

Cosmological parameter inference with Bayesian statistics

Luis E. Padilla,^{1,2,*} Luis O. Tellez,¹ Luis A. Escamilla,¹ and J. Alberto Vazquez^{3,1,†}

¹*Departamento de Física, Centro de Investigación y de Estudios Avanzados del IPN,
A.P. 14-740, 07000 México D.F., México.*

²*Department of Astronomy and Texas Cosmology Center,
University of Texas, Austin, TX, 78712-1083, U.S.A.*

³*Instituto de Ciencias Físicas, Universidad Nacional Autónoma de México,
Apdo. Postal 48-3, 62251 Cuernavaca, Morelos, México.*

(Dated: March 28, 2019)

Bayesian statistics and MCMC (Markov Chain Monte Carlo) algorithms have found their place in the field of Cosmology. They have become important mathematical and numerical tools, especially in parameter estimation and model comparison. In this paper we review some of the fundamental concepts to understand Bayesian statistics, to then introduce the MCMC algorithms and samplers that allow us to perform the parameter inference procedure. We also provide a general description of the standard cosmological model, known as the Λ CDM model, along with several alternatives to it; and current datasets coming from astrophysical and cosmological observations. Finally, with the tools acquired we use a MCMC algorithm implemented in python -called SimpleMC- to test the cosmological models and find out the combination of parameters that best describes the universe.

* epadilla@fis.cinvestav.mx

† javazquez@icf.unam.mx

1. INTRODUCTION

The beginning of the standard cosmology as it is known today emerged after 1920 when the Shapley-Curtis debate was carried out [1]. This debate was held between the astronomers Harlow Shapley and Heber Curtis, resulting in a revolution for astronomy at that time by reaching an important conclusion: “The universe had a larger scale than the Milky Way”. Several observations at that epoch established that the size and dynamics of the cosmos could be explained by Einstein’s General Theory of Relativity. In its childhood, cosmology was a speculative science based only on a few data sets, and it was characterized by a dispute between two cosmological models: the steady state model and the Big Bang (BB) theory. It was not until 1990 when the amount of data increased enough to discriminate and rule out compelling theories, being the BB model awarded as the most accepted theory. During the same decade, David Schramm heralded the “Golden Age of Cosmology” at a National Academy of Sciences colloquium [2].

Once the new age of cosmological observations arrived with a large variety of data, it was necessary to confront the cosmological models with such data. This was usually done through statistics. It is important to notice that, since we have only one universe, we cannot rely on a frequentist interpretation of statistics (we are not able to create multiple universes and make a frequentist inference of our models). An alternative interpretation to help us in our task is Bayesian statistics. In Bayesian statistics the probability is interpreted as a “degree of belief” and it may be useful when repetitive processes are complicated to reproduce.

The main aim of this work is to provide an introduction of Bayesian parameter inference and its applications to cosmology. We assume the reader is familiarized with the basic concepts of statistics, but not necessarily with Bayesian statistics. Then, we provide a general introduction to this subject, enough to work out some examples. This review is written in a generic way so the reader interested on the parameter inference may apply the theory to any subject, in particular we put into practice the Bayesian concepts on the cosmology branch of physics.

The paper is organized as follows. We start in Section 2 by pointing out the main differences between the Bayesian and Frequentist approaches to statistics. Then, in Section 3 we explain the basic and necessary mathematical concepts in Bayesian statistics to perform the parameter estimation procedure for a given model. Once we have the mathematical background, we continue in Section 4 with some of the numerical resources available to simplify our task, such numerical tools may become important given the fact that, in general, it is not possible to derive analytical results, specially when a model contains several parameters that need to be confronted with data. We then provide an example of some of these methods and tools applied to the simple problem of fitting a straight line in Section 5. In Section 6 we present an introduction to cosmology and applications of the tools given in previous sections in cosmology, then in Section 7 the focus is in some of the codes available to perform this work. Then, in Section 8, we apply these techniques to constrain the parameter space that describes the standard cosmological model, namely the Λ CDM model, and several alternatives to it. Finally, in Section 9 we present our conclusions.

2. BAYESIAN VS FREQUENTIST STATISTICS

Fundamentally, the main difference between Bayesian and Frequentist statistics is on the definition of probability. From a Frequentist point of view, probability has meaning in limiting cases of repeated measurements

$$P = \frac{n}{N}, \quad (1)$$

where n denotes the number of successes and N the total number of trials. Frequentist statistics defines probability as the limit for the number of independent trials going to infinity. Then, **for Frequentist statistics, probabilities are fundamentally related to frequencies of events.** On the other hand, in Bayesian statistics the concept of probability is extended to cover degrees of certainty about a statement. **For Bayesian statistics, probabilities are fundamentally related to our knowledge about an event.**

Here we introduce some key concepts to understand the consequences this difference entails; for an extended review see [3–7]. Let x be a random variable related to a particular event and $P(x)$ its corresponding probability distribution, for both cases the same rules of probabilities apply¹:

$$P(x) \geq 0, \quad (2a)$$

$$\int_{-\infty}^{\infty} dx P(x) = 1. \quad (2b)$$

For *mutually exclusive* events we have

$$P(x_1 \cup x_2) = P(x_1) + P(x_2), \quad (2c)$$

but in general

$$P(x_1 \cup x_2) = P(x_1) + P(x_2) - P(x_1 \cap x_2).$$

These rules are summed up as follow. The first condition (2a) is necessary due to the probability of having an event is always positive. The second rule (2b) is a normalized relation, which tells us that we are certain to obtain one of the possible outcomes. Now, in the third point (2c) we have that the probability of obtaining an observation, from a set of mutually exclusive events, is given by the individual probabilities of each event. Finally, and in general, if one event occurs *given* the occurrence of another then the probability that both x_1 and x_2 happen is equal to the probability of x_1 times the probability of x_2 given that x_1 has already happened

$$P(x_1 \cap x_2) = P(x_1)P(x_2|x_1). \quad (2d)$$

If two events x_1 and x_2 are mutually exclusive then

$$P(x_1 \cap x_2) = 0 = P(x_2 \cap x_1). \quad (3)$$

The rules of probability distributions must be fulfilled by both Frequentist and Bayesian statistics. However, there are some consequences derived by the fact these two scenarios have a different definition of probability, as we shall see.

¹ These rules are defined for a continuous variable; however, the corresponding discrete definition can be given immediately by replacing $\int dx \rightarrow \sum$.

2.1. Frequentist statistics

Any frequentist inferential procedure relies on three basic ingredients: the data, the model and an estimation procedure. The main assumption in Frequentist statistics is that the data has a definite, albeit unknown, underlying distribution to which all inference pertains.

The **data** is a measurement or observation, denoted by X , that can take any value from a corresponding sample space. A **sample space** of an observation X can be defined as a measurable space (x, \hat{B}) that contains all values that X can take upon measurement. In Frequentist statistics it is considered that there is a probability function $P_0 : \hat{B} \rightarrow [0, 1]$ in the sample space (x, \hat{B}) representing the “true distribution of the data”

$$X \sim P_0.$$

Now there is the model. For Frequentist statistics the **model** Q is a collection of probability measurements $P_\theta : \hat{B} \rightarrow [0, 1]$ in the sample space (x, \hat{B}) . The distributions P_θ are called *model distributions*, with θ as the model parameters; in this approach θ is unchanged. A model Q is said to be well-specified if it contains the true distribution of the data P_0 , i.e.

$$P_0 \in Q.$$

Finally, we need a point-estimator (or estimator) for P_0 . An *estimator* for P_0 is a map $\hat{P} : x \rightarrow Q$, representing our “best guess” $\hat{P} \in Q$ for P_0 based on the data X .

Hence, the Frequentist statistics is based on trying to answer the following questions: “what the data is trying to tell us about P_0 ?” or “considering the data, what can we say about the mean value of P_0 ?”.

2.2. Bayesian statistics

In Bayesian statistics, data and model are two elements of the same space [3], i.e. no formal distinction is made between measured quantities X and parameters θ . One may envisage the process of generating a measurement’s outcome $Y = y$ as two draws, one draw for Θ (where Θ is a model with associated probabilities to the parameter θ) to select a value of θ and a subsequent draw for P_θ to arrive at $X = x$. This perspective may seem rather absurd in view of the definitions for a Frequentist way of thinking, but in Bayesian statistics where probabilities are related to our own knowledge, it results natural to associate probability distributions to our parameters. In this way an element P_θ of the model is interpreted simply as the distribution of X given the parameter value θ , i.e. as the conditional distribution $X|\theta$.

2.3. Comparing both descriptions

Table I provides a short summary of the most important differences between the two statistics. To understand these differences let us review a typical example. Here we present an experiment and, since we are interested in comparing both descriptions, we show only the basic results from both points of view: Frequentist and Bayesian.

Example.- Let us assume we have a coin that has a probability p to land as heads and a probability $1 - p$ to land as tails. In the process of trying to estimate p (which must be $p = 0.5$ since we have only two possible states) we flip the coin 14 times, obtaining heads in 10 of the trials. Now we are interested in the next two possible events. To be precise: “What is the probability that in the next two tosses we will get two heads in a row?”.

Frequentist	Bayesian
Data are a repeatable random sample. There is a frequency.	Data are observed from the realized sample.
Underlying parameters remain constant during this repeatable process.	Parameters are unknown and described probabilistically.
Parameters are fixed.	Data are fixed.

TABLE I: Main differences between the Bayesian and Frequentist interpretations.

- *Frequentist approach.* As mentioned previously, in Frequentist statistics probability is related to the frequency of events, then our best estimate for p is $P(\text{head}) = p = \frac{\#\text{ of heads}}{\#\text{ of events}} = 10/14$. So, the probability of having 2 heads in a row is $P(2\text{heads}) = P(\text{head})P(\text{head}) \simeq 0.51$.
- *Bayesian approach.* In Bayesian statistics p is not a value, it is a random variable with its own distribution, and it must be defined by the existing evidence. In this example a good distribution for p is a binomial distribution

$$P(D|p) = \binom{14}{10} p^{10} (1-p)^4, \quad (4)$$

where D is our data set (14 trials and 10 successes). Then, by considering a non-informative prior (beforehand we do not know anything about p) and averaging over all possible values of p we have that the probability of having two heads is

$$P(2\text{heads}|D) = \frac{B(13, 5)}{B(11, 5)} = 0.485, \quad (5)$$

where $B(x, y)$ is the beta function. This Bayesian example will be expanded in detail during the following section, but for now we just want to stress out that both approximations arrive at different results.

In the Frequentist approach, since we adopt the probability as a frequency of events (the probability of having a head was fixed by $p = 10/14$), hence the final result was obtained by only multiplying each of these probabilities (since we assume the events are independent of each other). On the other hand, in the Bayesian framework it was necessary to average over all possible values of p in order to obtain a numerical value. However, in both cases, the probability differs from the real one ($P(2\text{heads}) = 0.25$) because we don't have enough data for our estimations.

Note: If you are unfamiliar with Bayesian statistics, do not be scared of the last example. In the next section we review the basic concepts and get back to this example to use the new tools learned.

3. A FIRST LOOK AT BAYESIAN STATISTICS

Before we start with the applications of Bayesian statistics in cosmology it is necessary to understand the most important mathematical tools in the Bayesian procedure. In this section, we present an informal revision but encourage the reader to look for the formal treatment in the literature, cited in each section.

3.1. Bayes theorem, priors, posteriors and all that stuff

When anyone is interested on the Bayesian framework, there are several concepts to understand before presenting the results. In this section we quickly review these concepts and then we take back the example about the coin toss given in the last section.

The Bayes theorem. The Bayes theorem is a direct consequence of the axioms of probability shown in Eqs. (2). From Eqn. (2d), without loss of generality, it must be fulfilled that $P(x_1 \cap x_2) = P(x_2 \cap x_1)$. In such case the following relation applies

$$P(x_2|x_1) = \frac{P(x_1|x_2)P(x_2)}{P(x_1)}. \quad (6)$$

As already mentioned, in the Bayesian framework data and model are part of the same space. Given a model (or hypothesis) H , considering $x_1 \rightarrow D$ as a set of data, and $x_2 \rightarrow \theta$ as the parameter vector of said hypothesis, we can rewrite the above equation as

$$P(\theta|D, H) = \frac{P(D|\theta, H)P(\theta|H)}{P(D|H)}. \quad (7)$$

This last relation is the so-called **Bayes theorem** and the most important tool in a Bayesian inference procedure. In this result, $P(\theta|D, H)$ is called the **posterior** probability of the model. $P(D|\theta, H) \equiv L(D|\theta, H)$ is called the **likelihood** and it will be our main focus in future sections. $P(\theta|H) \equiv \pi(\theta)$ is called the **prior** and expresses the knowledge about the model before acquiring the data. This prior can be fixed depending on either previous experiment results or the theory behind. $P(D|H) \equiv \mathcal{Z}$ is the evidence of the model, usually referred as the **Bayesian Evidence**. We notice that this evidence acts only as a normalizing factor, and is nothing more than the average of the likelihood over the prior

$$P(D|H) = \int d^N \theta P(D|\theta, H)P(\theta|H), \quad (8)$$

where N is the dimensionality of the parameter space. This quantity is usually ignored, for practical reasons, when testing the parameter space of a unique model. Nevertheless, the Bayesian evidence plays an important role for selecting the model that best “describes” the data, known as *model selection*. For convenience, the ratio of two evidences

$$K \equiv \frac{P(D|H_0)}{P(D|H_1)} = \frac{\int d^{N_0} \theta_0 P(D|\theta_0, H_0)P(\theta_0|H_0)}{\int d^{N_1} \theta_1 P(D|\theta_1, H_1)P(\theta_1|H_1)} = \frac{\mathcal{Z}_0}{\mathcal{Z}_1}, \quad (9)$$

or equivalently the difference in log evidence $\ln \mathcal{Z}_0 - \ln \mathcal{Z}_1$ if often termed as the **Bayes factor** $\mathcal{B}_{0,1}$:

$$\mathcal{B}_{0,1} = \ln \frac{\mathcal{Z}_0}{\mathcal{Z}_1}, \quad (10)$$

where θ_i is a parameter vector (with dimensionality N_i) for the hypothesis H_i and $i = 0, 1$. In Eqn. (10), the quantity $\mathcal{B}_{0,1} = \ln K$ provides an idea on how well model 0 may fit the data when is compared to model 1. Jeffreys provided a suitable guideline scale on which we are able to make qualitative conclusions (see Table II).

We can see that Bayes theorem has an enormous implication with respect to a statistical inferential point of view. In a typical scenario we collect some data and hope to interpret it with a

$ \mathcal{B}_{0,1} $	Odds	Probability	Strength
< 1.0	< 3 : 1	< 0.750	Inconclusive
1.0-2.5	$\sim 12 : 1$	0.923	Significant
2.5-5.0	$\sim 150 : 1$	0.993	Strong
> 5.0	> 150 : 1	> 0.993	Decisive

TABLE II: Jeffreys guideline scale for evaluating the strength of evidence when two models are compared.

given model, however, we usually do the opposite. That is, first we have a set of data and then we can confront a model considering the probability that our model fits the data. Bayes theorem provides a tool to relate both scenarios. Then, thanks to the Bayes theorem, in principle, we are able to select the model that best fits the data.

Example.- We go back to the example shown in the last section: the coin toss. We are interested in the probability of obtaining two heads in a row given the data $P(2\text{heads}|D)$ (D = the previous 14 coin tosses acting as data). First of all let us assume that we have a model with parameter p to define the probability of obtaining the two heads given our model $P(2\text{heads}|p)$. This parameter will have a probability distribution $P(p|D)$ depending on the data we already have. Therefore the probability can be obtained by averaging over all the possible parameters with its corresponding density distribution

$$P(2\text{heads}|D) = \int_0^1 P(2\text{heads}|p)P(p|D)dp. \quad (11)$$

For simplicity we do not update p between the two tosses and we assume that both are independent from each other. With this last assumption we have

$$P(2\text{heads}|p) = [P(\text{head}|p)]^2, \quad (12)$$

where $P(\text{head}|p)$ is the probability of obtaining a head given our model. We assume a simple description of $P(\text{head}|p)$ as

$$P(\text{head}|p) = p \Rightarrow P(2\text{heads}|p) = p^2. \quad (13)$$

On the other hand, notice that we do not know *a priori* the quantity $P(p|D)$ but $P(D|p)$ (i.e. we know the probability of obtaining a dataset by considering a model as correct). A good choice for experiments that have two possible results is a binomial distribution

$$P(x|p, n) = \binom{n}{x} p^x (1-p)^{n-x}, \quad (14)$$

with n the number of trials (this case = 14) and x the number of successes (here =10). Hence, we have an expression for $P(D|p)$ [Eqn. (4)]. Now we need to compute $P(p|D)$. Using the Bayes formula we have

$$P(p|D) = \frac{P(D|p)P(p)}{P(D)}. \quad (15)$$

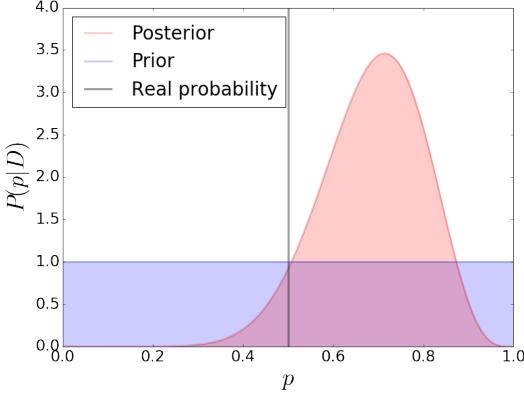


FIG. 1: The coin example: blue figure displays the prior distribution $P(p)$ which is updated, using the data, to get the posterior distribution $P(p|D)$, (red). The vertical black line corresponds to the real value, $p = 0.5$.

