Journal of Physics: Conference Series **325** (2011) 012006 doi:10.1088/1742-6596/325/1/012006

FABADA: a Fitting Algorithm for Bayesian Analysis of DAta

L. C. Pardo[1], M. Rovira-Esteva[1], S. Busch[2], M. D. Ruiz-Martin[1], J. Ll. Tamarit[1]

[1]Grup de Caracterització de Materials, Departament de Física i Enginyieria Nuclear, ETSEIB, Universitat Politècnica de Catalunya, Diagonal 647, 08028 Barcelona, Catalonia, Spain
[2]Physik Department E13 and Forschungs-Neutronenquelle Heinz Maier-Leibnitz (FRM II), Technische Universität München, Lichtenbergstr. 1, 85748 Garching, Germany

Abstract. The fit of data using a mathematical model is the standard way to know if the model describes data correctly and to obtain parameters that describe the physical processes hidden behind the experimental results. This is usually done by means of a χ^2 minimization procedure. Although this procedure is fast and quite reliable for simple models, it has many drawbacks when dealing with complicated problems such as models with many or correlated parameters. We present here a Bayesian method to explore the parameter space guided only by the probability laws underlying the χ^2 figure of merit. The presented method does not get stuck in local minima of the χ^2 landscape as it usually happens with classical minimization procedures. Moreover correlations between parameters are taken into account in a natural way. Finally, parameters are obtained as probability distribution functions so that all the complexity of the parameter space is shown.

1. Introduction

Science is based on the success of an hypothesis to describe experimental results, i. e., is based on the amount of "truth" and "falsity" of an hypothesis when contrasted with experimental results [1]. In order to find a quantitative method to determine this "amount of truth", hypotheses in science should at the end be reduced to a mathematical expression depending on a set of parameters with some physical meaning. The "amount of truth" is then determined by fitting the mathematical model to some experimental data. To quantify that, a figure of merit χ^2 can be defined as

$$\chi^2 = \sum_{k=1}^{n} \frac{(H_k\{P_i\} - D_k)^2}{\sigma_k^2} \tag{1}$$

where n is the number of experimental points, D_k $(k = 1, \ldots, n)$ are the experimental data, $H_k\{P_i\}$ $(k = 1, \ldots, n)$ are the values obtained from our hypothesis (the mathematical model) using the $\{P_i\}$ $(i = 1, \ldots, m)$ set of parameters contained in the model, m is the number of parameters, and σ_k $(k = 1, \ldots, n)$ are the experimental errors associated with the respective measured points D_k.

Data fitting is usually done by minimizing χ^2 (equation 1) using the Levenberg-Marquardt algorithm, which aims to find the minimum of the $\chi^2\{P_i\}$ hypersurface. This fit procedure has a twofold goal: first, to find the set of parameters $\{P_i\}$ which best describes the experimental data within their errors, and second, using this set of parameters, to define a figure of merit

which quantifies the "amount of truth" of the proposed hypothesis, taking into account how well it describes the data. In order to be able to compare different hypotheses with different numbers of parameters, it is reasonable to define a figure of merit which penalizes the addition of parameters such as the *reduced* χ^2 defined as $\chi^2_\nu = \frac{\chi^2}{n-m}$. In this equation, n is the number of experimental points and m is the number of parameters, so $n-m$ is the number of degrees of freedom.

This way to quantify how well experimental data are described by a hypothesis is based on what is called a frequentist approximation of the problem [2], and has many drawbacks associated with both the fit procedure (it usually gets stuck in local minima of the χ^2 hypersurface when the model is complex) and the way to quantify the correctness of the hypothesis describing experimental data. The final result using this method is characterized by a set of parameters with an associated error $(P_k \pm \varepsilon_k)$ and the figure of merit χ^2_ν. This way of quantifying the best fit to the data is based on the supposition that there is only one minimum in the $\chi^2\{P_i\}$ hypersurface within the data error, and that the functional dependence of $\chi^2\{P_i\}$ is quadratic on each parameter i (i. e., one can stop at the second term of a Taylor expansion of the obtained minimum), and thus allowing only symmetric errors. Moreover, errors are usually calculated disregarding possible correlations between parameters and are thus generally underestimated.

