number of unique reflections with negative observations	746	486
% of negative observations	3·8%	2·7%
Overall merging R factor*	10·9%	8·7%

* The merging R factor is defined as

$$\sum_{\substack{\text{unique}\\ \text{reflections}}} \left(\sum_{i=1}^{N} |I_i - \bar{I}| \right) \Big/ \sum_{\substack{\text{unique}\\ \text{reflections}}} \left(\sum_{i=1}^{N} I_i \right)$$

where I_i is the ith observation on a set of N equivalent terms, $\bar{I}$ is the weighted mean intensity for this set.

7. Results

The Bayesian method of producing sensible estimates of structure factor moduli whatever the observed intensity is expected to affect significantly only those terms which are small. As a preliminary test of its usefulness, we describe here its application to two sets of data for the protein phosphorylase b (Johnson, Weber, Wild, Wilson & Yeates, 1977). The first set is for the native protein and the second for its complex with the pseudo-substrate maltotriose. An impression of the quality of the data can be gained from Table 2.

For each set of data the equivalent observations were merged to provide a unique set of intensities with associated standard deviations. The data were then treated in two different ways. (a) The negative observations were set to zero and the intensities square-rooted to give estimates of the structure factor moduli, both $|F_p|$ for the protein and $|F_{ps}|$ for the protein plus pseudo-substrate complex. The $|F_{ps}|$ data set was subsequently scaled to the $|F_p|$ set. (b) The unique sets of intensity data were independently processed by our Bayesian method to produce posterior means for $|F_p|$ and $|F_{ps}|$. The two sets were then scaled together as above. The respective effects on the isomorphous differences, $||F_{ps}| - |F_p||$, are shown in Figs. 4 and 5. The results of the two treatments differ only when $|F_p|$ and/or $|F_{ps}|$ are small, and hence also when $4(\sin^2\theta)/\lambda^2$ is large. The isomorphous difference for the smallest terms is systematically lower for the data sets processed by the Bayesian method. The isomorphous differences for the data processed as in (a) above show a spurious rise for reflections with a small $|F_p|$ or $|F_{ps}|$.

The discrepancy between difference Fourier syntheses computed from the data treated in the two different ways is negligible. This is not surprising. The errors in such syntheses are dominated by the errors in the phases; the same set of phases was used in the computation of both syntheses. It is our intention to produce two different sets of phases by the isomorphous replacement method based on the two sets of isomorphous differences. The resulting Fourier syntheses will lead to a fairer test of our procedure, but the computational effort involved is great.

8. Conclusions

We should point out that to set negative intensity observations to zero, as in the previous section, is the worst possible way to treat such information, and will introduce maximal bias into the distribution. Other methods provide a more sensible means of treating negative observations. For example, all observations less than three standard deviations ('unobserved data') may be set to some fraction of the smallest 'observed' intensity. Such methods may lead to a reasonable approximation to the true intensity distribution for the complete set of data and reduce the bias introduced.

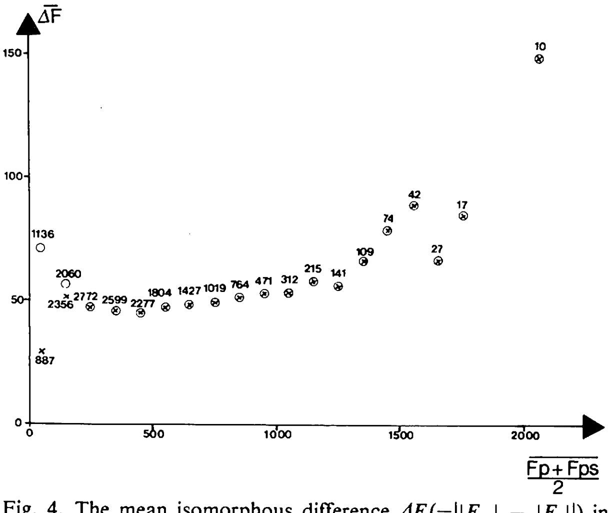

Fig. 4. The mean isomorphous difference $\Delta F(=||F_{ps}| - |F_p||)$ in ranges of $(|F_p| + |F_{ps}|)/2$. The data are those described in Table 2. O processed by method (*a*), § 7, × processed by the Bayesian method (*b*), § 7.

