Acta Cryst. (1973). A**29**, 510

Treating Weak Reflexions in Least-Squares Calculations

BY F. L. HIRSHFELD AND D. RABINOVICH

Department of Structural Chemistry, Weizmann Institute of Science, Rehovot, Israel

(*Received* 24 *October* 1972; *accepted* 16 *February* 1973)

In photographic intensity measurements, the non-observability of a reflexion establishes a conditional probability distribution for the true value of F that is undefined below an observational threshold. By contrast, counter methods yield values of F_o^2 in which the errors can be expected to approach a normal distribution even for the weakest reflexions. In a least-squares refinement, preferably based on F^2, this normal distribution should not be distorted by the arbitrary exclusion, or other maltreatment, of 'unobserved' reflexions. Such selective tampering with weak intensities biases the input data and so risks systematic error in the refined parameters. The expected effect is illustrated by a numerical experiment.

Among the more bewildering features of crystallographic data processing is the variety of prescriptions in vogue for the doctoring of reflexions too weak to be accurately measured. Two fallacies seem to lie at the root of much of the perplexity. One is a specious attempt to augment the information conveyed by the absence of a detectable intensity on an X-ray film with extraneous considerations of non-experimental origin. The other is an inertial tendency to use a diffractometer as a tool for simulating photographic data. Both continue to inflict an unknown burden of avoidable error on many crystallographic refinements.

Unobserved reflexions in photographic data

On an X-ray diffraction photograph, weak reflexions may produce no detectable mark and so deserve the label 'unobserved' or 'accidentally absent'. Several schemes have arisen for handling this situation, including the total neglect of all such reflexions. Hamilton (1955), Cruickshank, Pilling, Bujosa, Lovell & Truter (1961), and Arnott (1965) have fallen back on *a priori* probability distributions to supply expected amplitudes and variances for unobserved reflexions to be used in least-squares calculations, though Hamilton (1955) implicitly acknowledges the irrelevance of such *a priori* estimates when values of F_c are available from a refined model. A more basic flaw is that such a proposal seeks to graft onto F_o properties that belong to F_c and so undermines in advance the vital confrontation, *via* least-squares matching of F_c to F_o, between the structural model and the experimental evidence.

Dunning & Vand (1969), demonstrating the inconsistency of this approach, assign to unobserved reflexions a uniform – we would rather call it undefined – probability distribution between zero and an estimated observational threshold $F_{\rm lim}$, which requires that such reflexions be disregarded so long as F_c is smaller than $F_{\rm lim}$. These authors evidently take the probability density to fall abruptly to zero above $F_{\rm lim}$ and so offer no definitive recipe for dealing with reflexions whose F_c, at the end of a refinement, exceeds $F_{\rm lim}$. (They value unobserved reflexions mainly for promoting the convergence of a crude model, a practical objective justifying a pragmatic solution; once convergence has been assured, a more fundamental standard becomes appropriate.) An answer favoured by some workers, and ourselves, is to give these reflexions, *i.e.* when $F_o < F_{\rm lim} < F_c$, the same amplitude and statistical weight as would be given to an observed reflexion having $F_o = F_{\rm lim}$. This amounts to postulating a *conditional* probability distribution, supposing the true value of F to exceed $F_{\rm lim}$, that falls off in the usual way with increasing $F - F_{\rm lim}$. Thus, the knowledge that a reflexion is unobserved says nothing about the relative probabilities of different values of F below the threshold value $F_{\rm lim}$, nor does it assign a numerical probability to the proposition that the true value lies below $F_{\rm lim}$ rather than above;* what it tells us, provided $F_{\rm lim}$ has been suitably chosen, is that increasing values of F above $F_{\rm lim}$ are increasingly improbable. Accordingly, we should set $F_{\rm lim}$ at the point where the postulated probability function begins to decline and fix its weight in keeping with the steepness of this decline. But unanimity has yet to be achieved in the way crystallographers actually handle their data, and this question, like the related one of attaching weights to the observed reflexions, still offers ample scope for individual taste or contagious fashion.