The main difference of Bayesian inference from the previously exposed frequentist method is the absence of *any* supposition on the $\chi^2\{P_i\}$ landscape which will rather be explicitly explored taking into account experimental data. The method results in a different way to express fitted parameters and the figure of merit showing all the complexity of the final solution: they become Probability Distribution Functions (PDFs) obtained directly from exploring the $\chi^2\{P_i\}$ hypersurface.

Although Bayesian methods are widely used in astronomy or biology [3], they are scarcely used in condensed matter and usually for very specific tasks such as in the analysis of QENS data [4], and the analysis of diffraction data [5, 6]. We present in this work a general method to perform fittings and to analyze results based exclusively on probability by using Bayesian inference.

Although the presented Bayesian method is general, it is specially useful in three situations. Firstly, when the classical fitting procedure gets stuck in a local minimum of the chi squared hypersurface, i.e. when the present parameter set does not correspond to the best obtainable fit but any small parameter value change even decreases the fit quality. This may happen for example when fitting the intramolecular structure to diffraction data [7] but is a well known problem in basically every fit normally surpassed by a careful choice of the initial parameter values. Secondly when an intricate model selection shall be performed, such as in the case of models that describe molecular motions using QENS [8] or dielectric data [9]. Finally when the model is ill defined and more than one combination of parameters is able to describe data, or when data only allows to limit the range of parameters but not to obtain a best fitting value [10].

2. Data analysis using the Bayesian method

2.1. What is behind the ubiquitous χ^2?

The objective of the so called *Bayesian methods* [4, 14] is to find the probability that a hypothesis is true given some experimental evidence. This is done by taking into account both our prior state of knowledge concerning the hypothesis, and the likelihood that the data is described by the proposed hypothesis. Using probability notation, and considering the case that the experiment consists of a series of data D_k and that the hypothesis is represented by H_k, we can relate the aforementioned probabilities using the Bayes theorem [13, 14]:

$$P(H_k \,|\, D_k) = \frac{P(D_k \,|\, H_k) P(H_k)}{P(D_k)} \tag{2}$$

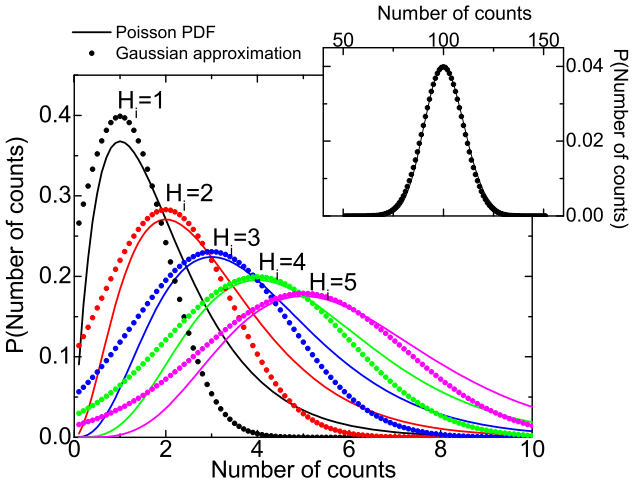

Figure 1. Poisson statistics followed by data in a counting experiment such as a scattering one (lines), and its usual Gaussian approximation (points) for an increasing number of expected counts. For an increasing number of counts, the Poisson distribution (line) can be approximated by a Gaussian function (points) with $\sigma = \sqrt{n}$, being the number of counts n. The inset shows that for a number of counts as great as 100 the approximation works quite well.

