Acta Cryst. (1985). A**41**, 113-122

Maximum Entropy Theory

By Alastair K. Livesey

Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, England

and John Skilling*

Department of Applied Mathematics and Theoretical Physics, Silver Street, Cambridge CB3 9EW, England

(*Received* 13 *December* 1983, *accepted* 6 *August* 1984)

Abstract

A maximum entropy method (MEM) map of electron density is obtained by maximizing $S=-\sum p\log p$ subject to whatever data are available. MEM is derived as the only reconstruction technique that is consistent with simple and general requirements. The method is very widely applicable, but, in this paper, attention is focused on the problem of producing electron density maps in crystallography. The entropy formula can also be derived by analogy with a thermodynamic system of quanta, but it is shown that this model can be misleading, and can break down in practice. MEM applied to a different problem related to quantum fluctuations in the thermodynamic model is shown to lead to formulae equivalent to the maximum determinant method. It is argued that direct MEM will produce superior maps.

1. Introduction

Crystallographers wish to produce and use images of atomic assemblies of the highest possible quality. Generally, these appear as maps of electron density, which are then interpreted appropriately. At the highest resolution, a map would be interpreted in terms of bonding and anti-bonding orbitals, at lower resolution as time-averaged atomic positions, and at yet lower resolution as indicating larger structures such as nucleic acids and α-helices. The various techniques of *ab initio* structure determination, phase extension and refinement, and least-squares refinement, are all aimed at producing or improving the quality of a map. Although crystallographers have traditionally viewed phase extension and refinement differently to the *ab initio* problem, we see these as merely different aspects of a continual process of improving the quality of the electron density maps.

There is growing interest in using variational techniques, in particular maximum entropy (Gull & Daniell, 1978), to select a single map for display and subsequent interpretation. Such a map is the 'best' in the sense of maximizing a given functional. Much effort has been expended in the search for suitable improvement formulae, either in reciprocal space (Navaza, Castellano & Tsoucaris, 1983) or in real space (Hoppe & Gassmann, 1968).

We argue that one should *always* select a single 'best' map by maximizing the entropy (Shannon, 1948; Jaynes, 1968). It is an accident of technique that crystallographic observations are made almost entirely in reciprocal space but this should not obscure the fact that the final map is to be displayed and used in real space. Many other types of data, including direct-space measurements, may also be available. Logically, each single datum is an extra constraint on the allowed maps of positive electron density. To avoid inconsistencies, the same map selection criterion should be used regardless of the type of data and of the resolution at which the map is to be interpreted. Several authors (Diamond, 1963; Hosoya & Tokonami, 1967; de Rango, Tsoucaris & Zelwer, 1974; Wilkins, Varghese & Lehmann, 1983) have already introduced the ideas of maximum entropy to crystallography, and Bricogne (1984*a*) has related the maximum entropy principle to notional probability distribution functions of electron density.

In this paper we present our understanding of the maximum entropy principle, and why it should be of use to crystallographers. We also relate it to other approaches and interpretations, and try to clear up some of the confusion that has arisen because, unfortunately, maximum entropy is easy to misapply and difficult to program.

Although this paper was written with the phase problem, as encountered routinely by crystallographers, specifically in mind, we believe the ideas expressed here are completely general and apply to any problem where a positive map or image is produced from limited and noisy data. Consequently, the maximum entropy formalism will need no revision as novel experimental techniques are introduced to help to solve crystal structures.

First, we set up the general problem of reconstructing a map from imperfect data, and present proposed solutions. Then, in § 3, we show that maximizing the Shannon/Jaynes entropy is the only selection procedure that satisfies simple consistency requirements. In § 4 we explore the 'thermodynamic' derivation of

---

* Author for correspondence.

0108-7673/85/020113-10$01.50

the entropy formula from a model of discrete classical quanta. We also point out (§ 5) some of the dangers of taking this model too seriously. In § 6, we show how fluctuations within the thermodynamic model can be characterized by a different expression (Burg, 1967, 1972) for the entropy, and that this is equivalent to the maximum determinant formula (Tsoucaris, 1970). Different entropy formulae give different maps, and we discuss the differences in § 7. Finally, in § 8 we summarize our conclusions and outline the present state of maximum entropy computing.

2. Selecting a feasible map

Solving crystal structures from X-ray diffraction data is a supremely difficult inverse problem. Given a crystal structure in the form of an electron density map $\rho(\mathbf{x})$, a Fourier transform on the appropriate unit cell gives the structure factors

$$F_{\mathbf{h}}=\int \rho(\mathbf{x}) \exp(2\pi i \mathbf{h} \,.\, \mathbf{x}) \,\mathrm{d} V. \tag{1}$$

The inverse problem of calculating ρ from observed structure factors is complicated, not only by inaccuracies in the data and by incomplete measurements (missing values of F), but also by the lack of phase information: diffraction data yield only the amplitude of F. For notational simplicity, we restrict ourselves to the pure problem of totally undetermined phases, though the maximum entropy formalism is equally applicable to extended data sets such as arise from isomorphous replacement.

Of all the possible electron density maps only some are consistent with the measured diffraction data: we call these the 'feasible' maps. Suppose that a structure factor is observed as

$$D_{\mathbf{h}}=|F_{\mathbf{h}}|^2+\alpha\sigma_{\mathbf{h}}, \tag{2}$$

where $\sigma_{\mathbf{h}}$ is the standard deviation of the measurement and α is a random variable, maybe Gaussian, of zero mean and unit variance. The likelihood of observing a particular data set D from a given crystal ρ is

$$\operatorname{prob}(D|\rho)\propto\exp(-\chi^2/2) \tag{3}$$

$$\chi^2=\sum_{\mathbf{h}}(D_{\mathbf{h}}-|F_{\mathbf{h}}|^2)^2/\sigma_{\mathbf{h}}^2. \tag{4}$$

When a large number M of data is observed, values of χ^2 substantially (more than a few $\sqrt{M}$) different from M are exponentially improbable, and would indicate that the map ρ failed to fit the data D. In practical terms, the feasible set is defined by $\chi^2=M$, and Gull & Daniell (1978) illustrated the dangers of failing to satisfy this condition.

Nevertheless, there are still very many feasible maps that are consistent with typical diffraction data. The feasible set allows a range of values for each measured but noisy amplitude, has two degrees of freedom for each unmeasured (complex) structure factor and has a circular degree of freedom for each unknown phase. It is thus high-dimensional with an awkward toroidal topology. The true crystal structure will, of course, be a feasible map, but there is no way of extracting it without further measurements or extra physical insight.