A very convenient prior distribution for this scenario is the *beta distribution* $Beta(p; a, b)$ ² defined as

$$Beta(p; a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}, \quad (16)$$

where Γ is the gamma function. So

$$P(p) = Beta(p; a, b). \quad (17)$$

We are interested in the explicit form of $P(p|D)$ and in such case we need to compute $P(D)$. Plugging Eqn. (4) and Eqn. (17) into the integral of Eqn. (8) we have

$$P(D) = B(10 + a, 4 + b) \equiv \frac{\Gamma(10 + a)\Gamma(4 + b)}{\Gamma((10 + a) + (4 + b))}, \quad (18)$$

and therefore

$$P(p|D) = \frac{p^{10+a-1} (1-p)^{4+b-1}}{B(10 + a, 4 + b)}. \quad (19)$$

Now we need to know the values of a and b . If we assume that we know nothing about p , then we can assume the prior as an uniform distribution, this means $a = b = 1$. Notice from Fig. 1 that our posterior result (Red figure) described by Eqn. (19) does not exactly agree with the real value of p (black dashed vertical line). We would expect the posterior distribution be centered at $p = 0.5$ with a very narrow distribution. Nevertheless this value is recovered by increasing the experimental data.

Finally, solving the integral in Eqn. (11) using (13) and (19) we arrive at the result obtained in the previous section

$$P(2\text{heads}|D) = \frac{B(13, 5)}{B(11, 5)} = 0.485. \quad (20)$$

² It is chosen because it describes several statistical distributions, in particular the normal distribution defined as the non-informative one.

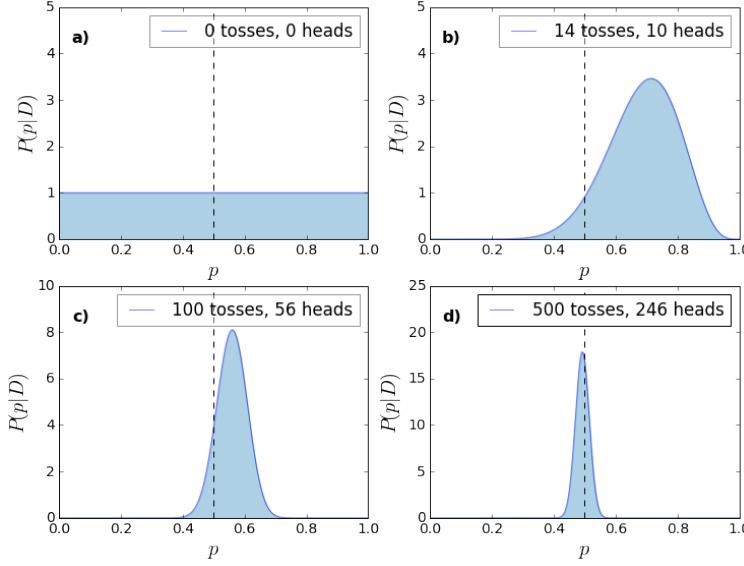


FIG. 2: Posterior distributions $P(p|D)$, when the data is increased. Notice that while we continue increasing the experimental results, the posterior distribution starts to be more localized near by the real value $p = 0.5$.

3.2. Updating the probability distribution

As seen in the coin example, we weren't able to get the real value of p because the lack of enough data. If we want to be closer, we would have to keep flipping the coin until the amount of data becomes sufficient. Let us continue with the example: suppose that after throwing the coin 100 times we obtain, let's say, 56 heads, while after throwing it 500 times we obtain 246 heads. Then, we expect to obtain a thinner distribution with center close to $p = 0.5$ (see Fig. 2). Given this, it is clear that in order to confront a parameter model and be more accurate about the most probable (or "real") value, it is necessary to increase the amount of data (and the precision) in any experiment. That is, if we take into account the 500 tosses – with 246 heads – the previous result is updated to $P(2\text{heads}|D) = 0.249$, much closer to the real value.

Then, we have some model parameters that have to be confronted with different sets of data. This can be done in two alternative ways: (a) by considering the sum of all datasets we have; or (b) by taking each data set as the new data, but our prior information updated by the previous information. The important point in Bayesian statistics is that it is indeed equivalent to choose any of these two possibilities. In the coin toss example it means that it is identical to start with the prior given in Fig. 2-a and then by considering the 500 datapoints we can arrive at the posterior in Fig. 2-d, or similarly start with the posterior shown in Fig. 2-c as our prior and consider only the last 400 datapoints to obtain the same posterior, displayed in Fig. 2-d.

In fact, if we rewrite Bayes theorem so that all probabilities are explicitly dependent on some prior information I [4]

$$P(\theta|DI, H) = \frac{P(\theta|I, H)P(DI|\theta, H)}{P(D|I, H)}, \quad (21)$$

and then we consider a new set of data D' , letting the old data become part of the prior information

$I' = DI$, we arrive at

$$P(\theta|D'I', H) = \frac{P(\theta|I, H)P(DD'I|\theta, H)}{P(DD'|I, H)} = P(\theta|[DD']I, H), \quad (22)$$

where we can explicitly see the equivalence of the two different options.

3.3. About the Likelihood

We mentioned that the Bayesian evidence is usually set apart when doing any inference procedure in the parameter space of a single model. Then, without loss of generality, we can fix it to $P(D|H) = 1$. If we ignore the prior³ we can identify the posterior with the likelihood $P(\theta|D, H) \propto L(D|\theta, H)$ and thus, by maximizing it, we can find the most probable set of parameters for a model given the data. However, having ignored $P(D|H)$ and the prior, we are not able to provide an absolute probability for a given model, but only relative probabilities. On the other hand, it is possible to report results independently of the prior by using the *Likelihood ratio*. The likelihood at a particular point in the parameter space can be compared with the best-fit value, or the maximum likelihood L_{max} . Then, we can say that some parameters are acceptable if the likelihood ratio

$$\Lambda = -2 \ln \left[\frac{L(D|\theta, H)}{L_{max}} \right], \quad (23)$$

is bigger than a given value.

Let us assume we have a Gaussian posterior distribution, which is single-peaked. We consider that $\hat{\theta}$ is the **mean** of the distribution

$$\hat{\theta} = \int d\theta \theta P(\theta|D, H). \quad (24)$$

If our model is well-specified and the expectation value of $\hat{\theta}$ corresponds to the real or most probable value θ_0 , we have

$$\langle \hat{\theta} \rangle = \theta_0, \quad (25)$$

then we say that $\hat{\theta}$ is *unbiased*. Considering a Taylor expansion of the *log likelihood* around its maximum

$$\ln L(D|\theta) = \ln L(D|\theta_0) + \frac{1}{2} (\theta_i - \theta_{0i}) \frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} (\theta_j - \theta_{0j}) + \dots, \quad (26)$$

where θ_0 corresponds to the parameter vector of the real model. In this manner, we have that the likelihood can be expressed as a multi-variable likelihood given by

$$L(D|\theta) = L(D|\theta_0) \exp \left[-\frac{1}{2} (\theta_i - \theta_{0i}) H_{ij} (\theta_j - \theta_{0j}) \right], \quad (27)$$

where

$$H_{ij} = -\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}, \quad (28)$$

³ It is expected that the real value of any given parameter for a large enough dataset is independent of the prior.

is called the **Hessian matrix** and it controls whether the estimates of θ_i and θ_j are correlated. If it is diagonal, these estimates are uncorrelated.

The above expression for the likelihood is a good approximation as long as our posterior distribution possesses a single-peak. It is worth mentioning that, if the data errors are normally distributed, then the likelihood for the data will be a Gaussian function as well. In fact, this is always true if the model is linearly dependent on the parameters. On the other hand, if the data is not normally distributed we can resort to the central limit theorem. In this way, the central limit theorem tell us that the resulting distribution will be best approximated by a multi-variate Gaussian distribution [6].

3.4. Letting aside the priors

In this section we present an argument for letting aside the prior in the parameter estimation. For this, we follow the example given in [5]. In this example there are two people, A and B, that are interested in the measurement of a given physical quantity θ . A and B have different prior beliefs regarding the possible value of θ . This discrepancy could be given by the experience, such as the possibility that A and B have made the same measurement at different times. Let us denote their priors by $P(\theta|I_i)$, ($i = A, B$), and assume they are described by two Gaussian distributions with mean μ_i and variance Σ_i^2 . Now, A and B make a measurement of θ together using an apparatus subject to a Gaussian noise with known variance σ . They obtain the value $\theta_0 = m_1$. Therefore they can write their likelihoods for θ as

$$L(D|\theta, HI) = L_0 \exp \left[-\frac{1}{2} \frac{(\theta - m_1)^2}{\sigma^2} \right]. \quad (29)$$

By using the Bayes formula, the posterior of the model A (and B) becomes

$$P(\theta|m_1) = \frac{L(m_1|\theta I_i)P(\theta|I_i)}{P(m_1|I_i)}, \quad (30)$$

where we have skipped writing explicitly the hypothesis H and used the notation given in Eqn. (21). Then, the posterior of A and B are (again) Gaussian with mean

$$\hat{\mu}_i = \frac{m_1 + (\sigma/\Sigma_i)^2 \mu_i}{1 + (\sigma/\Sigma_i)^2}, \quad (31)$$

and variance

$$\tau_i^2 = \frac{\sigma^2}{1 + (\sigma/\Sigma_i)^2}, \quad (i = A, B). \quad (32)$$

Thus, if the likelihood is more informative than the prior i.e. $(\sigma/\Sigma_i) \ll 1$ the posterior mean of A (and B) will converge towards the measured value, m_1 . As more data are obtained one can simply replace the value of m_1 in the above equation by the mean $\langle m \rangle$ and σ^2 by σ^2/N . Then, we can see that the initial prior μ_i of A and B will progressively be overridden by the data. This process is illustrated in Figure 3 where the green (red) curve corresponds to the probability distribution of θ for person A (B) and the blue curve corresponds to their likelihood.

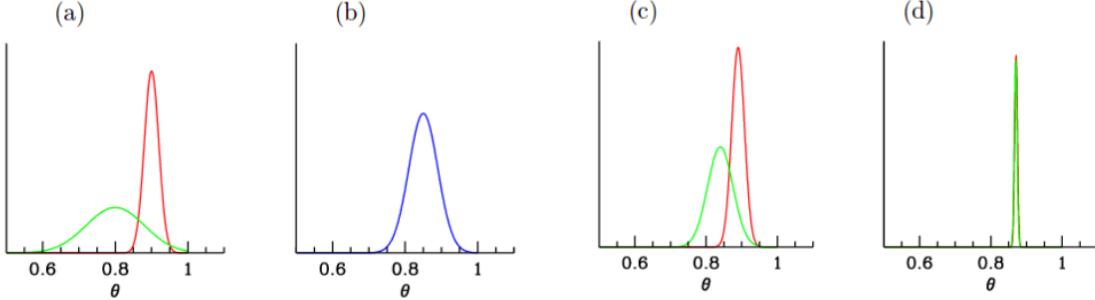


FIG. 3: Converging views in Bayesian inference (taken from [5]). A and B have different priors $P(\theta|I_i)$ for a value θ (panel (a)). Then, they observe one datum with an apparatus subject to a Gaussian noise and they obtained a likelihood $L(\theta; HI)$ (panel (b)), after which their posteriors $P(\theta|m_1)$ are obtained (panel (c)). Then, after observing 100 data, it can be seen how both posteriors are practically indistinguishable (panel (d)).

3.5. Chi-square and goodness of fit

We mentioned the main aim of parameter estimation is to maximize the likelihood in order to obtain the most probable set of model parameters, given the data. If we consider the Gaussian approximation given in Eqn. (27) we can see the likelihood will be maximum if the quantity

$$\chi^2 \equiv (\theta_i - \theta_{0i})H_{ij}(\theta_j - \theta_{0j}), \quad (33)$$

is minimum. The quantity χ^2 is usually called **chi-square** and is related to the Gaussian likelihood via $L = L_0 e^{-\chi^2/2}$. Then, we can say that maximizing the Gaussian likelihood is equivalent to minimizing the chi-square. However, as we mentioned before, there are some circumstances where the likelihood cannot be described by a Gaussian distribution, in these cases the chi-square and the likelihood are no longer equivalent.

The probability distribution for different values of χ^2 around its minimum, is given by the χ^2 distribution for $v = n - M$ degrees of freedom, where n is the number of independent data points and M the number of parameters. Hence, we can calculate the probability that an observed χ^2 exceeds by chance a value $\hat{\chi}^2$ for the correct model. This probability is given by $Q(v, \hat{\chi}^2) = 1 - \Gamma(v/2, \hat{\chi}^2/2)$ [8], where Γ is the incomplete Gamma function. Then, the probability that the observed χ^2 (even the correct model) is less than a given value $\hat{\chi}^2$ is $1 - Q$. This statement is strictly true if the errors are Gaussian and the model is a linear function of the likelihood, i.e., for Gaussian likelihoods.

If we evaluate the quantity Q for the best-fit values (minimum chi-square) we can have a measure of the goodness of fit. If Q is small (small probability) we can interpret it as:

- The model is wrong and can be rejected.
- The errors are underestimated.
- The error measurements are not normally distributed.

On the other hand, if Q is too large there are some reasons to believe that:

- Errors have been overestimated.
- Data are correlated or non-independent.
- The distribution is non-Gaussian.

		$\Delta\chi^2$		
σ	p	$M = 1$	$M = 2$	$M = 3$
1	68.3%	1.00	2.30	3.53
2	95.4%	4.00	6.17	8.02
3	99.73%	9.00	11.8	14.20

TABLE III: $\Delta\chi^2$ for the conventional 68.3%, 95.4% and 99.73% as a function of the number of parameters (M) for the joint confidence level.

3.6. Contour plots and confidence regions

Once the best fit parameters are obtained we would like to know the confidence regions where values could be considered good candidates for our model. The most logical election is to take values inside a compact region around the best fit value. Then, a natural choice are regions with constant χ^2 boundaries. When the χ^2 possesses more than one minimum, it is said that we have non-connected confidence regions, and for multi-variate Gaussian distributions (as the likelihood approximation in Eqn. (27)) these are ellipsoidal regions. In this section we exemplify how to calculate the confidence regions, following [6].

We consider a little perturbation from the best fit of chi-square $\Delta\chi^2 = \chi^2 - \chi_{best}^2$. Then we use the properties of χ^2 distribution to define confidence regions for variations on χ^2 to its minimum. In Table III we see the typical 68.3%, 95.4% and 99.73% confidence levels as a function of number of parameters M for the joint confidence level. For Gaussian distributions (as likelihood) these correspond to the conventional 1, 2 and 3 σ confidence levels. As an example we plot in Figure 4 the corresponding confidence regions associated to the coin example.

The general recipe to compute constant χ^2 confidence regions is as follows: after finding the best fit by minimizing χ^2 (or maximizing the likelihood) and checking that Q is acceptable for the best parameters, then:

1. Let M be the number of parameters, n the number of data and p be the confidence limit desired.
2. Solve the equation:

$$Q(n - M, \min(\chi^2) + \Delta\chi^2) = p. \quad (34)$$

3. Find the parameter region where $\chi^2 \leq \min(\chi^2) + \Delta\chi^2$. This defines the confidence region.

3.7. Marginalization

It is clear that a model may (in general) depend on more than one parameter. However, some of these parameters θ_i may be of less interest. For example, they may correspond to nuisance parameters like calibration factors, or it may be the case that we are interested in only one of the parameter constraints rather than the joint of two or more of them simultaneously. Then we **marginalize** over the uninteresting parameters by

$$P(\theta_1, \dots, \theta_j, H|D) = \int d\theta_{j+1} \dots d\theta_m P(\theta, H|D), \quad (35)$$

where m is the total number of parameters in our model and $\theta_1, \dots, \theta_j$ denote the parameters we are interested in.

3.8. Fisher Matrix

Once we have a dataset it is important to know the accuracy for which we can estimate parameters. Fisher suggested a way 70 years ago [9]. In this section we review the main results of his work.

First of all, consider again a Gaussian likelihood. As we notice, the **Hessian matrix** H_{ij} has information on the parameter errors and their covariance. More specifically, when all parameters are fixed except one (e.g. the i -th parameter), its error is $1/\sqrt{H_{ii}}$. These errors are called conditional errors, although they are rarely used.

A quantity to forecast the precision of a model, that arises naturally with Gaussian likelihoods, is the so-called **Fisher information matrix**

$$F_{ij} = -\left\langle \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} \right\rangle, \quad (36)$$

where

$$\mathcal{L} = \ln L. \quad (37)$$

It is clear that $F = \langle H \rangle$, where the average is made with observational data.

As we can see from Eqn. (2c), for independent data sets the complete likelihood is the product of the likelihoods, and the Fisher matrix is the sum of individual Fisher matrices. A pedagogical and easy case is having one-parameter θ_i with a Gaussian likelihood. In this scenario

$$\Delta \mathcal{L} = \frac{1}{2} F_{ii} (\theta_i - \theta_{0i})^2, \quad (38)$$

when $2\Delta \mathcal{L} = 1$ and identifying the $\Delta \chi^2$ corresponding to 68% confidence level, we notice that $1/\sqrt{F_{ii}}$ yields the $1 - \sigma$ displacement for θ_i . In the general case

$$\sigma_{ij}^2 \geq (F^{-1})_{ij}. \quad (39)$$

Thus, when all parameters are estimated simultaneously from the data, the marginalized error is

$$\sigma_{\theta_i} \geq (F^{-1})_{ii}^{1/2}. \quad (40)$$

The beauty of the Fisher matrix approach is that there is a simple prescription for setting it up by only knowing the model and measurement uncertainties, and under the assumption of a Gaussian likelihood the Fisher matrix is the inverse of the covariance matrix. So, all we have to do is set up the Fisher matrix and then invert it to obtain the covariance matrix (that is, the uncertainties on the model parameters). In addition, its fast calculations also enables one to explore different experimental setups and optimize the experiment.

