avantage de tenir compte de la géométrie moléculaire et de la réalité physique du problème, en termes de liaisons et d'interactions entre celles-ci.

Les auteurs remercient Messieurs les Professeurs J. Toussaint et A. Van de Vorst pour l'intérêt manifesté à ce travail.

Références

AMOS, A. T. & ROBERTS, H. G. FF. (1969*a*). *J. Chem. Phys.* **50**, 2375–2381.

AMOS, A. T. & ROBERTS, H. G. FF. (1969*b*). *Theor. Chim. Acta* **13**, 421–427.

BAUDET, J., GUY, J. & TILLIEU, J. (1960). *J. Phys. Radium*, **21**, 600–608.

CARALP, L. & HOARAU, J. (1971). *J. Chim. Phys.* **68**, 63–72.

CARALP, L. & HOARAU, J. (1972). *J. Chim. Phys.* **69**, 774–782.

CARALP, L., HOARAU, J. & PESQUER, M. (1969). *C. R. Acad. Sci.* **269**, 480–482.

COULSON, C. A., GOMEZ, J. A. N. F. & MALLION, R. B. (1975). *Mol. Phys.* **30**, 713–732.

DAVIES, D. W. (1963). *Mol. Phys.* **6**, 489–492.

HABERDITZL, W. (1976). *Theory and Applications of Molecular Diamagnetism*, pp. 59–233. New York: Wiley – Interscience.

HALEY, L. V. & HAMEKA, H. F. (1974). *J. Am. Chem. Soc.* **96**, 2020–2024.

JONATHAN, N., GORDON, S. & DAILEY, B. P. (1962). *J. Chem. Phys.* **36**, 2443–2448.

KRISHNAN, K. S. & BANERJEE, S. (1935). *Trans. R. Soc. London Ser. A*, **234**, 265–298.

O'SULLIVAN, P. S. & HAMEKA, H. F. (1970). *J. Am. Chem. Soc.* **92**, 1821–1824.

POPLE, J. A. (1962*a*). *J. Chem. Phys.* **37**, 53–59.

POPLE, J. A. (1962*b*). *J. Chem. Phys.* **37**, 60–66.

POPLE, J. A. (1963). *J. Chem. Phys.* **38**, 1276–1278.

POPLE, J. A. (1964). *J. Chem. Phys.* **41**, 2559–2560.

PULLMAN, A. & PULLMAN, B. (1952). *Les Théories Électroniques de la Chimie Organique*. Paris: Masson.

SOBRY, R. & VAN DEN BOSSCHE, G. (1982). *Acta Cryst.* A**38**, 286–287.

VAN DEN BOSSCHE, G. & SOBRY, R. (1974). *Acta Cryst.* A**30**, 616–625.

VAN DEN BOSSCHE, G. & SOBRY, R. (1981). *Acta Cryst.* A**37**, 211–219.

Acta Cryst. (1982). A**38**, 537–549

A Profile-Fitting Method for the Analysis of Diffractometer Intensity Data

BY STUART OATLEY AND SIMON FRENCH*

Laboratory of Molecular Biophysics, University of Oxford, South Parks Road, Oxford OX1 3PS, England

(*Received* 5 *June* 1981; *accepted* 24 *February* 1982)

Abstract

For diffractometer data collected using a step-scan method, various procedures have been proposed by which the integrated peak intensity may be estimated from the measured reflection profile. However, these all ignore some of the information available in the data, thereby reducing the accuracy of the estimation. Moreover, some make assumptions about structure present in the sequence of counts and so produce a large positive bias in their estimation of weak reflections. A profile-fitting approach based upon the Bayesian three-stage regression model is presented, which avoids these failings. The underlying theory is discussed, its implementation for off-line data reduction and its potential for on-line diffractometer control is described and its application to various protein data sets collected using both single- and multiple-counter diffractometers is reported.

* Present address: Department of Decision Theory, University of Manchester, Manchester M13 9PL, England.

1. Introduction

For a diffractometer operating in a step-scan mode, each reflection is recorded by measuring a sequence of counts as the machine steps across the peak and its local background. Each count c_i is an observation on the true (mean) count λ_i at the ith step. Thus

$$c_i \sim P_{c_i}(.|\lambda_i), \quad i = 1, 2, \ldots, N, \tag{1.1}$$

where the notation indicates that each c_i is drawn from a distribution with parameter λ_i. [A fuller explanation of the notation used in this paper may be found in French (1978).] The distributions $P_{c_i}(.|\lambda_i)$ are approximately Poisson ('counting statistics') with means λ_i but are perturbed slightly through instrument instability, such as small variations in the strength of the incident beam and slight breakdowns in the counting chains, and counting losses through saturation of the detector for intense reflections. The latter effect is unlikely to be of importance in protein crystallography.

Each λ_i is the sum of two elements: a contribution

0567-7394/82/040537-13$01.00

from the intensity of the reflection and a contribution from the background scatter. Thus

$$\lambda_i = J\pi(x_i) + b(x_i), \quad i = 1, 2, \ldots, N, \qquad (1.2)$$

where J is the integrated intensity of the reflection; $\pi(x)$ is the peak shape function, so $\int \pi(x)\,dx = 1$; x_i is the position in the scan of step i; and $b(x_i)$ is the background scatter at x_i. Thus the problem is to obtain an estimate of J together with some indication of the precision of this estimate.

The data available for the estimation of J clearly include the sequence of measured counts $(c_1, c_2, \ldots, c_N)$, but there are other sources of information which are often overlooked. In short, these are: (i) the local behaviour of the background; this may be predicted from the collection geometry, *e.g.* for ω scans $b(x)$ will be approximately constant throughout the scan; (ii) the properties expected in the peak shape, *e.g.* $\pi(x)$ is continuous and, in most cases, unimodal; (iii) the shape of the peaks already analysed; it has been observed that peak shape tends to vary only slowly through reciprocal space (Diamond, 1969); (iv) the position within the scan of the last measured reflection and the reliability of the diffractometer in moving from one reflection to the next; (v) for a multiple-counter diffractometer, the relative positions of the peaks within the simultaneously collected scans; this is defined by the setting geometry and, moreover, the peak shapes and backgrounds on these scans will usually be very similar.

Various methods have been proposed for the estimation of J. The majority base their estimate on the sequence of measured counts only, ignoring the sources of information (i) to (v). The simplest method is background–peak–background integration (BPB). In this it is assumed that for every reflection the entire peak lies in a window of W steps beginning at the N_Wth step, *i.e.* that

$$\int_{x_{N_W}}^{x_{N_W+W-1}} \pi(x)\,dx = 1.$$

Letting

$$B_1 = \sum_{i=1}^{N_W-1} c_i, \quad P = \sum_{i=N_W}^{N_W+W-1} c_i$$

and

$$B_2 = \sum_{i=N_W+W}^{N} c_i,$$

the BPB estimate of the integrated intensity is

$$I(\mathbf{c}) = P - [W/(N-W)](B_1 + B_2). \qquad (1.3)$$

If instrument instability and counting losses are ignored and the distributions in (1.1) taken to be independent Poisson, an estimate of the variance is

$$\mathrm{Var}[I(\mathbf{c})|J] = P + [W/(N-W)]^2 (B_1 + B_2). \qquad (1.4)$$

The main problem with BPB estimation is that in order to allow for crystal slippage during data collection, the window width must be chosen much larger than the peak width. This has two serious effects: firstly, the possible precision of the intensity estimate is decreased; and secondly, the crystal exposure and consequent irradiation damage are increased.

Various attempts have been made to overcome this difficulty. The ordinate analysis method of Watson, Shotton, Cox & Muirhead (1970), the method of Lehmann & Larsen (1974) and the centroid method of Tickle (1975) all use the vector of measured counts to centre the peak within the scan, *i.e.* N_W becomes a function of $\mathbf{c}$. Thus they compensate for any slight crystal slippage and hence enable a smaller window width to be used. There are, however, two problems. Firstly, for weak reflections noise may dominate the scan, resulting in the window being miscentred and a positive bias in the estimation [see French (1975), Tickle (1975) and Table 1 and Figs. 3–5 below]. It should be noted that Tickle's (1975) method is much less susceptible to this effect than the others. Secondly, since N_W is now a random quantity, *i.e.* because the allocation of counts to B_1, P and B_2 is now random, the variance estimate (1.4) no longer applies. This is shown in Table 1 for ordinate analysis.

None of the foregoing methods use any information from sources (i) to (iv) above. When multiple-counter data are analysed, part of source (v) may be used; the measured counts are notionally shifted by their calculated relative displacements on the scanning axis and then summed to produce a combined profile whose peak centre may be determined by either the ordinate analysis or centroid methods (Banner, Evans, Marsh & Phillips, 1977). Use of this combined profile improves the signal-to-noise ratio and reduces the problem of miscentring, but difficulties remain where all the reflections measured together are weak. This method still ignores the similarity between peak shapes and between backgrounds for the multiple counters. Thus the majority of the information (i) to (v) is wasted and as a result the estimation is not as accurate as is possible.

