RUFFA, A. (1963). *Phys. Rev.* **130**, 1412–1423.
RUFFA, A. (1967). *J. Chem. Phys.* **47**, 1874.
SCHMIDT, P. C. & WEISS, A. (1979). *Z. Naturforsch. Teil A*, **34**, 1471–1481.
SCHNEIDER, J. R., HAUSEN, N. K. & KRETSCHMER, H. (1981). *Acta Cryst.* A**37**, 711–722.
SHARMA, V. C. (1974*a*). *Acta Cryst.* A**30**, 278–280.
SHARMA, V. C. (1974*b*). *Acta Cryst.* A**30**, 299–300.
SHARMA, V. C. (1975). *Acta Cryst.* A**31**, 157.
SILVERMAN, J. N. & OBATA, Y, (1963). *J. Chem. Phys.* **38**, 1254–1255.
SLATER, J. C. (1974). *The Self-Consistent Field for Molecules and Solids – Quantum Theory of Molecules and Solids*, Vol. 4. New York: McGraw-Hill.
SUORTTI, P. & JENNINGS, L. D. (1977). *Acta Cryst.* A**33**, 1012–1027.
SYSIÖ, P. A. (1969). *Acta Cryst.* B**25**, 2374–2378.
TESSMAN, J. R., KAHN, A. H. & SHOCKLEY, W. (1953). *Phys. Rev.* **92**, 890–895.
TOGAWA, S., INKINEN, O. & MANNINEN, S. (1971). *J. Phys. Soc. Jpn*, **30**, 1132–1135.
TOSI, M. P. & FUMI, F. G. (1964). *J. Phys. Chem. Solids*, **25**, 45–52.
WATSON, R. E. (1958). *Phys. Rev.* **111**, 1108–1110.
WILLIAMS, A. R., KÜBLER, J. & GELATT, C. D. (1979). *Phys. Rev. B*, **19**, 6094–6118.
WILLIS, B. T. M. & PRYOR, A. W. (1975). *Thermal Vibrations in Crystallography*. Cambridge Univ. Press.
YAMASHITA, J. (1952). *J. Phys. Soc. Jpn*, **7**, 284–286.
YAMASHITA, J. & ASANO, S. (1970). *J. Phys. Soc. Jpn*, **28**, 1143–1150.
YODER, D. R. & COLELLA, R. (1982). *Phys. Rev. B*, **25**, 2545–2549.
ZUNGER, A. & FREEMAN, A. J. (1977). *Phys. Rev. B*, **16**, 2901–2926.

Acta Cryst. (1983). A**39**, 407–410

A Test of a Robust/Resistant Refinement Procedure on Synthetic Data Sets

BY E. PRINCE

National Measurement Laboratory, National Bureau of Standards, Washington, DC 20234, USA

AND W. L. NICHOLSON

Battelle Pacific Northwest Laboratories, Richland, WA 99352, USA

(*Received* 31 *August* 1982; *accepted* 6 *December* 1982)

Abstract

The conventional crystallographic least-squares procedure has been compared with a robust/resistant modification in which the weight of each reflection is multiplied by a function of the ratio of its residual to a resistant measure of the width of the residual distribution on the previous cycle. Three synthetic data sets were created by adding random errors, according to various probability distributions, to the calculated structure factors for a known crystal structure. A set with a Gaussian error distribution was refined with two sets of weights: one assigned correctly in proportion to the reciprocals of the variances of the data points, the other using unit weights throughout. The second error distribution was Gaussian contaminated by 10% drawn from another Gaussian distribution with its variance nine times greater. The third distribution was a long-tailed distribution derived by dividing a random variable with a Gaussian distribution by an independent random variable with a uniform distribution. Each of the first three cases was refined to convergence using both conventional and robust/resistant procedures, with the modified procedure leading to a result at least as close to the known structure as the conventional procedure. In the fourth case, the conventional procedure gave a poor fit, but the robust/resistant procedure converged to a reasonable approximation to the correct structure.

