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

ARTICLE INFO

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.

© 2017 Elsevier B.V. All rights reserved.

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 arises when different hypotheses are compared, represented for

example by two possible modeling functions F_A and F_B, in view of the acquired data. The determination of the most adapted model describing the data generally done with goodness-of-fit tests like the χ^2-test, the likelihood-ratio test, etc. [1–6]. In the unfortunate case where there is no clear propensity to a unique model and we are interested on the value of a parameter common to all models (as the position of the a peak with undefined shape), no sort of weighted average can be computed from goodness-of-fit test outcomes. To do this, the assignment of a probability $P(\mathcal{M})$ to each model is mandatory, which cannot be calculated in the classical statistics framework.

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 [7,8].

* Corresponding author.
E-mail address: trassinelli@insp.jussieu.fr

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 [9–12] and is commonly called *Bayesian statistics*.

Bayesian methods are routinely used in many fields: cosmology [13–15], particle physics [16], nuclear physics, …. In atomic physics their implementation is still limited (e.g. in atomic interferometry [17,18], quantum information [19], ion trapping [20], ion–matter interaction [21], etc.) with almost no use in atomic spectroscopy, even if in some cases it would be strongly 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 spectroscopy and to introduce the program `Nested_fit` for the calculation of distribution probabilities and related quantities (mean values, standard deviation, confidence intervals, etc.) 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. [11,14,22–24]. 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 and compare them to classical goodness-of-fit tests. 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' theorem and *Bayesian evidence* are present in Section 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 Section 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 Section 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 $P(X)$ 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 [22,25,12,24,23]

$$0 \leqslant P(X|I) \leqslant 1, \tag{1}$$
$$P(X|X,I) = 1, \tag{2}$$
$$P(X|I) + P(\bar{X}|I) = 1, \tag{3}$$
$$P(X,Y|I) = P(X|Y,I) \times P(Y|I). \tag{4}$$

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. For example, I can represent the ensemble of the physics laws describing a certain phenomenon, e.g. the thermodynamics laws, and X, Y can represent two quantitative measurements related to this phenomenon, e.g. two temperature measurements at different times of a cooling body. The joint probability $P(X,Y|I)$ means that both "X AND Y" are true (equivalent to the logical conjunction '∧'). The deduction of these axioms have been obtained for the first time in 1946 by Richard Cox using Boolean logic [22]. 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 [23]

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

Here the symbol '+' in the notation $X + Y$ means 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 $Y_{i+1} - Y_i \to dY$ 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 [9] formulated theorems of conditional probability, and rediscovered in 1774 and further developed by Pierre-Simon Marquis de la Laplace [10].

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 its 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 hypotheses are considered and compared (see next section).