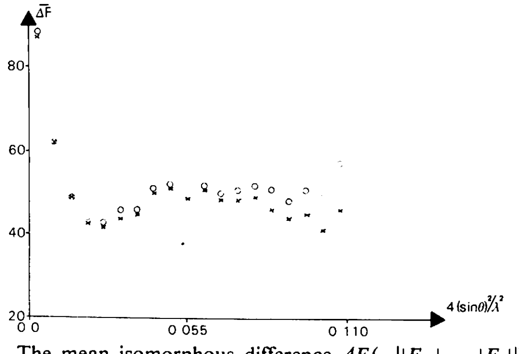

Fig. 5. The mean isomorphous difference $\Delta F(=||F_{ps}| - |F_p||)$ in ranges of $4(\sin^2\theta)/\lambda^2$. The data are those described in Table 2. O processed by method (*a*), § 7, × processed by the Bayesian method (*b*), § 7.

Nonetheless, we do suggest that all other methods suffer from two disadvantages relative to our own. Firstly, there is a discontinuity in their treatment of the observations with a cut-off applied at some *ad hoc* point, *e.g.* three standard deviations. Secondly, and more importantly, all observations in some class are accorded the same intensity. This means that information is certainly lost. The present Bayesian treatment is subject to neither of these criticisms.

Our experimental results strongly suggest that the Bayesian method of estimating structure factor moduli from intensities is more reliable and sensible than previous methods. We believe that a complete structure determination based on a Bayesian treatment of the data will yield a significant improvement, especially for a structure where the available data are weak in intensity.

We are grateful for financial support to ICI (KSW) and the Hayward Foundation (SF). We were greatly helped by conversations with Professor A. J. C. Wilson and were much encouraged by all in the Laboratory of Molecular Biophysics, University of Oxford.

APPENDIX A
Evaluation of posterior moments

Remember that our posterior belief in J is given by Bayes's theorem as [see (1.0)]:

$$p_J(J|I) \underset{J}{\propto} p_I(I|J)\,p_J(J). \qquad (A.0)$$

Further, remember that we are assuming that our conditional belief in the observation I given the true intensity J is normal, *viz*

$$I \sim N(J, \sigma^2). \qquad (A.1)$$

This assumption has the implication that should we observed an intensity such that $I < -4\sigma$, then something rather untoward has happened. In such cases we always reject the reflection, that is omit it from the data set.

Our task in this Appendix is to find algorithms for evaluating the posterior moments $E_J(J|I)$, $\mathrm{var}_J(J|I)$, $E_J(F|I)$ and $\mathrm{var}_J(F|I)$ [see (1.2), (1.3), (1.4) and (1.5)]. These evaluations will depend on our choice of prior.

Consider first the prior (2.0):

$$p_J(J)\begin{cases} = 1 & \text{if } J \geq 0 \\ = 0 & \text{if } J < 0. \end{cases} \qquad (A.2)$$

Using Bayes's theorem (A.0) gives immediately:

$$p_J(J|I)\begin{cases}\underset{J}{\propto}\exp[-\tfrac{1}{2}(J-I)^2/\sigma^2] & \text{if } J\ge 0\\ =0 & \text{if } J<0.\end{cases}\qquad (A.3)$$

Thus our posterior belief is a truncated normal distribution. To evaluate the required moments for this we have found the following procedure simplest. For the range $-4\sigma \le I < 3\sigma$, we linearly interpolate within tables calculated by numerical integration routines. For $I \ge 3\sigma$, the truncated normal distribution (A.3) is negligibly different from the proper normal

$$J \sim N(I, \sigma^2).$$

Thus

$$E_J(J|I) \simeq I$$

$$\operatorname{var}_J(J|I) \simeq \sigma^2$$

$$E_J(F|I) \simeq \sqrt{I}$$

$$\operatorname{var}_J(F|I) \simeq \sigma^2/4I$$

are perfectly satisfactory approximations.

Consider now the prior (2.2), *i.e.* Wilson's distribution for an acentric reflection:

$$p_J(J)\begin{cases}=(\Sigma)^{-1}\exp(-J/\Sigma) & \text{if } J\ge 0\\ =0 & \text{if } J<0.\end{cases}\qquad (A.4)$$

An application of Bayes's theorem (A.0) gives:

$$p_J(J|I)\begin{cases}\underset{J}{\propto}\exp[-\tfrac{1}{2}(I-J)^2/\sigma^2]\exp(-J/\Sigma) & \text{if } J\ge 0\\ =0, & \text{if } J<0.\end{cases}$$

On 'completing the square' in the exponent:

$$p_J(J|I)\begin{cases}\underset{J}{\propto}\exp\{-\tfrac{1}{2}[J-(I-\sigma^2/\Sigma)]^2/\sigma^2\} & \text{if } J\ge 0\\ =0 & \text{if } J<0.\end{cases}\qquad (A.5)$$

This posterior distribution is again truncated normal, but, instead of the underlying distribution being $N(I, \sigma^2)$, it is $N[(I - \sigma^2/\Sigma), \sigma^2]$. The moments of this distribution are calculated by the algorithm given above with the incorporation of this simple change in the underlying mean.

