Bayesian Inference in Surface Physics

Udo v. Toussaint and V. Dose
Max-Planck-Institut für Plasmaphysik,
EURATOM Association, Boltzmannstr. 2,
D-85748 Garching, GERMANY
email: udo.v.toussaint@ipp.mpg.de

Bayesian data analysis provides a consistent method for the extraction of information from physics experiments. The approach provides a unified rationale for data analysis, which both justifies many of the commonly used analysis procedures and reveals some of the implicit underlying assumptions. This paper introduces the general ideas of the Bayesian probability theory with emphasis on the application to the evaluation of experimental data, namely the deconvolution of the apparatus function for improving the energy resolution, the reconstruction of depth profiles from Rutherford backscattering measurements, handling of discordant data sets and mixture modelling for background estimation of Auger data.

PACS numbers: 82.80.Pv, 02.50Tt, 82.80Yc
Published with Appl. Phys. A: Received 11 March 2005
Accepted for publication 11 July 2005
Published online 16 September 2005
Appl. Phys. A **82** (2006), p.403-413

I. INTRODUCTION

Physics experiments are always affected by instrumental restrictions, limited measurement time and inaccurate standards. Therefore the typical problems encountered in the analysis of physics measurements are incomplete and noisy data. The reasoning about the interesting quantities is hampered by the the ill-posed nature of the underlying inversion problem. As a consequence, there is usually a huge number of widely scattered solutions consistent with the experimental data within the experimental errors. The experienced physicist has no problems in rejecting the 'unphysical' solutions among this manifold. This indicates that he disposes of prior information about the general nature of the 'physical' solutions. Bayesian Probability Theory (BPT) provides a general and consistent frame for data analysis problems. It allows for the inclusion of information prior to considering the data, which in turn permits to access ill-posed or underdetermined problems. Of paramount importance is the possibility to rank different candidate models for the explanation of a given set of data. In this paper we will present practical examples to which the Bayesian Probability theory can be applied successfully. Guided by these examples we will discuss typical features of the BPT. The paper is organized as follows. In section II we give a concise outline of the formalism with sufficient information for the reader to be able to apply the BPT. In section III to section V we present recent applications of the theory in order to illustrate its power. Section VI provides a summary and an outlook.

II. BAYESIAN CONCEPT

This section serves to state the theory and define the terminology employed in this paper. A more in depth coverage of the theory is provided by eg. [1, 2]. In Bayesian probability theory (BPT), the viability of a hypothesis H is assessed by calculating the probability of the hypothesis given the observed data D and any background information I. Following Jeffreys [3] we write such a probability as $P(H|D,I)$. The BPT rests on two rules [4] for manipulating conditional probabilities. The sum rule states that the probabilities of a proposition H and the proposition that H is false (signified by $\overline{H}$) add up to unity:

$$P(H|I) + P\left(\overline{H}|I\right) = 1. \tag{1}$$

Throughout this work, we will be concerned with exclusive and exhaustive hypotheses, so that if one particular hypothesis is true, all the others are false. For such hypotheses the normalization rule

$$\sum_i p(H_i|I) = 1 \tag{2}$$

holds. The second rule is the product rule which states that a joint probability or probability density function $P(H,D|I)$ can be factorized such that one of the propositions becomes part of the condition (i.e. moves right of the vertical bar). Due to the symmetry with respect to H and D, this can be done in two ways:

$$P(H,D|I) = P(H|I)\,P(D|H,I) = P(D|I)\,P(H|D,I). \tag{3}$$

Comparison of the two equivalent decompositions in (3) yields Bayes' theorem

$$P(H|D,I) = \frac{P(H|I)\,P(D|H,I)}{P(D|I)}. \tag{4}$$

Bayes' theorem relates the likelihood $P(D|H,I)$ to the posterior probability $P(H|D,I)$. The posterior probability distributions provide the full description of our state of knowledge about H. It is often necessary to summarize the distribution in terms of a few numbers. The most common description is given by the mean value of the posterior. Other possible choices are the position of the most probable value of the posterior (also termed maximum a posteriori (MAP) estimate) or the median of the distribution. For a symmetric distribution the mean value and the median coincide. All those numbers may be strongly misleading in the case of skew or multimodal distributions. Eq. 4 reveals also that the maximum-likelihood (ML) estimate is usually different from the posterior estimate except for the special case of a constant prior. The maximum-likelihood estimate obtained by maximizing the likelihood function is often mistaken as the most probable estimate given the data. This is not so: The obtained hypothesis is the one that would make the observed data most probable. This is logically quite different. An example taken from [5] highlights this distinction. The probability of rain given that there are clouds overhead and the probability of clouds overhead given that it is raining are clearly not the same. The quantity that is required (the most probable estimate given the data) is instead given by the posterior probability $P(H|D,I)$. It is related to the likelihood function through the prior probability $P(H|I)$. From a different point of view Bayes' theorem is a recipe for learning. Initially available prior knowledge about the hypothesis H coded in the distribution $P(H|I)$ is modified by the new information provided by the measured data D to its posterior distribution $P(H|D,I)$. The last quantity to be explained in Eq. 4 is $P(D|I)$. It follows from the marginalization rule which is itself a consequence of the sum and product rule. The extremely important marginalization rule tells how to remove an 'unwanted' nuisance variable from a Bayesian calculation:

$$P(D|M,I) = \sum_i P(H_i|M,I)\, P(D|H_i,M,I) \qquad (5)$$

Here we have split off the model M which specifies the bunch of hypotheses H_i we are considering from the general background information I. That is, the denominator of Bayes' theorem, which does not depend on H plays the role of a normalization constant. An additional significance of the evidence derived from Eq.5 is the probability of the data averaged over all hypotheses in the class specified by M. Therefore the evidence is of vital importance for model comparison (see IV).

The Bayesian formalism has been known for more than two centuries and it is extensively used in many fields as in robotics [6] or astronomy [7]. The routine use of Bayesian methods in the analysis of physics data, however, has still to come [8, 9]. The formalism is simple although the application may sometimes be computationally demanding. But the development of the last 10-15 years provided adequate computing power available to experimenters. So there are no longer any obstacles which prevent the use of BPT. Nevertheless the experience of many physicists trained in orthodox statistics that tackling realistic problems requires an (hardly available) arcane expert knowledge is in some areas only slowly superseded by the insight that there is a straightforward and general method for the evaluation of physics measurements.

A. Prior probability distributions

All of the rules we have written down so far show how to manipulate known probabilities to find the values of other probabilities and skipped the problem of how to formulate a distribution given certain *prior* knowledge. But to be useful in applications, we need rules that assign numerical values or functions to the initial probabilities that will be manipulated. Indeed, one of the advantages of Bayesian analysis is that it explicitly admits the existence of prior information. In other types of analysis it is often not easy to recognize the specific assumptions made by the analyst and (even worse) the implicit assumptions of the method (the latter assumptions are often unknown to the average practitioner). Prior information can consist of numerical values for the maximum, width or moments as mean or variance. Alternatively prior information can consist of properties which we expect for the posterior distribution of a problem. Essentially there are two different principles to derive a prior distribution.

1. *Transformation invariance*

E.T. Jaynes [10] (but also others eg. [11]) applied group theoretical methods to the problem of assigning priors. He demonstrated for a number of simple but practically important cases that, even if one is completely ignorant about the numerical values of the estimated parameters, the symmetry of the problem determines the prior unambiguously. Prominent examples are priors for scale parameters, location parameters or even priors for hyperplanes which are essential for Bayesian Neural Networks [12, 13]. We shall consider for concreteness in more detail the specific case of a prior for straight lines through the origin $y = ax$. A possible, naive prior for the slope of the straight lines would be $P(a|I) = const.$ On the other hand, the only sensible transformation of the coordinate system is in our specific case a rotation. $P(a)\, da$ is then an element of probability mass whose value must be independent from the system of coordinates which we use to evaluate its numerical value. Hence, for a different system of coordinates a' we must require that

$$P(a)\, da = P(a')\, da' \qquad (6)$$

yielding the Transformation Invariance equation [12]

$$\frac{\partial}{\partial \epsilon} \left\{ p\left(T_\epsilon\left(a\right)\right) \left| \frac{\partial T_{\epsilon(a)}}{\partial a} \right| \right\}_{\epsilon=0} = 0 \qquad (7)$$

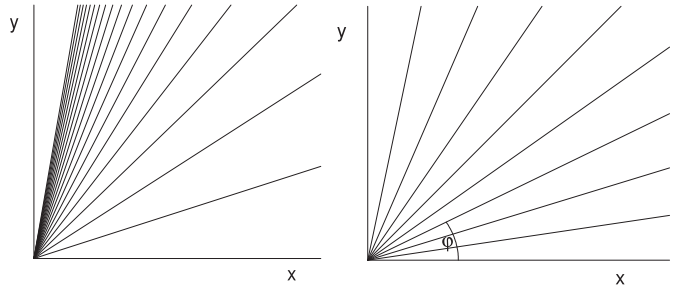

FIG. 1: Left panel: The density of the slope is constant. This results in a non-constant distribution for the angle between the straight lines and the x-axis. Right panel: The angular density is kept constant (from [14]).

where the infinitesimal transformation which maps a onto a' is denoted as $T_\epsilon(a)$. In our case this functional equation is solved by

$$p(a) = \frac{1}{\pi}\frac{1}{1+a^2} \tag{8}$$

which is not an obvious prior for the slope. But a visualization of both priors (Fig.1) shows that the prior Eq.8 is in agreement with our intuition: The angles of the straight lines with the x-axis are equally distributed (Fig.1, right hand panel), whereas the constant slope prior strongly favors large angles (Fig.1, left hand panel) [14].

