Cosmological Parameter Inference with Bayesian Statistics

Luis E. Padilla [1,2], **Luis O. Tellez** [1,3], **Luis A. Escamilla** [1,3] and **Jose Alberto Vazquez** [1,3,*]

1   Departamento de Física, Centro de Investigación y de Estudios Avanzados del IPN, A.P. 14-740,
    07000 Mexico City, Mexico; epadilla@fis.cinvestav.mx (L.E.P.); ltellez@fis.cinvestav.mx (L.O.T.);
    luis.escamilla@icf.unam.mx (L.A.E.)
2   Department of Astronomy and Texas Cosmology Center, University of Texas, Austin, TX 78712-1083, USA
3   Instituto de Ciencias Físicas, Universidad Nacional Autónoma de Mexico, Apdo. Postal 48-3,
    62251 Cuernavaca, Morelos, Mexico
*   Correspondence: javazquez@icf.unam.mx

Abstract: Bayesian statistics and Markov Chain Monte Carlo (MCMC) 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 fundamental concepts to understand Bayesian statistics and then introduce MCMC algorithms and samplers that allow us to perform the parameter inference procedure. We also introduce a general description of the standard cosmological model, known as the ΛCDM model, along with several alternatives, and current datasets coming from astrophysical and cosmological observations. Finally, with the tools acquired, we use an MCMC algorithm implemented in python to test several cosmological models and find out the combination of parameters that best describes the Universe.

Keywords: Observational Cosmology; Bayesian statistics; parameter inference; model comparison; dark energy; dark matter

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 datasets, 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, that the BB model awarded was the most accepted. 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, usually done through statistics. It is important to stress out that, since we have a unique 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 approach that will help in our task is the 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 that the parameter inference theory may be applicable to any subject, in particular we put into practice the Bayesian concepts into the field of cosmology.

The paper is organized as follows. In Section 2, we point out the main differences between the Bayesian and Frequentist approaches. Then, in Section 3, we introduce the basic 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 numerical resources that are able to simplify our task, especially for models with several parameters that need to be tested with many datasets. With these methods and tools in place, we provide the example of fitting a straight line in Section 5. In Section 6, we present an introduction to cosmology, and then, in Section 7, focus on some codes to compute the cosmological observables. In Section 8, we constrain the parameter space that describes the standard cosmological model, namely the ΛCDM model, along with several extensions. 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

$$P = \frac{n}{N}, \tag{1}$$

where n denotes the number of successes, and N the total number of trials. Frequentist statistics defines the probability P as the limit of 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 concerning an event.

Here, we introduce some key concepts to understand the consequences this difference entails; for an extended review see Reference [3–13] and references therein. 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 apply[1]:

$$P(x) \geq 0, \tag{2}$$

$$\int_{-\infty}^{\infty} dx\, P(x) = 1. \tag{3}$$

For mutually exclusive events, we have

$$P(x_1 \cup x_2) = P(x_1) + P(x_2), \tag{4}$$

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 follows: the first condition (2) is necessary due to the probability of having an event is always positive; the second rule (3) is a normalized relation, which tells us that we are certain to obtain one of the possible outcomes; now, in the third point (4) 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

$$P(x_1 \cap x_2) = P(x_1)P(x_2|x_1). \tag{5}$$

If two events x_1 and x_2 are mutually exclusive, then

$$P(x_1 \cap x_2) = 0 = P(x_2 \cap x_1). \tag{6}$$

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

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"

$$X \sim P_0.$$

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 being 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. Therefore, the Frequentist statistics is based on trying to answer the following questions: "what the data is trying to tell 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 when thinking in a Frequentist way, 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 1 provides a short summary of the most important differences between the two statistics. To understand these differences, let us review a basic 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.

Table 1. Main differences between the Bayesian and Frequentist interpretations.

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.

Example 1. *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. Our goal is to know whether this coin is fair ($p = 0.5$) or not. In the process of trying to estimate p, 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 approach. As mentioned previously, in Frequentist statistics probability is related to the frequency of events, then our best estimate for p is $P(head) = p = \frac{\# \, of \, heads}{\# \, of \, events} = 10/14$. So, the probability of having 2 heads in a row is $P(2heads) = P(head)P(head) \simeq 0.51$.*
- *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 (the 14 trials and 10 successes). Then, by considering that we do not know anything about p a priori and averaging over all possible values of p, we have that the probability of having two heads is*

$$P(2heads|D) = 0.485. \tag{7}$$

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 ($P(2heads) = 0.25$) because we do not have enough data for our estimations.