where $P(H_k\,|\,D_k)$ is called the *posterior*, the probability that the hypothesis is in fact describing the data. $P(D_k \mid H_k)$ is the likelihood, the probability that our data is well described by our hypothesis. $P(H_k)$ is called the prior, the PDF summarizing the knowledge we have beforehand about the hypothesis, and $P(D_k)$ is a normalization factor to assure that the integrated posterior probability is unity. In the following, we will assume no prior knowledge (maximum ignorance prior [14]), and in this special case Bayes theorem takes the simple form $P(H_k\,|\,D_k) \propto P(D_k\,|\,H_k) \equiv L$, where L is a short notation for likelihood.

In order to quantify the Bayes theorem, we need first to find the likelihood that one data point D_k is described by the mathematically modeled hypothesis H_k. In a counting experiment, this probability follows a Poisson distribution. It can be well approximated by a Gaussian distribution with $\sigma = \sqrt{D_k}$ (see also [14]) if the number of counts is high enough as it is shown in figure 1. Therefore for *one* experimental point ($k = i$, $i = 1, \ldots, n$):

$$P(D_{k=i}\,|\,H_{k=i}) = \frac{{H_k}^{D_k}\, e^{-H_k}}{D_k!} \approx \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{H_k - D_k}{\sigma_k}\right)^2\right] \tag{3}$$

Where on the right hand side of the expression it is not explicitly written that the equation is related to a single experimental point $i = k$ for simplicity. The likelihood that the set of data points D_k is correctly described by the hypothesis H_k can be therefore written as

$$P(D_k\,|\,H_k) \propto \prod_{k=1}^{n} \exp\left[-\frac{1}{2}\left(\frac{H_k - D_k}{\sigma_k}\right)^2\right] = \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\left(\frac{H_k - D_k}{\sigma_k}\right)^2\right] = \exp\left(-\frac{\chi 2}{2}\right) . \tag{4}$$

The figure of merit χ^2 is therefore related to the likelihood that the data is well described by the hypothesis H_k. The probability theory behind χ^2 also allows to deal with the case of experiments with only few counts where the Gaussian approximation, for which $\chi^2 = -2\ln L$, is not valid anymore and the Poisson distribution must be employed, simply by redefining χ^2 [14] as

$$\chi^2 = -2 \cdot \sum_{k=1}^{n} \ln\left[\frac{{H_k}^{D_k} e^{-H_k}}{D_k!}\right] \tag{5}$$

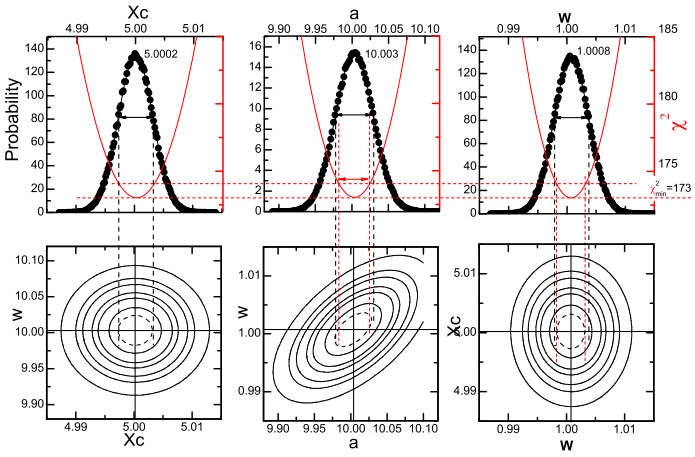

Figure 2. Upper row: PDFs associated with the center x_c, amplitude a and width w of a Gaussian function (solid circles) together with the $\chi^2(P_i)$ around its minimum value, fixing all parameters except P_i (lines). Bottom row: $\chi^2(P_i, P_j)$ plots showing the correlation between parameters, the contour lines have a distance of $\Delta\chi^2 = 1$.

