Journal of
Applied
Crystallography

ISSN 0021-8898

Received 1 February 2001
Accepted 9 March 2001

Background estimation using a robust Bayesian analysis

W. I. F. David* and D. S. Sivia

ISIS Facility, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England.
Correspondence e-mail: bill.david@rl.ac.uk

A novel method for the estimation of the background in a powder diffraction pattern has been developed using a robust Bayesian analysis. In formulating a probabilistic approach to background fitting, the diffraction peaks are considered to be nuisance data that must be taken into account. The underlying probability theory is discussed in terms of going beyond the Gaussian approximation normally associated with counting statistics and least-squares analysis. Various examples are presented that illustrate the general applicability of this approach.

1. Introduction

The successful removal of a background signal from a diffraction pattern without the availability of a suitable calibration measurement relies heavily on the ability to distinguish between broad and sharp features. While a procedure for the simultaneous estimation of the slowly varying contribution and deconvolution of the sharp structure has been put forward (Sivia, 1990), most of the work in this area has focused solely on the extraction of the background signal. One of the early proposals within crystallography was proffered by Steenstrup (1981), who developed an iterative procedure for fitting a low-order polynomial to the data using a least-squares analysis. The success of this method hinged on the successive removal of more and more data that were deemed to include a significant contribution from a Bragg peak. Other approaches have been suggested and a recent ingenious algorithm developed by Brückner (2000) that is based upon an adaptive low-pass filter is worthy of mention.

In this paper, a more formal probabilistic approach to the problem is developed, which correctly takes into account the error-bar on each data point while allowing for an unknown contribution from a Bragg peak that may or may not be present at that point in the diffraction pattern. This approach leads to a functional minimization that deviates from least-squares analysis in an asymmetrical manner. It is not iterative in the Steenstrup (1981) sense because no data are removed from the analysis. Indeed, the principal appeal of this algorithm is that all the measurements are treated on an equal footing. The approach developed in this paper bears a close similarity to the one described by Fischer *et al.* (2000) and, though not as rigorous in the full optimization of the ideal background parameterization, it is, we believe, more intuitively accessible.

2. Dealing with outlier data: beyond least squares

In a powder diffraction experiment, the uncertainty associated with a single observation in a diffraction pattern is generally considered to result solely from Poisson counting statistics. If more than around 20 counts have been measured at a particular point, then the Gaussian approximation to counting statistics is quite sufficient and least-squares analysis is the appropriate minimization procedure for fitting the diffraction pattern. [The issue of dealing with low count rates (<20 counts per point) has been treated elsewhere (Antoniadis *et al.*, 1990) and will not be discussed in this paper.] Least-squares minimization is often taken to be a basic tenet of data analysis. The almost universal use of least-squares analysis is because the underlying statistical assumption is a very general one and has its origins in the assumption that the uncertainties associated with an observation follow a Gaussian probability distribution. In other words, the observation of a data point, D, with an error bar, σ, can be stated mathematically in terms of a likelihood that follows a Gaussian probability distribution function:

$$p(D|\mu,\sigma) = [1/\sigma(2\pi)^{1/2}]\exp[-(D-\mu)^2/2\sigma^2]. \qquad (1)$$

The quoted error bar is usually based upon ideal conditions and in the case of a powder diffraction pattern is generally associated with counting statistics alone. Occasionally, however, rogue data points resulting from, for example, detector problems can occur. Based upon a Gaussian probability distribution derived solely from counting statistics, these points must be extremely unlikely. However, their very appearance testifies to the fact that conditions other than just counting statistics are causing uncertainties in the data value. When this occurs, an alternative approach would be to state that the error bar, σ, is a lower bound, $\sigma_{\min}$, on the uncertainty in the data point, D. With a suitably pessimistic upper bound, $\sigma_{\max}$, a Jeffreys (1939) probability distribution function can be assigned for the now uncertain error bar,

$$p(\sigma|\sigma_{\min}, \sigma_{\max}) = [1/\ln(\sigma_{\max}/\sigma_{\min})]\,(1/\sigma). \qquad (2)$$