Diffractometer data

More strangely, the habit of classifying reflexions as observed or unobserved, with the associated wealth of stratagems for handling the latter, has often survived the transition from photographic to diffractometric recording methods, creating confusion where none should have arisen. At least one popular text (Stout & Jensen, 1968) deplores this practice, asserting 'that

* One can in principle, by integrating between $F_{\rm lim}$ and some higher limit, obtain a lower bound to the absolute probability that F lies below this higher limit. Such a futile exercise has scant relevance to the least-squares method.

there is no theoretical basis for dropping reflexions, and that the 'best' results will be obtained from the complete data'. This dissent has recently gained reinforcement from a persuasive denunciation by Moore (1972) of the fraudulent suppression of weak intensities. Essentially the same lesson, it appears, has long been preached to more select audiences by such early converts as V. Schomaker and R. E. Marsh (personal communications). Our own late initiation was endorsed by Hamilton (1972), who ignited a lively debate by publicly championing our arguments in an unscheduled addendum to his prepared lecture. Our present aim, then, is to amplify this protest, for which we can claim authoritative support rather than priority of discovery.

The pertinent difference between cameras and diffractometers is that, unlike the photographic film, quantum counters are mainly limited in sensitivity, not by a lack of instrumental response to weak intensities, but by background noise of various kinds. In diffractometric measurements, the observed intensity that is taken as proportional to F_o^2 typically appears as a difference $I-B$ between total and background counts,

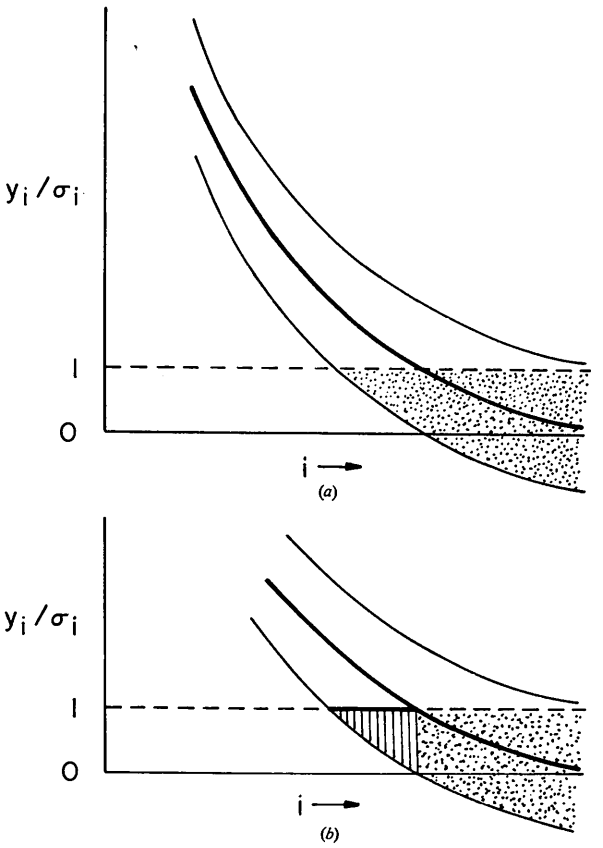

Fig. 1. Schematic distribution of measurable quantities y_i, indexed in order of decreasing y_i^T/σ_i. Heavy curves represent true values y_i^T, lighter curves bound likely ranges of observed values. (*a*) weak terms ($y_i < \sigma_i$) rejected; (*b*) weak terms assigned threshold values $y_i^m = \sigma_i$.

each of which may be made large enough, with reasonable counting times, to approach a normal error distribution. Their difference, whatever its magnitude, will then likewise have normally distributed errors – Hamilton (1964) has emphasized that this is theoretically and practically advantageous but not strictly essential – and be eminently suited, with no preliminary tampering, as input data to a least-squares refinement. Statistical integrity demands that all measured reflexions be treated impartially, whether F_o^2 is greater or smaller than $\sigma(F^2)$ or even negative. An algebraic demonstration of the propriety of faithfully including the net observed intensity in least-squares calculations even when it happens to be negative has been given by Schomaker (1969). (To reject negative values of $I-B$ may be regarded as another instance of the imposition on the experimental data of constraints drawn from our model rather than from the observational evidence.)

