Acta Cryst. (1984). A**40**, 705-708

Scaling by Entropy Maximization

BY DOUGLAS M. COLLINS

*Department of Chemistry, Arhus University, DK-8000 Arhus C, Denmark and Department of Chemistry, Texas A & M University, College Station, Texas, 77843, USA**

(*Received* 24 *January* 1984; *accepted* 3 *July* 1984)

Abstract

Entropy maximization is an alternative to least-squares or squares minimization for the relative scaling of intersecting data sets. The iterative calculations converge rapidly, even for difficult cases, and the scaling is truly relative in contrast to that of other widely used methods.

1. Introduction

The relative scaling of X-ray intensities from photographic records is customarily based on the method of Hamilton, Rollett & Sparks (1965). Their least-squares method solved the major computational problems which existed at that time. By its general acceptance the method has provided a standard approach for scaling calculations that is based on the uniquely useful principle of least squares.

Entropy maximization is another principle which can be used in relative scaling calculations. Although the method presented here is rooted in information theory and its iterative formalism has no apparent similarity to that of Hamilton, Rollett & Sparks (1965; HRS hereafter), its results provide excellent numerical corroboration of those from HRS. The similarity of results for the two very different approaches strengthens the claim of each as a correct procedure. The present method based on entropy maximization, although it applies only to positive numbers, appears to be more general. Two notable points are that the calculations do not require intervention to avoid singularities, and the scale factors are truly relative, that is, multiplying all scale factors by an arbitrary positive number does not change any average of scaled intensities. Certain characteristics of the entropy-maximization formalism are also present in the logarithmic scaling procedures of Nordman (1960) and Rae (1965; Rae & Blake, 1966).

2. Formalism

In order to bring the scaling problem into the framework of information theory, each observation is interpreted as the prior probability of a particular event. It may be said that the collection of possible events corresponds to a communications apparatus for which the task at hand is its adjustment to make any mismatch of actual and prior distributions carry as little information as possible. This task exactly parallels the discovery of die unfairness by a maximum-entropy criterion (Jaynes, 1979). In the present application the engineering analogy turns into a determination of scale factors by entropy maximization.

The entropy which measures the difference between a (normalized) prior distribution m and another (normalized) distribution p is formulated (Jaynes, 1968; Collins, 1982) as

$$H = -\sum_j p_j \ln [p_j/m_j]. \tag{2.1}$$

An intensity identified by h and observed on film i is denoted I_{hi}. The corresponding average intensity is J_h and the scale factor G_i puts J_h on the scale of I_{hi}. Thus, (2.1) may be put in the form

$$Q = -\sum_{h,i} (G_i J_h)' \ln [(G_i J_h)'/(I_{hi})'], \tag{2.2}$$

where the prime denotes normalization. Q is to be maximized by setting

$$\partial Q/\partial G_i = 0. \tag{2.3}$$

Straightforward algebraic manipulation results in

$$\ln G_i = A - \sum_h \frac{J_h}{\left(\sum_h J_h\right)} \ln \frac{J_h}{I_{hi}}, \tag{2.4}$$

where

$$A = \sum_{h,i} \frac{G_i J_h}{\left(\sum_{h,i} G_i J_h\right)} \ln \frac{G_i J_h}{I_{hi}}. \tag{2.5}$$

Note that both summations in (2.4) include only those terms for which there is an observation of intensity from film i. The proper form for computation of the best maximum-entropy value for J is found by setting

$$\partial Q/\partial J_h = 0. \tag{2.6}$$

* Permanent address.

0108-7673/84/060705-04$01.50

Straightforward algebraic manipulation results in

$$\ln J_h = A - \sum_i \frac{G_i}{\left(\sum_i G_i\right)} \ln \frac{G_i}{I_{hi}}, \qquad (2.7)$$

where A is given by (2.5). Note that both summations in (2.7) include only those terms for which a film carries an observation of intensity I.

Although it is not certain how weights are to be incorporated in entropy expressions (Jaynes, 1968), inclusion of customary weights can be rationalized as follows. Expand the logarithm in (2.1) to get

$$H = -\sum_j p_j[(p_j - m_j)/p_j + \tfrac{1}{2}(p_j - m_j)^2/p_j^2 + \ldots];$$

$$p_j/m_j > \tfrac{1}{2}. \qquad (2.8)$$

If as desired $p/m \simeq 1$, the higher-order terms in (2.8) may be ignored to give

$$H \simeq -\sum_j (p_j - m_j) + \tfrac{1}{2}(p_j - m_j)^2/p_j, \qquad (2.9)$$

and when $(p-m)$ is well distributed about zero

$$H \simeq -\sum_j \tfrac{1}{2}(p_j - m_j)^2/p_j. \qquad (2.10)$$

Clearly, within the approximations (*cf.* Jaynes, 1979) maximization of entropy has the same character as minimization of squares.