This holds for $\sigma_{\min} \leq \sigma \leq \sigma_{\max}$ and is zero otherwise. (The Jeffreys probability distribution function has the attractive property that it is scale invariant.) The likelihood for the data can then be expressed in terms of the two probability distribution functions discussed above as follows:

$$p(D|\mu, \sigma_{\min}, \sigma_{\max}) = \int_0^\infty p(D|\mu, \sigma)\, p(\sigma|\sigma_{\min}, \sigma_{\max})\, \mathrm{d}\sigma, \qquad (3)$$

where the product rule has been used to expand the joint probability distribution function for D and σ on the right-hand side. Substituting equations (1) and (2) into (3) and performing the resulting integration leads to

$$\begin{aligned} p(D|\mu, \sigma_{\min}, \sigma_{\max}) = {} & [1/(D-\mu)\ln(\sigma_{\max}/\sigma_{\min})] \\ & \times \big\{\mathrm{erf}[(D-\mu)/\sigma_{\min}2^{1/2}] \\ & - \mathrm{erf}[(D-\mu)/\sigma_{\max}2^{1/2}]\big\}. \qquad (4) \end{aligned}$$

Consideration of equation (4) indicates that when $\sigma_{\max}$ is more than an order of magnitude larger than $\sigma_{\min}$ (and this effectively describes an outlier) then the second error function may be dropped and the probability distribution function may be well approximated by the simpler function

$$p(D|\mu, \sigma \geq \sigma_{\min}) \propto [1/(D-\mu)]\, \mathrm{erf}[(D-\mu)/\sigma_{\min}2^{1/2}]. \qquad (5)$$

This likelihood function is illustrated in Fig. 1(*a*) (as a dashed line) where it is compared with the equivalent Gaussian probability distribution function (dotted line). The new robust function has the two important features that (i) the width of the central bump is principally controlled by $\sigma_{\min}$ and (ii) the decay in the tails is determined by allowed values of σ that are much larger than $\sigma_{\min}$. Note that the new distribution function is not simply a stretched Gaussian distribution. For small differences, the distribution is Gaussian in nature with a central bump that is about twice the width of the equivalent 'least-squares' Gaussian, whereas the tails of the distribution decay like a Jeffreys prior. The log (likelihood) distributions are shown in Fig. 1(*b*). The quadratic 'least-squares' behaviour of the conventional Gaussian distribution is clearly seen (dotted line), while the tails of the robust distribution (dashed line) decay logarithmically.

Consider the data shown in Fig. 2(*a*). The points appear to lie on a straight line and indeed both least-squares (dotted line) and robust statistics (dashed line) fitting give similar, acceptable results. Imagine now that the experiment has been repeated but unfortunately through problems with instrumentation several rogue points have now appeared in the data (see Fig. 2*b*). While manual removal of these outlier points is a practical procedure for small data sets, the subjective challenge of manual removal for very large data sets is substantial. Least-squares fitting of a straight line to the data presented in Fig. 2(*b*) leads to a poor fit to the data. The result, however, is correct within the least-squares approximation since all the data points have been considered to have similar errors resulting solely from standard measurement errors and counting statistics. However, the rogue data points have occurred because there are other problems in the data collection that have not been considered. Mathematically, this can be tackled if the error bars shown in Fig. 2(*b*) are presumed to be the lower bound on the uncertainty. The robust minimization then has no problem in fitting the full data set and finding a solution that passes through the 'good' data points. This arises because the logarithmic tails in the robust likelihood are much more tolerant of substantial discrepancies than the least-squares analysis; the penalty associated with having a fitted line well away from the rogue data is much smaller than with traditional least-squares analysis.