In particularly simple inverse problems, for which the desired 'map' is just a single number ρ, the feasible set can be presented and used directly as a confidence interval, usually of the form $a<\rho<b$. In crystallography, the feasible set is far too large and complicated to be used directly. As a matter of practical necessity, one must present just one feasible map (or at most a few). The selection is simply a pragmatic requirement, which can, in principle, be satisfied in many different ways. The following are three suggestions.

(1) Positivity: ρ must be non-negative throughout the unit cell. Clearly this is physically necessary in crystallography, where ρ represents an electron density. In reciprocal space this constraint is represented by the Harker-Kasper inequalities (Harker & Kasper, 1948) and by the condition that the Karle-Hauptman determinants (Karle & Hauptman, 1950) must be non-negative. However, although inequality constraints reduce the range (and may be easy to program; Papoulis, 1975; Gerchberg, 1974; Saxton, 1980) they do not usually reduce the dimensionality of the feasible set, so that positivity alone does not normally produce a well-defined single feasible map.

To produce a single map it is best to use a variational principle and maximize some functional of ρ (Tikhonov & Arsenin, 1977).

(2) Maximum entropy method (MEM) (Frieden, 1972; Ables, 1974; Gull & Daniell, 1978; Burch, Gull & Skilling, 1983): maximize

$$S=-\int p(\mathbf{x}) \log p(\mathbf{x}) \,\mathrm{d} V, \qquad p(\mathbf{x})=\rho(\mathbf{x})/\textstyle\int\rho \,\mathrm{d} V \tag{5}$$

over the feasible maps. This is the technique we recommend. It favours smooth maps, since the global (unconstrained) maximum of S is given by the uniform map $\rho=$ constant. Where there might be confusion with other formulae sometimes called 'entropy', we shall call S the Shannon/Jaynes entropy.

(3) Maximize

$$B=\int \log \rho \,\mathrm{d} V. \tag{6}$$

We call B the Burg entropy since, with suitable reinterpretation of symbols, B is the entropy of a time series in spectral analysis (Burg, 1967, 1972; Ulrych & Bishop, 1975). This technique is closely related to the maximum determinant method (Tsoucaris, 1970).

Many other physical constraints have been used to help interpret diffraction data. Examples include Sayre's equation (which models atomicity), solvent flattening and non-crystallographic symmetry. In so far as they introduce new and correct knowledge about the electron density, such techniques are clearly beneficial. Indeed, we hope eventually to incorporate such extra knowledge into a suitable generalized maximum entropy algorithm.

3. Why maximum entropy?

MEM is the only variational method consistent with simple general conditions that we consider a selected map ought to obey. A formal mathematical proof has given by Shore & Johnson (1980), and we give a simplified justification in more physical terms. We impose four requirements.

(1) We contend that the form $R(\rho)$ of the variation function should be independent of the type of data being analysed. For example, we wish to use the same function for X-ray diffraction intensities as for blurred direct-space data obtained by electron microscopy, particularly as both types of data can be used together to obtain better results than from either technique alone. With increasing reliance on a variety of measurement techniques, this requirement will become even more important in future.

(2) The units of ρ should not affect the shape of the construction and hence we seek a variation function of the form

$$R = R(p), \qquad p(\mathbf{x}) = \rho(\mathbf{x}) / \textstyle\int \rho \, \mathrm{d}V, \tag{7}$$

which depends only on the relative proportions p of the map ρ.

(3) For a discrete representation of an object $\mathbf{x} = i(i = 1, 2, \ldots, m)$, knowledge about the proportions in one set of cells should not affect the relative proportions in the remaining cells (unless we have prior knowledge of correlations between cells). Suppose that $p_j + p_k$ is fixed by observation, leaving the remaining proportions p_l to be assigned by maximizing R over whatever data are available on cells $l(l \neq j, l \neq k)$. Since the ratio of p_j to p_k does not affect the net amount of proportion to be assigned to the other cells l, the result of maximizing R over the p_l must be unaffected if the ratio p_j/p_k is changed. Accordingly, the gradient $\partial/\partial p_l$ of R over the other cells l (strictly, its projection onto $\sum p_l = \text{constant}$) must remain parallel to its original direction if p_j is increased and p_k correspondingly decreased. Since the result of maximizing R over constraints is determined by the direction of the gradient of R and not the magnitude, we may without loss of generality take the gradient over l to be completely unaffected by the operation $(\partial/\partial p_j - \partial/\partial p_k)$ of increasing p_j and decreasing p_k. A typical component of this gradient (preserving $\sum p$) is obtained by applying

$$(\partial/\partial p_l - \partial/\partial p_m); \qquad l \neq j, k; \qquad m \neq j, k. \tag{8}$$

Thus we have

$$(\partial/\partial p_l - \partial/\partial p_m) R_{jk} = 0 \quad \text{for all } l, m \neq j, k, \tag{9}$$

where $R_{jk} = (\partial R/\partial p_j - \partial R/\partial p_k)$. Hence, $\partial R_{jk}/\partial p_l$ is independent of l for all $l \neq j, k$. A differential $\partial/\partial p_l$ independent of the index l can always be annihilated by adding an appropriate function of $\sum p$ to the function being differentiated. This has no effect on calculations carried out at fixed normalization $\sum p = 1$, so that without loss of generality we may take

$$\partial R_{jk}/\partial p_l = 0 \quad \text{for all } l \neq j, k. \tag{10}$$

Rewriting this as

$$(\partial/\partial p_j - \partial/\partial p_k) \partial R/\partial p_l = 0 \quad \text{for all } j, k \neq l \tag{11}$$

shows that $(\partial/\partial p_j)\,(\partial R/\partial p_l)$ is independent of j for all $j \neq l$. Again without loss of generality, we may use another appropriate function of $\sum p$ to set

$$(\partial/\partial p_j)(\partial R/\partial p_l) = 0 \quad \text{for all } j \neq l. \tag{12}$$

It follows that $\partial R/\partial p_l$ depends only on p_l, so that R must separate into the form

$$R = \sum_l r_l(p_l). \tag{13}$$

In the absence of prior knowledge favouring some cells more than others, we must take the individual functional forms r to be independent of the index l, so that R becomes a symmetric sum

$$R = \sum_l r(p_l). \tag{14}$$

(4) Now consider an object in which the number m of cells factorizes, $m = ab$, or $m = abc$ *etc.* Data about the structure within one subclass (say a) should not impose any structure in any other subclass (such as b or c), unless one has specific prior knowledge of such correlations.

