Isomorphism replacement.

Criteria for sign determination

Introduction

The general problem is: A vector $\mathbf{a}$ is given.

The sort of problem one is concerned with is, for example:

"given a set of isomorphism replacement data - suppose not real amplitude - how do I decide whether a sign of a particular term is determined or not?" or more generally "what is the best way to handle the data if there are experimental errors present?" (as of course they always are). This little note covers my first tentative attempt at this problem.

Apart from a few remarks on the end it only covers centric (i.e. for protein, $= 2D\frac{1}{2}$ case), and it only deals with one of the many sources of error, i.e. with random errors of measurement. In other words, I have assumed it, for simplicity, the other possible sources of error such as
The uncertainty of the position of the heavy atom, lack of complete isomorphism, and an incorrect scaling factor are negligible, which is, of course, unlikely.

Again for simplicity, I have chosen a convenient form for the error distribution function. I have assumed that the total error in estimating $F_i$, the heavy atom contribution for some particular reflection, $F_i$, has a gaussian distribution with mean value $E_i$ and that $E_i$ is independent of both $F_i$ and $F_0$ (an heavy atom, protein). This seems a rather odd way of doing it, but it is really quite reasonable. For reflection having the same heavy atom factor $F_i$, it is highly equivalent to assuming that the experimental error is Poisson, neither a constant nor proportional to the observed intensity, but proportional to the square root of the observed intensity. This is probably a good approximation to the actual error, but it except for the very spot is not much greater than the background. For a given
...corresponds to the counting error rate counting for a fixed time for all reflections (ie error proportional to $\sqrt{\text{counts}}$).

In the problem I arrived at what I now regard as the best formulae. I made a number of preliminary attempts. I have written these on (as this is a NEC Internal note) as other people may otherwise waste time covering the same ground, and it is possible they may be more appropriate for the other source of error.

**Terminology**

I have usually formulated the problem as follows.

$F_0$ = calculated (assumed correctly) value of heavy atom contribution to the reflections being considered.

$F_0$ = observed change in $F_0$ this naturally includes the error.

Now we determine a sign by assuming that the sign of $F_0$ is the...
calculated
same as the sign of $F_H$. Occasionally, errors will reverse the
sign of $F_0$ and our assumption will be incorrect, and we shall therefore
deduce the sign of $F_0$ ($F_0 = \text{protein carbazole} \& F$) incorrectly.

Then I shall draw the curve of distribution of $F_0$, for a given
$F_H$, as

and shall always consider, for convenience, that $F_H$ is in the $(H)$
region and simply reverse the sign of the $(H) F_0$ axis. My assumption that
$E$ (the RMS error) is constant with $F_H$ and $F_0$ means that this
curve always has the same shape, but is shifted depending on the
value $F_H$, i.e., for a reflexion with a smaller $F_H$ it would look like

It is as well to spend a minute or
two making sure one has understood
these very simple considerations.
The curve of $F_o$ is then simply
\[ \frac{1}{E S_{2N}} \exp\left(-\frac{(F_o - F_n)^2}{2E^2}\right) \]

The curve being normalised to make the area under it unity.

Note that in practical use of $F_o$ this correction, $E_{2N}$, observed,
not $F_o$ but $|F_o|$. 

\[=\]
First Attempt

This was as follows.

Gives $F_H$, $E$, and $\frac{1}{E}$: when is

\[
\begin{align*}
\frac{\text{chance of being right wrong}}{\text{chance of being right}} &= 1
\end{align*}
\]

This is clearly

\[
\begin{align*}
\exp \left( \frac{F_H + 1/E}{2E} \right) &= \exp \left( \frac{-4F_H/1/1}{2E} \right) \\
\exp \left( \frac{F_H - 1/E}{2E} \right)
\end{align*}
\]

Thus, if we want \[
\frac{\text{chance wrong}}{\text{chance right}} \leq \frac{1}{100},
\]
we shall require \[
\frac{F_H}{1/E} \geq 2.3 \times 10^9.
\]

and if we plot \[
\frac{F_H}{1/E} \text{ against } \frac{1/E}{1/E}
\]
we shall get a hyperbola for the case corresponding to the equality sign. Different values of the chance ratio being considered will give different hyperbolas. In Fig. 11, curve A, the chance ratio is less than 1 in 100. Clearly in Fig. 11, where the chance of being wrong is less than 1 in 100, we shall be close the curve, and hence for much smaller than 1 in 100 will lie below it.
In general the chance of being right of being wrong is \( \frac{1}{1+k} \), then

\[
\frac{F_0 \cdot F_1}{E^2} = \log \left( \frac{1}{1+k} \right)
\]

Suppose, therefore, we adopt the rule that we shall "decide" the sign whenever the chance of being wrong is 1 in 100 or less. Then it is a question of occasion for reflection with a given \( F_0 \) and \( E \), shall we be able to decide?