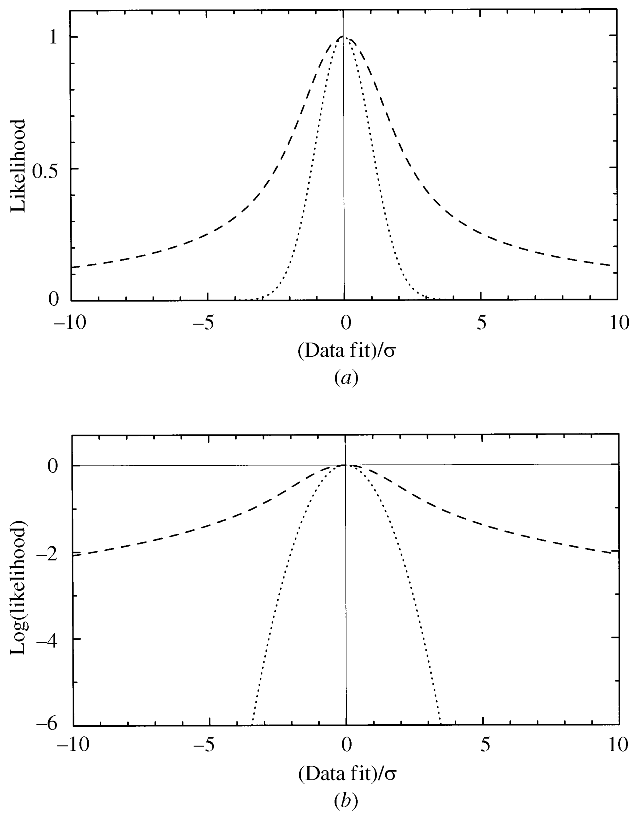

Figure 1
The likelihood functions (*a*) and their logarithms (*b*) of equations (1) (dotted line) and (4) (dashed line) with a mean of zero ($\mu = 0$) and a standard deviation of one ($\sigma = 1$). The dotted line is the standard Gaussian probability distribution function that penalizes outliers severely, while the dashed line corresponding to the robust probability distribution has long tails.

3. Background determination: discriminating against Bragg peaks

Although the outliers shown in Fig. 2(*b*) are single rogue values, it is clear that, with some imagination, the figure bears a passing resemblance to a high-resolution powder diffraction pattern. Given that the order of appearance of the outliers is unimportant, the robust fit to the data can be seen to have

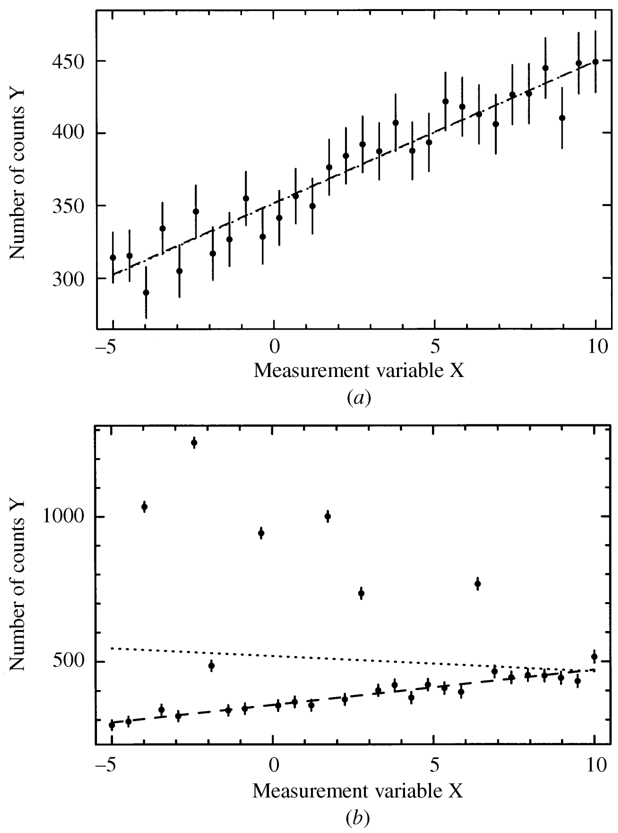

Figure 2
This figure shows a straight line fit to data that should generally follow a straight-line behaviour but occasionally contain a large rogue signal. The dotted line shows the traditional least-squares solution, while the dashed line is the result of a simple Bayesian analysis that allows for the possibility of outliers.

similarities to the fitting of a background in a powder diffraction pattern. Taking this analogy further, if Bragg peaks are considered to fulfil the role of positive outliers, then the probability distribution function associated with background estimation may be approximated to a standard Gaussian distribution for data points, D, below the background, B. In other words, with $z = (D - B)/\sigma$, then

$$p(D|z \le 0) \propto \exp(-z^2/2). \qquad (6a)$$

For points above the background, the robust probability distribution [equation (4)] will discriminate against Bragg peaks. This gives

$$p(D|z \ge 0) \propto (1/z\sigma)\,\mathrm{erf}(z/2^{1/2}). \qquad (6b)$$

Maximizing this asymmetrical probability distribution function is equivalent to least-squares minimization for points below the background and robust minimization for points above the background. The log (likelihood) distribution is shown in Fig. 3 as a dashed line. Specifically, the function minimized is

$$f(z) = \begin{cases} z^2 & \text{for } z \le 0, \\ 6\ln[2z/\pi^{1/2}\,\mathrm{erf}(z/2^{1/2})] & \text{for } z \ge 0. \end{cases} \qquad (7)$$

Although the above argument is based upon a discussion of modified probability distribution functions, the approach is still rather heuristic. A logical Bayesian analysis may be developed from considering the available information associated with a particular data point, D. The Bragg peaks are not considered to be outliers. Indeed, it is assumed that there are no rogue data and therefore that the error bar, σ, is known with confidence. This is equivalent to stating that the mean square difference between a good model, M, and the data value, D, should be given by