If a location parameter is to be estimated, for instance the mean μ of a Gaussian, the prior must be invariant under a shift b of the location. The solution of Eq.7 is in this case a constant prior $P(\mu) = const.$ If we are indifferent about a scale parameter σ such as the decay length of an exponetial or the width of a Gaussian the appropriate prior satisfying transformation invariance is Jeffrey's prior [3]

$$P(\sigma) = \frac{1}{\sigma}. \tag{9}$$

Both priors $P(\mu) = const$ and $P(\sigma) = \frac{1}{\sigma}$ are called improper because they are not normalizable on their respective supports $-\infty < \mu < \infty$ and $0 \leq \sigma < \infty$. Improper priors should always be used with care in Bayesian calculations. The proper procedure is to consider eg. Jeffrey's prior as the limit of properly normalized priors on the support $\frac{1}{B} \leq \sigma \leq B$

$$P(\sigma) = \frac{1}{2\ln B}\frac{1}{\sigma}. \tag{10}$$

Inferences from the posterior are then considered for $B \to \infty$. If the inference depends on B in this limit the prior Eq. 9 leads to inconsistensies and the whole problem must be reassessed.

2. The Maximum Entropy Principle

A principle-based approach for coding numerical information into prior probability densities is the *Maximum Entropy* (ME) principle [15–17]. It is a rule for converting certain types of information called testable information to a probability assignment. The information $Q\left(\vec{\theta}\right)$ is testable if, given a probability distribution $p\left(\vec{\theta}|M,I\right)$, we can determine unambiguously if the distribution $p\left(\vec{\theta}|M,I\right)$ is consistent with the information $Q\left(\vec{\theta}\right)$. Q may be the already mentioned maximum or mean of a distribution. But in general, there may be many distributions consistent with the given testable information Q. For example we may know the mean value of many roles of a die was 2.5 and want to use this knowledge to assign probabilities to the six possible outcomes of the next role of the die. This information is testable - we can calculate the mean value of any probability distribution for the six possible outcomes of a roll and see if it is 2.5 or not - but it does not single out one distribution. The basic idea is to choose the prior probability distribution that is compatible with the given information yet has minimal information content otherwise. A functional satisfying this requirement is the entropy

$$S = -\sum_i p_i \ln p_i \tag{11}$$

subject to the constraining information. If the only information at hand is that the probability distribution is normalized to one in a interval $[a,b]$ then the ME principle provides a uniform distribution over the interval

$$P(\theta|Q_0 = 1, M, I) = \frac{1}{b-a} \tag{12}$$

If additionally the expectation value θ_0 of the distribution is given then the most uninformative distribution for positive variables $0 \leq \theta < \infty$ compatible with those constraints is

$$P(\theta|Q_0 = 1, Q_1 = \theta_0, M, I) = \frac{1}{\theta_0}\exp\left(-\frac{\theta}{\theta_0}\right). \tag{13}$$

As a final example assume we know the point estimate θ_0 of θ and also its variance $< \Delta\theta^2 >= \sigma^2$. In this case maximum entropy selects as the least informative distribution a Gaussian in $-\infty < \theta < \infty$

$$p\left(\theta|Q_0 = 1, Q_1 = \theta_0, Q_2 = \sigma^2, M, I\right) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(\theta-\theta_0)\right). \tag{14}$$

III. APPLICATIONS I: RUTHERFORD BACKSCATTERING

In the following sections will apply the BPT to various ill-posed inversion problems, encountered in analyz-

ing experimental data from surface physics experiments. We begin with the deconvolution of RBS data which is also of importance for the improved depth resolution of RBS measurements in the next paragraph. Rutherford backscattering (RBS) is a surface analytical technique which is routinely used to determine surface compositions and depth profiles[18]. Its importance derives from its quantitative nature. In RBS, the energy distribution $d\,(E)$ of backscattered ions is measured for a fixed scattering angle ϕ. In the lower MeV range an elastic Coulomb collision model can be employed, in which the energy of the backscattered ions, usually either protons or helium nuclei is determined by the incident energy E_0, the scattering angle ϕ and the mass ratio $r_i = m_0/m_i$ of projectile ion m_0 and target atoms m_i. Since the projectile target interaction is coulombic, the scattering cross section is the quantitatively known Rutherford scattering cross section and only the mass ratio is unknown. The energy E of the of the backscattered ions is given by

$$E = E_0 \left(\frac{\sqrt{1 - r_i^2 \sin^2 \phi} + r_i \cos \phi}{1 + r_i} \right)^2 \qquad (15)$$

From Eq. 15 it follows that ions undergoing a collision with a heavy target atom loose less energy than ions colliding with target atoms of lower atomic mass. In an ideal RBS experiment the energy distribution of a thin film sample $\tilde{d}(E)$ would be composed of delta peaks for the different masses. In the real world we have to deal with a limited resolution due to the apparatus function and also with thick samples. In a thick sample both primary ions and scattered ions loose energy on their way through the sample, depending on the stopping power. This enables RBS to be depth sensitive but may also give rise to overlapping peaks in the spectrum.

A. Deconvolution of apparatus functions

Small, cheap and easy-to-use semiconductors are used in most RBS experiments for the energy analysis of the backscattered particles. Their performance is hampered by the energy-loss straggling in the Au entrance electrode of the detector and in the dead layer of the detector and by the statistics of the electron-hole pair creation. Together with additional contributions to the energy broadening, namely energy spread of the incident beam, electronic noise of the detector-preamplifier system and (for higher fluxes) pile-up the achievable resolution is limited. Therefore the energy distribution for fixed target mass is rather broad. As long as the masses, or rather the respective backscattering energy distributions, are well separated, it is straightforward to extract the mass composition from the bare experimental data. If, however, the masses are similar, particularly in the case of isotopes, the information is not readily accessible. The different contributions to the energy broadening can be summarized in a transfer function of the whole system, the *apparatus function* $A(E)$. The measured spectrum d(E) is given by the convolution of the ideal spectrum $\tilde{d}(E)$ with the apparatus function:

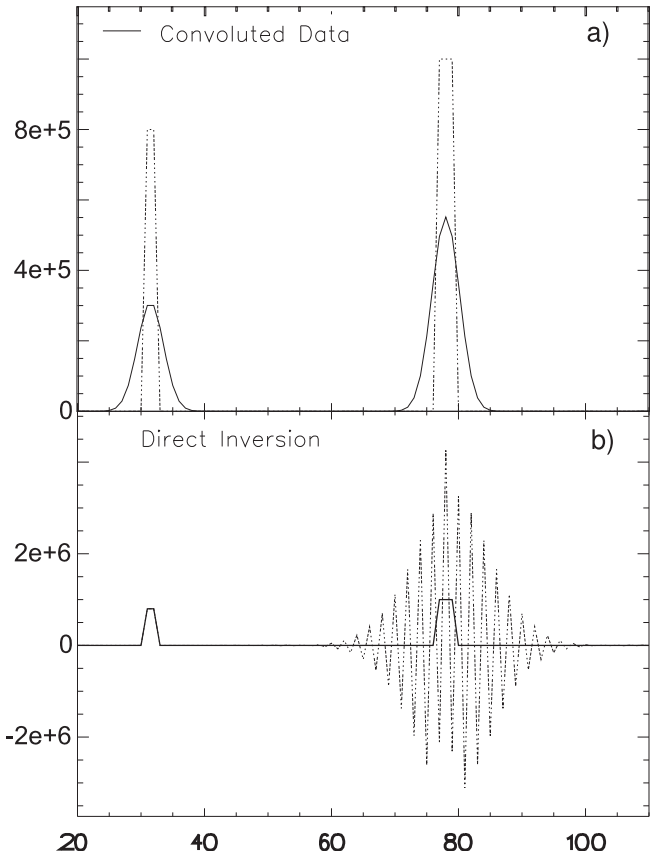

FIG. 2: a) Amplitude versus position. The initial spectrum (- -)is convolved with a Gaussian to yield the solid line. Before the inversion one count from the right-hand peak has been removed. b) The inversion result showing the dramatic influence of noise on the stability of deconvolution.