In this example, we have seen the forward application of statistics: using a mathematical model to relate measured quantities to an unknown quantity of interest, in this case, the probability of getting two heads in a row. This can also be applicable to the inverse problem. That is, by the having the data, we would like to obtain information about the parameters of a given model [4,14]. In Section 8, we will illustrate this point by using different cosmological observations and finding out the best-fit values of the parameters that describe a given model.

3. A First Look at Bayesian Statistics

Before we start with the applications of Bayesian statistics in cosmology, it is advisable to understand the important mathematical tools within the Bayesian procedure. In this section, we present a basic 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 in the Bayesian framework, there are several concepts to understand before presenting the results. In this section, we review these concepts, and then we get back to the example of 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 Equations (2)–(5). From Equation (5), without loss of generality, it must be fulfilled that $P(x_1 \cap x_2) = P(x_2 \cap x_1)$. In such a case, the following relation applies:

$$P(x_2|x_1) = \frac{P(x_1|x_2)P(x_2)}{P(x_1)}. \tag{8}$$

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

$$P(\theta|D, H) = \frac{P(D|\theta, H)P(\theta|H)}{P(D|H)}. \tag{9}$$

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. $L(D|\theta, H) \equiv P(D|\theta, H)$ is called the likelihood, and it will be our main focus in future sections, $\pi(\theta) \equiv P(\theta|H)$ 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), $\mathcal{Z} \equiv P(D|H)$ is the evidence of the model, usually referred to as the Bayesian Evidence.

The prior refers to the information one has a priori of the model. It can be defined in various ways; however, a common one is the uniform prior (also referred to as a flat prior):

$$\pi(\theta) \propto c, \tag{10}$$

with c being a constant. This type of prior is telling us that every parameter value is equally probable a priori, as seen in Figure 1. Using this prior also means that the posterior probability will be proportional to the likelihood (since the Bayesian Evidence is a constant). Another convenient prior distribution is the *beta distribution* $B(\theta; a, b)$ since it contains several statistical distributions by varying its parameters a and b (in particular the flat prior is obtained when $a = b = 1$). It is defined as

$$B(\theta; a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1 - \theta)^{b-1}, \tag{11}$$

where Γ is the gamma function. These are just two examples of useful priors, and it is evident that the choice of a prior will influence the posterior distribution, although its effect is reduced as more data are collected, as we shall see later in this section.

Now, regarding the Bayesian Evidence, we notice that it acts as a normalizing factor, and is nothing more than the average of the likelihood over the

$$P(D|H) = \int d^N\theta \, P(D|\theta, H)P(\theta|H), \tag{12}$$

where N is the dimensionality of the parameter space. This quantity is usually ignored, for practical reasons, i.e., 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, this process being known as *model selection*. For convenience, the ratio of two evidences

$$K \equiv \frac{P(D|H_0)}{P(D|H_1)} = \frac{\int d^{N_0}\theta_0 \, P(D|\theta_0, H_0)P(\theta_0|H_0)}{\int d^{N_1}\theta_1 \, P(D|\theta_1, H_1)P(\theta_1|H_1)} = \frac{\mathcal{Z}_0}{\mathcal{Z}_1}, \tag{13}$$

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

$$\mathcal{B}_{0,1} = \ln \frac{\mathcal{Z}_0}{\mathcal{Z}_1}, \tag{14}$$

where θ_i is a parameter vector (with dimensionality N_i) for the hypothesis H_i and $i = 0, 1$. In Equation (14), the quantity $\mathcal{B}_{0,1} = \ln K$ provides an idea on how well model 0 may fit the data when compared to model 1. Jeffreys provided a suitable guideline scale on which we are able to make qualitative conclusions (see Table 2 [15]).

Table 2. Jeffreys guideline scale for evaluating the strength of evidence when two models are compared.

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

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 then interpret it with a given model; however, we usually do the opposite. That is, first we have a set of data, and then we confront a model considering the probability that our model fits the data. Bayes' theorem provides a tool to relate both scenarios. Then, based on the Bayes' theorem, we are able to select the model that best fits the data.

