Bayesian data analysis tools for atomic physics

Martino Trassinelli

Bayesian data analysis tools for atomic physics

Martino Trassinelli

Institut des NanoSciences de Paris, CNRS, Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France

Abstract

We present an introduction to some concepts of Bayesian data analysis in the context of atomic physics. Starting from basic rules of probability, we present the Bayes' theorem and its applications. In particular we discuss about how to calculate simple and joint probability distributions and the Bayesian evidence, a model dependent quantity that allows to assign probabilities to different hypotheses from the analysis of a same data set. To give some practical examples, these methods are applied to two concrete cases. In the first example, the presence or not of a satellite line in an atomic spectrum is investigated. In the second example, we determine the most probable model among a set of possible profiles from the analysis of a statistically poor spectrum. We show also how to calculate the probability distribution of the main spectral component without having to determine uniquely the spectrum modeling. For these two studies, we implement the program `Nested_fit` to calculate the different probability distributions and other related quantities. `Nested_fit` is a Fortran90/Python code developed during the last years for analysis of atomic spectra. As indicated by the name, it is based on the nested algorithm, which is presented in details together with the program itself.

Keywords: Bayesian data analysis, atomic physics, nested sampling, model testing

1. Introduction

Commonly, a data analysis is based on the comparison between a function $F(\boldsymbol{a})$ used to model the data that depends on a set of parameters $\boldsymbol{a}$ (ex. $a_1 \rightarrow$ amplitude, $a_2 \rightarrow$ position, etc.) and the data them-self that consist in recorded number of counts y_i at each channel x_i. The estimation of the parameter values describing at best the data is generally obtained by the maximum likelihood method (and its lemma, the method of the least squares), which consists to find the values $\boldsymbol{a}^{best}$ that maximize the product of the probabilities for each channel x_i to observe y_i counts for a given expected value $F(x_i, \boldsymbol{a}^{best})$.

Even if very successfully in many cases, this method has some limitations. If some function parameter is subject to constraints on its values (as ex. one model parameter could be a mass of a particle, which cannot be negative), the corresponding boundary conditions cannot be taken into account in a well defined manner. With the likelihood function we are in fact calculating the probabilities to observe the data $\{x_i, y_i\}$ for given parameter values and not the probability to have certain parameter values for given experimental data.

An additional difficulty for the maximum likelihood method is the impossibility to assign probabilities to different hypothesis, represented for example by two possible modeling functions F_A and F_B, in view of the acquired data. In simple cases, the determination of the best models can be obtained with the maximum likelihood ratio (or chi-square ratio) or other criteria (Akaike information criterion [1], deviance information criterion [2], etc.), but no quantified probability can be assigned to each model.

Another important and fundamental problem of the common data analysis approach is the requirement of repeatability for the definition of *probability* itself. In classic data analysis manuals we can find sentences as:

> *"Suppose we toss a coin in the air and let it land. There is 50% probability that it will land heads up and a 50% probability that it will land tails up. By this we mean that if we continue tossing a coin repeatedly, the fraction of times that it lands with heads up will asymptotically approach 1/2 ..."* [3]

This definition is completely inadequate to rare processes as those encountered for example in cosmology, where several models are considered to describe one unique observation, our universe, and more recently in gravitational-wave astronomy, where at present only two observations are available [4, 5].

To overcome these problems, a different approach has to be implemented with a new and more general definition of probability, This approach is the result of the work of Th. Bayes, P.-S. Laplace, H. Jeffreys and of many others [6, 7, 8, 9] and is commonly called *Bayesian statistics*.

Bayesian methods are routinely used in many field: cosmology [10, 11, 12], particle physics, nuclear physics, particle physics, ... In atomic physics the implementation of these data analysis approach is not yet common. Nevertheless, in some cases their use is required. For example, when we want to determine the correct shape of a instrumental response function we are actually testing hypotheses, as in the case of the determination of the presence or not of possible line contributions in a complex or statistically poor spectrum.

The goal of this article is to present a basic introduction of Bayesian data analysis methods in the context of atomic physics

Email address: `trassinelli@insp.jussieu.fr` (Martino Trassinelli)

spectroscopy and to introduce the program `Nested_fit` for the calculation of distribution probabilities and related quantities from the application of these methods. The introduction to Bayesian statics is based in the extended literature on this domain, and in particular on Refs. 11, 8, 13, 14, 15. For a clear and practical presentation, we will present two simple applications of data analysis where we implemented a Bayesian approach using the `Nested_fit` program. The first example is about the probability evaluation of the presence of a satellite peak in a simple atomic spectrum. The second one deals with the analysis of a statistically poor spectrum in which one or multiple peaks contributions has to be considered and where possible aberrations in the response function have also to be taken into account. We will in particular show how to assign probabilities to the different models from the experimental data analysis. Moreover, we will see how to extract the probability distribution of the main peak position without the need to uniquely choose between the different models.

The article is organized as following. A general definition of probability and Bayesian statistic concepts as the Bayes' theoreme and *Bayesian evidence* are present in Sec. 2, together with a very general and axiomatic definition of probability deduced from simple logic arguments. In Section 3 we present in details the nested algorithm for the calculation of the Bayesian evidence and in Sec. 4 we will see its implementation in the program `Nested_fit`, which is also presented. These two sections are quite technical and they could be skipped in a first reading. Section 5 is dedicated to the Bayesian data analysis applications to two real data sets and Sec. 6 is our conclusion. Two appendixes are also proposed: one about the introduction of information and complexity concepts in the context of Bayesian statistics, and a second about the evaluation of the uncertainty of the Bayesian evidence calculated by the nested sampling method.

2. Probability

2.1. Probability axioms

A very general definition of probability $P(X, I)$ can be obtained by trying to assign real numbers to a certain degree of plausibility or believe than assertions X, Y, etc., would be true. X and Y assertions are very general. They can be assertions of specific statements (ex. "In the next toss the coin will land heads") or implying values (ex. the parameter b is in a certain range $[b_{min}, b_{max}]$). When basic logic and consistency are required, the form of the probability P is ensured by the axioms [13, 16, 9, 15, 14]

$$0 \leq P(X|I) \leq 1, \tag{1}$$

$$P(X|X, I) = 1, \tag{2}$$

$$P(X|I) + P(\bar{X}|I) = 1, \tag{3}$$

$$P(X, Y|I) = P(X|Y, I) \times P(Y|I). \tag{4}$$

In the equations above, $\bar{X}$ indicates the negation of the assertion X (not-X); the vertical bar "|" means "given" and where I represents the current state of knowledge. The joint probability $P(X, Y|I)$ means that both "X AND Y" are true (equivalent to the logical conjunction '$\wedge$'). The deduction of these axioms have been obtained for the first time in 1946 by Richard Cox using Boolean logic [13]. The first three axioms are compatible with the usual probability rules. Here we have an additional axiom that, as we will see, plays a very important role.

From these axioms the following rule (sum rule) is deduced [14]

$$P(X + Y|I) = P(X|I) + P(Y|I) - P(X, Y|I). \tag{5}$$

Here the symbol '+' means here the logical disjunction ($X+Y \equiv X \vee Y \equiv$ "X OR Y is true").

The fourth axiom determines the rule for inference probabilities (product rule) for conditional cases. If X and Y are independent assertions, this is reduced to the classical probability property

$$P(X, Y|I) = P(X|I) \times P(Y|I). \tag{6}$$

When a set of mutual exclusive assertions are considered $\{Y_i\}$, with $P(Y_i|Y_{j\neq i}) = 0$, we have the so-called *marginalization rule*

$$P(X|I) = \sum_i P(X, Y_i|I) \tag{7}$$

that in the limit of continuous case $Y_{i+1} - Y_i \rightarrow dY$ becomes

$$P(X|I) = \int_{-\infty}^{\infty} P(X, Y|I) dY. \tag{8}$$

2.2. Bayes' theorem and posterior probability

Another important corollary can be derived from the fourth axiom (Eq. (4)) and the similar expression with exchange between X and Y:

$$P(X|Y, I) = \frac{P(Y|X, I) \times P(X|I)}{P(Y|I)}. \tag{9}$$

This is what is called the Bayes' Theorem, named after Rev. Thomas Bayes, who first [6] formulated theorems of conditional probability and rediscover in 1774 and further developed by Pierre-Simon Marquis de la Laplace [7].

For a better insight in the implication of this theorem, we consider the case where X represent the hypothesis that the parameter values set $\boldsymbol{a}$ truly describes the data (via the function $F(x, \boldsymbol{a})$) and where Y correspond to the recorded data $\{x_i, y_i\}$. In this case Eq. (9) becomes

$$P(\boldsymbol{a}|\{x_i, y_i\}, I) = \frac{P(\{x_i, y_i\}|\boldsymbol{a}, I) \times P(\boldsymbol{a}|I)}{P(\{x_i, y_i\}|I)} = \frac{L(\boldsymbol{a}) \times P(\boldsymbol{a}|I)}{P(\{x_i, y_i\}|I)}, \tag{10}$$

where I includes our available background information and where $P(\{x_i, y_i\}|\boldsymbol{a}, I)$ is by definition the likelihood function $L(\boldsymbol{a})$ for the given set of data. Differently from the common statistical approach where only the likelihood function is considered, we have here the additional term $P(\boldsymbol{a}|I)$ that includes the prior knowledge on the parameters $\boldsymbol{a}$ or possible boundaries. The denominator term $P(\{x_i, y_i\}|I)$ can be considered for the moment as a normalization factor but it plays an important role when different hypothesis are considered and compared (see next section).