$$d(E_i) = \int_{-\infty}^{\infty} dE'\, \tilde{d}(E') A(E_i - E') \approx \sum_{j=1}^{N_d} A_{ij} \tilde{d}(E_j) \quad (16)$$

The matrix A_{ij} represents the discretized apparatus function, taking into account that the measured spectrum is binned. The convoluted spectrum d(E) can be calculated easily if $\tilde{d}(E)$ and A are known. The inversion of (16) yields the ideal spectrum. Unfortunately, the inversion is frequently utterly ill-posed if the eigenvalue spectrum of A_{ij} varies over orders of magnitude [19]. This is generally the case for Gaussian apparatus functions and entails a strong amplification of experimental errors. An illustrative example is shown in Fig.2. An assumed spectrum shown as the dashed curve in Fig.2a turns into the solid curve by convolution with a Gaussian of approximately the same width. Note that the signal (solid curve) in this artificial counting experiment has a maximum of about $5\text{x}10^5$ counts in the right-hand side spectral density. To

demonstrate the effect of noise, we changed the counts in the right-hand side peak channel by just one count. Fig.2b shows the dramatic influence of noise on the stability of the deconvolution. While the left-hand peak is restored correctly because its data were exact, the effect of noise on the level of 10^{-6} in the right-hand peak is disastrous. The reconstruction oscillates wildly and attains even unphysical negative values. To overcome this problem the statistical nature of the error has to be taken into account properly in conjunction with the intrinsic properties of the objective solution. The goal is to determine the posterior probability density $P\left(\vec{f}|\vec{d},\vec{\sigma},I\right)$ for the RBS spectrum f_j at the N energies E_j, given N_d experimental data d_i and the respective errors σ_i. Prior knowledge to be incorporated is that there may be correlations between neighboring channels: A random permutation of the energy channels results in a spectrum not accepted as RBS spectrum by any expert. The correlations are imposed on $\vec{f}$ through a convolution of a *hidden* density $\vec{h}$ with a smoothing kernel B with spatially varying widths. The image f is then obtained from

$$f\left(x,h,b\right)=\int dy\, B\left(\frac{x-y}{b\left(y\right)}\right)h\left(y\right). \qquad (17)$$

In [20] a gaussian kernel

$$B\left(\frac{x-y}{b\left(y\right)}\right)=\frac{1}{b\left(y\right)\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-y}{b\left(y\right)}\right)^2\right] \qquad (18)$$

has been used. In the Bayesian approach, since the interest is in $\vec{f}$, the nuisance parameters $\vec{h}$ and $\vec{b}$ have to be marginalized,

$$P\left(\vec{f}|\vec{d},\vec{\sigma},I\right)=\int d^Nh\, d^Nb\, P\left(\vec{f},\vec{h},\vec{b}|\vec{d},\vec{\sigma},I\right). \qquad (19)$$

Bayes theorem relates the yet unknown $P\left(\vec{f},\vec{h},\vec{b}|\vec{d},\vec{\sigma},I\right)$ to known quantities, namely the likelihood $P\left(\vec{d}|\vec{f},\vec{\sigma},I\right)$ and the prior probability densities $P\left(\vec{h}|I\right)$ and $P\left(\vec{b}|I\right)$ via

$$P\left(\vec{f},\vec{h},\vec{b}|\vec{d},\vec{\sigma},I\right)\propto P\left(\vec{d}|\vec{f},\vec{\sigma},I\right)P\left(\vec{f}|\vec{h},\vec{b}\right)P\left(\vec{h}|I\right)P\left(\vec{b}|I\right). \qquad (20)$$

The uninformative prior $P\left(\vec{h}|I\right)$ for a PAD is the entropic prior [21] and the prior $P\left(\vec{b}|I\right)$ constrains the kernel widths to a sensible range. Finally, the probability density $P\left(\vec{f}|\vec{h},\vec{b}\right)$ is given by $\delta\left(\vec{f}\left(\vec{x}\right)-\vec{f}\left(\vec{x},\vec{h},\vec{b}\right)\right)$ because the knowledge of $\vec{h}$ and $\vec{b}$ uniquely determines the value of $\vec{f}$.

The application of the adaptive deconvolution method is shown in Fig.3. The spectrum was measured with 2.6 MeV ^{4}He at a scattering angle of 165°. The apparatus function (left peak) was determined by measuring an RBS spectrum of a thin cobalt layer of about 0.75 nm thickness on a silicon substrate (cobalt is isotopicaly pure). The width of the Co peak is about 19 keV FWHM which reflects the limited resolution since the intrinsic energy spread due to energy-loss and energy-loss straggling in the thin Co layer is only about 3 keV. The apparatus function is slightly asymmetric. Using this measured apparatus function $\mathbf{A}$ with its pointwise uncertainty $\vec{\sigma}_A$ due to the counting statistics, the likelihood function $P\left(\vec{d}|\vec{f},\vec{\sigma},I\right)$ of counting experiments obeys the Poisson statistics. Since we deal with large numbers of counts the Poisson distribution is well approximated by a Gaussian distribution

$$P\left(\vec{d}|\vec{f},I\right)=\frac{1}{\prod_{i=1}^{N_d}\sqrt{2\pi\sigma_{eff,i}^2}}\cdot \exp\left[-\frac{1}{2}\sum_{i=1}^{N_d}\frac{\left(d_i-\sum_{j=1}^{N}A_{ij}f_j\right)^2}{\sigma_{eff,i}^2}\right] \qquad (21)$$

with $\sigma_{eff,i}^2=\sigma_i^2+\sum_{j=1}^{N}\sigma_{A,ij}^2f_j^2$ [23]. Using Eq.19 and the measured apparatus function for cobalt the copper signal on the right hand side was deconvolved[22]. After deconvolution, the two isotopes ^{63}Cu and ^{65}Cu are clearly resolved. The FWHM of the dominant ^{63}Cu peak after deconvolution is 3.0keV, which is about 6 times better than the achieved experimental resolution and far beyond any conceivable experimental resolution with the available setup. The measured abundances of the isotopes are 70.1% ^{63}Cu and 29.9% ^{65}Cu. This compares

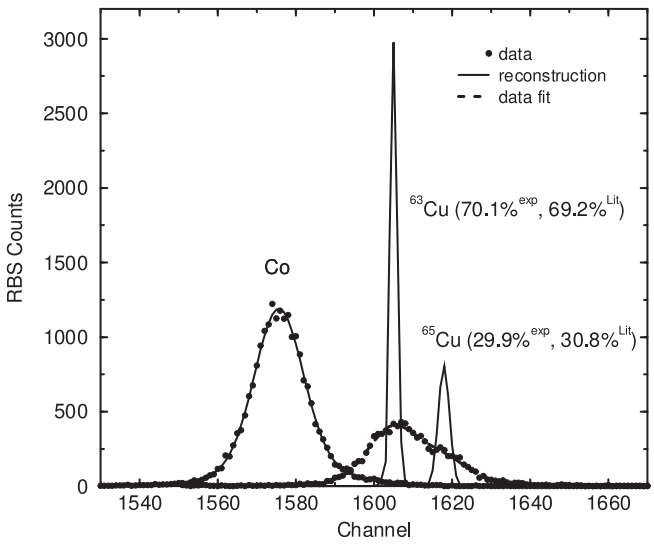

FIG. 3: RBS spectra of thin Co and Cu films on a Si substrate, measured with 2.6 MeV He. The Cu spectrum (right hand side) is deconvolved with the apparatus function obtained from the Co spectrum (left hand side). The two Cu isotopes are clearly resolved with measured abundances close to the natural abundances (from [22]).

favourably to the natural abundance of 69.2 % ^{63}Cu and 30.8 % ^{65}Cu.