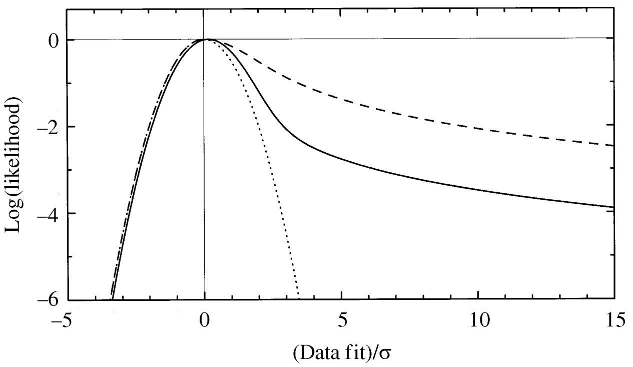

Figure 3
This figure shows three log(likelihood) distributions associated with background estimation. The dotted line is a quadratic function associated with a Gaussian probability distribution function and standard least squares analysis. The dashed line is a combination of least squares (for points below the fitted line) and robust outlier estimation (for points above the fitted line). The solid line is the log(likelihood) distribution obtained from a Bayesian analysis that marginalizes out the effects of Bragg peaks.

$$\langle (M - D)^2 \rangle = \sigma^2. \qquad (8)$$

In the present analysis, all that is known is that the model is equal to the sum of the (always positive) peak contribution, A, and the background, B. Although, it is an anathema for crystallographers to treat Bragg peaks as a nuisance, this is precisely what needs to be performed since it is the background that is of interest and not the Bragg peak contribution. Thus the probability distribution function for the background given the data, D, and all other information (labelled I), such as the Bragg peak positivity, is simply the integral over all possible peak profile values:

$$p(B|D, I) = \int\limits_0^\infty p(A, B|D, I)\,\mathrm{d}A. \qquad (9)$$

Invoking Bayes' theorem and separating peak and background distribution gives

$$\begin{aligned} p(B|D, I) &\propto \int\limits_0^\infty p(D|A, B, I)\,p(A, B|I)\,\mathrm{d}A \\ &= p(B|I) \int\limits_0^\infty p(D|A, B, I)\,p(A|I)\,\mathrm{d}A. \qquad (10) \end{aligned}$$

A priori, it is difficult to scale the Bragg peak contribution relative to the background and a reasonable assumption for the probability distribution for A is again the scale-invariant Jeffreys distribution:

$$p(B|D, I) \propto \int_0^\infty A^{-1} \exp[-(A + B - D)^2/2\sigma^2]\, dA. \qquad (11)$$

This equation may be simplified by writing $D - B$ as Δ and putting $u = A/\sigma 2^{1/2}$ and $t = \Delta/\sigma 2^{1/2}$. The probability distribution for the background reduces to

$$p(B|D, I) \propto \int_0^\infty u^{-1} \exp[-(u - t)^2]\, du. \qquad (12)$$

This integral behaves pathologically as $u \to 0$, which is a consequence of the scale invariance of the Jeffreys distribution for small u. The integral is, however, well behaved if the lower limit is chosen to be close to zero. This is equivalent to the reasonable presumption that the lower limit of A is finite and positive but much smaller than σ. The precise choice of lower limit is unimportant since the function, to within a scale factor, varies only slowly as a function of the lower cut-off point of the integral. The log (likelihood) distribution associated with marginalizing out the Bragg peaks is shown in Fig. 3 as a solid line. It appears to be intermediate with respect to the standard least-squares analysis and the robust outlier distribution. It is straightforward to show that this log (likelihood) also has the desirable characteristics of falling off logarithmically for large positive Δ and quadratically for positive Δ. Moreover, for $\pm 2\sigma$, the distribution behaves very similarly to a quadratic least-squares model and thus for regions where there is only background the fit should essentially be equivalent to a least-squares analysis.

In implementing this new log (likelihood) function, the logarithm of equation (12) was evaluated using the trapezium rule. The resulting log (likelihood) function was calculated by interpolation from a table and minimized using a simplex procedure. The optimized spline and log (likelihood) routines amount to some 200 lines of Fortran code.