The priors can look as an unsuitable input due to the possible subjectivity in their choice; this is actually the main critics to *Bayesian statistics*. On the contrary, the priors reflects our knowledge or ignorance in a quantify way. If two scientists have different choices of priors, and uses some common experimental data, the posterior probability distributions are generally not significantly different. If the posteriors are different because of the different choice of priors, this means that the data are not sufficient to analyze the problem.

From $P(\boldsymbol{a}|\{x_i, y_i\}, I)$, the probability distribution of each parameter $P(a_j|\{x_i, y_i\}, I)$ or joint probabilities $P(a_j, a_k|\{x_i, y_i\}, I)$ can be obtained from the marginalization (Eq. (8)), i.e. the integration of the posterior probability on the unconcerned parameters.

2.3. Model testing and Bayesian evidence

An important consequence of the Bayes' theorem is to have the possibility to assign probabilities to different hypothesis (models) with a simple and well-defined procedure. In this case, in Eq. (9) X represent the hypothesis that the model $\mathcal{M}$ describes well the observations and Y represent the data, as in the previous section. From Bayes' theorem we have that the posterior probability of the model $\mathcal{M}$ is [9, 15, 11]

$$P(\mathcal{M}|\{x_i, y_i\}, I) \propto P(\{x_i, y_i\}|\mathcal{M}, I) \times P(\mathcal{M}|I), \qquad (11)$$

where the first term of the right part is the so-called *Bayesian evidence* E of the model and the second term is the prior probability assigned to the model from our background knowledge. Using the marginalization rule to the parameter values and the probability properties (Eqs. (1–4)), we have

$$\begin{aligned} E \equiv P(\{x_i, y_i\}|\mathcal{M}, I) = \\ = \int P(\{x_i, y_i\}|\boldsymbol{a}, \mathcal{M}, I) P(\boldsymbol{a}|\mathcal{M}, I) d^J\boldsymbol{a} = \\ = \int L^{\mathcal{M}}(\boldsymbol{a}) P(\boldsymbol{a}|\mathcal{M}, I) d^J\boldsymbol{a}, \end{aligned} \qquad (12)$$

where J is the number of the parameters of the considered model, and where we explicitly show the likelihood function $L^{\mathcal{M}}(\boldsymbol{a})$ relative to the model $\mathcal{M}$. The Bayesian evidence, also called *marginal likelihood* or *model likelihood*, is the the integral of the likelihood function over the J-dimensional parameter space under the priors for a specific model choice. The evidence is also the denominator of Eq. (9), which now assumes a more clearly signification than a simple normalization factor (with $\mathcal{M}$ included in I). Considering equal priors, the probability of a model is higher if the evidence is higher, which means that the average of the likelihood function over the model parameter space is higher. To note, this does not implies that the maximum of the likelihood function is larger, as in the case of the likelihood ratio test used to compare the goodness of fit of two models (where however we do not assign probabilities to the models themselves but where we define only a criterion to choose between two models). Models with higher number of parameters are generally penalized because of the higher dimensionality of the integral that corresponds to a larger parameter volume $V_{\boldsymbol{a}}$ (and then to a lower average value of the likelihood function). In fact, the calculation of the model probability via the Bayesian evidence includes, in some sense, the Ockham's razor[1] favoring simpler models when the values of the likelihood function are similar.

If we have to choose among only two different models $\mathcal{M}_1, \mathcal{M}_2$, the comparison between model probabilities is related to the calculation of the simple ratio

$$\frac{P(\mathcal{M}_1|\{x_i, y_i\}, I)}{P(\mathcal{M}_2|\{x_i, y_i\}, I)} = \frac{P(\{x_i, y_i\}|\mathcal{M}_1, I)}{P(\{x_i, y_i\}|\mathcal{M}_2, I)} \times \frac{P(\mathcal{M}_1|I)}{P(\mathcal{M}_2|I)}. \qquad (13)$$

If the prior probabilities of the models are equal, this probability ratio is given by the *Bayes factor* $B_{12} = E_1/E_2$ that is nothing else than the ratio of the evidences [8, 15, 11]. Values of B_{12} larger or smaller than one indicate a propensity for $\mathcal{M}_1$ or $\mathcal{M}_2$, respectively. In the literature several tables are available to assign, in addition to probabilities, degree of propensity of favor to one or other model [8, 17] with a correspondence to the p-value and the standard deviation [18].

For models with similar values of evidence, another criterium to decide between them is the Bayesian complexity C, which measures the number of model parameters that the data can support [11]. This quantity is related to the gain of information (in the Shannon sense) and it is discussed in Appendix A. When E are similar, we should favour the simplest model, i.e. the model with the smallest value of C.

The possibility to assigning probabilities to models has another important advantage. In the case we are interested to determine the probability distribution of a common parameter a_j without the need to identify the correct model among the available choices $\mathcal{M}_\ell$, we can obtain the probability distribution $P(a_j|\{x_i, y_i\}, I)$ from the weighted sum

$$P(a_j|\{x_i, y_i\}, I) = \sum_\ell P(a_j|\{x_i, y_i\}, \mathcal{M}_\ell, I) \times P(\mathcal{M}_\ell|, I), \qquad (14)$$

where $P(a_j|\{x_i, y_i\}, \mathcal{M}_\ell, I)$ are the probability distributions of a_j for each model and $P(\mathcal{M}_\ell|, I)$ are the probabilities of the different models. As we will see in Sec. 5.2, this capability plays an important role in the case where the models have comparable probabilities.

3. The nested sampling algorithm

3.1. The evidence calculation problem

The major difficulty to calculate hypothesis probabilities is the substantial computational power required for the evaluation of the Bayesian evidence. Contrary to the maximum likelihood method, where only the maximum of a function has to be

[1] *"Non sunt multiplicanda entia sine necessitate"* , "Entities must not be multiplied beyond necessity" from William of Ockham's (1287-1347), which can be interpreted in a more modern form as "Among competing hypotheses, the one with the fewest assumptions should be selected".

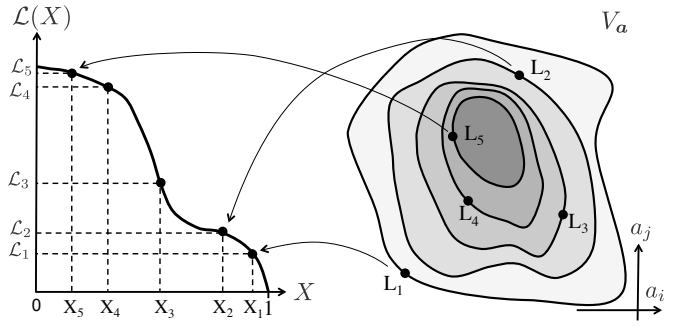

Figure 1: Visualisation of the integral of $\mathcal{L}(X)$ and corresponding volumes on the parameter space (two parameters only, a_i, a_j are considered with a 2D representation).

found, we have to calculate an integral over the J-dimensional space of parameters. Except in very few cases, there is not analytical solution of Eq. (12). The numerical integration by quadrature is not efficient due to the span of different order of magnitude of the likelihood function and the high dimensionality of the problem. The calculation of the evidence is generally done via the Monte Carlo sampling of the product $P(\{x_i, y_i\}|\boldsymbol{a}, \mathcal{M}, I)P(\boldsymbol{a}|\mathcal{M}, I)$.

A common approach to produce good sampling is the use of the Markov chain Monte Carlo (MCMC) technique. A Markov chain is a sequence of random variables such that the probability of the n^{th} element in the chain only depends on the value of the $(n-1)^{\text{th}}$ element. The purpose of the Markov chain is to construct a sequence of points $\boldsymbol{a}_n$ in the parameter space whose density is proportional to the posterior probability distribution. Different probabilistic algorithms are applied to build these chains like Metropolis-Hasting algorithm, Gibbs sampling, Hamiltonian Monte Carlo, etc. (see as example Ref. [19] and references their-in). Another method is the nested sampling where a subdivision of nested volumes in the parameter space is used to calculate the multi-dimensional integral. On this method is based the program `Nested_fit` we present in this article.

The *nested sampling* algorithm is based on the subdivision of the parameters space volume $V(\boldsymbol{a})$, delimited by the parameters prior probabilities, into J-dimensional nested volumes that get closer and closer to the maxima of the likelihood function. With this method, the calculation of the evidence (Eq. 12) is reduced to one-dimensional integral from the original J-dimensional problem. This method has been originally developed by John Skilling in 2004 [20, 15, 21].

To reduce to an one-dimensional integral, we define the variable X (real and positive) corresponding to the volume of the parameter space, weighted by the priors, for which the likelihood function is larger than a certain value $\mathcal{L}$:

$$X(\mathcal{L}) = \int_{L(\boldsymbol{a})>\mathcal{L}} P(\boldsymbol{a}|I)d^J\boldsymbol{a}, \tag{15}$$

where I is the background information we have. A schematic visualisation of this relation is presented in Fig. 1. $X(\mathcal{L})$ is by construction monotonic and invertible, with $\mathcal{L} = \mathcal{L}(X)$. When $\mathcal{L} = 0$, the whole parameter volume $V_{\boldsymbol{a}}$ is considered and then $X = 1$ because of the prior probability normalization. When $\mathcal{L} \geq \max[L(\boldsymbol{a})]$, X is equal to zero. The infinitesimal volume dX is

$$dX = P(\boldsymbol{a}|I)d^J\boldsymbol{a}, \tag{16}$$

where $P(\boldsymbol{a}|I)d^J\boldsymbol{a}$ corresponds to the infinitesimal weighted volume of the parameter space where $\mathcal{L}(X) < L(\{x_i, y_i\}, \boldsymbol{a}) < \mathcal{L}(X + dX)$.

