*International Tables for Crystallography* (2006). Vol. C, Section 8.2.2, pp. 689–691.

# 8.2. Other refinement methods

By E. Prince and D. M. Collins

Chapter 8.1 discusses structure refinement by the method of least squares, which has a long history of successful use in data fitting and statistical analysis of results. It is an excellent technique to use in a wide range of practical problems, it is easy to implement, and it usually gives results that are straightforward and unambiguous. If a set of observations, $y_i$, is an unbiased estimate of the values of model functions, $M_i(\mathbf{x})$, a properly weighted least-squares estimate is the best, linear, unbiased estimate of the parameters, $\mathbf{x}$, provided the variances of the p.d.f.s of the populations from which the observations are drawn are finite. This assumes, however, that the model is correct and complete, an assumption whose validity may not necessarily be easily justified. Furthermore, least squares tends to perform poorly when the distribution of errors in the observations has longer tails than a normal, or Gaussian, distribution. For these reasons, a number of other procedures have been developed that attempt to retain the strengths of least squares but are less sensitive to departures from the ideal conditions that have been implicitly assumed. In this chapter, we discuss several of these methods. Two of them, maximum-likelihood methods and robust/resistant methods, are closely related to least squares. A third one uses a function that is mathematically related to the entropy function of thermodynamics and statistical mechanics, and is therefore referred to as the maximum-entropy method. For a discussion of the particular application of least squares to structure refinement with powder data that has become known as the Rietveld method (Rietveld, l969), see Chapter 8.6.

## 8.2.1. Maximum-likelihood methods

In Chapter 8.1, structure refinement is presented as finding the answer to the question, ‘given a set of observations drawn randomly from populations whose means are given by a model, $\boldsymbol{M}(\mathbf{x})$, for some set of unknown parameters, $\mathbf{x}$, how can we best determine the means, variances and covariances of a joint probability density function that describes the probabilities that the true values of the elements of $\mathbf{x}$ lie in certain ranges?’. For a broad class of density functions for the observations, the linear estimate that is unbiased and has minimum variances for all parameters is given by the properly weighted method of least squares. The problem can also be stated in the slightly different manner, ‘given a model and a set of observations, what is the *likelihood* of observing those particular values, and for what values of the parameters of the model is that likelihood a maximum?’. This set of parameters is the *maximum-likelihood estimate*.

Suppose the $i$th observation is drawn from a population whose p.d.f. is $\Phi_i(\Delta_i)$, where $\Delta_i = [y_i - M_i(\mathbf{x})]/s_i$, $\mathbf{x}$ is the set of ‘true’ values of the parameters, and $s_i$ is a measure of scale appropriate to that observation. If the observations are independent, their joint p.d.f. is the product of the individual, marginal p.d.f.s:

$$\Phi_J(\Delta) = \prod_{i=1}^{n} \Phi_i(\Delta_i). \qquad (8.2.1.1)$$

The function $\Phi_i(\Delta_i)$ can also be viewed as a conditional p.d.f. for $y_i$ given $M_i(\mathbf{x})$, or, equivalently, as a likelihood function for $\mathbf{x}$ given an observed value of $y_i$, in which case it is written $l_i(\mathbf{x}|y_i)$. Because a value actually observed logically must have a finite, positive likelihood, the density function in (8.2.1.1) and its logarithm will be maximum for the same values of $\mathbf{x}$:

$$\ln[l(\mathbf{x}|\mathbf{y})] = \sum_{i=1}^{n} \ln[l_i(\mathbf{x}|y_i)]. \qquad (8.2.1.2)$$

In the particular case where the error distribution is normal, and $\sigma_i$, the standard uncertainty of the ith observation, is known, then

$$\Phi_i(\Delta_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-(1/2)\{[y_i - M_i(\mathbf{x})]/\sigma_i\}^2\right), \qquad (8.2.1.3)$$

and the logarithm of the likelihood function is maximum when

$$S = \sum_{i=1}^{n} \{[y - M_i(\mathbf{x})]/\sigma_i\}^2 \qquad (8.2.1.4)$$

is minimum, and the maximum-likelihood estimate and the least-squares estimate are identical.