Example 2. *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 $P(2heads|D)$ (D being the previous 14 coin tosses acting as data). First, let us assume that we have a model with a parameter p that defines the probability of obtaining the two heads, that is $P(2heads|p)$. This parameter p will have a probability distribution $P(p|D)$ depending on the data in place. Therefore, the probability can be obtained by averaging over all the possible parameters with its corresponding density distribution*

$$P(2heads|D) = \int_0^1 P(2heads|p)P(p|D)dp. \tag{15}$$

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

$$P(2heads|p) = [P(head|p)]^2, \tag{16}$$

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

$$P(head|p) = p \quad \Rightarrow \quad P(2heads|p) = p^2. \tag{17}$$

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 outcomes is the binomial distribution

$$P(x|p, n) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{18}$$

with n the number of trials (in this case, = 14) and x the number of successes (here =10). Hence, we have an expression for $P(D|p)$. Using the Bayes' formula, we have

$$P(p|D) = \frac{P(D|p)P(p)}{P(D)}. \tag{19}$$

For the prior, we will use the beta distribution (11), so

$$P(p) = B(p; a, b). \tag{20}$$

In order to get the explicit form of $P(p|D)$, we still need to compute $P(D)$. That is, by plugging Equations (18) *and* (20) *into the integral of Equation* (12) *yields to*

$$P(D) = B(10+a, 4+b) \equiv \frac{\Gamma(10+a)\Gamma(4+b)}{\Gamma((10+a)+(4+b))}; \tag{21}$$

therefore,

$$P(p|D) = \frac{p^{10+a-1}(1-p)^{4+b-1}}{B(10+a, 4+b)}. \tag{22}$$

If we know nothing about p, then we can assume the prior is a uniform distribution; this means $a = b = 1$. Notice from Figure 1 *that our posterior result (Red figure) described by Equation* (22) *does not exactly agree with the real value of p (black dashed vertical line). We would expect the posterior distribution to 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 Equation (15) *using* (17) *and* (22)*, we arrive at the result obtained in the previous section*

$$P(2heads|D) = \frac{B(13,5)}{B(11,5)} = 0.485. \tag{23}$$

It is important to clarify that the inferred value for the parameter p is merely the probability of said value given our data D. This value can change and will generally be a better estimation if more data are collected (as will be seen in detail in the next section).
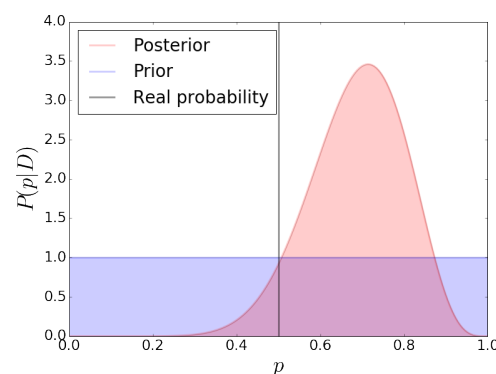

Figure 1. The coin example: blue figure displays the prior distribution $P(p)$ which is updated, when the data is taken into account, to get the posterior distribution $P(p|D)$ (red). The vertical black line corresponds to the real value, $p = 0.5$.

3.2. Updating the Probability Distribution

As seen in the coin example, we were not able to get the real value of p because of the lack of enough data. If we want to get 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 us say 56 heads, while after throwing it 500 times, we obtain 246 heads. Then, we expect to obtain a thinner distribution with its center close to $p = 0.5$ (see Figure 2). Given this, it is clear that in order to confront a parameter 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 $P(2heads|D) = 0.249$, 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 set of all datasets; or (b) by taking each dataset 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 Figure 2a, and then, by considering the 500 data points, we can arrive at the posterior in Figure 2d, or similarly start with the posterior shown in Figure 2c as our prior and consider only the last 400 data points to obtain the same posterior, displayed in Figure 2d.
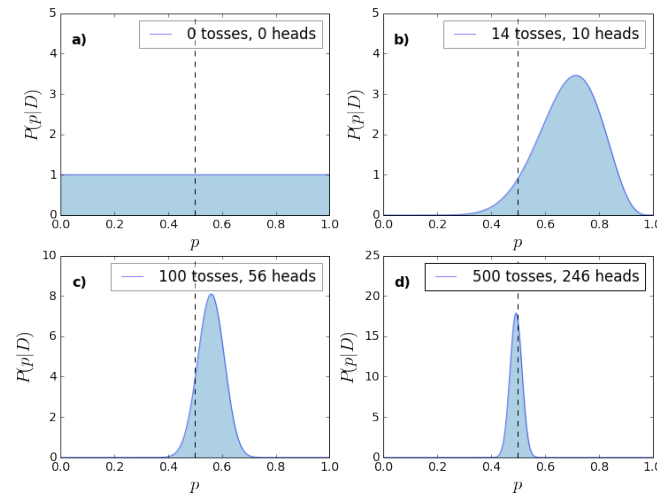

Figure 2. Posterior distributions $P(p|D)$, when: (**a**) there is no available data, (**b**) after 14 coin tosses of which 10 were heads, (**c**) after 100 coin tosses of which 56 were heads, (**d**) after 500 tosses of which 246 were heads. Notice that, while we continue increasing the experimental results, the posterior distribution starts to be more localized near by the real value $p = 0.5$.

