Background estimation in experimental spectra

R. Fischer[a], K. M. Hanson[a,b], V. Dose[a], and W. von der Linden[c]

[a]*Max-Planck-Institut für Plasmaphysik, EURATOM Association*
D-85740 Garching bei München, Germany
e-mail: Rainer.Fischer@ipp.mpg.de
[b]*Los Alamos National Laboratory,*
MS P940, Los Alamos, New Mexico 87545, USA
e-mail: kmh@lanl.gov
[c]*Institut für Theoretische Physik, Technische Universität Graz,*
Petersgasse 16, A-8010 Graz, Austria
e-mail: von_der_Linden@itp.tu-graz.ac.at

A general probabilistic technique for estimating background contributions to measured spectra is presented. A Bayesian model is used to capture the defining characteristics of the problem, namely that the background is smoother than the signal. The signal is allowed to have positive and/or negative components. The background is represented in terms of a cubic spline basis. A variable degree of smoothness of the background is attained by allowing the number of knots and the knot positions to be adaptively chosen on the basis of the data. The fully Bayesian approach taken provides a natural way to handle knot adaptivity and allows uncertainties in the background to be estimated. Our technique is demonstrated on a PIXE spectrum from a geological sample and an Auger spectrum from iron, which contains signals with both positive and negative components.

I. INTRODUCTION

Quantitative spectral analysis often relies on being able to subtract from the data the contribution from the background. In a previous paper, von der Linden et al. [1] presented a general approach to estimating a background contained in spectral data that was based on the assumption that the signal varies much more rapidly than the background. In that work the background was represented by a sequence of cubic splines with equally spaced knots. The minimum knot spacing was determined by the width of the signal structure that one wishes to exclude from the background curve.

This paper extends the earlier work in two important directions; first by employing adaptive splines to represent the background, which is achieved by allowing the number of spline knots to vary in accordance with the requirements of the data, and secondly, by handling bipolar signals, i.e., signals with positive and negative components. We also address several calculational issues, including the improvement in the convergence procedure to determine the spline amplitudes.

We motivate our improvements by referring to a graph of the results from Ref. [1] showing a Particle Induced X-ray Emission (PIXE) spectrum and the estimated background function. The data in Fig. 1 are displayed on a logarithmic scale to exhibit a deficiency in the previous results, already pointed out in Ref. [2]. At the high-energy end of the spectrum, which contains no apparent signal structure, the estimated background has many oscillations. These oscillations do not appear to be supported by the data, given their large uncertainties. Although the wiggles in this tail region of the spectrum do not pose a problem for interpreting this data set, they demonstrate an inherent problem in the previous approach, which could degrade its estimates underneath signal peaks. Our primary goal here is to avoid this spurious behavior in the estimated background. The approach we take is to allow the number of knots and their placement to adapt to the requirements of the data, similar to what was used before for deblurring [3].

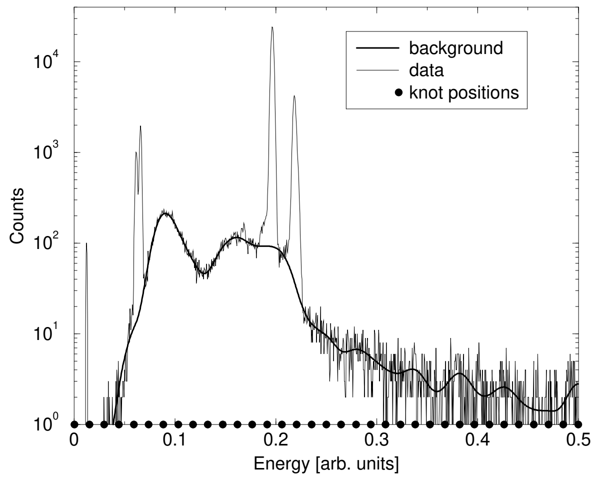

FIG. 1. A PIXE spectrum for a geological sample with the background estimate obtained in Ref. 1 using 35 evenly spaced spline knots. The oscillations in the estimated background above the energy of 0.25 seem unwarranted, given the large uncertainties in the measurements in this region.

Another objective of this paper is to demonstrate that, with a minor modification to the method presented in Ref. [1], it is possible to cope with signals with positive and negative components. We demonstrate this capability on an Auger spectrum.

We refer the reader to the earlier paper [1] for details that we omit here.

II. BAYESIAN APPROACH TO BACKGROUND ESTIMATION

The general idea that we wish to capture with our Bayesian model is that a spectrum consists of a smooth background with additive signal peaks that are relatively compact. We seek a curve $b(x)$, defined over an interval from x_{min} to x_{max}, that describes the background under a spectrum, which is discretely sampled at positions x_i over the same interval. The measured values of the spectrum at these points are designated d_i, collectively referred to as the vector $\boldsymbol{d}$. To cover a wide range of applications, we identify the background by the fact that it is smoother than the signal. More restrictive specifications are certainly possible for a restricted class of problems and can be dealt with in a similar fashion. The smoothness of the background is ensured by expanding it in terms of a set of cubic spline functions

$$b_i = \sum_{\nu=1}^{E} \phi(x_i, \xi_\nu) c_\nu = \sum_{\nu=1}^{E} \Phi_{i,\nu} c_\nu \qquad (1)$$

or in vector notation $\boldsymbol{b} = \boldsymbol{\Phi c}$. The c_ν are the spline values at the E knot positions ξ_ν. The transformation $\phi(x_i, \xi_\nu)$ depends on the vectors $\boldsymbol{\xi}$ and $\boldsymbol{x}$ and, hence, the matrix $\boldsymbol{\Phi}$ depends on these vectors. Without going into detail, we use the results of spline theory [4–6] to determine the elements of $\boldsymbol{\Phi}$. An implicit assumption must be made about the curve at the endpoints. We choose the natural spline condition, that is, assume that the second derivatives of $b(x)$ are zero at the ends of the interval. Other possible boundary conditions are given in [4,5]. Although the basis set that we consider consists of cubic splines, our approach can be easily adopted to other smooth basis functions.

In our Bayesian approach we focus on the probability of the background having a value b_i at each measurement position x_i represented by $p(b_i|\boldsymbol{d}, \mathcal{M}, \mathcal{I})$. This probability depends on the full data set $\boldsymbol{d}$, an as-yet-unspecified model for the background, summarized here simply as $\mathcal{M}$, and all relevant information $\mathcal{I}$ concerning the nature of the physical situation and knowledge of the experiment. We will include in $\mathcal{I}$ knowledge of the noise in the experimental measurements. Also included in $\mathcal{I}$ is the knowledge of the signal structure that we wish to exclude from background, summarized in our spline model by the parameter Δx, the minimum distance between spline knots. Both of these specifications play a crucial role since they provide the information that the model uses to discriminate the signal from the background.

Equation (1) allows us to focus on the $\boldsymbol{c}$ as the fundamental set of parameters to be estimated. According to Bayes law, [7–9] the desired probability for $\boldsymbol{c}$ can be expressed as

$$p(\boldsymbol{c}|\boldsymbol{d}, \boldsymbol{\xi}, E, \mathcal{I}) = \frac{p(\boldsymbol{d}|\boldsymbol{c}, \boldsymbol{\xi}, E, \mathcal{I})\, p(\boldsymbol{c}|\boldsymbol{\xi}, E, \mathcal{I})}{p(\boldsymbol{d}|\boldsymbol{\xi}, E, \mathcal{I})} \ . \qquad (2)$$

The likelihood, $p(\boldsymbol{d}|\boldsymbol{c}, \boldsymbol{\xi}, E, \mathcal{I})$, expresses the probability of the measurements, given their uncertainties. The prior, $p(\boldsymbol{c}|\boldsymbol{\xi}, E, \mathcal{I})$, is a probabilistic statement of what we know about the quantities of interest, $\boldsymbol{c}$ in this case, independent of the experimental data. The denominator, $p(\boldsymbol{d}|\boldsymbol{\xi}, E, \mathcal{I}) = \int d^E c\, p(\boldsymbol{d}|\boldsymbol{c}, \boldsymbol{\xi}, E, \mathcal{I})\, p(\boldsymbol{c}|\boldsymbol{\xi}, E, \mathcal{I})$, called the evidence, guarantees that the posterior has the correct normalization: $\int d^E c \ p(\boldsymbol{c}|\boldsymbol{d}, \boldsymbol{\xi}, E, \mathcal{I}) = 1$. As we shall see, the evidence plays a central role in determining the number of spline knots, E, in our adaptive model.