This is simply the integral under the appropriate part of the \( F_0 \) distribution curve, and in reality one to be, for the 1 in 100 case,

\[
\Delta F_0 = \frac{1}{2} \left[ \text{erf}(y + \frac{115}{7}) - \text{erf}(y - \frac{115}{7}) \right]
\]

where \( y = \frac{F_0}{\sqrt{E}} \), \( \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} \, dt \), and the \( \pm \) sign is taken depending upon \( y \) if \( y < \frac{115}{7} \)

\[
\begin{align*}
\{ & 1 & \text{if} & y \geq \frac{115}{7} \\
& -1 & \text{if} & y \leq \frac{115}{7}
\}
\end{align*}
\]

This curve is plotted in Fig. 2, curve A.
Second attempt

In the first attempt I adopted the criterion that I should "decide" a right if the chance of being wrong was 1 in 100 (say) or less. In the second I decided to use a less stringent criterion, namely that on the average one would be wrong 1 in 100 times. This means one is often then one has to consider the area under the curve. One then easily obtains the relation

\[
1 - \frac{\text{erf} \left( \frac{E - \langle E \rangle}{\sigma E} \right)}{\sqrt{\pi} \sigma E} \leq 1 - \frac{\text{erf} \left( \frac{E + \langle E \rangle}{\sigma E} \right)}{\sqrt{\pi} \sigma E}.
\]

Fraction right = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{E - \langle E \rangle}{\sigma E} \right) \right]

Fraction wrong = \frac{1}{2} \left[ 1 - \text{erf} \left( \frac{E + \langle E \rangle}{\sigma E} \right) \right]

Fraction undecided = \frac{1}{2} \left[ \text{erf} \left( \frac{E - \langle E \rangle}{\sigma E} \right) - \text{erf} \left( \frac{E + \langle E \rangle}{\sigma E} \right) \right]

And then if we put fraction wrong = \frac{k}{1 - k'}, we can obtain for any given fraction right a graph here a curve which shows whether to decide a case or not. This is shown above in Fig. 1. All cases above
\[ I = (I_0^t + F)^t = I_0^t + 2FE + F^t \]
\[ = F^t + 2FE \]

\[ I_{obs} = k(F + E)^t \]
\[ = k(F^t + 2FE + E^t) \]
\[ = k(F^t + 2FE) \]

\[ I_{obs} = I \pm I_E \]

We assume \( I_E = k \sqrt{I} \)

\[ I_{obs} = I \pm k \sqrt{I} \]
\[ F^t \pm 2FE \]
\[ F^t \pm 2F \sqrt{E} \]

\[ 2E = k \]
the line should be "decided": all cases below the line we should
be left "undecided" [The curves should have been calculated for
$k' = 1$ in 100, but was accidentally done for $k' = 100$. The values
were obtained with tables and successive approximation] Normally to

A "decide" a term we need to know not only the calculated
$F_0$, but also the observed difference $F_0$. None knows that if
$F_0$ is big enough ($F_0 > 3.165$) we do not need to know the
magnitude of $F_0$ to "decide" a term (and naturally if it is zero
we cannot find $F_0$). This paradoxical result is simply the result of
our assumption of the form of the distribution curve, and it simply

means here if $F_0$ is big enough then small value of $\frac{F_0}{\sigma_0^2}$, or

large value of $F_0$ becoming $\frac{F_0}{\sigma_0^2}$. giving the wrong sign, will be rare.