For an error distribution that is not normal, the maximum-likelihood estimate will be different from the least-squares estimate, but it will, in general, involve finding a set of parameters for which a sum of terms like those in (8.2.1.2) is a maximum (or the sum of the negatives of such terms is a minimum). It can thus be expressed in the general form: find the minimum of the sum

$$S = \sum_{i=1}^{n} \rho(\Delta_i), \qquad (8.2.1.5)$$

where $\rho$ is defined by $\rho(x) = -\ln[\Phi(x)]$, and $\Phi(x)$ is the p.d.f. of the error distribution appropriate to the observations. If $\rho(x) = x^2/2$, the method is least squares. If the error distribution is the Cauchy distribution, $\Phi(x) = [\pi(1 + x^2)]^{-1}$, $\rho(x) = \ln(1 + x^2)$, which increases much more slowly than $x^2$ as $|x|$ increases, causing large deviations to have much less influence than they do in least squares.

Although there is no need for $\rho(x)$ to be a symmetric function of $x$ (the error distribution can be skewed), it may be assumed to have a minimum at $x = 0$, so that $\mathrm{d}\rho(x)/\mathrm{d}x = 0$. A series expansion about the origin therefore begins with the quadratic term, and

$$\rho(x) = (x^2/2)\left(1 + \sum_{k=1}^{\infty} a_k x^k\right). \qquad (8.2.1.6)$$

This procedure is thus equivalent to a variant of least squares in which the weights are functions of the deviation.

## 8.2.2. Robust/resistant methods

Properly weighted least squares gives the best *linear* estimate for a very broad range of distributions of random errors in the data and the maximum-likelihood estimate if that error distribution is normal or Gaussian. But the best linear estimator may nevertheless not be a very good one, and the error distribution may not be well known. It is therefore important to address the question of how good an estimation procedure may be when the conditions for which it is designed may not be satisfied. Refinement procedures may be classified according to the extent that they possess two properties known as robustness and resistance. A procedure is said to be *robust* if it works well for a broad range of error distributions and *resistant* if its results are not strongly affected by fluctuations in any small subset of the data. Because least squares is a linear estimator, the influence of any single data point on the parameter estimates increases without limit as the difference between the observation and the model increases. It therefore works poorly if the actual error distribution contains large deviations with a frequency that substantially exceeds that expected from a normal distribution. Further, it has the undesirable property that it will make the fit of a few wildly discrepant data points better by making the fit of many points a little worse. Least squares is therefore neither robust nor resistant.

Tukey (1974) has listed a number of properties a procedure should have in order to be robust and resistant. Because least squares works well when the error distribution is normal, the procedure should behave like least squares for small deviations whose distribution is similar to the normal distribution. It should de-emphasize large differences between the model and the data, and it should connect these extremes smoothly. A procedure suggested by Tukey was applied to crystal structure refinement by Nicholson, Prince, Buchanan & Tucker (1982). It corresponds to a fitting function $\rho(\Delta)$ [equation (8.2.1.5)] of the form

$$\begin{aligned}\rho(\Delta) &= (\Delta^2/2)(1-\Delta^2+\Delta^4/3) \qquad |\Delta|<1,\\ \rho(\Delta) &= 1/6 \qquad\qquad\qquad\qquad\quad |\Delta|\geq 1,\end{aligned} \tag{8.2.2.1}$$

where $\Delta_i = [y_i - M_i(\mathbf{x})]/s_i$, and $s$ is a resistant measure of scale.

In order to see what is meant by a resistant measure, consider a large sample of observations, $y_i$, with a normal distribution. The sample mean,

$$\bar{y} = (1/n)\sum_{i=1}^{n} y_i, \tag{8.2.2.2}$$

is an unbiased estimate of the population mean. Contamination of the sample by a small number of observations containing large, systematic errors, however, would have a large effect on the estimate. The median value of $y_i$ is also an unbiased estimate of the population mean, but it is virtually unaffected by a few contaminating points. Similarly, the sample variance,