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

$$P(\theta|DI, H) = \frac{P(\theta|I, H)P(DI|\theta, H)}{P(D|I, H)}, \tag{24}$$

and then consider a new set of data D', letting the old data become part of the prior information $I' = DI$, we arrive at

$$
\begin{aligned}
P(\theta|D'I', H) &= \frac{P(\theta|I, H)P(DD'I|\theta, H)}{P(DD'|I, H)} \\
&= P(\theta|[DD']I, H), \tag{25}
\end{aligned}
$$

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 prior (a motivation for doing this will be expanded upon in the next section), we can identify the posterior with the likelihood $P(\theta|D, H) \propto L(D|\theta, H)$; 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

$$\Lambda = -2\ln\left[\frac{L(D|\theta, H)}{L_{\max}}\right], \tag{26}$$

is bigger than a given value.

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

$$\hat{\theta} = \int d\theta\, \theta P(\theta|D, H). \tag{27}$$

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

$$\langle \hat{\theta} \rangle = \theta_0, \tag{28}$$

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

$$
\begin{aligned}
\ln L(D|\theta) \;=\; & \ln L(D|\theta_0) \\
&+\; \frac{1}{2}(\theta_i - \theta_{0i})\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}(\theta_j - \theta_{0j}) \\
&+\; ...,
\end{aligned}
\tag{29}
$$

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

$$L(D|\theta) = L(D|\theta_0) \exp\left[-\frac{1}{2}(\theta_i - \theta_{0i})H_{ij}(\theta_j - \theta_{0j}) \right], \tag{30}$$

where

$$H_{ij} = -\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j}, \tag{31}$$

is called the Hessian matrix. It controls whether the estimates of θ_i and θ_j are correlated, and 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 tells 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 Reference [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 a similar 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 measure 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

$$L(D|\theta, HI) = L_0 \exp\left[-\frac{1}{2}\frac{(\theta - m_1)^2}{\sigma^2} \right]. \tag{32}$$

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

$$P(\theta|m_1) = \frac{L(m_1|\theta I_i)P(\theta|I_i)}{P(m_1|I_i)}, \tag{33}$$

where we have skipped writing explicitly the hypothesis H and used the notation given in Equation (24). Then, the posterior of A and B are (again) Gaussian with mean

$$\hat{\mu}_i = \frac{m_1 + (\sigma/\Sigma_i)^2 \mu_i}{1 + (\sigma/\Sigma_i)^2}, \tag{34}$$

and variance

$$\tau_i^2 = \frac{\sigma^2}{1 + (\sigma/\Sigma_i)^2}, \quad (i = A, B). \tag{35}$$

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.
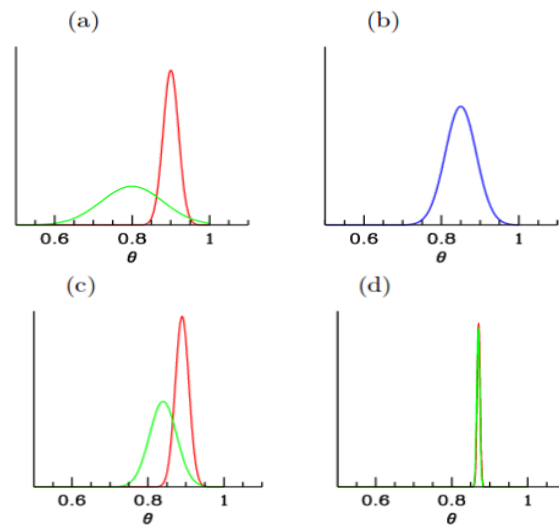

Figure 3. Converging views in Bayesian inference (taken from Reference [5]). A and B have different priors $P(\theta|I_i)$ for a value θ (panel (**a**)). Then, they observe one datum with an apparatus subject to a Gaussian noise and they obtained a likelihood $L(\theta; HI)$ (panel (**b**)), after which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

3.5. Chi-Square and Goodness of Fit

We mentioned that 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 Equation (30), we can see the likelihood will be maximum if the quantity

$$\chi^2 \equiv (\theta_i - \theta_{0i}) H_{ij} (\theta_j - \theta_{0j}), \tag{36}$$

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 $\hat{\chi}$ for the correct model. This

probability is given by $Q(v, \hat{\chi}) = 1 - \Gamma(v/2, \hat{\chi}/2)$ [16], 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.

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 Equation (30)) these are ellipsoidal regions. In this section, we exemplify how to calculate the confidence regions, following Reference [6].