Preserving the normal distribution

If the expected errors in the intensity data are indeed normally distributed, all care should be taken to keep them so, including the choice to refine on F^2 rather than on F. (It is not possible for both F and F^2 to have normal error distributions when $F/\sigma(F)$ is small.) In this respect the exclusion, whether absolute or contingent, of weak reflexions merely because they are weak is far more damaging than simply reducing the number of observations. When we selectively reject weak intensities, we discard more underestimated than overestimated reflexions and so destroy the assumed symmetry in the error distributions of the surviving data. The resulting bias will inevitably lead to systematic error in those parameters, principally the thermal parameters, that depend strongly on the average intensities of many weak reflexions.

Suppose we have a set of measurable quantities y_i – which might stand for F^2 – whose experimental errors are normally distributed with standard deviations σ_i. In fact the only property of the normal distribution we care about here is its symmetry, *i.e.* that positive and negative errors of equal magnitude are equally likely. Ordering our data by decreasing y_i^T/σ_i, where y_i^T is the true value of y_i, we might expect a distribution such as that illustrated in Fig. 1, in which the hypothetical assembly of discrete points has been idealized to a continuous thick curve. This is bracketed between two parallel thinner curves, drawn to represent the expected ranges of observed values $y_i^o = y_i^T + \delta_i$, where the errors δ_i are depicted as contained largely between the limits $\pm\sigma_i$.

One regrettable, but not uncommon, practice is to drop all measurements giving $y_i^o < \sigma_i$. These are represented by the shaded region in Fig. 1(*a*). Throughout a range of data corresponding roughly to $y_i^T < 2\sigma_i$ the accepted values are seen to include a preponderance of positive errors. A least-squares refinement, which should produce calculated values y^c tightly

clustered about the true values y^T, will instead tend to make y^c, in the affected range, systematically larger than y^T. Of course, deletion of a large proportion of the least accurate data will invariably improve the average agreement between y^o and y^c. This effect may account for the popularity of the habit, since lower agreement factors are easier to recognize than systematic errors in the parameters.

An alternative scheme, closely akin to that recommended above for photographic diffraction data, is to replace y^o by $y^m=\sigma$ whenever $y^o<\sigma$ but to include such a term in the least-squares equations only if y^c exceeds y^m. Analysis of this procedure is more complex, largely because the selection of admitted reflexions depends on the variable quantities y^c. But we can imagine a hypothetical refinement cycle in which the initial values of all parameters happen to equal their true values. We should then (assuming a perfect model) have all $y_i^c=y_i^T$ and the situation would be as shown schematically in Fig. 1(*b*). In the range $y^T<\sigma$ all terms with $y^o<\sigma$ are again rejected. But for $y^T>\sigma$, terms with $y^o<\sigma$ are now included with, however, y^o adjusted upward to $y^m=\sigma$. The resulting bias is qualitatively the same as in the former scheme but somewhat mitigated in severity. On the final refinement cycle, which is the only one that ultimately matters, we expect that many values of y^c, in the range of interest, will have risen above y^T in an attempt to match the biased sample of y^o. To the extent that the model parameters favour parallel variations in the values of neighbouring y_i^c – the way shifts in thermal parameters raise or lower many F_c values in unison – this may move a proportion of the y^m terms from the rejected to the accepted category, but Fig. 1(*b*) should still describe fairly the overall effect.

Clearly, setting the qualifying hurdle, as many do, at some multiple of σ rather than σ itself is a change in detail, not in substance.

Numerical test

The predicted effects are readily demonstrated by a simple numerical experiment. For purposes of illustration we have constructed a hypothetical crystal structure containing a single atom at the origin of a cubic unit cell of 1 Å edge. Selecting the 50 independent lowest-angle reflexions, from 100 to 533, we have calculated structure factors using a constant scattering factor $f=1$ as for a point charge (or a neutron-scattering length $b=1$) and an isotropic thermal parameter $U=0{\cdot}0036$ Å^2, chosen arbitrarily to give a ratio near 20:1 between the largest and the smallest structure amplitudes. The calculated structure factors define our true values F_T^2. A random-number routine was then used to generate 40 independent sets of random 'errors', for all 50 reflexions, having, verifiably, normal probability distributions and standard deviations $\sigma(F^2)=0{\cdot}1F_T$. (The actual moments of the 2000 values of $t_i=\delta_i/\sigma_i$ were: $\langle t\rangle=-0{\cdot}037$, $\langle t^2\rangle=1{\cdot}006$, $\langle t^3\rangle=-0{\cdot}104$, $\langle t^4\rangle=3{\cdot}019$, as against the ideal values: 0, 1, 0, 3.)