$$s^2 = [1/(n-1)]\sum_{i=1}^{n}(y_i-\bar{y})^2, \tag{8.2.2.3}$$

is an unbiased estimate of the population variance, but, again, it is strongly affected by a few discrepant points, whereas $[0.7413r_q]^2$, where $r_q$ is the *interquartile range*, the difference between the first and third quartile observations, is an estimate of the population variance that is almost unaffected by a small number of discrepant points. The median and the interquartile range are thus resistant quantities that can be used to estimate the mean and variance of a population distribution when the sample may contain points that do not belong to the population. A value of the scale parameter, $s_i$, for use in evaluating the quantities in (8.2.2.1), that has proved to be useful is $s_i = 9|\delta_m|\sigma_i$, where $|\delta_m|$ represents the median value of $|[y_i - M_i(\mathbf{x})]/\sigma_i|$, the *median absolute deviation*, or MAD.

Implementation of a procedure based on the function given in (8.2.2.1) involves modification of the weights used in each cycle by

$$\begin{aligned}\varphi(\Delta) &= (1-\Delta^2)^2, \qquad |\Delta|<1,\\ \varphi(\Delta) &= 0, \qquad\qquad\quad |\Delta|\geq 1.\end{aligned} \tag{8.2.2.4}$$

Because of this weight modification, the procedure is sometimes referred to as 'iteratively reweighted least squares'. It should be recognized, however, that the function that is minimized is more complex than a sum of squares. In a strict application of the Gauss–Newton algorithm (see Section 8.1.3) to the minimization of this function, each term in the summations to form the normal-equations matrix contains a factor $\omega(\Delta_i)$, where $\omega(\Delta) = \mathrm{d}^2\rho/\mathrm{d}\Delta^2 = 1 - 6\Delta^2 + 5\Delta^4$. This factor actually gives some data points a negative effective 'weight', because the sum is actually reduced by making the fit worse. The inverse of this normal-equations matrix is not an estimate of the variance–covariance matrix; for that the unmodified weights, equal to $1/\sigma_i^2$, must be used, but, because more discrepant points have been deliberately down weighted relative to the ideal weights, the variances are, in general, underestimated. A recommended procedure (Huber, 1973; Nicholson *et al.*, 1982) is to calculate the normal-equations matrix using the unmodified weights, invert that matrix, and premultiply by an estimate of the variance of the residuals (Section 8.4.1) using modified weights and $(n-p)$ degrees of freedom. Huber showed that this estimate is biased low, and suggests multiplication by a number, $c^2$, greater than one, and given by

$$c = [1 + ps_\omega^2/n\overline{\omega}^2]/\overline{\omega}, \tag{8.2.2.5}$$

where $\overline{\omega}$ is the mean value, and $s_\omega^2$ is the variance of $\omega(\Delta_i)$ over the entire data set. The conditions under which this expression is derived are not well satisfied in the crystallographic problem, but, if $n/p$ is large and $\overline{\omega}$ is not too much less than one, the value of $c$ will be close to $1/\overline{\omega}$. $\overline{\omega}$ plays the role of a 'variance efficiency factor'. That is, the variances are approximately those that would be achieved with a least-squares fit to a data set with normally distributed errors that contained $n\overline{\omega}$ data points.

Robust/resistant methods have been discussed in detail by Huber (1981), Belsley, Kuh & Welsch (1980), and Hoaglin, Mosteller & Tukey (1983). An analysis by Wilson (1976) shows that a fitting procedure gives unbiased estimates if

$$\sum_{i=1}^{n}\left[\left(\frac{\partial w_i}{\partial y_{ci}}\right)\left(\frac{\mathrm{d}y_{ci}}{\mathrm{d}x}\right)\sigma_i^2\right] = 2\sum_{i=1}^{n}\left[\left(\frac{\partial w_i}{\partial y_{oi}}\right)\left(\frac{\mathrm{d}y_{ci}}{\mathrm{d}x}\right)\sigma_i^2\right], \tag{8.2.2.6}$$

where $y_{oi}$ and $y_{ci}$ are the observed and calculated values of $y_i$, respectively. Least squares is the case where all terms on both sides of the equation are equal to zero; the weights are fixed. In maximum-likelihood estimation or robust/resistant estimation, the effective weights are functions of the deviation, causing possible introduction of bias. Equation (8.2.2.6), however, suggests that the estimates will still be unbiased if the sums on both sides are zero, which will be the case if the error distribution and the weight modification function are both symmetric about $\Delta = 0$.