B. Depth profiles

Backscattering spectroscopy using ion beams with energies in the MeV range is used extensively to determine the distribution of target elements in the sample as a function of depth below the surface. The ideal RBS spectra f is a linear superposition of the spectra of the individual elemental depth profiles $c_i(x)$. But the energy of the penetrating and backscattered particles depends in a complicated, nonlinear way on the sample composition and morphology. Therefore simulation codes are required to simulate an RBS spectrum given the sample. State-of-the-art software considers the energy-dependent electronic and nuclear stopping for the ions and energy loss straggling[24]. The depth profiles are then obtained by varying the sample parameters until a minimum quadratic misfit is achieved (Maximum-likelihood solution). Although this approach, guided by the experimentalists experience often succeeds it has the severe shortcoming that it does not solve the inverse problem[25]. A good fit is a necessary but not a sufficient condition: different depth profiles can result in very similar fits (see eg.Fig.2). The posterior expectation (the mean) for the concentration c is

$$\langle c \rangle = \int dc\, c\, p\,(c|d, I) \tag{22}$$

and the variance

$$\left\langle \delta c^2 \right\rangle = \int dc\, \left(c - \langle c \rangle\right)^2 p\,(c|d, I)\,. \tag{23}$$

The analysis is completely analogous to the one in the section III A. Only the linear relationship given by Eq.16 is now replaced by the forward calculation of the simulation codes

$$d'\,(E_i) = g\,(c\,(x)) \tag{24}$$

given a depth profile $c\,(x)$. A prior which incorporates the knowledge of the concentrations being larger than 0 and allows in addition for the inclusion of a default model is given by the entropic prior [21]. An example is provided by a study of first-wall materials for fusion experiments. Carbon is considered as a favorable first wall material for fusion reactors, in particular for plasma facing components subject to exceptionally high thermal heat loads. Apart from the lifetime of a material under such conditions, a critical issue in the case of carbon is the possible formation of significant tritium inventories by codeposition with redeposited carbon atoms [26, 27]. Both issues are mainly determined by the carbon erosion rate resulting from physical sputtering and chemical erosion[28]. To estimate the carbon erosion rates in the divertor of the fusion experiment ASDEX Upgrade

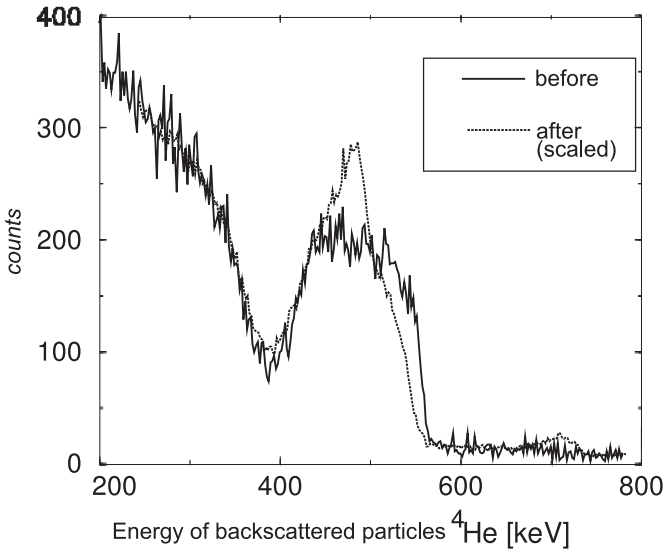

FIG. 4: RBS data of the sample before and after plasma exposition

graphite probes were covered with a 150nm layer of ^{13}C and exposed to a single plasma discharge. ^{13}C was used because the chemical erosion is unaffected by isotope substitution and to allow the measurement of redeposited ^{12}C eroded at other plasma facing components. The RBS spectrum of the sample before exposure is shown in Fig.4 as the solid line. The right peak indicates the ^{13}C layer on top of the ^{12}C sample. After exposure the high-energy edge is shifted towards lower energies, indicating the absence of ^{13}C at the surface. The increased intensity in the channels at 500keV indicates a mixture of ^{12}C and ^{13}C but no further information is easily extracted from the spectrum. Additionally a small amount of oxygen is present after exposure (peak at 700keV). The results of the Bayesian depth profile reconstruction are given in Fig.5[29]. Before exposure a ^{13}C layer can be seen, approximately 2.2x10^{18}atoms/cm^2 thick, but with an average contribution of 20% ^{12}C. After exposure most of the ^{13}C is still present but there is an additional layer of ^{12}C deposited on top of it. Oxygen has been codeposited. The RBS spectrum calculated from the estimated depth profiles agrees with the experimental data within the error bars [30]. The surprising result is the coexistence of erosion and deposition at the area where the outermost closed magnetic surface intersects the divertor. This so called 'strike-point' area experiences extremely high thermal loads and was considered as erosion dominated. At the same time this measurement shows that conclusions based on net changes in sample thickness may strongly underestimate the dynamical modifications.

IV. APPLICATIONS II: MODEL COMPARISON

So far we have seen the Bayesian approach to parametric estimation: Compute the posterior probability distri-

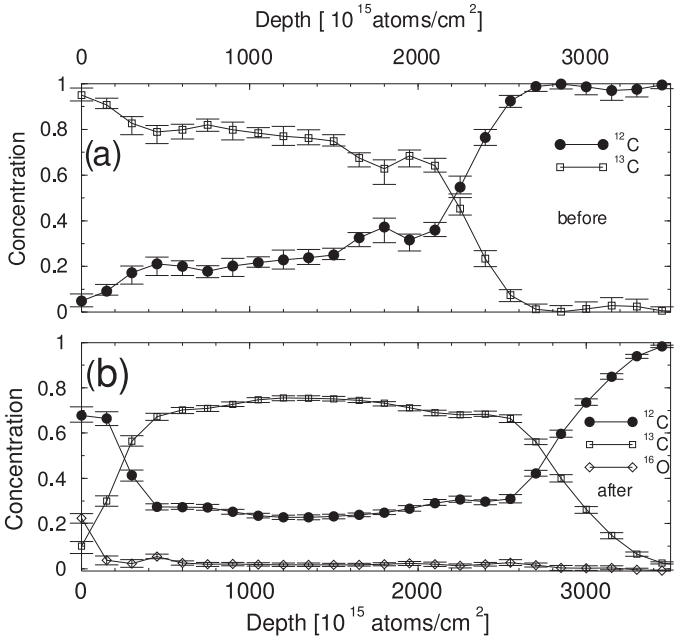

FIG. 5: Reconstructed depth profiles and asymmetric confidence intervals from the RBS spectra shown in Fig.4. The upper panel shows the sample composition before exposure, the lower panel after exposure (from [29]).

bution (and derived quantities like mean or variance) of the parameters of a given model. A more complex situation arises when we have several models M_i, each of which might depend on several, possibly different parameters. The formal Bayesian approach is identical to the one of parameter estimation, where the previous evidence plays now the role of the likelihood:

$$P\left(M|D,I\right) = \frac{P\left(D|M,I\right)P\left(M|I\right)}{P\left(D|I\right)}, \tag{25}$$

where $P\left(D|M,I\right)$ can be computed with the help of Eq.5. If we have no reason to prefer a model we can assign equal prior probabilities to all of them $P\left(M_i|I\right) = P\left(M_j|I\right)$. This is a frequently arising situation, but it should be kept in mind that more precise prior knowledge can be incorporated and should be used if available. With an indifferent state of knowledge about the prior model probability $P\left(M_i|I\right)$ the ratio of the posterior model probabilities $P\left(M_i|D,I\right)$ and $P\left(M_j|D,I\right)$ reduces to the ratio of the evidences, the so called Bayes factor

$$B_{ij} = \frac{P\left(D|M_i,I\right)}{P\left(D|M_j,I\right)}. \tag{26}$$

We can provide a simple interpretation of the evidence and the way Ockham's razor (avoiding unnecessarily complex models) is incorporated in the model comparison, as follows [31]. First, we write the marginal likelihood $P\left(D|M,I\right)$ in the form

$$P\left(D|M,I\right) = \int d\vec{\theta}\, P\left(D|M,\theta,I\right) P\left(\vec{\theta}|I\right). \tag{27}$$

In the by far most common cases the prior is much more diffuse than the likelihood. It's variations over the range where the likelihood peaks can then be neglected. Therefore we can move the prior term taken at θ_{ML}, the point where the likelihood attains its maximum value, outside the integral

$$P\left(D|M,I\right) \approx P\left(\vec{\theta}_{ML}|I\right) \int d\vec{\theta}\, P\left(D|M,\vec{\theta},I\right) \tag{28}$$

The remaining θ-integral over the likelihood may be further approximated by

$$P\left(D|M,I\right) \approx P\left(D|M,\vec{\theta}_{ML},I\right) P\left(\vec{\theta}_{ML}|I\right) \left(\Delta\theta_{like}\right)^{N_\theta} \tag{29}$$

where $\left(\Delta\theta_{like}\right)^{N_\theta}$ is the approximate likelihood volume. If we take the prior to be approximately uniform over some interval $\Delta\theta_{prior}$ larger than the posterior peak and note that the prior is normalized to one then we can estimate $P\left(\vec{\theta}_{ML}|I\right) \approx 1/\left(\Delta\theta_{prior}\right)^{N_\theta}$. Eq.29 becomes

$$P\left(D|M,I\right) \approx P\left(D|M,\theta_{ML},I\right) \left(\frac{\Delta\vec{\theta}_{like}}{\Delta\theta_{prior}}\right)^{N_\theta} \tag{30}$$

Under these assumptions the evidence is approximately equal to the maximum likelihood solution penalized by the second term, which is referred to as an Ockham factor. Since by assumption $\Delta\theta_{like} \ll \Delta\theta_{prior}$ the Ockham factor is $\ll 1$. With an increasing number of model parameters N_θ the improvements in the likelihood will eventually be counterbalanced by the decreasing second term in Eq.30 thus defining an optimal model complexity.

