Signal and background separation

W. von der Linden, V. Dose

Max-Planck-Institut für Plasmaphysik, EURATOM Association

D-85740 Garching b. München, Germany

e-mail: wvl@ipp.mpg.de

J. Padayachee, V. Prozesky

Van de Graaff Group National Accelerator Centre

P.O.Box 72, FAURE 7131, South Africa

e-mail: padayachee@nac.ac.za

(26.8.98)

Abstract

In many areas of research the measured spectra consist of a collection of 'peaks' – the sought-for signals – which sit on top of an unknown background. The subtraction of the background in a spectrum has been the subject of many investigations and different techniques, varying from filtering to fitting polynomial functions, have been developed. These techniques yield results which are not always satisfactory and often even misleading. Based upon the rules of probability theory, we derive a formalism to separate the background- from the signal-part of a spectrum in a rigorous and self-consistent manner. We compare the results of the probabilistic approach to those obtained by two commonly used methods in an analysis of Particle Induced X-ray Emission (PIXE) spectra.

02.50.-r,02.50.Rj,07.60.-j,29.30.Kv

I. INTRODUCTION

The analysis of spectra is generally hampered by the presence of noise and an unknown background. This is especially true for particle induced X-ray emission (PIXE) spectra where projectile and secondary electron bremsstrahlung leads to a significant background. We will therefore use PIXE spectra as real-world test examples.

The central problem to be tackled in this paper is best explained using the PIXE spectrum depicted in Fig.1. It consists of a fairly smooth background plus signal with both very large as well as small peaks which are comparable in height to the background. The goal is to determine the background part of the spectrum and by eliminating it from the data-points to infer the desired signal. The problem falls into the realm of inductive logic which tells us how to deal with partial truth: We have experimental data, a vague theoretical conception and additional prior knowledge. The information is, however, not stringent enough for a unique result. (Bayesian) probability theory (BPT) [8,9], as E.T. Jaynes [10] appropriately called it 'THE LOGIC OF SCIENCE', provides the consistent frame to exploit all bits of information rigorously. A nice tutorial type of introduction to BPT is given by D.S. Sivia in his book '*Data Analysis - A Bayesian Tutorial*' [11].

Before discussing the details of the probabilistic approach, we want to mention that several techniques are available for determining the background in a spectrum. The traditional method is to estimate the background by fitting a set of 'semi-empirical' polynomial functions [1,2] to the data while limiting the expansion order to keep the fit from being influenced by sharp peaks in the spectrum. These techniques have been optimized [3] by the use of orthonormal basis sets for the fitting functions, with initialization routines that improve initial guesses of the functions and their coefficients for the final fitting during the full non-linear least squares fitting process. Modern methods tend towards the elimination of the background by digital filtering [4,5] or stripping [6], avoiding any assumption regarding the functional form of the background.

In basic polynomial fitting routines the biggest problems were the unstable nature of the fits, mainly due to the addition of more degrees of freedom to the non-linear fitting program. These problems were alleviated by using polynomials with some physical basis, as well as by empirical improvements to the functions. Non-linear fitting programs simultaneously determine the signal peaks and the coefficients of the background polynomials based on iterative least squares methods. The expansion order of these polynomials were typically less than 10, increasing the number of degrees of freedom by that number. This often led to unstable fits, with a resulting increase in analysis time with human input to correct the procedure.

Digital filtering is based on the convolution of the spectrum with a top-hat filter, as in the GUPIX program [4], or a frequency differentiated non-linear digital filter ("rolling ball"), as in the program used at the Schonland Research Centre (SRC) in Johannesburg [5]. The top-hat filter has a central upper lobe consisting of a number of positive coefficients, and two negative outer lobes each having a number of negative coefficients. The convolution replaces the data content of every channel by this convolving filter. The width of the filter is based on the energy resolution of the measurement, and therefore is effective in areas where the background changes linearly over any region of around 400 eV, which is generally a good approximation.

In the "rolling ball" method [5] the differentiated non-linear digital filter is the equivalent of rolling a ball below the data points and marking out the background as the locus of the ball at each point. To account for the varying peak width of the energy spectra, the diameter of the ball gradually increases as a function of energy. The mathematical functions involved are relatively simple arithmetic functions that can be efficiently computed. There is some distortion of the peaks, but this does not adversely affect the results.

The stripping of the peaks in GeoPIXE [6] is based on estimating the background by iterative suppression of the counts in channels containing peak information. These techniques are less efficient in areas where the background shape changes relatively quickly, leading to over-filtering or stripping in that case. The advantage of filtering and stripping is the elimination of extra free parameters in the non-linear fit routine in the analysis program, and the relative robustness of these techniques.

II. THE BAYESIAN APPROACH

All techniques discussed so far are ad-hoc and yield results of unpredictable reliability. A first consistent probabilistic approach had been suggested by us [7] to separate the signal from the noise and the background and at the same time deconvolve the apparatus function within quantified maximum entropy (QME). The main goal there was to derive a formalism which still allows the use of standard (commercial) QME packages with merely modified input. In return we had to put up with a couple of approximations. Here we will present the full probabilistic approach leaving the framework of QME. The basic idea is best explained guided by the PIXE spectrum mentioned before in Fig.1. Intuitively, one has a fairly clear conception on the essence of the background. What our intuition does is the following: Find the regions of the spectrum that have no signal contribution and fit a smooth function to the so-determined background data points. How can this idea by quantified? The answer is provided by BPT [8,9,11]. Here, we are seeking the probability for the background having a value b_i at point x_i, i.e. $p(b_i|\boldsymbol{d}, \boldsymbol{\sigma}, \boldsymbol{\Pi}, \mathcal{I})$ in the light of all data points d_i (summaries in $\boldsymbol{d}$), the respective experimental noise $\boldsymbol{\sigma}$, a yet unspecified set of parameters $\boldsymbol{\Pi}$ and all background information $\mathcal{I}$ that uniquely defines the problem. The latter plays a crucial role since it provides the information which we need to discriminate the signal from the background. It has to do with correlations either in the signal or the background and is strongly problem dependent. To cover a wide range of applications, we identify the background by the fact that it is smoother then 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 an appropriate set of basis functions $\phi_\nu(x)$

$$b_i = \sum_{\nu=1}^{E} \phi_\nu(x_i) c_\nu = \sum_{\nu=1}^{E} \Phi_{i,\nu} c_\nu \tag{1}$$