Introduction

In a previous paper (Nicholson, Prince, Buchanan & Tucker, 1982) we have described the application of a robust/resistant (hereafter designated R/R) refinement algorithm to refinement of the multiple data sets collected from L-(+)-tartaric acid (formerly known as D(+)-tartaric acid) in the International Union of Crystallography's Single Crystal Intensity Project (Abrahams, Hamilton & Mathieson, 1970). The procedure proved to be a very efficient means of separating from the data sets small numbers of data points which were inconsistent with the body of the data, and convergence was thereby achieved for several of the data sets in which least-squares (LS) refinement was unstable in the previous study carried out by Hamilton & Abrahams (1970). In addition, some, but not all, of the variability in refined parameters from the LS refinement was removed.

0567-7394/83/030407-04$01.50

The results of the study on the L-(+)-tartaric acid data provided strong evidence that the alternative procedure is 'resistant', *i.e.*, that it is insensitive to variations in small subsets of the data. To show that it is also 'robust', *i.e.*, that it precisely estimates the correct model over a wide range of conditions where the error distributions are not Gaussian and the weights are not proportional to the reciprocals of the variances of the data points, would require an exact knowledge of the true structure parameters, knowledge that is not available in an experimental situation. In order to study the robustness of the procedure, therefore, we have made use of several synthetic data sets in which the 'observed' structure factors were actually the calculated structure factors for a known model to which random errors had been added according to various probability distributions.

Creation of data sets

Three synthetic data sets were created; one of them was refined using two different weighting schemes, making four cases altogether. The starting point in each case was a set of 233 calculated structure factors corresponding to the refined structure of ammonium azide (Prince & Choi, 1978). For case I a random-number generator was used to generate a list of numbers that had a Gaussian distribution with zero mean and unit variance. These numbers were multiplied by 2·5% of the calculated F and then added to the calculated F to produce a list of 'observed' F's with a Gaussian error distribution. In case I each value of F received a weight equal to $1/(0{\cdot}025F)^2$. In case II the list of synthetic observed F's was identical to case I, but all F's were assigned unit weights. For case III 10% of the errors were multiplied by three before adding to the calculated F's. Finally, in case IV each number in the Gaussian list was divided by another random number drawn from a distribution that is uniform over the range from 0 to 1. This produces a probability density function $\Phi(y) = (2\pi)^{-1/2}[1 - \exp(-y^2/2)]/y^2$. It has long tails similar to the Cauchy (or Lorentzian) distribution, but its density near the middle is less sharply peaked than is a Cauchy distribution. In cases III and IV, the weights used were the same as those in case I.

Refinement procedure

Each of the four cases was refined twice, using a modification of the least-squares refinement program *RFINE4* (Finger & Prince, 1975). The starting model was the structure from which the data set was derived. The function minimized was $\sum w(|F_o| - |F_c|)^2$. In each case the refinement was first run with fixed weights, as in conventional least squares. The refinement was then repeated using the R/R algorithm. In practice this amounts to modifying the weights, in each cycle after the first, by $W_i' = W_i\,\varphi(R_i/aS)$, where $R_i = W_i^{1/2}(|F_{oi}| - |F_{ci}|)$, a is a constant chosen to exclude the most extreme data, and S is a resistant measure of scale. For the weight-modifying function, we have used Tukey's (1974) 'biweight' function, defined by $\varphi(x) = (1 - x^2)^2$ for $|x| \leq 1$; $\varphi(x) = 0$ for $|x| > 1$. The measure of scale for the $(j + 1)$th cycle is given by a formula suggested by Huber (1973),

$$S_{j+1} = \left((1/\beta) \sum_{i=1}^{n} \{\varphi[R_i(\boldsymbol{\theta}_j)/aS_j]\, R_i(\boldsymbol{\theta}_j)\}^2/(n-p)\right)^{1/2},$$

where n is the number of data points, p is the number of parameters, and $\boldsymbol{\theta}_j$ designates the vector of estimated parameters used in the jth cycle. β is the expected value of $z^2\,\varphi(z)$ if Z is distributed according to the true error-distribution function. If the error distribution is Gaussian and $a = 6$, then $\beta = 0{\cdot}72767$.