This is easiest to see when the object is two-dimensional, with a and b being the number of cells in the vertical and horizontal directions respectively (Fig. 1). Data constraining the vertical marginals

$$u_i = \sum_{j=1}^{b} p_{ij} \qquad (i = 1, \ldots, a) \tag{15}$$

should tell us nothing about the horizontal structure, and neither should data constraining the horizontal marginals

$$v_j = \sum_{i=1}^{a} p_{ij} \qquad (j = 1, \ldots, b) \tag{16}$$

tell us anything about the vertical structure. Although

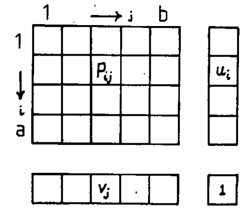

Fig. 1. Marginal sums u and v of a two-dimensional object p.

there will usually be many maps p_{ij} consistent with the given marginals, the *only* map in which the subclasses decouple is

$$p_{ij} = u_i v_j. \tag{17}$$

This requirement allows us to fix the functional form of R. With suitable Lagrange multipliers λ_i for u_i, μ_j for v_j and ν for the normalization $\sum p_{ij} = 1$, the constrained maximum of R occurs at the extremum of

$$\sum_{ij} [r(p_{ij}) - (\lambda_i + \mu_j + \nu) p_{ij}]. \tag{18}$$

Differentiating and setting $p_{ij} = u_i v_j$ yields

$$r'(u_i v_j) = \lambda_i + \mu_j + \nu \qquad \text{for all } i, j. \tag{19}$$

The multipliers λ_i can be written as a function $\lambda(u_i)$ and similarly μ_j as $\mu(v_j)$, so that

$$r'(u_i v_j) = \lambda(u_i) + \mu(v_j) + \nu \quad \text{for all } i, j. \tag{20}$$

If two v_j happen to be close in value, differencing gives

$$u_i r''(u_i v_j) = \mu'(v_j) \quad \text{for all } i, \tag{21}$$

which can only be satisfied in general if r obeys

$$x r''(x) = \text{constant}. \tag{22}$$

This integrates to

$$r(x) = A x \log x + B x + C, \tag{23}$$

giving

$$R(p) = A \sum p \log p + B \sum p + C. \tag{24}$$

Additive constants (including $\sum p = 1$) may be disregarded, and the scale of R is irrelevant. On choosing a minus sign to let R have a maxumum rather than a minimum, we may thus set

$$R(p) = -\sum p \log p = S(p). \tag{25}$$

The flexibility disregarded in the proof above corresponds merely to allowing R to be a monotonic function of S.

We have now shown that any function other than the Shannon/Jaynes entropy S is liable to give misleading results in which data relevant to only one dimension corrupt the reconstruction in other dimensions.

It is easy to check that in general S succeeds in separating independent dimensions. Let the data be (scalar or vector) functions $U(u)$ and $V(v)$ of the marginals u and v. The u structure may be found by maximizing $S(u)$ over the given value of U and normalization. Differentiating the Shannon/Jaynes formula gives

$$\log u_i = -\lambda \partial U / \partial u_i - \alpha - 1 \tag{26}$$

for Lagrange multipliers λ and α. Likewise, the v structure is determined by

$$\log v_j = -\mu \partial V / \partial v_j - \beta - 1. \tag{27}$$

The full map ought to be $p_{ij} = u_i v_j$, obeying

$$\begin{aligned} \log p_{ij} &= \log u_i + \log v_j \\ &= -\lambda \partial U / \partial u_i - \mu \partial V / v_j - \alpha - \beta - 2. \end{aligned} \tag{28}$$

This is indeed exactly what we would reach if we maximized $S(p)$ over U, V and the normalization simultaneously in a direct determination of p. In an optimization problem, constraint differentials combine additively *via* Lagrange multipliers. If individual maps are to combine multiplicatively ($p = uv$), the additive property must relate to the map logarithm

$$\log p = \log u + \log v. \tag{29}$$

Thus the differential $\partial R / \partial p_i$ must behave like $\log p$, so that R takes the '$p \log p$' form.

Of course, we do not suggest that crystallographers observe two-dimensional objects by measuring their marginals. It just happens that such data are relatively easy to analyse. However, suppose a crystallographer had used a variational function to obtain the density of a $P1$ crystal from ($hk0$) and ($00l$) data. He would be disturbed if the density projected down z depended on the ($00l$) data, or if the density projected onto z depended on the ($hk0$) data. Only the '$p \log p$' form succeeds in separating the dimensions. Inconsistency will certainly not disappear just because practical data are more complicated, although its effects will doubtless be more obscure: the only way to avoid it is to use S.

The theorem that S is the only consistent variational function is of fundamental importance because it addresses directly the problem of selecting a single map from the feasible set. By insisting that the variational technique be self-consistent, we have in fact retraced the steps that led Shannon to his proof of the uniqueness of $-\sum p \log p$ as an information measure.

It follows that our derivation is very general. Entropy is a property of *any* vector p_j ($j = 1, \ldots, m$) that is positive and additive. The argument applies equally to probability distributions in coding theory (Shannon, 1948), to optical intensity images (Burch, Gull & Skilling, 1983) and to the electron density in a crystal unit cell.

Our derivation does not rely on any probabilistic interpretation: in fact we were careful to use the word 'proportion' and not 'probability' specifically to avoid this confusion. We are led to *prefer* the maximum entropy map on grounds of consistency, but in this approach we are *not* led to quantify this preference probabilistically. We do not believe that the maximum entropy map is any more likely than any other map that fits the data.

4. Thermodynamic derivation of *S*

The classical derivation of S is Boltzmann's work in statistical thermodynamics. This does have a probabilistic interpretation, and several authors, most recently Bricogne (1984*a*), have carried this over into crystallography, contending that the maximum entropy map is indeed more probable than other maps.

Place N classical quanta at random in the m cells of a map, and suppose that $n_j = Np_j$ quanta are observed to fall in cell j. The probability of obtaining a full pattern $\{n\}$ is proportional to the degeneracy

$$\Omega = N!/\prod n_j! \qquad (30)$$

As N becomes large, Stirling's approximation yields

$$\log \Omega \simeq NS, \qquad (31)$$

where

$$S = -\sum p_j \log p_j \qquad (32)$$

measures the uncertainty in the location of any one quantum within the map. Thus, maximum entropy is equivalent to maximum degeneracy within a physical model in which a map density $\rho(\mathbf{x})$ induces a proportional density of classical quanta. Other probabilistic derivations (Wilkins, Varghese & Lehmann, 1983) do not explicitly require N to become large.