A. The Prior Probabilities

The distinguishing characteristic of the background that we wish to exploit is its smoothness. In the earlier work [1], the prior on the background used to express its smoothness was based on the integral of the square of the slope of the background. That prior is inconsistent with the cubic splines used to represent the background, which are known to minimize the integral of the square of the second derivative. Therefore, we now use the more appropriate prior

$$p(\boldsymbol{b}|\mu, \mathcal{I}) = \frac{1}{Z} \exp\left\{ -\mu \int dx\, |b''(x)|^2 \right\} \ , \qquad (3)$$

where $b''(x)$ is the second derivative of the background function at x. This prior has the additional advantage over the previous one that it does not penalize linear backgrounds. The factor Z is included for normalization. The positive parameter μ controls the width of this prior distribution.

The expansion in Eq. (1) yields for the prior

$$p(\boldsymbol{c}|\mu, \boldsymbol{\xi}, E, \mathcal{I}) = \pi^{-E/2} \mu^{E/2} (\widetilde{\det} \boldsymbol{D})^{1/2} \exp\left\{ -\mu \boldsymbol{c}^{\mathrm{T}} \boldsymbol{D} \boldsymbol{c} \right\} \ , \qquad (4)$$

where $D_{\nu_1,\nu_2} = \int dx \ \phi''_{\nu_1}(x) \ \phi''_{\nu_2}(x)$. The matrix D can be evaluated analytically or numerically.

The determinant of $\boldsymbol{D}$ provides the volume factor needed for the proper normalization of the Gaussian. The tilde over the determinant symbol indicates the need for a special treatment of the determinant evaluation. Because both constant and linear eigenvectors have zero eigenvalue, $\boldsymbol{D}$ has two zero eigenvalues. Thus the actual determinant of $\boldsymbol{D}$ is zero, which would make Eq. (4) useless. The proper interpretation is achieved through the addition of $-\epsilon\mu\, \boldsymbol{c}^{\mathrm{T}}\boldsymbol{c}$ to the exponent in (4), which adds a very small ϵ to the diagonal elements of $\boldsymbol{D}$. The modified determinant is $\det \boldsymbol{D} = \epsilon^2 \cdot \widetilde{\det}\boldsymbol{D}$, with the understanding that $\widetilde{\det}\boldsymbol{D}$ is the product of the $E-2$ nonzero eigenvalues of $\boldsymbol{D}$. For parameter estimation, ϵ is an unimportant proportionality factor and for model comparison the term drops out. Thus one obtains the same results as if one had started with Eq. (4).

Since μ is a nuisance parameter for our problem, according to the rules of probability, it should be integrated out, that is, $p(\boldsymbol{c}|\bullet) = \int d\mu\, p(\mu\,, \boldsymbol{c}|\bullet) = \int d\mu\, p(\boldsymbol{c}|\mu\,\bullet)\, p(\mu|\bullet)$. The dot indicates any applicable conditionals that do not need to be specified. This parameter can be dealt with straight away. The appropriate prior for a scale parameter, such as μ, is Jeffreys' prior $p(\mu|\mathcal{I}) \propto 1/\mu$, with the usual caveats [1]. The integration yields the multivariate Student's t distribution

$$p(\boldsymbol{c}|\boldsymbol{\xi}, E, \mathcal{I}) = \pi^{-E/2} (\widetilde{\det}\boldsymbol{D})^{1/2} \Gamma(E/2) (\boldsymbol{c}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{c})^{-E/2} \quad . \quad (5)$$

In this paper we allow the positions of spline knots ξ_ν to vary, except for ξ_1 and ξ_E, which are fixed at x_{min} and x_{max}, respectively. The objective is to allow a variable degree of smoothing for the background. Since the ξ_ν are now parameters that are subject to a probabilistic treatment, we need a prior for them. We pick a general noncommittal prior by assuming it is uniform over the phase space available to the ξ_ν [3]. For the interval from $\xi_1 \equiv x_{min}$ to $\xi_E \equiv x_{max}$, taking into account the minimum spacing Δx and the required ordering of the knot positions, that is $\xi_1 + \Delta x \le \xi_2; \xi_2 + \Delta x \le \xi_3; \cdots \xi_{E-1} + \Delta x \le \xi_E$, the prior on $\boldsymbol{\xi}$ is $p(\boldsymbol{\xi}|E, \mathcal{I}) = Z^{-1} \prod_{k=2}^{E} \theta[\xi_{k-1} + \Delta x \le \xi_k]$, where the function θ is unity when its argument conditions are met and zero otherwise. The normalization integral

$$Z = \int_{x_{min}+\Delta x}^{x_{max}-(E-2)\Delta x} d\xi_2 \int_{\xi_2+\Delta x}^{x_{max}-(E-3)\Delta x} d\xi_3 \cdots \int_{\xi_{E-2}+\Delta x}^{x_{max}-\Delta x} d\xi_{E-1} \quad (6)$$

is easily done, resulting in

$$p(\boldsymbol{\xi}|E, \mathcal{I}) = \frac{(E-2)! \prod_{k=2}^{E} \theta[\xi_{k-1} + \Delta x \le \xi_k]}{[x_{max} - x_{min} - (E-1)\Delta x]^{(E-2)}} \quad . \quad (7)$$

The denominator is simply the total volume of the space in which the $(E-2)$ ξ_ν parameters can vary. The factorial in the numerator accounts for the ordering requirement.

The number of spline knots E is also variable. The prior on E is chosen to have a uniform value of $[E_{max} - E_{min} + 1]^{-1}$ for all integer values of E between the minimum number, $E_{min} = 2$, and the maximum number, $E_{max} = \text{integer}[(x_{max} - x_{min})/\Delta x] + 1$, where the output of the integer function is the integral part of its argument. It is zero elsewhere.

B. The Likelihood

The first factor in the numerator of Eq. (2), $p(\boldsymbol{d}|\boldsymbol{c}, \boldsymbol{\xi}, E, \mathcal{I})$, is the likelihood of the experimental data. The data generally consist of the sum of signal and background components, plus a contribution from noise. The innovative idea presented in Ref. [1] is to treat data points containing contributions from the signal as outliers when attempting to fit the background. By incorporating it probabilistically and considering it to be a nuisance variable, the signal is removed from the analysis by integrating over it. This idea grew out of recent Bayesian approaches to the treatment of outlying data in which it was recognized that the presence of a wide nonGaussian tail in the likelihood function effectively reduces the influence of outliers [10–13].