The main point of the Fisher matrix formalism is to predict how well the experiment will be able to constrain the parameters, of a given model, before doing the experiment and perhaps even without simulating it in any detail. We can then forecast the results of different experiments and look at trade-offs such as precision versus cost. In other words, we can engage in experimental design. The inequality in Eqn. (39) is called the Kramer-Rao inequality. One can see that the

Fisher information matrix represents a lower bound of the errors. Only when the likelihood is normally distributed, the inequality is transformed into an equality. However as we saw in Sec. 3.3 a Gaussian likelihood is only applicable to some circumstances, being generally impossible to be applied, so the key is to have a good understanding of our theoretical model in such a way that we can construct a Gaussian likelihood.

3.8.1. Constructing Fisher Matrices: A simple description

Let us construct Fisher matrices in a simple way. Suppose we have a model that depends of N parameters $\theta_1, \theta_2, \dots, \theta_N$. We consider M observables f_1, f_2, \dots, f_M each one related to the model parameters by some equation $f_i = f_i(\theta_1, \theta_2, \dots, \theta_N)$. Then the elements of the Fisher matrix can be computed as

$$F_{ij} = \sum_k \frac{1}{\sigma_k^2} \frac{\partial f_k}{\partial \theta_i} \frac{\partial f_k}{\partial \theta_j}, \quad (41)$$

where σ_k are the errors associated to each observable and we have considered them Gaussianly distributed.

Here, instead of taking the real data values (which could be unknown) it is possible to recreate the data with a fiducial model. The errors associated to the mock data can be taken as the expected experimental errors, and then be possible to calculate the above expression.

To complement the subject, there is also the **Figure of Merit** used by the Dark Energy Task Force (DETF) [10] which is defined as the reciprocal of the area in the plane enclosing the 95% confidence limit of two parameters. The larger the figure of merit the greater accuracy one has measuring said parameters. As an example let us take a look at Figure 15 and right panel of Figure 16, the area of the error ellipse with only Hubble Data (HD) is clearly bigger than the error ellipse using HD plus several data sets. Then, for this case the figure of merit would be bigger than with only HD data since its area is smaller, making it more accurate for measuring the parameters Ω_m and h . The DETF figure of merit can also be used to see how different experiments break degeneracies. It can also be used to predict accuracy in future experiments (experimental design).

3.9. Importance Sampling

We call **Importance Sampling** (IS) to different techniques of determining properties of a distribution by drawing samples from another one. The main request of this idea, is that the distribution one samples from should be representative of the distribution of interest (for a larger number of samples). In such case, we should infer different quantities out of it. In this section we review the basic concepts necessary to understand the IS, following [11].

Suppose we are interested in computing the expectation value $\mu_f = E_p[f(X)]$, where $f(X)$ is a probability density of a random variable X and the sub-index p means average over the distribution p . Then, if we consider a new probability density $q(x)$ that satisfies $q(x) > 0$ whenever $f(x)p(x) \neq 0$, we can rewrite the mean value μ_f as

$$\mu_f = \int f(x)p(x)dx = \int f(x) \frac{p(x)}{q(x)} q(x)dx = E_q[f(X)w(x)], \quad (42)$$

where $w(x) = p(x)/q(x)$, and now we have an average over q . So, if we have a collection of different

draws $x^{(1)}, \dots, x^{(m)}$ from $q(x)$, we can estimate μ_f using these draws as

$$\hat{\mu}_f = \frac{1}{m} \sum_{j=1}^m w(x^{(j)}) f(x^{(j)}). \quad (43)$$

If $p(x)$ is known only up to a normalizing constant, the above expression can be calculated as a ratio estimate

$$\hat{\mu}_f = \frac{\sum_{j=1}^m w(x^{(j)}) f(x^{(j)})}{\sum_{j=1}^m w(x^{(j)})}. \quad (44)$$

For the strong law of large numbers, in the limit when $m \rightarrow \infty$ we will have that $\hat{\mu}_f \rightarrow \mu_f$.

Another useful quantity to compute in Bayesian analysis is the ratio between evidences for two different models

$$\frac{P'(D)}{P(D)} = E \left[\frac{P'(\theta, D)}{P(\theta, D)} \right]_{P(\theta|D)} \simeq \frac{1}{N} \sum_{n=1}^N \frac{P'(D|\theta_n) P'(\theta_n)}{P(D|\theta_n) P(\theta_n)}, \quad (45)$$

where the samples $\{\theta_n\}$ are drawn from $P(\theta|D)$.

An important result for importance sampling is that, if we have a new set of data which is broadly consistent with the current data (in the sense that the posterior only shrinks), we can make use of importance sampling in order to quickly calculate a new posterior including the new data.

3.10. Combining datasets: Hyperparameter method

Suppose we are dealing with multiple datasets $\{D_1, \dots, D_N\}$, coming from a collection of different surveys $\{S_1, \dots, S_N\}$. Sometimes it is difficult to know, *a priori*, if all our data are consistent with each other, or whether there could be one or more that are likely to be erroneous. If we were sure that all datasets are consistent, then it should be enough to update the probability as seen in Sec. 3.2 in order to calculate the new posterior distribution for the parameters we are interested in. However, since there is usually an uncertainty about this, a way to know how useful a data may be is by introducing the **hyperparameter method**. This method was initially introduced by [12, 13] in order to perform a joint estimation of cosmological parameters from combined datasets. This method may be used as long as every survey is independent from each other. In this section we review the main steps necessary to understand the hyperparameter method. If the reader is interested in a more extended explanation, we encourage to consult [12, 13].

The main feature of this process is the introduction of a new set of “hyperparameters” α in the Bayesian procedure to allow extra freedom in the parameter estimation. These hyperparameters are equivalent to nuisance parameters in the sense that we need to marginalize over them in order to recover the posterior distribution, i.e.

$$P(\theta|D, H) = \frac{1}{P(D|H)} \int P(\theta|\alpha, H) P(\alpha|D, H) d\alpha, \quad (46)$$

where we have used the Bayes theorem. Now, for the method it is necessary to assume the hyperparameters α and the parameters of interest θ are independent, i.e. $P(\theta, \alpha, H) = P(\alpha)P(\theta, H)$,

it is also necessary to assume that each hyperparameter α_k is independent from each other, i.e. $P(\alpha) = P(\alpha_1)P(\alpha_2)\dots P(\alpha_N)$. In this way we can rewrite the above expression as

$$P(\theta|D, H) = \frac{P(\theta, H)}{P(D|H)} \left[\prod_{k=1}^N \int P(D_k|\theta, \alpha_k, H) P(\alpha_k) d\alpha_k \right]. \quad (47)$$

Here, the quantity inside the square brackets is the marginalized likelihood over the hyperparameters. We can identify the quantity inside the integration as the individual likelihood $L(D_k|\theta, \alpha_k, H)$, for every α_k and the data set D_k ; $P(D|H)$ is the evidence and, similarly to a parameter inference procedure, it works as a normalizing function, i.e. $P(D|H) = \int d\theta P(\theta, H) L(D|\theta, H)$. Notice that, by considering $P(\alpha_k) = \delta(\alpha_k - 1)$, we rely on the standard approach, where no hyperparameters are used.

We add these α_k in order to weight every dataset and take away the data that does not seem to be consistent with other ones. Then, we would like to know whether the data supports the introduction of hyperparameters or not. A way to address this point is given by the Bayesian evidence K defined in Eqn. (9). If we consider a Gaussian likelihood with maximum entropy prior, and assuming that in average the hyperparameters' weight are unity, we can rewrite the marginalized likelihood function $L(D|\theta, H_1)$ for model H_1 as

$$P(D|\theta, H_1) = \prod_{k=1}^N \frac{2\Gamma(\frac{n_k}{2} + 1)}{\pi^{n_k/2} |V_k|^{1/2}} (\chi_k^2 + 2)^{-(\frac{n_k}{2} + 1)}, \quad (48)$$

obtaining an explicit functional form for K , given by

$$K = \prod_{k=1}^N \frac{2^{n_k/2+1} \Gamma(n_k/2 + 1)}{\chi_k^2 + 2} e^{-\chi_k^2/2}. \quad (49)$$

Here, χ_k^2 is given by (33) for every dataset and n_k is the number of points contained in D_k . In equation (48) V_k is the covariance matrix for the k -data. Suppose we have two models, one with hyperparameters, called H_1 , and a second one without them, called H_0 . The Bayesian evidence $P(D|H_i)$ is the key quantity for making a comparison between two different models. In fact, by using the Bayes factor K from Eqn. (49) we can estimate the necessity to introduce the hyperparameters to our model using the criteria given in Table II. Notice that, if we have a set of independent samples for H_0 , we can compute an estimate for K with the help of equation (45).

4. NUMERICAL TOOLS

In typical scenarios it results very difficult to compute the posterior distribution analytically. For these cases the numerical tools available play an important role during the parameter estimation task. There exist several options to carry out this work, nevertheless in this section we focus only on the Markov Chain Monte Carlo (MCMC) with the Metropolis Hastings algorithm (MHA). Additionally, in this section we present some useful details we take into account to make more efficient our computation.

4.1. MCMC techniques for parameter inference

The purpose of a MCMC algorithm is to build up a sequence of points (called “**chain**”) in a parameter space in order to evaluate the posterior of Eqn. (7). In this section we review the basic results for this procedure in a simplistic way, but for curious readers it is recommendable to check [14–17] for the Markov chain theory.

A **Monte Carlo** simulation is assigned to algorithms that use random number generators to approximate a specific quantity. On the other hand, a sequence X_1, X_2, \dots of elements of some set is a **Markov Chain** if the conditional distribution of X_{n+1} given X_1, \dots, X_n depends only on X_n . In other words, a Markov Chain is a process where we can compute subsequent steps based only in the information given at the present. An important property of a Markov Chain is that it converges to a stationary state where successive elements of the chain are samples from the target distribution, in our case it converges to the posterior $P(\theta|D, H)$. In this way we can estimate all the usual quantities of interest out of it (mean, variance, etc).

The combination of both procedures is called a **MCMC**. The number of points required to get good estimates in MCMCs is said to scale linearly with the number of parameters, so this method becomes much faster than grids as the dimensionality increases.

The target density is approximated by a set of delta functions

$$p(\theta|D, H) \simeq \frac{1}{N} \sum_{i=1}^N \delta(\theta - \theta_i), \quad (50)$$

being N the number of points in the chain. Then, the posterior mean is computed as

$$\langle \theta \rangle = \int d\theta P(\theta, H|D) \simeq \frac{1}{N} \sum_{i=1}^N \theta_i, \quad (51)$$

where \simeq follows because the samples θ_i are generated out of the posterior by construction. Then, we can estimate any integrals (such as the mean, variance, etc.) as

$$\langle f(\theta) \rangle \simeq \frac{1}{N} \sum_{i=1}^N f(\theta_i). \quad (52)$$

As mentioned before, in a Markov Chain it is necessary to generate a new point θ_{i+1} from the present point θ_i . However, as it is expected, we need a criteria for accepting (or refusing) this new point depending on whether it turns out to be better for our model or not. If this new step is worse than the previous one, we may accept it, since it could be the case that, if we only accept steps with better probability, we could be converging into a local maximum in our parameter space and, therefore, not completely mapping all of it. The simplest algorithm that contains all this information in its methodology is known as the Metropolis-Hastings algorithm.

4.1.1. Metropolis-Hastings algorithm

In the **Metropolis-Hastings algorithm** [18, 19] it is necessary to start from a random initial point θ_i , with an associated posterior probability $p_i = p(\theta_i|D, H)$. We need to propose a candidate θ_c by drawing from a **proposal distribution** $q(\theta_i, \theta_c)$ used as a generator of new random steps. Then, the probability of acceptance the new point is given by

$$p(\text{acceptance}) = \min \left[1, \frac{p_c q(\theta_c, \theta_i)}{p_i q(\theta_i, \theta_c)} \right]. \quad (53)$$

If the proposal distribution is symmetric the algorithm is reduced to the *Metropolis algorithm*

$$p(\text{acceptance}) = \min \left[1, \frac{p_c}{p_i} \right]. \quad (54)$$

In this way the complete algorithm can be expressed by the following steps:

1. Choose a random initial condition θ_i in the parameter space and compute the posterior distribution.
2. Generate a new candidate from a proposal distribution in the parameter space and compute the corresponding posterior distribution.
3. Accept (or not) the new point with the help of the Metropolis-Hastings algorithm.
4. If the point is not accepted, repeat the previous point in the chain.
5. Repeat steps 2-4 until you have a large enough chain.

4.1.2. A first example of parameter inference

In order to exemplify the numerical tools learned in this section, let us go back to the coin toss example seen in Sec. 3.1. Since our main interest is that the reader understands the basic procedure given in this section, let us try to estimate the value of p (or region of values for p) that best matches our data (hence, we assume only the 14 times that the coin was thrown). To calculate the posterior distribution (17) we use the MHA.

As mentioned before, we consider a likelihood given by a binomial distribution (4) and a normal distributed prior (16) ($a = b = 1$). As our first “guess” for p we consider $p_i = 0.1$. We generate a new candidate p_c as $p_c = p_{cu} + G(p_{cu}, \hat{\sigma})$, where $G(p_{cu}, \hat{\sigma})$ is our proposed Gaussian distribution centered at p_{cu} with variance $\hat{\sigma} = 0.1$; p_{cu} is the current value of p , for our first step is $p_{cu} = p_i$. Then, we introduce the MHA in a Python code, as can be seen in Appendix A. Our final result, (shown in Fig. 4), is a posterior distribution that matches very well with the results calculated analytically (shown in Figure 1). Numerically we obtained $p = 0.695^{+0.123}_{-0.107}$, where the upper and lower values for p correspond to the 1σ standard deviation. Notice that we have plotted the width of our 1σ , 2σ and 3σ confidence regions in the same figure.

To complement the example we also show in the right panel of Figure 4 the Markov Chain generated by our code where we have considered 5000 steps in our chain. It is easy to see that the chain oscillates with a large amplitude around a middle value. This amplitude is expected because we do not have enough data to constrain more accurately the value of p .

Remark: In appendix A we include the MCMC algorithm using an explicit code for the MCMC process. However, in Python there are some modules that can simplify this task. For example, PyMC3 [20] is a Python module that implements statistical models and fitting algorithms, including the MCMC algorithm. We use this module at the end of this section by applying the tools already learned.

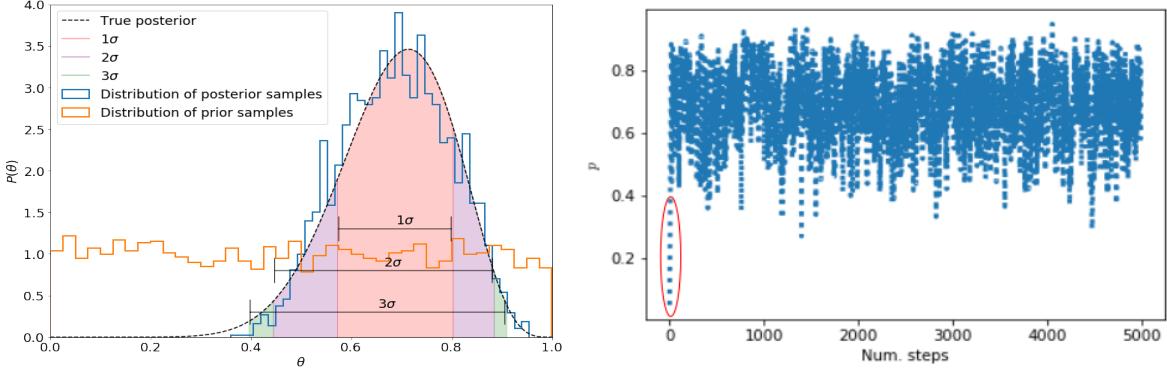


FIG. 4: Left panel: 1D posterior distribution for our example. We plot the prior distribution (red), true posterior (dashed-black) and the posterior calculated by the MHA (blue). We plot 1,2 and 3 σ confidence regions for the estimation of p . Right panel: associated Markov chain. We use $p_i = 0.1$ as our first “guess” for p .

4.1.3. Convergence test

It is clear that we need a test to know when our chains have converged. We need to verify that the points in the chain are not converging to a “false convergent point” or to a local maximum point. In this sense, we need that our algorithm takes into account this possible difficulty. The simplest way (the informal way) to know if our chain is converging to a global maximum is by running several chains starting with different initial proposals for the parameters we are interested in. Then, if we see by naked eye, that all the chains seem to converge into a single region of the possible value for our parameter, we may say that our chains are converging to that region.

Taking yet again the example of the coins, we can run several chains for the above example and try to estimate if the value (region) of p that we found is a stationary value. In Figure 5 we plot 5 different Markov chains with initial “guess” conditions $p = 0.2, 0.3, 0.5, 0.7, 0.9$. As we expected from the analytical result, after several steps all the chains seem to concentrate near by the same value.