It is easy to take the analogy too seriously. Clearly, N must be finite, even though one uses the continuum formula $-\sum p \log p$, otherwise one would be practically certain to produce an extremely uniform map. Should N be the number of electrons per unit cell? If so, should classical statistics be replaced by Fermi–Dirac statistics? Since the electrons cluster into atoms, should N count the atoms? What if the atoms are bosons? Since the atoms cluster into amino acids (in proteins), should N count the amino acid units? How many unit cells are being observed anyway? Questions like these have often confused discussion of maximum entropy (Kikuchi & Soffer, 1977; Frieden, 1983).

We contend that N does not always have a useful meaning, and that the analogy with thermodynamics can be dangerously misleading. The thermodynamic analogy indicates that exp (NS) is a prior probability prob(p). The rules of probability then require one to modulate this by the likelihood exp $(-\chi^2/2)$ to produce the Bayesian posterior probability distribution

$$\text{prob}\,(\rho|D) \propto \text{prob}\,(D|\rho)\ \text{prob}\,(\rho) \propto \exp\,(NS - \chi^2/2). \qquad (33)$$

For any given N, the maximum entropy map would be the 'most probable' one, which maximizes

$$Q = NS - \chi^2/2. \qquad (34)$$

At this maximum, ρ obeys $2N\,\partial S/\partial p_i = \partial\chi^2/\partial p_i$, so that

$$N = |\nabla\chi^2|/2|\nabla S|. \qquad (35)$$

The parameter N would be chosen so that at this maximum χ^2 would take its correct value close to the number of data. Thus, N would be chosen *a posteriori*, which is not entirely consistent with using exp (NS) as a prior probability.

In any case, the underlying thermodynamic model is often not appropriate to scientific inference. Consider the following counter-argument due to Gull (1983). One is given a mixture of three gases, say A^{40}, Kr^{84}, and ^{132}Xe. The number of atoms N (say 10^{24}) and the mass M (say $80N$ nucleons) are observed. What is the composition of the mixture, in terms of the proportions (p_A, p_{Kr}, p_{Xe}) of atomic populations? The feasible compositions are parameterized as $[1/11+(120/253)\theta,\ 10/11-(10/11)\theta,\ (10/23)\theta]$, where $0 \leqq \theta \leqq 1$. Among these the one with maximum entropy has $\theta = 0{\cdot}6352$. It would, however, be rash to interpret exp (NS) as a probability distribution over θ, because with $N = 10^{24}$ the maximum is very sharply peaked and θ would be effectively determined to within about $\pm 9 \times 10^{-12}$. Obviously, one must not draw such strong conclusions from such limited data. It would only be legitimate to do so if the thermodynamic quanta were chosen from a single exchangeable population. This would be true if the A, Kr and Xe were continually transmuting into each other, and the equilibrium was reproducibly determined by the given constraints N and M (Jaynes, 1957*a*, *b*).

In scientific inference, one is not usually observing such an exchangeable system. For example, a crystallographer is given a specific crystal from the preparation laboratory, from which he proceeds to observe certain parameters such as structure factors. The data fix the feasible set, possibly in terms of the likelihood exp $(-\chi^2/2)$, but that is all. The thermodynamic model contains hidden assumptions that are not applicable to crystallography. It cannot make the maximum entropy solution more probable than other members of the feasible set.

5. Practical difficulties with N

Regardless of any theoretical weaknesses, the use of a model with some number N of quanta may lead to practical difficulties. Specifically, maximizing $Q = NS - \chi^2/2$ does not always give the maximum entropy solution. Consider the one-dimensional three-cell object (ρ_1, ρ_2, ρ_3). Suppose that the two Fourier amplitudes are given exactly by

$$\rho_1 + \rho_2 + \rho_3 = 1 \qquad (36)$$

and

$$\rho_1^2 + \rho_2^2 + \rho_3^2 = 0{\cdot}5, \qquad (37)$$

leaving the single free phase to be determined. On the $\rho_1 + \rho_2 + \rho_3 = 1$ plane, the allowed maps form a circle (Fig. 2). Let the phase be found indirectly through a logarithmic measurement of ρ_1/ρ_2, say

$$\log\,(\rho_1/\rho_2) = -10 \pm \sigma. \qquad (38)$$

Thus,

$$\chi^2 = [\log(\rho_1/\rho_2) + 10]^2/\sigma^2, \qquad (39)$$

whilst the entropy is

$$S = -\rho_1 \log \rho_1 - \rho_2 \log \rho_2 - \rho_3 \log \rho_3. \qquad (40)$$

We seek to maximize S over $\chi^2 = 1$.

As the phase varies between 0 and 2π, it parameterizes the $S(\chi^2)$ curve shown in Fig. 3. Maximizing S at constant χ^2 corresponds to choosing the uppermost curve in Fig. 3.

Maximizing Q, on the other hand, requires us to pre-select a value of N. The loci $Q = \text{constant}$ are straight lines in the (χ^2, S) plane of slope $1/2N$. Q is maximized by the map that lies on the highest such line (*i.e.* largest intercept on the S axis). Larve values of N give us maps along part AB of the trajectory. Small values give us CD. At the critical value of N corresponding to the slope of the chord BC, the solution jumps from B straight to C. Intermediate points BC are not obtained. A similar jump with purely phaseless data was first noticed by Bryan (1980). We give our simpler example, not because it is typical of large crystallographic data sets, but because it can be analysed fully and because it illustrates the danger of using a Lagrange multiplier method (based on N) to solve a non-convex constrained optimization problem.

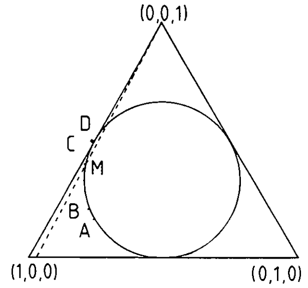

Fig. 2. Maps consistent with given amplitude data lie on the circle. The dashed line is the $\chi^2 = 1$ condition for $\sigma = 6{\cdot}5$, along which the maximum entropy solution is M. A, B, C, D correspond to points in Fig. 3.