or in vector notation $\boldsymbol{b} = \Phi \boldsymbol{c}$. Here E is the expansion order, which is part of the parameter list $\boldsymbol{\Pi}$. The basis set which we will employ in the examples are either Legendre polynomials or cubic splines. The formalism is, however, valid for any basis set and for other applications a different basis might be more appropriate. The probability for the background vector $\boldsymbol{b}$ is

$$
\begin{aligned}
p(\boldsymbol{b}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I}) &= \int d^E c\; p(\boldsymbol{b}|\boldsymbol{c},\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})\, p(\boldsymbol{c}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I}) \\
&= \int d^E c\; \delta(\boldsymbol{b}-\Phi\boldsymbol{c}) p(\boldsymbol{c}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I}) \qquad , \qquad (2)
\end{aligned}
$$

where we have used the probabilistic marginalization rule [11]. According to Bayes theorem [12,11] the desired probability $p(\boldsymbol{c}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})$ can be expressed as

$$
p(\boldsymbol{c}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I}) = \frac{1}{Z} p(\boldsymbol{d}|\boldsymbol{c},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})\, p(\boldsymbol{c}|\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I}) \qquad . \qquad (3)
$$

Bayes theorem splits the problem into the likelihood $p(\boldsymbol{d}|\boldsymbol{c},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})$, i.e. the error statistics of the experiment and the prior $p(\boldsymbol{c}|\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})$. The latter is the place to feed in all we know about the solution irrespective of the experimental data. The normalization Z can most easily be derived in the end by making sure that $\int d^E c\, p(\boldsymbol{c}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I}) = 1$.

A. The prior probability

We assumed, that the discriminating feature of the background is its smoothness. Consequently, the global first derivative is a characteristic quantity appropriate as testable information [13,14]. Hence, the prior probability according to the Maximum Entropy principle [13,14] becomes

$$
p(\boldsymbol{b}|\mu,\mathcal{I}) = \frac{1}{Z} e^{-\mu \sum_i (b'(x_i))^2} \qquad , \qquad (4)
$$

where $b'(x_i)$ stands for the slope of the background at point x_i. The hyper-parameter μ will be marginalized later on. The expansion in Eq. (1) yields

$$
\begin{aligned}
p(\boldsymbol{c}|\mu,\mathcal{I}) &= \pi^{-E/2} \mu^{E/2} (\det D)^{1/2} e^{-\mu \boldsymbol{c}^T D \boldsymbol{c}} \\
D_{l_1,l_2} &= \sum_i^N \phi'_{l_1}(x_i)\phi'_{l_2}(x_i) \qquad , \qquad (5)
\end{aligned}
$$

where we have included the normalization factor. Obviously, $\phi'_l(x) \equiv 0$ for a constant function, i.e. D has one eigenvector ψ_0 with eigenvalue zero. In order to deal with a normalizable prior we modify $D \to D + \epsilon\psi_0\psi_0^T$ and let ϵ tend to zero in the final results. It can be shown that this is equivalent to replacing $\det(D)$ in Eq. (5) by $\det(D')$

$$\det(D') := \prod_{\substack{i \\ \lambda_i > 0}} \lambda_i \qquad , \tag{6}$$

where λ_i are the eigenvalues of D. The regularization parameter μ is a hyperparameter to be integrated out according to the rules of probability $p(\boldsymbol{c}|\mathcal{I})=\int d\mu p(\boldsymbol{c}|\mu,\mathcal{I})p(\mu|\mathcal{I})$. The prior for the scale parameter μ is Jeffreys' prior $p(\mu|\mathcal{I}) = 1/\mu$ and we obtain the multivariate student-t distribution

$$p(\boldsymbol{c}|\mathcal{I}) = \pi^{-E/2}(\det D')^{1/2}\Gamma(E/2)(\boldsymbol{c}^T D\boldsymbol{c})^{-E/2} \qquad , \tag{7}$$

It should be pointed out, that Jeffreys' prior is not normalizable. It can, however, be considered as limiting distribution of a sequence proper priors. Since the posterior probability will be proper, the missing normalization constant of Jeffreys' prior drops out.

B. The Marginal Likelihood

The term $p(\boldsymbol{d}|\boldsymbol{c},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})$ in Eq. (3) is actually a *marginal* likelihood since the experimental data consist of signal plus background plus noise $\boldsymbol{d} = \boldsymbol{s} + \boldsymbol{b} + \boldsymbol{\eta}$ and the signal-part is not specified in $p(\boldsymbol{d}|\boldsymbol{c},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})$. Hence, the signal has been integrated by the marginalization means of BPT [15]. The idea we put forth is an extension of a suggestion made by Sivia and Press [16,17] on how to deal with *duff* data, a collection of data-points which are inconsistent in the sense that they lay far outside each other's error bars. Recalling the idea exposed in the beginning, the spectrum consists of regions which have no signal contribution and those which have a signal contribution and can be identified with the *duff* data in the Sivia-Press approach. We introduce the propositions B_i: '*data point i is purely background*' and the complement $\overline{B_i}$: '*data point i contains a signal contribution*'. In the first case the likelihood is simply the statistics of the experiment. We allow for two common situations:

uncorrelated Gaussian or Poisson distribution. For data point i the likelihood reads

$$p(d_i|B_i, b_i, \boldsymbol{\sigma}, \mathbf{\Pi}, \mathcal{I}) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(d_i-b_i)^2/2\sigma_i^2} \; ; & \text{Gaussian} \\ e^{-b_i} \frac{b_i^{d_i}}{d_i!} \; ; & \text{Poisson} \end{cases} , \tag{8}$$

In the second case, the data point contains a signal contribution s_i which easily predominates the background part. In this case the likelihood is again given by

$$p(d_i|s_i, b_i, \boldsymbol{\sigma}, \mathbf{\Pi}, \mathcal{I}) = \begin{cases} \frac{1}{\sqrt{2\sigma_i^2}} e^{-(d_i-b_i-s_i)^2/2\sigma_i^2} \; ; & \text{Gaussian} \\ e^{-(b_i+s_i)} \frac{(b_i+s_i)^{d_i}}{d_i!} \; ; & \text{Poisson} \end{cases} . \tag{9}$$

So far so good, but what is the value s_i for the signal? Here the marginalization rule [15] saves the day:

$$p(d_i|\overline{B_i}, \boldsymbol{c}, \boldsymbol{\sigma}, \mathbf{\Pi}, \mathcal{I}) = \int_0^\infty ds_i \; p(d_i|s_i, \boldsymbol{c}, \boldsymbol{\sigma}, \mathcal{I}) \, p(s_i|\mathbf{\Pi}, \mathcal{I}) \qquad . \tag{10}$$