A. Mass Spectroscopy

Plasma-based surface processing is widely used in the microchip and display industry, where many manufacturing processes occur in plasma reactors. The identification and quantification of plasma products for processing control have become one of the urgent topics for plasma physicists. Detailed knowledge of concentrations of reactive particles like free radicals is needed to understand the underlying microprocesses [32]. Mass spectroscopy is a convenient technique to directly monitor the particle fluxes at the substrate sites. Traditional quadrupole spectrometers are widely used due to high sensitivity, reasonable stability and low costs. To be filtered in the quadrupole field, neutral gases have to be ionized, most commonly by electron impact. At a typical electron energy of 50-100 eV (used to achieve a high ionization efficiency) stable molecules decompose in a variety of fragment ions leading to the so called *cracking pattern*. For overlapping cracking patterns subtraction methods have been devised to disentangle the measured spectra[33]. These methods suffer from excessive error buildup and

are not applicable, when unstable constituents like radicals are assessed, due to the lack of knowledge of cracking patterns. Furthermore the fragmentation is also an instrument specific property and thus requires an instrument specific calibration. A rigorous analysis of composite mass spectra employs BPT which also succeeds without exact cracking patterns [34, 35].

Assuming a linear response of the mass spectrometer the mass signal vector of measurement j, $\vec{d}_j$ is the sum of the contributions of all species in the mixture

$$\vec{d}_j = \mathbf{C}\vec{x}_j + \vec{\epsilon}_j \tag{31}$$

with gaussian noise $\vec{\epsilon}$. The goal is to determine the posterior distribution of the cracking matrix elements $\mathbf{C}$, the vector $\vec{x}_j$ of species concentrations in measurement j and also the number of species E. $\vec{\epsilon}_j$ is the vector of measurement errors associated with the signal vector $\vec{d}_j$. The cracking column vectors are normalized to sum up to one. The likelihood is given as

$$P\left(\mathbf{D}|\mathbf{C},\mathbf{X},\{\mathbf{S}\},E,I\right) = \tag{32}$$
$$\prod_j \frac{1}{\prod_i \sqrt{2\pi} s_{ij}} \exp\left(-\frac{1}{2}\left(\vec{d}_j - \mathbf{C}\vec{x}_j\right)^T \mathbf{S}_j^{-1}\left(\vec{d}_j - \mathbf{C}\vec{x}_j\right)\right).$$

$\{\mathbf{S}\}$ denotes the ensemble of diagonal matrices $\mathbf{S}_j$ with components $(\mathbf{S}_j)_{ii} = s_{ij}^2$, given by the measurement error of j-th measurement in the i-th mass channel. The only components which still need to be specified to start the Bayesian inference are the prior distributions for the number of components $P(E|I)$, the concentration matrix $P(\mathbf{X}|E,I)$ and finally the cracking matrix elements $P(\mathbf{C}|E,I)$. For the prior probability a constant prior is chosen $P(E|I) = 1/E_{\max}$. Cracking patterns of stable molecules are listed as point estimates, e.g. in the tables of Cornu and Massot [36]. Together with the requirement that the cracking coefficients are confined to the interval $[0,1]$ this allows the computation of an exponential prior for the cracking coefficients $P(\mathbf{C}|E,I)$. Note however that this prior, though still exponential, is more complicated than Eq.13 since the support of the cracking coefficients is not infinite but rather confined to the interval $[0,1]$(Ref. [34]. Prior knowledge about the components of a CH_4 plasma is chosen from experimental experience. Common knowledge is that H_2 and CH_4 are the main neutral constituents and all other species remain below a few percent with declining intensity as the carbon content of a species rises. This allows again the assignment of exponential prior distributions for the concentrations. The probability for a particular set of E species in the model is given in terms of the data $\mathbf{D}$ and variances $\{\mathbf{S}\}$ by Bayes theorem

$$P\left(E|\mathbf{D},\{\mathbf{S}\},I\right) = \frac{P(E|I)\,P\left(\mathbf{D}|\{\mathbf{S}\},E,I\right)}{P\left(\mathbf{D}|\{\mathbf{S}\},I\right)} \tag{33}$$

The marginal likelihood $P\left(\mathbf{D}|\{\mathbf{S}\},I\right)$ is obtained from

$$P\left(\mathbf{D}|\{\mathbf{S},E\},I\right) = \tag{34}$$
$$\int d\mathbf{C}\,d\mathbf{X}\,P(\mathbf{C}|E,I)\,P(\mathbf{X}|E,I)\,P\left(\mathbf{D}|\mathbf{C},\mathbf{X},\{\mathbf{S}\},E,I\right).$$

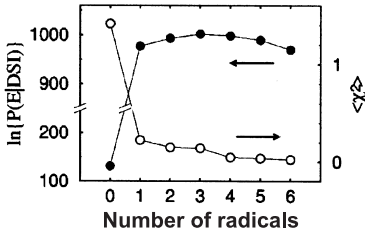

FIG. 6: Natural logarithm of $P(E|\mathbf{D},\mathbf{S},I)$ and the misfit between data and model for combinations of six free radicals (C_2H_5, CH_3,H,C_2H_3,CH,CH_2). x axis shows the number of radicals involved in the model, which were taken in the given order (from [37]).

The dimension of the integral is high and increases with the number of data sets represented by $\mathbf{D}$ and the number of species chosen to model the observations. Such high dimensional integrals (for interpretation of the spectrum shown in Fig. 7 the dimension exceeds 400) can be tackled either by Markov Chain Monte Carlo techniques (using thermodynamic integration for a faster convergence) or approximately by saddle point approximations which may not always exist in the analysis of mass spectra.
A low temperature methane plasma was analyzed with respect to H_2 and C_1H_x and C_2H_y molecules. In particular the identification of the relevant radicals and their concentrations was of interest. As can be seen from Fig.6 the misfit decreases monotonously as more radicals are incorporated into the model. By contrast, the evidence attains a maximum for inclusion of three radicals (C_2H_5, CH_3) and H and decays slowly for more complicated models. This result is rather reasonable since these radicals are produced by breaking just one atomic bond from the stable and abundant molecules H_2, CH_4 and C_2H_6. The next step after the identification of the number of species contributing to the set of measurements is the estimation of the concentrations and the cracking coefficients. The required posterior probability distribution is given by

$$P\left(\mathbf{X},\mathbf{C}|\mathbf{D},\mathbf{S},E,I\right) = \tag{35}$$
$$\frac{P(\mathbf{X}|E,I)\,P(\mathbf{C}|E,I)\,P\left(\mathbf{D}|\mathbf{C},\mathbf{X},\mathbf{S},E,I\right)}{P\left(\mathbf{D}|\mathbf{S},E,I\right)}.$$

Detection and quantification of radicals is one attractive result of the Baysian analysis of a beam shutter (on/off) experiment in the diagnostic of a low temperature plasma. Equally important and equally demanding is the analysis of the neutral gas mass spectra in particular for plasmas with hydrocarbon fuel gases. Fig 7 displays a result from a comprehensive data set from 34 mass channels for 27 different plasma conditions of an inductively coupled pulsed plasma discharge, together with calibration measurements for 11 species. Two models with a different number of hydrocarbon molecules are compared. The modeled data agree extremely well

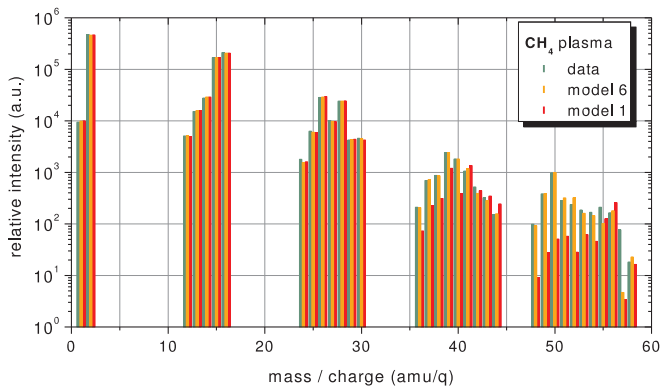

FIG. 7: Comparison of data computed by 2 different models with the measured mass spectrometer values.

with the measurements for masses below 35. For higher masses there is a discrepancy between model 1 and the data whereas model 6 gives a nearly perfect match, indicating the presence of C_4H_2 and C_4H_6 in the plasma. The large number of different species possibly present in the plasma lead to a huge number of different models to be compared. A detailed discussion of the results is therefore beyond the scope of this paper and will be given elsewhere. Nevertheless the algorithmic implementation of the Bayesian method is so efficient that CPU time is no longer a valid argument to digress to less-reliable methods, except for monitoring purposes [38].