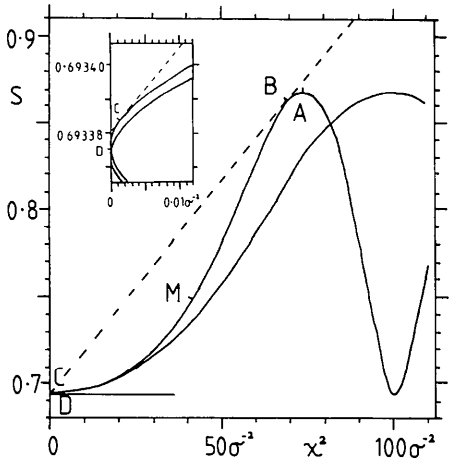

Fig. 3. Entropy S as a function of χ^2, together with supporting chord BC (the insert is an enlargement of the region near C and D). Maximizing Q gives points along AB and along CD only, so that point M cannot be reached.

A physical thermodynamic system constrained by χ^2 (='volume') would normally split into two phases (B = 'gas', C = 'liquid') with proportions determined by the constraint value. When reconstructing a map, this prescription is wrong. It is certainly possible to find a linear combination of map B and map C that fits $\chi^2 = 1$ for any intermediate value of σ: such a map lies along the straight line between B and C in Fig. 2. But this is not a maximum entropy map: it does not maximize S over that (or any other) value of σ and it does not even fit the observed constraints.

Maximizing Q has not worked because not all maximizations can be accomplished by Lagrange multipliers. Only those curves that are globally convex [*i.e.* $d^2S/d(\chi^2)^2 < 0$] can be reliably treated by maximizing Q. All experiments that give data linear in ρ happen to fall into this class, but crystallographic X-ray amplitude data do not.

Another difficulty with maximizing Q is that the feasible set could well be defined not as a probabilistic construct but by rigid YES/NO constrainst on maps. We have been implicitly assuming that the likelihood function $\text{prob}(D|\rho)$ is continuous, and have expressed it as $\exp(-\chi^2)^2$ in terms of normally distributed structure factor errors. If, on the other hand, the errors were rigidly constrained so that $\chi^2 <$ constant, then the likelihood would be discontinuous at the boundary. Its gradient $\nabla\chi^2$ would be undefined at this boundary, and so would the value of N (equation 35). This would make the prior probability $\exp(NS)$ undefined also. We cannot believe that the maximum entropy map should be favoured arbitrarily more or less strongly merely because the data happen to refer to a rigid inequality rather than being expressed in terms of normally distributed errors.

Geometrically, N refers to the relative behaviour of S and χ^2 *normal* to the constraint surfaces, $\chi^2 =$ constant, in the m-dimensional map space. N does not necessarily measure the relative behaviour of S and χ^2 *parallel* to the constraint surface, which is how any probabilistic preference for the maximum entropy map would be quantified. Without N, subsidiary constructions like partition functions (Jaynes, 1957*a*, *b*) also disappear from the formalism. We maintain that the thermodynamic analogy obscures the simplicity of the maximum entropy method. In particular, $\exp(NS)$ does *not* act as a modulating probability, which makes some maps more probable than others. The maximum entropy map is to be *preferred* on grounds of consistency and simplicity, but it is *no more likely* than any other feasible map.

6. Fluctuations and the maximum determinant rule

Let us return to the model in which a large number N of classical quanta (perhaps thought of as atoms) are placed in the m subdivisions of a unit cell. For any set of probabilities p, we have seen that

$$S = -\sum p_j \log p_j \tag{41}$$

measures the uncertainty in the location of any single quantum. Maximizing S makes this location as uncertain as possible, so that the map is as uniform as possible overall.

We can also (Navaza, Castellano & Tsoucaris, 1983) assign a *different* entropy to the *fluctuations* of individual occupation numbers n_j about their mean values

$$\langle n_j \rangle = N p_j. \tag{42}$$

The probability of a particular value (0, 1, 2, ...) for n_j is the Poisson formula

$$\text{prob}\,(n_j) = e^{-\langle n_j \rangle} \langle n_j \rangle^{n_j} / n_j!, \tag{43}$$

which for large N reduces to the Gaussian form

$$\text{prob}\,(n_j) = (2\pi \langle n_j \rangle)^{-1/2} \exp\,[-(n_j - \langle n_j \rangle)^2 / 2\langle n_j \rangle]. \tag{44}$$

The entropy of this distribution is

$$H_j = -\sum_{n_j=0}^{\infty} \text{prob}\,(n_j) \log \text{prob}\,(n_j) \tag{45}$$

$$= \tfrac{1}{2} \log\,(2\pi e \langle n_j \rangle). \tag{46}$$

Correspondingly, the probability of obtaining a full pattern n is

$$\text{prob}\,(n) = \left(\prod_j 2\pi \langle n_j \rangle \right)^{-1/2} \exp\left[-\sum_j (n_j - \langle n_j \rangle)^2 / 2\langle n_j \rangle \right] \tag{47}$$

with entropy

$$H = \tfrac{1}{2} \sum_j \log\,(2\pi e \langle n_j \rangle) \tag{48}$$

$$= \text{constant} + \tfrac{1}{2} \sum_j \log p_j. \tag{49}$$

Apart from constants, this is the Burg entropy $\sum \log p_j$, which we see is intimately connected with fluctuations in the map.

Traditionally, crystallographers have worked almost exclusively in the reciprocal space of structure factor data and phases rather than the direct space of the electron density ρ and its proportions p. For notational simplicity, we choose units such that $F_{000} = \sum \rho = \sum \langle \rho \rangle = 1$, so that $p = \rho$. The probability density of individual maps ρ is

$$\text{prob}\,(\rho) \propto \exp\left[-N \sum_j (\rho_j - \langle \rho_j \rangle)^2 / 2\langle \rho_j \rangle \right] \tag{50}$$

$$\propto \exp\left(-N \sum_j \rho_j^2 / 2\langle \rho_j \rangle \right). \tag{51}$$

In reciprocal space, we may define individual structure factors

$$E_h = \sum_j \rho_j \exp 2\pi i h j \tag{52}$$

and mean unitary structure factors

$$U_h = \sum \langle \rho_j \rangle \exp 2\pi i h j. \tag{53}$$

The probability distribution, now considered as a function of reciprocal space variables, is

$$\text{prob}\,(E) \propto \exp -\tfrac{1}{2} N \sum_{qr} E_q^* (U^{-1})_{qr} E_r, \tag{54}$$

where U^{-1} is the inverse of the circulant matrix $U_{qr} = U_{q-r}$ of structure factors. (To avoid subsidiary indexing, we use one-dimensional notation with all indices ranging from 0 to $m-1$.) The entropy of this, obtained most easily by transforming (48), is

$$H = \text{constant} + \tfrac{1}{2} \log D_m, \tag{55}$$

where $D_m = \det\,[U]$ is the Karle-Hauptman determinant of structure factors U_{qr}. Maximizing the Burg entropy is thus identical to maximizing the corresponding determinant $\det\,[U]$.