We omitted all irrelevant entries in the conditional part of the probabilities. Again employing the maximum entropy principle along with the least committal testable information, namely the first moment, we have

$$p(s_i|\mathbf{\Pi}, \mathcal{I}) = \frac{1}{\xi} e^{-\frac{s_i}{\xi}} \tag{11}$$

In other words we introduce a scale ξ for the signal. This quantity is part of the parameter list $\mathbf{\Pi}$. There are as usual two possibilities, either the scale is known due to prior knowledge about the experimental setup or it is not and has to be inferred by the rules of probability theory. Now the marginal likelihood for the case '*an unspecified signal is included in the data*' can be evaluated analytically yielding

$$p(d_i|\overline{B_i}, \boldsymbol{c}, \boldsymbol{\sigma}, \mathbf{\Pi}, \mathcal{I}) = \begin{cases} \frac{1}{2\xi}(1 + \mathrm{erf}(\frac{d_i - b_i - \sigma_i^2/\xi}{\sqrt{2\sigma_i^2}})) \; ; & \text{Gaussian} \\ \frac{e^{\frac{b_i}{\xi}} Q(d_i+1, b_i(1+1/\xi))}{\xi(1+1/\xi)^{d_i+1}} \; ; & \text{Poisson} \end{cases} , \tag{12}$$

where $Q(a,x)$ is the regularized incomplete gamma-function $\Gamma(a,x)/\Gamma(a)$. The entire marginal likelihood of Eq. (3) constitutes a mixture model [18] of the two cases Eq. (8) and Eq. (12) with respective weights $\beta = p(B_i|\mathbf{\Pi},\mathcal{I})$ and $(1-\beta) = p(\overline{B_i}|\mathbf{\Pi},\mathcal{I})$. The parameter β completes the list $\mathbf{\Pi} = \{E, \xi, \beta\}$.

$$p(d|\boldsymbol{c},\boldsymbol{\sigma},\mathbf{\Pi},\mathcal{I}) = \prod_i \left[(1-\beta)\, p(d_i|B_i,\boldsymbol{c},\boldsymbol{\sigma},\mathbf{\Pi},\mathcal{I}) \;+\; \beta\, p(d_i|\overline{B_i},\boldsymbol{c},\boldsymbol{\sigma},\mathbf{\Pi},\mathcal{I})\right]. \tag{13}$$

The two likelihood functions of the mixture model are plotted in Fig.2. For each data point d_i there are two typical scales in the likelihood terms: the individual error σ_i and the global signal scale ξ. In both models, the smaller scale σ_i prevents the background from rising significantly above the data points. As far as deviations to lower values are concerned, the two likelihood functions behave differently: In the the first case Eq. (8) the background is tightly bound to the data points on a scale set by σ_i. Whereas the second likelihood Eq. (12) decays merely exponentially on a much larger scale set by ξ. So in the peak regions of the spectrum the second likelihood takes over since it penalizes only weakly the large discrepancy between the data and the background. As desired, the approach almost entirely ignores the signal-regions of the spectrum in the determination of the background parameters $\boldsymbol{c}$. This part of the spectrum is, however, important as far as inference of the parameters Π are concerned. An integral part of the probabilistic approach is the so-called Ockham's razor [19,20], which tries to keep the model as simple as possible. The first likelihood function, the one for the background-only cas, is simpler since it is less flexible. The measure of complexity that enters the formalism is the Ockham factor, the ratio of the amplitudes of the two likelihood functions, i.e. $\sqrt{\frac{\pi}{2}}\frac{\sigma_i}{\xi}$, which is a very small quantity. This factor penalizes individually those data points which are described by the more complex model, the one containing a signal contribution. The complex model wins only if there is no decent chance to interpret the data as background-only points. Qualitatively, the behavior is as follows: for $\beta = 1$ the entire spectrum is assumed to be background and the approach would minimize the misfit between the data and the background model. The resulting background would be much too large. In the opposite case ($\beta = 0$) it is assumed that all data points contain a signal contribution and therefore the resulting background estimate would be too small. In the intermediate case ($0 < \beta < 1$) Ockham's factor is active and tries to treat as many points as consistent with the data as background-only.

Given the posterior probability for the background we can easily compute the expectation value and confidence interval

$$
\begin{aligned}
\overline{b_i} &= \int b_i p(\boldsymbol{b}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})db \\
\overline{\Delta^2 b_i} &= \int (b_i - \overline{b_i})^2 p(\boldsymbol{b}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})db \ .
\end{aligned}
\tag{14}
$$

These quantities will be given for several PIXE spectra in a later section of the paper.

C. Probabilities for Π and B_i

The posterior probability still depends on yet unknown parameters $\boldsymbol{\Pi}$. Since the marginal likelihood and the prior are given, it is an easy matter of applying BPT to determine the joint probability for the parameters in the list $\boldsymbol{\Pi}$, namely E, β and ξ:

$$
p(\boldsymbol{\Pi}|\boldsymbol{d},\boldsymbol{\sigma},\mathcal{I}) = \frac{1}{Z} p(\boldsymbol{\Pi}|\mathcal{I}) \int d^E c \ \underbrace{p(\boldsymbol{d}|\boldsymbol{c},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})\, p(\boldsymbol{c}|\boldsymbol{\Pi},\mathcal{I})}_{\exp(-\psi)} \qquad . \tag{15}
$$

Unnecessary conditions have been omitted. We will employ the standard Laplace approximation to evaluate the integral analytically. To this end we expand ψ to second order around the maximum value $\hat{\mathbf{c}}$ yielding a Gaussian. In addition, since the Gaussian is restricted to a very narrow region, we can ignore the positivity constraint imposed on the background and the integral Eq. (15) can be evaluated analytically

$$
p(\boldsymbol{\Pi}|\boldsymbol{d},\boldsymbol{\sigma},\mathcal{I}) \approx \quad \frac{1}{Z} p(\boldsymbol{\Pi}|\mathcal{I}) p(\boldsymbol{d}|\hat{\mathbf{c}},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})\, p(\hat{\mathbf{c}}|\boldsymbol{\Pi},\mathcal{I})\ (2\pi)^{E/2} \det(\nabla\nabla^T \psi|_{\hat{\mathbf{c}}})^{-1/2} \qquad . \tag{16}
$$

The argument of the determinant is the Hessian. The prior for the parameters factors into $p(\boldsymbol{\Pi}|\mathcal{I}) = p(E|\mathcal{I})\, p(\beta|\mathcal{I})\, p(\xi|\mathcal{I})$ since expansion order E, prior background probability β and the signal scale ξ are logically independent. The expansion order is certainly restricted to a moderate upper limit E^* yielding a flat prior $p(E|\mathcal{I}) = \theta(E \le E^*)/(E^*)$. The uninformative prior for β is $p(\beta|\mathcal{I}) = \theta(0 \leq \beta \leq 1)$ while for the scale parameter ξ it is Jeffreys' [8,9] prior $p(\xi|\mathcal{I}) = 1/\xi$. For the expansion order E it is again Ockham's razor that implicitly tends to keep the model as simple as possible, i.e. it favors small expansion orders. The driving force for the Ockham factor is in the prior for $\boldsymbol{c}$ Eq. (7).