The LS refinement of case I was carried through seven cycles to ensure complete convergence. That the LS procedure gives results close to those expected for this idealized case is shown by the fact that the

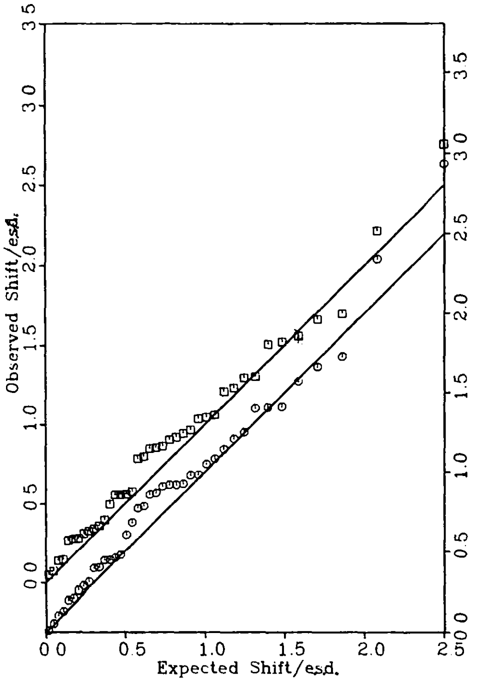

Fig. 1. A half-normal quantile–quantile plot of the observed absolute shifts of the refined parameters from the initial structure, in each case divided by the estimated standard deviation, for a synthetic data set with a Gaussian error distribution. The squares are for a conventional least-squares refinement, and the ordinate is labeled on the left. The circles are for a robust/resistant refinement, and the ordinate is labeled on the right. Straight lines passing through the origin with unit slope are shown for comparison.

weighted agreement index, R_w, is 0·022 and the estimated standard deviation of an observation of unit weight is 0·99.

Among the 40 parameters in the model, the sensitivity of the calculated structure factors to variations in the parameters varies widely from parameter to parameter. In order to put all shifts on a common scale the estimated standard deviation of each parameter as calculated in the LS refinement of case I was taken as a measure of relative precision for that parameter for all refinements. A half-normal quantile–quantile plot (Abrahams & Keve, 1971) of the ordered absolute differences between the refined parameters and their 'correct' values divided by the standard deviation is shown in Fig. 1 for case I. The fact that almost all of the points in this plot lie close to the line passing through the origin with unit slope is further confirmation that the LS procedure gives results close to those expected in the idealized case, in spite of the fact that the linearized model is only an approximation to the true one.

The other seven refinements were carried through four cycles, satisfactory convergence being achieved in every case, with no shift in the final cycle greater than 0·05 of a standard deviation. Table 1 is a summary of the agreement indices of the eight cases. For the R/R refinement, the estimated standard deviation of an observation of unit weight, S, is computed from Huber's formula, given above.

Discussion of results

Figs. 1 through 4 are half-normal quantile–quantile plots (Abrahams & Keve, 1971) of the differences between the refined parameters and the 'correct' parameters. The ordinate of each plotted point is $|x_r - x_c|/s_x$, where s_x is the estimated standard deviation of parameter x calculated from the LS refinement of case I and x_r and x_c are, respectively, the refined and correct values of each parameter. The values of this quantity for the 40 parameters are

Table 1. *Summary of agreement index information for refinement of various data sets*

N is the number of reflections included in the R/R refinement. The number in the LS refinement was 233 for all cases.

E.s.d. denotes the estimated standard deviation of an observation of unit weight. For the R/R refinement, S is the estimated standard deviation computed using Huber's formula.