Lastly, consider the prior given by (2.3), *i.e.* Wilson's distribution for a centric reflection. Bayes's theorem (A.0), gives, after 'completing the square' in the exponent:

$$p_J(J|I)\begin{cases}\underset{J}{\propto}J^{-1/2}\exp\{-\tfrac{1}{2}[J-(I-\sigma^2/2\Sigma)]^2/\sigma^2\} & \text{if } J\ge 0\\ =0 & \text{if } J<0.\end{cases}\qquad (A.6)$$

This, unfortunately, is not a truncated normal distribution, as can be seen from the presence of the $J^{-1/2}$ term. However, it is nonetheless fairly easy to approximate the moments of this distribution. Again for $-4\sigma \le (I - \sigma^2/2\Sigma) < 4\sigma$ the moments have been tabulated by numerical integration. Linear interpolation is used in these tables for this range. Outside this range we use an approximation developed as follows.

Let Y have the normal distribution $N[(I - \sigma^2/2\Sigma), \sigma^2]$, then we have immediately:

$$E_J(J|I) \simeq E_Y(Y^{1/2})/E_Y(Y^{-1/2}) \qquad (A.7)$$

$$E_J(F|I) \simeq 1/E_Y(Y^{-1/2}) \qquad (A.8)$$

$$\operatorname{var}_J(F|I) = E_J(J|I) - [E_J(F|I)]^2 \qquad (A.9)$$

$$\operatorname{var}_J(J|I) = \operatorname{var}_J(F|I)\, 4E_J(J|I). \qquad (A.10)$$

Instead of (A.10) we could have noted that

$$E_J(J^2|I) \simeq E_Y(Y^{-3/2})/E_Y(Y^{-1/2}) \qquad (A.11)$$

and calculated $\operatorname{var}_J(J|I)$ accordingly, but this is unnecessary to our level of approximation. There is still the problem of calculating $E_Y(Y^{1/2})$ and $E_Y(Y^{-1/2})$. Note that if Z is normally distributed $N(\mu, 1)$ and $\mu > 4$, then (French, 1978, Appendix B):

$$E_Z[f(Z)] \simeq f(\mu) + \tfrac{1}{2}f''(\mu) + f^{iv}(\mu). \qquad (A.12)$$

Combining (A.7), (A.8) with (A.12) gives, on writing $h = (I/\sigma - \sigma/2\Sigma)$:

$$E_J(J|I) \simeq h\sigma(1 - h^{-2}/2 - 3h^{-4}/4) \qquad (A.13)$$

$$E_J(F|I) \simeq (h\sigma)^{1/2}(1 - 3h^{-2}/8 - 87h^{-4}/128). \qquad (A.14)$$

The required variances can then be developed by using (A.9) and (A.10).

References

ARNDT, U. W., CHAMPNESS, J. N., PHIZACKERLY, R. P. & WONNACOTT, A. J. (1973). *J. Appl. Cryst.* **6**, 457–463.

BARNETT, V. (1973). *Comparative Statistical Analysis.* London: Addison-Wesley.

BOX, G. E. P. & TAIO, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Reading, Mass.: Addison-Wesley.

DIAMOND, R. (1969). *Acta Cryst.* A**25**, 43–55.

DIANANDA, P. H. (1953). *Proc. Cambridge Philos. Soc.* **49**, 239–246.

DIANANDA, P. H. (1954). *Proc. Cambridge Philos. Soc.* **50**, 287–292.

DIANANDA, P. H. (1955). *Proc. Cambridge Philos. Soc.* **51**, 92–95.

DODSON, E. (1976). *Crystallographic Computing Techniques*, edited by F. R. AHMED, pp. 205–212. Copenhagen: Munksgaard.

FRENCH, S. (1975). D. Phil. Thesis, Univ. of Oxford.

FRENCH, S. (1977). *Probability: Some Interpretations.* Internal Paper, Department of Decision Theory, Univ. of Manchester.

FRENCH, S. (1978). *Acta Cryst.* To be published.