2.2. The Bayesian method

The probabilistic understanding of χ^2 makes it possible to define a unique method, first to fit the experimental data, and then to analyze the obtained results, using a Markov Chain Monte Carlo (MCMC) technique where a set of parameters P_i^{new} is generated from an old set P_i^{old} by randomly changing one of the parameters, i. e., $P_i^{\text{new}} = P_i^{\text{old}} + (\text{RND} - 0.5) \cdot 2\Delta P_i^{\max}$. In the last equation $\Delta P_i^{\max}$ is the maximum change allowed for the parameter and will be called parameter jump for short, and RND is a random number between 0 and 1. The probability to accept the new set of parameters is given by

$$\frac{P(H(P_i^{\text{new}})\,|\,D_k)}{P(H(P_i^{\text{old}})\,|\,D_k)} = \exp\left(-\frac{\chi^2_{\text{new}} - \chi^2_{\text{old}}}{2}\right) \tag{6}$$

where χ^2_{new} and χ^2_{old} correspond to the χ^2 (as defined in equation 1) for the new and old set of parameters. Both fitting and analysis consist therefore in the successive generation of parameter sets $\{P_i\}$ (Markov Chains) with the successive acceptations ruled by equation 6.

3. Two academic examples

3.1. Fitting a Gaussian

In order to test the fit algorithm, a standard function such as a Gaussian

$$y(x) = a/\sqrt{2\pi w}\exp\left[-(x - x_c)^2/(2w^2)\right] + b \tag{7}$$

was generated with the parameter set $\{a, w, x_c\} = \{10.0, 1.0, 5.0\}$ and being b fixed to zero. The data were generated with a Normal distributed error associated with each point of 0.05 and subsequently fitted by the presented algorithm using the same formula (with $b = 0$).

The calculated PDFs associated with each parameter P_i are shown in the top row of figure 2 together with the χ^2 dependence on this parameter, calculated by varying only the parameter P_i and leaving the others fixed, i. e., making a cut of the hypersurface $\chi^2\{a, w, x_c\}$. As one may expect, the minimum of χ^2 coincides with the maximum probability of each parameter PDF.

The most probable parameter *values* – the ones where the PDF is maximal – coincide very nicely with the original values as can be seen in the top row of figure 2. In the following, the discussion will focus on the determination of the parameter *errors*. There are two ways to determine the parameter errors: (i) the commonly used definition of the error as the value of the parameter that increases χ^2 by one unit ($\Delta P = |P(\chi^2_{\min}) - P(\chi^2_{\min} + 1)|$) and (ii) the width of the Gaussian associated with the PDF (the width of the Gaussian at $y = e^{-0.5} \cdot y^{\max}$).

The obtained PDFs can be well described by a Gaussian function in the present examples (solid line in the PDFs shown in figure 2). This proves that in this simple case the minimum of $\chi^2\{P_i\}$ is quadratic in each parameter, and therefore the frequentist definitions of errors can be used – the two measures of the parameter errors should coincide. The error is the defined in such a way that P_i has a 68% probability to be within $P_i - \sigma_{P_i}$ and $P_i + \sigma_{P_i}$ (see [15]).

In figure 2 it can be seen that errors calculated from the PDFs are equal (for x_c) or larger (for a and w) than those calculated using the method of incrementing χ^2. This discrepancy can be explained by parameter correlations seen in the contour plots of the two-dimensional cuts through the $\chi^2(P_i, P_j)$ hypersurface shown in the bottom row of figure 2: from the symmetry of the contours involving x_c it can be concluded that x_c is independent from both, a and w, whereas these two parameters are correlated, causing the main axis of the contour ellipsoids to be not parallel to the parameter axes. For the parameters a and w, the error calculated from the PDF coincides with the limits of the contour $\chi^2 = \chi^2_{\min} + 1$. The error calculated from χ^2 coincides with the intersection of the contour with the x axis, thus underestimating its value.