Two useful results follow immediately from the form of (2.10). The parallel in maximum-entropy methods to the diagonal weight matrix or ordinary term-by-term weighting of least-squares methods is again term-by-term weighting. Secondly, to make each term in H of equivalent value in some average sense, the weight for a value of p should be proportional to

$$w = p/\sigma^2(p). \qquad (2.11)$$

Because of the normalization in (2.2), (2.4), (2.5) and (2.7), the scale of weights is inconsequential. In these equations every term is to be weighted with

$$w_{hi} = I_{hi}/\sigma^2(I_{hi}) \qquad (2.12)$$

for calculations including weights.

Equations (2.4), (2.5) and (2.7) constitute an iterative solution to the scaling problem. A singular property of this set of equations is that multiplication of the scale factors by an arbitrary positive constant has no effect on the average intensities; the scale factors are necessarily independent of the intensity scale as well.

3. Examples from HRS

In the following tests the true relativity of scale is used to set a scale factor at 1·0 after each iteration. Also, after each iteration, the overall intensity scale is set independently by a least-squares criterion to correspond with the scale of the observations. To get started the scale factors were set to make the average intensity the same for each film.

Table 1. *Data for HRS, example* 1

	Film 1	Film 2	Film 3
I_1	100	2	—
I_2	—	1	2
I_3	100	3	—

Table 2. *Data for HRS, example* 2

	Film 1	Film 2	Film 3	Film 4
I_1	100	2	—	—
I_2	—	1	2	—
I_3	100	3	—	—
I_4	—	—	1	4
I_5	25	—	—	1

The pathological data sets of HRS, restated in Tables 1 and 2, were used without change. Four sets of calculations were carried out to obtain scale factors and best maximum-entropy values of J for each of the two examples of HRS, both with and without weights. Weighting is introduced in (2.2)–(2.7) through multiplication of every $G_i J_h$ or I_{hi} by w_{hi}. The weighting scheme used was

$$\sigma(I) \sim I;\ w \sim 1/I. \qquad (3.1)$$

The lack of weights, that is, unit weighting, corresponds to

$$\sigma(I) \sim I^{1/2};\ w = \text{a constant}. \qquad (3.2)$$

Formal error estimation was carried out by evaluation of the r.m.s. deviations

$$\sigma(J_h) = [\langle (J_h - I_{hi}K_i)^2 \rangle]^{1/2}, \qquad (3.3)$$

and

$$\sigma(K_i) = [\langle (K_i - J_h/I_{hi})^2 \rangle]^{1/2}, \qquad (3.4)$$

where $K \equiv G^{-1}$, and the angle brackets signify an average value.

Fifteen iterations for each of the four sets of calculations showed no divergence. It will be clear that formally estimated r.m.s. deviations are of uncertain significance when the data are so few and incongruous as in these tests. In any case, by a criterion of all subsequent shifts in K less than $0{\cdot}1\sigma(K)$, convergence was achieved in an average of five cycles; the worst case was convergence in seven cycles. The results of the tests are given in Tables 3 and 4.

4. Discussion

The HRS method for scaling together intersecting data sets has been successful and satisfactory for the crystallographic community. The fixing of one scale factor was the principal basis for objections by Fox

Table 3. *Results of scaling for HRS, example* 1

Formal r.m.s. deviations are given in parentheses.

	Weighted	Unweighted	HRS
I_1	4·36 (53)	4·85 (4)	5·00
I_2	1·99 (1)	1·99 (1)	2·00
I_3	5·34 (45)	4·90 (3)	5·00
K_1	0·0490 (49)	0·0488 (2)	0·050
K_2	2·00 (19)	1·99 (38)	2·00
K_3	1·000 (4)	1·000 (3)	1·000
Q^*	$-0{\cdot}34\times10^{-2}$	$-0{\cdot}47\times10^{-3}$	—

* Entropy from the appropriately weighted equation (2.2)

Table 4. *Results of scaling for HRS, example* 2

Formal r.m.s. deviations are given in parentheses.