An interesting quantity is the 'background probability' i.e. the probability for proposition B_i, $p(B_i|\boldsymbol{d},\boldsymbol{\sigma},\mathcal{I})$. In order to determine the 'background probability' we employ the sum-rule to introduce the missing bits of information, i.e. parameter set $\boldsymbol{\Pi}$ and background coefficients $\boldsymbol{c}$

$$p(B_i|\boldsymbol{d},\boldsymbol{\sigma},\mathcal{I}) = \int d\Pi \int d^E c\, p(B_i|\boldsymbol{d},\boldsymbol{c},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})\, p(\boldsymbol{c}|\boldsymbol{d},\boldsymbol{\sigma},\boldsymbol{\Pi},\mathcal{I})\, p(\boldsymbol{\Pi}|\boldsymbol{d},\boldsymbol{\sigma},\mathcal{I}). \tag{17}$$

The probabilities for $\boldsymbol{c}$ and $\boldsymbol{\Pi}$, respectively, are sharply peaked at $\hat{\mathbf{c}}$ and $\hat{\boldsymbol{\Pi}}$. Since the background probability is just a diagnostic tool, it suffices to replace Eq. (17) by $p(B_i|\boldsymbol{d},\boldsymbol{\sigma},\mathcal{I}) = p(B_i|\boldsymbol{d},\hat{\mathbf{c}},\boldsymbol{\sigma},\hat{\boldsymbol{\Pi}},\mathcal{I})$. The next step uses Bayes' theorem

$$\begin{aligned}
p(B_i|\boldsymbol{d},\hat{\mathbf{c}},\boldsymbol{\sigma},\hat{\boldsymbol{\Pi}},\mathcal{I}) &= \frac{1}{Z} p(d|B_i,\hat{\mathbf{c}},\boldsymbol{\sigma},\hat{\boldsymbol{\Pi}},\mathcal{I})\ \underbrace{p(B_i|\hat{\boldsymbol{\Pi}},\mathcal{I})}_{1-\hat{\beta}} \\
&= \frac{(1-\hat{\beta})}{Z} \prod_{j\neq i} p(d_j|B_i,\hat{\mathbf{c}},\boldsymbol{\sigma},\hat{\boldsymbol{\Pi}},\mathcal{I})\ p(d_i|B_i,\hat{\mathbf{c}},\boldsymbol{\sigma},\hat{\boldsymbol{\Pi}},\mathcal{I}) \\
&= \frac{(1-\hat{\beta})\ p(d_i|B_i,\hat{\mathbf{c}},\boldsymbol{\sigma},\hat{\boldsymbol{\Pi}},\mathcal{I})}{\hat{\beta}\ p(d_i|\overline{B_i},\hat{\mathbf{c}},\boldsymbol{\sigma},\hat{\boldsymbol{\Pi}},\mathcal{I})\ +\ (1-\hat{\beta})\ p(d_i|B_i,\hat{\mathbf{c}},\boldsymbol{\sigma},\hat{\boldsymbol{\Pi}},\mathcal{I})}
\end{aligned} \tag{18}$$

In the last line we inserted the normalization which is the sum over the two cases of B_i being true or false. We define an average background probability

$$\overline{p(B_i|\boldsymbol{d},\boldsymbol{\sigma},\mathcal{I})} = \frac{1}{N}\sum_{i=1}^{N} p(B_i|\boldsymbol{d},\boldsymbol{\sigma},\mathcal{I})\ . \tag{19}$$

It is expedient to express the signal scale ξ in units of the average data value

$$\overline{d} = \frac{1}{N}\sum_{i=1}^{N} d_i\ . \tag{20}$$

This completes the formalism needed to determine all quantities of interest. The remaining task is the numerical evaluation of the formalism, i.e. first the determination of the most probable values for the parameters E, β, ξ and the computation of the expectation values for the background $\boldsymbol{b}$ and the respective confidence intervals.

III. RESULTS

We begin the discussion of the results with the ivory PIXE spectrum depicted in Fig.1. The afore-mentioned intuitive approach would first identify 'background-only' points and then fit a smooth function to these data points. The Bayesian analysis allows to quantify the detection of background-only points. In Fig.3 those data points are marked by solid circles for which the background probability $p(B_i|\boldsymbol{d}, \boldsymbol{\sigma}, \mathcal{I})$ as determined in Eq. (18) is greater then 90%. The latter value for the threshold in Fig.3 is arbitrary and serves merely a diagnostic purpose. For the present example the average background probability $\overline{p(B_i|\boldsymbol{d}, \boldsymbol{\sigma}, \mathcal{I})} = 0.9$, which agrees fairly well with our intuitive conception. The expectation value for the background $\overline{b_i}$ as defined in Eq. (14) is depicted in Fig.4 and it is compared with the results obtained by the rolling-ball and the GeoPIXE method, respectively. We observe that both, the Bayesian as well as the GeoPIXE result are reasonable and in close agreement, while the rolling ball result is somewhat disappointing. The signal contribution, which is defined here as simply the difference between PIXE spectrum and background is shown in Fig.5. The computed confidence intervalls are comparable to the original errorbars and the five small peaksof approximately 50 counts amplitude are all significant. We have omitted the errorbars in Fig.5 in order not to overload the figure. In the present example the Bayesian approach and the GeoPIXE method yield comparably good results, while the rolling ball result is somewhat disappointing. The most probable value for the signal scale is $\xi = 1.9\overline{d}$. A few additional remarks are in order. The abscissa values of the support-points of the cubic splines have been chosen equidistantly with a spacing set by twice the width of the narrowest signal peak. This is part of our prior-information $\mathcal{I}$, which allows us to distinguish signal from background.