The convergence method used above is very informal and we would like to have a better way to ensure that our result is correct. The usual test is the *Gelman-Rubin* convergence criterion [21, 22]. That is, by starting with M chains with very different initial points and N points per chain, if θ_i^j is a point in the parameter space of position i and belonging to the chain j , we need to compute the mean of each chain

$$\langle \theta^j \rangle = \frac{1}{N} \sum_{i=1}^N \theta_i^j, \quad (55)$$

and the mean of all the chains

$$\langle \theta \rangle = \frac{1}{NM} \sum_{i=1}^N \sum_{j=1}^M \theta_i^j. \quad (56)$$

Then, the chain-to-chain variance B is

$$B = \frac{1}{M-1} \sum_{j=1}^M (\langle \theta^j \rangle - \langle \theta \rangle)^2, \quad (57)$$

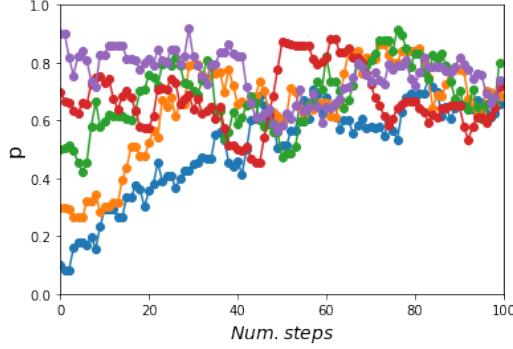


FIG. 5: Multiple MCMC. We use five Markov Chains to estimate the convergence.

and the average variance of each chain is

$$W = \frac{1}{M(N-1)} \sum_{i=1}^N \sum_{j=1}^M (\theta_i^j - \langle \theta^j \rangle)^2. \quad (58)$$

If our chains converge, W and B/N must agree. In fact we say that the chains converge when the quantity

$$\hat{R} = \frac{\frac{N-1}{N}W + B(1 + \frac{1}{M})}{W}, \quad (59)$$

which is the ratio of the two estimates, approaches unity. A typical convergence criteria is when $0.97 < \hat{R} < 1.03$.

4.1.4. Some useful details

The proposal distribution. The choice of a proposal distribution q is crucial for the efficient exploration of the posterior. In our example we used a Gaussian-like distribution with a variance (step) $\hat{\sigma} = 0.1$. This value was taken because we initially explored, by hand, different values for $\hat{\sigma}$ and we select the quickest that approaches the analytic posterior distribution of p . However, if the scale of q is too small compared to the scale of the target (in the sense that the typical jump is small), then the chain may take very long to explore the target distribution which implies that the algorithm will be very inefficient. As we can see in Figure 6 (left panel), considering an initial step $p_i = 0.6$ and a variance for the proposal distribution $\hat{\sigma} = 0.002$, the number of points are not enough for the system to move to its “real” posterior distribution. On the other hand, if the scale of q is too large, the chain gets stuck and it does not jump very frequently (right panel of the figure with $\hat{\sigma} = 0.8$) so we will have different “peaks” in our posterior.

In order to fix this issue in a more efficient way, it is recommendable to run an exploratory MCMC, compute the covariance matrix from the samples, and then re-run with this covariance matrix as the covariance of a multivariate Gaussian proposal distribution. This process can be computed a couple of times before running the “real” MCMC.

The burn-in. It is important to notice that at the beginning of the chain we will have a region of points outside the stationary region (points inside the ellipse in the right panel of Figure 4). This early part of the chain (called “burn-in”) must be ignored, this means that the dependence

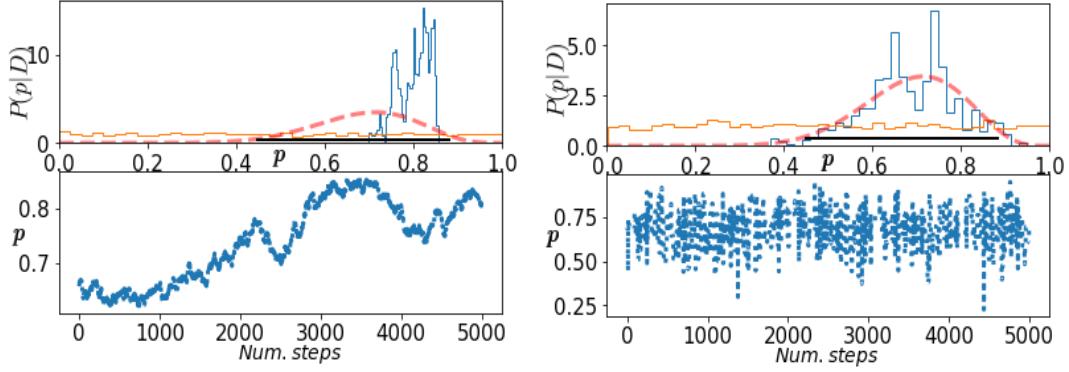


FIG. 6: Two Markov Chains considering different variance for our Gaussian proposal distribution. Left panel corresponds to $\hat{\sigma} = 0.002$, while right panel corresponds to $\hat{\sigma} = 0.8$.

on the starting point must be lost. Thus, it is important to have a reliable convergence test.

Thinning. There are several Bayesian statisticians that usually thin their MCMC, this means that they do not prefer to save every step given by the MCMC; instead, they prefer to save a new step each time n steps have taken place. An obvious consequence of thinning the chains is that the amount of autocorrelation is reduced. However, as long as the chains are thinned, the precision for the estimated parameters is reduced [23]. Thinning the chains can be useful in other kind of circumstances, for example, if we have limitations in memory. Notice that thinning a chain does not yield incorrect results; it yields correct results but less efficient than using the full chains.

Autocorrelation probes. A complementary way to look for convergence in a MCMC estimation is by looking for the autocorrelation between the samples. The autocorrelation *lag* k is defined as the correlation between every sample and the sample k steps before. It can be quantified as [24, 25]

$$\rho_k = \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t)Var(X_{t+k})}} = \frac{E[(X_t - \bar{X})(X_{t+k} - \bar{X})]}{\sqrt{E[(X_t - \bar{X})^2]E[(X_{t+k} - \bar{X})^2]}}, \quad (60)$$

where X_t is the t -th sample and \bar{X} is the mean of the samples. This autocorrelation should become smaller as long as k increases (this means that samples start to become independent).

More samplers

Gibbs sampling. The basic idea of the Gibbs sampling algorithm [26] is to split the multidimensional θ into blocks and sample each block separately, conditional on the most recent values of the other blocks. It basically breaks a high-dimensional problem into low-dimensional problems.

The algorithm reads as follows:

1. θ consists of k blocks $\theta_1, \dots, \theta_k$. Then, at step i
2. Draw θ_1^{i+1} from $p(\theta_1 | \theta_2^i, \dots, \theta_k^i)$
3. Draw θ_2^{i+1} from $p(\theta_2 | \theta_1^{i+1}, \theta_3^i, \dots, \theta_k^i)$
4. ...

5. Draw θ_k^{i+1} from $p(\theta_k | \theta_1^{i+1}, \theta_2^{i+1}, \dots, \theta_{k-1}^{i+1})$
6. Repeat the above steps for the wished iterations with $i \rightarrow i + 1$.

The distribution $p(\theta_1 | \theta_2, \dots, \theta_k) = \frac{p(\theta_1, \dots, \theta_k)}{p(\theta_2, \dots, \theta_k)}$ is known as the *full conditional distribution of θ_1* . This algorithm is a special case of MHA where the proposal is always accepted.

Metropolis Coupled Markov Chain Monte Carlo (MC^3). It is easy to see that it could be a little problematic if our likelihood has local maxima. The MC^3 is a modification of the standard MCMC algorithm that consists of running several Markov Chains in parallel to explore the target distribution for different “temperatures”. This simplifies the way we sample our parameter space and help us to avoid this local maxima. Here we exemplify the basic idea of this algorithm, however if you are interested in a more extensive explanation, or a modification to make the temperature of the chains dynamical, please consult the reference [27].

We consider a tempering version of the posterior distribution $P(\theta, T | D, H)$

$$P(\theta, T | D, H) \propto L(\theta, D)^{1/T} P(\theta, H), \quad (61)$$

where L is the likelihood and $P(\theta, H)$ the prior. Notice that, for higher T , individual peaks of L become flatter, making the distribution easier to sample with a MCMC algorithm. Now, we have to run N chains with different temperatures assigned in a ladder $T_1 < T_2 < \dots < T_N$, usually taken with a geometrically distributed division, with $T_1 = 1$. The coldest chain T_1 samples the posterior distribution more accurately and behaves as a typical MCMC. Then, we define this chain as the main chain. The rest of the chains are running such that they can cross local maximum likelihoods easier and transport this information to our main chain.

The chains explore independently the landscape for a certain number of generations. Then, in a pre-determined interval, the chains are allowed to “swap” its actual position with a probability

$$A_{i,j} = \min \left\{ \left(\frac{L(\theta_i)}{L(\theta_j)} \right)^{1/T_j - 1/T_i}, 1 \right\}. \quad (62)$$

In this way, if a swap is accepted, chains i and j must exchange their current position in the parameter space, then chain i has to be in position θ_j and chain j has to move to position θ_i .

We can see that, since the hottest chain T_{max} can access easier to all the modes of $P(\theta, H, T_{max} | D)$, then it can propagate its position to colder chains, to be precise, it can propagate its position to the coldest chain $T = 1$. At the same time, the position of colder chains can be propagated to hotter chains, allowing them to explore the entire prior volume.

Affine Invariant MCMC Ensemble Sampler. The main property of this algorithm relies on its invariance under affine transformations. Let's consider a highly anisotropic density

$$p(x_1, x_2) \propto \exp \left(\frac{-(x_1 - x_2)^2}{2\epsilon} - \frac{(x_1 + x_2)^2}{2} \right), \quad (63)$$

which is difficult to calculate for small ϵ . But by making the affine transformation

$$y_1 = \frac{x_1 - x_2}{\sqrt{\epsilon}}, \quad y_2 = x_1 + x_2, \quad (64)$$

we can rewrite the anisotropic density into the easier problem

$$p(y_1, y_2) \propto \exp \left(\frac{-(y_1^2 + y_2^2)}{2} \right). \quad (65)$$

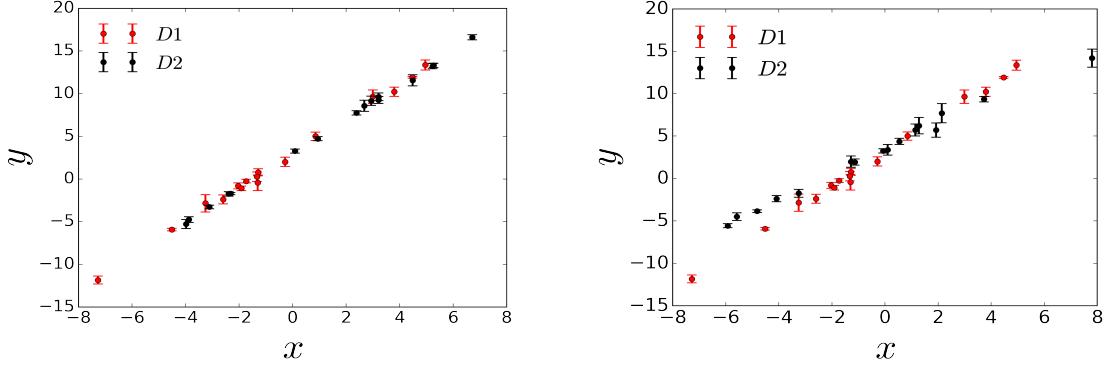


FIG. 7: Datasets D_1 and D_2 measured by our straight-line theory. Case 1 (left) and case 2 (right).

A MCMC sampler has the form $X(t+1) = R(X(t), \psi(t), p)$, where $X(t)$ is the sample after t iterations, R is the sampler algorithm, ψ is the sequence of independent identically distributed random variables and p is the density. A sampler is said to be affine invariant if, for any affine transformation $Ax + b$,

$$R(AX(t) + b, \psi(t), p_{A,b}) = AR(X(t), \psi(t), p) + b. \quad (66)$$

There are already several algorithms that are affine invariant, one of the easiest is known as the *stretch move* [28]. An algorithm fully implemented in Python under the name **EMCEE** [29] is also affine invariant, and there are also some other algorithms that can be found in [30].

Even more samplers. The generation of the elements in a Markov chain is probabilistic by construction and it depends on the algorithm we are working with. The MHA is the easiest algorithm used in Bayesian inference. However, there are several algorithms that can help us to fulfill our mission. For instance, some of the most popular and effective ones, are the Hamiltonian Monte Carlo (see e.g. [31, 32]) or the Adaptative Metropolis-Hastings (AMH) (see e.g. [33]).

5. FITTING A STRAIGHT-LINE

In this section we apply the tools learned so far to the simplest example: fitting a straight-line. That is, we assume that we have a certain theory where our measurements should follow a straight line. Then, in order to apply our techniques, we simulate several datasets along this line. One of the principal topics we want to analyse is the hyperparameter method and how it works, so we will apply our analysis to two different cases (Figure 7):

1. Consider two datasets taken from the same straight-line but with different errors.
2. Consider two datasets but now we simulate both of them from different straight-lines and different errors.

In our analysis we used the PyMC3 module implemented in Python. Our complete code can be downloaded from the git repository [34]. This code is simple to use and can be modified easily for any model to be tested. We recommend to use the file called “new model” where the reader can find a blank project. Here the data and model can be added up and, by running all the notebook, obtain all the analysis we present in this section. One can find as well several notes that will help in programming the model with PyMC3, even if the model contains functions that are not defined in PyMC3.

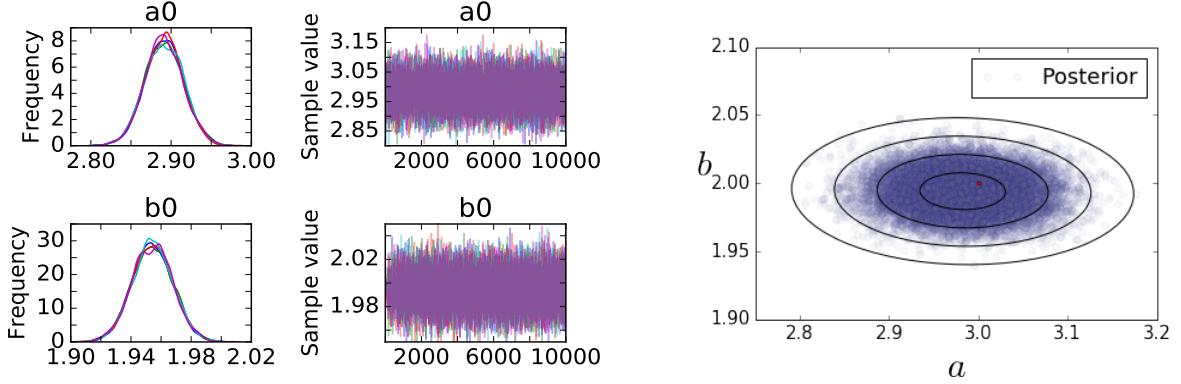


FIG. 8: Left panel: 1D marginalized posterior distributions for our samples and the Markov chains for model H_0 . Right panel: 2D marginalized posterior distributions along with 1-4 confidence regions for our parameters for model H_0 . The red point corresponds to the true value.

5.1. Case 1

In this example we start by considering that our measurements for a given theory (a straight-line $y = a + bx$) are given by the data shown in left panel of Figure 7. These two datasets, D1 and D2, were generated from the line $y = 3 + 2x$, adding a gaussian error to each point. For D1 we add an error with a standard deviation $\sigma_1 = 0.3$, while for D2 we use $\sigma_2 = 0.2$. Then, we would like to estimate the parameters of the model, i.e. a and b . We will analyse this data with and without the hyperparameter method and discuss in detail our results.

Without hyperparameters. Model H_0 .

Before we make a Bayesian estimation, it is necessary to specify our priors. As we have seen, a good prior is a non informative one. Suppose we only know some limits for a and b (we can see them by eye in our data). Then we consider the flat priors

$$a \propto U[0, 5] \quad \text{and} \quad b \propto U[0, 3], \quad (67)$$

where $U[\alpha, \beta]$ are uniform distributions with lower limit α and upper limit β .

From equation (27) we can write our likelihood as

$$L(D; \text{line}) \propto \exp \left[- \sum_d \frac{(y_d - y)^2}{2\sigma_d^2} \right], \quad (68)$$

where y_d is our data taken from the dataset $D = D_1 + D_2$ and σ_d its errors.

We use the MHA to generate our MCMC. In our analysis we ran 5 chains with 10,000 steps for each one. We ran each chain with a temperature $T = 2$ and we thinned them every 50 steps. The results we obtained correspond to $a = 2.982 \pm 0.047$ and $b = 1.994 \pm 0.013$, and their posterior distributions are plotted in Figure 8. Notice that there are some regions where the frequency of events in our sample is increased. So we can say that such parameter regions seem to more likely match the data. Additionally we compute the Gelman-Rubin criterion for each variable in order to verify that our results converged, i.e. for a is 1.000017 and for b is 1.000291. We see that this number is very close to 1, so our convergence criterion is fulfilled. Right panel of Figure 8 displays

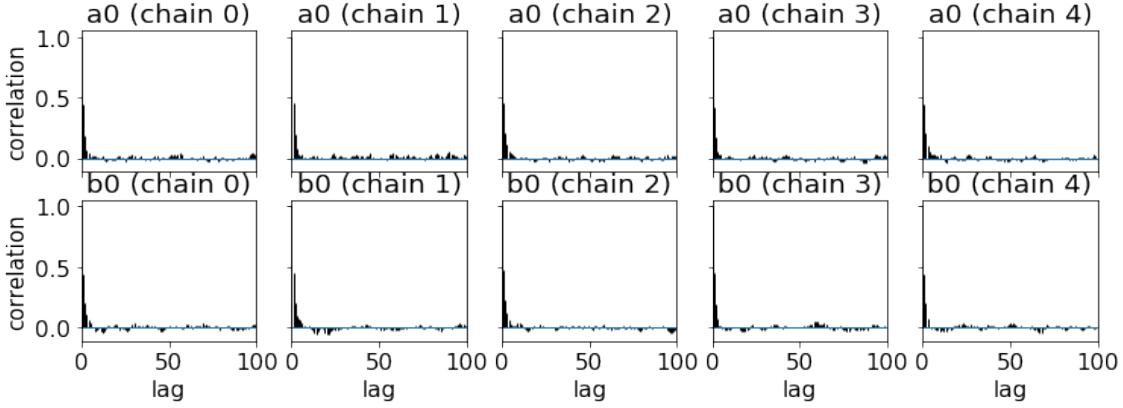


FIG. 9: Autocorrelation plots for model H_0 .

the $1 - 4 \sigma$ confidence regions. We also add a point in red to show the real value for our parameters. The real value for a and b are inside of the curve corresponding to one standard deviation of our estimations in the inferential method.