The analysis presented in this section can be viewed as a background-biased least-squares analysis. In order to evaluate the background, a smoothly varying function must be fitted using the algorithm described above. A variety of options may be implemented. Steenstrup (1981) used Ralston polynomials, which are a particularly appropriate set of polynomials since they are orthogonal with respect to the weights associated with the individual data points. Other more common orthogonal polynomials, such as the Chebyshev polynomials, may also be used (Press *et al.*, 1992). In this paper, cubic splines have been used to describe the background function. Cubic splines may be viewed as an extension of linear interpolation (Press *et al.*, 1992) and are piecewise continuous functions defined by ordinate and coordinate values at knot points across the diffraction pattern. The important advantage that splines have over linear interpolation is that the background function is not only continuous but also continuously smooth across the diffraction pattern. Splines can offer an advantage over orthogonal polynomials in that a full diffraction pattern that has a highly structured background in a limited region may be fitted with fewer spline points than orthogonal polynomial values.

In implementing the spline interpolation algorithm, the number of knots may be varied and their positions as well as their coordinate values may be refined. In the original version of the program, knot positions were indeed refined. However, when the knot values were fixed at a constant spacing across the diffraction pattern and only coordinate values refined, optimization of the code allowed an order of magnitude increase in speed without any discernible degradation in background estimation. In the final implementation of the code, therefore, the user only needs to decide upon the number of knots. Typically, across a complete diffraction pattern, the number of knots varies between five, for little structure, and twenty, for a highly structured background. The precise value is not crucial and is left for the user to assess by eye.

4. Examples

Powder diffraction patterns from different materials and diffractometers can manifest very different characteristics. The background in a high-resolution synchrotron diffraction pattern can generally be easily evaluated, whereas measurements of a poorly crystalline phase may consist of very broad Bragg peaks that make the background and Bragg peaks difficult to distinguish from one another. Although backgrounds are generally slowly varying, the presence of an amorphous phase or glass capillary may produce a background (from the viewpoint of the Rietveld refinement of a crystalline phase) that is quite highly structured. Similarly, very high counting statistics can show pronounced features, such as diffuse scattering, that are not visible in a rapid measurement. On the other hand, poor counting statistics may make it difficult to distinguish between peak and background. In the remainder of this section, four examples are presented that cover most of these eventualities and indicate that the consistent Bayesian approach to background analysis is broadly applicable to powder diffraction.

4.1. Orthorhombic zopiclone

The powder diffraction pattern shown in Fig. 4 represents a high-resolution X-ray diffraction data set of the orthorhombic form of the pharmaceutical compound zopiclone. The data were collected on beamline BM16 (ESRF, Grenoble) at a wavelength of 0.8 Å in a 1 mm capillary over a period of 30 min. The relatively short counting time means that, although the Bragg peaks are sharp, the background is rather noisy and thus weak peaks are relatively difficult to distinguish from the background. At low angles, the background determination is good, with weak peaks clearly visible above the calculated background. At higher angles, the ability to distinguish between background and weakly overlapping Bragg peaks is more difficult and although the background estimation is generally excellent, there are two regions at 2θ = 17 and 19° where the background has been slightly overestimated. This can be clearly seen in the expanded view in Fig. 4(b) of the background region of the diffraction pattern. It