Diamond (1969) has developed a profile-fitting method which does attempt to use all the available information. His results show it to be considerably more successful in the estimation of integrated intensities than any of the other methods. However, it makes use of the information (i) to (v) heuristically, not probabilistically, and thus the estimates of variance do not truly reflect the precision of his estimated intensities. Also no account is taken of the continuity in the peak shape, or of the positions of the peaks within previous scans.

More recently, Hanson, Watenpaugh, Sieker & Jensen (1979) have reported a method which may be thought of as an approximation to Diamond's (1969)

method and to the one that we present below. It is a two-pass procedure which fits Gaussian peak shapes to the vectors of counts. On the first pass only the intense reflections are fitted and the variation of peak width over reciprocal space determined. On the second pass this variation is used to predict the peak width of all reflections and their intensities are integrated. The method is computationally fast and the results are encouraging. However, it ignores much of the information from sources (i) to (v) and the estimated variance is not a complete reflection of the precision of the integrated intensity. Since it is a two-pass procedure, it does not have the potential for on-line diffractometer control.

The profile-fitting method described here we believe overcomes the various criticisms of the earlier methods. The theory underlying the method has already been presented briefly (French, 1978; French and Oatley, 1981), and is given in greater detail in the next section. § 3 contains analyses of simulated data and describes practical experience of the method for both single- and five-counter data, in particular its successful treatment of weak reflections. Some of the approximations used in implementing the method on a computer, and summary flow diagrams of the main algorithms employed, are given in Appendix *A*.

2. The Bayesian three-stage model

Expressions (1.1) and (1.2) indicate that the expected value of c_i,

$$E(c_i) = \lambda_i = J\pi(x_i) + b(x_i) \tag{2.1}$$

for $i = 1, 2, \ldots, N$. Our approach to estimating J is to fit to the vector of observed counts a function of the form $[J\pi(x,\boldsymbol{\alpha}) + b(x,\boldsymbol{\beta})]$, where $\pi(x,\boldsymbol{\alpha})$ and $b(x,\boldsymbol{\beta})$ are parametric approximations to the true peak shape and background functions respectively. Since the peak shape function should integrate to unity obvious candidates for $\pi(x,\boldsymbol{\alpha})$ are probability density functions. We have found that Johnson's (1949) suggestion for transforming the normal density curve leads to families of curves which well model the peak shapes that arise in protein crystallography. These have four parameters $\boldsymbol{\alpha} = (\mu,\sigma,\gamma,\delta)$, such that

$$\pi(x,\boldsymbol{\alpha}) = (\delta/\sigma\sqrt{2\pi})\,(\partial g/\partial y)\exp\{-\tfrac{1}{2}[\gamma + \delta g(y)]^2\}, \tag{2.2}$$

where $y = (x - \mu)/\sigma$ and $g(y)$ is a monotonic increasing function (see below). In all cases, $\pi(x,\boldsymbol{\alpha})$ is continuous and unimodal. μ determines the location of the peak within the scan and σ its width; the parameters γ, δ and the choice of $g(y)$ together determine the peak shape. We have used two choices of $g(y)$:

$$g(y) = \sinh^{-1}(y) \quad \text{and} \quad g(y) = \sinh(y), \tag{2.3}$$

which lead to families of curves which are respectively more sharply peaked and more flat-topped than the normal. In both cases δ is related to the departure from normality and γ determines the skewness of the curve, symmetry resulting when $\gamma = 0$. Fig. 1 illustrates the wide variety of peak shapes that can result.

We have found that the peaks arising from a single crystal are either consistently sharply peaked or consistently flat-topped over large regions of reciprocal space. Thus the function $g(y)$ need not be chosen for each reflection, but may be determined by examination of the first few peaks to be processed.

For the backgrounds we have found that a linear approximation is adequate:

$$b(x,\boldsymbol{\beta}) = \beta_1 + \beta_2 x. \tag{2.4}$$

Usually the background is approximately constant across the scan, so that $\beta_2 \simeq 0$.

Other choices of $\pi(x,\boldsymbol{\alpha})$ and $b(x,\boldsymbol{\beta})$ may be more appropriate in other branches of crystallography or for other collection geometries. However, it should be noted that our profile-fitting method remains applicable whatever choices are made.

As discussed in French (1978) and French & Oatley (1981), the Bayesian three-stage model provides a natural structuring of estimation problems. In particular, it was shown how the problem of estimating J may be so structured, using the following three-stage model:

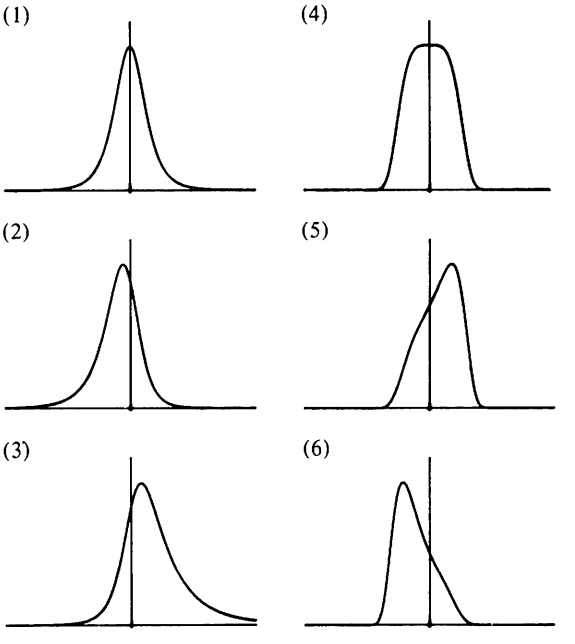


Fig. 1. Examples of peak shapes obtainable from (2.2) and (2.3). Each peak is normalized and has $\mu = 0{\cdot}0$ and $\sigma = 1{\cdot}0$; for (1) to (3), $g(y) = \sinh^{-1}(y)$ and for (4) to (6), $g(y) = \sinh(y)$: (1) $\gamma = 0{\cdot}0$, $\delta = 1{\cdot}5$; (2) $\gamma = 0{\cdot}6$, $\delta = 1{\cdot}5$; (3) $\gamma = -0{\cdot}8$, $\delta = 1{\cdot}2$; (4) $\gamma = 0{\cdot}0$, $\delta = 1{\cdot}0$; (5) $\gamma = -0{\cdot}5$, $\delta = 1{\cdot}0$; (6) $\gamma = 0{\cdot}8$, $\delta = 1{\cdot}0$.

Stage I. *Observation error*

This stage models the counting and instrument instability errors in the observations.

$$\sqrt{c_i} \sim N\{\sqrt{\lambda_i}, 0{\cdot}25[1 + \sigma_1^2 \lambda_i]\}. \qquad (2.5)$$

Here σ_1 relates specifically to the instrument instability errors and it is assumed that errors at different steps in the scan are uncorrelated.

Stage II. *Modelling error*

This stage describes how well it is expected that the parametric approximation $[J\pi(x,\boldsymbol{\alpha}) + b(x,\boldsymbol{\beta})]$ will model $[J\pi(x) + b(x)]$.

$$\sqrt{\lambda_i} \sim N(\sqrt{v_i}, 0{\cdot}25\sigma_2^2 v_i), \qquad (2.6)$$

where $v_i = J\pi(x_i, \boldsymbol{\alpha}) + b(x, \boldsymbol{\beta})$. σ_2^2 is the relative variance of the modelling error. Again it is assumed that errors at different steps are uncorrelated. For the reasons behind this assumption see French (1978).

Stage III. *Prior knowledge*

Here information about $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, learned from the previously fitted reflections, is introduced. This procedure is described in detail below.

To solve this model, and hence estimate J and determine its variance, σ_1, σ_2 and the structure at the third stage must be specified. First we consider σ_1 and σ_2.

Note from (2.5) and (2.6) that $E(\sqrt{c_i}) = \sqrt{\lambda_i}$ and in turn $E(\sqrt{\lambda_i}) = \sqrt{v_i}$. The relation between the first two stages is therefore linear and they may be combined:

$$\sqrt{c_i} \sim N(\sqrt{v_i}, 0{\cdot}25[1 + \sigma_1^2 \lambda_i + \sigma_2^2 v_i]). \qquad (2.7)$$

Moreover, v_i is a good approximation to λ_i so we may approximate the variance in (2.7) by

$$0{\cdot}25[1 + (\sigma_1^2 + \sigma_2^2)v_i]. \qquad (2.8)$$

Thus, σ_1^2 and σ_2^2 enter the calculations only through their sum, $e^2 = (\sigma_1^2 + \sigma_2^2)$. Providing that the diffractometer operates consistently, we may expect σ_1^2 to be constant over the data set. Similarly, unless some severe change in the crystal occurs during data collection, we may expect σ_2^2 to be constant. We estimate e^2 by repeatedly fitting the first few peaks in the data set using different values of e^2. Its value is correct when standardized residuals have approximately unit variances, *i.e.* if v_i are calculated from the fitted parameters

$$\frac{1}{N}\sum_{i=1}^{N} (c_i - v_i)^2/[(1 + e^2 v_i)v_i] \simeq 1. \qquad (2.9)$$

(To be strictly correct the residuals between $\sqrt{c_i}$ and $\sqrt{v_i}$ should be considered, but there are computational advantages in using this approximately equivalent form.)