We continue with the autocorrelation plots. As we mentioned, we need these plots to be small as k increases in order to consider that our analysis is converging. We see in Figure 9 such plots and notice that our convergence criteria is fulfilled. Then, in Case 1 we can see that the model H_0 looks to be a very good estimation procedure.

With hyperparameters. Model H_1 .

Now let us consider the Hyperparameter method. In this case our likelihood can be written as Eq. (48). Similarly to the last procedure, we compute the posterior with flat priors and using 5 chains with 10,000 steps for each one, and check for autocorrelations. Our results are as followed: $a = 2.97 \pm 0.038$ with Gelman-Rubin of 1.000113 and $b = 1.995 \pm 0.010$ with Gelman-Rubin 1.000155. Comparing both procedures we observe they provide similar results. In fact, the confidence regions for both approximations, Fig. 8 and left panel of Figure 10, are similar as well. So, which method is better? We could say that the method with hyperparameters is as good as the one without them, but in order to be sure we compute the evidence ratio K between both models. We obtained from Eqn. (49)

$$K = 3. \quad (69)$$

Then, comparing with Table II we can say that the evidence for H_1 to be better than H_0 is weak. In such case it should be equally better to work with H_0 as to H_1 , as we explained before.

Finally, in order to exemplify our results, let us plot in the right panel of Figure 10 our data with the straight-line inferred by the mean parameters of both models. As we expected our estimation fits well the data for both cases.

5.2. Case 2

Here we consider that we have the same theory for the straight-line but different measurements. The data points are given in the right panel of Figure 7. These correspond to our dataset D_1

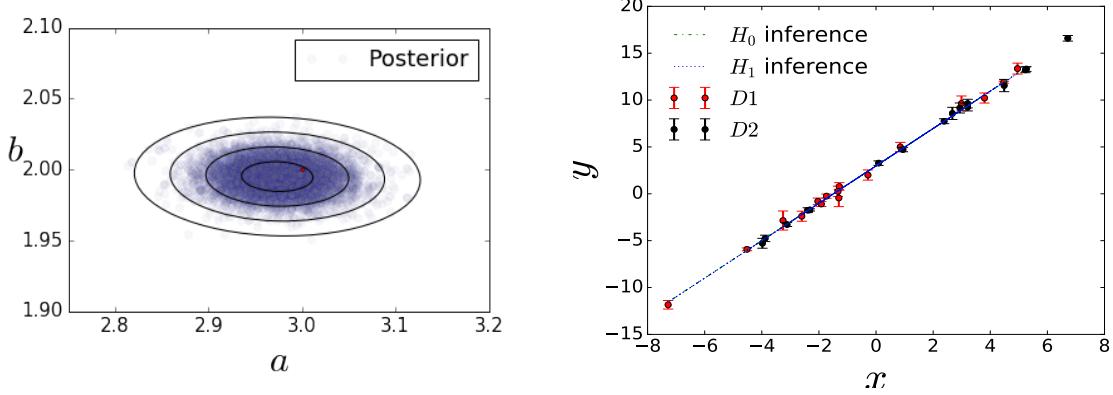


FIG. 10: Left panel: confidence regions for the parameters in model H_1 . Right panel: the best-fit for the straight-lines inferred by the data.

and D_2 , but now changing D_2 by 16 new points generated around the line $y = 3.5 + 1.5x$ with a Gaussian noise and standard deviation $\sigma = 0.5$. So, our datasets are not auto-consistent with each other. Let us make again a parameter estimation for the parameters a and b and look for the differences in both procedures.

Without hyperparameters. Model H_0 .

We follow the same procedure as in Case 1. We computed our posterior and verified that our results converged with the help of the Gelman-Rubin criterion and the autocorrelation plots. Our results are the following: $a = 3.528 \pm 0.056$ and $b = 1.795 \pm 0.014$. Then we plotted our $1 - 4\sigma$ confidence regions in left panel of Figure 11. It is easy to see that our estimation differs so much from the real parameters in our datasets (red points). Of course this is because we are trying to fit a model with non auto-consistent datasets and therefore we arrive at incorrect results. Now, let us see what happens in the hyperparameters procedure.

With hyperparameters. Model H_1 .

In the top right panel of Figure 11 we plotted our posterior distribution. We see immediately that both approximations are very different. While for model H_0 we obtained a single region far away of the real values of our data, for model H_1 we obtained two local maximum regions near the real values for our datasets (red dots). For this example we do not calculate the typical mean and standard deviation for our results.

As the last example we compare both methods. Given the fact that we know *a priori* the real values of our parameters for this example, we could immediately say that the method with hyperparameters is a better approximation than the case without them. However, we confirm this assumption by calculating the ratio K between both models. We obtain

$$K = 37, \quad (70)$$

which means that we have a very strong evidence that H_1 is better than H_0 .

Finally, we can plot the straight-line inferred by model H_0 and the two inferred by model H_1 . Considering parameters inside the two regions in the top right panel of Figure (11) we obtain the bottom panel of Figure 11.

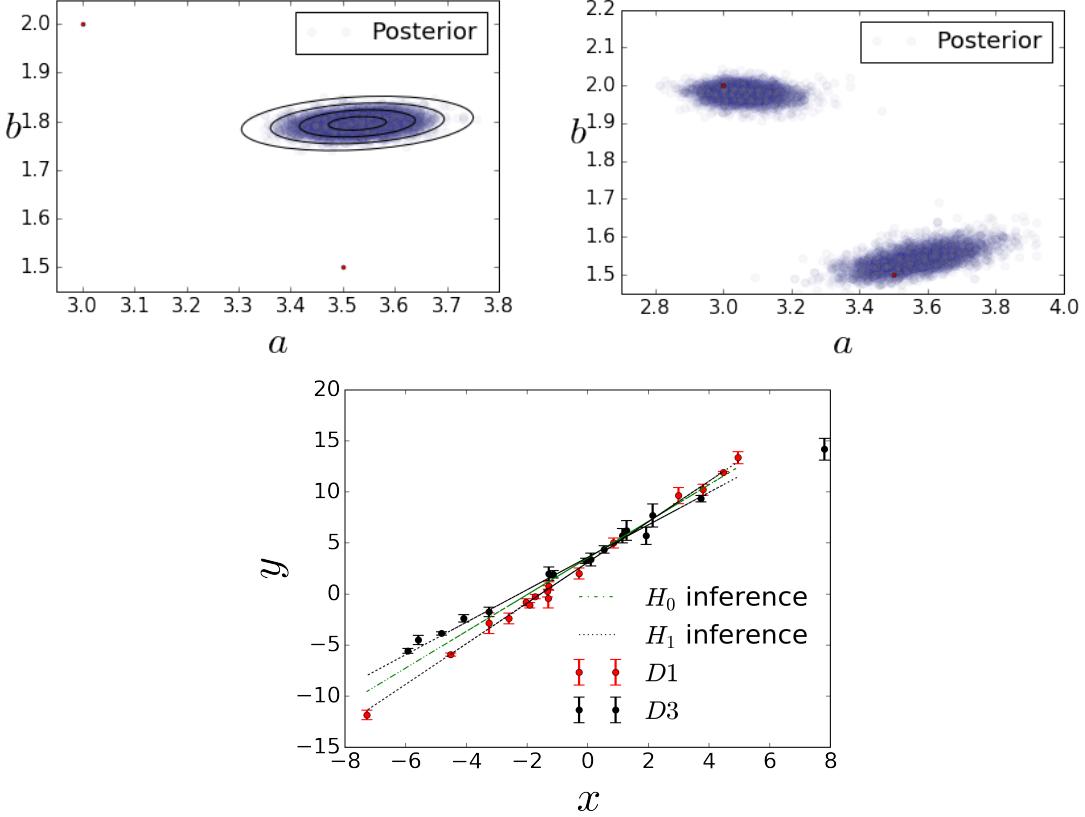


FIG. 11: Top left panel: confidence regions for the parameters in model H_0 . Top right panel: confidence regions for the parameters in model H_1 . Bottom panel: Best-fit values for the straight-lines for Case 2 inferred by our with data.

6. BAYESIAN STATISTICS IN COSMOLOGY

6.1. Theoretical Background

Bayesian statistics is a very useful tool in Cosmology to determine for instance the combination of model parameters that best describes the Universe. In this section we present the basics of Cosmology necessary to apply the Bayesian statistics. In our examples we will focus only on the background Universe –for the moment we avoid perturbations–, since the main purpose of this article is the applications of these techniques rather than the cosmology by itself. It should be clear, however, that the extension to consider perturbations is immediate, i.e. there is only an increment in the number of parameters, and the expressions turn out to be just a little more complicated.

6.1.1. Einstein Field equations

In order to specify the geometry of the Universe, an essential assumption is the **Cosmological Principle**: for a particular time and on sufficiently large scales the observable Universe can be considered homogeneous and isotropic, with great precision. For example, at scales greater than 100 Mega-parsecs the distribution of galaxies observed on the celestial sphere justifies the

assumption of isotropy. The uniformity observed in the temperature distribution (one part in 10^5) measured through the Cosmic Microwave Background (CMB) is the best observational evidence we have in favor of a universal isotropy. Therefore, if isotropy is taken for granted and by taking into account that our position in the Universe has no preference over any another –known as the **Copernican Principle**– the homogeneity follows when considering isotropy in each point.

Homogeneity establishes that the universe is observed equally in each point of space.

Isotropy establishes that the Universe is observed equally in all directions.

The formalism of General Relativity establishes the relationship between the geometry of space-time and the matter on it. That is, the curvature of the spacetime produces physical effects on the matter and these effects are associated to the gravitational field. Additionally, the curvature is related to the matter, described by an energy-momentum tensor $T_{\mu\nu}$. The above expressions can be summarized by paraphrasing Wheeler: “*matter tells space-time how to curve and, in turn, the geometry of this curvature tells matter how to move*”. We can write this sentence down by the Einstein equations

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}. \quad (71)$$

Where $G_{\mu\nu}$ is the Einstein tensor (geometry of the spacetime) and G is the gravitational Newton constant [35, 36]. Throughout this review we use natural units $c = \hbar = 1$.

The distance between two points in a curved space-time can be measured as

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu, \quad (72)$$

where $g_{\mu\nu}$ is the metric tensor that contains all the information about the geometry of the space-time. From now on, and unless otherwise, greek letters μ, ν, \dots denote spacetime indices ranging from 0 to 3, while latin letters i, j, \dots denote spatial coordinates ranging from 1 to 3.

The geometry that best describes a homogeneous, isotropic and expanding Universe is given by the Friedmann-Lemaître-Robertson-Walker metric (FLRW), with a line element

$$ds^2 = dt^2 - a^2(t) \gamma_{ij} dx^i dx^j, \quad (73)$$

where

$$\gamma_{ij} \equiv \delta_{ij} + \kappa \frac{x_i x_j}{1 - \kappa (x_k x^k)}. \quad (74)$$

In Eqn. (73), a represents the scale factor of the Universe which only depends on time, and by convention is normalized to today $a(t_0) \equiv 1$. Similarly, in the expression (74), x^i labelled the spatial coordinates (also called comoving coordinates), δ_{ij} is the Kronecker delta and κ describes the curvature of the space-time.

6.1.2. Friedmann and continuity equations

The content of the Universe needs to satisfy homogeneity and isotropy as well, and hence here it is described by the energy-momentum tensor of a perfect fluid

$$T_{\mu\nu} = (\rho + P) U_\mu U_\nu - P g_{\mu\nu}, \quad (75)$$

where ρ is the energy density, P is the fluid pressure and U_μ is the 4-velocity relative to the observer. If we take the velocity as $U^\mu = (1, 0, 0, 0)$ (comoving observer), the energy-momentum tensor reduces to

$$T_\nu^\mu = g^{\mu\lambda} T_{\lambda\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & -P & 0 & 0 \\ 0 & 0 & -P & 0 \\ 0 & 0 & 0 & -P \end{pmatrix} \quad (76)$$

Using equations (71) and (76), with the FLRW metric, we can obtain the Friedmann equations

$$H^2 \equiv \left(\frac{\dot{a}}{a} \right)^2 = \frac{8\pi G}{3} \sum_i \rho_i - \frac{\kappa}{a^2}, \quad \frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \sum_i (\rho_i + 3P_i). \quad (77)$$

In this expression, H accounts for the rate of expansion/contraction of the the universe, named as the Hubble parameter. Subindex i labels all the components that believe the Universe is made of, as we will see in the next section. These equations describe the evolution of the Universe. By combining equations (77) we can obtain the continuity equation given by

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + P) = 0. \quad (78)$$

The meaning of (78) is the conservation of the energy-momentum tensor ($\nabla_\mu T^{\mu\nu} = 0$). In order to close the system we need to include an equation-of-state that relates pressure and energy density for a given fluid. In particular, we are interested on barotropic fluids which generally have the form of $P = \omega\rho$.

6.1.3. Content of the Universe

Once the equations that define the dynamics of the Universe are known, it is necessary to specify the content of it. The standard cosmological model, also known as Λ Cold Dark Matter (Λ CDM) is one of the most accepted models to describe the Universe. It assumes the content is provided with:

- **Dust:** It has no pressure and its energy density takes the form of $\rho \propto a^{-3}$. Dust is conformed by baryons (ordinary matter).
- **Dark matter:** It is proposed to explain several astrophysical observations like the dynamics of galaxies in the Coma cluster or the rotation curves of galaxies [37, 38]. This type of matter only interacts gravitationally with the rest of the Universe and its energy density evolves in the same fashion as dust. The Λ CDM model assumes dark matter is conformed by weakly interacting massive particles (WIMPs).
- **Radiation:** These correspond to relativistic particles that follow the relation $P = \frac{1}{3}\rho$. This implies a density with a behaviour $\rho \propto a^{-4}$. We consider photons ρ_γ and massless neutrinos ρ_ν as radiation, so the total radiation energy density in the Universe is given by

$$\rho_r = \rho_\gamma + \rho_\nu. \quad (79)$$

The relation between these quantities is

$$\rho_\nu = N_{eff} \times \frac{7}{8} \times \left(\frac{4}{11} \right)^{4/3} \rho_\gamma, \quad (80)$$

where N_{eff} is the effective number of relativistic degrees of freedom, with standard value $N_{eff} = 3.046$ [39].

- **Dark Energy:** It is introduced to explain the current accelerated expansion of the Universe. In the Λ CDM model, dark energy is given by the cosmological constant Λ or equivalently by an equation-of-state $\omega = -1$.

Component	ω
Dust	0
Radiation	1/3
Cosmological Constant	-1

TABLE IV: Equation of state associated to each component of the Universe.

Each of these components can be described by its equation of state shown in Table IV, and defining the density parameter ⁴

$$\Omega_i \equiv \frac{\rho_i}{\rho_{crit}}, \quad \text{with} \quad \rho_{crit} = \frac{3H^2}{8\pi G}, \quad (81)$$

we can rewrite (77) as

$$\frac{H^2}{H_0^2} = \Omega_{r,0}a^{-4} + \Omega_{m,0}a^{-3} + \Omega_{k,0}a^{-2} + \Omega_{\Lambda,0}. \quad (82)$$

Where $\Omega_{r,0}$ is the radiation density parameter, $\Omega_{m,0} \equiv \Omega_{b,0} + \Omega_{DM,0}$ corresponds to the total matter, $\Omega_{b,0}$ to baryons, $\Omega_{DM,0}$ to dark matter, $\Omega_k \equiv -\kappa/(aH)^2$ the curvature density parameter and $\Omega_{\Lambda} \equiv \Lambda/3H^2$ associated to the Cosmological Constant; where the subscript zero indicates they are evaluated today ($a(t_0) = 1$).

6.1.4. Alternatives to the Λ CDM model

The Λ CDM model has had great success in modeling a wide range of astronomical observations. However, it is in apparent conflict with some observations on small-scales within galaxies (e.g. cuspy halo density profiles, overproduction of satellite dwarfs within the Local Group, amongst many others, see for example [38, 40]). In addition, all attempts to detect WIMPs either directly in the laboratory, or indirectly by astronomical signals of distant objects have failed so far. Also, a large range of the particle parameters – predicted to be detectable – have thereby been ruled out. For some of these reasons, it seems necessary to explore alternatives to the standard Λ CDM model. With this in mind, several alternatives have been suggested. For instance the Scalar Field Dark Matter (SFDM) model proposes the dark matter is a spin 0 boson particle [41–45]; or the Self Interacting Dark Matter, as its name states, it relies on the cold dark matter to be made of self interacting particles [46]. On the other hand, in order to explain the accelerated expansion of the universe there exist different modifications to the theory of General Relativity, i.e. $f(R)$ theories [47, 48], braneworld models [49, 50]. There are also several candidates to be the dark energy of the

⁴ ρ_{crit} is the condition to have a flat Universe or equivalently zero curvature.

universe – alternatives to the cosmological constant –, i.e. scalar fields (quintessence, K-essence, phantom, quintom, non-minimally coupled scalar fields [51–53] ; or many more alternatives i.e. anisotropic universes [55–57]. Finally, if the dark energy is assumed to be a perfect fluid, then one of the most popular time-evolving parameterization for its equation of state consists of expanding ω in a Taylor series, for example the Chevallier-Polarski-Linder (CPL) $\omega = \omega_0 + \omega_a(1 - a)$, with two free parameters ω_0, ω_a [58, 59]. It may also be expanded into Fourier series [60] or many more Bayesian approaches have been suggested to account for a dynamical dark energy [61].