We introduce the proposition B_i: '*datum* d_i *is purely background*' and its complement $\overline{B_i}$: 'd_i *contains some signal contribution*'. The likelihood is the probability distribution corresponding to the measurement uncertainty, given the expected measurement, y_i. When B_i is true, the likelihood for the ith measurement is

$$p(d_i|B_i, y_i, \mathcal{I}) = \begin{cases} (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp\left[-(d_i - y_i)^2/2\sigma_i^2\right] ; & \text{Gaussian} \\ \dfrac{y_i^{d_i}}{d_i!} \exp[-y_i] \,, \quad (y_i \ge 0) ; & \text{Poisson} \,, \end{cases} \quad (8)$$

where the expected value is just the background function at x_i, namely $y_i = b_i$. The parameters E and $\boldsymbol{\xi}$ do not appear here because their dependence is implicitly contained in b_i. We allow for the two most common types of measurement noise corresponding to the uncorrelated Gaussian or Poisson distributions. When the measurement contains a contribution from the signal, the likelihood $p(d_i|s_i, \overline{B_i}, b_i, \mathcal{I})$ is given by the same formula, but with $y_i = b_i + s_i$.

Similar to what was done in Ref. [1], rather than treating the signal as a variable to be estimated, we describe the signal probabilistically in terms of a prior. We provide for the possibility of signals with both positive and negative components by writing the prior as a two-sided exponential function

$$p(s_i|\lambda_+, \lambda_-, \mathcal{I}) = \begin{cases} \lambda_+^{-1} \exp\left[-\frac{s_i}{\lambda_+}\right] \,, & s_i \ge 0 \\ \lambda_-^{-1} \exp\left[+\frac{s_i}{\lambda_-}\right] \,, & s_i < 0 \end{cases} \quad (9)$$

with the restrictions $\lambda_+ > 0$ and $\lambda_- > 0$. In other words, we introduce two different scales for the signal, dependent on its sign. According to the Maximum-Entropy principle the exponential prior is the least informative prior being constrained only to a given scale length $\lambda_{+/-} = < s_{+/-} >$.

The likelihood for the case $\overline{B_i}$ is obtained by marginalizing over the signal

$$p(d_i|\overline{B_i}, b_i, \mathcal{I}) = \int_{-\infty}^{\infty} ds_i \; p(d_i|s_i, \overline{B_i}, b_i, \mathcal{I})\, p(s_i|\lambda_+, \lambda_-, \mathcal{I}) \quad . \tag{10}$$

For the Poisson case, the lower limit must be set to $-b_i$ to respect the nonnegativity constraint of the Poisson likelihood. This integral can be evaluated analytically, yielding for the positive part of the exponential of Eq. (9), i.e., $(\lambda = \lambda_+)$

$$p(d_i|\overline{B_i}, b_i, \lambda, \mathcal{I}) = \begin{cases} \dfrac{1}{2|\lambda|}\left\{1 + \mathrm{erf}\left[\frac{\lambda(d_i-b_i)-\sigma_i^2}{|\lambda|\sqrt{2\sigma_i^2}}\right]\right\} \times \\ \qquad \exp\left[\frac{-\lambda(d_i-b_i)+\sigma_i^2/2}{\lambda^2}\right]; \qquad \text{Gaussian} \\ \dfrac{\exp\left[b_i/\lambda\right]}{|\lambda|(1+\lambda^{-1})^{d_i+1}} \times \\ \qquad \dfrac{\Gamma[(d_i+1), b_i(1+\lambda^{-1})]}{\Gamma(d_i+1)}; \qquad \text{Poisson} \quad , \end{cases} \tag{11}$$

where $\Gamma[a, x] = \int_x^\infty e^{-t}\, t^{a-1}\, dt \quad (a > 0)$ is the incomplete Gamma-function and $\Gamma[a] = \Gamma[a, 0]$ is the Gamma function ($\Gamma[n + 1] = n!$). For the combined positive and negative signals in Eq. (9), the likelihood is the sum of two contributions, one obtained by substituting λ_+ for λ in Eq. (11) and the other by substituting $-\lambda_-$. In the latter substitution for the Poisson case, one must replace $\Gamma[(d_i+1), b_i(1+\lambda^{-1})]$ by $\Gamma(d_i+1) - \Gamma[(d_i+1), b_i(1-\lambda_-^{-1})]$ to account for the finite lower limit of integration.

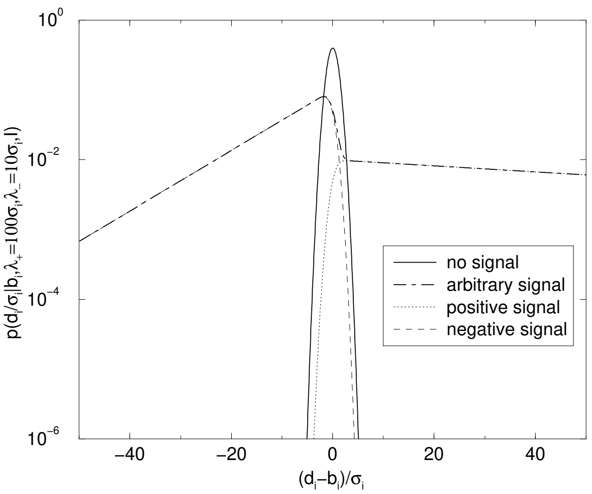

FIG. 2. The likelihood functions for the cases that there is no signal present, and for positive and negative signals of scales $\lambda_- = 10\sigma$ and $\lambda_+ = 100\sigma$. The relative contribution of the later to the mixture model (12) for the likelihood is weighted by $1 - \beta$, and the former by β.

To complete the specification of the likelihood, we employ a mixture model [9], which effectively combines the probability distributions for the two possibilities, B_i and $\overline{B_i}$,

$$p(\boldsymbol{d}|\boldsymbol{c}, \boldsymbol{\xi}, E, \beta, \lambda, \mathcal{I}) = \prod_i [\, \beta\; p(d_i|B_i, \boldsymbol{c}, \boldsymbol{\xi}, E, \beta, \lambda, \mathcal{I}) \; + \\ (1-\beta)\; p(d_i|\overline{B_i}, \boldsymbol{c}, \boldsymbol{\xi}, E, \beta, \lambda, \mathcal{I})] \quad , \tag{12}$$

where β is the probability that a data point contains no signal contribution. We will consider the parameters E, β, λ_+, and λ_- as auxiliary parameters for the adaptive spline problem, whose specifications will be addressed in Sect. II C. The likelihood functions contributing to the mixture model are plotted in Fig. 2. The sum of the two types of likelihood in the mixture model for each datum results in a likelihood function with a central peak plus a long tail. The presence of such a long tail has the effect of reducing the influence of outlying data points when several data points are combined [10–13]. In the case of background estimation, the result is to reduce the influence of points that lie outside the uncertainty band of the measurement errors, which presumably contain significant signal contributions. Without this tail, the resulting curve would be drawn significantly toward the signal structure and not be representative of the background.

C. Determining Auxiliary Parameters

There are numerous parameters, Δx, $\boldsymbol{\sigma}$, E, β and λ's, that have so far been assumed to be fixed. These must be specified to perform the data analysis. It is our view that as many of these parameters as possible should be determined from information about the experiment. Other parameters may have preferred values, based on general arguments, and still others are appropriately determined from the data.

In the present background estimation situation, it is imperative that the minimum knot spacing Δx be determined from knowledge of the experimental situation or by examination of the spectrum. This parameter should be set on the basis of the physicist's experience with the experiment and is certainly no less than the instrumental resolution. Similarly, the experimentalist must choose between Poisson and Gaussian likelihood functions, and, in the latter case, specify the rms deviation of the noise, which may depend on the measured spectral amplitude. The scale of the signal expressed by the λ's should also be set by the physicist on the basis of the expected signal amplitudes. If the signals are expected to be of one sign, that information should obviously be incorporated. It is important to specify all these parameters, because they play a major role in helping the spline model distinguish between background and signal.