Tsoucaris (1970) used the probability distribution (54) in his 'maximum determinant rule' for phase extension. Treating U as fixed, he advised adjusting one or more phases of individual structure factors E in order to minimize the quadratic form

$$Q_m = \sum_{qr} E_q^* (U^{-1})_{qr} E_r \tag{56}$$

and hence maximize the individual probability prob (E). Then the phases of U could be set to the new phases of E, and E re-adjusted. (Technically, he also introduced another determinant Δ_{m+1} in order to avoid explicit matrix inversions in the formula for Q_m.) When the iterative phase adjustment is complete, the most probable individual structure factors E coincide with the assumed values U in both amplitude and phase. This means that there are no extra constraints (incorrect phase assignments) imposed on the model of N random quanta. Just as in a real thermodynamic system, the entropy H will then take its maximum value consistent with actual (measured amplitude) constraints. Tsoucaris's rule thus leads to phase assignments that maximize H and $\det\,[U]$.

Alternatively, the probability distribution (54), or equivalently (51), may be obtained directly as the maximum entropy distribution over variance constraints

$$N \int p(\rho_j) \rho_j^2 \, d\rho_j = \langle \rho_j \rangle \leftarrow \text{data}. \tag{57}$$

This suggests a connection with Sayre's equation

$$N\rho^2 = \rho \tag{58}$$

for N identical point atoms (with the unit cell now subdivided sufficiently finely that the atoms are scattered sparsely with occupation numbers 0 or 1 only). Indeed, Tsoucaris first derived his rule from Sayre's equation, intuitively interpreting the left-hand side $N\rho^2$ in probabilistic terms using fluctuating structure factors E, and the right-hand side ρ in terms of data constraints on mean structure factors U.

Whichever derivation is preferred, the maximum determinant rule is equivalent to maximizing the Burg entropy over measured constraints on the proportions p. This has been noted previously by Britten & Collins (1982), Narayan & Nityananda (1982), Piro (1983) and Steenstrup (1984). Collins & Mahar (1983) also explicitly pointed out that the Shannon/Jaynes and Burg entropies produce different maps. The maximum determinant procedure produces a map that allows the underlying fluctuations to be as large as possible. This would be an appropriate technique if one were interested in individual samples of occupation numbers, because their distribution would then be as wide and non-committal as possible.

We maintain that this is not the crystallographic problem. Crystallographers wish to determine the proportions p of electron density through the unit cell. They do not wish to determine an electron density that maximizes the allowed fluctuations within an underlying model of dubious physical validity.

7. Comparison of maximum entropy with maximum determinant

Since positivity of ρ is built into the Burg entropy, the maximum determinant procedure was a considerable advance over earlier simpler reconstruction techniques. However, it does not give an optimal map.

To demonstrate this, we take a simple example from Nityananda & Narayan (1982) as presented by Skilling & Gull (1984). Suppose we have structure factor data on a unit cube.

$$F_{000} = 1, \qquad F_{100} = F_{010} = F_{001} = 0{\cdot}5, \tag{59}$$

where

$$F_{hkl} = \int_{-1/2}^{1/2} \mathrm{d}V \rho(x, y, z) \exp 2\pi i(hx + ky + lz). \tag{60}$$

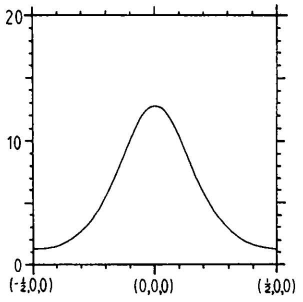

Fig. 4. Cross-section of Shannon/Jaynes map.

This is consistent with many distributions ρ, among them eight point atoms at $(\pm\frac{1}{6}, \pm\frac{1}{6}, \pm\frac{1}{6})$, two point atoms at $\pm(\frac{1}{6}, \frac{1}{6}, \frac{1}{6})$, one point atom at $(0, 0, 0)$ on a uniform background *etc.*

The Shannon/Jaynes solution is straightforward:

$$\rho(\mathbf{x}) = A \exp(\lambda_1 \cos 2\pi x + \lambda_2 \cos 2\pi y + \lambda_3 \cos 2\pi z). \tag{61}$$

To fit the data values, we need $A = 0{\cdot}391$ and $\lambda_1 = \lambda_2 = \lambda_3 = 1{\cdot}161$, as shown in cross section in Fig. 4. Note that the x, y and z data correctly factorize out independently in the final solution.

The Burg solution is less straightforward. Variational maximization of $\int \log \rho \, \mathrm{d}V$ under the constraints yields

$$\rho(\mathbf{x}) = A/(1 - \lambda_1 \cos 2\pi x - \lambda_2 \cos 2\pi y - \lambda_3 \cos 2\pi z). \tag{62}$$

This solution fails to factorize, so that the x data are interfering with y and z, and conversely, but there is also a more serious difficulty. We need $\lambda_1 = \lambda_2 = \lambda_3$ by symmetry, and then the most non-uniform allowed reconstruction ($\rho \geq 0$) has $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$. The map cannot become more non-uniform (larger λ) without becoming negative near the origin. Unfortunately, this critical reconstruction has $F_{100}/F_{000} = 0{\cdot}34$, and it is not possible to fit the required ratio $F_{100}/F_{000} = 0{\cdot}5$ woth any of the purely variational solutions. We recall that not all maximization problems can be solved with Lagrange multipliers.