6.2. Cosmological parameters

6.2.1. Base parameters

These parameters, also known as *standard parameters*, are the main quantities used in the description of the Universe. They are not predicted by a fundamental theory but their values must be adjusted to provide the best description of the current astrophysical and cosmological observables. To explain the homogeneous and isotropic Universe we can use the density parameter of each component $\Omega_{i,0}$ and the Hubble parameter H_0 related by (82). In particular, the radiation contribution is measured with great precision so that Ω_γ is pinned down very accurately and hence no need to fit this parameter. In a similar way for neutrinos, as long as they maintain a relativistic behavior, they can be related to the density of the photons through (80).

On the other hand, the existence of strong degeneracies from different combinations of parameters is also notorious. In particular the geometric degeneracy involving Ω_m , Ω_Λ and the curvature parameter $\Omega_k = 1 - \Omega_m - \Omega_\Lambda$. To reduce these degeneracies it is common to introduce a combination of cosmological parameters such that they have orthogonal effects in the measurements.

6.2.2. Derived parameters

The above standard set of parameters provides an adequate description of the cosmological models. However, this parameterization is not unique and some other can be as good as this one. Various parameterizations make use of the knowledge of the physics or the sensitivity of the detectors and can therefore be interpreted more naturally. In general, other parameters could have been used to describe the universe, for example: the age of the universe, the current temperature of the neutrino background, the epoch of equality matter-radiation or the epoch of reionization. In the standard cosmological model (Λ CDM), in order to decrease degeneracies, the physical energy densities $\Omega_{DM,0}h^2$ and $\Omega_{b,0}h^2$ are used as base parameters [39].

6.3. Cosmological observations

In this section we review some of the most common experiments and observables used to constrain the cosmological models.

Baryon Acoustic Oscillations (BAO): The BAO is a statistical property, a feature in the correlation function of galaxies or in the power spectrum. The best description of the early Universe considers that it was made of by a plasma of coupled photons and matter (baryons and dark matter). The interaction between the gravitational force due to matter and the radiation pressure formed spherical waves in the plasma. When the Universe cooled down enough, the protons and

electrons were able to join together forming hydrogen atoms, and therefore this process allowed photons to decouple from the rest of the baryons. The photons began to travel uninterrupted while the gravitational field attracted matter towards the center of the spherical wave. The final configuration is an overdensity of matter in the center and a shell of baryons of fixed radius called sound horizon. This radius, used as a standard ruler, is the maximum distance that sound waves could have travelled through the primordial plasma before recombination. The sound horizon r_d is given by

$$r_d = \int_{z_d}^{\infty} \frac{c_s(z)}{H(z)} dz, \quad (83)$$

where the sound speed (in terms of redshift z) in the photon-baryon fluid is $c_s(z) = 3^{-1/2} c [1 + \frac{3}{4} \rho_b(z)/\rho_\gamma(z)]^{-1/2}$, and z_d is the redshift when photons and baryons decouple.

The BAO scale is determined by adopting a fiducial model to be able to translate the angular and redshift separations at comoving distances. The information of the measurement is found in the ratio (α) of the measured BAO scale and that predicted by the fiducial model (“fid”). In an anisotropic fit two ratios are used, one perpendicular α_{\perp} and one parallel α_{\parallel} to the line of sight. A measurement of α_{\perp} constrains the ratio of the comoving angular diameter distance to the sound horizon [62]

$$\frac{D_M(z)}{r_d} = \alpha_{\perp} \frac{D_{M,fid}(z)}{r_{d,fid}}, \quad (84)$$

where the comoving angular diameter distance is given by

$$D_M(z) = \frac{c}{H_0} S_k \left(\frac{D_c(z)}{c/H_0} \right). \quad (85)$$

The the line-of-sight comoving distance is defined as

$$D_c(z) = \frac{c}{H_0} \int_0^z dz' \frac{H_0}{H(z')}, \quad (86)$$

and $S_k(z)$

$$S_k(x) = \begin{cases} \sinh(\sqrt{\Omega_k}x) / \sqrt{\Omega_k} & \Omega_k > 0, \\ x & \Omega_k = 0, \\ \sin(\sqrt{-\Omega_k}x) / \sqrt{-\Omega_k} & \Omega_k < 0. \end{cases} \quad (87)$$

The Hubble parameter can be constrained by measuring α_{\parallel} using an analogous quantity

$$\frac{D_H(z)}{r_d} = \alpha_{\parallel} \frac{D_{H,fid}(z)}{r_{d,fid}}, \quad (88)$$

with $D_H(z) = c/H(z)$.

If redshift-space distortions are weak⁵, an isotropic analysis measures an effective combination of (84) and (88), and the volume averaged distance $D_V(z)$ [62]

$$\frac{D_V(z)}{r_d} = \alpha \frac{D_{V,fid}(z)}{r_{d,fid}}, \quad (89)$$

⁵ This approach is valid for luminous galaxy surveys but not for the Ly- α .

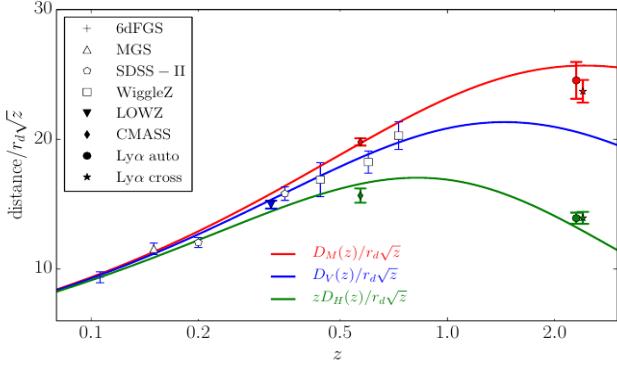


FIG. 12: BAO Hubble diagram. BAO measurements of D_V/r_d , D_M/r_d and zD_H/r_d from the sources indicated in the legend. The scaling factor \sqrt{z} is included for a better display of the error bars. Solid lines are plotted by using the best-fit values obtained by the Planck satellite [64]. The Ly α cross-correlation points have been shifted in redshift; auto-correlation points are plotted at the correct effective redshift. Figure taken from [62].

with $D_V(z) = [zD_H(z)D_M^2(z)]^{1/3}$.

The BAO measurements constrain the cosmological parameters through the radius of the sound horizon r_d , Hubble distance $D_H(z)$ and the comoving angular diameter distance $D_M(z)$ – see Figure 12 –. The data used in this work to constrain D_V/d_r is obtained from the 6dF Galaxy Survey (6dFGS, [63]) from UK Schmidt Telescope, the Main Galaxy Sample (MGS, [65]) and the BOSS LOWZ Sample [66] from SDSS. On the other hand the data from BOSS CMASS Sample [66], BOSS Lyman α auto-correlation (Ly α auto, [67]), BOSS Lyman- α cross correlation (Ly α cross, [68]) are used to constrain D_M/r_d , D_H/r_d and r_d .

Supernovae type Ia (SNIa): A supernova is the explosion of a star. There are two types of supernova. The first occurs in a binary star system, one of which is a white dwarf that steals matter from the star that accompanies it. When the white dwarf accumulates a lot of matter it explodes producing the supernova. The second type occurs when a star runs out of fuel and matter flows into its core. When the core becomes so heavy that it cannot withstand its own gravitational force, it collapses and the star explodes producing the supernova [69]. Empirically, the peak luminosity of the type Ia supernovas (SNIa) can be used as a distance indicator from the relation between redshift and distance [39]. From several analysis of SNIa the Supernova Cosmology Project and High-z Supernova Search Team both found evidence that the Universe is currently speeding up [70–72].

These stars allow us to make relative distance measurements using the luminosity distance given by

$$D_L \equiv \sqrt{\frac{L}{4\pi S}}, \quad (90)$$

where L is the luminosity defined as the energy emitted per unit solid angle per second and S is the radiation flux density defined as the energy received per unit area per second [37]. The observable quantity is the radiation flux density received and it cannot be translated into the luminosity density unless the absolute luminosity of the object is known. To solve these problem we use supernovae because they have the same luminosity being at different distances. Even if luminosity is unknown, it will appear as a scaling factor [37]. The relation between D_L and the cosmological parameters is given by

$$D_L = D_M(1 + z), \quad (91)$$

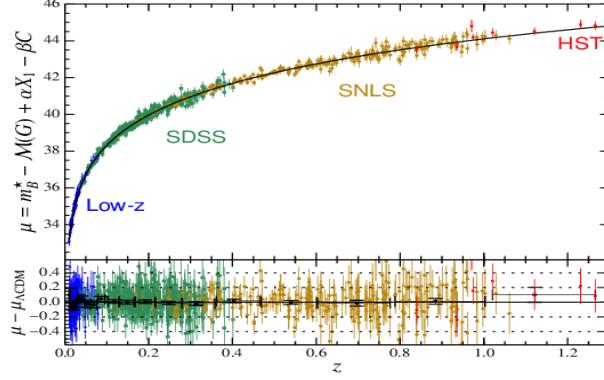


FIG. 13: Joint Light-curve Analysis data (JLA). Vertical axis is the standardized distance modulus μ (luminosity distance function) and the horizontal axis is the redshift z . Source: [74].

where D_M is provided by equation (85). Another important quantity in the observation of supernovas is the standardized distance modulus

$$\mu = m_B^* - M_B + \alpha X_1 - \beta C, \quad (92)$$

where m_B^* is the observed peak magnitude in the restframe of blue band (B), α , β and M_B are parameters that depend on host galaxy properties [73]. X_1 is the time stretching of the lightcurve. And C is the supernova color at maximum brightness. The relation between the standardized distance modulus and the luminosity distance is

$$\mu = 5\log_{10}\left(\frac{D_L}{10pc}\right). \quad (93)$$

The data used in this paper (showed in Figure 13) is obtained from the Joint Light-curve Analysis (JLA). It is a collaboration to analyze the data of 740 stars from the SDSS-II (previous version of BOSS), the SuperNova Legacy Survey (SNLS, [75]) experiment that used the Canada-France-Hawaii telescope (CFHT), the Calán/Totolo Survey, the Carnegie Supernova Project, the Harvard-Smithsonian Center for Astrophysics, La Silla Observatory, Fred Lawrence Observatory and Hubble Space Telescope (HST) [73]. For simplicity we compress all the information into a linear function fit over 30 bins (31 nodes) spaced evenly in $\log(z)$ with a 31×31 covariance matrix [62].

Cosmic Microwave Background (CMB): Corresponds to the radiation that permeates all the Universe; discovered in 1965. Before recombination, baryons and photons were tightly coupled, and once photons decouple from the rest of the matter, they travelled uninterrupted until reach us. The temperature radiation measured at different parts of the sky contains information of the last scattering epoch, gravitational lensing, among others. Here, the CMB displays the primordial anisotropies studied in the angular power spectrum. One of the most important collaboration that studies the CMB corresponds to the Planck satellite⁶. It is an European Space Agency mission which main objective is to measure the temperature, polarization and anisotropies of the CMB over the entire sky. These results would allow to determine the properties of the Universe at large scales, the nature of dark matter and dark energy, as well as test inflationary theories, determine whether the Universe is homogeneous or not and obtain maps of galaxies in the microwave [76–78].

⁶ Previous probes are like COBE [85] and WMAP [86].

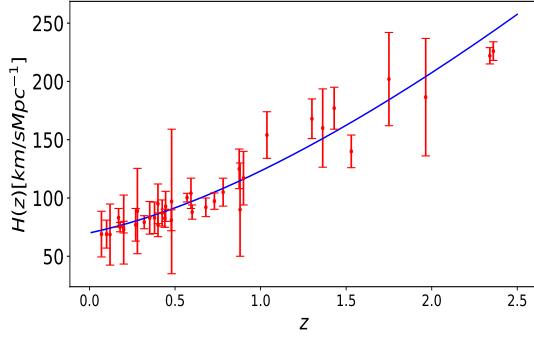


FIG. 14: Hubble data as a function of redshift z for Cosmic Chronometers. The solid line corresponds to the best-fit by using Λ CDM model.

In this work we use the CMB information as a BAO located at redshift $z = 1090$, measuring the angular scale of the sound horizon $D_M(1090)/r_d$. Also to calibrate the absolute length of the BAO ruler through the determination of $\Omega_b h^2$ and $\Omega_{cb} h^2$ ⁷.

Large Red Galaxies: (Cosmic Chronometers) These are the most massive galaxies for each redshift z and contain the oldest star population. These kind of galaxies are used to estimate the Hubble factor because they contain little stardust which makes it easier to get their spectrum. The way these chronometers work is by selecting two galaxies at different redshift between $z \sim 0 - 2$ and compare their upper cut in its age distributions. By doing this, it is possible to obtain the difference of ages Δt and redshifts Δz such that the expression $\frac{dz}{dt}$ can be approximated as $\frac{dz}{dt} \simeq \frac{\Delta z}{\Delta t}$. This quantity is related with the Hubble factor via [79, 80]

$$H(z) = \frac{-1}{1+z} \frac{dz}{dt} \simeq \frac{-1}{1+z} \frac{\Delta z}{\Delta t}. \quad (94)$$

In this work, the data used to constrain $H(z)$ was obtained from [81–84]. In Figure 14 a compilation of these data is shown.

7. PARAMETERS INFERENCE IN COSMOLOGY

The simplest way to understand how all the concepts of Bayesian statistics can be applied to cosmology is by an example. We consider the typical example in Cosmology, which is the estimation of the Hubble parameter H_0 at the present time and the matter density of the Universe Ω_m , assuming a Λ CDM standard model. In this section we use our own Python code using the PyMC3 Python's module in the same way we did with the straight-line example. For interested readers, the code can be found in [34]. Notice that this can be used not only in cosmology but also with the theory of your preference. The main modification is to specify a new model in the method “pm.model()” and upload your own data. All information needed is contained within the repository.

The data.- Let us start by considering the observation we want to analyze. For this particular example we focus only on the Cosmic Chronometers, as shown in Figure 14.

⁷ Details of how they are used can be seen in [62].

The model.- Our interest, over this example, is to fit the density parameters for each component of the Universe as well as the value of the Hubble parameter at present time. For this purpose we use equation (82) in terms of redshift $1 + z = 1/a$, and for simplicity we assume a flat Universe and do not consider radiation. Then the model we use in the fitting is given by

$$H^2 = H_0^2[\Omega_{m,0}(1+z)^3 + \Omega_{\Lambda,0}], \quad (95)$$

constrained by the relation

$$\Omega_{m,0} + \Omega_{\Lambda,0} = 1. \quad (96)$$

Notice that the above relation implies that we can get rid of a parameter – $\Omega_{m,0}$ or $\Omega_{\Lambda,0}$ – in the analysis. Then, the parameters we decided to estimate are H_0 and $\Omega_{m,0}$. By assuming the Gaussian approximation, we construct the likelihood given by

$$L \propto \exp \left[- \sum_i \frac{(H_i - H(z_i))^2}{2\sigma_i^2} \right], \quad (97)$$

where $H(z_i)$ is described by equation (95) evaluated at each redshift z_i , H_i is the value of the Hubble parameter measured at z_i and σ_i the error of the i -th measurement.

Now, as part of the model construction, it is necessary to specify our priors. The only apriori information we have about the free parameters is that each component of the Universe $\Omega_{i,0}$ must satisfy the relation $0 \leq \Omega_i \leq 1$, while for the present Hubble parameter its conservative prior can be obtained by observing the data at our disposal. In such cases, a good prior choice is an Uniform distribution with limits $\Omega_m \in [0, 1]$ and $H_0 \in [10, 100]$. Therefore, we have as our priors

$$\Omega_{m,0} \sim U[0, 1], \quad H_0 \sim U[10, 100]. \quad (98)$$

Numerical estimation.- We follow the same procedure than in the straight-line example. In the left panel of Figure 15 we have plotted the chains obtained for our estimations and its corresponding 1D posterior distribution. Similarly than in the above examples, we have also plotted the $1 - 4\sigma$ confidence regions in the right panel of the same figure. Additionally we have obtained the mean, standard deviation and the Gelman Rubin criteria for each parameter: $H_0 = 67.77 \pm 3.13$ with convergence 1.00045 and $\Omega_m = 0.331 \pm 0.0628$ with convergence 1.00044. We can see that our estimations are very similar than the values reported in the literature [64]. The minor difference can be understood because we have simplify our analysis.

7.1. Cosmological and statistical codes

As we have shown until now, during the process of fitting free parameters of a given model there are three steps we have followed: first, obtain the data we would like to confront with the model parameters. Then, construct the likelihood associated with the theory we are working with. Depending on the nature of the data, the likelihood can depend on different ways of the parameters. For example, in the last exercise the likelihood in terms of the parameters is via (95) and (97). Finally, it is necessary to program all the numerical tools in order to obtain the parameter inference. Such programming can be done in PyMC3, for example, as we saw before.