The parameter β, which is the probability that a data point contains just background, is one that can be specified by a general argument. Clearly $\beta = 0.5$ is the noncommittal value, stating that each datum is equally likely to contain a signal contribution or not. This choice can also be motivated by an argument given in Ref. [12]. It was shown there that if a separate β_i is associated with each data point, marginalization over the βs results in an integral of the form $\int_0^1 d\beta_1[(1-\beta_1)p(d_1|\overline{B_1}\bullet) + \beta_1 p(d_1|B_1\bullet)] \int_0^1 d\beta_2[(1-\beta_2)p(d_2|\overline{B_2}\bullet) + \beta_2 p(d_2|B_2\bullet)]\cdots$. This integral can be done analytically to obtain $[\frac{1}{2}p(d_1|\overline{B_1}\bullet) + \frac{1}{2}p(d_1|B_1\bullet)][\frac{1}{2}p(d_2|\overline{B_2}\bullet) + \frac{1}{2}p(d_2|B_2\bullet)]\cdots$. The effect is the same as setting all the β_i equal to $\frac{1}{2}$.

The last parameter to deal with is the number of spline knots, E. This parameter obviously cannot be set beforehand, since we want the spline model to adapt to the data. However, E is a nuisance parameter, that is, we don't care what its value is, except to estimate the $\boldsymbol{c}$ and $\boldsymbol{\xi}$ parameters. Probability theory requires that one integrates the joint distribution over nuisance parameters. Beginning with the joint probability distribution in $\boldsymbol{c}, \boldsymbol{\xi}$, and E, we integrate over the first two parameters to obtain

$$\begin{aligned} p(E|\boldsymbol{d},\mathcal{I}) &= \int d^E c\; d^{E-2}\xi\; p(\boldsymbol{c},\boldsymbol{\xi},E|\boldsymbol{d},\mathcal{I}) \\ &\propto \int d^E c\; d^{E-2}\xi\; p(\boldsymbol{d}|\boldsymbol{c},\boldsymbol{\xi},E,\mathcal{I})\, p(\boldsymbol{c},\boldsymbol{\xi},E|\mathcal{I}) \\ &= p(E|\mathcal{I}) \int d^E c\; d^{E-2}\xi\; p(\boldsymbol{d}|\boldsymbol{c},\boldsymbol{\xi},E,\mathcal{I})\, p(\boldsymbol{c},\boldsymbol{\xi}|E,\mathcal{I}) \ , \end{aligned} \tag{13}$$

where we have assumed that the priors on $\boldsymbol{c}$ and $\boldsymbol{\xi}$ are logically independent from that on E. The leading factor is the prior for E, given in Sect. II A. This integral is the same as the denominator of Bayes law for estimating the parameters, given in Eq. (2), which is called the evidence. We define the scalar

$$\psi(\boldsymbol{c},\boldsymbol{\xi}) = -\log[p(\boldsymbol{d}|\boldsymbol{c},\boldsymbol{\xi},E,\mathcal{I})\, p(\boldsymbol{c},\boldsymbol{\xi}|E,\mathcal{I})] \ , \tag{14}$$

which is the minus-logarithm of the integrand in the previous equation.

We approximate ψ by expanding it to second order in $\boldsymbol{c}$ around its maximum value at $\hat{\boldsymbol{c}}$ yielding a Gaussian for its exponential. Because the Gaussian is restricted to a narrow region, the integration can be extended to $-\infty < c < \infty$, so that the integral over $d^E c$ can be evaluated analytically. Equation (13) becomes

$$\begin{aligned} p(E|\boldsymbol{d},\mathcal{I}) \approx& \frac{1}{Z} p(E|\mathcal{I}) \int d^{E-2}\xi\; p(\boldsymbol{d}|\hat{\boldsymbol{c}},\boldsymbol{\xi},E,\mathcal{I}) \times \\ & p(\hat{\boldsymbol{c}},\boldsymbol{\xi}|E,\mathcal{I})\; (2\pi)^{E/2} \det(\boldsymbol{H_c}(\boldsymbol{\xi}))^{-1/2} \ . \end{aligned} \tag{15}$$

The argument of the determinant is the Hessian, $\boldsymbol{H_c}(\boldsymbol{\xi}) = \nabla_{\boldsymbol{c}} \nabla_{\boldsymbol{c}}^{\mathrm{T}} \psi|_{\hat{\boldsymbol{c}}}$, the E by E matrix of second partial derivatives of ψ with respect to $\boldsymbol{c}$, evaluated at its maximum with respect to $\boldsymbol{c}$. Because ψ is a function of both $\boldsymbol{c}$ and $\boldsymbol{\xi}$, $\boldsymbol{H_c}$ is a function of $\boldsymbol{\xi}$. We will use this technique to approximate integrals several more times.

D. Variance in Background

The expectation value of the second moment matrix of $\boldsymbol{b}$ is obtained by integrating over the posterior probability of the parameters $\boldsymbol{c}$ and $\boldsymbol{\xi}$,

$$\begin{aligned} <\boldsymbol{b}\,\boldsymbol{b}^{\mathrm{T}}> &= \int d^E c\; d^{E-2}\xi\; \boldsymbol{\Phi}(\boldsymbol{\xi})\boldsymbol{c}\,\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\xi})\, p(\boldsymbol{c},\boldsymbol{\xi}|\boldsymbol{d}\bullet) \\ &= \int d^E c\; d^{E-2}\xi\; \boldsymbol{\Phi}(\boldsymbol{\xi})\boldsymbol{c}\,\boldsymbol{c}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\xi})\, p(\boldsymbol{c}|\boldsymbol{\xi},\boldsymbol{d}\bullet)\, p(\boldsymbol{\xi}|\boldsymbol{d}\bullet) \\ &= \int d^{E-2}\xi\; \boldsymbol{\Phi}(\boldsymbol{\xi}) \left[\int d^E c\; \boldsymbol{c}\,\boldsymbol{c}^{\mathrm{T}} p(\boldsymbol{c}|\boldsymbol{\xi},\boldsymbol{d}\bullet)\right] \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\xi})\, p(\boldsymbol{\xi}|\boldsymbol{d}\bullet) \\ &\approx \int d^{E-2}\xi\; \boldsymbol{\Phi}(\boldsymbol{\xi}) \left[\boldsymbol{H_c}^{-1}(\boldsymbol{\xi}) + \hat{\boldsymbol{c}}\,\hat{\boldsymbol{c}}^{\mathrm{T}}\right] \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\xi})\, p(\boldsymbol{\xi}|\boldsymbol{d}\bullet) \\ &= \int d^{E-2}\xi\; \left[\boldsymbol{\Phi}(\boldsymbol{\xi})\boldsymbol{H_c}^{-1}(\boldsymbol{\xi})\,\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\xi}) + \hat{\boldsymbol{b}}(\boldsymbol{\xi})\hat{\boldsymbol{b}}^{\mathrm{T}}(\boldsymbol{\xi})\right] p(\boldsymbol{\xi}|\boldsymbol{d}\bullet) \ , \end{aligned}$$

where $\hat{\boldsymbol{c}}$ is estimated as the mean value of $p(\boldsymbol{c}|\boldsymbol{\xi},\boldsymbol{d}\bullet)$ for a fixed $\boldsymbol{\xi}$. The covariance matrix expressing the uncertainties in the estimated background is then

$$\begin{aligned} <\Delta\boldsymbol{b}\,\Delta\boldsymbol{b}^{\mathrm{T}}> &= <\boldsymbol{b}\,\boldsymbol{b}^{\mathrm{T}}> - <\boldsymbol{b}><\boldsymbol{b}^{\mathrm{T}}> \\ &= \int d^{E-2}\xi\; [\boldsymbol{\Phi}(\boldsymbol{\xi})\boldsymbol{H_c}^{-1}(\boldsymbol{\xi})\,\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\xi}) + \\ &\qquad \Delta\hat{\boldsymbol{b}}(\boldsymbol{\xi})\Delta\hat{\boldsymbol{b}}^{\mathrm{T}}(\boldsymbol{\xi})] p(\boldsymbol{\xi}|\boldsymbol{d}\bullet) \ , \end{aligned} \tag{16}$$