The variational equation (62) only holds where the solution is differentiable. Beyond the critical ratio of 0·34, the variational solution can no longer accommodate the required non-uniformity without becoming negative over some finite volume. Any extra non-uniformity must appear at the origin (where the variational solution becomes singular) as a non-differentiable δ function condensation. For the required ratio of 0·5, the constrained maximum of B is (Fig. 5)

$$\rho(\mathbf{x}) = \tfrac{1}{2}[1 - \tfrac{1}{3}(\cos 2\pi x + \cos 2\pi y + \cos 2\pi z)]^{-1} + 0{\cdot}242\delta(\mathbf{x}). \tag{63}$$

Almost a quarter of the reconstruction is forced into

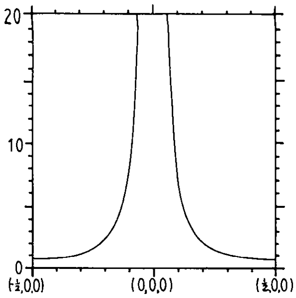

Fig. 5. Cross section of Burg map. This map contains a δ function at the origin, as well as the r^{-2} singularity apparent from the graph.

a single point, without any evidence for this in the data.

This difficulty is likely to recur often with crystallographic data. A Burg reconstruction from structure factor data $F_{\mathbf{h}}$ is the reciprocal of a band-limited function

$$\rho(\mathbf{x}) = 1/\sum_{\mathbf{h}} \lambda_{\mathbf{h}} R_{\mathbf{h}}(\mathbf{x}). \qquad (64)$$

Generally, in three (or more) dimensions, this may *not* be able to fit a set of data $F_{\mathbf{h}}$, even though the data are consistent with smooth positive (though substantially non-uniform) reconstructions. This is because singularities can appear in ρ while the data remain finite. At a small distance r from a zero (minimum because $\rho \geq 0$) of $\sum \lambda_{\mathbf{h}} R_{\mathbf{h}}(\mathbf{x})$, $\rho(\mathbf{x}) = O(r^{-2}) \rightarrow \infty$. This does *not* give a large contribution to the data because these are *volume* integrals $\int \mathrm{d}r\, 4\pi r^2 \ldots \rho$.

δ functions in ρ do not give a large contribution to $\int \log \rho \,\mathrm{d}V$ either, because of the logarithm, so that the Burg entropy imposes no direct penalty on singular reconstructions. For arbitrary data, it may well be that the only way of keeping ρ positive everywhere is to allow part of the solution to condense into δ functions, which do not affect the Burg entropy.

On the other hand, δ functions in ρ give infinite negative contribution to the Shannon/Jaynes form $S = -\int \rho \log \rho \,\mathrm{d}V$, so that they only appear in the solution if the data categorically demand them - in which case they will presumably be correct.

Of course there may be δ functions in the true map. The point is that crystallographers will find spiky reconstructions more difficult to interpret, especially when the spikes are in the wrong places! The maximum determinant (Burg entropy) method gives extra structure to ρ without proper supporting evidence from the data.

8. Conclusions

The data come first. They tell us which maps are consistent with the data, and which are inconsistent and to be discarded. The data have then done all they can.

We have shown that maximizing the Shannon/Jaynes entropy $-\int p \log p \,\mathrm{d}V$ is then the *only* way of selecting a reconstruction that satisfies simple consistency requirements. The argument rested on an absence of prior knowledge favouring any one part of the unit cell over any other: if such knowledge is available, it can and should be inserted as a prior model $m(\mathbf{x})$ in a revised 'relative' entropy

$$S = -\int p \log (p/m) \,\mathrm{d}V \qquad (65)$$

(Jaynes, 1968).

The maximum entropy method will produce a high-quality map of electron density, which, on grounds of simplicity, consistency and general applicability, is to be *preferred* over all others that fit the given data. However, it is *not* any more *likely* than any other. Such claims can only rest on probabilistic calculations based on the thermodynamic analogy with a system of discrete quanta. Not only have we shown that the model based on N quanta can mislead, but also that it can break down completely when nonlinear constraints, such as unphased structure factor amplitudes, are imposed on it.

An analysis of fluctuations within the quantum thermodynamic model leads to consideration of the Burg form of entropy, which is equivalent to the maximum determinant formula. Maps produced by this method are liable to contain sharp features for which there is no compelling evidence in the data. We believe that true (Shannon-Jaynes) maximum entropy maps will prove to be clearer and easier to interpret.

However, we are not yet ready to demonstrate this with realistic crystallographic data. The Cambridge maximum entropy algorithm (Bryan, 1980; Skilling, 1981) was designed principally for linear data, such as phased structure factors. It deals easily with phased Fourier data in astronomical interferometry (Scott 1981) and nuclear magnetic resonance spectroscopy (Sibisi, 1983; Sibisi, Skilling, Brereton, Laue & Staunton, 1984); with convolution data in optics (Daniell & Gull, 1980; Burch, Gull & Skilling, 1983) and high-energy astronomy (Skilling, Strong & Bennett, 1983); with tomographic data (Kemp, 1980) in medicine and elsewhere; with extended X-ray absorption fine structure data (Livesey, 1984) and with other linear data.

Maximum entropy reconstructions have been achieved from two-dimensional phaseless Fourier data (Skilling, 1983). In some of the most powerful three-dimensional work so far, Bryan, Bansal, Folkhard, Nave & Marvin (1983) have developed the Cambridge algorithm to reconstruct an image of a filamentous bacterial virus using incomplete fibre diffraction data. Other work has been reported by Wilkins (1983) and by Collins (1982), and Bricogne (1983, 1984*b*) has presented impressive work on three-dimensional data for crambin. However, a robust program for crystallographic data has yet to be written.

We owe a great debt to Professor E. T. Jaynes, whose clear and profound papers on entropy (collected as Jaynes, 1983) have deeply influenced our thinking. It is a pleasure to thank him, Dr S. F. Gull and the numerous members of the Cambridge maximum entropy group, and Dr P. H. Gaskell for many enlightening discussions over the years. It is with equal pleasure that we thank the organizers and participants of the 1982 CECAM Paris meeting and the 1983 Medical Foundation of Buffalo meeting for encouraging our interest in crystallographic direct methods.

Two referees have given our paper their careful attention, and we acknowledge their constructive and helpful comments. One of us (AKL) also thanks St John's College, Cambridge, and the Cavendish Laboratory Trust Fund for support.

References

Ables, J. G. (1974). *Astron. Astrophys. Suppl. Ser.* **15**, 383-393.

Bricogne, G. (1983). Paper presented at Medical Foundation of Buffalo (NY) School on Direct Methods and Macromolecular Crystallography, and private communication.

Bricogne, G. (1984*a*). *Acta Cryst.* **A40**, 410-445.

Bricogne, G. (1984*b*). *Methods and Applications in Crystallographic Computing*, edited by T. Ashida & S. Hall. Oxford Univ. Press. in the press.