It is notorious that the above process can be a tremendous task for programming. For instance, when the theories involve a large number of parameters, or some of them must be marginalized in

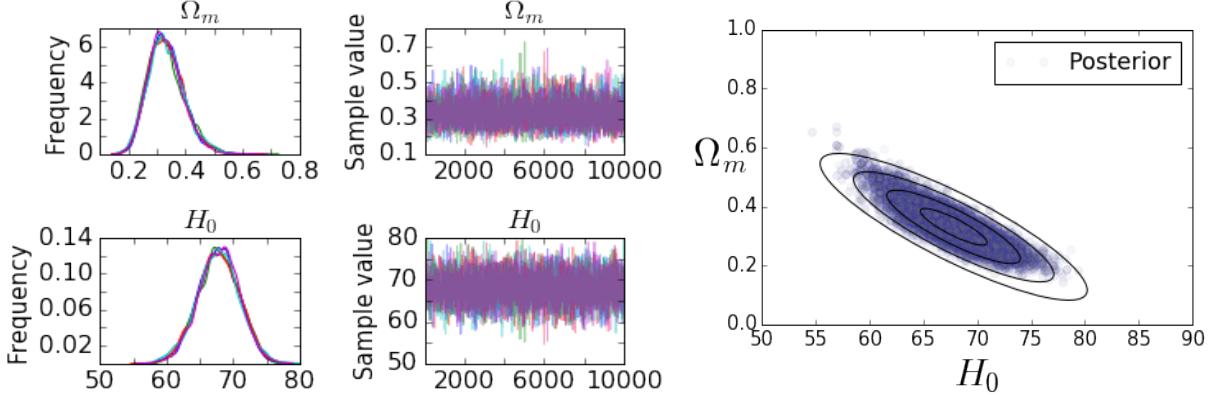


FIG. 15: Results for the Λ CDM model. Left panel: 1D posterior distributions for the parameters H_0 and Ω_m along with its chains. Right panel: joint 2D posterior distributions with 1-4 confidence levels.

order to ignore the non-interesting ones (like nuisance parameters), or/and when the models depend on the parameters of interest in a difficult way (like by solving differential equations, integrals, etc). Then we can proceed in two different ways: First is by accepting the challenge and creating our own code. Developing a new code may be a good option rather than using some others as it would require a long time to learn the implementation and modification (specially if the theory we are dealing with it is somehow simple). Otherwise, we need to rely on existing codes, principally when the theory of interest is quite complicated, as it is usually the case when perturbations are taken into account on the cosmological models. In this section we present some cosmological and statistical codes available and ready to test cosmological models.

7.1.1. Cosmological codes

Nowadays there are several cosmological Boltzmann codes freely on the web. Some of them include: CMBFAST (written in FORTRAN 77; [87]), CMBEASY (C++; [88]), CAMB (FORTRAN 90; [89, 90]), CLASS (C; [91, 92]), COSMOSIS [93] (It is written in Python and it works as an interface between CLASS, CAMB, MontePython, CosmoMC, and more). All of them are used for calculating the linear CMB anisotropy spectra, based on integrations over the sources along the photon line of sight.

7.1.2. Statistical codes

Once the cosmological model is established we need a statistical code to estimate the free parameters of our model. Fortunately there are several MCMC codes free to download that can make this homework easy to handle. Some of them are:

Monte Python.- It is a Monte Carlo code for Cosmological Parameter extraction that can be downloaded in [94]. It contains likelihoods of most recent experiments, and interfaces with the Boltzmann code Class for computing the cosmological observables. The code has several sampling methods available: Metropolis-Hastings, Nested Sampling (through MultiNest), EMCEE (through CosmoHammer) and Importance Sampling. If you are interested on this code some help can be found at [92, 95].

Parameter	Λ CDM	$o\Lambda$ CDM	ω CDM	$o\omega$ CDM	$\omega_0\omega_a$ CDM	$o\omega_0\omega_a$ CDM
Ω_m	0.299 ± 0.007	0.298 ± 0.007	0.303 ± 0.009	0.299 ± 0.009	0.307 ± 0.010	0.306 ± 0.010
$\Omega_b h^2$	0.0224 ± 0.0002	0.0227 ± 0.0003	0.0224 ± 0.0003	0.0227 ± 0.0003	0.0224 ± 0.0003	0.0226 ± 0.0003
h	0.684 ± 0.006	0.679 ± 0.007	0.677 ± 0.108	0.676 ± 0.010	0.674 ± 0.011	0.670 ± 0.011
Ω_k	...	-0.004 ± 0.002	...	-0.003 ± 0.003	...	-0.006 ± 0.003
ω_0	-0.96 ± 0.05	-0.97 ± 0.05	-0.91 ± 0.10	-0.83 ± 0.11
ω_a	-0.17 ± 0.41	-0.52 ± 0.51
χ^2_{min}	73.57	71.59	73.13	71.5	72.8	69.7

TABLE V: Cosmological parameter constraints from BAO data combined with our compressed description of CMB from Planck, the JLA SN and Hubble data (BBAO+Planck+SN+HD). Two-tailed distributions are shown along with 1σ C.L. Entries for which the parameter is fixed are marked with dash ($\Omega_k = 0$, $\omega = \omega_0 = -1$, $\omega_a = 0$).

CosmoMC.- is a Fortran 2003 MCMC engine for exploring cosmological parameter space [96]. It contains Monte Carlo samples and importance sampling. It also has by default several likelihoods of the most recent experiments, and interfaces with CAMB.

SimpleMC.- It is a MCMC code for cosmological parameter estimation where only the expansion history of the Universe matters. This code solves the cosmological equations for the background parameters in the same way that CLASS or CAMB and it contains the statistical parameter inference that CosmoMC or MontePython. An advantage of this code is that it is completely written in Python, being a programming language easy to handle. It was written by Anže Slosar and José A. Vázquez and can be downloaded on [97].

8. EXAMPLES WITH SIMPLEMC

The main interest of this section is to test several cosmological models through the SimpleMC code. We selected this code since we feel more confident to use it, however, the results we present can also be obtained in any of the aforementioned codes.

Throughout these examples we consider Gaussian likelihoods for each dataset with the following form

$$L \propto \exp \left[- \sum_i \frac{(T(z_i) - T_i)^2}{2\sigma_i} \right], \quad (99)$$

where $T(z_i)$ is the theoretical value related to the observation T_i ; and σ_i are the corresponding errors associated for each measurement. In our estimation $T(z_i)$ is given by (91) for Supernovas; (84) for CMB; (84), (88) and (89) for BAO and (94) for Cosmic Chronometers.

We use the BAO data mentioned in section 6.3 (labeled as BBAO), for Supernovas we use the Joint Light-Curve Analysis compressed data denoted as SN, the Planck data (denoted as Planck) for CMB and the Cosmic Chronometers data (HD)⁸.

⁸ See section 6.3.

8.1. Models of the Universe

The base example corresponds to the flat Λ CDM model, already explored in Section 7, but now including the full observations presented in Section 6.3 and using the SimpleMC code. The SimpleMC code is easily manageable and its documentation can be found in [97]. Here, in order to test the Friedmann equation (82) with the data, we consider as free parameters (along with flat priors): the total matter dimensionless density parameter $\Omega_m \in [0.05, 1.5]$, the baryon physical density $\Omega_b h^2 \in [0.02, 0.025]$ and the dimensionless Hubble constant $h \in [0.4, 1]$. Then, assuming the base Λ CDM model we let the curvature of the universe be a free parameter (model $\text{o}\Lambda\text{CDM}$), with its corresponding flat prior $\Omega_k \in [-1.5, 1.5]$. Moreover, because the cosmological constant is only a particular case for the dark energy equation-of-state $\omega = -1$, we let ω be a free parameter with flat priors $\omega \in [-2.0, 0.0]$ and labelled it as model ωCDM . We may combine the addition of curvature and constant ω to define the $\text{o}\omega\text{CDM}$ model. In order to go even further and describe a dynamical dark energy, we use the CPL parameterization for the equation of state with flat priors on the parameters $\omega_0 \in [-2.0, 0.0]$ and $\omega_a \in [-2.0, 2.0]$; and labelled it as $\omega_0\omega_a\text{CDM}$ model. Again, we can incorporate the curvature of the Universe to the CPL parameterization, named it as $\text{o}\omega_0\omega_a\text{CDM}$.

By using the combined dataset BBAO+Planck+SN+HD, Table V shows the best fit values, along with 1σ confidence levels. The first important result to highlight is how the constraints have shrunk once new information is taken into account. That is, in Section 7, for the Λ CDM model and by using only Hubble data, we had $h = 0.677 \pm 0.313$ and $\Omega_m = 0.331 \pm 0.0628$. Now with the inclusion of BAO, Planck and SN data the constraints have improved considerably to $h = 0.684 \pm 0.006$ and $\Omega_m = 0.299 \pm 0.007$. Figure 15 is updated with new data in order to get Figure 16. Here, the left panel displays the chain for the parameter h , with 9000 steps. In the right panel of the same figure we plot the 2D posterior distribution along with 1 and 2σ confidence regions obtained from our estimations.

From Table V we also observe the best fit of most of the new parameters – additional to the base Λ CDM model – remained well inside the 1σ confidence level. However the exception is for the $\text{o}\omega_0\omega_a\text{CDM}$ model, where Ω_k , ω_0 and ω_a lay down right outside the 1σ region. This main feature is better observed in Figure 17, where the 2D posterior distributions – along with 1 and 2σ confidence regions are shown. Here, the standard Λ CDM values are marked with a dash line ($\Omega_k = 0$, $\omega = \omega_0 = -1$, $\omega_a = 0$). The inclusion of extra parameters improves the fit to the data, observed through the minimum χ^2_{min} – last row of Table V –. However, it also carries out a penalisation factor that affects directly the model selection, as seen in [54, 60, 61].

9. CONCLUSIONS AND DISCUSSION

The fact that the number of cosmological observations have increased impressively over the last decade would have allowed us to obtain a better description of the Universe. However, since we still have the limitation of only one Universe, a Frequentist approach may not be the best to rely on and hence the Bayesian statistics came across. In this work we provide a review of the Bayesian statistics and present some of its applications to cosmology, mainly throughout several examples.

The Bayesian statistics rests on the rules of probability which yield to the Bayes theorem. Given a model or hypothesis H for some data D , Bayes' theorem tells us how to determine the probability distribution of the set of parameters θ on which the model depends. Bayes' theorem states that

$$P(\theta|D, H) = \frac{P(D|\theta, H)P(\theta|H)}{P(D|H)},$$

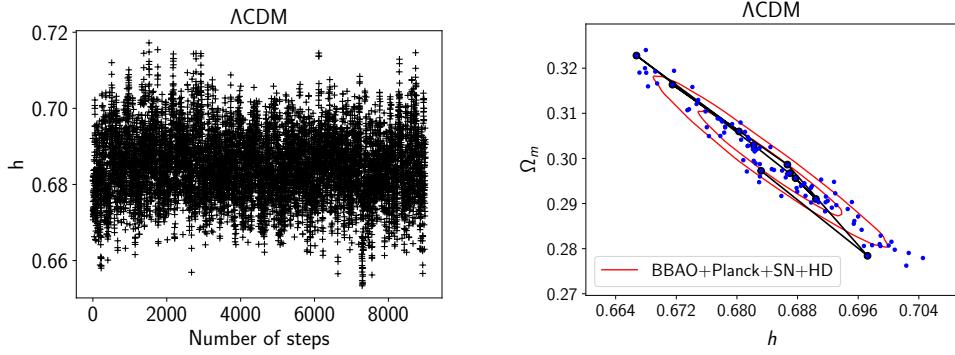


FIG. 16: Left panel: Markov chain for the parameter h , with 9000 steps. Right panel: 2D posterior distribution with confidence regions 1 and 2 σ for the joint parameters h and Ω_m .

where the *prior* probability $P(\theta|H)$ – the state of knowledge before acquiring the data – is upgraded through the *likelihood* $P(D|\theta, H)$ when experimental data D are considered. The aim for parameter estimation is to then obtain the posterior probability $P(\theta|D, H)$ which represents the state of knowledge once we have taken into account the new information.

We noticed that if we ignore the prior probability, we can identify the posterior probability with the likelihood $P(\theta|D, H) \propto L(D|\theta, H)$, and thus by maximizing it, we can find the most probable set of parameters for a model given the data. Moreover, if we assume a Gaussian approximation for the likelihood, then the chi-squared quantity is related to the Gaussian likelihood via $L = L_0 e^{-\chi^2/2}$. Therefore, maximizing the Gaussian likelihood is equivalent to minimizing the chi-squared. Once the posterior distribution for a set of parameters, of a given model, is calculated we show the results in the form of confidence regions of said parameters. Also for this particular case, in which likelihoods are Gaussian, Fisher's matrix can be calculated according to the Hessian matrix, where the latter contains information about the errors of the parameters and their covariances. The Fisher matrix gives information about the accuracy of the model and allows to predict how well an experiment will be able to constrain the set of parameters for a given model.

On the other hand, sometimes it is difficult to know, a priori, if multiple datasets are consistent with each other, or whether there could be one or more that are likely to be erroneous. Since there is usually this uncertainty, a way to know how useful a dataset may be is by introducing the hyperparameter method. These hyperparameters act as ‘weights’ for every data set in order to take away data that does not seem to be consistent with the rest of them. Here, the key quantity to estimate the necessity of introducing hyperparameters to our model is given by the Bayesian evidence $P(D|H)$.

The estimation of the posterior distribution is a very computationally demanding process, since it requires a multidimensional exploration of the likelihood and prior. To carry out the exploration of the cosmological parameter space we focus on Markov Chain Monte Carlo methods with the Metropolis Hastings algorithm. A Markov process is a stochastic process (that aims to describe the temporal evolution of some random phenomenon) where the probability distribution of the immediate future state depends only on the present state. Any computational algorithm that uses random numbers is called Monte Carlo. Thus, the Metropolis Hastings algorithm uses a transition kernel to construct a sequence of points (called ‘chain’) in the parameter space in order to evaluate the posterior distribution of said parameter. The generation of the elements in a Markov chain is probabilistic by construction and it depends on the algorithm we are working with. The MHA is

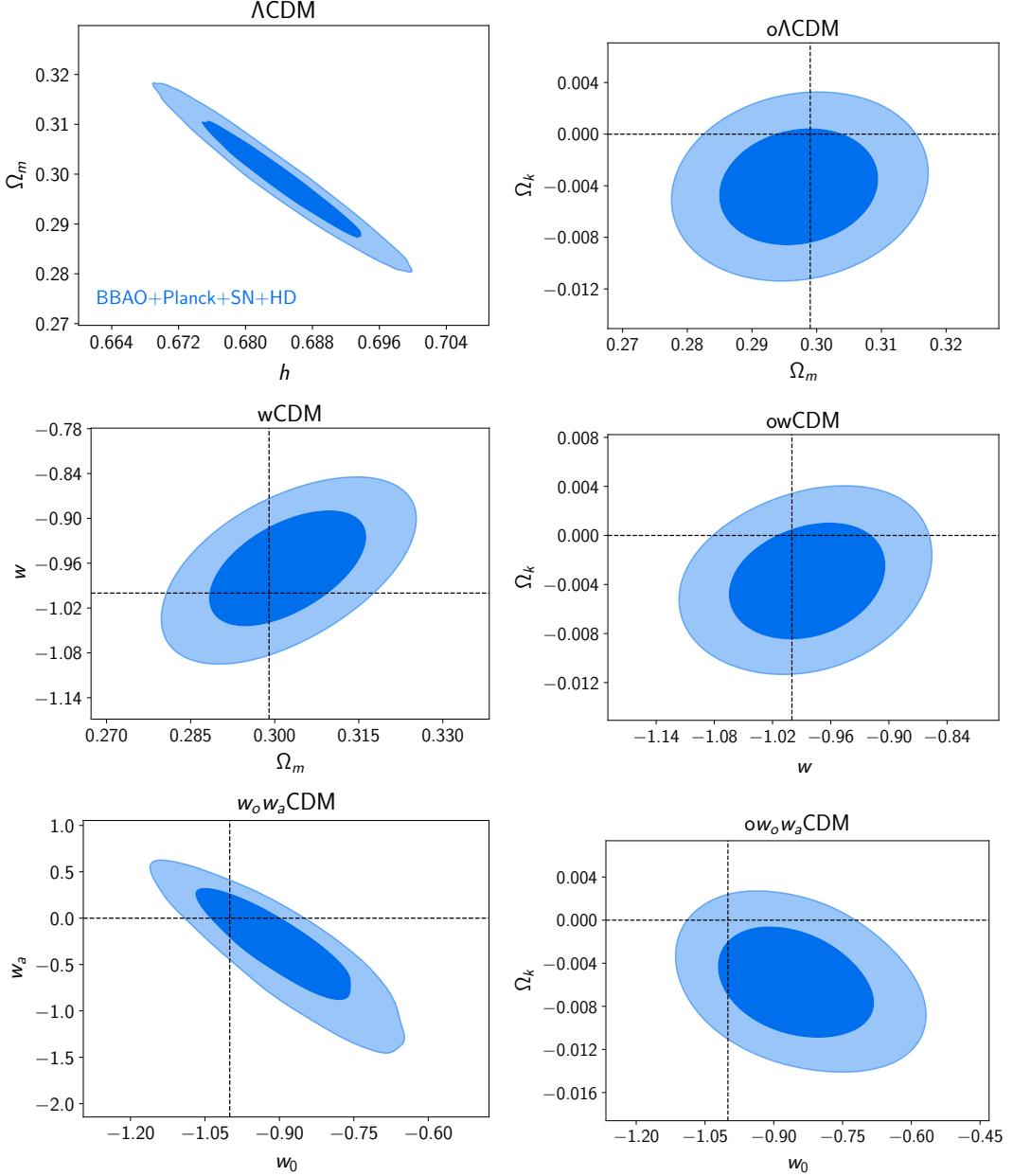


FIG. 17: This figure shows the 2D posterior distributions with 1,2- σ confidence regions for different models. $\text{o}\Lambda\text{CDM}$ refers to ΛCDM with curvature, ωCDM is a flat universe with the dark energy equation of state as a free parameter, ωCDM generalizes to non-zero curvature, $w_0 w_a \text{CDM}$ uses the CPL parameterization and $\text{o}w_0 w_a \text{CDM}$ generalizes to non-zero curvature. The dashed lines show the standard ΛCDM values.

the easiest algorithm used in Bayesian inference, however to explore complex posterior distributions more efficiently we provide a brief description of several samplers.