where $\Delta\boldsymbol{b} = \boldsymbol{b} - <\boldsymbol{b}>$ and $\Delta\hat{\boldsymbol{b}} = \hat{\boldsymbol{b}}(\boldsymbol{\xi}) - <\boldsymbol{b}>$. We have again introduced a Gaussian approximation for the integrand to do part of the integral analytically. The first term within the square brackets stems from the covariances of $\boldsymbol{c}$ around $\hat{\boldsymbol{c}}$ given by $\boldsymbol{H_c}$, the Hessian of ψ with respect to $\boldsymbol{c}$. The second term describes the covariance of the $\hat{\boldsymbol{b}}(\boldsymbol{\xi})$ due to the variation of $\boldsymbol{\xi}$. Since the $\boldsymbol{c}$ integration is treated analytically, only the $\boldsymbol{\xi}$ integration needs to be done numerically, for example, by MCMC sampling [14] from $p(\boldsymbol{\xi}|\boldsymbol{d})$, as explained in Sect. III C.

III. CALCULATIONAL PROCEDURE

We describe in this section the separate steps in a complete calculation for any particular data set. In the innermost loop, we need to be able to find the spline values that maximize the posterior (2), namely $\hat{\boldsymbol{c}}$. The next higher level involves finding the best knot locations for a fixed E and the highest level loop is over E to marginalize over E.

A. Estimation of Spline Values

The most basic calculation is to find the spline values $\boldsymbol{c}$ that maximize the posterior (2), assuming particular values for the knot positions $\boldsymbol{\xi}$ and the auxiliary parameters $(E, \beta, \lambda_+, \lambda_-)$. The denominator in (2) can be ignored at this point because it does not depend on $\boldsymbol{c}$. What is actually done is to minimize ψ, defined in Eq. (14), with respect to the knot values $\boldsymbol{c}$, which is a nonlinear optimization problem. To evaluate ψ, we use the likelihood given in Eq. (12), inserting the appropriate expression in Eq. (11) and the prior is given in Eq. (5). Both the gradient (first derivative) and curvature matrix (second derivative) of ψ are evaluated analytically. A gradient-based quasi-Newton optimization algorithm is employed to minimize ψ. The optimization algorithm we use can impose a nonnegativity constraint of the background curve. We find that the optimization occasionally stalls and the appropriate global minimum in $\boldsymbol{c}$ is not reached because of the existence of local minima.

We have developed a new technique to enhance the convergence behavior of the optimization algorithm. Our technique is based on artificially broadening the background only part of the likelihood function during the early part of the optimization process, which effectively eliminates local minima by forcing all data points to be belong to the background. This broadening is easily accomplished for the Gaussian likelihood by increasing the value of the $\boldsymbol{\sigma}$ in the likelihood for the background term in Eq. (11). We do not find it necessary to resort to this technique for our Poisson examples, the PIXE data. However, a similar scheme might be used for the Poisson case, e.g., by dividing the expected number of counts y_i and the measured counts d_i in the likelihood Eq. (8) by a common factor. The effect of our approach is to increase the reach of the function being minimized, which is quadratic in the case of the Gaussian likelihood, and promote larger steps in the Newton-type optimization algorithm. In a little more detail, we begin the optimization by multiplying $\boldsymbol{\sigma}$ by a common factor, which is chosen to make the rms value of $\boldsymbol{\sigma}$ the same as λ. After convergence, $\boldsymbol{\sigma}$ is divided by two and the optimization is resumed from the last operating point. This process is repeated until the nominal values for $\boldsymbol{\sigma}$ are reached. We find that this procedure, which resembles a multiscale approach used to solve geometrical optimization problems [15], yields very robust and speedy convergence to the global minimum.

B. Estimation of Knot Positions

The knot positions $\hat{\boldsymbol{\xi}}$ are to be found by minimizing ψ, given in Eq. (14). This optimization problem is somewhat harder than the one associated with finding $\hat{\mathbf{c}}$. The reason lies in the numerous constraints on the knot positions, namely that they must be ordered and they must be no closer to each other than a specified Δx. Furthermore, there are many local minima in ψ. Therefore, we use another optimization strategy, that of simulated annealing [16], to find the most probable knot positions. Throughout this process, the number of knots E is held fixed.

The simulated annealing technique is based on a Markov Chain Monte Carlo algorithm (MCMC) [14], described in more detail in the next section. The widths of the Cauchy distribution for calculating the Markov steps are fixed throughout the cooling process. The probability distribution is flattened by dividing ψ by T, a fictitious temperature. The initial temperature is $T = 500$. When a step is accepted, T is decreased by multiplying it by 0.95 if the new value of ψ is smaller than any previous value, or by 0.995 if it is not. At the end of the full annealing sequence, the estimated knot position vector $\hat{\boldsymbol{\xi}}$ is the one that had the smallest value for ψ.

C. Marginalization over Number of Knots

In probability theory, as explained in Sect. II C, it is proper to marginalize over nuisance parameters that we don't care about knowing, such as E. The probability of E is given in Eq. (15). Again the integrand is approximated as a Gaussian in $\boldsymbol{\xi}$

$$p(E|\boldsymbol{d},\mathcal{I}) \propto p(E|\mathcal{I})\, p(\boldsymbol{d}|\hat{\mathbf{c}},\hat{\boldsymbol{\xi}},E,\mathcal{I}) \times \\ p(\hat{\mathbf{c}}|E,\mathcal{I})\,(2\pi)^{E/2}\det(\boldsymbol{H}_{\boldsymbol{c}})^{-1/2} \times \\ \int d^{E-2}\xi \;\; p(\boldsymbol{\xi}|E,\mathcal{I})\;\exp[-\frac{1}{2}(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}})^{\mathrm{T}}\boldsymbol{H}_{\boldsymbol{\xi}}(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}})] \;, \tag{17}$$

where $\boldsymbol{H}_{\boldsymbol{\xi}}$ is the $(E-2)$ by $(E-2)$ Hessian matrix for ψ with respect to the variable $\boldsymbol{\xi}$, calculated at the optimal knot positions $\hat{\boldsymbol{\xi}}$. The prior probability in (15) $p(\boldsymbol{c},\boldsymbol{\xi}|E,\mathcal{I})$ has been replaced with the product of the prior on $\boldsymbol{c}$ and the prior on $\boldsymbol{\xi}$, which is valid because these are logically independent priors. The integration here is complicated by the ordering restrictions placed on the ξ_is by the prior on $\boldsymbol{\xi}$ given in Eq. (7). Thus, the integration is over a restricted volume V defined by the ordering requirement. The integral can not be evaluated analytically because it is impossible to simply extend the integration limits to infinity. Therefore, $\boldsymbol{H}_{\boldsymbol{\xi}}$ is replaced by an effective Hessian $\boldsymbol{H}^*_{\boldsymbol{\xi}}$, which must reflect the complicated integration volume V,

$$p(E|\boldsymbol{d},\mathcal{I}) \approx p(E|\mathcal{I})\, p(\boldsymbol{d}|\hat{\mathbf{c}},\hat{\boldsymbol{\xi}},E,\mathcal{I}) \times \\ p(\hat{\mathbf{c}}|E,\mathcal{I})\,(2\pi)^{E/2}\det(\boldsymbol{H}_{\boldsymbol{c}})^{-1/2} \times \\ p(\hat{\boldsymbol{\xi}}|E,\mathcal{I})\,(2\pi)^{(E-2)/2}\det(\boldsymbol{H}^*_{\boldsymbol{\xi}})^{-1/2} \;. \tag{18}$$

The effective Hessian $\boldsymbol{H}^*_{\boldsymbol{\xi}}$ is actually estimated using MCMC to draw knot positions from the probability distribution in the integral in Eq. (15), i.e., $p(\boldsymbol{d}|\hat{\mathbf{c}},\boldsymbol{\xi},E,\mathcal{I})\, p(\boldsymbol{\xi}|E,\mathcal{I}) \;\; (2\pi)^{E/2}\det(\boldsymbol{H}_{\boldsymbol{c}}(\boldsymbol{\xi}))^{-1/2}$. The covariance matrix $(\boldsymbol{H}^*_{\boldsymbol{\xi}})^{-1}$ is estimated as the matrix of second moments of the resulting set of MCMC samples of $\boldsymbol{\xi}$.