We consider a small perturbation from the best fit of chi-square $\Delta\chi^2 = \chi^2 - \chi^2_{\text{best}}$. Then, we use the properties of χ^2 distribution to define confidence regions for variations on χ^2 to its minimum. In Table 3, 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, 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 with the coin example.

Table 3. $\Delta\chi^2$ for the conventional 68.3%, 95.4%, and 99.73% as a function of the number of parameters (M) for the joint confidence level.

			$\Delta\chi^2$	
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

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 the confidence limit desired.
2. Solve the equation:

$$Q(n - M, min(\chi^2) + \Delta\chi^2) = p. \tag{37}$$

3. Find the parameter region where $\chi^2 \leq min(\chi^2) + \Delta\chi^2$. This defines the confidence region.
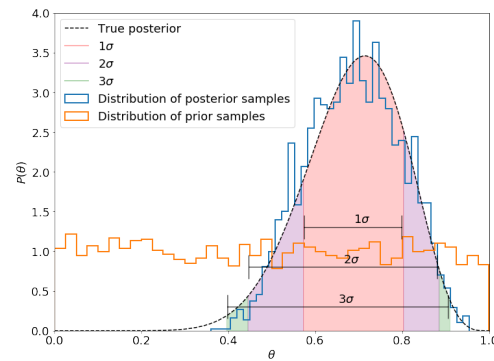
Figure 4. One-dimensional posterior distribution for our example. We plot the prior distribution (red), true posterior (dashed-black), and the posterior calculated by the MHA (blue). We plot 1, 2 and 3σ confidence regions for the estimation of p.

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

$$P(\theta_1,...,\theta_j, H|D) = \int d\theta_{j+1}...d\theta_m P(\theta, H|D), \tag{38}$$

where m is the total number of parameters in our model, and $\theta_1,...,\theta_j$ 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 [17]. Let us start by considering 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

$$F_{ij} = -\left\langle \frac{\partial^2 \mathcal{L}}{\partial \theta_i \partial \theta_j} \right\rangle, \tag{39}$$

where

$$\mathcal{L} = \ln L. \tag{40}$$

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

As we can see from Equation (4), for independent datasets, 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,

$$\Delta\mathcal{L} = \frac{1}{2} F_{ii} (\theta_i - \theta_{0i})^2, \tag{41}$$

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,

$$\sigma_{ij}^2 \geq (F^{-1})_{ij}. \tag{42}$$

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

$$\sigma_{\theta_i} \geq (F^{-1})_{ii}^{1/2}. \tag{43}$$

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 calculation 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 Equation (42) 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 Section 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.

Constructing Fisher Matrices: A Simple Description

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

$$F_{ij} = \sum_k \frac{1}{\sigma_k^2} \frac{\partial f_k}{\partial \theta_i} \frac{\partial f_k}{\partial \theta_j}, \tag{44}$$

where σ_k are the errors associated with each observable, and we have considered they are 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 with the mock data can be taken as the expected experimental errors, and then it is 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) [18], 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. The figure of merit can also be used to see how different experiments break degeneracies, and 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 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 the Reference [19].

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

$$
\begin{aligned}
\mu_f = \int f(x)p(x)dx &= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\
&= E_q[f(X)w(x)],
\end{aligned}
\tag{45}
$$

where $w(x) = p(x)/q(x)$, and now we have an average over q. So, if we have a collection of different draws $x^{(1)}, ..., x^{(m)}$ from $q(x)$, we can estimate μ_f using these draws as

$$
\hat{\mu}_f = \frac{1}{m}\sum_{j=1}^{m} w(x^{(j)})f(x^{(j)}).
\tag{46}
$$

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

$$
\hat{\mu}_f = \frac{\sum_{j=1}^{m} w(x^{(j)})f(x^{(j)})}{\sum_{j=1}^{m} w(x^{(j)})}.
\tag{47}
$$

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

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

$$
\frac{P'(D)}{P(D)} = E\left[\frac{P'(\theta, D)}{P(\theta, D)}\right]_{P(\theta|D)} \simeq \frac{1}{N}\sum_{n=1}^{N} \frac{P'(D|\theta_n)P'(\theta_n)}{P(D|\theta_n)P(\theta_n)},
\tag{48}
$$

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 $\{D_1, \ldots, D_N\}$, coming from a collection of different surveys $\{S_1, \ldots, S_N\}$. 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 Section 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 the data may be is by introducing the hyperparameter method. This method was initially introduced by Reference [20,21] 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 of each other. In this section, we review the main steps necessary to understand the hyperparameter method.