We can also evaluate for this case the fraction of cases determined
($5\alpha$ of them will be right and $1/\alpha$ will be wrong), for and given value

of $\frac{F_0}{\sigma_0^2}$, and then in Figure 20 of curve B.
Third Attempt

As this, pour I paused, and reflected there after. All I was now setting up as a Bookie, and that this business of quoting odds needed some justification. What, after all, am I really aiming at? Having put the question, a reasonable answer suggested itself. What we appear to require from any given set of data is:

1. the best possible Fourier

and

2. an estimate of how bad it is in.

Now here is one rather obvious criteria of badness: This is the root mean square (R.M.S) deviation from the "perfect" Fourier, & that is one to with all the spins right—and, perhaps, one should also say "with all the . . ."

Is revealed core without experimental error. But this last point.
is probably not as important one as the first stage.

Notice that in this criteria is not necessarily the best one - it is merely the most convenient. For example if one can deal with a structure in which one could see the atoms, a criterion to which stressed this particular aspect of the Fourier might be more useful. For the proteins, and especially for less resolution projections, the simple criteria would seem to be adequate.

Of course a degree of smoothing can be included, but this needs not upset the underlying mathematical formulation.

We then define as our measure of "fuzziness"

\[ \frac{1}{V} \int (f - e)^2 \, dv \]

where \( e \) = "true" electron density is with all signs correct.

\( f \) = experimental electron density, with some sign errors. One sign error = 1.

Thus, scales "perfect" such that \( E = 0 \)

scale "no information" such that \( E = 1 \)
So that, by Bonville theorem, we have

\[ \text{Badness} = \frac{\sum (F - \overline{F})^2}{\sum F^2} \]

Note that one could just as reasonably define the badness as \( \frac{1}{n} \) of this expression. (This is for the central case, but these will be an analogous expression for the general case.)

We now group the terms into three classes.

1. Right sign = \( R \)
2. Wrong sign = \( W \)
3. Undecided = \( U \)
   (a term omitted in Fourier)

and then so we have \( \sum (F - \overline{F})^2 = \sum 4F_w^2 + \sum F_o^2 \)

Now when we require in a case criteria to enable us to decide "shall we include this particular term, with its apparent sign, or shall we omit it?" and the criteria will depend on \( F_w \) and \( |F_o| \).
In other words, given \( F_0 \) and \( F \), we need to define a

"choice of decision value" = \( \frac{F_0}{F} \), and then if

\[ F_0 > \frac{F_0}{F} \]

we put the term in, and if

\[ F_0 < \frac{F_0}{F} \]

we leave it out.

Then, the number of cases of

\( R, W \) and \( U \) we get will be

proportional to the area under the curve above the sheet.

So the choice of \( W = \frac{1}{2} \left[ 1 - \text{erf} \left( \frac{F_0 + F}{\sqrt{2}E} \right) \right] \)

and choice of \( U = \frac{1}{2} \left[ \text{erf} \left( \frac{F_0 - F}{\sqrt{2}E} \right) - \text{erf} \left( \frac{F_0 - F}{\sqrt{2}E} \right) \right] \)

(for \( F_0 \geq 2F \)).

Now, to minimize the "sudden" we need with

\( R, W \) or \( U \) or choice of \( F_0 \), we need

\[ 4 \frac{d(\text{chance of } W)}{dF_0} + \frac{d(\text{chance of } U)}{dF_0} = 0 \]

And carrying out the algebra we easily obtain the conclusion

\[ \exp \left( \frac{2F_0}{F} \right) = 3 \quad \text{or} \quad \frac{F_0}{F} = 0.549. \]
This is plotted in Fig. 1 curve C. Comparing it with our known cases from two cases we see that (over most of the range) our previous criteria had been much too cautious, and seem to have sprung from a sincere desire for certainty, rather than reality, together with a regrettable tendency (in the higher range of the second case) to gamble on a dead "dead end."

There is no difficulty in obtaining the curve for the fraction of terms undecided, using the above criteria, i.e., evaluating the expression for \( u \) on the previous page, and this is shown in the curve C, Figure 2.

In this case however we have a further measure of our method to compute, namely the mean contribution to the goodness of
reme with a given value of \( \frac{F_f}{F_p} \). This again is
easily done, as since it is \( 4W + V \), using the
formulas given on page 13. This has been computed and
plotted in Fig. 3 curve C.