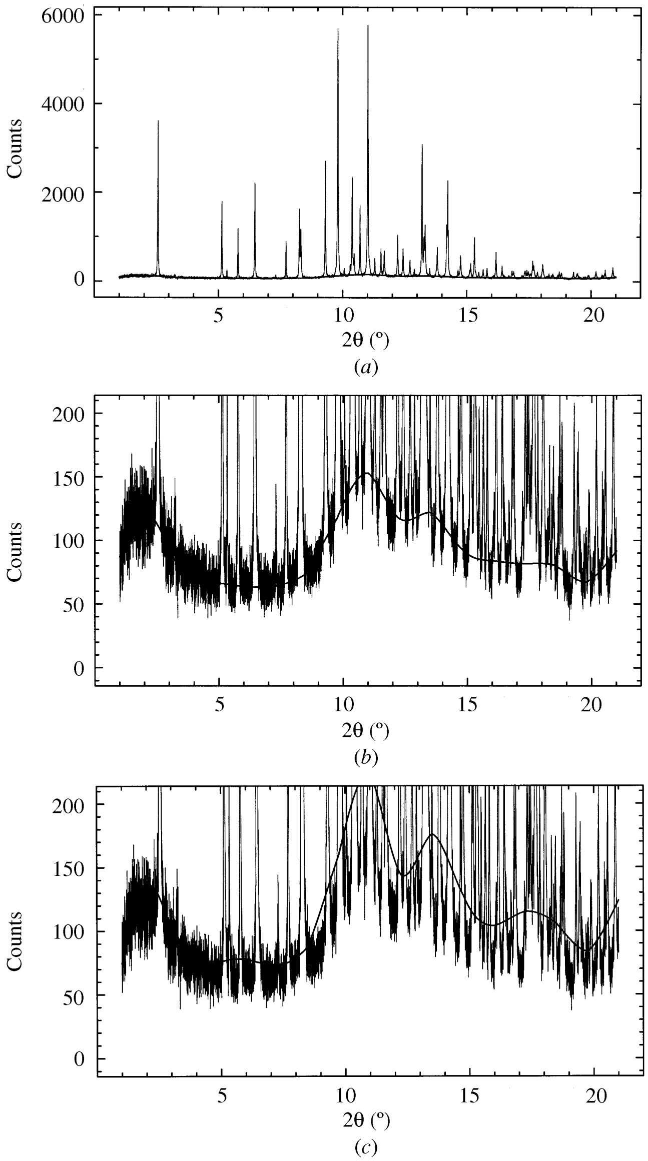

Figure 4
Diffraction data from a sample of orthorhombic zopiclone collected on BM16 at the ESRF, Grenoble, using a wavelength of 0.8 Å. The full diffraction pattern is shown in (*a*). Comparison of (*b*) (Bayesian background analysis using equation (6) and (*c*) (standard least-squares analysis) highlights the excellent fit to the background that is obtained by using an appropriate probability distribution function. Fifteen knots were used to describe the cubic spline function.

is worth comparing this background-biased fit to the data with the traditional least-squares fit to the same data, which is shown in Fig. 4(*c*). At low angles (below 5°), the least-squares fit follows the background well for the simple reason that the region is almost entirely background. However, at higher angles, the least-squares analysis fits neither the background nor the Bragg peaks, but the appropriate weighted average. Large deviations from the least-squares fitted line are severely penalized; this has the effect of lifting the fitted line above the background. With the background-biased fit, large deviations from fitting the Bragg peaks are only logarithmically penalized and thus the optimized function more closely follows the true background value.

4.2. Tetracycline hydrochloride

This material was one of two compounds that were used as part of an informal round robin test for structure solution from powder diffraction data (Le Bail & Cranswick, 1998). The data, represented in Fig. 5, were collected on beamline 9.1 at the SRS Daresbury for 5 h in a 0.5 mm glass capillary at a wavelength of 0.692 Å using a developmental image-plate system (Roberts *et al.*, 1998). The line shape is essentially symmetrical with a minimum full width at half-maximum of around 0.05°. The combination of long counting time and image-plate system leads to extremely good counting statistics, which in turn means that the background from the glass capillary is very highly structured. This is a difficult background problem to tackle automatically and in fact a better approach would be to fit the background arising from the glass capillary analytically using an appropriate pair distribution function. Nevertheless, the approach outlined in this paper gives a very good assessment of the correct background position with only a small number of systematic errors at higher angles.

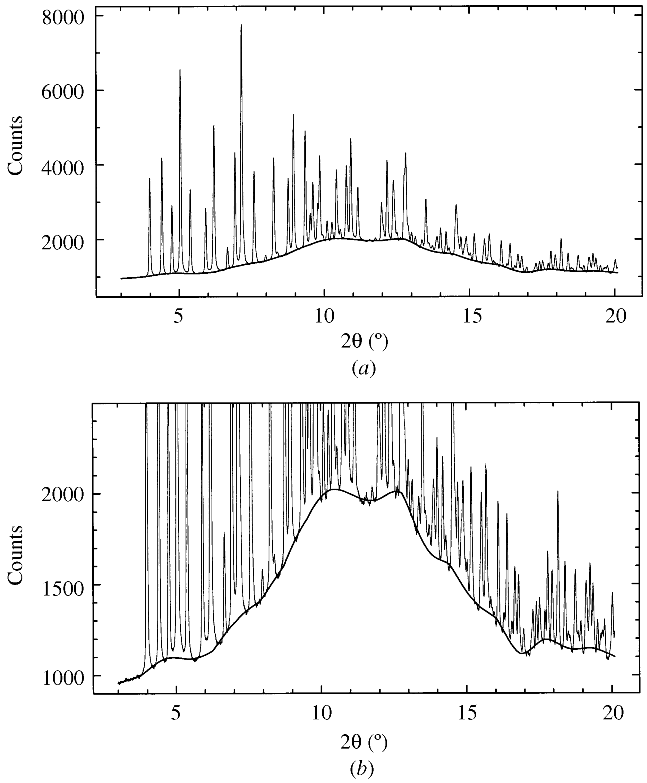

Figure 5
Diffraction data from a sample of tetracycline hydrochloride collected using an image plate on beamline 9.2 at SRS Daresbury (see http://sdpd.univ-lemans.fr/SDPDRR/). The full diffraction pattern is shown in (*a*) and expanded to highlight the background in (*b*). Twenty knots were used to describe the cubic spline function.