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.,

$$
P(\theta|D, H) = \frac{1}{P(D|H)}\int P(\theta|\alpha, H)P(\alpha|D, H)d\alpha,
\tag{49}
$$

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

of each other, i.e., $P(\alpha) = P(\alpha_1)P(\alpha_2)\ldots P(\alpha_N)$. In this way, we can rewrite the above expression as

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

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 dataset 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 Equation (13). If we consider a Gaussian likelihood with maximum entropy prior, and assuming that on average the hyperparameters' weights are unity, we can rewrite the marginalized likelihood function $L(D|\theta,H_1)$ for model H_1 as

$$P(D|\theta,H_1) = \prod_{k=1}^{N} \frac{2\Gamma(\frac{n_k}{2}+1)}{\pi^{n_k/2}|V_k|^{1/2}} (\chi_k^2 + 2)^{-\left(\frac{n_k}{2}+1\right)}, \tag{50}$$

obtaining an explicit functional form for K, given by

$$K = \prod_{k=1}^{N} \frac{2^{n_k/2+1}\Gamma(n_k/2+1)}{\chi_k^2 + 2} e^{-\chi_k^2/2}. \tag{51}$$

Here, χ_k^2 is given by (36) for every dataset, and n_k is the number of points contained in D_k. In Equation (50), 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 Equation (51), we can estimate the necessity to introduce the hyperparameters to our model using the criteria given in Table 2. 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 (48).

4. Numerical Tools

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

4.1. MCMC Techniques for Parameter Inference

The purpose of the MCMC algorithm is to build up a sequence of points (called **chain**) in a parameter space in order to evaluate the posterior of Equation (9). In this section, we review the basic steps for this procedure in a simplistic way; however, it is recommendable to check [22–26] for a more formal version of the MCMC 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 $X_1, X_2, \ldots$ of elements of some set is a Markov Chain if the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ 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 $P(\theta|D, H)$. Hence, we can estimate all the usual quantities of interest out of the posterior (mean, variance, etc.).

The combination of both procedures is called an 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:

$$p(\theta|D, H) \simeq \frac{1}{N} \sum_{i=1}^{N} \delta(\theta - \theta_i), \tag{52}$$

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

$$\langle \theta \rangle = \int d\theta \theta P(\theta, H|D) \simeq \frac{1}{N} \sum_{i=1}^{N} \theta_i, \tag{53}$$

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

$$\langle f(\theta) \rangle \simeq \frac{1}{N} \sum_{i=1}^{N} f(\theta_i). \tag{54}$$

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 criterion for accepting (or rejecting) this new point depending on whether it turns out to be better for our model or not. If the new step is worse than the previous one, we may still accept it since, if we only accept steps with better probability, we could be converging into a local maximum in our parameter space and, therefore, just exploring a small region of the entire space. 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 [27,28], 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 of the new point is given by

$$p(acceptance) = min\left[1, \frac{p_c q(\theta_c, \theta_i)}{p_i q(\theta_i, \theta_c)}\right]. \tag{55}$$

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

$$p(acceptance) = min\left[1, \frac{p_c}{p_i}\right]. \tag{56}$$

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 reject) the new point with the help of the Metropolis-Hastings algorithm.
4. If the point is rejected, 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 put into practice the numerical tools, we take again the coin toss example seen from Section 3.1. Let us try to estimate the value of p (or region of values for p) that best matches our data (we assume only the 14 times that the coin was thrown). To calculate the posterior distribution (20), we use the MHA.

As mentioned before, we consider a likelihood given by the binomial distribution (18) and a normal distributed prior (11) ($a = b = 1$). 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. Our final result (shown in Figure 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 (and Figure 4), we also show in Figure 5 the Markov Chains generated by our code. It is easy to see that the chains oscillate with small amplitude around the mean value.
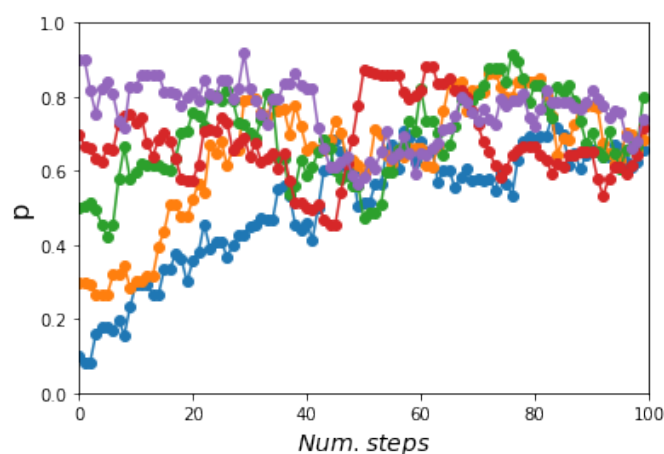

Figure 5. The five Markov Chains used to estimate the posterior distribution. We use $p = 0.1, 0.3, 0.5, 0.7, 0.9$ for starting points.