Next we turn to a geological grain sample where the detector had a dip between 8 and 10keV. The PIXE spectrum in Fig.6 is again decomposed into background-only and signal-carrying data points. The average background-probability is again $\overline{p(B_i|\boldsymbol{d}, \boldsymbol{\sigma}, \mathcal{I})} = 0.9$. The signal scale is here $\xi = 1.0\,\overline{d}$. It is noteworthy that we generally found fairly good results for the background if we use $\xi \approx= \overline{d}$ and $\beta \approx 0.5$ instead of the correct most probable values. The inferred background is shown in Fig.7. It is again compared with the results obtained by the rolling-ball and the GeoPIXE method, respectively. In this example the Bayesian approach yields again a satisfactory result, while the outcome of the other methods is really disappointing. This is particularly obvious in the extracted signal which is shown in Fig.8. Apart from the main peaks, all other structures are zero within the confidence intervals.

We have compared our method with GeoPIXE and rolling ball for a variety of PIXE spectra. We have seen in the presented examples that the Bayesian approach furnishes good results. The same quality was observed in all analized spectra and the Bayesian results are used as standards to measure the quality of the other approaches. Instead of presenting a large collection of graphs we introduce a figure of merit $\kappa^{\mathrm{m}} = \sum_i (b_i^{\mathrm{m}} - b_i^{\mathrm{bayes}}) / \sum_i |b_i^{\mathrm{bayes}}|$, to assess the quality of the two common methods, GeoPIXE and rolling balls. The sum extends over all data points and b_i^{m} stands for the background at data point i obtained with method m. As pointed out before, the Bayesian result serves as standard. The results are listed in Table I. The samples are identified by there filenames. The table contains two PIXE spectrum of the ivory tusk (ivo) of an African elephant, the study was to determine whether you can localize the tusks from the trace element signature, this has implications in poaching control. 'Oto' stands for geological grains. This sample is particularly interesting for us, since the detector used for the measurement was showing a strange efficiency behaviour, which makes the background separation more challenging. The 'Geop' is the spectrum of a geopolymer.

Neither of the standard methods is clearly preferable over the other. In both methods it happens that the background goes far into the signal part or is much too small, which is indicated by large positive or negative κ_1-values. The first row corresponds to the spectrum of Fig.3 and the second row to the spectrum of Fig.6.

Finally, we consider an example for which the spline basis is disadvantageous and which is best treated in the Legendre basis. It is the mock spectrum depicted in Fig.9 which consists of two signal peaks, broadened by a rather broad apparatus function. The signal sits on a parabolic background. This example had been introduced in ref. [7] to illustrate the importance of background elimination in the quantified maximum entropy scheme. In Fig.9 the inferred background is given for expansion order 1-4, which corresponds to polynomials of degree 0-3. We see that we need expansion order 3 to describe the background. Expansion to higher orders does, however, not improve the quality of the fit through the background-only data points. The probability for the expansion order, which is depicted in Fig.10, is indeed sharply peaked at $E = 3$. We again encounter the interplay of data-constraint and simplicity. An expansion order less than 3 is not sufficient to fit the data while an expansion order beyond 3 does not pay off fit-wise and is therefore penalized by Ockham's factor. Fig.10 also shows the typical asymmetry in this probability. The left flank is described by a Gaussian due to the likelihood while the right flank follows a power-law decay dictated by Ockham's razor. For the optimal expansion order (3) the background is subtracted from the mock data furnishing the desired bare signal also shown in Fig.9. The signal still contains the experimental broadening which can now easily be deconvoluted by standard QME, since there is no background left which could give rise to ringing or other artificial structures.

IV. SUMMARY

We have demonstrated how the rules of probability theory can be used to separate the signal- from the background-part of a spectrum. The probabilistic approach has been tested in a variety of cases and it appeared that the results were in all cases satisfactory. Of course, the probabilistic approach is superior to any other ad-hoc method, since it provides the frame to consistently and rigorously exploit all bits of information available for a given problem [8,9]. If, contrary to expectation, another approach leads to 'better' results that would merely mean that this method employs information which has been withheld from the probabilistic approach. The probabilistic approach has only one drawback, it is slightly more laborious than ad-hoc methods, which are usually geared to be computationally simple and fast. The reader, who is mainly interested in a quick and dirty approach, can, however, simplify the formalism by setting $\beta \approx .5$ and $\xi \approx \frac{1}{N}\sum_{i=1}^{N} d_i$. The remaining task is merely the maximization of the marginal likelihood Eq. (13) with respect to the background-expansion coefficients c_i, which can be accomplished by standard library packages. The computational effort is then comparable to that of non-linear least-squares problems. In general, the parameters Π have to be optimized according to their posterior probability. But even for a data-set comprising 2000 data points the entire analysis takes approximately one minute on a medium-sized workstation.

REFERENCES

[1] P. V. Espen, K. Janssens, and J. Nobels *Cheamom. Int. Lab. Syst.*, **1**, p. 109, 1986.

[2] G. Johannsson *X-ray Spectrom.*, **11**, p. 194, 1982.

[3] B. Vekemans, K. Jannssens, L. Vincze, F. Adams, and P. V. Espen *Spectrochimica Acta*, **50B(2)**, p. 149, 1995.

[4] J. Maxwell, J. Campbell, and W. Teesdale *Nucl. Instr. and Meth.*, **B43**, p. 218, 1989.

[5] M. Kneen and H. Annegarn *Nucl. Instr. and Meth.*, **B109/110**, p. 209, 1996.

[6] C. Ryan, D. Cousens, S. Sie, W. Griffin, G. Suter, and E. Clayton *Nucl. Instr. and Meth.*, **B47**, p. 55, 1990.

[7] W. von der Linden, V. Dose, and R. Fischer, "How to separate the signal from the background," in *MAXENT96 - Proceedings of the Maximum Entropy Conference 1996*, M. Sears, V. Nedeljkovic, N. E. Pendock, and S. Sibisi, eds., p. 146, NMB Printers, Port Elizabeth, South Africa, 1996.

[8] H. Jeffreys, *Theory of Probability*, Oxford University Press, London, 1939.

[9] M. Tribus, *Rational Descriptions, Decisions and Design*, Pergamon, New York, 1969.

[10] E. Jaynes, *Probability Theory: The Logic of Science.* http://omega.albany.edu:8008/JaynesBook.html.

[11] D. Sivia, *Data Analysis - A Bayesian Tutorial*, Clarendon Press, Oxford, 1996.

[12] R. T. Bayes, "Essay toward solving a problem in the doctrine of chances," *Philos. Trans. R. Soc. London*, **53**, pp. 370–418, 1763.

[13] E. Jaynes, "Prior probabilities," in *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, R. Rosenkrantz, ed., p. 114, Reidel, Dordrecht, 1983.

[14] J. Kapur and H. Kesavan, eds., *Entropy Optimization Principles with Applications*,

Academic Press, Inc., San Diego, 1992.