The resulting 40 sets of data, each comprising simulated values of F_o^2 for the 50 reflexions, were successively presented as input to a least-squares routine for the refinement of two parameters: the scale factor k and the isotropic thermal parameter U. Five refinement cycles were applied to each data set, minimizing the residual

$$\Delta=\sum w(F_o^2-k^2F_c^2)^2$$

with $w=\sigma^{-2}=100/F_T^2$. All 40 refinements were performed according to three alternative recipes. Scheme (*a*) omitted from the summation for Δ all reflexions with $F_o^2<\sigma(F^2)$, producing the situation depicted in Fig. 1(*a*). Scheme (*b*) corresponded to Fig. 1(*b*), assigning to reflexions with $F_o^2<\sigma(F^2)$ threshold values $F_m^2=\sigma(F^2)$ but including them in Δ only when $k^2F_c^2>F_m^2$. Scheme (*c*) accepted all 50 observational data as given.

The outcome of the experiment is summarized in Table 1. For each of the three schemes we list: the average number n of reflexions included in the final refinement cycle; the average discrepancy indices (with summation over included reflexions only) $R=\sum||F_o|-kF_c|/\sum|F_o|$ and $r=(\Delta/\sum wF_o^4)^{1/2}$; the root-mean-square figure of merit $d=[\Delta/(n-2)]^{1/2}$; the average refined values of the parameters k and U; and their estimated standard deviations and correlation coefficient. For these last quantities, we compare averages of the conventional estimates derived from the inverse least-squares matrix $\mathbf{A}^{-1}$, *i.e.* $\sigma_\mu=(A^{\mu\mu})^{1/2}d$ and $\varrho(\mu,\nu)=A^{\mu\nu}/(A^{\mu\mu}A^{\nu\nu})^{1/2}$, with direct estimates based on the sample variances of the final parameter values from the 40 independent 'experiments', *i.e.* $\sigma'_\mu=[\sum(\mu-\bar{\mu})^2/(40-1)]^{1/2}$ and $\varrho'(k,U)=\sum(k-\bar{k})(U-\bar{U})/[\sum(k-\bar{k})^2\sum(U-\bar{U})^2]^{1/2}$.

Table 1. *Comparison of three refinement schemes applied to the same* 40 *sets of 'experimental' data*

We use the notation $\langle x\rangle=\bar{x}$ for the arithmetic mean of x_i, $\{x\}$ for the root-mean-square value $\langle x^2\rangle^{1/2}$.

Scheme	*a*	*b*	*c*	'True' values
$\langle n\rangle$	35·7	42·4	50·0	
$\langle R\rangle\times10^3$	95	103	130	
$\{r\}\times10^3$	207	217	265	
$\{d\}$	0·938	0·902	1·006	
$\langle k\rangle\times10^4$	9664	9765	9974	10000
$\{\sigma(k)\}\times10^4$	273	260	293	
$\sigma'(k)\times10^4$	313	298	291	
$\langle U\rangle\times10^6$Å^{-2}	3336	3432	3611	3600
$\{\sigma(U)\}\times10^6$Å^{-2}	131	121	141	
$\sigma'(U)\times10^6$Å^{-2}	132	118	124	
$\langle\varrho(k,U)\rangle$	0·76	0·75	0·74	
$\varrho'(k,U)$	0·87	0·83	0·78	

As the Table confirms, rejection of weak reflexions has only a cosmetic virtue; it produces smaller discrepancy indices R and r. In column (*a*) we note that rejection of all reflexions with $F_o^2<\sigma(F^2)$ has caused a

systematic underestimation of both k and U by amounts exceeding their apparent standard deviations. Scheme (*b*) is, as expected, not quite as bad as (*a*) but similarly faulty. On the other hand, admission, in scheme (*c*), of all 50 data, including those – an average of 5·3 per set – with negative values of F_o^2, leads to average values of k and U in close agreement with their 'true' values.