The aim of an MCMC algorithm [14] is to generate a sequence of parameters $\boldsymbol{y}_k, (k = 1, 2, \cdots, K)$ that represent random draws from a specified probability density distribution, let's say $\pi(\boldsymbol{y})$. To add a new member to the sequence $\boldsymbol{y}_{k+1}$, the Metropolis algorithm consists of trying a proposed step away from the present $\boldsymbol{y}_k$. The proposed step $\Delta\boldsymbol{y}$ is drawn randomly from a symmetric distribution, and is either accepted or rejected on the basis of the value of π at the new position compared to the old position. If the step is rejected, $\boldsymbol{y}_{k+1}$ is set equal to $\boldsymbol{y}_k$. For the step distribution we use a Cauchy distribution, i.e., $\propto [1+(|\Delta\boldsymbol{y}|/W)^2]^{-1}$, where W is the full-width at half-maximum (FWHM) of the distribution [17]. With its wide tails, the Cauchy distribution occasionally proposes large steps, which can be useful for getting out of local minima. In our algorithm, only one knot position is moved at a time. When a knot is moved to within Δx of another knot, the move is rejected. When a knot is moved past other fixed knots, the knots are renumbered to maintain the required knot ordering.

The FWHM of the Cauchy distribution is started at a value of about one tenth the interval width $(x_{max} - x_{min})/(E-1)$ and the width for each knot position is adaptively adjusted during a training run to obtain an approximate 50% acceptance rate for proposed steps. For the PIXE spectrum in Fig. 4, the final FWHM values ranges from 10^{-4} to 0.02. For the MCMC runs to draw samples from the probability distribution of $\boldsymbol{\xi}$ cited above, on the order of 10^5 cycles through the full parameter set are taken. We check the performance of our MCMC procedure by calculating the autocorrelation function for each knot position [14]. The estimated correlation lengths range from 10 to 1000 MCMC iterations. The pivotal knot position is chosen randomly. From this, the number of effectively independent samples from the probability density function for a run of 10^5 iterations is from 50 to 5000. The simulated annealing procedure used to find the most likely knot positions described in the preceding section proceeds similarly, but with the introduction of the artificial temperature.

As we shall see in our results, there are competing factors in Eq. (18). The likelihood factor $p(\boldsymbol{d}|\hat{\boldsymbol{c}}, \hat{\boldsymbol{\xi}}, E, \mathcal{I})$ should always increase with increasing E because the data must always be matched better by the spline model with more knots when the knots are allowed to move. The Ockham factors for $\hat{\boldsymbol{c}}$, $p(\hat{\boldsymbol{c}}|\boldsymbol{\xi}, E, \mathcal{I})(2\pi)^{E/2}\det(\boldsymbol{H}_{\boldsymbol{c}})^{-1/2}$ (Eq. (5)) and for $\boldsymbol{\xi}$, $p(\hat{\boldsymbol{\xi}}|E, \mathcal{I})(2\pi)^{(E-2)/2}\det(\boldsymbol{H}_{\boldsymbol{\xi}})^{-1/2}$ typically decrease as E increases. This competition between likelihood and the priors is the action of Ockham's razor [18–20], named after William of Ockham, whose principle states that models should be no more complex than necessary to explain the available data. The overall effect is that there will be a maximum in the probability of E beyond which the addition of more knots does not help represent the background significantly better.

D. Estimation of Uncertainties in Background

The uncertainty bound on the estimated background function may be calculated as described in Sect. II D. Equation (16) shows how the covariance in the estimates for $\boldsymbol{b}$ is obtained by splitting the covariance into two terms, one arising from the uncertainties in $\boldsymbol{c}$ for fixed $\boldsymbol{\xi}$, and the other from uncertainties in $\boldsymbol{\xi}$. The contribution from the first term is based on the analytic expression for the Hessian $\boldsymbol{H}_{\boldsymbol{c}}$, which can be evaluated for any $\boldsymbol{\xi}$. The rest of the calculation involves randomly drawing samples from $p(\boldsymbol{\xi}|\boldsymbol{d})$ using the Markov Chain Monte Carlo (MCMC) technique described above. For each $\boldsymbol{\xi}$ drawn, the optimum $\hat{\boldsymbol{c}}$ has to be found using the minimization procedure described above. Then, the spline values at the data points are obtained: $\hat{\boldsymbol{b}} = \boldsymbol{\Phi}\hat{\boldsymbol{c}}$. The integration in Eq. (16) is accomplished by averaging the quantity within the square brackets in the integrand over the $\boldsymbol{\xi}$ samples.

IV. RESULTS

We now describe the results of applying the analysis outlined in the preceding section to the PIXE data shown in Fig. 1. For this analysis the underlying auxiliary parameters, described in Sect. II C, are the same as used in the previous analysis shown in Fig. 1. The minimum distance between knots is $\Delta x = 0.015$, the approximate width of the conspicuous signal peaks at their base. Because we know that the signal peaks in the PIXE spectrum must be positive, we exclude the contribution of negative signals to the likelihood, in effect setting $\lambda_- = 0$. The scale λ_+ should be derived from the signal [1]. As the signal is much larger than the background, we set λ_+ equal to the average value of the data set, about 270 in this case. Figure 3 shows the probability distribution for E given in Eq. (18). Note the extremely large dynamic range of this plot. The likelihood, $p(\boldsymbol{d}|\hat{\boldsymbol{c}}, \hat{\boldsymbol{\xi}}, E, \mathcal{I})$ increases monotonically with E since the fit to the data always improves with more knots. The Ockham factor for $\hat{\boldsymbol{c}}$, $p(\hat{\boldsymbol{c}}|\boldsymbol{\xi}, E, \mathcal{I})(2\pi)^{E/2}\det(\boldsymbol{H}_{\boldsymbol{c}})^{-1/2}$ (Eq. (5)) decreases gradually over the range of E shown. The corresponding factor for $\boldsymbol{\xi}$, $p(\hat{\boldsymbol{\xi}}|E, \mathcal{I})(2\pi)^{(E-2)/2}\det(\boldsymbol{H}_{\boldsymbol{\xi}})^{-1/2}$ decreases substantially. The net result is a strong peak in the probability at $E = 14$, which contains a probability of 80%. Since most of the probability falls into the single $E = 14$ bin, we may legitimately fix E at 14, instead of marginalizing over E, to obtain the final background estimates.