With the above definitions, we can then rewrite Eq. (12) as a simpler one-dimensional integral in X:

$$E = \int_0^1 \mathcal{L}(X)dX. \tag{17}$$

3.2. The algorithm for the numerical integration

The one-dimensional integral in the above equation and represented on the left part of Fig. 1 can be numerically calculated using the rectangle integration method subdividing the [0, 1] interval in $M + 1$ segments with an ensemble $\{X_m\}$ of M ordered points $0 < X_M < ... < X_2 < X_1 < X_0 = 1$. Equation (17) is approximated by the sum

$$E \approx \sum_m \mathcal{L}_m \Delta X_m, \tag{18}$$

where $\mathcal{L}_m = \mathcal{L}(X_m)$ and $\Delta X_m = X_m - X_{m+1}$. The difficulty is now the determination of $\mathcal{L}_m$ and ΔX_m because we do not know a priori the relation between X and $\mathcal{L}$.

The evaluation of $\mathcal{L}_m$ values is obtained by the exploration of the likelihood function via a Monte Carlo sampling. For this, we use a collection of K parameter values $\{\boldsymbol{a}_k\}$ that we call *live points*. At the beginning, these values are randomly chosen from the prior probability distribution $P(\boldsymbol{a}_k|I)$ and they evolve during the computation steps described in the following paragraphs.

To clearly present the different stages of the algorithm, we consider a real analysis of a very simple case. We assume a Gaussian peak plus a flat background (four parameters in total) as model and a very statistical poor data set. The data refer to a high-resolution X-ray spectrum of the helium-like uranium $1s2p\ ^3P_2 \rightarrow 1s2s\ ^3S_1$ intrashell transition and is obtained from Bragg diffraction from a curved crystal [22]. The data and the best guess (maximum likelihood) of Gaussian peak profile are shown in Fig. 2.

For each computation step m of the algorithm we indicate with $\{\boldsymbol{a}_{m,k}\}$ the live points of the step, with $k = 1, \ldots, K$. The corresponding likelihood function values are indicated by $\mathcal{L}_{m,k} = L(\boldsymbol{a}_{m,k})$ and we define $\mathcal{L}_m = \min(\mathcal{L}_{m,k})$. The related X values are indicated by $\xi_{m,k} = X(\mathcal{L}_{m,k})$ and we define $X_m = \max(\xi_{m,k})$. Considering Eq. (15) and Fig. 1, we see that X_m is equal to the integral of the volume where all $\{\boldsymbol{a}_{m,k}\}$ are contained. In other words, the volume $V_{L\geq\mathcal{L}_m}$ in the parameter space corresponds to the segment $[0, X_m]$ in the X axis. Let us see the different steps in details.

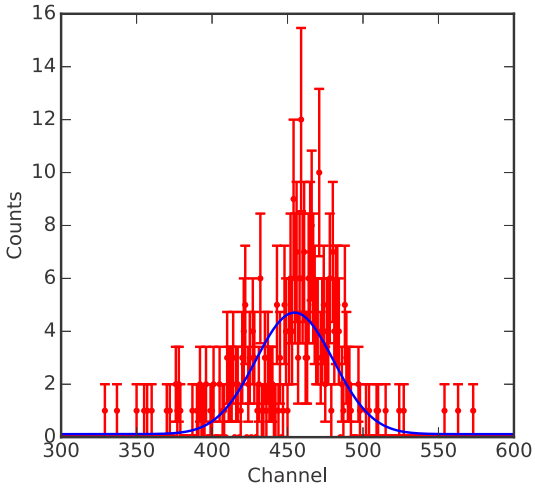

Figure 2: High-resolution X-ray spectrum of the helium-like uranium $1s2p\ ^3P_2 \rightarrow 1s2s\ ^3S_1$ intrashell transition from Ref. [22] and corresponding fit with one Gaussian peak (plus a flat background).

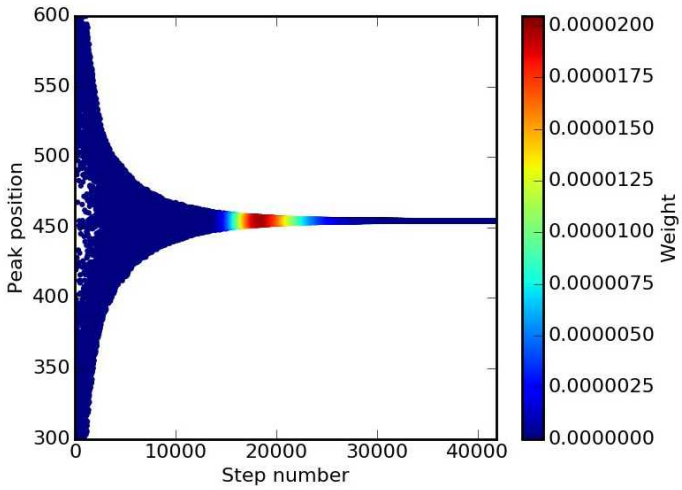

Figure 3: Evolution of the sampled parameter value relative to the peak over the algorithm step.

Step 1: The initial $\{\boldsymbol{a}_{1,k}\}$ live points are sorted considering $P(\boldsymbol{a}_k|I)$ and $\mathcal{L}_1 = \min(\mathcal{L}_{1,k})$ is found. From $\xi_{m,k} = X(\mathcal{L}_{1,k})$ we have $X_1 = \max(\xi_{1,k})$ and the $\Delta X_1 = X_0 - X_1$, where $X_0 = 1$. We have now our first pair of values for the sum in Eq. (18).

Step 2: We built now a new ensemble of live points $\{\boldsymbol{a}_{2,k}\}$, which is the same as $\{\boldsymbol{a}_{1,k}\}$ but where we remove the k'-th element with the lower value of likelihood (corresponding to the higher value of X, i.e. where $\mathcal{L}_1 = L(\boldsymbol{a}_{1,k'})$ with $X_1 = \xi_{1,k'}$ and we store its value with the name $\tilde{\boldsymbol{a}}_1 = \boldsymbol{a}_{1,k'}$. We replace this point with a new $\boldsymbol{a}$ value, randomly chosen with the only condition $L(\boldsymbol{a}) > \mathcal{L}_1$. With this requirement we impose that this point is inside the volume $V_{L\geq\mathcal{L}_1}$. From this new ensembles $\{\xi_{2,k}\}$ and $\{\mathcal{L}_{2,k}\}$ we define $X_2 = \max(\xi_{2,k})$. The interval $[0, X_2]$ correspond to the volume of the parameter space $V_{L\geq\mathcal{L}_2}$ nested in the volume $V_{L\geq\mathcal{L}_1}$ (see Fig. 1). We have then the element $\mathcal{L}_2, \Delta X_2$ of the sum in Eq. (18) and we store the value of the discarded live points.

Step m: We continue the iteration as in the step 2, storing at each step the values $\mathcal{L}_m, \Delta X_m$ and $\tilde{\boldsymbol{a}}_m$. All new live points $\{\boldsymbol{a}_{m,k}\}$ are enclosed in smaller and smaller parameter volumes defined by $L(\boldsymbol{a}) > \mathcal{L}_m$ that correspond to the intervals $[0, X_m]$ (see Fig. 1) with $X_m = \max(\xi_{m,k})$.

Step M, the end: After M iterations, the estimated error Err_M on the evidence E evaluation due to the truncation of the sum in Eq. (18) is less than the target accuracy ΔE and we stop the calculation. For each step m, Err_m is upper limited by the product $L_{max}X_m$ where $L_{max} = \max[L(\boldsymbol{a}_{m,k})]$. When $L_{max}X_m < \Delta E$, we have $Err_m < \Delta E$, the main iteration loop of the nested algorithm stops and the main calculation is finalized. The likelihood function value associated to the last live points is the average $\mathcal{L}_M = \langle L(\boldsymbol{a}_{M,k})\rangle$.

In addition to the final live points $\{\boldsymbol{a}_{M,k}\}$ and their likelihood function values, all intermediate $\mathcal{L}_m, \Delta X_m, \tilde{\boldsymbol{a}}_m$ are stored and used for the calculation of the posterior probability distributions as presented in Sec. 3.3.

For the specific example where we consider the analysis of the data in Fig 2 and a Gaussian peak as model, we show in Fig. 3 the evolution of the values of the $\tilde{\boldsymbol{a}}_m$ component relative to the peak position as function of the algorithm step number. Starting from a sampling range corresponding to our priors (here a flat distribution between channel 300 and 600), the algorithm explores smaller and smaller ranges corresponding to nested volumes of the parameter space. The product $\mathcal{L}_m\Delta X_m$ relative to each steps are shown in both plots of Fig. 4 via the value $weight = \mathcal{L}_m\Delta X_m/E$ (see next section for further explanation).