In French (1978) a general approach to setting the prior distributions at stage III was discussed. Here we consider the setting of these distributions for our choices (2.2) and (2.4) of $\pi(x,\boldsymbol{\alpha})$ and $b(x,\boldsymbol{\beta})$ respectively.

A vague prior distribution is one which states that nothing is known about the relevant quantity other than the information contained in the experimental data. Suppose J, the integrated intensity, can be maximally 10^5 but is typically of the order of a few hundred. The prior distribution

$$J \sim N(0, 10^{20}) \qquad (2.10)$$

has an effectively constant density over the range of J and thus does not differentially weight the possible values; hence the data alone will determine the estimate (posterior mean) of J.

The prior distribution for the parameters $\boldsymbol{\alpha} = (\mu, \sigma, \gamma, \delta)$ need not be so vague. Suppose we are setting the prior for the sth reflection, after successfully fitting the profile at the $(s-1)$th reflection, adjacent to it in reciprocal space. The posterior distribution for the parameters after the $(s-1)$th reflection is:

$$\boldsymbol{\alpha}_{s-1} = \begin{pmatrix} \mu_{s-1} \\ \sigma_{s-1} \\ \gamma_{s-1} \\ \delta_{s-1} \end{pmatrix} \sim N(\mathbf{m}_{s-1}, W_{s-1}). \qquad (2.11)$$

For our collection geometries, there are no predictable changes in $\boldsymbol{\alpha}$ between reflections, but it is known that, firstly, the peak shape is likely to vary slowly across reciprocal space and, secondly, even if there is crystal slippage, the peak position within one scan will be very close to that for the previous reflection. Hence we may use $\mathbf{m}_{s-1}$ as the prior mean of $\boldsymbol{\alpha}_s$, but increase the diagonal terms of W_{s-1} to form the prior variance of $\boldsymbol{\alpha}_s$:

$$\boldsymbol{\alpha}_s = \begin{pmatrix} \mu_s \\ \sigma_s \\ \gamma_s \\ \delta_s \end{pmatrix} \sim N\left[\mathbf{m}_{s-1}, W_{s-1} + \begin{pmatrix} u_\mu^2 & 0 & 0 & 0 \\ 0 & u_\sigma^2 & 0 & 0 \\ 0 & 0 & u_\gamma^2 & 0 \\ 0 & 0 & 0 & u_\delta^2 \end{pmatrix}\right]. \qquad (2.12)$$

Thus to define the prior for $\boldsymbol{\alpha}_s$ we need to specify u_μ^2, u_σ^2, u_γ^2, and u_δ^2. Consider, say, μ. If the average shift in peak position between adjacent scans is 0·1 (one-tenth of a step in the scan), then taking $u_\mu^2 = (0{\cdot}1)^2$ is reasonable. The other shift variances may be set similarly.

Finally, a prior distribution must be set for β_1 and β_2

as given by (2.4). The background is expected to be constant across most scans, thus

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10^{20} & 0 \\ 0 & 10^{-10} \end{pmatrix} \right]. \quad (2.13)$$

The variance of 10^{20} corresponds to a vague prior for the background level β, while the variance 10^{-10} forces the background slope to remain very close to zero. If it is believed that the background actually slopes, a more appropriate value should be used, *e.g.* β_2 might be in the range -1 to $+1$, where a prior variance of 1 would be reasonable.

Distributions (2.10), (2.12) and (2.13) together define the prior distribution for the third stage. Given these we may solve the three-stage model and so estimate J. This is not in fact accomplished by the method suggested in Appendix A of French (1978) because for large data sets the computation would become prohibitive. Instead we use a fast approximate solution method which is described in the Appendix of this paper.

Only the setting of the prior distribution for α in the simplest case was considered above; a complete flow chart of the routine controlling the profile fitting is given in Fig. 2. On reading the observed profile for a reflection the prior distributions for J, β_1 and β_2 are set according to (2.10) and (2.13). If the fitting at the last reflection was successful and if the current reflection is adjacent in reciprocal space, then the prior for α is set as given by (2.12). Usually a single cycle of the non-linear fitting routine is sufficient to obtain a good fit and thus the number of these cycles, maxcyc, is set to 1. There are two distinct ways in which the fitting may be said to have failed. Firstly, numerical convergence may not have been reached, *i.e.* after maxcyc cycles the shifts in the parameters may still be significant. Secondly, numerical convergence may have been obtained but, nonetheless, the fitted profile may poorly represent the data. We check the quality of the fitting by a modified Smirnov statistic:

$$G_f = \max_{n=1}^{N} \left| \sum_{i=1}^{n} (c_i - v_i) \right| \Big/ \sum c_i. \quad (2.14)$$

[*N.B.* χ^2 is a very poor statistic for checking the goodness of fit in this case (French, 1975).] The fit is considered adequate provided G_f is not too large; this criterion depends on the strength of the data, the number of steps in the scan and the reliability of the counters, but it is easy with experience of the method to set a suitable limiting value.

When either type of failure is detected, the prior for α is reset to 'free' values (see below), and the fitting reattempted, now allowing up to ncyc non-linear cycles rather than just 1.

It would be incorrect to weaken the prior for α as in (2.12) when processing has moved to a new region of reciprocal space or when, despite resetting the prior for α and allowing ncyc non-linear cycles, the fitting at the previous reflection failed completely. In these cases, the prior for α is set to 'free' values, being either the last values satisfactorily fitted or those fitted at the beginning of the last row, depending on whichever is nearer in reciprocal space, and a diagonal variance matrix used which allows the parameters considerable freedom during convergence. These variances should be typically ten times greater than the shift variances u_μ^2, u_σ^2, u_ν^2 and u_δ^2. For the first reflection in the data to be fitted, the prior means of α must be specified; suitable values can be obtained during the trial fitting of the first few peaks undertaken to determine e^2 and the curve type.

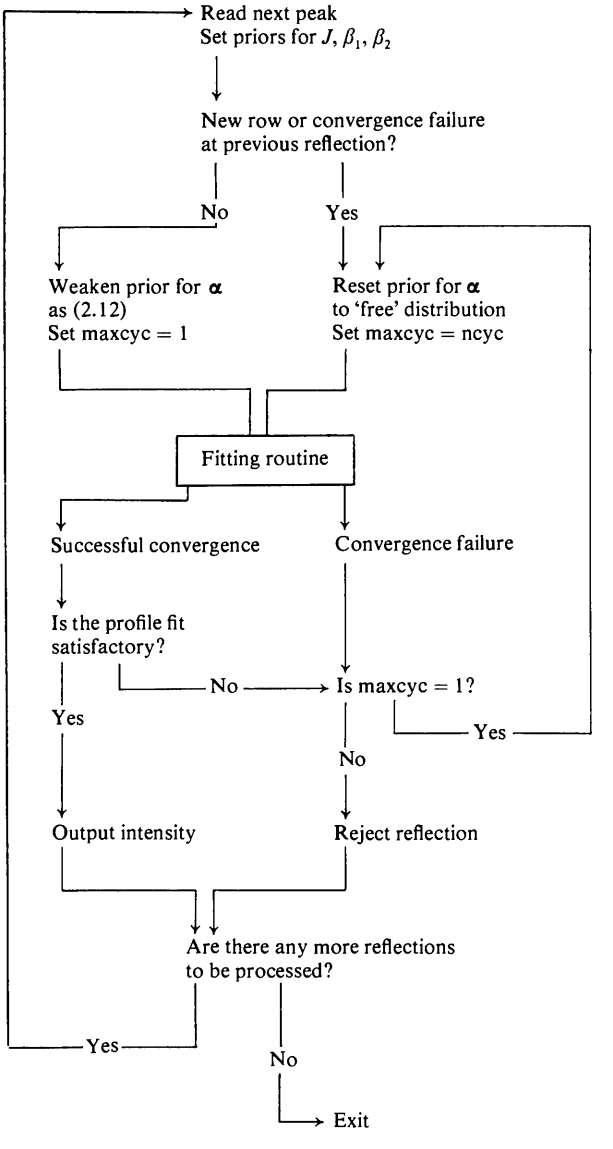


Fig. 2. Flow chart of the routine controlling profile fitting.