I should insert here a remark I ought to have
made earlier. This is that one is assuming that \( F_f \) and
\( F_p \) are uncorrelated, which is only true for the average
over all possible structures, though it is likely to be
true except for our purposes in any particular case. These
will be difficulties here when \( F_p \) is very small. This
assumption means that on the average one is deciding a
risk the "badness" does not depend on \( F_p \), but does the
decrease does depend on \( F_f \). This point has been made
made in outline earlier, but it needs emphasis here.
Fourth Attempt

Having completed my third attempt I thought I had
the problem in a satisfactory shape when it occurred to me
for my formulation could be improved. For it is not true that
the only choice we have is to include a term or leave it
out. For there is nothing to prevent us putting it in
with reduced weight, and adjusting the weights to express
our a uncertainty of [in the choice of n]. In other
words, we are in our Fourier P F where p is
equal to 1 for cases where the weight is certainly determined,
equal to zero when there is no information, and will in
general lie between 0 and 1. Our problem now is to
find the an expression which allows us to calculate p,
Given $F_1$, $|F_0|$ and $E$, let the value of $p$ being chosen to minimize the "badness". This presents no difficulty.

Let

\[
\frac{\text{chance of being wrong}}{\text{chance of being right}} = k
\]

\[
\text{Our "badness"} = (1-p) \left( \frac{1}{1+k} \right) + (1+p) \left( \frac{k}{1+k} \right)
\]

We now have that \( \frac{\text{d(expected badness)}}{dp} = 0 \)

and easily obtain

\[
k = \frac{1-p}{1+p}
\]

or

\[
p = \frac{1-k}{1+k}
\]

which is the root of being right??

Now

\[
k = \exp\left(- \frac{(F_1 + |F_0|)^2}{2E^2}\right) = \exp\left(- \frac{2|F_1|F_0}{E^2}\right)
\]

\[
p = \frac{1 - \exp\left(- \frac{2|F_1|F_0}{E^2}\right)}{1 + \exp\left(- \frac{2|F_1|F_0}{E^2}\right)}
\]

which is plotted in Figure 4.
In this case, as there is no meaning in plotting a curve of the number of terms as undetermined, since in a sense we determine them all, though with varying degrees of uncertainty (they would all be hyperbolas) we should here be to plot a family of curves, for different values of \( p \). In increasing to note that curve C of Figure 1 (which applied to our third case) corresponds to this formulation to the curve for weight \( = \frac{1}{2} \). In other words, our third case can be regarded as a simplification of the present one, in which all terms which should have weight between \( 2 \) and \( 1 \) are given weight 1, and those with weight between 0 and 1 are omitted from the formula.

Curves B of Figure 1 can easily be taken & the appropriate hypothesis for the weight \( p = 0.98 \).
It remains to calculate the "average b tournament", given $F_n, F_0$, and $E$. This turns our to be somewhat more difficult, as it leads to an awkward integral.

Consider how the contribution to the b tournament when the sign is determined correctly. The chance of this occurring is

$$\frac{1}{\pi} \exp \left( -\frac{E_n}{E} \right)$$

and the b tournament factor is $(1 - p)^{E_n}$. Substituting

and using our previous values for $p$, we easily obtain for the total and contribution of this part to the b tournament

$$\int_0^\infty \frac{1}{\pi} \exp \left( -\frac{E_n}{E} \right) \left[ 1 + \exp \left( -\frac{2E_n}{E} \right) \right]^{-\frac{1}{2}}$$

And nothing for convenience $x = \frac{F_n}{E}$ and $x_0 = \frac{F_0}{E}$. We obtain

$$4 \int_0^\infty \frac{\exp \left[ - (x - x_0)^2 \right] \exp \left( -8x_0 x \right)}{(1 + \exp \left( -4x_0 x \right))^{1/2}} dx = \frac{16}{\pi} \int_0^\infty \frac{\exp \left[ - (x + x_0)^2 \right]}{(1 + \exp (2x + 2x_0))} dx$$
Therefore when the sign is determined incorrectly, the

\[ \frac{1}{E} \exp \left( -\frac{(E_x + E_0)^2}{2E^2} \right) \]

and carrying on the same steps and substituting a.

before we obtain for the total contribution from cases of the rest

\[ \int_{-\infty}^{\infty} \frac{\exp \left[ -(x + x_0)^2 \right]}{\cosh^2(2x x_0)} \, dx \]

So that, adding the two contributions together, we obtain

for the average density

\[ \int_{-\infty}^{\infty} \frac{\exp \left[ -(x + x_0)^2 \right]}{\cosh^2(2x x_0)} \, dx \]

I have been unable to evaluate this integral

and it will have to be done numerically.
Total Backmen

Notice that so far we have only computed the average contribution to the backmen for terms all keeping the same value of $F_n$. This is not what we really require. Rather we want to know the total value of the backmen from all terms, gives a heavy atom of a certain size and the appropriate value of $E$.