We have now a recipe for calculating $\mathcal{L}_m$ values but not the X_m. In the previous paragraphs we defined $X_m = \max[X(\mathcal{L}_{m,k})]$ using Eq. (15). But we do not know the function $X(\mathcal{L})$ and neither its inverse $\mathcal{L}(X)$. We can, however, estimate the values of X_m from some statistical consideration. The extraction of a set of K live points $\boldsymbol{a}_{m,k}$ in the parameter volume $V_{L(\boldsymbol{a})>\mathcal{L}_m}$ correspond to sort K random numbers in the interval $[0, X_m]$ (with $\xi_{m,k} = X[L(\boldsymbol{a}_{m,k})]$). For each step, when we pass from the $[0, X_{m-1}]$ interval to the $[0, X_m]$ interval, we shrink the volume (one-dimensional here) by a factor $t_m = X_m/X_{m-1}$. The probability distribution for each t_m is equal to the probability for having a maximum value t given K random numbers $\in [0, 1]$. The statistical distribution of t is (see Appendix B for more details)

$$P(t) = Kt^{K-1}, \quad \text{with} \quad \langle \ln t\rangle = -1/K. \tag{19}$$

For the first and second step we have $X_1 = t_1$ ($X_0 = 1$) and $X_2 = t_2X_1 = t_1t_2$. For a generic step, considering Eq. (19), X_m

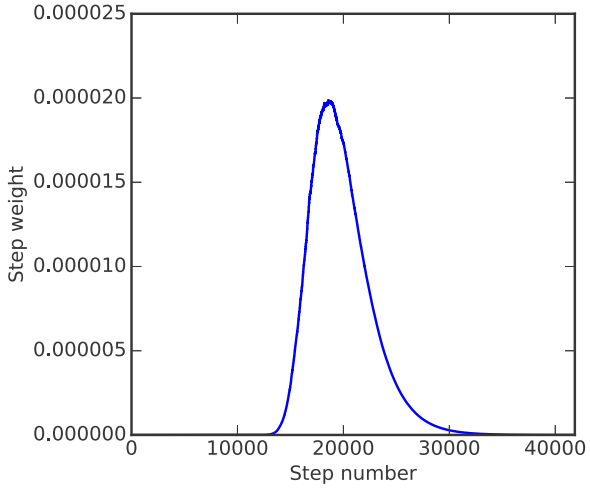

Figure 4: Weights associated to discarded value at each step, which are proportional to the product $\mathcal{L}_m \Delta X_m$)].

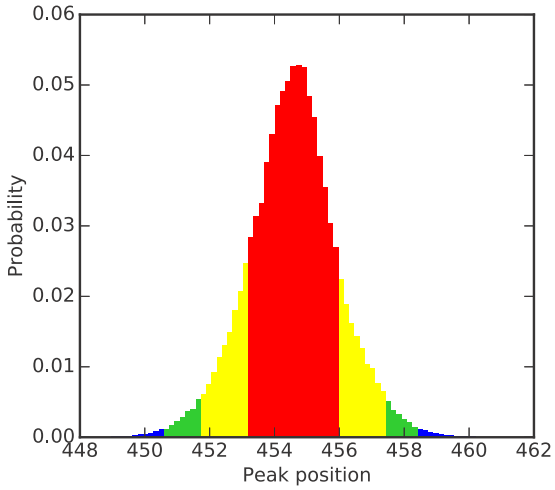

Figure 5: Histogram of the Gaussian peak position built from the values $\mathcal{L}_m$, ΔX_m, and $\tilde{\boldsymbol{a}}_m$ (see text). Red, yellow and green regions indicate 68%, 95% and 99% confidence intervals (credible intervals).

is given by the product

$$X_m = \prod_i^m t_i \quad \text{and then} \quad X_m \approx e^{-m/K}. \tag{20}$$

From this equation, the values of ΔX_m can be estimated, with $\Delta X_M = e^{-M/K}$ for the last live points. This approximation introduces an error in the evidence calculation that is proportional to $K^{-1/2}$, where K is the number of the employed live points. A detailed discussion of the evidence uncertainty is presented in Appendix B.

We note that for the final calculation of the evidence, the terms $\mathcal{L}_m, \Delta X_m$ in Eq. (18) are not equally important. ΔX_m values are monotonically decreasing with m where $\mathcal{L}_m$ values are increasing. As we can see from Fig. 3, the product $\mathcal{L}_m \Delta X_m$ (which defines the step weight as we will see next section) has a maximum. $\tilde{\boldsymbol{a}}_m$ corresponding to this maximum will strongly influence the posterior probability distributions and the value of the evidence.

The bottleneck of the nested sampling algorithm is the search of new points within the J-dimensional volume defined by $L > \mathcal{L}_m$. Different methods are commonly employed to accomplish this difficult task. One efficient method is the ellipsoidal nested sampling [23]. It is based for each step on the approximation of the iso-likelihood contour defined by $L = \mathcal{L}_m$ by a J-dimensional ellipsoid calculated from the covariance matrix of the live points. The new point is then selected within the ellipsoidal volume (times an user-defined enlargement factor). This methods, well adapted for unimodal posterior distribution has been also extended to multimodal problems [24, 12], i.e. with the presence of distinguished regions of the parameter space with high values of the likelihood function. Other search algorithms are based on Markov chain Monte Carlo (MCMC) methods [25], as in particular the *lawn mower robot* method, developed by L. Simons [26], and the recent *Galilean Monte Carlo* [27, 28], particularly adapted to explore the regions close to the boundary of $V_{L>\mathcal{L}_m}$ volumes. `Nested_fit` program is based in an evolution of Simons' algorithm and is presented in Sec. 4.

Additional material on the nested sampling can be found in Refs. 20, 15, 21, 23, 12, 25. In particular in Ref. 29, the different search algorithms, their efficiency and accuracy are discussed.

3.3. Posterior probability distributions

The posterior probability distributions are built from the discarded live points $\tilde{\boldsymbol{a}}_m$, the final set of K live points $\boldsymbol{a}_{M,k}$ and their associated $\mathcal{L}_m, \Delta X_m$ values.

Once the evidence $E \equiv P(\{x_i, y_i\}|I)$ is determined, posterior inference can be easily generated from the $\{\tilde{\boldsymbol{a}}_m\}$ and $\{\boldsymbol{a}_k\}_M$ values. Each $\tilde{\boldsymbol{a}}_m$ is in the infinitesimal parameter volume $\Delta V_{\mathcal{L}_m < L(\tilde{\boldsymbol{a}}_m) < \mathcal{L}_{m+1}}$ that correspond to the interval ΔX_m. Considering the discrete form of Eq. (16) and Eq. (10), we can calculate the probability associated to the parameter values $\tilde{\boldsymbol{a}}_m$, in other words the step *weight* named in the previous sections:

$$P(\tilde{\boldsymbol{a}}_m|\{x_i, y_i\}, I) = P(X_m) \approx \frac{\mathcal{L}_m \Delta X_m}{E}. \tag{21}$$

From Eq. (21), the probability distribution of any single parameter a_j is obtained by marginalization (Eq.(8)), i. e. integrating of the posterior probability $P(\boldsymbol{a}|\{x_i, y_i\}, I)$ over the other parameters. In our case, if the parameter of interest corresponds to the j^{th} component, its probability distribution can be built from $(\tilde{a}_m)_j$ values and their corresponding weights defined by Eq. (21).

For our specific example with a Gaussian distribution as a model, this corresponds to take $(\tilde{a}_m)_j$ values showed in Fig. 3 (top) and built a weighted histogram (with the weights $\mathcal{L}_m \Delta X_m / E$ showed by the different color intensities in Fig. 3 and in Fig. 4). From the marginalization on $J - 2$ parameters

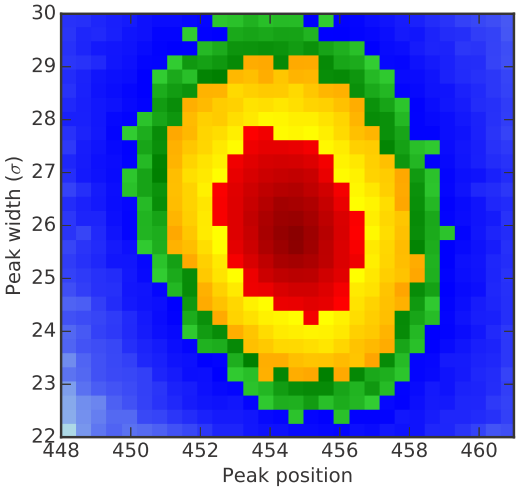

Figure 6: Joint probability distribution of the parameters relative to the peak position and width (in terms of sigmas of the Gaussian profile). Red, yellow and green regions indicate 68%, 95% and 99% confidence intervals (credible intervals).

also joint probabilities can be built, as that one presented in Fig. 6 corresponding to the position and width distribution of the peak.

4. The program

4.1. General considerations

`Nested_Fit` has been developed in Fortran90 for the calculation of the Bayesian evidence and posterior parameters probability distributions for a given set of data and selected model. The core of `Nested_Fit` is the algorithm used for the calculation of the Bayesian evidence which is, as indicated by its name, the nested sampling developed by Skilling and presented in Sec. 3, but with an original method to find new live points.

Due to the large span of the values of the different calculated element (likelihood function, evidence, X_m, etc.), all computations are done with respect to their logarithms, as many programs based on the nested sampling. The data input are provided in the form $\{x_i, y_i\}$, where x_i are real numbers and y_i are necessarily counts detected at the channels x_i. To analyze statistical poor (but also not-poor) data sets, the likelihood function is built considering a Poissonian statistics (which tend to the normal distribution for large number of counts) for each channel, leading to

$$L(\boldsymbol{a}) = \prod_i \frac{F(x_i, \boldsymbol{a})^{y_i} e^{-F(x_i,\boldsymbol{a})}}{y_i!}, \qquad (22)$$

where for each channel, y_i is the recorded number of counts and $F(x_i, \boldsymbol{a})$ is the expected value of the modeling function that depends on the parameters $\boldsymbol{a}$. A large library of functions is available and new *ad hoc* functions can easily be added.

Outputs of `Nested_Fit` include the evaluation of the Bayesian evidence, the corresponding information gain and complexity and the information to build parameter probability distributions. The different probability histograms and other plots are produced via a series of functions of a dedicated python library. The figures of this article are examples of their typical outcomes. Additionally to the graphically outputs, `Python` library functions can be used to recursively modify the input file `nf_input.dat` and to read the results in the output files. These functions are particularly useful for automated analysis and systematic surveys.

Several set of data can analyzed at the same time. For example, distinct spectra with a same response function can be analyzed, and common parameters such as the profile width can be extracted by correctly taking into account the correlations between data sets.