It is worth returning to the sources of information (i) to (v) and indicating how these are assimilated. The prior distribution (2.13) and the choice of $b(x, \boldsymbol{\beta})$ allows us to represent our knowledge of the behaviour of the background, source (i). Our choice of $\pi(x,\boldsymbol{\alpha})$ represents our knowledge of the peak shape (ii), while the prior distribution for $\boldsymbol{\alpha}$ introduces our knowledge of its similarity for adjacent reflections, sources (iii) and (iv). Finally, for multiple counter data, information source (v) is modelled by using the same peak parameters to fit the reflections on all counters, after having defined the origin of x on each scan so that the peak positions share the same numerical value. The modelling variance σ_2^2, and hence e^2, now contains a contribution from the variation expected between the simultaneously collected peaks. This approach has the computational advantage of reducing the number of parameters but, were the variation between reflections greater, it would be better to allow different peak parameters for each and introduce the prior information on their similarity exchangeably. For a full treatment of exchangeable prior information, see Lindley & Smith (1972).

3. Tests and practical experience of the method

The reliability of our profile-fitting method in the estimation of the true intensity of a reflection, and in particular its ability to analyse weak data successfully, was initially examined using simulated reflection profiles. Model Gaussian peaks corresponding to various integrated intensity values, J, were added to a constant background; for each intensity, 250 scans of 20 steps were generated by superimposing random noise according to 'counting statistics' and normally distributed instrument instability error. Thus,

$$\left.\begin{aligned} z_i &\sim J(1/8\pi)^{1/2}\exp[-0{\cdot}5(i-10)^2/4] + 50 \\ \lambda_i &\sim N(Z_i, 0{\cdot}0009Z_i^2);\ c_i \sim P(\lambda_i) \end{aligned}\right\}$$
$$i = 1, 2, \ldots, 20. \qquad (3.1)$$

The net integrated intensity was obtained both by ordinate analysis and by profile fitting. In order to ensure that the peak was completely contained within the ordinate analysis window, a width of 14 steps was used. For the profile fitting, the curve type was that given by $g(y) = \sinh^{-1}(y)$ in (2.3), with priors, reset for each fit, of $E(\mu) = 10{\cdot}0$, $\mathrm{Var}(\mu) = 1{\cdot}0$; $E(\sigma) = 10{\cdot}0$, $\mathrm{Var}(\sigma) = 1{\cdot}0$; $E(\gamma) = 0{\cdot}0$, $\mathrm{Var}(\gamma) = 0{\cdot}5$; $E(\delta) = 5{\cdot}0$, $\mathrm{Var}(\delta) = 0{\cdot}5$; $E(J) = 0{\cdot}0$, $\mathrm{Var}(J) = 10^{20}$; $E(\beta_1) = 0{\cdot}0$, $\mathrm{Var}(\beta_1) = 10^{20}$; $E(\beta_2) = 0{\cdot}0$, $\mathrm{Var}(\beta_2) = 10^{-10}$. Instrument instability error was set at 3%. For each generated scan, the posterior probability distribution converged to less than 2% within ten cycles.

The results are summarized in Table 1. For each integration method we give three columns. The first and second give respectively the mean and standard deviation of the sample of 250 integrated intensities. The third column gives the mean of their predicted standard deviation *i.e.* the mean of the 250 values given by expression (1.4) in the case of ordinate analysis and the mean of the 250 values given by the corresponding expression for profile fitting. The variation in these predicted standard deviations was less than 5% over each sample of 250 scans. Ideally the predicted standard deviations should be equal to the standard deviation of the sample of integrated intensities, *i.e.* the second and third columns should be roughly equal. It is clear that ordinate analysis produces a very strong positive bias for small intensities, which is still significant at a net intensity of 500, a '10σ intensity'. The standard deviations predicted by (1·4) over-estimate the true variation for small intensities, as discussed in § 2. Profile fitting clearly gives greatly improved results, although there is evidence of a slight positive bias, which is here probably due to the modelling of the Gaussian curve in the simulated data by one derived from (2.2) and (2.3). It is, however, of a small enough magnitude to have negligible effect on the quality of a reduced data set. The predicted standard deviations describe the true variation well, and they are con-

Table 1. *Comparison of the results of ordinate analysis and profile fitting on the simulated data*

	Ordinate analysis			Profile fitting		
True intensity	Mean integrated intensity	Sampled standard deviation	Mean predicted standard deviation [from (1.4)]	Mean integrated intensity	Sampled standard deviation	Mean predicted standard deviation
0	48·6	37·0	47·8	−1·0	31·0	24·3
10	57·9	38·1	47·8	10·6	28·9	24·6
25	71·2	37·3	48·0	25·8	29·4	25·0
50	96·3	36·8	48·2	56·3	25·8	25·8
75	121·9	36·4	48·5	78·6	24·7	26·6
100	148·9	35·4	48·7	106·9	28·4	27·3
250	290·9	39·6	50·4	253·9	30·1	31·0
500	534·3	47·2	52·8	505·5	40·8	36·4

Table 2. *Protein crystal parameters*

	Space group	Cell dimensions (Å)	Resolution (Å)
Prealbumin	$P2_12_12$	$a = 43{\cdot}5, b = 85{\cdot}7, c = 66{\cdot}0$	1·8
Orthorhombic hen egg-white lysozyme	$P2_12_12_1$	$a = 59{\cdot}0, b = 68{\cdot}6, c = 30{\cdot}4$	6
2-Zinc insulin	R_3	$a = b = 82{\cdot}5, c = 34{\cdot}0$	1·5
Cubic insulin	$I2_13$	$a = b = c = 78{\cdot}9$	1·7
Hagfish insulin	$P4_12_12$	$a = b = 38{\cdot}4, c = 85{\cdot}3$	1·9

siderably reduced for small reflections compared with the ordinate analysis values.

We will now describe the results of the application of the method to a number of data sets collected on five proteins, whose crystal parameters are given in Table 2. For the cubic insulin data, the overall merging residual, R_m, was 0·071 for the 10 100 reflections profile fitted to 1·7 Å resolution (Dodson, Dodson, Lewitova & Sabesan, 1978):

$$R_m = \sum_h \sum_i |I_i - \bar{I}| \Big/ \sum_h \sum_i I_i, \qquad (3.2)$$

where I_i is the ith observation on a set of equivalent terms and $\bar{I}$ is the weighted mean intensity of this set. Fig. 3 illustrates the profile fitting of three peaks from a sequence of eight consecutive reflections in these data.

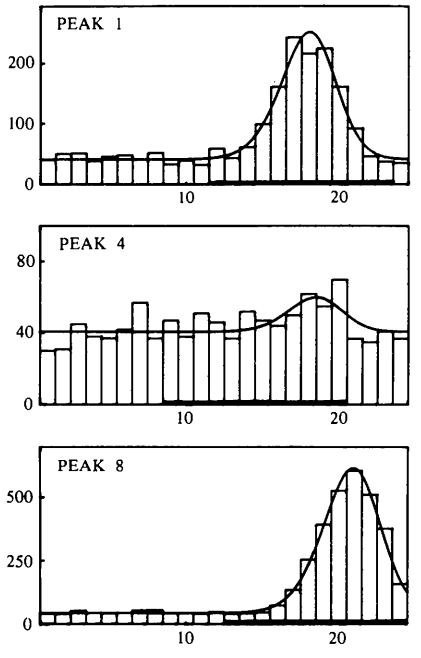


Fig. 3. Profile fitting of the three reflections from the cubic insulin data. The window positions determined by ordinate analysis are indicated by the thick line at the base of each diagram. Peak 1: ordinate analysis: intensity 938·0, e.s.d. 38·0; profile fitting: intensity 951·7, e.s.d. 41·2, $\mu = 19{\cdot}76$, $\sigma = 6{\cdot}53$, $\gamma = 1{\cdot}05$, $\delta = 3{\cdot}76$; $\beta_1 = 41{\cdot}2$; $G_f = 0{\cdot}0240$. Peak 4: ordinate analysis: intensity 132·0, e.s.d. 24·5; profile fitting: intensity 87·5, e.s.d. 21·6, $\mu = 20{\cdot}30$, $\sigma = 6{\cdot}48$, $\gamma = 1{\cdot}15$, $\delta = 3{\cdot}84$; $\beta_1 = 40{\cdot}5$; $G_f = 0{\cdot}0262$. Peak 8: ordinate analysis: intensity 2623·0, e.s.d. 56·3; profile fitting: intensity 2723·2, e.s.d. 61·1, $\mu = 22{\cdot}58$, $\sigma = 6{\cdot}75$, $\gamma = 0{\cdot}95$, $\delta = 3{\cdot}67$, $\beta_1 = 43{\cdot}3$; $G_f = 0{\cdot}0068$.

The fitted parameters show little divergence in the course of processing. The ability of the method, guided by its prior knowledge of background level and peak shape and position, to distinguish signal from noise is illustrated in peak 4 where ordinate analysis has defined a peak window too near the start of the scan, resulting in an over-estimation of the intensity. It can also be seen that as a result of crystal movement the peak position has shifted considerably during this sequence and it is encouraging to note how well this has been tracked by the fitting. This movement has resulted in the last peak, and also the sixth and seventh in the sequence, being seriously 'clipped'; this would invalidate other methods of step-scan integration, but our profile-fitting method, since it can calculate the total area under the fitted peak, is able to extract valid intensity information from the scan.