Although this fact is well known in the frequentist approximation [15], to take correlations between parameters into account would involve diagonalizing the covariant matrix. That is scarcely done and in any case is useful only in simple cases as the one presented when errors are symmetric, i. e., when $\chi^2\{P_i\}$ is quadratic in P_i. The Bayesian approach takes in a natural way any correlation between parameters into account, and can also treat non-Gaussian PDFs being much more powerful than the frequentist approximation.

3.2. Fitting with Poisson statistics

The standard way of fitting data using the minimization of χ^2 (as defined in equation 4) is no longer valid when the number of counts is low. However, as we have seen, simply by redefining χ^2 using equation 5, we can perform the fits when the number of counts is arbitrarily low.

To test our algorithm we have generated a series of random numbers D_K following a Poisson distribution around different fixed values H_K. We have then fitted these series of randomly generated points using the usual definition of χ^2, equation 4, therefore wrongly assuming that the numbers were generated following a Gaussian PDF (see Fig. 1) and with the definition given in equation 5, that is, correctly assuming that D_K follows a Poisson PDF around H_K.

In figure 3 we show the relative discrepancy between the fitted value and the value H_K used to generate the series of points using both methods as a function of the value H_K on a logarithmic scale. As expected, for H_K greater than about 10^3 both methods yield the same result. On the contrary for smaller values the discrepancy increases, reaching 30% for $H_K = 1$.

It is therefore important to take into account that in the limit of low counts the usual approximation between Poisson and Gaussian statistics should not be used. In figure 3 it is displayed the fit of a Gaussian function (equation 7) with parameters $\{a, w, x_c, b\} = \{20, 5, 1, 2\}$ each point D_K being generated following a Poisson distribution. The fit using Poisson statistics is closer to the generated function, i. e., unaffected by the error, as it can be seen in the figure proving that the proposed algorithm is also useful to fit in the case of low count rates.

4. Conclusions

We have proposed a general Bayesian method to fit data and analyze results from the fit [12]. The classical frequentist approach makes some assumptions concerning the χ^2 landscape: there is only a minimum of $\chi^2\{P_i\}$ able to describe data within its error, this minimum has a quadratic dependence on the parameters, and the parameters are not correlated. The here proposed method avoids such problems, sampling the parameter space only with the guide of probability rules. This method has already been successfully used to analyze experiments coming from diffraction experiments [16], quasielastic neutron scattering [17, 10, 8] and dielectric spectroscopy [9]. We finally summarize the main advantages of the proposed method:

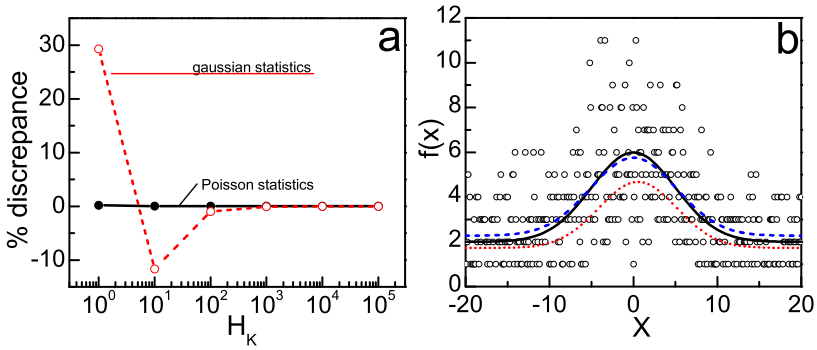

Figure 3. (a) Discrepancy between the fits to a constant value H_K set to 10^i, $i = 0, 1, \ldots, 5$ using a Gaussian (empty circles) and a Poisson statistic (full circles), being D_K values generated using a Poisson statistics. (b) Gaussian function generated assuming a low-count experiment (circles), i. e., assuming a Poisson statistics for each point. The solid line is the generated function unaffected by the error, and the dashed line the fitted function. Dotted line is the fit assuming a normal distribution of errors.