	Weighted	Unweighted	HRS
I_1	5·95 (35)	5·65 (3)	6·00
I_2	3·59 (100)	3·61 (120)	3·83
I_3	6·97 (73)	5·70 (5)	6·00
I_4	2·83 (94)	3·49 (60)	3·80
I_5	1·12 (44)	1·39 (4)	1·50
K_1	0·0630 (79)	0·0568 (3)	0·060
K_2	3·22 (77)	2·59 (69)	2·56
K_3	2·54 (73)	2·51 (86)	2·54
K_4	1·00 (33)	1·00 (17)	1·00
Q^*	$-0{\cdot}20\times10^{-1}$	$-0{\cdot}19\times10^{-2}$	—

* Entropy from the appropriately weighted equation (2.2).

& Holmes (1966) and by Monahan, Schiffer & Schiffer (1967). The latter authors presented a system of equations which allows for variation of all scale factors without explicit constraint and in each iteration provides a value for a scale factor itself rather than a scale-factor increment. Although the three papers differ significantly in their final equations, they are three expressions of the basic method set out by HRS (Monahan, Schiffer & Schiffer).

In contrast to the HRS and cognate methods, the present method yields truly relative scale factors. This can be seen by inspection of (2.4), (2.5) and (2.7). If one supposes that their iterative evaluation has been completed, then it is clear that multiplication of all scaled intensities by an arbitrary positive constant leaves the scale factors unchanged. Moreover, the scale factors may be multiplied by an arbitrary positive constant and the scaled intensities are unaffected. The overall scale of the intensities and the scale of the scale factors are completely independent. This independence does not characterize the logarithmic scaling of Nordman (1960) although it is independent of overall scale with respect to adjustment of a scale-factor logarithm.

It is important that entropy maximization yields essentially the same results as the squares minimization of HRS. Nevertheless, the entropy maximization procedure appears superior in that it does not require any intervention, either to guard against unreasonable results, or to set or reset any value. Computational economy is probably not a significant issue for scaling calculations, but it is worth noting that the maximum-entropy calculations converge very rapidly. To be definite, a test was set which involved 27 data, each recorded twice for a total of 54 observations in two sets with a mean difference of ~2%. The scale of one set was misset by an order of magnitude. It was returned to its correct value in one iteration. Against any advantage in convergence speed, the maximum-entropy method does require the evaluation of two logarithms in addition to normal arithmetic operations for each observation in every iteration.

Robust/resistant procedures work well over a broad class of error distributions (robust) and are not strongly influenced by any small subset of data (resistant) (Nicholson, Prince, Buchanan & Tucker, 1982). It seems clear on the basis of the examples given in the preceding section that the entropy-maximization procedure is well described as robust. Following Nicholson *et al.* we consider that the procedure is resistant in that large discrepancies between observations and modeled equivalents are de-emphasized in (2.4), (2.5) and (2.7). The de-emphasis follows from the comparison of observed and modeled data through the logarithm of their ratio so long as observations of extraordinarily small magnitude are not used. Anisotropic scaling as described by Rossmann, Leslie, Abdel-Meguid & Tsukihara (1979) or other similarly exacting calculations would appear to benefit from use of entropy maximization rather than least squares which is neither robust nor resistant (Nicholson, Prince, Buchanan & Tucker).

Nielsen (1977) gave a maximum-entropy method for optimizing weights in least-squares analysis. It is precisely in the spirit of his method to assert that the entropies calculated from (the appropriately weighted) (2.2) for different weighting schemes may be used to judge among them. By the criterion of maximum entropy the unweighted calculations summarized in Tables 3 and 4 are decisively favored because the entropies are an order of magnitude more negative for the alternatives.

Entropy maximization is a completely workable alternative to squares minimization for the scaling problem. Although the procedure was constructed to deal with data records which are at least potentially interpretable as positive-definite probability distributions, the only practical limitation is that the data all be nonzero magnitudes. In fact the method can be made completely general through replacement of magnitudes by suitably parameterized probabilities which reflect the proper range of a random variable. The generalizations will be taken up in future papers.

This work has been supported in part by the Robert A. Welch Foundation through grant A-742 and by the Research Corporation through a Cottrell research grant.