Examination of the individual refinements shows the same pattern in the converged values of k and U repeated for each of the 40 data sets; scheme (*a*) consistently yields the lowest values for both parameters, scheme (*c*) the highest. This is just as Fig. 1 predicts. The more weak reflexions we exclude, the greater the bias tending to raise the calculated values F_c^2 of the weak reflexions. In our model these are all high-angle reflexions and the best way to increase their magnitudes is to decrease U. Because of the strong positive correlation between k and U [$\varrho(k, U)$ is consistently near 0·75] a drop in U entails a concomitant drop in k.

The estimated standard deviations $\sigma(k)$ and $\sigma(U)$ show no appreciable differences among the three schemes tested. Insofar as schemes (*a*) and (*b*) appear to give slightly smaller standard deviations, as deduced from the least-squares matrix, than scheme (*c*), R. E. Marsh (personal communication) has noted that this is mainly an artefact due to multiplication by the questionable factor d, which is artificially depressed below unity by the selective exclusion, or 'adjustment', of underestimated reflexions. Accordingly, this trend in the least-squares estimates of $\sigma(k)$ and $\sigma(U)$ is not reflected in the corresponding 'experimental' estimates σ' from the sample variances. In any case, the observed variations are too small and irregular to suggest any clear difference among the three recipes either in the apparent magnitudes of the standard deviations or in the validity of the least-squares estimates of these standard deviations for predicting the expected scatter in the parameter values from replicated experiments. The mischief caused by biased observational data shows up mainly in systematic, not random, errors in the derived parameters.

Conclusions

Without doubt, our computer experiment is highly artificial in enabling us to know both F_T and $\sigma(F^2)$ exactly and to verify directly the normal distribution of our experimental errors. In a real experiment, we never know F_T, we customarily depend on a rough estimate of $\sigma(F^2)$, and we content ourselves with a pious hope that our observational errors are not too abnormally distributed. But none of these difficulties justify the gratuitous affront of an *a posteriori* selection among the measured values of F_o^2. Should technical limitations or economic pressure dictate the wholesale elimination of weak reflexions, fairness to the observational data demands that the discrimination, on the last cycle, should be based on F_c, not on F_o.

Our chosen model is unrealistic also in its utter triviality. In the more typical structure with many atoms in the asymmetric unit, weak reflexions are not confined to high angles and the vulnerability of the several parameters, both positional and vibrational, to systematic error is less easily predicted. This very unpredictability, however, is a reason for greater, not lesser, care in the proper handling of the input data. This is not to claim that the majority of published refinements, including our own, require urgent reexamination. On the contrary, our limited experience indicates that in real situations the effect of biased data on the structurally interesting parameters is rarely large enough to matter. But it is plainly easier to avoid the error altogether than to detect the exceptional circumstances where it might be significant.

We are grateful for helpful comments by R. A. Young on an earlier draft of this paper. We also benefited from instructive and illuminating correspondence by R. E. Marsh, by F. H. Moore, and by V. Schomaker, who kindly sent us his unpublished notes on the subtraction of background counts in least-squares calculations. And we are glad to acknowledge a small part of our debt to the late W. C. Hamilton, who welcomed our criticism of his own former opinions and even defended our conclusions in public, who led us to relevant background material and did all he could to encourage the prompt publication of our results.

References

ARNOTT, S. (1965). *Acta Cryst.* **18**, 297–298.

CRUICKSHANK, D. W. J., PILLING, D. E., BUJOSA, A., LOVELL, F. M. & TRUTER, M. R. (1961). *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, p. 46, Edited by R. PEPINSKY, J. M. ROBERTSON and J. C. SPEAKMAN. Oxford: Pergamon Press.

DUNNING, A. J. & VAND, V. (1969). *Acta Cryst.* A**25**, 489–491.

HAMILTON, W. C. (1955). *Acta Cryst.* **8**, 185–186.

HAMILTON, W. C. (1964). *Statistics in Physical Science*, p. 127. New York: Ronald Press.

HAMILTON, W. C. (1972). Invited lecture at Summer School on the Experimental Aspects of X-ray and Neutron Single-Crystal Diffraction Methods, Aarhus.

MOORE, F. H. (1972). Paper presented to Ninth International Congress of Crystallography, Kyoto (see abstract, *Acta Cryst.* A**28**, S256).

SCHOMAKER, V. (1969). Unpublished notes.

STOUT, G. H. & JENSEN, L. H. (1968). *X-ray Structure Determination*, p. 183, New York: Macmillan.