Britten, P. L. & Collins, D. M. (1982). *Acta Cryst.* **A38**, 129-132.

Bryan, R. K. (1980). PhD thesis. Univ. of Cambridge, England.

Bryan, R. K., Bansal, M., Folkhard, W., Nave, C. & Marvin, D. A. (1983). *Proc. Natl Acad. Sci. USA*, **80**, 4728-4731.

Burch, S. F., Gull, S. F. & Skilling, J. (1983). *Comput. Graph. Image Process.* **23**, 113-128.

Burg, J. P. (1967). Paper presented at 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City.

Burg, J. P. (1972). *Geophysics*, **38**, 375-376.

Collins, D. M. (1982). *Nature* (*London*), **298**, 49-51.

Collins, D. M. & Mahar, M. C. (1983). *Acta Cryst.* **A39**, 777-783.

Daniell, G. J. & Gull, S. F. (1980). *Proc. Inst. Electr. Eng.* **127E**, 170-172.

Diamond, R. (1963). *Acta Cryst.* **16**, 627-639.

Frieden, B. R. (1972). *J. Opt. Soc. Am.* **62**, 511-518.

Frieden, B. R. (1983). *J. Opt. Soc. Am.* **73**, 927-938.

Gerchberg, R. W. (1974). *Opt. Acta*, **21**, 709-720.

Gull, S. F. (1983). Paper presented at the 3rd Workshop on Maximum Entropy, Univ. of Wyoming.

Gull, S. F. & Daniell, G. J. (1978). *Nature* (*London*), **272**, 686-690.

Harker, D. & Kasper, J. S. (1948). *Acta Cryst.* **1**, 70-75.

Hoppe, W. & Gassmann, J. (1968). *Acta Cryst.* **B24**, 97-107.

Hosoya, S. & Tokonami, M. (1976). *Acta Cryst.* **23**, 18-25.

Jaynes, E. T. (1957*a*). *Phys. Rev.* **106**, 620-630.

Jaynes, E. T. (1957*b*). *Phys. Rev.* **108**, 171-190.

Jaynes, E. T. (1968). *IEEE Trans. Syst. Sci. Cybern.* SSC-**4**, 227-241.

Jaynes, E. T. (1983). (Collected Works.) *Papers on Probability, Statistics, and Statistical Physics*, edited by R. D. Rosenkrantz. Dordrecht, Holland: Reidel.

Karle, J. & Hauptman, H. (1950). *Acta Cryst.* **3**, 181-187.

Kemp, M. C. (1980). *Med. Radionuclide Imaging* (*IAEA, Vienna*), **1**, 313-323.

Kikuchi, R. & Soffer, B. H. (1977). *J. Opt. Soc. Am.* **67**, 1656-1665.

Livesey, A. K. (1984). PhD thesis, Univ. of Cambridge, England. Submitted.

Narayan, R. & Nityananda, R. (1982). *Acta Cryst.* **A38**, 122-128.

Navaza, J., Castellano, E. E. & Tsoucaris, G. (1983). *Acta Cryst.* **A39**, 622-630.

Nityananda, R. & Narayan, R. (1982). *J. Astrophys. Astron.* **3**, 419-450.

Papoulis, A. (1975). *IEEE Trans. Circuits Syst.* CAS-**22**, 735-742.

Piro, O. E. (1983). *Acta Cryst.* **A39**, 61-67.

Rango, C. de, Tsoucaris, G. & Zelwer, C. (1974). *Acta Cryst.* **A30**, 342-353.

Saxton, W. O. (1980). In *Computer Processing of Electron Microscope Images*, edited by P. W. Hawkes. *Topics in Current Physics* 13, pp. 35-87. Berlin: Springer-Verlag.

Scott, P. F. (1981). *Mon. Notices R. Astron. Soc.* **194**, 23P-29P.

Shannon, C. E. (1948). *Bell Syst. Tech. J.* **27**, 379-423, 623-656.

Shore, J. E. & Johnson, R. W. (1980). *IEEE Trans. Inf. Theory*, IT-**26**, 26-37; correction: *IEEE Trans.* (1983). IT-**29**, 942-943.

Sibisi, S. (1983). *Nature* (*London*), **301**, 134-136.

Sibisi, S., Skilling, J., Brereton, R. G., Laue, E. D. & Staunton, J. (1984). *Nature* (*London*), **311**, 446-447.

Skilling, J. (1981). Paper presented at 1st Workshop on Maximum Entropy Estimation and Data Analysis, Univ. of Wyoming.

Skilling, J. (1983). Paper presented at Opt. Soc. Am. meeting on Signal REcovery and Synthesis with Incomplete Information and Partial Constraints, Incline Village, Nevada.

Skilling, J. & Gull, S. F. (1984). *SIAM* (*Soc. Ind. Appl. Math.*) *Proc. Am. Math. Soc.* **14**, 167-189.

Skilling, J., Strong, A. W. & Bennett, K. (1979). *Mon. Notices R. Astron. Soc.* **187**, 145-152.

Steenstrup, S. (1984). *Acta Cryst.* **A40**, 66-67.

Tikhonov, A. & Arsenin, V. (1977). *Solutions of Ill-posed Problems.* London: Wiley.

Tsoucaris, G. (1970). *Acta Cryst.* **A26**, 492-499.

Ulrych, T. J. & Bishop, T. N. (1975). *Rev. Geophys. Space Phys.* **13**, 183-200.

Wilkins, S. W. (1983). *Acta Cryst.* **A39**, 892-895, 896-898.

Wilkins, S. W., Varghese, J. N. & Lehmann, M. S. (1983). *Acta Cryst.* **A39**, 47-60; erratum: *Acta Cryst.* (1983), **A39**, 594.

Acta Cryst. (1985). **A41**, 122-128

Round-off Errors in Inter-experimental Comparisons

By Robin Taylor and Olga Kennard

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

(*Received* 6 *April* 1984; *accepted* 17 *September* 1984)

Abstract

Two independent determinations of the same structure may be compared by means of statistical techniques such as normal probability plots and χ^2 hypothesis tests. Computer simulations show that errors may arise in the application of these techniques if rounded estimates of structural parameters and their e.s.d.s are used in the calculations. Round-off errors are particularly serious in goodness-of-fit hypothesis tests, since they increase the probability of making type I errors, *i.e.* falsely rejecting the null hypothesis.

0108-7673/85/020122-07$01.50