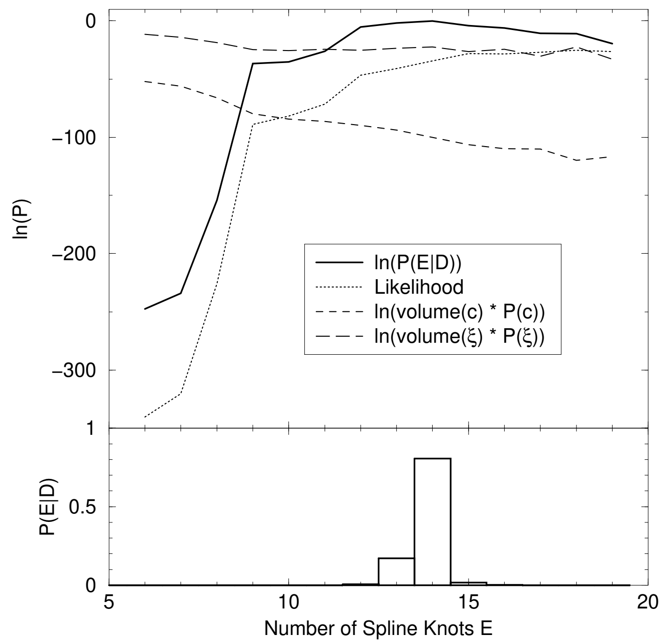

FIG. 3. The probability for the parameter E (the number of spline knots) given by Eq. (15), shown as the solid curve, with its various contributions. The maximum probability occurs at $E = 14$ knots.

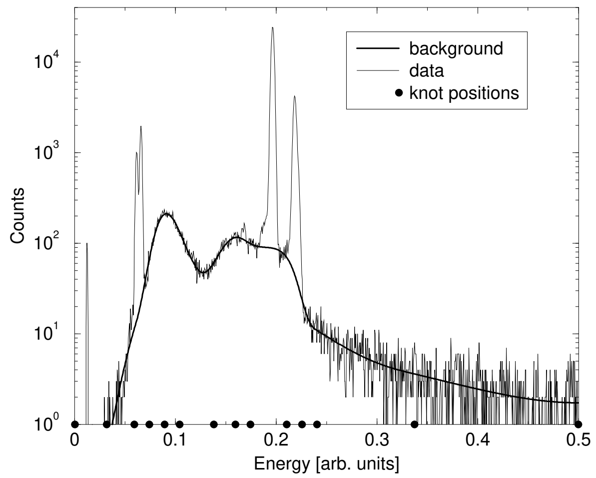

FIG. 4. The same PIXE spectrum as in Fig. 1, showing the most probable background estimate obtained using adaptive splines in which the optimal number of knots is found to be 14. In the energy region above 0.25, the estimated background is now smooth, indicating a lack of evidence in the data for the oscillations visible in Fig. 1.

The background estimate with the highest posterior probability obtained in the simulated annealing search for the most probable knot position is shown in Fig 4. The high-energy portion of the spectrum is now fit with a smooth background, consistent with a physicist's expectation. It is remarkable that our model requires only one additional spline knot to fit the energy region above 0.25. It is also interesting to note that the background under the first significant peak at an energy of approximately 0.06 is smoother and more plausible than for the previous analysis. The placement of the knots is of interest. The highest knot density occurs in the vicinities of the three major peaks in the background. While these seem like fairly smooth sections of the background on this semi-log plot, the curve varies somewhat more rapidly in the linear space in which it is modeled. These adaptive background estimates are very plausible.

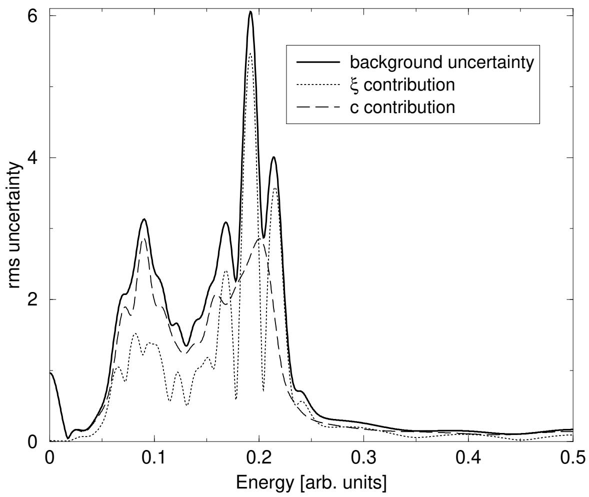

FIG. 5. The uncertainties in the background function displayed in Fig. 4. The separate contributions to the rms deviation of the background values are shown; from the uncertainties in the $\boldsymbol{c}$ and the variance arising from the knot positions $\boldsymbol{\xi}$.

The rms uncertainties in the estimated background curve are summarized in Fig. 5 as uncertainty bounds. These are derived from Eq. (16) by combining the variances from uncertainties in $\boldsymbol{c}$ using the analytic part for fixed knots plus uncertainties arising from the knot positions $\boldsymbol{\xi}$, obtained by numerical integration over the possible knot positions. First of all, we see that the uncertainties are quite small compared with the background itself, on the order of a few percent in the peak regions and about an order of magnitude smaller in the high-energy end of the spectrum. The uncertainties due to those in $\boldsymbol{c}$ dominate at the first significant peak and in the high-energy tail. However, the uncertainties arising from knot placement are most important around the two signal peaks in the spectrum around an energy of 0.2. Clearly, no simple formula based on a single contribution to the total uncertainty applies.

The uncertainty bands shown in Fig. 5 actually correspond to the square root of the diagonal terms of the covariance of $\boldsymbol{b}$ given in Eq. (16). These are useful for showing the limits of uncertainty of the curve, but are not applicable for estimating the consequences of these uncertainties in the background on further computation, e.g., on the areas under a signal peak. For that, the full covariance matrix is required because one expects a significant degree of correlation in the uncertainties from one position to another. For example, when two points lie near each other in the same spline interval, there is a strong positive correlation in their uncertainties because their estimates both rely on the same cubic spline curve. It is feasible to calculate the full covariance matrix using Eq. (16), but not so easy to display it.

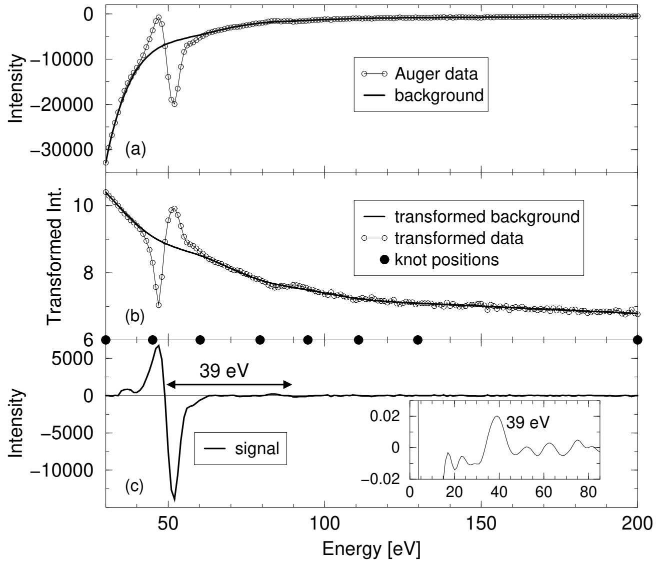

FIG. 6. (a) An MVV Auger spectrum for iron. The estimated background shown is that obtained for the transformed spectrum shown in (b). (b) A logarithmic transformation of the Auger spectrum shown in (a) reduces the curvature of the background, rendering it suitable for the general approach presented here. The estimated background is shown. (c) The signal determined by subtracting the estimated background from the original spectrum. The inset in (c) shows the autocorrelation of the signal vs. energy difference. A significant secondary peak is seen at an energy offset of 39 eV.

To demonstrate how well our background method works for signals with both positive and negative contributions, we turn to the Auger spectrum shown in Fig. 6a. This spectrum was obtained for an iron sample using a four-grid low-energy electron diffraction (LEED) optics, operated in the retarding-field mode. Harmonic modulation of the retarding potential and lock-in detection of the transmitted current on the second harmonic of the modulation frequency results in spectra as shown in Fig. 6a. Such spectra constitute the energy derivative of the sum of the Auger electron energy distribution, the signal, and the slowly varying, much larger secondary electron energy distribution, the background. The signal contains both positive and negative components. For quantitative Auger analysis it is mandatory to separate the two contributions to the total signal [21,22]. The principal signal seen at around 50 eV comes from an $M_{2,3}VV$ Auger transition.