References

COLLINS, D. M. (1982). *Nature (London)*, **298**, 49–51.
FOX, G. C. & HOLMES, K. C. (1966). *Acta Cryst.* **20**, 886–891.
HAMILTON, W. C., ROLLETT, J. S. & SPARKS, R. A. (1965). *Acta Cryst.* **18**, 129–130.
JAYNES, E. T. (1968). *IEEE Trans.* SSC-4, pp. 227–241.
JAYNES, E. T. (1979). In *The Maximum Entropy Formalism*, edited by R. D. LEVINE & M. TRIBUS, pp. 15–118. Cambridge: Massachusetts Institute of Technology.
MONAHAN, J. E., SCHIFFER, M. & SCHIFFER, J. P. (1967). *Acta Cryst.* **22**, 322.
NICHOLSON, W. L., PRINCE, E., BUCHANAN, J. & TUCKER, P. (1982). In *Crystallographic Statistics: Progress and Problems*, edited by S. RAMASESHAN, M. F. RICHARDSON & A. J. C. WILSON, pp. 230–363. Bangalore: Indian Academy of Science.
NIELSEN, K. (1977). *Acta Cryst.* A**33**, 1009–1010.
NORDMAN, C. E. (1960). *Acta Cryst.* **13**, 535–539.
RAE, A. D. (1965). *Acta Cryst.* **19**, 683–684.
RAE, A. D. & BLAKE, A. B. (1966). *Acta Cryst.* **20**, 586.
ROSSMANN, M. G., LESLIE, A. G. W., ABDEL-MEGUID, S. S. & TSUKIHARA, T. (1979). *J. Appl. Cryst.* **12**, 570–581.

Acta Cryst. (1984). A**40**, 708–712

Comparison of Methods of Matching Protein Structures

BY C. E. KENKNIGHT

Biochemistry Department, Biosciences West, University of Arizona, Tucson, AZ 85721, *USA*

(*Received* 24 *February* 1984; *accepted* 3 *July* 1984)

Abstract

Two fast methods of superposing two sets of atomic coordinates by least-squares refinement are described and related to two earlier fast methods. A Newton method is applied to rotations of a 3×3 outer product matrix used previously by Ferro & Hermans [*Acta Cryst.* (1977), A**33**, 345–347] and by McLachlan [*Acta Cryst.* (1972), A**28**, 656–657]. Three of the methods work better if one molecule has its inertial matrix aligned with *xyz*. A Newton–Gauss method that rotates the coordinates can converge rapidly after a rough orientation using three strategic atoms. The average superposition takes about 0·003 s on a Cyber 175 with the best method, rotations about the *xyz* axes in turn. Experience with reliability is reported for large residuals.

Introduction

This paper describes experience with several methods for calculating the rigid-body rotations that are needed for matching similar molecular structures. Two new methods are presented and compared with published methods. In applications to proteins there is a systematic search for a likeness between a fragment of structure anywhere in one protein, *A*, to any part of a second protein, *B* (Rao & Rossmann, 1973; Rossmann & Argos, 1976, 1977; Remington & Matthews, 1978; McLachlan, 1979). A typical search involves more than a million structure matches, so a fast method is essential. The frequency of matches that do not reach the global minimum for the least-squares search is also of concern for interpretation of supposed likenesses.

Three of the methods under discussion here are, in principle, equivalent. Thus, aspects of the theory given next are like that given by McLachlan (1972) and by Ferro & Hermans (1977). Performances of the associated algorithms are not equivalent and effort was directed to understanding why not. Running times were studied for a variety of conditions: magnitude of the residuals, magnitude of the relative rotations, orientation of one of the coordinate sets, and use of several tricks to speed or ensure convergence. The theory set forth here was helpful in understanding what gave minimal running times.

The problem and the Newton method

Let $\mathbf{a}_k$, $\mathbf{b}_k$ ($k=1$ to N) be the position vectors of two sets of N atoms from the molecular fragments *A* and *B*. Let w_k be a weight for each atom. We want to minimize the residual *E*, an inner product,

$$E=\tfrac{1}{2}\sum w_k(R\mathbf{a}-\mathbf{b})'_k(R\mathbf{a}-\mathbf{b})_k. \qquad (1)$$

Here the prime signifies a transpose so that $\mathbf{a}'$ is a row vector. If we were to interpret w_k as the strength of a linear spring joining the atoms numbered k (McLachlan, 1982), then *E* would have the interpretation of a potential energy. Such a system is static if the net force and torque on, say, *A* due to *B* vanishes. The vanishing force requires that the centroids of *A* and *B* coincide (McLachlan, 1972; Remington & Matthews, 1978) while the vanishing torque requires that the weighted vector cross product of the structures vanish:

$$\mathbf{g}=\sum w_k(\mathbf{b}_k\times\mathbf{a}_k)=0. \qquad (2)$$

Thus it is reasonable that a unique orthogonal proper rotation matrix *R* with determinant +1 exists (see McLachlan, 1979), which transforms $\mathbf{a}$ referred to the centroid as origin to $\mathbf{r}=R\mathbf{a}$ and minimizes the residual

0108-7673/84/060708-05$01.50