[15] E. Jaynes, "Marginalization and prior probabilities," in *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, R. Rosenkrantz, ed., p. 337, Reidel, Dordrecht, 1983.

[16] D. Sivia, "Dealing with duff data," in *MAXENT96 - Proceedings of the Maximum Entropy Conference 1996*, M. Sears, V. Nedeljkovic, N. E. Pendock, and S. Sibisi, eds., p. 131, NMB Printers, Port Elizabeth, South Africa, 1996.

[17] W. H. Press, "Understanding data better with bayesian and global statistical methods," in *Unsolved Problems in Astrophysics, Proceedings of Conference in Honor of John Bahcall*, J. P. Ostriker, ed., Princeton University Press, Princeton, 1996. preprint in astro-ph/9604126.

[18] A. O'Hagan in *Kendall's advanced theory of statistics, Bayesian Inference*, John Wiley & Sons, New York, 1st ed., 1994.

[19] A. Garrett, "Ockham's razor," *Physics World*, **Mai**, p. 39, 1991.

[20] A. Garrett, "Ockham's razor," in *Maximum Entropy and Bayesian Methods*, W. Grandy, Jr. and L. Schick, eds., p. 357, Kluwer Academic Publishers, Dordrecht, 1991.

FIGURES

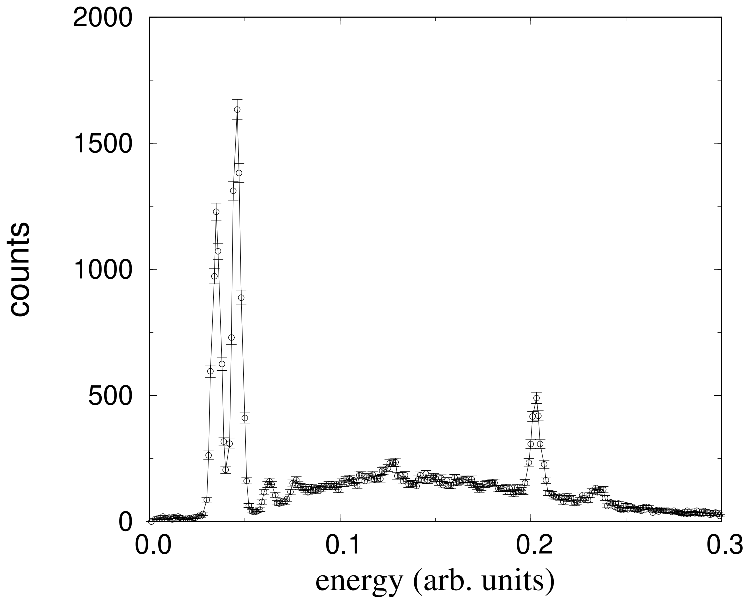

FIG. 1. *Illustrative PIXE spectrum of ivory.*

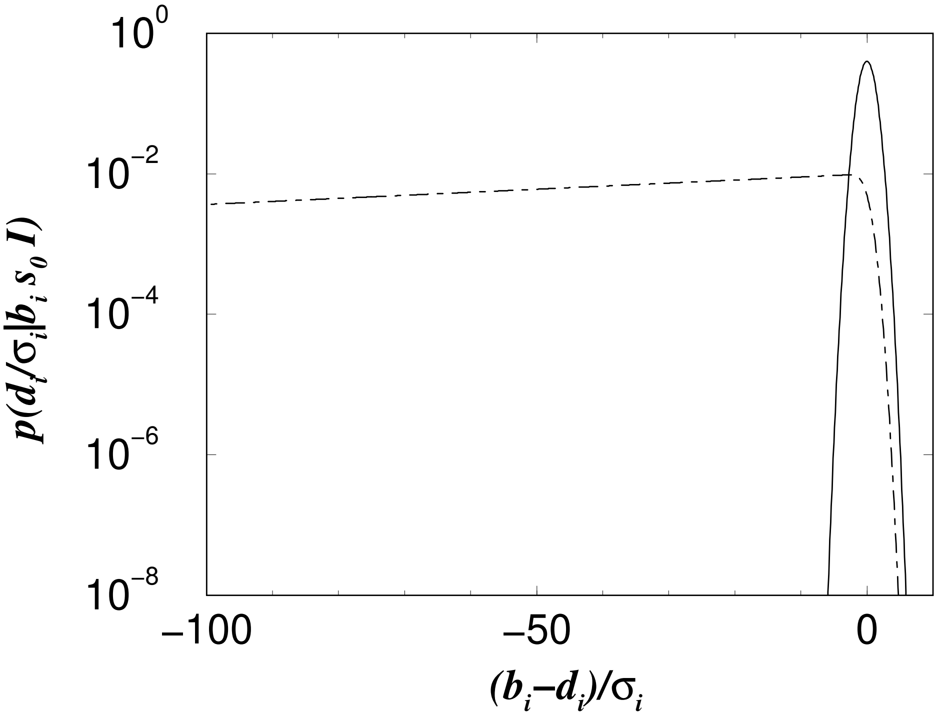

FIG. 2. *The two contributions to the marginal likelihood Eq. (13) terms for the cases bare background (solid line) and background plus marginalized signal (dashed line) for* $\xi = 100\sigma$. *Here only the Gaussian likelihood is shown, the curves for for the Poisson likelihood look qualitatively the same.*

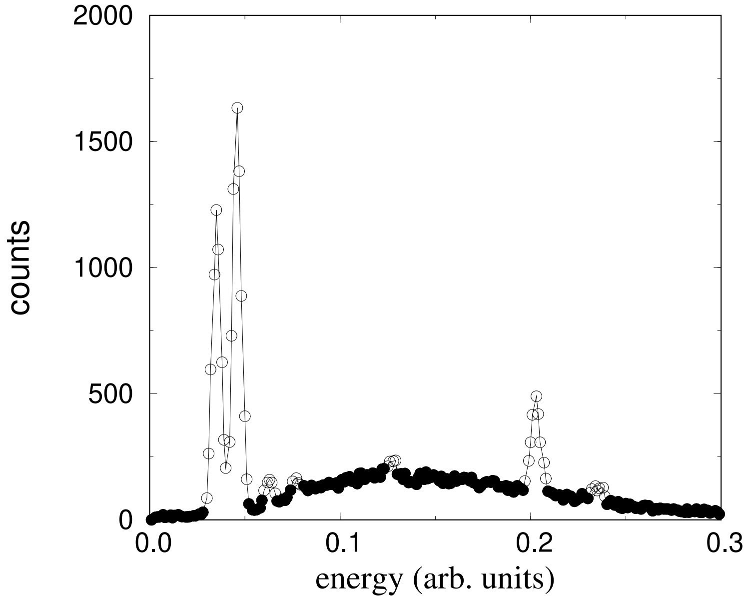

FIG. 3. *Same ivory PIXE spectrum as in Fig.1. Data points which are identified to carry a signal contribution are marked by open circles while the 'background-only' points are marked by solid circles.*