Data have been collected on human prealbumin (Oatley & Burridge, 1981) with the hormones 3,5,3′-triiodo-L-thyronine (T_3) and L-thyroxine (T_4) bound, to 5·4 and 5·8 Å resolution respectively. The ordinates were analysed during data collection by Tickle's (1975) centroid method and were subsequently profile fitted. The resulting values of R_m for each data set are given in Table 3, where it can be seen that the profile fitting has significantly improved the internal consistency of the data. Subsequently, the high-resolution data sets for the

Table 3. *Comparison of the merging residuals for equivalent reflections in the prealbumin data sets*

	Number of reflections	Centroid method	Profile fitting
Prealbumin + T_3	993	0·049	0·036
Prealbumin + T_4	786	0·043	0·030

Table 4. *Analysis of the* 1·5 *to* 1·7 Å *resolution shell of* 2-Zn *insulin data*

$\sin^2\theta/\lambda^2$ range	Number of reflections	Ordinate analysis $\langle F \rangle$	Ordinate analysis R	Profile fitting $\langle F \rangle$	Profile fitting R
0·0845–0·0910	1635	33·82	0·304	36·43	0·294
0·0910–0·0975	1069	34·06	0·308	33·54	0·301
0·0975–0·1040	1123	36·22	0·317	32·47	0·310
0·1040–0·1111	1152	33·18	0·327	29·30	0·322

two hormone complexes were profile fitted, yielding R_m values of 0·060 for T_3 to 1·9 Å resolution and 0·066 for T_4 to 1·8 Å resolution.

The high-resolution (1·5 Å) data for 2-Zn insulin (Dodson, Dodson, Hodgkin & Reynolds, 1979) were originally collected using ordinate analysis, and it is clear from Table 4 that in the highest-resolution shell, 1·7 to 1·5 Å, there is a severe positive bias in the data, since the mean value of F does not decrease with increasing resolution. These data were recollected using just ten steps across peak and background, and then subjected to profile analysis. Three reflections close together in reciprocal space are illustrated in Fig. 4. For these the centroid and ordinate analysis methods produced identical results; profile fitting gives essentially the same results for the first and third reflections but a much more reasonable value for the intensity of the second reflection. Overall, the improvement in the behaviour of $\langle F \rangle$ is shown in Table 4, which also gives the improvement obtained in the crystallographic residual R,

$$R = \sum_h |F_o - F_c| / \sum_h F_o, \qquad (3.3)$$

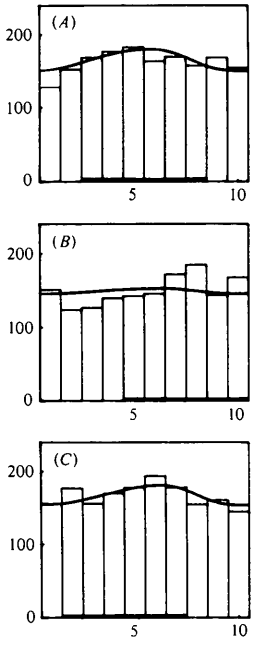


Fig. 4. Profile fitting of three reflections from the 1·7 to 1·5 Å shell of the 2-Zn insulin data. The window positions determined by both ordinate analysis and the centroid method are indicated by the thick line at the base of each diagram. (*A*) Ordinate analysis: intensity 113·0, e.s.d. 52·0; profile fitting: intensity 125·0, e.s.d. 54·0, $\mu = 4{\cdot}70$, $\sigma = 3{\cdot}30$, $\gamma = -0{\cdot}32$, $\delta = 1{\cdot}64$, $\beta_1 = 151{\cdot}0$, $G_f = 0{\cdot}0132$. (*B*) Ordinate analysis: intensity 145·0, e.s.d. 47·0; profile fitting: intensity 35·0, e.s.d. 49·0, $\mu = 4{\cdot}70$, $\sigma = 3{\cdot}27$, $\gamma = -0{\cdot}47$, $\delta = 1{\cdot}63$, $\beta_1 = 146{\cdot}0$, $G_f = 0{\cdot}0455$. (*C*) Ordinate analysis: intensity 127·0, e.s.d. 49·0; profile fitting: intensity 128·0, e.s.d. 56·0, $\mu = 4{\cdot}70$, $\sigma = 3{\cdot}23$, $\gamma = -0{\cdot}52$, $\delta = 1{\cdot}62$, $\beta_1 = 154{\cdot}0$, $G_f = 0{\cdot}0166$.

with the incorporation of the new data. These R values were calculated for the same protein model, *i.e.* same F_c, before further refinement. Similar coarse steps across peak and background were also used in the data collection for hagfish insulin (Cutfield, Cutfield, Dodson, Dodson, Emdin & Reynolds, 1979).

Integrated intensities to 6 Å resolution were obtained for native orthorhombic hen egg-white lysozyme and a $Pt(CN)_4$ derivative both by ordinate analysis and our profile method (Artymiuk, Blake, Rice & Wilson, 1981). The R_m for equivalent reflections, 0·021 for native and 0·015 for the derivative (centrics only) were the same for both methods. These values may not fairly indicate the relative quality of the data sets since the ordinate analysis value will tend to be reduced by the positive bias inherent in the method. However, an indication of the improvement in the quality of the reduced data may be obtained from the behaviour of the $F_{\rm HLE}$ refinement (Dodson & Vijayan, 1971) of the heavy-atom derivative. Here the value of $R_{F_{\rm HLE}}$,

$$R_{F_{\rm HLE}} = \sum_h |F_{\rm HLE} - F_{H_{\rm calc}}| / \sum_h |F_{\rm HLE}|, \qquad (3.4)$$

was reduced overall from 0·378 for the ordinate analysis set to 0·358 for the profile fitted set; in particular, the highest range of $\sin^2\theta/\lambda^2$, from 0·0056 to 0·0071, which contained the weakest data, and was therefore the region where ordinate analysis was likely to produce the poorest isomorphous and anomalous differences, showed a reduction from 0·459 to 0·406.

Both the ordinate analysis and centroid methods have been implemented in a modified form on a five-counter five-circle diffractometer (Banner, Evans, Marsh & Phillips, 1977). As described earlier, these analyse a combined profile, and therefore, since the peak position is more likely to be correct, suffer less from the over-estimation of weak reflections than do the single-counter applications. However, we have found that our profile method can offer significant advantages here too.

Fig. 5 illustrates some results of fitting quintuplets from the high-resolution data sets for native prealbumin (Oatley, 1976; Blake, Geisow, Oatley, Rérat & Rérat, 1978) and for prealbumin with T_3 and T_4 bound (Oatley & Burridge, 1981). Fig. 5(*a*) shows a quintuplet from a shell of 2·0–2·25 Å resolution; here the centroid method has found the peak position well and the profiles are good fits to the measurements. The fitting has ignored the noise at steps 18 and 19 in counter 1. Fig. 5(*b*) shows a very weak quintuplet from a 1·8–2·0 Å shell where the centroid method has failed completely and placed the window in the first half of the scan. Profile fitting is not misled by the generally higher counts here since the resulting peak positions would be too far from those previously fitted, and produces reasonable results from an extremely low signal-to-noise ratio.