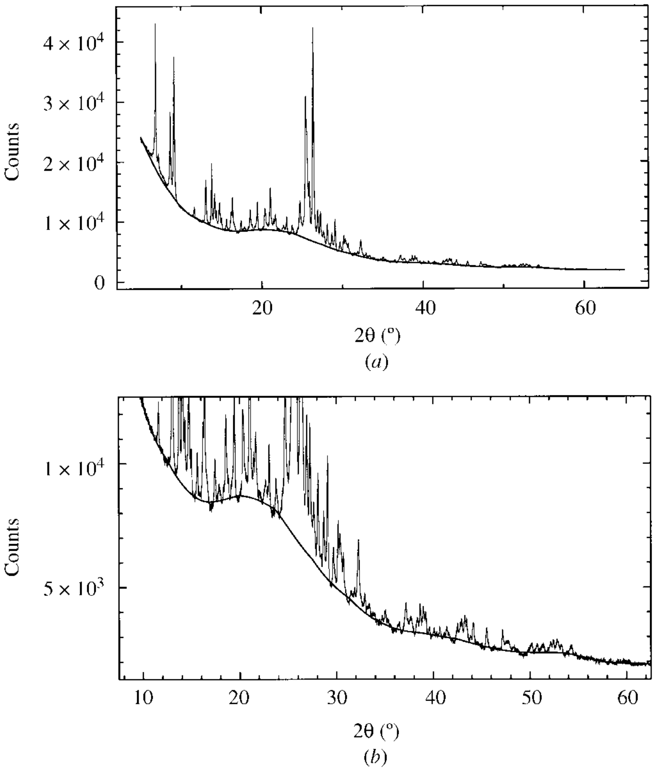

Figure 6
Laboratory diffraction data from a sample of berberine chloride obtained using a Bruker D8 diffractometer and Cu $K\alpha_1$ radiation, and fitted with a position-sensitive detector. The full diffraction pattern is shown in (*a*) and expanded to highlight the background in (*b*). Fifteen knots were used to describe the cubic spline function.

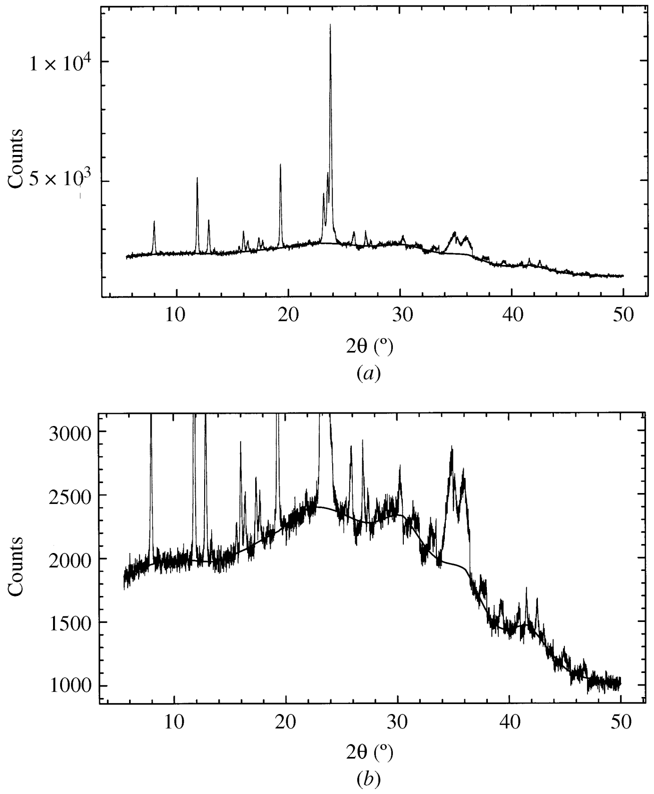

Figure 7
Diffraction data from a sample of 1,4-diethynyl-2,5-bis(octyloxy)benzene collected on a Stoe X-ray powder diffractometer equipped with a linear position-sensitive detector using Cu $K\alpha_1$ radiation (1.5406 Å) (from Tedesco *et al.* 2001). Fifteen knots were used to describe the cubic spline function. The full diffraction pattern is presented in (*a*) and the background region is highlighted in (*b*).