4.2. Computation algorithm of the Bayesian evidence

The calculation of the Bayesian evidence is made with the nested sampling following the steps presented in Sec. 3, similarly to other programs based to the same algorithm [15, 23, 24, 12, 25]. Even if the basic structure is practically identical to existing codes, the algorithm for the search of new live points is substantially different. The searching algorithm is a Makov chain Monte Carlo method to explore the parameter volume $V_{L>\mathcal{L}_m}$. To cancel the correlation between the starting point and the final point, a series of N jumps are done in this volume. The different stages of the algorithm are

1. Choose randomly a starting point $\boldsymbol{a}_{n=0} = \boldsymbol{a}_0$ from the available live points $\{\boldsymbol{a}_{m,k}\}$ as starting point of the Markov chain where n is the number of the jump. The number of tries n_t (see below) is set to zero.
2. From the values $\boldsymbol{a}_{n-1}$, find a new parameter sets $\boldsymbol{a}_n$ where each j^{th} parameter is calculated by $(a_n)_j = (a_{n-1})_j + f\, r\, \sigma_j$, where σ_j is the standard deviation of the live points $\{\boldsymbol{a}_{m,k}\}$ relative to the j^{th} parameter, $r \in [-1, 1]$ is a sorted random number and f is a parameter defined by the user.
   (a) If $L(\boldsymbol{a}_n) > \mathcal{L}_m$ and $n < N$, go to the beginning of step 2 with an increment of the jump number $n = n + 1$.
   (b) If $L(\boldsymbol{a}_n) > \mathcal{L}_m$ and $n = N$, $\boldsymbol{a}_{n=N}$ is new *live point* to be included in the new set $\{\boldsymbol{a}_{m+1,k}\}$.
   (c) If $L(\boldsymbol{a}_n) < \mathcal{L}_m$ and $n < N$ and the number of tries n_t is less than the maximum allowed number N_t, go back to beginning of step 2 with an increment of the number of tries $n_t = n_t + 1$.
   (d) If $L(\boldsymbol{a}_n) < \mathcal{L}_m$ and $n < N$ and $n_t = N_t$ a new parameter set $\boldsymbol{a}_0$ has to be selected. Instead than choosing one of the existing live points, $\boldsymbol{a}_0$ is built from distinct j^{th} components from different live points: $(a_0)_j = (a_{m,k})_j$ where k is randomly chosen between 1 and K for each j. Then $\boldsymbol{a}_{n=0} = \boldsymbol{a}_0$ and go to the beginning of step 2.

The last step makes the algorithm well adapted to problems with multimodal parameter distributions allowing to easily jump between high-likelihood regions. An example of presence of several maximal likelihood regions is presented in Fig. 7 where we plot the joint probability of the position and amplitude of one of the four Gaussian peaks of the considered model.

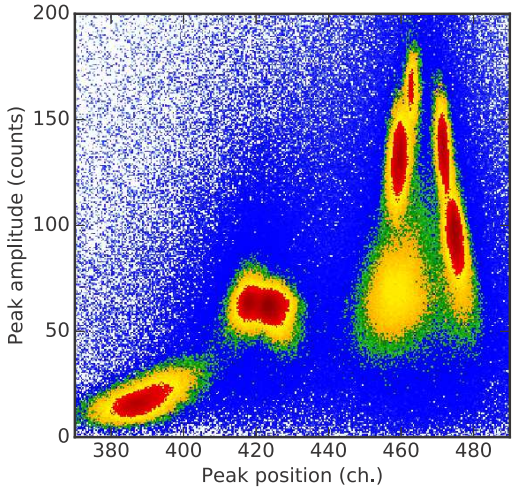

Figure 7: Joint probability distribution of the parameters relative to the position and amplitude of one peak in a four Gaussian peaks model. Eight distinct likelihood maxima can be identified. Red, yellow and green regions indicate 68%, 95% and 99% confidence intervals.

The value of N_t is fixed in the code ($N_t = 10000$ in the present version). The other parameters can be provided by the input file.

4.3. Inputs

All input parameter required by `Nested_fit` are provided in the file `nf_input.dat` The most important inputs are:

The maximum number of jumps N and the real number f: These parameters are important for efficiency of the search of the new live points and for the non-correlated and efficient exploration of the parameter space. Higher value of f guarantee a better independence between the current live points and the new point but a minor efficiency for finding it because of the higher probability to jump in the volume region $L < \mathcal{L}_m$. The same reasoning is applied for the total number of jumps N.

The number of live points K: The choice of K influence directly the expected accuracy of the evidence $\delta E \propto 1/\sqrt{K}$, and also provides a better sampling of the parameter volume. As counterpart, an increasing of K increases the computation time.

The required final evidence accuracy ΔE: A too large value of the accuracy will bias the evidence calculation. A too small value can make the evidence computation significantly long. For a given problem, the optimal value is obtained by looking a posteriori at the evolution of $\mathcal{L}_m \Delta X_m$. The caclulation has to stop significantly far from the region where the product $\mathcal{L}_m \Delta X_m$ is large, i. e. far from the most influent values of $X \sim \exp(-\mathcal{H})$ where $\mathcal{H}$ is the extracted information (in the sense of Shannon, see Appendix A and Appendix B). Good and efficient values are generally between 10^{-3} and 10^{-5} as also discussed in Ref. 25.

The number of trials sets of live points N_{LPS}: Besides theoretical considerations, the best strategy to estimate the evidence accuracy is to calculate E several times with different starting sets of live points (with different seed for the random generator) and to extract the mean and standard deviation of the logarithmic values of the computed evidence, which is the pertinent quantity for the uncertainty evaluation (see Appendix B). In addition this method provides more sampling points of the parameter space for a better evaluation of the posterior probability distributions, especially important when multimodal distributions are present.

The parameter priors Priors of the different parameters can be selected between two options: (i) an uniform prior where the parameter value boundaries have to be provided or (ii) a normal distribution where a main value and the associated standard deviation have to be provided (as example from a past experiment).

Except for the priors, for each case the different parameters have to be tuned by looking the output in order to have valuable results (to uniformly and randomly cover the entire parameter space) but also to have a fast calculation (a good efficiency to find new live points). For this goal, the most sensitive parameters are the number of live points K, the number of jumps N and the real number f.

4.4. Outputs

Once ended, the program provides four major output files described below.

- `nf_output_res.dat` contains the details of the computation (n. of live points trials, n. of total iteration), the final evidence value and its uncertainty $E \pm \delta E$, the parameter values $\hat{\boldsymbol{a}}$ corresponding to the maximum of the likelihood function, and the mean, the median, the standard deviation and the confidence intervals (68%, 95% and 99%) of the posterior probability distribution of each parameter. Moreover, the information gain $\mathcal{H}$, the Bayesian complexity C and the theoretical minimal value of iteration deduced from the extracted information value are also provided. δE is calculated only if $N_{LPS} > 2$.

- `nf_output_data.dat` contains the original input data together with the model function values corresponding to the parameters $\hat{\boldsymbol{a}}$ with the highest likelihood function value, the residuals and the uncertainty associated to the data.

- `nf_output_tries.dat` is present only if $N_{LPS} > 2$. For each live points trial, it contains the final evidence, the number of iterations and the maximum value of the likelihood function.

- `nf_output_points.dat` contains all discarded and final live points values $\tilde{\boldsymbol{a}}_m$ and $\{\boldsymbol{a}_{M,k}\}$, their associated likelihood values $L(\{x_i, y_i\}, \boldsymbol{a})\}$ and posterior probabilities $P(\boldsymbol{a}|\{x_i, y_i\}, I) \approx \mathcal{L}_m \Delta X_m / E$. From them, the different parameter probability distributions, as shown in Fig. 5, or

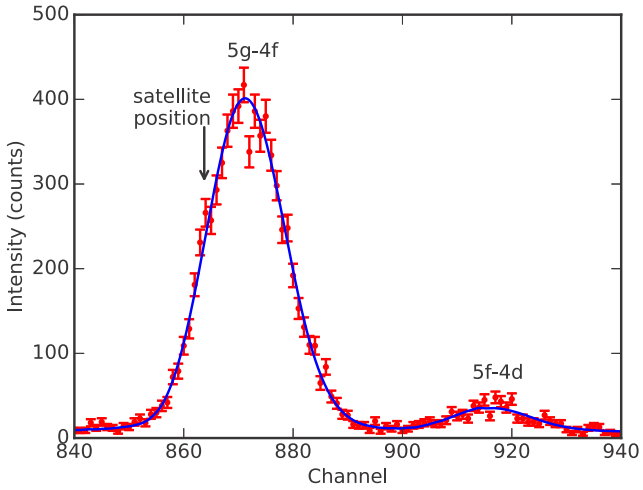

Figure 8: Pionic nitrogen 5 – 4 transitions. Possible additional transitions from the presence of one remaining electron in the K shell are indicated

joint probabilities, as shown in Figs. 6 and 7, can be built from the marginalization (Eq. 8) of the unretained parameters.

5. Two examples

In this section we will present two practical applications of the of the statistical analysis methods described above. In the first one, we calculate the probability of the presence or not of a satellite line in a spectrum at a well defined position but with unknown intensity. The second, more complex, consists in the analysis of a statistically poor set of data for which we would like to determine the most probable model among different possibilities and to extract the position of the main component.

5.1. Satellite line contamination

We consider a common case in spectroscopy where we would like to test the presence or not of an unresolved weak spectral line close to a intense line. In this specific example, we consider the $5g - 4f$ transition in pionic nitrogen, an hydrogen-like atom formed by a nitrogen nucleus and a negatively charged pion. During the formation of the pionic atoms, all electrons are expected to be ejected. The presence of a remaining electron in the K shell cannot completely be excluded. Its presence can cause a shift of the main tradition energy due to the Coulomb screening and then an appearance of a new component in the spectrum. To determine the probability of such a scenario, we have to calculate the evidence for the two possible models: Model 1 without remaining electrons (a pure H-like pionic atom) and Model 2 with the possible presence of one remaining electron. More details on the physics case can be found in Refs. 30, 31.