We must then sum over all values of $F_n$. This is more easily done numerically. The results are for the case of one heavy atom, $E$ (related to a diamond) are shown for case $C$ in Figure 5. (Case $D$ has not yet been worked out). Here is here the maximum contribution from the heavy atom, and $E$ (the total error on $F_n$) is averaged "corrected" over the reflections considered. The value of
$F_p$ have been assumed to be within error.

Eventually such cases I hope to compute a

and curves for all four cases, as this will enable one to see how much one gains by using one criterion rather than another.

How much badness can one tolerate?

The answer to this would clearly be given a

"that depends."

But to fix ideas remember one might

ask how much straightforward experimental error, after the badness. From the definition we see that an RMS error $\sigma$ |

$\frac{1}{1000}$ sign as bad as or the measured inaccuracy, given

a badness of $\frac{\sigma}{2^4}$ (for $\varepsilon$ small) : note rows, if it

is $\frac{\sigma}{2^4}$, not $\varepsilon$ (this point is worth checking).
30% of the intensity measurement is only giving acceptable errors in the Fourier series. This means that we are prepared to accept a standard of about 0.02. It remains to be seen whether this is tolerable for the case of proteins. Generically, the formulation above may still, as by chance, prove more important than current applications in structure determination.
Non-cancer care

The Cure itself approach has one great formal advantage - it can be carried over into the non-cancer care. I have done some preliminary work on this.

The mathematics is irreducible, but an approximate treatment looks promising. One of their letters.
Decision curve

When point falls above the line - include
When point falls below the line - exclude

\[ \frac{F_{o_b}}{E} \]

\[ F_{o_b} = \text{observed } F \text{ due to heavy atom} \]

\[ F_n = \text{calculated } F \text{ due to heavy atom} \]

\[ E = \text{joint RMS error in determining } F_{o_b} \]

Case A (No lead)

Case B (No lead)

Case C
Fraction of noise "determined" for a given uncertainty $F/n$. 

Fraction of cases undefined.
Average contribution to the "Badness" from a term of given $F_k$.
\[ y = \frac{1 - \exp(\cdot x)}{1 + \exp(\cdot x)} \]
Fig 5

Bedrock factor (a fraction of depth to bedrock)

[Graph with labeled axes and points, indicating relationships between bedrock factor and depth to bedrock.]

(E - 0.055 River + 0.6)
\[ E = P u \quad E^2 = 2.7060 \quad P = 3.500 \]

We have the formula:
\[ p = 1 - \exp \left( -\frac{2(\sqrt{E}F_F)}{E} \right) \]

where:
\[ y = \frac{1 - \exp(-x)}{1 + \exp(-x)} \]

<table>
<thead>
<tr>
<th>x</th>
<th>\exp(-x)</th>
<th>1 - \exp(-x)</th>
<th>1 + \exp(-x)</th>
<th>1 - \exp(-x) \cdot 1 + \exp(-x)</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.353</td>
<td>0.8647</td>
<td>1.4173</td>
<td>0.1353</td>
<td>0.762</td>
</tr>
<tr>
<td>2.5</td>
<td>0.0821</td>
<td>0.9179</td>
<td>1.0821</td>
<td>0.0821</td>
<td>0.948</td>
</tr>
<tr>
<td>3.0</td>
<td>0.0498</td>
<td>0.9502</td>
<td>1.0498</td>
<td>0.0498</td>
<td>0.965</td>
</tr>
<tr>
<td>3.5</td>
<td>0.0302</td>
<td>0.9698</td>
<td>1.0302</td>
<td>0.0302</td>
<td>0.941</td>
</tr>
<tr>
<td>4.0</td>
<td>0.0183</td>
<td>0.9817</td>
<td>1.0183</td>
<td>0.0183</td>
<td>0.964</td>
</tr>
</tbody>
</table>