4.3. Berberine chloride

The laboratory data shown in Fig. 6 were collected on a Siemens D8 diffractometer in capillary geometry using a position-sensitive detector with monochromatic Cu $K\alpha_1$ radiation. The large differences between the high background at low angles and low background at high angles are in general well handled by a fifteen-knot spline function. The only significant misfitting occurs around the first two clumps of peaks where highly structured diffuse scattering has too much curvature to be fitted by only a fifteen-knot spline.

4.4. 1,4-Diethynyl-2,5-bis(octyloxy)benzene

Diffraction data from a sample of 1,4-diethynyl-2,5-bis(octyloxy)benzene were collected on a Stoe X-ray powder diffractometer equipped with a linear position-sensitive detector using Cu $K\alpha_1$ radiation (1.5406 Å) (from Tedesco *et al.*, 2001). The full diffraction pattern is presented in Fig. 7(*a*) and the background region is highlighted in Fig. 7(*b*). This is the worst of the four examples and the data quality is poor. The counting statistics are low, there are few Bragg peaks and the background levels are high compared with the Bragg peaks, particularly above $2\theta = 25°$. Nevertheless, using fifteen knots, the background has been well determined, even above $2\theta = 30°$ where the broad Bragg peaks almost merge with the background. Although the statistics are poor, there is still substantial structure in the background and background evaluation with a small number of knots (less than ten) gives relatively poor results.

5. Conclusions

Background estimation in a powder diffraction pattern is not an exact science. Without a detailed knowledge of all the contributions to both the background and the Bragg peaks, it is impossible to determine precisely the true background. However, if all the physical causes of Bragg peak area and shape are understood and correctly accounted for, then the background may, in principle, be determined. Alternatively, if all the contributions to the background are understood and accounted for, then the background may be determined using the appropriate physically based equations. It is important to re-emphasize that the Bayesian process outlined in this paper is one of background estimation rather than determination. The challenge with all estimation techniques is to make the estimation as close to a correct determination as possible. Without a physical model, however, there is no complete guarantee that the background is correct, particularly in areas of broad peaks or substantial peak overlap. Interestingly, despite the development of sophisticated estimation algorithms (see below and Steenstrup, 1981) and ingenious filters (Brückner, 2000), estimation by eye is still one of the best approaches. However, when a degree of automation is required, the algorithm presented in this paper along with the developments of other authors such as Steenstrup (1981) and Brückner (2000) offer practical approaches to the background estimation process.

We acknowledge Drs K. Shankland and N. Shankland for help in the provision of the diffraction data for zopiclone, Dr A. Florence for the berberine chloride data and Dr J. P. Attfield for the diffraction data of 1,4-diethynyl-2,5-bis(octyloxy)benzene. The data set for tetracycline hydrochloride was part of a structure determination round robin (Le Bail & Cranswick, 1998) and was obtained from http://www.cristal.org/SDPDRR/download.html.

References

Antoniadis, A., Berruyer, J. & Filhol, A. (1990). *Acta Cryst.* A**46**, 692–711.

Brückner, S. (2000). *J. Appl. Cryst.* **33**, 977–979.

Fischer, R., Dose, V., Hanson, K. M. & von der Linden, W. (2000). *Phys. Rev. E*, **61**, 1152–1160.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.

Le Bail, A. & Cranswick, L. M. D. (1998). *Structure Determination by Powder Diffractometry Round Robin*, http://sdpd.univ-lemans.Fr/SDPDRR/.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes*. Cambridge University Press.

Roberts, M. A., Finney, J. L. & Bushnell-Wye, G. (1998). *Mater. Sci. Forum*, **278–281**, 318–322.

Sivia, D. S. (1990). *Maximum Entropy and Bayesian Methods*, edited by P. F. Fougère, pp. 195–209. Dordrecht: Kluwer.

Steenstrup, S. (1981). *J. Appl. Cryst.* **14**, 226–229.

Tedesco, E., Marseglia, E. A., A-Mandhary, M. R. A., Al-Suti, M. K., Khan, M. S., David, W. I. F., Shankland, K., Feeder, N., Attfield, J. P. & Raithby, P. R. (2001). In preparation.