It is evident from Fig. 6 that, while the background may be smooth, it varies quite rapidly at low energies. This behavior is inconsistent with our general background model, whose prior is based on the second derivative of the background. However, a simple transformation of the measured spectrum brings the background into conformance with our background model and does not dilute the signal characteristics unduly. By taking the logarithm of the measured spectrum, the nearly exponential rise of the spectrum is transformed into an approximately linear dependence that is more easily accommodated by the background model. Furthermore, such a transformation of the ordinate does not change the width of the signal structure, leaving unchanged the minimum knot separation criterion. As a general principle for applying our model to a specific spectrum, it may be transformed to bring the background into conformance with the background model, provided the signal contributions do not loose their assumed rapid and localized characteristics. For example, we find that taking the square root of the horizontal scale, after a suitable offset, yields a data record that also provided reasonable estimates of the background.

Figure 6b shows the Auger spectrum after the transformation $z(k) = \log[a - y(k)]$, where $y(k)$ is the original spectral amplitude and a is a constant ($= 340$ in this case). The uncertainties in the transformed spectrum are obtained by dividing the uncertainties in the original spectrum, σ_i, by $a - y(k)$. σ_i is estimated to be approximately 35 over the entire spectrum. The transformed spectrum is analyzed using the background models described earlier. The minimum knot separation is set at $\Delta x = 15$ eV. In this analysis, λ_+ and λ_- are assumed to be equal because the positive and negative signals are expected to have approximately the same amplitudes. They are set to a typical value of about 0.1. The evidence evaluation of Eq. (18) shows that $p(E|\boldsymbol{d}, \mathcal{I})$ is rather flat for the number of nodes between $E = 8$ and $E = E_{max} = 12$. The lack of a strong peak in the evidence, as seen in the earlier PIXE analysis, may be explained as follows. The prior on $\boldsymbol{\xi}$, given in Eq. (7), increases considerably as E approaches E_{max} because of the decreasing available volume for knots. This effect is partly counteracted by the decreasing volume given by $\boldsymbol{H}_{\boldsymbol{\xi}}$, but not completely. Thus, the Ockham factor pertaining to $\boldsymbol{\xi}$ may effectively increase with increasing E, a behavior that is unexpected, but plausible. It is not the number of parameters that define the penalizing Ockham factor but the phase space of the prior covered by the high-likelihood region, which may increase when the parameters are highly correlated. As the likelihood probability increases insignificantly for $E \geq 8$, we show the background estimated for $E = 8$. The results for $E > 8$ lie within the line thickness of the results for $E = 8$. Thus marginalization over E would yield quite the same result. The estimated background is shown in Fig. 6b, and is transformed back into Fig. 6a for comparison with the original spectrum.

After plotting the difference between the original spectrum and its estimated background shown in Fig. 6c, a possible secondary peak is observed. This small peak is demonstrated in the autocorrelation of this background subtracted spectrum, shown as an inset in Fig. 6c. A secondary peak with an amplitude of about 2% of the main peak is convincingly shown at an energy offset of 39 eV, which corresponds to an M_1VV Auger transition for iron. In this case, a proper background subtraction even helps one identify the presence of less apparent signals in an Auger spectrum.

V. SUMMARY

We have developed a probabilistic model to separate the background from signals in spectra. The general assumptions are that the background varies smoothly and that each rapidly varying signal peak is confined to a well-defined interval. The background is represented by a cubic spline basis. In order to allow the smoothness of the background to accommodate the data, we have allowed the number of spline knots and their position to vary. Our Bayesian approach provides a straightforward way to deal with this adaptivity by marginalizing over the probability of the number of knots. The effect of Ockham's factor is to produce a maximum in this probability. We have further extended the earlier work by incorporating signals with either positive or negative components, or both. The uncertainties in the estimated background have also been shown.

ACKNOWLEDGMENTS

We thank J. Padayachee and V. Prozesky of the Van de Graff Group at the National Accelerator Centre, Faure, South Africa, for supplying the PIXE data used in our first example. The Auger spectrum was provided by H. Kang of the Max-Planck-Institut für Plasmaphysik in Garching, Germany. One of the authors (KMH) gratefully acknowledges support from the Max-Planck-Institut für Plasmaphysik (EURATOM Association) and the U.S. Dept. of Energy (Contract W-7405-ENG-36).

[1] W. von der Linden, V. Dose, J. Padayachee, and V. Prozesky, Phys. Rev. E **59**, 6527 (1999).
[2] J. Padayachee *et al.*, Nucl. Inst. Meth. B **150**, 129 (1999).
[3] W. von der Linden, V. Dose, and R. Fischer, in *MAXENT96 - Proceedings of the Maximum Entropy Conference 1996*, edited by M. Sears, V. Nedeljkovic, N. E. Pendock, and S. S. Sibisi (NMB Printers, Port Elizabeth, South Africa, 1997), pp. 154–163.
[4] J. Stoer, *Einführung in die Numerische Mathematik I* (Springer Verlag, Berlin, 1979).
[5] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis* (Springer Verlag, New York, 1980).
[6] C. de Boor, *A Practical Guide to Splines* (Springer Verlag, New York, 1978).
[7] R. T. Bayes, Philos. Trans. R. Soc. London **53**, 370 (1763), reprinted in S. J. Press, *Bayesian Statistics: Principles, Models, and Applications*, John Wiley & Sons, New York, 1989.
[8] D. S. Sivia, *Data Analysis - A Bayesian Tutorial* (Clarendon Press, Oxford, 1996).
[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis* (Chapman & Hall, London, 1995).
[10] D. S. Sivia, in *MAXENT96 - Proceedings of the Maximum Entropy Conference 1996*, edited by M. Sears, V. Nedeljkovic, N. E. Pendock, and S. Sibisi (NMB Printers, Port Elizabeth, South Africa, 1997), pp. 131–137.
[11] W. H. Press, in *Unsolved Problems in Astrophysics*, edited by J. N. Bahcall and J. P. Ostriker (Princeton Univ. Press, Princeton, 1997), pp. 49–60.
[12] V. Dose and W. von der Linden, in *Maximum Entropy and Bayesian Methods*, edited by V. Dose, W. von der Linden, R. Fischer, and R. Preuss (Kluwer Academic Publishers, Dordrecht, 1999).
[13] K. M. Hanson and D. R. Wolf, in *Maximum Entropy and Bayesian Methods*, edited by G. Heidbreder (Kluwer Academic Publishers, Dordrecht, 1996), pp. 255–263.
[14] *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Chapman & Hall, London, 1996).
[15] G. S. Cunningham, I. Koyfman, and K. M. Hanson, in *Medical Imaging: Image Processing*, M. H. Loew and K. M. Hanson, eds., Proc. SPIE **2710**, 145 (1996).
[16] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes*, 2nd ed. (Cambridge University Press, Cambridge, 1992).
[17] H. Szu and R. Hartley, Phys. Lett. A **122**, 157 (1987).
[18] A. J. M. Garrett, Physics World, May, 39 (1991).
[19] A. J. M. Garrett, in *Maximum Entropy and Bayesian Methods*, edited by W. T. Grandy, Jr. and L. H. Schick (Kluwer Academic Publishers, Dordrecht, 1991), p. 357.
[20] D. J. C. MacKay, Neural Computation **4**, 415 (1992).
[21] R. Weissmann and K. Müller, Surface Science Reports **105**, 251 (1981).
[22] P. Staib and J. Kirschner, Appl. Phys. **3**, 421 (1974).