Remark: In Figure 6, 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 [29] 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.

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. In this sense, we need that our algorithm takes into account this possible difficulty. The simplest way (the informal way) to know whether our chain is converging to a global maximum is by running several chains starting with different initial proposals for the parameters. Then, if we see by the naked eye that all 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 whether the value (region) of p that we found is a stationary value. In Figure 5, we plot 5 different Markov chains with initial conditions $p = 0.1, 0.3, 0.5, 0.7, 0.9$. As we expected from the analytical result, after several steps all the chains seem to concentrate nearby the same value.

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

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

```python
'''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)
```

```python
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
```

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

Figure 6. Here, we show our MCMC code, written in Python, from scratch.

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 [30,31]. 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:

$$\langle \theta^j \rangle = \frac{1}{N} \sum_{i=1}^{N} \theta_i^j, \tag{57}$$

and the mean of all the chains

$$\langle \theta \rangle = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \theta_i^j. \tag{58}$$

Then, the chain-to-chain variance B is

$$B = \frac{1}{M-1} \sum_{j=1}^{M} (\langle \theta^j \rangle - \langle \theta \rangle)^2, \tag{59}$$

and the average variance of each chain is

$$W = \frac{1}{M(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{M} (\theta_i^j - \langle \theta^j \rangle)^2. \tag{60}$$

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

$$\hat{R} = \frac{\frac{N-1}{N}W + B(1 + \frac{1}{M})}{W}, \tag{61}$$

which is the ratio of the two estimates, approaches unity. A typical convergence criterion 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 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 7 (top 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 (bottom panel of the figure corresponding to $\hat{\sigma} = 0.8$), so we will have different maxima in our posterior.
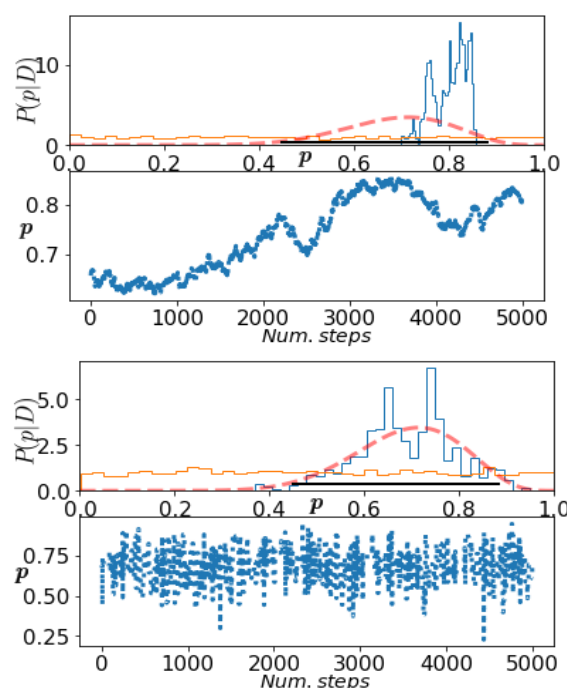

Figure 7. Two Markov Chains considering different variance for our Gaussian proposal distribution. The upper panel corresponds to $\hat{\sigma} = 0.002$, while the lower panel corresponds to $\hat{\sigma} = 0.8$.

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 final MCMC.

The burn-in: It is important to notice that, at the beginning of the chain, we have a set of points far outside the stationary region. This early stage of the chain (called burn-in)

must be ignored; this means that the dependence 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, which 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 [32]. 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 ones, than using the full chains.

Autocorrelation probes: A complementary way to look for convergence in an 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 [33,34]

$$
\begin{aligned}
\rho_k &= \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t)Var(X_{t+k})}} \\
&= \frac{E[(X_t - X)(X_{t+k} - X)]}{\sqrt{E[(X_t - X)^2]E[(X_{t+k} - X)^2]}},
\end{aligned}
\tag{62}
$$

where X_t is the t-th sample, and 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 [35] 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 $\theta_1, \ldots, \theta_k$. Then, at step i
2. Draw θ_1^{i+1} from $p(\theta_1 | \theta_2^i, \ldots, \theta_k^i)$
3. Draw θ_2^{i+1} from $p(\theta_2 | \theta_1^{i+1}, \theta_3^i, \ldots, \theta_k^i)$
4. $\ldots$
5. Draw θ_k^{i+1} from $p(\theta_k | \theta_1^{i+1}, \theta_2^{i+1}, \ldots, \theta_{k-1}^{i+1})$
6. Repeat the above steps for the wished iterations with $i \rightarrow i + 1$.