The examined data consist in seven distinct spectra similar to the one represented in Fig. 8 for a total of about 60000 recorded counts. Each spectra is obtained by a Bragg spectrometer equipped by a spherically bent crystal. The evidence and probability distributions of both models are computed with `Nested_fit` taking into account all seven spectra at the same time. For this specific propose we used $K = 1000$ live points and an accuracy requirement $\delta E = 10^{-5}$. For the search of the new points we choose the values $J = 20$ jumps and $f = 0.1$. These parameters insure an efficient and complete exploration of the parameter space and an accurate evaluation of the evidence. For a rough estimation of the evidence uncertainty we consider $N_{LPS} = 8$ different live point trial sets. For both models, we chose flat prior probability distributions for the different parameters. Compared to model 1, model 2 has only as additional free parameter the satellite line intensity whose relative position with respect to the main line has been fixed by the theory.

Since we have to choose among two models only, the relevant quantity to calculate is the Bayes factor B_{12}, defined in Sec. 2.3, from which we can determine the criterium in favor to one of the two hypothesis.

From the output of `Nested_fit`, we have $\ln B_{12} = 6.6 \pm 1.8$, which correspond to a probability of 99.98% in favor to the model without remaining electrons (between 99.86% and 100% when the Bayes factor uncertainty is taken into account). This Bayes factor value indicates a decisive support for the M_1 hypothesis considering any considered scale ("decisive" in the Jeffreys scale [8], "very strong" in the Kass scale[17] or "strong" in the Gordon-Trotta scale [18]) with an equivalent p-value of about 10^{-5} for Model 2 [18].

In conclusion, the presence of remaining electrons can be safely excluded and the main line position can be reliably evaluated. Additional discussion on this analysis can be found in Ref. 31.

5.2. A nasty peak

In this second example we consider the experimental data already presented in Sec. 4 corresponding to the helium-like uranium $1s2p\ ^3P_2 \rightarrow 1s2s\ ^3S_1$ intrashell transition obtained from a Bragg diffraction spectrometer equipped by a curved crystal [22]. As can be seen in Fig. 9, the experimental peak is statistically poor, quite broad and asymmetric. We do not know where this asymmetry comes from. Eventually, it might be related to the presence of several spectral components or from spectrometer's aberrations. From the Bayesian analysis we would like i) to determine the most probable model that describes the data and ii) to determine the probability distribution of the main spectral component position, independently on the choice of the model.

For each model, we calculate with `Nested_fit` the evidence, the probability distributions and the complexity using the same parameters as in the previous example except for the number of live points and the number of trial sets. Here we use $K = 2000$ live points and N_{LPS} between 8 and 32 depending on the model. For all models, we chose flat prior probability distributions.

First we consider the simple case where we can have only Gaussian peaks, between one and four, with the same width σ,

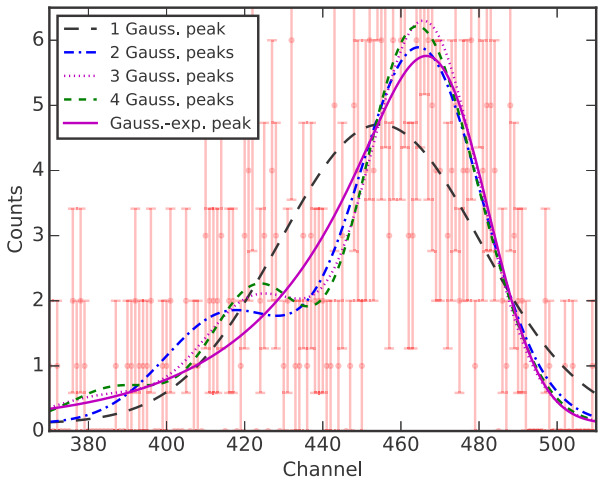

Figure 9: Profile curves corresponding to the likelihood maxima of the different models (1–4 Gaussian peaks and Gaussian-exponential peak).

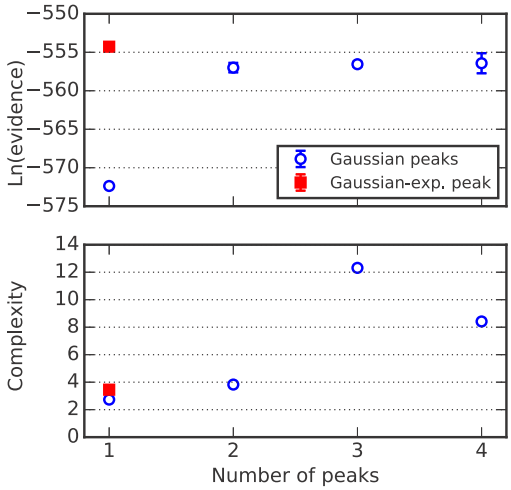

Figure 10: Evidence and complexity of the different considered models.

which we know to be a priori between 10 and 30 channels, and a flat background. From these working hypotheses, we would like to determine which model is the most probable. i.e. how many peaks are present, and what is the position of the main peak. To note, the model with four Gaussian peaks requires for any single trial set much more computation time than the single peak due to the presence of several high-value likelihood regions (see Fig. 7). This is in fact the practical reason why we consider a maximum of four components. Similar examples have been presented in the past by Sivia [32, 15]. With respect to these works, here we consider the analysis of a statistically poor data set from a real experiment instead of a simulation, where we do not know the real nature of the spectra.

To visually compare the outcome of the different models, we present in Fig. 9 the corresponding curves relative to each likelihood function maximum. As it can be observe, the profile maxima are close to each other except for the single Gaussian peak profile. In the particular case of the 4-peak model, two Gaussian component are unresolved (as suggested by Fig. 11).

The quantitative results obtained from `Nested_fit` are summarized in Table 1 and Fig. 10 where we report values of the evidence (in the logarithmic scale), of the model complexity and the probability of the model (in the table only). The model with a single Gaussian peak results to have a very low probability. From the results of the other hypotheses, we cannot clearly determine how many peaks are present because models with 2, 3 and 4 components have the same evidence (within the associated uncertainty). As suggested by Trotta [11], a criterium to choose between different models with similar evidence is the Bayesian complexity C value (see Sec. 2.3 and Appendix A). When different hypotheses have similar evidence values, we should choose the model with the lower value of C to favor once more simple models versus complex models, in agreement to the Ockham's razor principle. In our case, the two-peak model is the favorite with a low complexity value, only slightly higher

Table 1: Evidence, model probability and complexity of the different models: a series of n Gaussian peak or a Gaussian-exponential (left side) peak.

Model	$\ln E$	Relative probability	Complexity
1 G. peak	-572.37 ± 0.15	4.9×10^{-8}	2.7
2 G. peaks	-557.00 ± 0.62	23.1%	3.8
3 G. peaks	-556.55 ± 0.31	36.0%	12.3
4 G. peaks	-556.43 ± 1.30	40.9%	8.4
1 G.-exp. peak	-554.28 ± 0.21	–	3.5

than the one-peak model complexity, and high probability.

If we are not interested to determine the number of peaks, but only to the main peak position component μ_0, we can built the correspondent probability distribution $P(\mu_0|\{x_i, y_i\}, I)$ from the output of the each model analysis. As in Eq. (14), we can build $P(\mu_0|\{x_i, y_i\}, I)$ from the different $P(\mu_0|\{x_i, y_i\}, \mathcal{M}_\ell, I)$ distributions using as weight the model probabilities summarized in Table 1. The final probability distribution of the main peak position (around channels 450–480) is presented in Fig. 11. It is quite complex, with the presence of several maxima mainly due to the four-peak model contribution. These maxima correspond in fact to the high-likelihood regions visible in Fig. 7. Because of the low probability, the one-peak model does not contribute significantly to the final distribution. As comparison its contribution is presented in Fig. 11 with a magnification factor.

Alternatively to the presence of several Gaussian peaks, a valid hypothesis is the presence of some kind of aberration due to the spectrometer characteristics. A spectrometer with cylindrically bent crystal in the Johann geometry is in fact used. To take into account this possibility, we model the aberration effect by a line profile resulting from the convolution between a Gaussian and an exponential function [33]. As we can see in Fig. 9, the curve corresponding to the likelihood maximum reproduces well the data, with a maximum very close to the multi-Gaussian

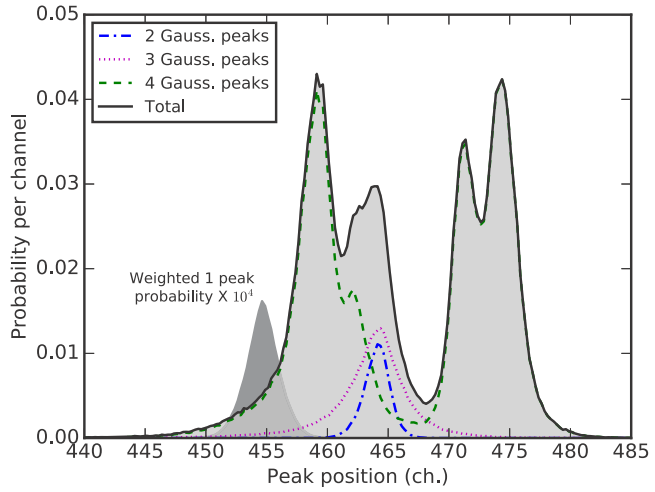

Figure 11: Probability distribution of the main peak position from the single probabilities of the models with one-to-four Gaussian peaks. For the single peak model, we magnify its weighted probability (in grey) to compare the distributions.

peaks models (with exclusion of the single peak model). From Table 1 and Fig. 10, we can observe more quantitatively that the associated evidence is significantly higher than any other model. With respect to the two-Gaussian peaks, the probability for the Gaussian-exponential profile is in fact 98.3%. At the same time the associated complexity remains small, intermediate between single and double Gaussian peak models, indicating that, together with the model probability, the presence of an aberration as explanation of the asymmetry experimental data distribution is the most valid hypothesis.