B. Discordant Data Sets

Experimental data from different sources may suffer from discordant calibrations and possibly cover different regions of the independent variables. Here we give an example how to treat unknown scale factors of different data sets.

Chemical erosion due to hydrogen ion bombardment is the dominant erosion process for carbon-based plasma facing materials in fusion experiments. In the low flux regime, i.e. $\theta < 10^{19}/\mathrm{m}^2\mathrm{s}$ the mechanism of chemical erosion is reasonably well understood. At high fluxes θ, such as experienced in fusion devices, there was indication from various data that the chemical sputtering yield decreases with ion flux [39]. Weight loss measurements are available for the low flux regime [40]. Those measurements are the most reliable ones, since these data require no further calibration factors. The function for the chemical erosion yield is taken from Ref.[41]. For the weight loss measurements we assume that the erosion yield $\phi(\theta, \theta_0)$ depends on flux through

$$\phi(\theta, \theta_0) = \mathrm{Y}_{\mathrm{chem}} \cdot \frac{1}{1 + \theta/\theta_0}. \tag{36}$$

In contrast, calibration factors are necessary for mass spectroscopy and optical emission spectroscopy. For the high flux data the eroded molecule flux was determined spectroscopically from the CH band intensity. The reduction of the CH band emission to a total erosion yield requires accurate knowledge of the CH optical transition rates. We allow here for an uncertainty of the measured erosion data by introducing an unknown calibration factor γ. However, with erosion data collected in fusion machines the situation may also be different. The optical system used to record hydrogen and CH band emissions may suffer from a calibration error which translates into a common recalibration factor γ for *both* the hydrogen flux θ and the erosion yield. In this case the appropriate model is given by

$$\phi(\theta, \theta_0, \gamma) = \mathrm{Y}_{\mathrm{chem}} \cdot \frac{1}{1 + \gamma\theta/\theta_0} \tag{37}$$

The first term $\mathrm{Y}_{\mathrm{chem}}$ varies very weakly with flux θ and shall be considered constant. In the end we have to distinguish between a data set from weight loss measurements δ considered to be scaled correctly

$$\delta_i = c \cdot \phi(\theta_i, \theta_0) + e_i \tag{38}$$

and the data sets from optical measurements Δ_j with a possible scale factor γ either only for the erosion yield

$$\gamma\Delta_j = c \cdot \phi(\Theta_j, \theta_0,) + E_j \tag{39}$$

or also for the incoming flux

$$\gamma\Delta_j = c \cdot \phi(\Theta_j, \theta_0, \gamma) + E_j. \tag{40}$$

Assuming the expectation value of the error to be zero and the variance given by s_i^2 (S_i^2) the likelihood functions for the two data sets read

$$P\left(\vec{\delta}|\vec{\theta}, \vec{s}, \theta_0, c\right) = \prod_i \frac{1}{s_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{\delta_i - c\phi(\theta_i, \theta_0)}{s_i}\right]^2\right\} \tag{41}$$

$$P\left(\vec{\Delta}|\vec{\Theta}, \vec{S}, \theta_0, \gamma, c\right) = \prod_j \frac{\gamma}{S_j\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{\gamma\Delta_j - c\phi(\Theta_j, \theta_0, (\gamma))}{S_j}\right]^2\right\} \tag{42}$$

Unfortunately, the experimental error estimates for the available data sets are not compatible with the observed scatter of the data (so called outliers are present). Outlier tolerance may be obtained in the following way [42]. Let us assume that the probability density for the true error σ is given by a distribution which allows for large discrepancies between scattered data and specified errors

$$P(\sigma_i|s_i, I) = \frac{2}{\pi}\left(\frac{s_i}{\sigma_i}\right)^2 \exp\left(\frac{s_i^2}{\sigma_i^2}\right) \tag{43}$$

but with mean $<\sigma>$=s of the error estimate. Marginalization of σ yields a modified likelihood (compare with Eq.41)

$$P\left(\vec{\delta}|\vec{\theta},\vec{s},\theta_0,c\right) = \prod_i \frac{1}{s_i 2\pi\sqrt{2}} \left\{\frac{1}{\pi} + \frac{1}{2}\left[\delta_i - c\phi\left(\theta_i,\theta_0\right)\right]\right\}^{-\frac{3}{2}} \quad (44)$$

and similarly for Eq.42. First the expectation value of the scale parameter γ is of interest. It is obtained by

$$\langle\gamma\rangle = \frac{\int d\gamma\, d\theta_0\, \gamma P\left(\gamma,\theta_0|\vec{\delta},\vec{\Delta},\vec{\theta},\vec{\Theta},\vec{s},\vec{S},c,I\right)}{\int d\gamma\, d\theta_0\, P\left(\gamma,\theta_0|\vec{\delta},\vec{\Delta},\vec{\theta},\vec{\Theta},\vec{s},\vec{S},c,I\right)} \quad (45)$$

and can be rewritten using Bayes' theorem

$$P\left(\gamma,\theta_0|\vec{\delta},\vec{\Delta},\vec{\theta},\vec{\Theta},\vec{s},\vec{S},c,I\right) = \frac{P\left(\gamma,\theta_0,c|I\right)}{P\left(\vec{\delta},\vec{\Delta}|\vec{\theta},\vec{\Theta},\vec{s},\vec{S},c,I\right)} P\left(\vec{\delta},\vec{\Delta}|\gamma,\theta_0,\vec{\theta},\vec{\Theta},\vec{s},\vec{S},c,I\right). \quad (46)$$

The last term in Eq.47 is the product of the two likelihoods. Assuming the independence of the two data sets $\vec{\delta}$ and $\vec{\Delta}$

$$P\left(\vec{\delta},\vec{\Delta}|\gamma,\theta_0,\vec{\theta},\vec{\Theta},\vec{s},\vec{S},c,I\right) = P\left(\vec{\delta}|\vec{\theta},\vec{s},\theta_0\right) P\left(\vec{\Delta}|\vec{\Theta},\vec{S},\theta_0,\gamma,c\right). \quad (47)$$

The prior distributions of

$$P\left(\gamma,\theta_0,c|I\right) = P\left(\gamma|I\right) P\left(\theta_0|I\right) P\left(c|I\right) \quad (48)$$

are taken to be flat, except for $P\left(\gamma|I\right)$ where we can assume an expectation value for the scale factor $\langle\gamma\rangle = 1$. Any other choice would imply a deliberately introduced bias in the calibrations used to obtain data set $\vec{\Delta}$. By virtue of the principle of maximum entropy this results in an exponential prior

$$P\left(\gamma|I\right) = \exp\left(-\gamma\right). \quad (49)$$

The Bayes factor for the two models is given by the ratio of the marginalized likelihoods

$$P\left(\vec{\delta},\vec{\Delta}|M_k,\vec{\theta},\vec{\Theta},\vec{s},\vec{S},I\right) = \int d\gamma\, d\theta_0\, dc\, P\left(\vec{\delta},\vec{\Delta}|M_k,\gamma,\theta_0,\vec{\theta},\vec{\Theta},\vec{s},\vec{S},c\right) P\left(\gamma,\theta_0,c|I\right) \quad (50)$$

when no model is preferred a priori. Computing the odds ratio reveals that the model given in Eq.36 for the low flux data $\left(<10^{20}/\mathrm{m^2s}\right)$ and Eq. 39 for the high flux data is to be preferred by a factor of 10 over the combination of Eq.36 and Eq.40 for the data sets shown in Fig. 8. This gives not any reason of deferring from the statement of the experimentalists, that the calibration for the incident hydrogen flux is quite reliable and that the correction factor should be applied to the eroded atom flux only, rather than to the incident hydrogen flux and eroded atom flux [43, 44]. Therefore the results depicted in Fig.8 refer to the first model. The mean of the threshold value is $\theta_0 = 28.8 \cdot 10^{-23}\mathrm{m^2s}$ with scale factors of $\gamma = 0.72$ and $\gamma = 0.32$ for the data from Ref.[45] and [46], respectively.

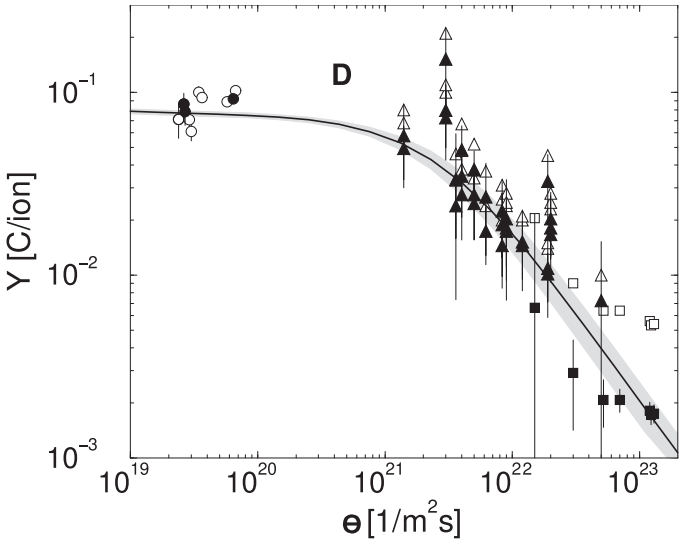

FIG. 8: Flux dependence of chemical erosion yield of graphite under hydrogen irradiation. The data set δ is represented by circles. Filled circles correspond to the subset for which the fitting curve (solid line) is valid (E_0 = 30eV,T=600K).Open triangles and squares represent data from Ref.[45] and [46], respectively, while the full symbols show the data sets after multiplication with the corresponding scale factor (0.72 and 0.32). Error bars show the assigned experimental error. The gray shaded area is the confidence range (from [43]).