Finally we show how Bayesian statistics is a very useful tool in Cosmology to determine for instance the combination of model parameters that best describes the Universe. In particular, we confront the standard cosmological model (ΛCDM) to current observations and compare it to different models. We found the model that best fit the data corresponds to a curved Universe with a dynamical dark energy, namely $\text{o}w_0 w_a \text{CDM}$.

ACKNOWLEDGMENTS

JAV acknowledges the support provided by SEP-CONACYT Investigación Básica A1-S-21925, and UNAM-PAPIIT IA102219. LEP, LOT and LAE were supported by CONACyT México.

Appendix A: A simple MCMC python code

Here we show our MCMC written in Python. This code is very simple and its purpose is to help the reader to understand how to program a MCMC code from scratch. However, if the reader is interested in more sophisticated algorithms PyMC3 module in python may be useful [20].

```

'''The neccesary packages for our MCMC code'''

import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as st

'''Calculate the posterior'''
def target(lik, prior, param, theta):
    if theta < 0 or theta > 1:
        return 0
    else:
        return lik(param[0], theta).pmf(param[1])*prior.pdf(theta)

param = 14,10 #our initial data: 14 trials and 10 heads
a = 1 #Params for the beta function
b = 1

lik = st.binom #Our likelihood
prior = st.beta(a,b) #Our prior
sigma = 0.1 #Standard desviation for the gaussian proposal distribution

theta = 0.05 #Our guess for p

niters = 5000 #How many iterations we want to do
samples = np.zeros(niters+1)
samples[0] = theta

for i in range(niters):
    theta_p = theta + st.norm(0, sigma).rvs() #New step
    rho = min(1, target(lik, prior, param, theta_p)/target(lik, prior, param, theta))
    u = np.random.uniform()
    '''Metropolis Hastings algorithm'''
    if u < rho:
        theta = theta_p
    samples[i+1] = theta
nmcmc = len(samples)//2

'''The chain generated by our algorithm'''
plt.plot(samples, ':',label="Chain")
plt.xlabel("p")
plt.ylabel("P(p|D)")
plt.legend()
plt.show()
```

-
- [1] *The Shapley - Curtis Debate in 1920.* Retrieved December 2017, from https://apod.nasa.gov/diamond_jubilee/debate_1920.html.
- [2] M. S. Turner, *David Norman Schramm, a biographical memoir*, National Academy of Sciences, (2009).
- [3] B.J.K. Kleijn, *Bayesian statistic, lecture notes 2015*, Universiteit van Amsterdam.
- [4] Alan Heavens, *Statistical techniques in cosmology*, Lect. Notes Phys. 800:147-177, (2010), arXiv:0906.0664.
- [5] Roberto Trotta, *Bayes in the sky: Bayesian inference and model selection in cosmology*, Contemp. Phys. 49:71-104, (2008), arXiv:0803.4089.
- [6] Licia Verde, *Statistical methods in cosmology*, Lect.Notes Phys.800:147-177, (2010), arXiv:0911.3105.
- [7] Roberto Trotta, *Bayesian Methods in Cosmology*, (2017), arXiv:1701.01467.
- [8] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P.Flannery. *Numerical Recipes, The Art of Scientific Computing. 3rd. edition, Cambridge University Press*, (1996).
- [9] Fisher R.A. *The logic of inductive inference J. Roy. Stat. Soc.* **98**, 39, (1935).
- [10] Albrecht A. et al., Report of the Dark Energy Task Force, June (2006).
- [11] S. T. Tokdar and R. E. Kass, *Importance sampling: A review*, WIREs Comp. Stat., 2.
- [12] Lahav O., Bridle S. L., Hobson M. P., Lasenby A. N., & Sodre L., *Bayesian ‘hyper-parameters’ approach to joint estimation: the Hubble constant from CMB measurements*, MN-RAS, 315, 45, (2000), arXiv:astro-ph/9912105.
- [13] Hobson M. P., Bridle S. L., & Lahav O., *Combining cosmological data sets: hyperparameters and Bayesian evidence*, MNRAS, 335, 377, (2002), arXiv:astro-ph/0203259.
- [14] Tanner, M., *Tools for Statistical Inference, Method for Exploration of Posterior Distributions and Likelihood Functions*, Spring Series in Statistics, 3rd Edition, (1993).
- [15] Gilks, W., Richardson, S. and Spiegelhalter, D., *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC Interdisciplinary Statistics series, 1st Edition, (1995).
- [16] Gelman A., Carlin J., Stern H and Rubin D., *Bayesian Data Analysis*, Chapman & Hall/CRC Texts in Statistical Science series, 2nd Edition, (1995).
- [17] Ross M. Sheldon, *Introduction to Probability Models*, 10th Edition, (1989).
- [18] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E., *Equations of state calculations by fast computing machines.*, *J. Chem. Phys.*, 21(6): 10871092, (1953).
- [19] Hastings, W., *Monte Carlo sampling methods using Markov chains and their application.*, *Biometrika*, 57: 97109, (1970).
- [20] Salvatier J., Wiecki T.V., Fonnesbeck C., *Probabilistic programming in Python using PyMC3*, PeerJ Computer Science 2:e55, (2016), arXiv:1507.08050.
- [21] Gelman, A., and D. B. Rubin, *Inference from Iterative Simulation Using Multiple Sequences*. *Statistical Science* 7: 457511, (1992).
- [22] Brooks, S. P., and A. Gelman, *General Methods for Monitoring Convergence of Iterative Simulations*. *Journal of Computational and Graphical Statistics* 7: 434455, (1997).
- [23] William A. Link and Mitchell J. Eaton, *On thinning of chains in MCMC*, *Methods in ecology and evolution*, 3, 112115, (2012).
- [24] Geweke J. *Evaluating the accuracy of sampling-based approaches to calculating posterior moments*. In Bernardo, pages 169193, (1992).
- [25] *Model checking and diagnostics*. Retrieved December 2017, from <https://pymc-devs.github.io/pymc/modelchecking.html>.
- [26] Geman S., Geman D., *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6, (6), pp. 721741, (1984).
- [27] W. D. Vousden, W. M. Farr and I. Mandel, *Dynamic temperature selection for parallel-tempering in Markov chain Monte Carlo simulations*, MNRAS Vol. 455 1919-1937, (2016), arXiv:1501.05823.
- [28] Foreman-Mackey D., *emcee: The MCMC Hammer*, Publications of the Astronomical Society of the Pacific Vol. 125, No. 925, (2013), arXiv:1202.3665.
- [29] *emcee*. Retrieved February 2019, from <http://dfm.io/emcee/current/>.
- [30] Goodman J., Ware J., *Ensemble Samplers with Affine Invariance*, Commun. Appl. Math. Comput. Sci. Volume 5, Number 1, (2010).

- [31] K. M. Hanson, *Markov Chain Monte Carlo posterior sampling with the Hamiltonian method*, Proc. SPIE4322, pp. 456-467, (2001).
- [32] Radford M. Neal, *MCMC using Hamiltonian dynamics*, Published as Chapter 5 of the Handbook of Markov Chain Monte Carlo, (2011), arXiv:1206.1901v1.
- [33] S. T. Tokdar and R. E. Kass, *Importance sampling: a review*, Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, no. 1, pp. 5460, (2010).
- [34] https://github.com/ja-vazquez/Cosmologia_observacional.git
- [35] D. Baumann, *Cosmology. Part III Mathematical Tripos*, University of Cambridge.
- [36] Wald R. M., *General Relativity*, The University of Chicago Press, (1984).
- [37] A. Liddle, *An Introduction to Modern Cosmology*, John Wiley and Sons Ltd, Second Edition, (2003).
- [38] Vázquez-González A., Matos T., *La materia oscura del universo: retos y perspectivas*, Rev. Mex. Fs. E 54 (2) (2008) 193202.
- [39] Lahav O., Liddle A. R., *The Cosmological Parameters 2014*, Article for The Review of Particle Physics 2014, (2014), arXiv:1401.1389v1.
- [40] Bull P. et al., *Beyond Λ CDM: Problems, solutions, and the road ahead*, Physics of the Dark Universe 12 (2016) 56-99, arXiv:1512.05356v2.
- [41] Ji S. U., Sin S. J., *Late-time phase transition and the galactic halo as a Bose liquid*, Phys. Rev. D, 50, 3650 (1994).
- [42] Guzmán F. S., Matos T., *Scalar fields as dark matter in spiral galaxies: comparison with experiments*, Astron. Nachr., 320, 97.
- [43] Guzmán F. S., Matos T., *Scalar Fields as Dark Matter in Spiral Galaxies*, Class. Quantum Grav., 17, L9-L16 (2000).
- [44] Lee J., Koh I., *Galactic halos as boson stars*, Phys. Rev. D, 53, 2236 (19) (1996).
- [45] Matos T., Vázquez-González A., Magaña J., ϕ^2 as dark matter, 2009, Mon. Not. R. Astron. Soc., 393, 1359-1369.
- [46] Spergel D.N., Steinhardt P.J., *Observational Evidence for Self-Interacting Cold Dark Matter*, Phys. Rev. Lett., 84, 3760 (2000).
- [47] A. De Felice, S. Tsujikawa, *$f(R)$ theories* Living Rev.Rel. 13 (2010) 3, arXiv:1002.4928
- [48] L. G. Jaime, M. Jaber, C. Escamilla-Rivera, *New parametrized equation of state for dark energy surveys*, Phys.Rev. D98 (2018) no.8, 083530, arXiv:1804.04284.
- [49] Nojiri S. et al., *Modified Gravity Theories on a Nutshell: Inflation, Bounce and Late-time Evolution*, Phys.Rept. 692 (2017) 1-104, arXiv:1705.11098v2.
- [50] M. A. García-Aspeitia, A. Hernandez-Almada, J. Magaña, M .H. Amante, V. Motta, C. Martínez-Robles. *Brane with variable tension as a possible solution to the problem of the late cosmic acceleration*, Phys.Rev. D97 (2018) no.10, 101301, arXiv:1804.05085.
- [51] Shinji Tsujikawa, *Quintessence: a review*, Class. Quantum Grav. 30 214003 (2013).
- [52] Yoo J., Watanabe Y., *Theoretical Models of Dark Energy*, Int. J. Mod. Phys. D 21, 1230002, (2012), arXiv:1212.4726.
- [53] Feng B., *The Quintom Model of Dark Energy* (2006), arXiv:astro-ph/0602156.
- [54] Zhao G. et al., *Dynamical dark energy in light of the latest observations*, Nature Astronomy, 1, 627-632, (2017), arXiv:1701.08165.
- [55] O. Akarsu, N. Katirci, N. Ozdemir and J. A. Vazquez, *Anisotropic massive Brans-Dicke gravity extension of standard Λ CDM model*, arXiv:1903.06679.
- [56] M. Sharif and Sehrish Azeem, *Dark Energy Models and Cosmic Acceleration with Anisotropic Universe in $f(T)$ Gravity*, Commun. Theor. Phys. 61 482 (2014).
- [57] Saadeh D. et al., *How isotropic is the Universe?*, Phys. Rev. Lett. 117, 131302 (2016), arXiv:1605.07178.
- [58] Linden S., Virey J., *A test of the CPL parameterization for rapid dark energy equation of state transitions*, Phys.Rev.D78:023526 (2008), arXiv:0804.0389.
- [59] Scherrer R. J., *Mapping the Chevallier-Polarski-Linder parametrization onto Physical Dark Energy Models*, Phys. Rev. D 92, 043001 (2015), arXiv:1505.05781.
- [60] Tamayo D. and Vazquez J.A., *Fourier series expansion of the dark energy equation of state*, (2019), arXiv:1901.08679.
- [61] S. Hee, J. A. Vazquez, W. J. Handley, M. P. Hobson, A. N. Lasenby, *Constraining the dark energy equation of state using Bayes theorem and the Kullback-Leibler divergence*, Mon. Not. Roy. Astron. Soc. 466 (2017) 369, arXiv:1607.00270.

- [62] Aubourg E. et al., *Cosmological implications of baryon acoustic oscillation (BAO) measurements*, Phys. Rev. D 92, 123516 (2015), arXiv:1411.1074.
- [63] Beutler F. et al., *The 6dF Galaxy Survey: Baryon Acoustic Oscillations and the Local Hubble Constant*, MNRAS 416, 3017 (2011), arXiv:1106.3366.
- [64] Planck Collaboration, *Planck 2013 results. XVI. Cosmological parameters*, A&A 571, A16 (2014), arXiv:1303.5076.
- [65] Strauss M.A. et al., *Spectroscopic Target Selection in the Sloan Digital Sky Survey: The Main Galaxy Sample*, AJ 124, 1810 (2002), arXiv:astro-ph/0206225.
- [66] Anderson L. et al., *The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations in the Data Release 10 and 11 galaxy samples*, MNRAS 441, 24 (2014), arXiv:1312.4877.
- [67] Delubac T. et al., *Baryon Acoustic Oscillations an the Ly α forest of BOSS DR11 quasars*, A&A 574, A59 (2015), arXiv:1404.1801.
- [68] Font-Ribera A. et al., *Quasar-Lyman α Forest Cross-Correlation from BOSS DR11: Baryon Acoustic Oscillation*, JCAP 5, 27, 027 (2014), arXiv:1311.1767.
- [69] What Is a Supernova? <https://www.nasa.gov/audience/forstudents/5-8/features/nasa-knows/what-is-a-supernova.html>
- [70] Riess A. G. et al., *Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant*, Astron.J.116:1009-1038 (1998), arXiv:astro-ph/9805201.
- [71] Garnavich P. et al., *Supernova Limits on the Cosmic Equation of State*, Astrophys.J. 509 (1998) 74-79, arXiv:astro-ph/9806396.
- [72] Perlmutter S. et al., *Measurements of Ω and Λ from 42 High-Redshift Supernovae*, Astrophys.J.517:565-586,(1999), arXiv:astro-ph/9812133.
- [73] Betoule M. et al., *Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples*, A&A 568, A22 (2014), arXiv:1401.4064.
- [74] http://supernovae.in2p3.fr/sdss_snls_jla/ReadMe.html#sec-4-1
- [75] http://irfu.cea.fr/en/Phocea/Vie_des_labos/Ast/ast_technique.php?id_ast=2289
- [76] Planck, European Space Agency. https://www.esa.int/Our_Activities/Space_Science/Planck
- [77] Planck Science Team Home, European Space Agency. <https://www.cosmos.esa.int/web/planck/home>
- [78] Looking back to the dawn of time Planck., European Space Agency. <http://people.na.infn.it/~barbarin/MaterialeDidattico/0+approfondimenti%20corso%20Fisica%20astroparticellare/1-CMB/Planck-div.pdf>
- [79] Stern D. et al., *Cosmic Chronometers: Constraining the Equation of State of Dark Energy. II. A Spectroscopic Catalog of Red Galaxies in Galaxy Clusters*, The Astrophysical Journal Supplement, Volume 188, Issue 1, pp. 280-289 (2010), arXiv:0907.3152.
- [80] Stern D. et al., *Cosmic Chronometers: Constraining the Equation of State of Dark Energy. I: H(z) Measurements*, JCAP02(2010) 008, arXiv:0907.3149.
- [81] Moresco M. et al., *Improved constraints on the expansion rate of the Universe up to $z \sim 1.1$ from the spectroscopic evolution of cosmic chronometers*, JCAP 1208 (2012) 006, arXiv:1201.3609.
- [82] Moresco M. et al., *New constraints on cosmological parameters and neutrino properties using the expansion rate of the Universe to $z \sim 1.75$* , JCAP07(2012) 053, arXiv:1201.6658.
- [83] Moresco M., *Raising the bar: new constraints on the Hubble parameter with cosmic chronometers at $z \sim 2$* , Mon.Not.Roy.Astron.Soc. 450 (2015) no.1, L16-L20, arXiv:1503.01116.
- [84] Moresco M., *A 6 % measurement of the Hubble parameter at $z \sim 0.45$: direct evidence of the epoch of cosmic re-acceleration*, JCAP 1605 (2016) no.05, 014, arXiv:1601.01701.
- [85] <https://science.nasa.gov/missions/cobe>
- [86] https://www.nasa.gov/pdf/594754main_Citrin_Liz.pdf
- [87] https://lambda.gsfc.nasa.gov/toolbox/tb_cmbfast_ov.cfm
- [88] <http://adsabs.harvard.edu/abs/2010ascl.soft07004D>
- [89] <http://camb.info/>
- [90] <http://camb.info/readme.html>
- [91] <http://class-code.net/>
- [92] Zumalacarregui Miguel; *CLASS, hi class and Monte Python basics IFT School on Cosmology Tools*; March 11, 17.

- [93] Zuntz J. et al., *CosmoSIS: modular cosmological parameter estimation*, A&C, (2015), 12, 45–59, arXiv:1409.3409.
- [94] <http://baudren.github.io/montepython.html>
- [95] Benjamin Audren, *Monte Python documentation, Release 2.2.0*; October 21, 2015
- [96] <http://cosmologist.info/cosmomc/>
- [97] <https://github.com/slosar/april>