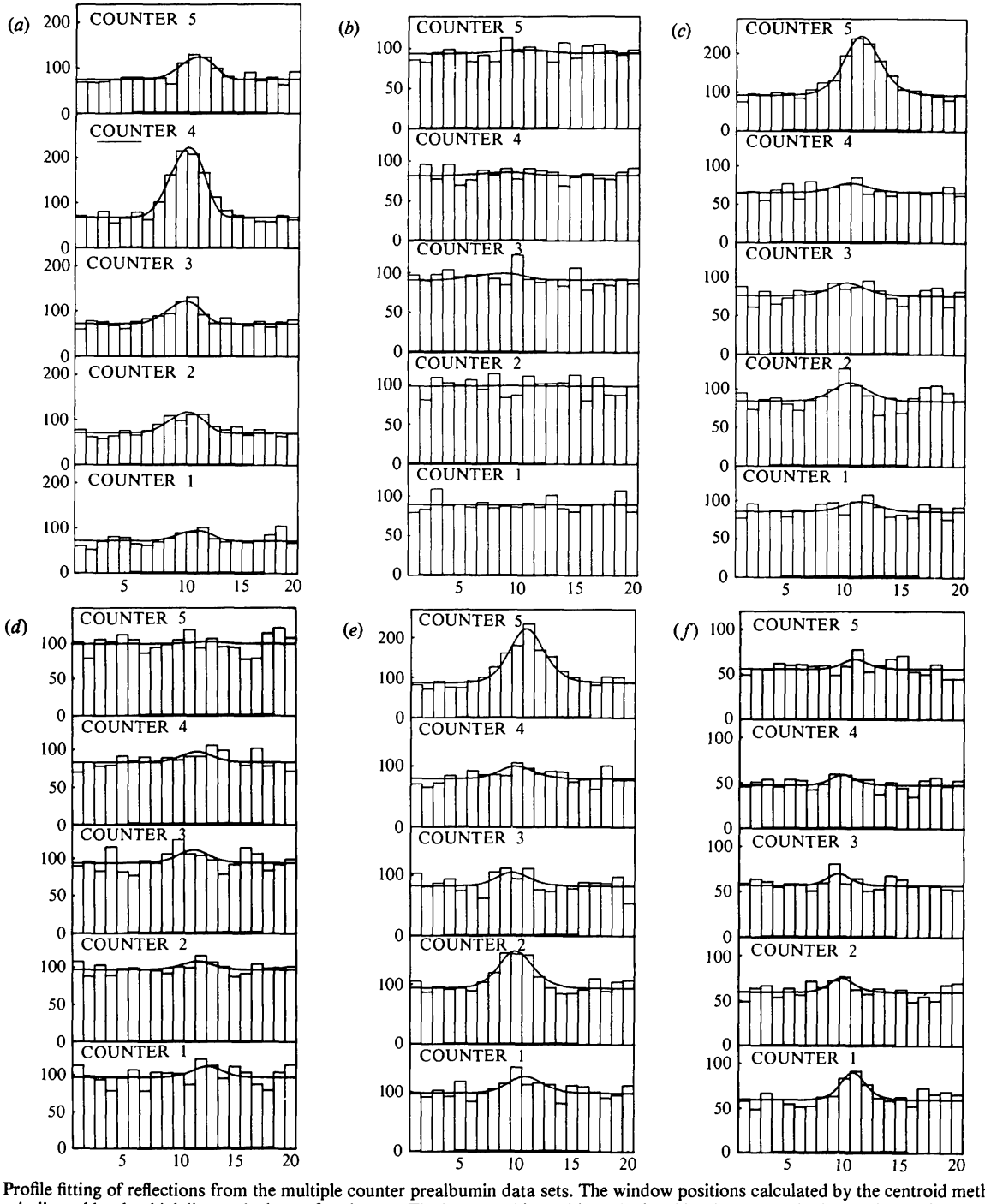


Fig. 5. Profile fitting of reflections from the multiple counter prealbumin data sets. The window positions calculated by the centroid method are indicated by the thick line at the base of each scan. The integrated intensities calculated by the two methods are given below:

Counter		(a)	1	2	3	4	5	(b)	1	2	3	4	5	(c)	1	2	3	4	5
Centroid method	:		65	229	189	581	175		11	−14	93	10	−2		17	−33	31	65	655
Profile fitting	:		86	174	182	567	182		0	4	44	22	24		55	96	70	48	622
			$\mu = 9{\cdot}24$, $\sigma = 3{\cdot}75$, $\gamma = -0{\cdot}76$, $\delta = 2{\cdot}43$, $G_f = 0{\cdot}0054$						$\mu = 8{\cdot}03$, $\sigma = 4{\cdot}25$, $\gamma = -0{\cdot}57$, $\delta = 1{\cdot}95$, $G_f = 0{\cdot}0040$						$\mu = 10{\cdot}63$, $\sigma = 3{\cdot}35$, $\gamma = -0{\cdot}11$, $\delta = 2{\cdot}07$, $G_f = 0{\cdot}0057$				
Counter		(d)	1	2	3	4	5	(e)	1	2	3	4	5	(f)	1	2	3	4	5
Centroid method	:		−18	15	86	135	−83		132	144	17	151	564		32	72	30	36	118
Profile fitting	:		56	42	64	53	12		112	247	95	84	545		91	47	40	34	33
			$\mu = 12{\cdot}04$, $\sigma = 2{\cdot}73$, $\gamma = 0{\cdot}08$, $\delta = 1{\cdot}89$, $G_f = 0{\cdot}0044$						$\mu = 9{\cdot}86$, $\sigma = 3{\cdot}38$, $\gamma = -0{\cdot}24$, $\delta = 2{\cdot}12$, $G_f = 0{\cdot}0045$						$\mu = 9{\cdot}91$, $\sigma = 2{\cdot}05$, $\gamma = -0{\cdot}14$, $\delta = 1{\cdot}75$, $G_f = 0{\cdot}0060$				

The further examples of profile fitting in Figs. 5(*c*)–(*f*), show that, even when the integration window is placed in the correct region of the scan, the random nature of counting events can cause the ordinate analysis and centroid methods to produce poor estimates of the integrated intensity. A careful comparison of these values, and those (frequently very different) values calculated by profile fitting, with the observed counts shows clearly the superiority of the profile method in describing the observed distributions. These comments apply equally to single-counter data.

A good test of the quality of the estimation of the integrated intensity is provided by the method used to determine scale factors between the five counters of the five-circle diffractometer. At the end of the main data collection, a separate batch of data is measured which consists of overlapping quintuplets of reflections; by stepping parallel to the collection axis, a particular reflection will be measured sequentially in each of the five counters (Evans, 1974). The scale factors between the counters are determined from the common reflections by the method of Fox & Holmes (1966) and usually deviate from unity by less than 5%. The merging residuals are evaluated for each of the ten combinations of counters (1 and 2, 1 and 3, *etc.*) and their weighted mean values are presented in Table 5 for four such batches of data measured on native and T_3-bound prealbumin. Profile fitting results in a considerable improvement for each data set. This is particularly marked for the T_3 c^* set which is not only weak but also has a high background level which is likely to introduce greater unrealiability into the centroid, or ordinate analysis, estimations.

4. Discussion

In the six years since this method was developed, it has become established in this laboratory as the standard method of evaluation of the integrated intensities of diffractometer data. We believe that the results described here amply demonstrate its power and potential.

Experience with our method suggests that it can provide reliable estimates of integrated intensities and their precision over the complete range of intensity values and that it is significantly better than some other methods for the weakest data. The accurate estimation of such data becomes increasingly important as more crystal studies are undertaken at very high resolution, or where only small crystals are available. Regrettably many workers regard weak reflections, usually those where $I/\sigma(I)$ is less than 2 or 3, to be 'not significant' and therefore exclude them from the data set. Such a criterion depends in any case on having a reliable estimate of the standard deviation, which, as we have pointed out earlier, is provided by our method but not by some others. However, although exclusion of these reflections may have the advantage of arbitrarily reducing the crystallographic residual, R, it is theoretically unjustified for diffractometer data (Hirshfeld & Rabinovich, 1973) and with suitable weighting it should be advantageous to include all measured reflections. In model calculations Hirshfeld & Rabinovich (1973) demonstrated how the exclusion of the weakest reflections introduces a bias which leads to underestimation of the thermal parameter and scale factor. Subsequently Arnberg, Hovmöller & Westman (1979) showed the improvement attainable in real structures through inclusion of all data, and also carried out a successful refinement of one structure using only data for which $I/\sigma(I) < 3{\cdot}3$, which had been originally discarded. Similarly, in protein refinement, underestimation of the thermal parameters occurs if data for which $I/\sigma(I) < 3$ are excluded (Oatley, 1981).

Since protein crystals are particularly sensitive to irradiation damage, one may well wish to collect data at the maximum possible rate. A number of methods have been proposed to accomplish this (*e.g.* Wyckoff, Doscher, Tsernoglou, Inagami, Johnson, Hardman, Allwell, Kelly & Richards, 1967; Hanson, Watenpaugh, Sieker & Jensen, 1979). These usually combine a limited number of counting steps in the region of the peak maximum with some form of averaged background; this is frequently assumed to be a function of 2θ only and is estimated by scanning between reciprocal-lattice rows, although local empirical corrections may be necessary (Hanson, Watenpaugh, Sieker & Jensen, 1979). Our method could operate similarly, but we prefer to measure some background in each scan to avoid assumptions about its behaviour; an accurate value is particularly important in the estimation of weak reflections. Rapid data collection may entail short count times or coarse stepping intervals; the use our method makes of its prior knowledge of peak shape and local background enables it to derive reliable intensities from such data, as in the

Table 5. *Comparison of the merging residuals for prealbumin counter–counter scaling data*

		Number of reflections	Centroid method $\langle I \rangle$	Centroid method $\langle R_m \rangle$	Profile fitting $\langle I \rangle$	Profile fitting $\langle R_m \rangle$
Native	b^* mount	740	2144	0·021	2106	0·016
	c^* mount	1695	2191	0·028	2196	0·019
T_3 complex	b^* mount	1090	761	0·037	746	0·028
	c^* mount	800	762	0·067	772	0·039

cases described in § 3 above. If our method has a fault, it is that it is much more time consuming than, say, ordinate analysis. On an ICL 1906A the present program requires 0·05 to 0·3 s to process a reflection, depending on the number of steps in the scan. Thus a complete data set may take 20 min to process. Our code is fairly efficient, but could undoubtedly be optimized further. However, we would argue that even at the present speed the improvement gained in the integrated intensities together with the more reliable indication of their precision more than justifies the computing costs.