	LS			R/R			
Case	R	R_w	e.s.d.	N	R	R_w	S
I	0·018	0·022	0·99	233	0·018	0·021	1·00
II	0·018	0·020	3·06	233	0·018	0·019	2·82
III	0·022	0·028	1·22	233	0·022	0·025	1·17
IV	0·055	0·130	5·77	224	0·031	0·034	1·66

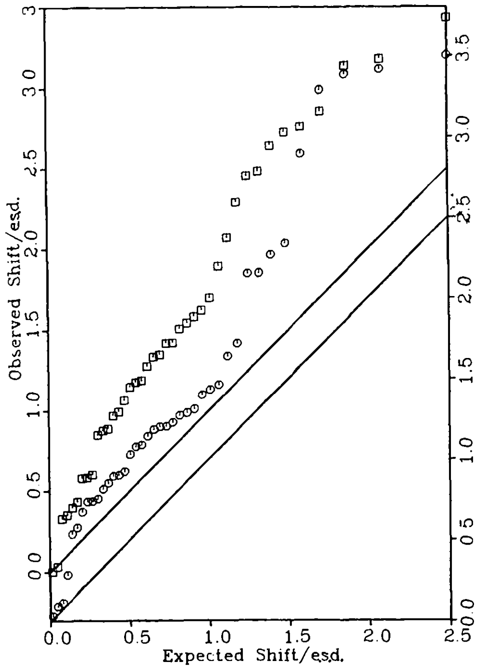

Fig. 2. A half-normal quantile–quantile plot, as in Fig. 1, for least-squares and robust/resistant refinements for the data set with Gaussian errors but with unit weights used throughout. The squares are for the LS refinement, and the circles are for the R/R refinement.

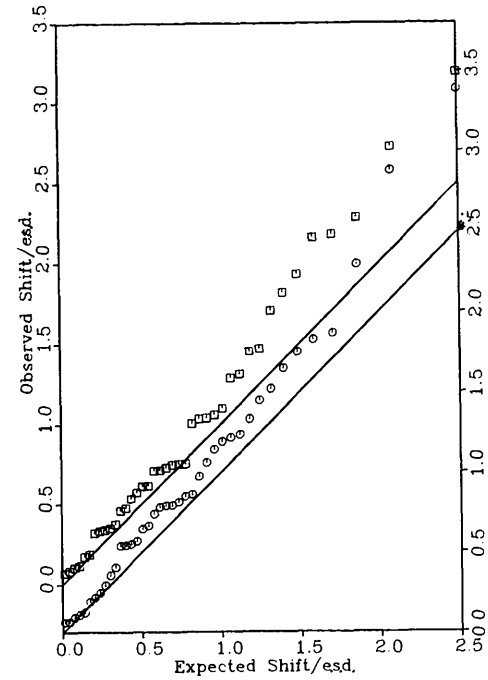

Fig. 3. A half-normal quantile–quantile plot, as in Fig. 1, for least-squares and robust/resistant refinements for a data set in which a Gaussian error distribution is contaminated by 10% of another Gaussian distribution with a standard deviation three times greater. The squares are for the LS refinement, and the circles are for the R/R refinement.

arranged in ascending order. The abscissa for the *i*th point is the value x_i for which $F(x_i) = (1/2) + (2i - 1)/80$, where $F(x)$ is the cumulative Gaussian distribution function. If the model is linear and the weights are properly assigned, this plot should be a straight line with zero intercept and unit slope. The ordinates for the LS and R/R refinements are displaced for clarity.

As can be seen in Fig. 1, in case I the R/R procedure gives results which are virtually identical to those given by LS with fixed weights. Fig. 2 shows, however, that the distortion of the error distribution caused by using unit weights throughout produced a noticeable increase in the discrepancy between the refined structure and the 'correct' structure. Using the robust/resistant procedure, however, the effect of the initially 'wrong' weights is partially compensated and the results are closer to the assumed model for most parameters. Case III, shown in Fig. 3, represents the sort of error distribution which may be fairly common in real experimental data, *i.e.* most data points are good, but a small minority are influenced by some unmodelled effect that introduces random errors with a broader distribution. In this case, also, the iterative R/R procedure gives a result closer to the known model. Case IV is an example of a pathological error distribution that rarely, if ever, appears in a real experiment. For the LS procedure there is a large difference between the refined model and the starting model. The R/R procedure, however, converged rapidly to a model that is recognizably similar to the starting model.