V. APPLICATIONS III: MIXTURE MODELLING

Mixture modelling is an ideal tool to solve the ubiquitous problem of background and source separation. Examples are PIXE measurements [47] and Auger data [48] but also x-ray images in high-energy astrophysics[49]. The basic idea is simple. The background is relatively slowly varying compared to the signal. Therefore the background is represented by a smooth function. Data points that are significantly separated from the background are considered as outliers, as data points containing background and signal contributions. Given an observed data set $\vec{d} = \{d_i\}$ we can formulate two complementary hypotheses

$$B_i : d_i = b_i + \epsilon_i \quad (51)$$

$$\overline{B}_i : d_i = b_i + s_i + \epsilon_i \quad (52)$$

Hypothesis B_i specifies that d_i consists only of background b_i and noise ϵ_i and hypothesis $\overline{B}_i$ that an additional source contribution is present. For counting experiments (and only positive signal contributions) the likelihood for the two distributions is given by the Poisson distribution:

$$P\left(d_i|B_i,b_i\right)=\frac{b_i^{d_i}}{d_i!}\exp\left(-b_i\right), \tag{53}$$

$$P\left(d_i|B_i,b_i\right)=\frac{\left(b_i+s_i\right)^{d_i}}{d_i!}\exp\left(-\left(b_i+s_i\right)\right), \tag{54}$$

Since we do not know the signal intensities we marginalize over all possibilities. The average signal intensity s_0 of the data set can be used as a reasonable expectation value of the prior distribution of the signal [50]

$$P\left(s_i|s_0\right)=\frac{\exp\left(-s_i/s_0\right)}{s_0}. \tag{55}$$

Then the marginal Poisson likelihood for the hypothesis $\overline{B}_i$ is given by

$$P\left(d_i|\overline{B}_i,b_i,s_0\right)= \tag{56}$$
$$\frac{\exp\left(\frac{b_i}{s_0}\right)}{s_0\left(1+\frac{1}{s_0}\right)^{(d_i+1)}}\frac{\Gamma\left[\left(d_i+1\right),b_i\left(1+\frac{1}{s_0}\right)\right]}{\Gamma\left(d_i+1\right)}$$

The two different likelihoods for the propositions B_i and $\overline{B}_i$ are combined in the likelihood for the mixture model

$$P\left(\vec{d}|b,s_0,\beta\right)= \tag{57}$$
$$\prod_i\left[\beta P\left(d_i|B_i,b_i\right)+\left(1-\beta\right)P\left(d_i|\overline{B}_i,b_i,s_0\right)\right]$$

where β is the probability that a data point contains no signal contribution. $\beta=0.5$ is a noncommittal but unrealistic value, stating that each datum is equally likely to contain signal contribution or not. So far we have not specified which basic functions are suited for the background model. An obvious choice in one dimension is to use cubic splines. In [48] the background is represented by a cubic spline together with a smoothness prior for the background

$$P\left(b|\mu,I\right)=\frac{1}{Z}\exp\left(-\mu\int\left|b''\right|^2\,dx\right). \tag{58}$$

and applied to an Auger spectrum obtained with a four-grid low-energy electron-diffraction (LEED) optics in the retarding field mode. Such spectra constitute the superposition of the energy derivative of the sum of the Auger electron energy distribution, the signal, and the much larger secondary electron energy distribution, the background. The latter is known to be rapidly varying in the low-energy region, as seen in Fig.9a. The peaks at 47 eV come from an $M_{2,3}VV$ Auger transition. While the background may be smooth, it varies quite rapidly at low energies. The variation of the data can be reduced by a logarithmic transformation of the signal $y'=\log\left(a-y\right)$. The estimated background is given in Fig.9b as solid line together with the transformed data. After plotting the difference between original spectrum and its estimated background shown in Fig.9c, a possible secondary peak is observed at (47+39) eV which is further substantiated by the autocorrelation of the background subtracted spectrum. The peak at 86 eV with an amplitude of about 2% of the main signal corresponds to the M_1VV Auger transition for iron. In this case a proper background subtraction reveals the presence of less apparent signals in the Auger spectrum.

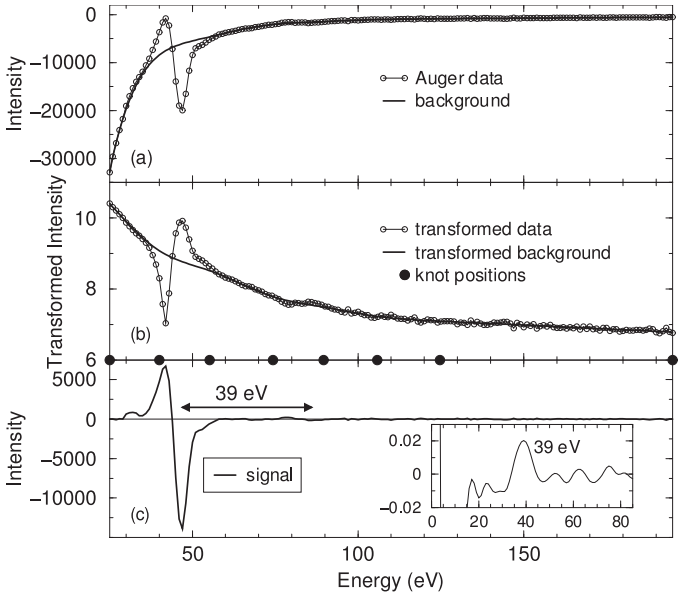

FIG. 9: An MVV Auger spectrum for iron. The estimated background shown is obtained for the transformed spectrum shown in (b). A logarithmic transformation of the Auger spectrum reduces the curvature of the background. The estimated background is shown as solid line. The eight support points of the spline are indicated by filled circles. (c) The signal obtained by subtracting the data and the background. A secondary peak is present at an energy of 86 eV, 39 eV above the $M_{2,3}VV$ Auger transition, substantiated by the autocorrelation of the signal vs. energy difference (see inset)(from [48]).

VI. CONCLUSION AND OUTLOOK

We have demonstrated that Bayesian Probability Theory is a powerful tool for inference from physical data. It allows the extraction of the most convincing conclusions implied by given data and any prior knowledge of the circumstances in a systematic way. This has been noticed in observational branches as in biometrics and astronomy where the data sets cannot be augmented at will

and have to be exploited as far as possible. Fortunately also in other branches of physics the situation expressed in [51]: *We use fantastic telescopes, best physical models and the best computers. The weak link in this chain is interpreting our data using 100-year-old-mathematics* is slowly improving.

But Bayesian Probability theory is not a magic black box guaranteed to compensate for badly designed experiments. Information absent in the data can not be revealed by any kind of data analysis. But how to find an optimal experimental setup? Here we enter one of the active areas of research which are beyond the scope of the article but should not go unmentioned: *Bayesian experimental design* is an increasingly important topic which is feasible due to the advent of modern computers [52]. Another area attracting more and more interest is *integrated data analysis*, the combined evaluation of information of different sources (eg. the variaty of diagnostics of a fusion experiment [53]) on a much larger scale than today. This is at the very heart of the Bayesian Probability theory because not only all kinds of different prior information can easily be incorporated but also the implicit interdependencies of the diagnostics are exploited resulting in superior (uncertainty) estimates.

[1] D.S. Sivia. *Data Analysis - A Baysian Tutorial*. Clarendon, Oxford, 1996.

[2] T. Leonard and J.S.J. Hsu. *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*. Cambridge University Press, Cambridge, 1999.

[3] H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, 1961.

[4] R. T. Cox. Probability, frequency and reasonable expectation. *Am. J. Phys.*, 14:1–13, 1946.

[5] D.S. Sivia and W.I.F. David. A bayesian approach to extracting structure-factor amplitudes from powder diffraction data. *Acta Cryst. A*, 50:703–714, 1994.

[6] S. Russell and P. Norvig. *Artifical Intelligence: A Modern Approach*. Prentice Hall, New Jersey, 2003.

[7] P.C. Gregory. Bayesian periodic signal detection. *ApJ*, 520:361–375, 1999.