Although our method was developed as, and is currently implemented as, an off-line data reduction procedure, its various features clearly give it the potential for on-line diffractometer control; indeed the solution method of the Appendix is derived from one developed in control theory (Aoki, 1967). We have shown that the method is able to model a wide variety of peak shapes. Furthermore it is able to account for changes in the width or the shape of the peaks which may occur through irradiation during data collection, and monitors any sudden transitions. Its ability to track crystal slippage would enable it to keep the peaks well centred in the scan. It can also identify and, if appropriate, make allowances for noise which may arise through deficiencies in the counting chains. Finally, because the method provides a good estimate of the precision of the integrated intensity, it would be easy to optimize data collection by measuring each reflection to constant relative precision or some similar criterion. [See Clegg (1981) for an on-line implementation of Diamond's (1969) profile fitting.]

We are grateful to the members of this laboratory for their data and their encouragement and to Professors D. V. Lindley and A. J. C. Wilson and Dr J. S. Rollett for helpful advice and criticism. SJO is a Mr and Mrs John Jaffé Donation Research Fellow of the Royal Society. Financial assistance was also provided by the Medical Research Council and the Hayward Foundation (SF).

APPENDIX
A brief description of the profile-fitting algorithm

As stated in the main body of the paper, we do not use the form of iterative linearization suggested in French (1978) to solve the three-stage non-linear model; the computation for large data sets would be prohibitive. Instead we use a block-diagonal approach, which we now develop. For generality in the following, suppose that there are M counters so that M reflections are scanned simultaneously. The algorithm uses the same peak parameters to fit the M reflections, and hence would need modification if the variation in shape between the reflections were sufficiently great to require different parameter sets for each. We shall use the subscript j to index the counters; thus c_{ji} is the count at step i of counter j and so forth. Moreover, x_{ji} is the position in the scan of step i of counter j, but with the origin of x chosen for each scan such that the peak positions have the same numerical value. (See Banner, Evans, Marsh & Phillips, 1977.) We shall also need the notation:

$k = K(L)M$	the index k increments from an initial value of K by intervals L while $k \leq M$;
$\nabla_{\alpha}\pi(x,\alpha)$	the gradient vector of $\pi(x,\alpha)$ with respect to α, x being fixed;
$\mathbf{q}_j$	the vector $(J_j,\beta_{j1},\beta_{j2})^T$ of true intensity and background parameters for counter j.

We begin by noting that we never need estimates of the λ_{ji} *per se*. Thus we may combine stages I and II to give [see (2.7) and (2.8)]:

$$\sqrt{c_{ji}} \sim N[\sqrt{v_{ji}}, 0{\cdot}25(1 + e^2 v_{ji})] \qquad (A.1)$$

for $j = 1(1)M$, $i = 1(1)N$. Next we note that

$$E(\sqrt{c_{ji}}) = \sqrt{v_{ji}} = [J_j \pi(x_{ji},\alpha) + \beta_{j1} + \beta_{j2} x_{ji}]^{1/2}.$$

Had the square root not been taken in order to stabilize the variance, this equation would have been linear in $\mathbf{q}_j = (J_j,\beta_{j1},\beta_{j2})^T$ for fixed α. This suggests an iterative procedure in each cycle of which first the $\mathbf{q}_j, j = 1(1)M$, are re-estimated for the fixed current estimate α and then α is re-estimated for the fixed updated estimates of $\mathbf{q}_j$. This is precisely what we do. The algorithm itself is summarized in Fig. 6. However, before we can discuss that in detail we need to summarize the solution of a Bayesian two-stage model. See Aoki (1967) for the proof of the following.

Lemma. Suppose that

$$\text{stage I:} \quad \mathbf{Y} \sim N(A_1\theta_1, C_1); \qquad (A.2)$$

$$\text{stage II:} \quad \theta_1 \sim N(A_2\theta_2, C_2). \qquad (A.3)$$

Then, after $\mathbf{Y} = \mathbf{y}$ has been observed, $P_{\theta_1}(.|\mathbf{y})$ is $N(\theta_1^*, D_1)$, where

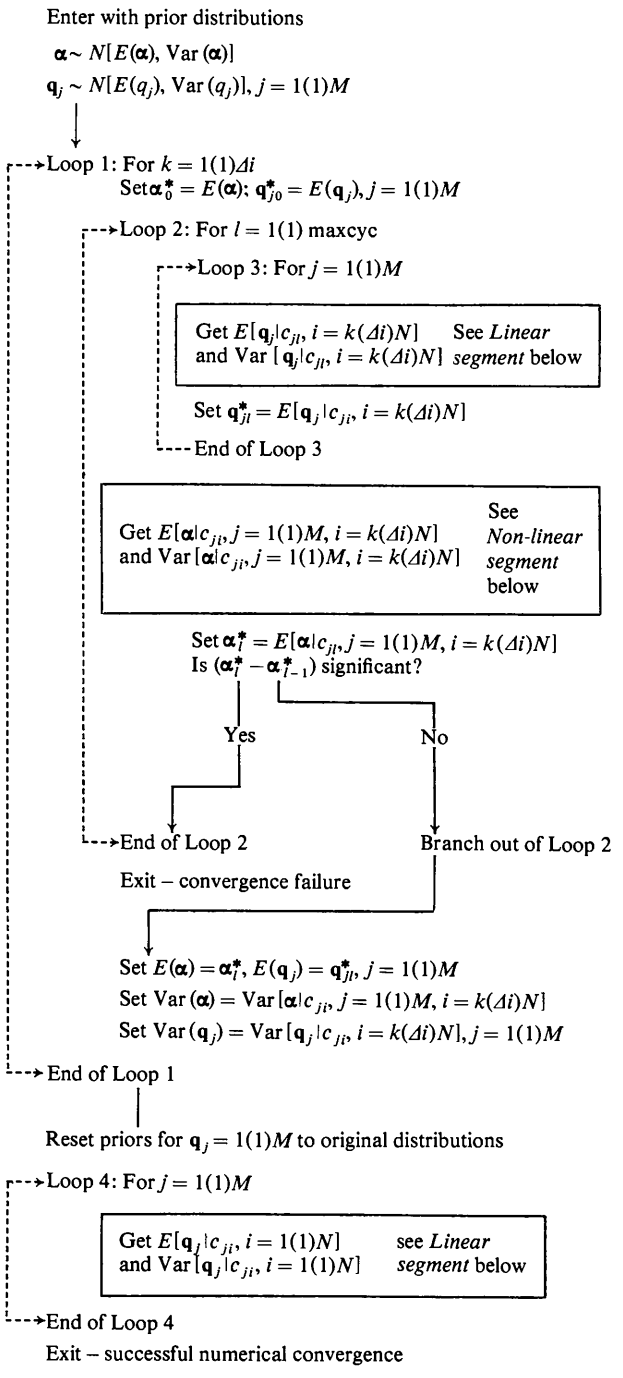
$$\theta_1^* = A_2\theta_2 + K(\mathbf{y} - A_1A_2\theta_2), \qquad (A.4)$$

$$K = D_1 A_1^T C_1^{-1}, \qquad (A.5)$$

$$D_1^{-1} = C_2^{-1} + A_1^T C_1^{-1} A_1. \qquad (A.6)$$

The matrix K is shown as the Kalman Filter because it 'filters out' the information contained in the discrepancy between the observation $\mathbf{y}$ and its prior expectation $A_1A_2\theta_2$.

It is easiest to explain the algorithm by initially ignoring the outer loop 1; whenever the expression $i = k(\Delta i)N$ occurs, read instead $i = 1(1)N$. The non-linear cycling is controlled by loop 2. If convergence does not occur in maxcyc iterations, the fitting is abandoned.