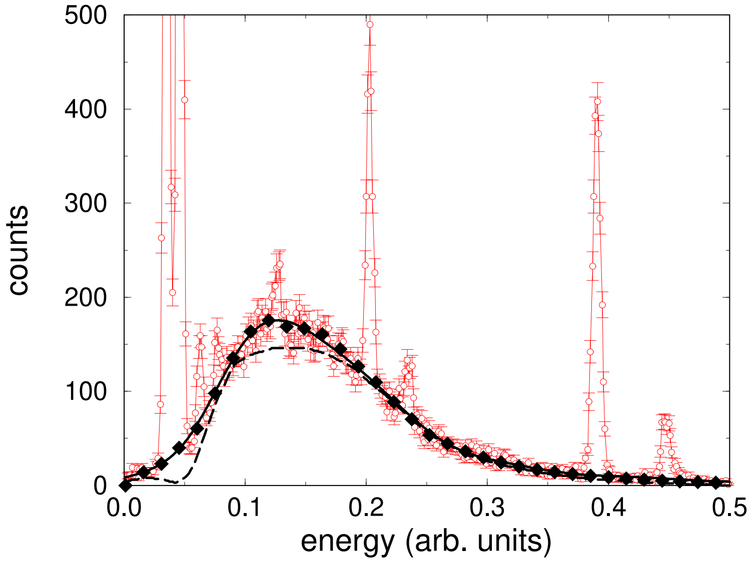

FIG. 4. *Comparison of the reconstructed background part of the ivory PIXE spectrum of Fig.1. Solid line: bayes, Dashed line: rolling ball, diamonds: G eoPIXE.*

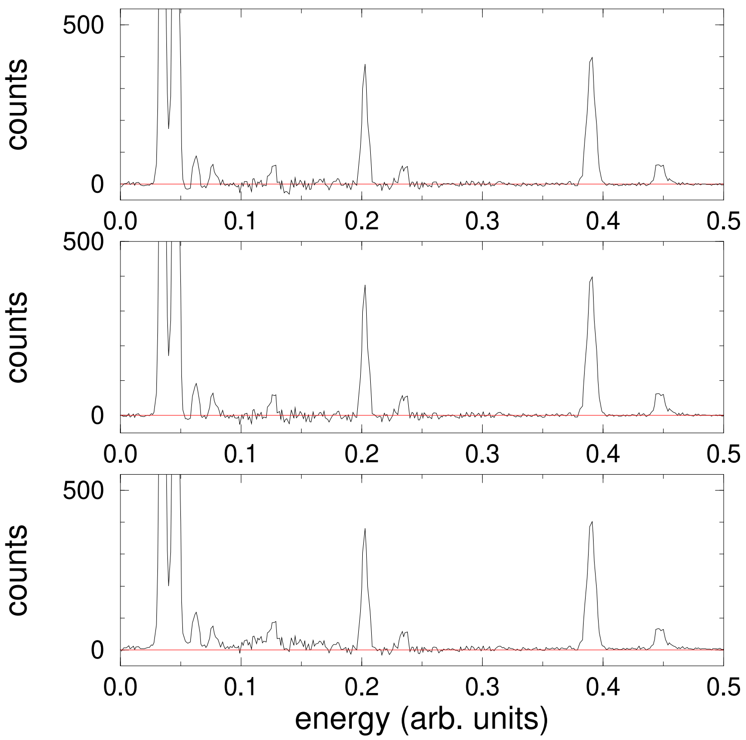

FIG. 5. *Signal contribution of the ivory PIXE spectrum of Fig.1. From top to bottom: Bayes, GeoPIXE, and rolling ball.*

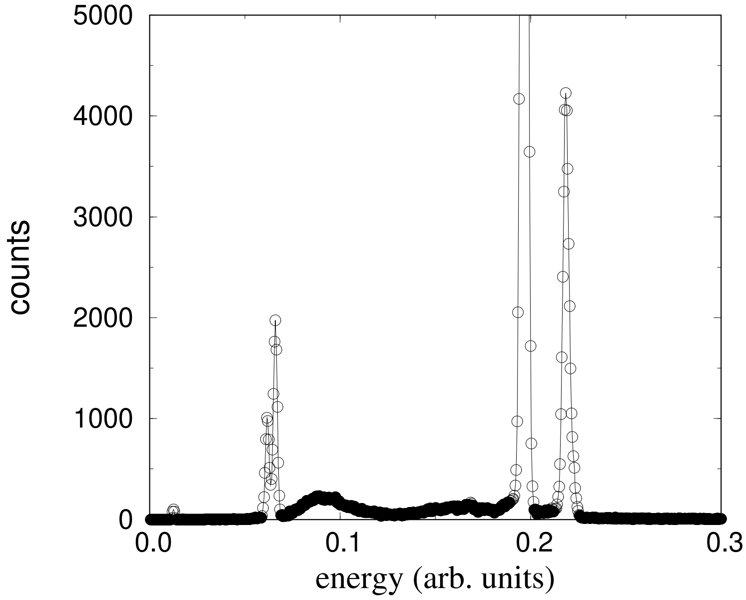

FIG. 6. *Geological-PIXE spectrum. Data points which are identified to carry a signal contribution are marked by open circles while the 'background-only' points are depicted as solid circles.*

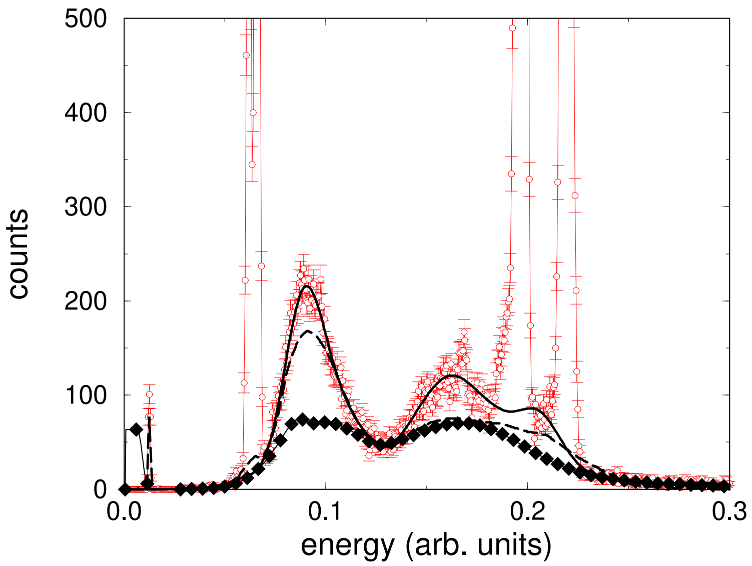

FIG. 7. *Comparison of the reconstructed background part of the PIXE spectrum of Fig.6. Solid line: bayes, dashed line: rolling ball, diamonds: G eoPIXE.*

