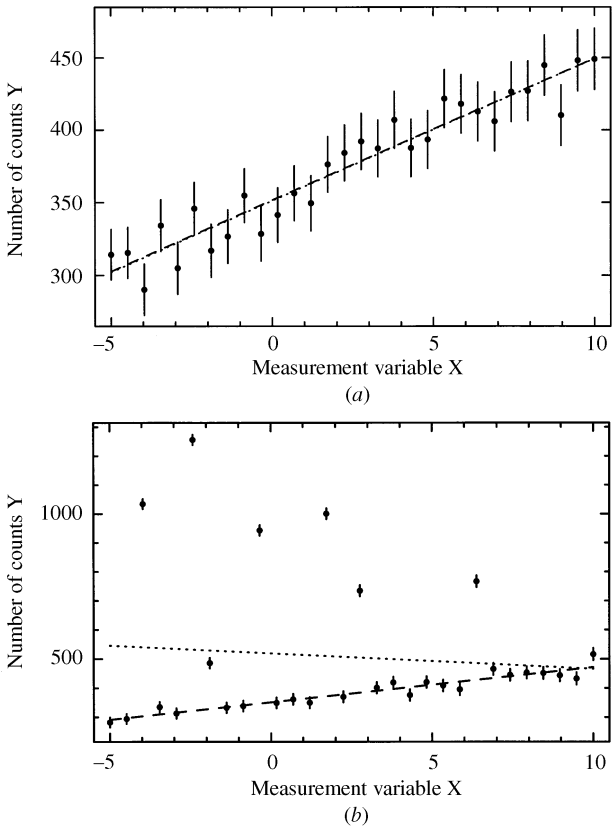

Figure 2
This figure shows a straight line fit to data that should generally follow a straight-line behaviour but occasionally contain a large rogue signal. The dotted line shows the traditional least-squares solution, while the dashed line is the result of a simple Bayesian analysis that allows for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

$$p(D|z \leq 0) \propto \exp(-z^2/2). \tag{6a}$$

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

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

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

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

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

$$\langle (M - D)^2 \rangle = \sigma^2. \tag{8}$$

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:

$$p(B|D, I) = \int_0^\infty p(A, B|D, I)\,\mathrm{d}A. \tag{9}$$

Invoking Bayes' theorem and separating peak and background distribution gives

$$\begin{aligned} p(B|D, I) &\propto \int_0^\infty p(D|A, B, I)\,p(A, B|I)\,\mathrm{d}A \\ &= p(B|I) \int_0^\infty p(D|A, B, I)\,p(A|I)\,\mathrm{d}A. \end{aligned} \tag{10}$$

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:

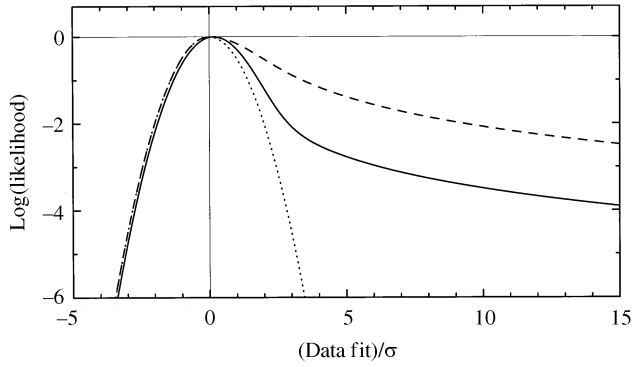

Figure 3
This figure shows three log(likelihood) distributions associated with background estimation. The dotted line is a quadratic function associated with a Gaussian probability distribution function and standard least squares analysis. The dashed line is a combination of least squares (for points 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.