Note that the fact that two different weighting schemes applied to the same data lead to different values for the estimate does not necessarily imply that either value is biased. As long as the observations represent unbiased estimates of the values of the model functions, any weighting scheme gives unbiased estimates of the model parameters, although some weighting schemes will cause those estimates to be more precise than others will. Bias can be introduced if a procedure *systematically* causes fluctuations in one direction to be weighted more heavily than fluctuations in the other. For example, in the Rietveld method (Chapter 8.6), the observations are counts of quanta, which are subject to fluctuation according to the Poisson distribution, where the probability of observing $k$ counts per unit time is given by

$$\Phi(k) = \lambda^k \exp(-\lambda)/k!. \tag{8.2.2.7}$$

The mean and the variance of this p.d.f. are both equal to $\lambda$, so that the ideal weighting should have $w_i = 1/\lambda_i$. However, $\lambda_i$ is not known *a priori*, and must be estimated. The usual procedure is to take $k_i$ as an estimate of $\lambda_i$, but this is an unbiased estimate only asymptotically for large $k$ (Box & Tiao, 1973), and, furthermore, causes observations that have negative, random errors to be weighted more heavily than observations that have positive ones. This correlation can be removed by using, after a preliminary cycle of refinement, $M_i(\widehat{\mathbf{x}})$ as an estimate of $\lambda_i$. This might seem to have the effect of making the weights dependent on the calculated values, so that the right-hand side of (8.2.2.6) is no longer zero, but this applies only if the weights are changed during the refinement. There is thus no conflict with the result in (8.1.2.9). In practice, in any case, many other sources of uncertainty are much more important than any possible bias that could be introduced by this effect.

### 8.2.3. Entropy maximization

#### 8.2.3.1. *Introduction*

Entropy maximization, like least squares, is of interest primarily as a framework within which to find or adjust parameters of a model. Rationalization of the name 'entropy maximization' by analogy to thermodynamics is controversial, but there is formal proof (Shore & Johnson, 1980) supporting entropy maximization as the unique method of inference that satisfies basic consistency requirements (Livesey & Skilling, 1985). The proof consists of discovering the consequences of four consistency axioms, which may be stated informally as follows:

(1) the result of the inference should be unique;
(2) the result of the inference should be invariant to any transformations of coordinate system;
(3) it should not matter whether independent information is accounted for independently or jointly;
(4) it should not matter whether independent subsystems are treated separately in conditional problems or collected and treated jointly.

The term 'entropy' is used in this chapter as a name only, the name for variation functions that include the form $\varphi\ln\varphi$, where $\varphi$ may represent probability or, more generally, a positive proportion. Any positive measure, either observed or derived, of the relative apportionment of a characteristic quantity among observations can serve as the proportion.

The method of entropy maximization may be formulated as follows: given a set of $n$ observations, $y_i$, that are measurements of quantities that can be described by model functions, $M_i(\mathbf{x})$, where $\mathbf{x}$ is a vector of parameters, find the prior, positive proportions, $\mu_i = f(y_i)$, and the values of the parameters for which the positive proportions $\varphi = f[M_i(\mathbf{x})]$ make the sum

$$S = -\sum_{i=1}^{n}\varphi_i'\ln(\varphi_i'/\mu_i'), \qquad (8.2.3.1)$$

where $\varphi_i' = \varphi_i/\sum\varphi_j$ and $\mu_i' = \mu_i/\sum\mu_j$, a maximum. $S$ is called the *Shannon–Jaynes entropy*. For some applications (Collins, 1982), it is desirable to include in the variation function additional terms or restraints that give $S$ the form

$$S = -\sum_{i=1}^{n}\varphi_i'\ln(\varphi_i'/\mu_i') + \lambda_1\xi_1(\mathbf{x},\mathbf{y}) + \lambda_2\xi_2(\mathbf{x},\mathbf{y}) + \ldots, \qquad (8.2.3.2)$$

where the $\lambda$s are undetermined multipliers, but we shall discuss here only applications where $\lambda_i = 0$ for all $i$, and an unrestrained entropy is maximized. A necessary condition for $S$ to be a maximum is for the gradient to vanish. Using