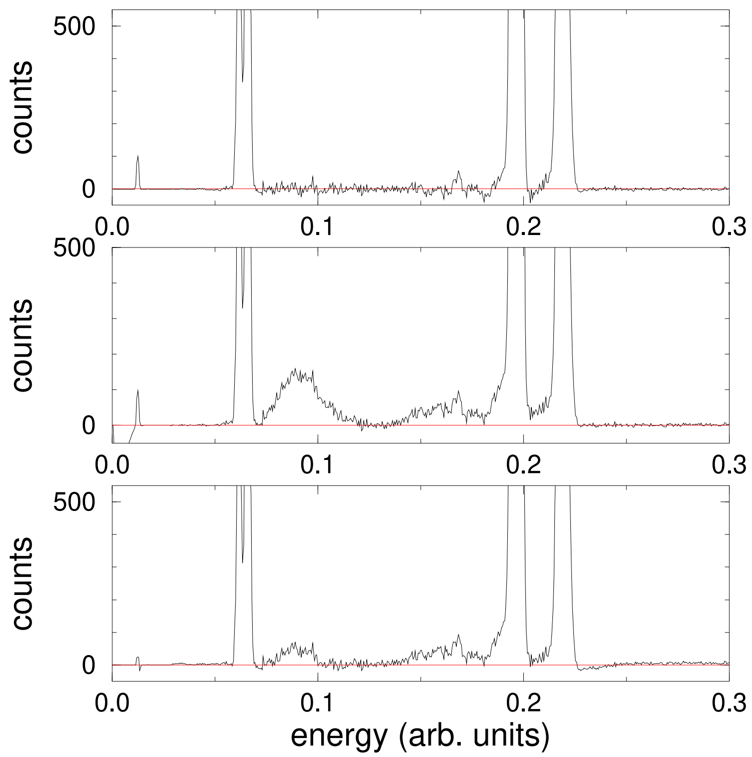

FIG. 8. *Signal contribution of the PIXE spectrum of Fig.6. From top to bottom: Bayes, GeoPIXE, and rolling ball.*

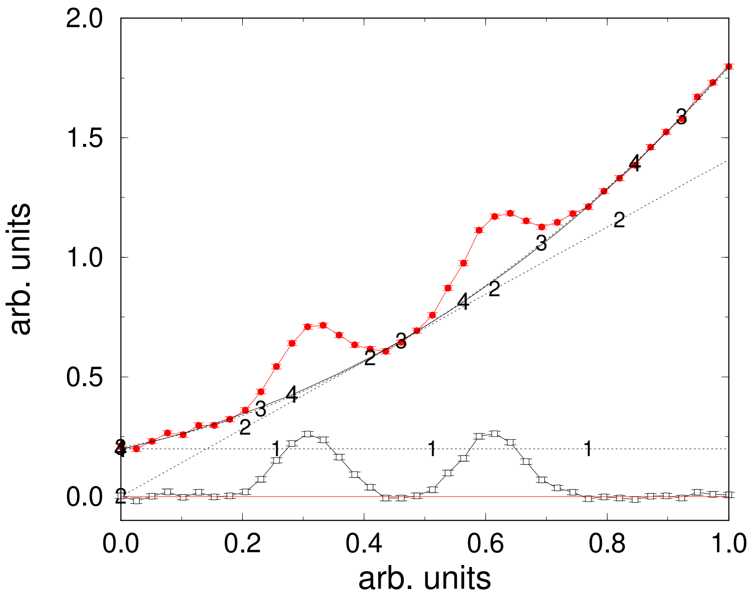

FIG. 9. *Result for the mock data of [7] for expansion order 1 to 4 as indicated on the curves. The result for expansion order 4 is depicted by a solid line. The mock spectrum is represented by solid circles plus error bars. The reconstructed signal along with the respective confidence intervals is shown at the bottom of the figure (squares).*

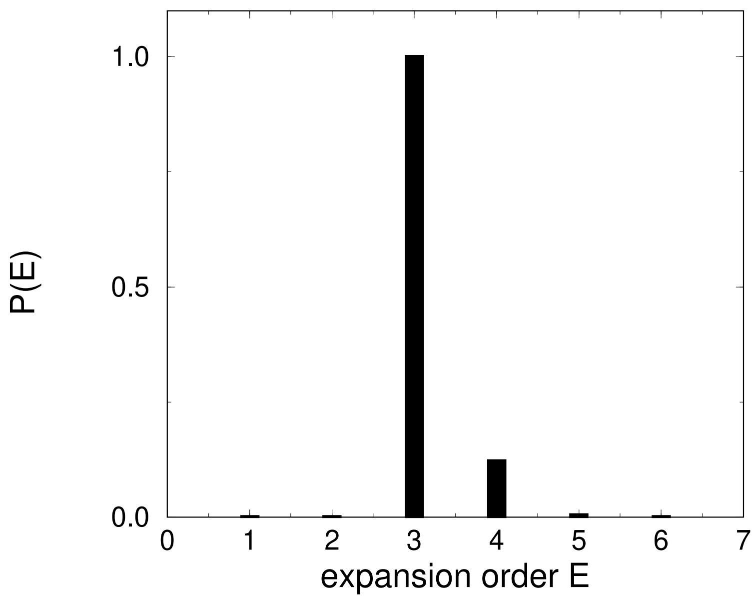

FIG. 10. *Probability for the expansion order E for the mock data set of [7].*

TABLES

TABLE I. Figure of merit κ for the GeoPIXE (gp) and rolling ball (rb) method computed for various spectra.

name	κ^{gp}	κ^{rb}
Ivo212	-0.02	-0.19
Oto100	-0.39	-0.22
Oto01	-0.37	0.00
Ivo100	1.56	0.03
Geop17	0.58	-0.11
AgZn18	0.74	3.65