Linear segment

Stage I: $c_{ji} \sim N[\mathbf{a}_{ji}^T \mathbf{q}_j, \mathrm{Var}(c_{ji})]$ for defined range of i
Stage II: $\mathbf{q}_j \sim N[E(\mathbf{q}_j), \mathrm{Var}(\mathbf{q}_j)]$

where $\mathbf{a}_{ji}^T = [\pi(x_{ji}, \alpha_l^*), 1, x_{ji}]$

$$\mathrm{Var}(c_{ji}) = v_{ji}(1 + e^2 v_{ji}) + r_{ji}$$

$$r_{ji} = \begin{cases} 0 & \text{(Loop 3)} \\ \tilde{J}_j^2[\nabla_\alpha^T \pi(x_{ji}, \alpha^*) \mathrm{Var}(\alpha) \nabla_\alpha \pi(x_{ji}, \alpha^*)] & \text{(Loop 4)} \end{cases}$$

$$v_{ji} = \mathbf{a}_{ji}^T \mathbf{q}_{j(l-1)}^*$$

Initially the current approximation, α_0^*, to the posterior expectation of the peak parameters is set to the prior expectation. In loop 2 the intensities and backgrounds for each counter are first fitted (loop 3) to the observed counts assuming that the peak parameters are fixed at the values α_{l-1}^*. This fitting is achieved by using the Kalman filter summarized above in (*A*.4), (*A*.5), and (*A*.6) with stages I and II given in (*A*.2) and (*A*.3) identified with those described in the *linear segment* of Fig. 6. Note that the observations are taken to be the counts themselves and not their square roots. This is because numerical experiments have shown that convergence is less affected by problems of variance dependence than by errors introduced by linearizing the square root. Also, of course, the computation is reduced. In calculating the variance at each cycle, it is important to use the current predicted count, v_{ji}, and not the observed count c_{ji}. If c_{ji} is used, low observed counts get too high a weight and high observed counts get too low a weight. This can result in serious under-estimation of the background. Thus $\mathrm{Var}(c_{ji})$ is taken to be $v_{ji}(1 + e^2 v_{ji})$, where $v_{ji} = \mathbf{a}_{ji}^T \mathbf{q}_{j(l-1)}^*$. Initially background–peak–background estimates of the intensity are used so that v_{ji} takes reasonable values in the first cycle.

Having fitted the intensities and backgrounds, a linearized two-stage non-linear model is used to update the current approximation to the posterior expectations of the peak parameters. Here the square roots of the observed counts are fitted and not the counts themselves. This has been done because stabilizing the weighting scheme seems to give a greater radius of convergence to the algorithm. The observations and variances given in the description of the non-linear segment are calculated using the current approximations to the posterior means $E[q_j|c_{ji}, i = k(\Delta i)N]$ just found in loop 3. Many of the quantities calculated during loop 3 may be saved and used here without recalculation. Again the Kalman filtering equations (*A*.4), (*A*.5), and (*A*.6) are used to solve the model. To determine whether the non-linear cycling has converged, the shift $(\alpha_l^* - \alpha_{l-1}^*)$ is expressed in posterior standard deviations as given by $\mathrm{Var}[\alpha|c_{ji}, i = k(\Delta i)N, j = 1(1)M]$. If this is less than a predefined constant, convergence is assumed. This comparison is equivalent under the assumption of complete normality to checking whether the posterior probability distribution has shifted by less than a predetermined percentage. When

Non-linear segment

Stage I: $y_{ji} \sim N[\nabla_\alpha^T\{v_{ji}^{1/2}\}\alpha, \mathrm{Var}(y_j)]$ $\quad j = 1(1)M, i = k(\Delta i)N$
Stage II: $\alpha \sim N[E(\alpha), \mathrm{Var}(\alpha)]$

where $y_{ji} = c_{ji}^{1/2} - v_{ji}^{1/2} + \nabla_\alpha^T\{v_{ji}^{1/2}\}\alpha_{l-1}^*$

$\mathrm{Var}(y_{ji}) = (1 + e^2 v_{ji})/4.$

Fig. 6. Flow diagram of fitting algorithm. The notation is explained in the text.

convergence occurs, processing branches out of loop 2 and the distributions are updated to take account of the fitted data. *N.B.* the prior distributions for $\mathbf{q}_j$ and $\boldsymbol{\alpha}$ have been unchanged until now. To update them iteratively in loop 2 would be to lose much of the information that they contain and thus be completely wrong.

It is now convenient to explain loop 1. If there are many steps in each scan this algorithm can require much core store. In order to reduce this we take advantage of a property of Bayesian estimation and, in particular, the Kalman filter; we assimilate the data in blocks. First steps 1, $(\Delta i + 1)$, $(2\Delta i + 1)$, ... in each scan are fitted; then steps 2, $(\Delta i + 2)$, $(2\Delta i + 2)$, ... and so on until all the counts have been fitted. It is important to pick the steps in each block in this manner; otherwise unfair sampling of the profile can lead to biases in the final fitted parameters.

Now that all the data have been fitted, the intensities and backgrounds are completely re-estimated in loop 4. Naturally to do so their prior distributions are reset to their original values. This re-estimation ensures that posterior distributions for the intensities contain *all* the available peak parameter information. Furthermore, it eliminates any bias that may have crept in through assimilating the data in blocks. There is one very important difference between the two-stage model used here and that in loop 3: the observations variance now includes a contribution from the uncertainty in the peak parameters and, hence, the uncertainty in $\pi(x_{ji}, \boldsymbol{\alpha})$. One of the advantages we claim for this profile fitting is that the posterior variance for J reflects all the available information faithfully. In partitioning the two-stage model into a linear and a non-linear segment, the very correlations that do this are lost. The term r_{ji} here recaptures this information. It is simply derived from the induced variance in the peak shape function. The term $\tilde{J}_j$ is the quick background–peak–background estimate of the intensity referred to above. Theoretically, it would appear to be better to use the current estimate from profile fitting. However, if something untoward has happened and, although numerically the process has converged, practically the result is insane, then inflating the variance according to the profile-fitted value can cause overflow problems before the controlling program can intervene. It should be noted that the observation variance may be similarly inflated in loop 3. Indeed doing so slightly improves the radius of convergence of the algorithm, but at the cost of extra computation.

We have compared the performance of this algorithm with the full solution of the three-stage non-linear model as described in French (1978). This algorithm is significantly faster and achieves essentially identical numerical results (French, 1975). In particular, the use of the quantity r_{ji} to allow for lost correlations is extremely successful.

References

AOKI, M. (1967). *Optimization of Stochastic Systems.* New York: Academic Press.

ARNBERG, L., HOVMÖLLER, S. & WESTMAN, S. (1979). *Acta Cryst.* A**35**, 497–499.

ARTYMIUK, P. J., BLAKE, C. C. F., RICE, D. W. & WILSON, K. S. (1982). *Acta Cryst.* B**38**, 778–783.

BANNER, D. W., EVANS, P. R., MARSH, D. J. & PHILLIPS, D. C. (1977). *J. Appl. Cryst.* **10**, 45–51.

BLAKE, C. C. F., GEISOW, M. J., OATLEY, S. J., RÉRAT, B. & RÉRAT, C. (1978). *J. Mol. Biol.* **121**, 339–356.

CLEGG, W. (1981). *Acta Cryst.* A**37**, 22–28.

CUTFIELD, J. F., CUTFIELD, S. M., DODSON, E. J., DODSON, G. G., EMDIN, S. F. & REYNOLDS, C. D. (1979). *J. Mol. Biol.* **132**, 85–100.

DIAMOND, R. (1969). *Acta Cryst.* A**25**, 43–55.

DODSON, E. J,. DODSON, G. G., HODGKIN, D. C. & REYNOLDS, C. D. (1979). *Can. J. Biochem.* **57**, 469–479.

DODSON, E. J., DODSON, G. G., LEWITOVA, A. & SABESAN, M. (1978). *J. Mol. Biol.* **125**, 387–396.

DODSON, E. J. & VIJAYAN, M. (1971). *Acta Cryst.* B**27**, 2402–2411.

EVANS, P. R. (1974). D.Phil. Thesis, Univ. of Oxford.

FRENCH, S. (1975). D.Phil. Thesis, Univ. of Oxford.

FRENCH, S. (1978). *Acta Cryst.* A**34**, 728–738.

FRENCH, S. & OATLEY, S. (1981). In *Problems and Progress in Crystallographic Statistics*, edited by A. J. C. WILSON & S. RAMASESHAN. In the press.

HANSON, J. C., WATENPAUGH, K. D., SIEKER, L. & JENSEN, L. H. (1979). *Acta Cryst.* A**35**, 616–621.

HIRSHFELD, F. L. & RABINOVICH, D. (1973). *Acta Cryst.* A**29**, 510–513.

JOHNSON, N. L. (1949). *Biometrika*, **36**, 149–168.

LEHMANN, M. S. & LARSEN, F. K. (1974). *Acta Cryst.* A**30**, 580–584.

LINDLEY, D. V. & SMITH, A. F. M. (1972). *J. R. Stat. Soc. B*, **341**, 1–41.

OATLEY, S. J. (1976). D.Phil. Thesis, Univ. of Oxford.

OATLEY, S. J. (1981). Unpublished results.

OATLEY, S. J. & BURRIDGE, J. M. (1981). Unpublished results.

TICKLE, I. J. (1975). *Acta Cryst.* B**31**, 329–331.

WATSON, H. C., SHOTTON, D. M., COX, J. M. & MUIRHEAD, H. (1970). *Nature* (*London*), **225**, 806–811.

WYCKOFF, H. W., DOSCHER, M., TSERNOGLOU, D., INAGAMI, T., JOHNSON, L. N., HARDMAN, K. D., ALLWELL, N. M., KELLY, D. M. & RICHARDS, F. M. (1967). *J. Mol. Biol.* **27**, 563–578.