HAMILTON, W. C. (1964). *Statistics in the Physical Sciences.* New York: Roland Press.

HIRSHFELD, F. L. & RABINOVICH, D. (1973). *Acta Cryst.* A**29**, 510–513.

HOWELLS, E. R., PHILLIPS, D. C. & ROGERS, D. (1950). *Acta Cryst.* **3**, 210–214.

IRWIN, J. O. (1937). *J. R. Stat. Soc.* **100**, 415–416.

IWASAKI, H. & IZO, T. (1977). *Acta Cryst.* A**33**, 227–229.

JOHNSON, L. N., WEBER, I. T., WILD, D. L., WILSON, K. S. & YEATES, D. G. R. (1977). 4th Eur. Crystallogr. Meet., Oxford, Abstr. PI. 111.

LINDLEY, D. V. (1965). *Probability and Statistics from a Bayesian point of View*, Vols. 1 and 2, Cambridge Univ. Press.

MCCANDLISH, L. E., STOUT, G. H. & ANDREWS, L. C. (1975). *Acta Cryst.* A**31**, 245–249.

SKELLAM, J. G. (1946). *J. R. Stat. Soc.* **109**, 296.

TICKLE, I. (1975). *Acta Cryst.* B**31**, 329–331.

WATSON, H. C., SHOTTON, D. M., COX, J. M. & MUIRHEAD, H. (1970). *Nature* (*London*), **225**, 806–811.

WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

WILSON, A. J. C. (1977). Private communication.

WILSON, K. S. & YEATES, D. G. R. (1978). In preparation.

Acta Cryst. (1978). A**34**, 525–528

The Extension Concept and Its Role in the Probabilistic Theory of the Structure Seminvariants*

BY HERBERT HAUPTMAN

Medical Foundation of Buffalo Inc., 73 *High Street, Buffalo, New York* 14203, *USA*

(*Received* 4 *November* 1977; *accepted* 17 *January* 1978)

By embedding the structure seminvariant T and symmetry-related variants of T in suitable structure invariants Q the values of which, because of the space-group-dependent relations among the phases, are related to T, one reduces the probabilistic theory of the structure seminvariants to that of the structure invariants, which is well developed. The structure invariants Q are said to be extensions of the structure seminvariant T.

1. Introduction

It is assumed that the reader is familiar with the idea of 'neighborhood of a structure invariant or seminvariant', the 'neighborhood principle', and the roles these concepts play in the probabilistic theory of the structure invariants and seminvariants (see, for example, Hauptman, 1975, 1976; Green & Hauptman, 1976). Systems of neighborhoods of the structure invariants are now well known (see, for example, Hauptman, 1977*a*,*b*), and neighborhoods for selected structure seminvariants have also been identified (see, for example, Green & Hauptman, 1978).

The major goal of the present paper is to show how to determine in a systematic and unambiguous way neighborhoods of the structure seminvariants in general by exploiting the symmetries deriving from the space groups. The method is to embed a given structure seminvariant T and its symmetry-related variants in suitable structure invariants Q to which T is related *via* the space-group symmetries. Then the neighborhoods of T are determined by the known neighborhoods of Q. The structure invariant Q is said to be an extension of the structure seminvariant T. Recently secured methods may then be employed to derive suitable conditional probability distributions leading to estimates of the structure seminvariants in terms of the magnitudes in their neighborhoods.

The method will be illustrated by examples in space groups $P1$, $P\bar{1}$, $P2_1$ and $P2_12_12_1$ but is clearly of sufficient generality to be applicable to structure seminvariants in general.

Although the idea of embedding a structure seminvariant in an appropriate structure invariant is not new (Giacovazzo, 1975; Hauptman, 1976; Green & Hauptman, 1978) the present paper appears to be the first in which the interplay between the space-group symmetries and the neighborhood concept is systematically exploited. However, see Hauptman (1976) and Giacovazzo (1977*b*) for different techniques for obtaining neighborhoods of the structure seminvariants.

2. The second neighborhoods of the three-phase structure invariant in $P1$ and $P\bar{1}$

The linear combination of three phases

$$T=\varphi_{\mathbf{h}}+\varphi_{\mathbf{k}}+\varphi_{\mathbf{l}}, \qquad (2.1)$$

is a structure invariant if

$$\mathbf{h}+\mathbf{k}+\mathbf{l}=0. \qquad (2.2)$$

[*] Presented at the Michigan State University Meeting of the American Crystallographic Association, August 7–12 1977, Abstract H3.