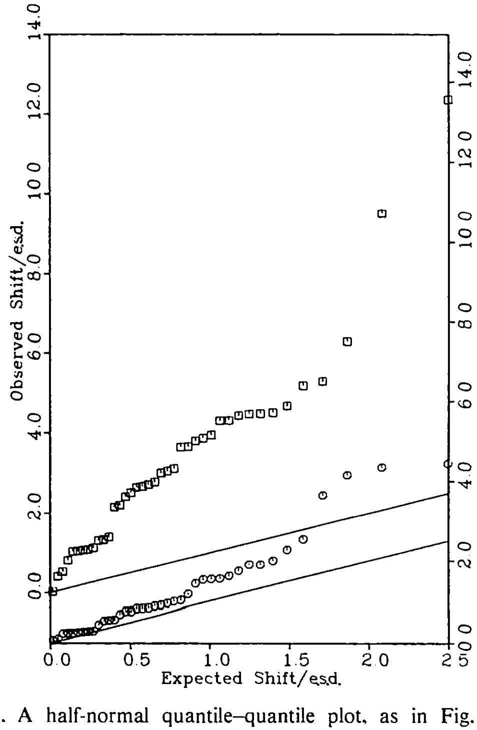

Fig. 4. A half-normal quantile–quantile plot, as in Fig. 1, for least-squares and robust/resistant refinements for a data set with a long-tailed 'slash' error distribution. Conventional least squares works poorly for long-tailed distributions. The squares are for the LS refinement, and the circles are for the R/R refinement.

While this study is not the sort of exhaustive Monte Carlo calculation that would be necessary to prove conclusively the robust character of the robust/resistant procedure, it does suggest that the modified procedure can be expected to give results as good as the conventional least-squares process in the refinement of good data and markedly better results for the refinement of data sets that contain large deviations with frequencies appreciably greater than are expected in a Gaussian error distribution. With real data the various factors contributing to the variance of the observations are never completely known, so that the ideal conditions for the use of least squares are rarely present. A possible practical procedure is to refine the structure by both methods and compare the results. If the results agree, it gives some confidence that the data are good and the structure is reliable. If the results do not agree, this study indicates that the R/R procedure is more likely to lead to a reliable result. For many scientifically interesting substances, it is difficult to obtain single crystals of the high quality necessary for collecting accurate data, and this procedure can be very useful in extracting meaningful structural information. However, the results must be examined with great care to make sure that the lack of fit in the conventional least-squares procedure is really due to a long-tailed error distribution rather than to an important deficiency in the model.

References

Abrahams, S. C., Hamilton, W. C. & Mathieson, A. McL. (1970). *Acta Cryst.* A**26**, 1–18.

Abrahams, S. C. & Keve, E. T. (1971). *Acta Cryst.* A**27**, 157–165.

Finger, L. W. & Prince, E. (1975). *RFINE4. A System of Fortran IV Computer Programs for Crystal Structure Computations, Nat. Bur. Stand. Tech. Note* 854. National Bureau of Standards, Washington, DC, USA.

Hamilton, W. C. & Abrahams, S. C. (1970). *Acta Cryst.* A**26**, 18–24.

Huber, P. J. (1973). *Ann. Stat.* **1**, 799–821.

Nicholson, W. L., Prince, E., Buchanan, J. A. & Tucker, P. E. (1982). *Crystallographic Statistics: Progress and Problems*, edited by S. Ramaseshan, M. F. Richardson & A. J. C. Wilson. Bangalore: Indian Academy of Sciences.

Prince, E. & Choi, C. S. (1978). *Acta Cryst.* B**34**, 2606–2608.

Rogers, W. H. & Tukey, J. W. (1972). *Stat. Ned.* **26**, 211–226.

Tukey, J. W. (1974). *Critical Evaluation of Chemical and Physical Structural Information*, pp. 3–14. National Academy of Sciences, Washington, DC, USA.