$$\frac{\partial S}{\partial x_j} = \sum_{i=1}^{n}\left(\frac{\partial S}{\partial\varphi_i}\right)\left(\frac{\partial\varphi_i}{\partial x_j}\right) \qquad (8.2.3.3)$$

and

$$\frac{\partial S}{\partial\varphi_i} = \sum_{k=1}^{n}\left(\frac{\partial S}{\partial\varphi_k'}\right)\left(\frac{\partial\varphi_k'}{\partial\varphi_i}\right), \qquad (8.2.3.4)$$

straightforward algebraic manipulation gives equations of the form

$$\sum_{i=1}^{n}\left\{\frac{\partial\varphi_i}{\partial x_j} - \varphi_i'\left(\sum_{k=1}^{n}\frac{\partial\varphi_k}{\partial x_j}\right)\right\}\ln\left(\frac{\varphi_i'}{\mu_i'}\right) = 0. \qquad (8.2.3.5)$$

It should be noted that, although the entropy function should, in principle, have a unique stationary point corresponding to the global maximum, there are occasional circumstances, particularly with restrained problems where the undetermined multipliers are not all zero, where it may be necessary to verify that a stationary solution actually maximizes entropy.

#### 8.2.3.2. *Some examples*

For an example of the application of the maximum-entropy method, consider (Collins, 1984) a collection of diffraction intensities in which various subsets have been measured under different conditions, such as on different films or with different crystals. All systematic corrections have been made, but it is necessary to put the different subsets onto a common scale. Assume that every subset has measurements in common with some other subset, and that no collection of subsets is isolated from the others. Let the measurement of intensity $I_h$ in subset $i$ be $J_{hi}$, and let the scale factor that puts intensity $I_h$ on the scale of subset $i$ be $k_i$. Equation (8.2.3.1) becomes

$$S = -\sum_{h=1}^{n}\sum_{i=1}^{m}(k_iI_h)'\ln\left[\frac{(k_iI_h)'}{J_{hi}'}\right], \qquad (8.2.3.6)$$

where the term is zero if $I_h$ does not appear in subset $i$. Because $k_i$ and $I_h$ are parameters of the model, equations (8.2.3.5) become

$$\sum_{i=1}^{m}k_i\ln\left[\frac{(k_iI_h)'}{J_{hi}'}\right] - \sum_{h=1}^{n}\sum_{i=1}^{m}(k_iI_h)'\left(\sum_{l=1}^{m}k_l\right)\ln\left[\frac{(k_iI_h)'}{J_{hi}'}\right] = 0, \qquad (8.2.3.7a)$$

and

$$\sum_{h=1}^{n}I_h\ln\left[\frac{(k_iI_h)'}{J_{hi}'}\right] - \sum_{h=1}^{n}\sum_{i=1}^{m}(k_iI_h)'\left(\sum_{l=1}^{n}I_l\right)\ln\left[\frac{(k_iI_h)'}{J_{hi}'}\right] = 0. \qquad (8.2.3.7b)$$

These simplify to

$$\ln I_h = Q - \sum_{i=1}^{m}k_i'\ln(k_i/J_{hi}) \qquad (8.2.3.8a)$$

and

$$\ln k_i = Q - \sum_{h=1}^{n}I_h'\ln(I_h/J_{hi}), \qquad (8.2.3.8b)$$

where

$$Q = \sum_{h=1}^{n}\sum_{i=1}^{m}(k_iI_h)'\ln[(k_iI_h)/J_{hi}]. \qquad (8.2.3.8c)$$

Equations (8.2.3.8) may be solved iteratively, starting with the approximations $k_i = \sum_{h=1}^{n}J_{hi}$ and $Q = 0$.

The standard uncertainties of scale factors and intensities are not used in the solution of equations (8.2.3.8), and must be computed separately. They may be estimated on a fractional basis from the variances of estimated population means $\langle J_{hi}/I_h\rangle$ for a scale factor and $\langle J_{hi}/k_i\rangle$ for an intensity, respectively. The maximum-entropy scale factors and scaled intensities are relative, and either set may be multiplied by an arbitrary, positive constant without affecting the solution.

references