In the previous paragraphs we show how evidence and complexity evaluations can help to determine the most plausible model to describe a set of data. In this specific example, we remember that we consider a strong assumption on the number of the possible Gaussian peaks to mainly limit the computational time. Other hypotheses could be considered but always taking into account our prior knowledge coming from previous experiments or general physical considerations. Formally this prior knowledge should be included in the model prior probability that, once multiplied to the evidence, gives the final probability for the different models. For this point, critics could be addressed about the objectivity. But again, the meaning of such dependency on the priors should be pragmatically be interpreted as a message saying that the data quality is not sufficient to correctly analyze the problem and choose among different hypotheses. Nevertheless, this approach provide a well defined procedure to exclude unrealistic models with the comparison with the data via the evidence computation (as for the single-peak model) or, via prior probabilities, models that are not consistent with our present knowledge of physics and simply common sense, on which our logic is based.

6. Conclusions

The main intent of this article is to provide an useful starting point for the atomic physics community to use Bayesian methods for data analysis. For this propose, we provide a very synthetic and basic introduction to Bayesian statistics. We show how, from basic logic reasoning with requirement of consistency, a very general definition of probability can be constructed. This definition automatically implies the Bayes' theorem, which plays the central role for the prior probability inclusion. From this approach, we see how posterior probabilities can be simply calculated as well as probabilities for different hypotheses.

To visualize the practical consequences of the use of these new concepts, we show two atomic spectra analysis examples. In the first one we see how we can determine the presence or not of an unresolved spectral line. In the second, more complex, we calculate the probability of different possible models (different number of peaks and peak shape) and we see how to extract valuable information (the main peak position in our case) from equiprobable hypotheses.

For hypothesis testing, the calculation of the Bayesian evidence from the experimental data is essential. Different methods are available in the literature to evaluate the Bayesian evidence. In this article we present in detail the nested sampling technique developed originally by J. Skilling in 2004 based on a particular for of Monte Carlo sampling of the model parameter space. We also present the newly developed program `Nested_fit` based on such method but with a new parameter exploration algorithm. We show its capabilities and typical inputs and outputs.

Acknowledgments

First of all we would like to express our deep gratitude to Leopold M. Simons who introduced us to the Bayesian data analysis and without whom this work could not have been started. We would like to sincerely thank also Nicolas Winckler for the numerous discussions about statistics and data analysis, and, together with Robert Grisenti, for the careful reading of the manuscript. This work have been developed in the context of several experiments; we would like to thank all members of the Pionic Hydrogen, FOCAL and GSI Oscillation collaborations and the ASUR group at the Institute of NanoSicence of Paris for support and discussions.

Appendix A. Information and complexity

The gain of knowledge we obtain from the analysis of experimental data can be quantified in terms of information $\mathcal{H}$, in the Shannon sense [34, 35], that we gain in the process by the comparison between the posterior probability $P(\boldsymbol{a}|\{x_i, y_i\}, I)$ and the prior probability $P(\boldsymbol{a}|I)$. The information gain, in units of nat[2],

[2] *nat* is the unit of information when the normal logarithm is used, similarly to the *bit*, the unit where the base-2 logarithm is employed.

is given by the so-called Kullback-Leibler divergence [36]

$$\mathcal{H} \equiv D_{KL} = \int P(\boldsymbol{a}|\{x_i, y_i\}, I) \ln \left[\frac{P(\boldsymbol{a}|\{x_i, y_i\}, I)}{P(\boldsymbol{a}|I)} \right] d^D\boldsymbol{a}. \quad \text{(A.1)}$$

Considering Eq. (10), D_{KL} can be written as

$$D_{KL} = -\ln E + \int P(\boldsymbol{a}|\{x_i, y_i\}, I) \ln L(\boldsymbol{a}) d^D\boldsymbol{a}, \quad \text{(A.2)}$$

which is nothing else that the negative logarithm of the evidence plus the average of the logarithmic value of the likelihood function.

From D_{KL} there is a interesting quantity can be derived that provides an additional criterion to compare models: the *Bayesian complexity* C. C is calculated from the difference between the D_{KL}, i.e. the average of $\ln(L)$, and the "expected surprise" [11] from the data represented by the value $\hat{D}_{KL}$, where

$$\hat{D}_{KL} = -\ln E + \ln L(\hat{\boldsymbol{a}}), \quad \text{(A.3)}$$

where $\hat{\boldsymbol{a}}$ usually correspond to the posterior parameter mean values, or other possible estimators (ex. the likelihood function maximum or the posterior distribution medians) depending on the details of the problem[3]. The complexity is then defined as

$$C = -2(D_{KL} - \hat{D}_{KL}) = -2\left[\langle \ln L(\boldsymbol{a}) \rangle - \ln L(\hat{\boldsymbol{a}})\right], \quad \text{(A.4)}$$

where the symbol $\langle\ \rangle$ indicates the mean value [2, 11]. C gives in practice a measurement of the number of parameters that the data can support for a certain model $\mathcal{M}$ for a defined set of data and parameter priors [2, 37].

For equiprobable models (similar evidence values), the comparison of Bayesian complexity could determine the choice in favor to one model or the other. Considering two different models $\mathcal{M}_1$ and $\mathcal{M}_2$ with $E_1 \approx E_2$ and different number of parameters $J_1 < J_2$, we can have to cases [11]:

$C_1 < C_2$: The quality of the data is sufficient to measure the additional parameters of the more complicated model, but they do not improve its evidence by much. We should prefer model with less parameters.

$C_1 \approx C_2$: The quality of the data is not sufficient to measure the additional parameters of the more complicated model and we cannot draw any conclusions as to whether extra parameters are needed.

Appendix B. Theoretical uncertainty of the evidence calculation by nested sampling

The main uncertainty of the final evaluation of the evidence calculated by the nested sampling is, as stated by the author of this method J. Skilling, related to the probabilistic nature of the terms ΔX_m in Eq. (18) [15, 21, 38, 25]. The choice of numerical integration of Eq. (17) (rectangle method, trapezoidal rule, etc.) does not influence very much the final result. Instead, the statistical glittering of ΔX_m in Eq. (18) introduces an error.

The interval values are calculated from $X_m = \prod_i^m t_i$ (Eq. (20)), where t_i are the shrinking of the considered interval of X. The statistical distribution of the shrinking values t_i can be obtained from simple probabilistic considerations. For each step m, the shrinking value is derived from the $\{\xi_{m,k}\}$ values of X that correspond to the K considered live points. The K randomly sorted live points correspond the K values $\{\xi_{m,k}\}$ that are uniformly distributed in the interval $[0, X_m]$. To pass to the $m+1$ step, we have to identify the maximum value of $\{\xi_{m,k}\}$ to determine the shrinking factor $t_{m+1} = \max(\xi_{m,k}/X_m)$. This correspond to find the maximum of K values $\{x_k\}$ uniformly distributed in the interval $[0, 1]$ (where $x_k = \xi_{m,k}/X_m$).

Considering a certain $x_{k'} = t$, the probability that all other values are less than t is $\prod_{k\neq k'} P(x_k \in [0, t]) = t^{K-1}$. Because this is valid for any $x_{k'} \in \{x_k\}$, then

$$P(t = \max\{x_k\}) = Kt^{K-1}. \quad \text{(B.1)}$$

This probability distribution has the following properties. The average and standard deviation of $\ln t$ are

$$\langle \ln t \rangle = -\frac{1}{K} \quad \text{and} \quad \sigma_{\ln t} = \frac{1}{K}. \quad \text{(B.2)}$$

From the above equation, we evaluate the X_m values

$$\ln X_m = -\frac{m}{K} \pm \frac{\sqrt{m}}{K}. \quad \text{(B.3)}$$

If the main value of X_m is taken into account (as in Sec. 3), we introduce an error of the order of $\sqrt{m}/K$ in the evidence evaluation via ΔX_m values.

As we can see in Fig. 4, not all m steps contribute equally to for the final value of E. The calculated evidence is dominated by the region where the product $\mathcal{L}_m \Delta X_m$ is maximal. The maximum position can correlate to the information gain $\mathcal{H}$ associated to the data (and the model) by Eq. (A.1).

To estimate this position, we have to make some approximation. Considering Eqs. (A.1), (16) and (17), we have that the information in terms of $\mathcal{L}(X)$ is

$$\mathcal{H} = \int_0^1 \frac{\mathcal{L}(X)}{E} \ln \left[\frac{\mathcal{L}(X)}{E} \right] dX = \int_0^1 P(X) \ln P(X) dX. \quad \text{(B.4)}$$

If we assume the extreme case of a likelihood function with a core with a constant value $\mathcal{L}(X) = \hat{\mathcal{L}}$ for $X < \hat{X}$ and zero elsewhere [38], we have that $E = \hat{L}\hat{X}$ and then $P(X) = 1/\hat{X}$ for $X < \hat{X}$ and zero otherwise. In this case we have

$$\mathcal{H} = \int_0^{\hat{X}} \frac{1}{\hat{X}} \ln\left(\frac{1}{\hat{X}}\right) dX = -\ln \hat{X} \quad \text{(B.5)}$$

and then $\hat{X} = e^{-\mathcal{H}}$ (see also Refs. 15, 21, 38, 25 for further considerations).