[8] V. Dose. Bayesian inference in physics: case studies. *Rep. Prog. Phys.*, 66:1421–1461, 2003.

[9] F.H. Fröhner. *Evaluation and Analysis of Nuclear Resonance Data, JEFF Report 18*. OECD Nuclear Energy agency, Paris, 2000.

[10] E.T. Jaynes. Prior probabilities. *IEEE Trans. Syst. Sci. Cybernetics*, SSC-4:227–241, (reprinted in Jaynes 1983), 1968.

[11] M.G. Kendall and P.A.P. Moran. *Geometrical Probability*. Griffin, London, 1963.

[12] V. Dose. Hyperplane priors. In C. J. Williams, editor, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume Conf. Proc. 659, pages 350–357. AIP, Melville, NY, 2003.

[13] U. v. Toussaint, S. Gori, and V. Dose. Bayesian neural-networks-based evaluation of binary speckle data. *Applied Optics*, 43:5356–5363, 2004.

[14] V. Dose. Bayes in five days. *CIPS Reprint Series*, 83:1–44, 2002. http://www.ipp.mpg.de/OP/Datenanalyse/Publications.

[15] E.T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, 1957.

[16] E.T. Jaynes. Information theory and statistical mechanics II. *Phys. Rev.*, 108:171–190, 1957.

[17] J.N. Kapur and H.K. Kesavan. *Entropy Optimization Principles with Applications*. Academic, Boston, 1992.

[18] J. Tesmer and M. Nastasi, editors. *Handbook of Modern Ion Beam Analysis*. Material Research Society, Pittsburgh, Pennsylvania, 1995.

[19] W. von der Linden. Maximum-entropy data analysis. *Appl. Phys. A*, 60:155, 1995.

[20] R. Fischer, M. Mayer, W. von der Linden, and V. Dose. Enhancement of the energy resolution in ion-beam experiments with the maximum-entropy method. *Phys. Rev. E*, 55:6667, 1997.

[21] J. Skilling. Maximum entropy in action. In B. Buck and V. Macauley, editors, *Fundamentals of maxent in data analysis*, page 19. Clarendon Press, Oxford, 1991.

[22] R. Fischer, M. Mayer, W. von der Linden, and V. Dose. Energy resolution enhancement in ion beam experiments with Bayesian probability theory. *Nucl. Inst. Meth. B*, 136-138:1140, 1998.

[23] V. Dose, R. Fischer, and W. von der Linden. Deconvolution based on experimentally determined apparatus functions. In G. Erickson, editor, *Maximum Entropy and Bayesian Methods*, pages 147–152. Kluwer Academic, Dordrecht, 1998.

[24] M. Mayer. SIMNRA, a simulation program for the analysis of NRA, RBS and ERDA. In J.L. Duggan and I. Morgan, editors, *Proceedings of the 15th International Conference on the Application of accelerators in Research and Industry*, volume Conf. Proc. 475, page 541. AIP, Melville, NY, 1999.

[25] M. Mayer, R. Fischer, S. Lindig, U. v. Toussaint, R. W. Stark, and V. Dose. Bayesian reconstruction of surface roughness and depth profiles. *Nucl. Inst. Meth. B*, 228:349–359, 2005.

[26] R. Behrisch. Modifications of solids due to the exposure to high temperature plasmas. *Phys. Res.*, 8:569–573, 1988.

[27] J.N. Brooks, D. Alman, G. Federici, D.N. Ruzic, and D.G. White. Erosion/redeposition analysis: status of modeling and code validation for semi-detached edge plasmas. *J. Nuc. Mater.*, 266-269:58, 1999.

[28] K. Krieger, U. v. Toussaint, and the ASDEX-Upgrade team. Direct measurement of carbon erosion rates in the divertor of ASDEX Upgrade. In *Proc. of the 26th EPS Conference on Controlled Fusion and Plasma Physics*, volume ECA 23J, pages 1529–1532. Europ. Phys. Soc., Maastricht, 1999.

[29] U. v. Toussaint, K. Krieger, R. Fischer, and V. Dose. Depth profile reconstruction from Rutherford backscattering data. In W. von der Linden, V. Dose, R. Fischer, and R. Preuss, editors, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, Dordrecht, 1999.

[30] U. v. Toussaint, R. Fischer, K. Krieger, and V. Dose.

Depth profile determination with confidence intervals from Rutherford backscattering data. *New J. Phys.*, 1:11, 1999.

[31] P.C. Greogory and T.J. Loredo. A new method for the detection of a periodic signal of unknown shape and period. *Astro. J.*, 398:146–168, 1992.

[32] A. von Keudell. Formation of polymer-like hydrocarbon films from radical beams of methyl and atomic hydrogen. *Thin Solid Films*, 402:1–37, 2002.

[33] R. Dobrozemsky and G. Schwarzinger. Mass spectroscopy of fusion-plasma gases. *J. Vac. Sci. Technology A*, 10(4):2661–2664, 1992.

[34] T. Schwarz-Selinger, R. Preuss, V. Dose, and W. von der Linden. Analysis of multicomponent mass spectra applying Bayesian probability theory. *J. Mass Spect.*, 36:866–874, 2001.

[35] R. Preuss, H. Kang, T. Schwarz-Selinger, and V. Dose. Quantitative analysis of multicomponent mass spectra. In R. L. Fry, editor, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume Conf. Proc. 617, pages 155–162. AIP, Melville, NY, 2002.

[36] A. Cornu and R. Massot. *Compilation of Mass Spectral Data*. Heyden, London, 1979.

[37] H. Kang and V. Dose. Radical detection in a methane plasma. *J. Vac. Sci. Technol. A*, 21(6):1978–1980, 2003.

[38] U. v. Toussaint, V. Dose, and A. Golan. Maximum entropy decomposition of quadrupole mass spectra. *J. Vac. Sci. Technol. A*, 22(2):401–406, 2004.

[39] J. Roth and C. Garcia-Rosales. Analytical description of the chemical erosion of graphite by hydrogen ions. *Nucl. Fus.*, 36:1647–1659 with corrigendum Nucl. Fus. 37 (1997) 897, 1996.

[40] M. Balden and J. Roth. New weight loss measurements of the chemical erosion yield of carbon materials under hydrogen ion bombardement. *J. Nucl. Mater.*, 280:39–44, 2000.

[41] J. Roth. Chemical erosion of carbon based materials in fusion devices. *J. Nucl. Mater.*, 266-269:51–57, 1999.

[42] V. Dose and W. von der Linden. Outlier tolerant parameter estimation. In W. von der Linden, V. Dose, R. Fischer, and R. Preuss, editors, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, Dordrecht, 1999.

[43] R. Preuss, P. Pecher, and V. Dose. Handling discordant data sets. In C. R. Smith J. Rychert, G. Erickson, editor, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume Conf. Proc. 567, pages 213–220. AIP, Melville, NY, 2001.

[44] V. Dose, R. Preuss, and J. Roth. Evaluation of chemical erosion data for carbon materials at high ion fluxes using Bayesian probability theory. *Journal of Nuclear Materials*, 288:153–162, 2001.

[45] H. Grote and W. Bohmeyer et al. Chemical sputtering yields of carbon based materials at high flux densities. *J. Nucl. Mater.*, 266-269:1059–1064, 1999.

[46] G.R. Tynan. APS-meeting. *New Orleans*, 1998.

[47] J. Padayachee, V. M. Prozesky, W. von der Linden, M. S. Nkwinika, and V. Dose. Bayesian PIXE background subtraction. *Nucl. Inst. Meth. B*, 150:129–135, 1999.

[48] R. Fischer, K. M. Hanson, V. Dose, and W. von der Linden. Background estimation in experimental spectra. *Phys. Rev. E*, 61:1152, 2000.

[49] F. Guglielmetti, R. Fischer, V. Dose, W. Voges, and G. Boese. Source detection with Bayesian inference on ROSAT all-sky survey data sample. In Mark G. Allen Francois Ochsenbein and Daniel Egret, editors, *ASP Conference Series Volumes, Astronomical Data Analysis Software and Systems (ADASS) XIII*, volume CS 314, page O03.3, Strasbourg, 2004.

[50] W. von der Linden, V. Dose, J. Padayachee, and V. Prozesky. Signal and background separation. *Phys. Rev. E*, 59:6527–6534, 1999.

[51] D. Mackenzie. Vital statistics. *New Scientist*, 2453:36–41, 2004.

[52] T.J. Loredo and D.F. Chernoff. Bayesian adaptive exploration. In E.D. Feigelson and G.J. Babu, editors, *Statistical Chalenges in Astronomy*, pages 57–70. Springer, Berlin, 2003.

[53] R. Fischer, A. Dinklage, and E. Pasch. Bayesian modelling of fusion diagnostics. *Plasma Phys. Control. Fusion*, 45:1095–1111, 2003.