The Bayesian method will not get stuck in local minima of the χ^2 hypersurface during the fit procedure if its barrier is smaller than the error associated with the experimental data set [16].

Parameters are obtained as PDFs and, because the whole parameter space is sampled, correlations between parameters are taken into account. Moreover, a natural way to define errors based on the PDF of parameters is obtained within this method [17, 10, 9].

The likelihood (which as we have seen is directly related to χ^2) obtained by this method is also a PDF hence revealing the whole complexity of the parameter landscape. Model selection is then performed taking into account all parameter combinations compatible with the experiment [8].

This work was supported by the Spanish *Ministerio de Ciencia e Innovación* (FIS2008-00837) and by the Catalonia government (2009SGR-1251). We would also like to thank helpful comments on the manuscript made from K. Kretschmer, T. Unruh and I. Pereyra.

References

[1] Popper K R 2002 *Conjectures and Refutations: The Growth of Scientific Knowledge* ed. Routledge Classics
[2] The frequentist description defines probability of a certain event A ($P(A)$) as the limiting frequency with which the event A is observed when a great number of events A is taken into account.
[3] Trotta R 2008 *Contemp. Phys.* **49(2)** 71; Huelsenbeck J P, Ronquist F, Nielsen R and Bollback J P 2001 *Science* **294** 5550; McVean G A T, Myers S R, Hunt S, Deloukas P, Bentley D R and Donnelly P 2004 *Science* **304** 5670; Diehl R et al. 2006 *A&A* **449** 1025
[4] Sivia D, Carlile C J and Howells W S 1992 *Physica B* 182 **4** 341
[5] Bermejo F J, Santoro J, Mompean F J and Dore J C 1988 *Nucl. Instrum. Meth. Phys. Res. B*, **34** 505; Bacallado S, Chodera J D and Pande V 2009 *J. Chem. Phys.* **131(4)** 45106
[6] O Gereben, P Jóvári, L Temleitner, L Pusztai 2007 J. *Optoelectron. Adv. Mater.* 9 **10** 3021; Sz Pothoczki, L Temleitner, P Jóvári, S Kohara and L Pusztai 2009 *J. Chem. Phys.* 130 064503
[7] Adyaa A K and Bianchi L 2000 *J. Chem. Phys.* 112**9** 4231; Leclercq F, Damay P, Foukani M, Chieux P, Bellisent-Funel M C, Rassat A, Fabre C 1993 *Phys. Rev. B* 48**4** 2748
[8] Busch S, Smuda C, Pardo L C and Unruh T 2010 *J. Am. Chem. Soc.* **132(10)** 3232

[9] Martinez-Garcia J C, Tamarit J Ll, Pardo L C, Barrio M, Rzoska S J and Droz-Rzoska A 2010 *J. Phys. Chem. B* **114** 6099
[10] Pardo L C, Rovira-Esteva M, Busch S, Ruiz-Martin M D, Tamarit J Ll and Unruh T 2009 Bayesian Analysis of QENS data: From parameter determination to model selection *Preprint* arXiv:0907.3711
[11] Hastings W K 1970 *Biometrika* **57(1)** 97
[12] FABADA program (Fit Algortihm for Bayesian Analysis of DAta) can be found in http://gcm.upc.edu/members/luis-carlos/bayesiano
[13] Bayes T 1764 *Phil. trans. Roy. Soc.* **53** 370
[14] Sivia D 2006 *Data Analysis: A bayesian tutorial* Oxford University Press
[15] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 *Numerical Recipes in Fortran 77: the art of scientific computing* Ed. Cambridge University Press, 2nd edition
[16] Rovira-Esteva M et al. Interplay of intramolecular and intramolecular structure: a Bayesian analysis *To be published*
[17] Rovira-Esteva M et al. 2010 *Phys. Rev. B* **81(9)** 092202