From Eqs. (B.3) and (B.5), we see that the m value associated to this region, the most influent region for the value of E, is $m = K\mathcal{H}$ and

$$\ln \hat{X} = \mathcal{H} \pm \sqrt{\frac{\mathcal{H}}{K}}. \quad \text{(B.6)}$$

[3] For multimode posterior probability distributions, the likelihood function maximum is more adapted. In fact the mean value can easily be far from the parameter region corresponding to high values of the likelihood function.

The dominant uncertainty associated to the evidence is then

$$\delta(\ln E) \approx \delta\left[\ln\left(\sum_m \Delta X_m\right)\right] \approx \sqrt{\frac{\mathcal{H}}{K}}. \tag{B.7}$$

Many approximations in this evaluation have been done but the dependency of $\delta(\ln E) \propto 1/\sqrt{K}$ emerges. This dependency has been confirmed by computational studies [25] that also investigate the influence of the search algorithm parameters for the new live points in the nested sampling.

A more pragmatic and practical way to evaluate the accuracy of E, which is employed in `Nested_Fit` program (see Sec. 4), is to calculate the evidence for different trials with different sets of live points and calculate then the average and the standard deviation of the different values of $\ln E$. From the consideration above, this is in fact the natural estimation to study the uncertainty of E [38, 39].

References

[1] H. Akaike, A new look at the statistical model identification, IEEE Trans. Automat. Contr. 19 (6) (1974) 716–723. doi:10.1109/TAC.1974.1100705.

[2] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. Van Der Linde, Bayesian measures of model complexity and fit, J. R. Stat. Soc. B 64 (4) (2002) 583–639. doi:10.1111/1467-9868.00353.
URL http://dx.doi.org/10.1111/1467-9868.00353

[3] P. Bevington, D. Robinson, Data reduction and error analysis for the physical sciences, McGraw-Hill, 2003.

[4] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Observation of gravitational waves from a binary black hole merger, Phys. Rev. Lett. 116 (6) (2016) 061102.
URL http://link.aps.org/doi/10.1103/PhysRevLett.116.061102

[5] B. P. Abbott *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), Gw151226: Observation of gravitational waves from a 22-solar-mass binary black hole coalescence, Phys. Rev. Lett. 116 (24) (2016) 241103.
URL http://link.aps.org/doi/10.1103/PhysRevLett.116.241103

[6] M. Bayes, M. Price, An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, Philosophical Transactions 53 (1763) 370–418. doi:10.1098/rstl.1763.0053.
URL http://rstl.royalsocietypublishing.org/content/53/370.short

[7] P. Laplace, Essai philosophique sur les probabilités, Bachelier, 1825.

[8] H. Jeffreys, Theory of Probability, 3rd Edition, Oxford University Press, Oxford, U.K., 1961.

[9] E. Jaynes, G. Bretthorst, Probability Theory: The Logic of Science, Cambridge University Press, 2003.

[10] A. Lewis, S. Bridle, Cosmological parameters from cmb and other data: A monte carlo approach, Phys. Rev. D 66 (10) (2002) 103511.

[11] R. Trotta, Bayes in the sky: Bayesian inference and model selection in cosmology, Contemp. Phys. 49 (2) (2008) 71–104.

[12] F. Feroz, M. P. Hobson, M. Bridges, Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics, Mon. Not. R. Astron. Soc. 398 (4) (2009) 1601–1614. doi:10.1111/j.1365-2966.2009.14548.x.
URL http://dx.doi.org/10.1111/j.1365-2966.2009.14548.x

[13] R. T. Cox, Probability, frequency and reasonable expectation, Am. J. Phys. 14 (1) (1946) 1–13. doi:doi:http://dx.doi.org/10.1119/1.1990764.
URL http://scitation.aip.org/content/aapt/journal/ajp/14/1/10.1119/1.1990764

[14] L. E. Ballentine, Probability theory in quantum mechanics, Am. J. Phys. 54 (10) (1986) 883–889. doi:doi:http://dx.doi.org/10.1119/1.14783.
URL http://scitation.aip.org/content/aapt/journal/ajp/54/10/10.1119/1.14783

[15] D. S. Sivia, J. Skilling, Data analysis: a Bayesian tutorial, 2nd Edition, Oxford University Press, 2006.

[16] R. Cox, Algebra of Probable Inference, Johns Hopkins University Press, 1961.

[17] R. E. Kass, A. E. Raftery, Bayes factors, J. Am. Stat. Assoc. 90 (430) (1995) 773–795. doi:10.1080/01621459.1995.10476572.
URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.

[18] C. Gordon, R. Trotta, Bayesian calibrated significance levels applied to the spectral ti Mon. Not. R. Astron. Soc. 382 (4) (2007) 1859–1863. doi:10.1111/j.1365-2966.2007.12707.x.
URL http://mnras.oxfordjournals.org/content/382/4/1859.abstract

[19] C. Robert, G. Casella, Monte Carlo Statistical Methods, Springer New York, 2013.

[20] J. Skilling, Nested sampling, AIP Conf. Proc. 735 (1) (2004) 395–405.

[21] J. Skilling, Nested sampling for general bayesian computation, Bayesian Anal. 1 (4) (2006) 833–859.

[22] M. Trassinelli, A. Kumar, H. F. Beyer, P. Indelicato, R. Märtin, R. Reuschl, Y. S. Kozhedub, C. Brandau, H. Bräuning, S. Geyer, A. Gumberidze, S. Hess, P. Jagodzinski, C. Kozhuharov, D. Liesen, U. Spillmann, S. Trotsenko, G. Weber, D. F. A. Winters, T. Stöhlker, Observation of the $2p_{3/2} \rightarrow 2s_{1/2}$ intra-shell transition in He-like uranium, Eur. Phys. Lett. 87 (6) (2009) 63001.

[23] P. Mukherjee, D. Parkinson, A. R. Liddle, A nested sampling algorithm for cosmological model selection, Astrophys. J. Lett. 638 (2) (2006) L51.

[24] F. Feroz, M. P. Hobson, Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses, Mon. Not. R. Astron. Soc. 384 (2) (2008) 449–463.

[25] J. Veitch, A. Vecchio, Bayesian coherent analysis of in-spiral gravitational wave signals with a detector network, Phys. Rev. D 81 (6) (2010) 062003.

[26] M. Theisen, Analyse der linienform von röntgenübergängen nach der bayesmethode, Diplomarbeit, Fakultät für Mathematik, Informatik uns Naturwisssenschaften der RWTH Aachen (2013).

[27] J. Skilling, Bayesian computation in big spaces-nested sampling and galilean monte carlo, AIP Conf. Proc. 1443 (1) (2012) 145–156.

[28] F. Feroz, J. Skilling, Exploring multi-modal distributions with nested sampling, AIP Conf. Proc. 1553 (1) (2013) 106–113.

[29] J. Buchner, A statistical test for nested sampling algorithms, Stat. Comp. 26 (1) (2016) 383–392.

[30] M. Trassinelli, D. F. Anagnostopoulos, G. Borchert, A. Dax, J. P. Egger, D. Gotta, M. Hennebach, P. Indelicato, Y. W. Liu, B. Manil, N. Nelms, L. M. Simons, A. Wells, Measurement of the charged pion mass using x-ray spectroscopy of exotic atoms, Phys. Lett. B 759 (2016) 583–588. doi:http://dx.doi.org/10.1016/j.physletb.2016.06.025.
URL http://www.sciencedirect.com/science/article/pii/S0370269316

[31] M. Trassinelli, D. F. Anagnostopoulos, G. Borchert, A. Dax, J. P. Egger, D. Gotta, M. Hennebach, P. Indelicato, Y. W. Liu, B. Manil, N. Nelms, L. M. Simons, A. Wells, Measurement of the charged pion mass using a low-density target of light atoms, EPJ web conf. (2016) accepted.

[32] D. S. Sivia, C. J. Carlile, Molecular spectroscopy and bayesian spectral analysis—hov J. Chem. Phys. 96 (1) (1992) 170–178. doi:doi:http://dx.doi.org/10.1063/1.462505.
URL http://scitation.aip.org/content/aip/journal/jcp/96/1/10.106

[33] Y. Kalambet, Y. Kozmin, K. Mikhailova, I. Nagaev, P. Tikhonov, Reconstruction of chromatographic peaks using the exponentially modified gaussian function, J. Chemometr. 25 (7) (2011) 352–356. doi:10.1002/cem.1343.

[34] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (4) (1948) 623–656. doi:10.1002/j.1538-7305.1948.tb00917.x.

[35] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

[36] S. Kullback, R. A. Leibler, On information and sufficiency, Ann. Math. Stat. 22 (1951) 79–86. doi:10.1214/aoms/1177729694.
URL http://projecteuclid.org/euclid.aoms/1177729694

[37] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. Van Der Linde, The deviance information criterion: 12 years on, J. R. Stat. Soc. B 76 (3) (2014) 485–493. doi:10.1111/rssb.12062.
URL http://dx.doi.org/10.1111/rssb.12062

[38] J. Skilling, Nested sampling's convergence, AIP Conf. Proc. 1193 (1) (2009) 277–291. doi:doi:http://dx.doi.org/10.1063/1.3275625.

URL http://scitation.aip.org/content/aip/proceeding/aipcp/10.1063/1.3275625
[39] N. Chopin, C. P. Robert, Properties of nested sampling, Biometrika 97 (3) (2010) 741–755. doi:10.1093/biomet/asq021.
URL http://biomet.oxfordjournals.org/content/97/3/741.abstract