The distribution $p(\theta_1 | \theta_2, \ldots, \theta_k) = \frac{p(\theta_1, \ldots, \theta_k)}{p(\theta_2, \ldots, \theta_k)}$ 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. The temperature controls the height of the peaks in the likelihood; this simplifies the way we sample our parameter space and helps us to avoid local maxima. If a distribution has a temperature $T < 1$, it is said to be tempered.

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

$$
P(\theta, T|D, H) \propto L(\theta, D)^{1/T} P(\theta, H),
\tag{63}
$$

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 an MCMC algorithm. Now, we have to run N chains with different temperatures assigned in a ladder $T_1 < T_2 < \ldots < T_N$, 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 probability

$$A_{i,j} = min\left\{ \left(\frac{L(\theta_i)}{L(\theta_j)}\right)^{1/T_j - 1/T_i}, 1 \right\}. \tag{64}$$

In this way, if a swap is accepted, chains i and j must exchange their current position in the parameter space, and 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. For an extensive explanation, or modification to make the temperature of the chains dynamical, see Reference [36–38].

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

$$p(x_1, x_2) \propto \exp\left(\frac{-(x_1 - x_2)^2}{2\epsilon} - \frac{(x_1 + x_2)^2}{2} \right), \tag{65}$$

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

$$y_1 = \frac{x_1 - x_2}{\sqrt{\epsilon}}, \quad y_2 = x_1 + x_2, \tag{66}$$

we can rewrite the anisotropic density into the easier problem

$$p(y_1, y_2) \propto \exp\left(\frac{-(y_1^2 + y_2^2)}{2} \right). \tag{67}$$

An 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 $Ax + b$,

$$R(AX(t) + b, \psi(t), p_{A,b}) = AR(X(t), \psi(t), p) + b. \tag{68}$$

There are already several algorithms that are affine invariant, one of the easiest is known as the *stretch move* [39]. An algorithm fully implemented in Python under the name EMCEE [40] is also affine invariant, and some others that can be found in Reference [41].

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 to accomplish our mission, for instance, some of the most popular and effective ones, are the Hamiltoninan Monte Carlo (see, e.g., Reference [42,43]) or the Adaptative Metropolis-Hastings (AMH) (see, e.g., Reference [19]).

5. Fitting a Straight-Line

In this section, we carry out an standard example: fitting a straight-line. That is, we assume that we have a certain theory where our measurements should follow a straight line. Then, we simulate several datasets and focus on two different cases (Figure 8):

1. Consider two datasets coming from the same straight-line but having different errors.
2. Consider two datasets coming from different straight-lines and also having different errors.

In our analysis, we used the PyMC3 module, and the complete code can be downloaded from the Git repository [44].
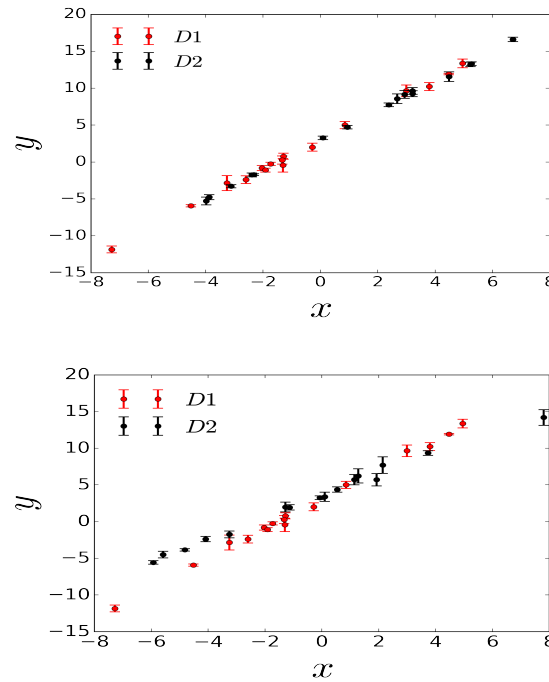

Figure 8. Datasets D_1 and D_2 measured by the straight-line theory. Top: case 1. Bottom: case 2.

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 the upper panel of Figure 8. The two datasets, D1 and D2, were generated from the same line $y = 3 + 2x$, adding a gaussian error to each point. For D1, we added 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 analyze this data with and without the hyperparameter method and discuss in detail our results.

Model H_0: without hyperparameters.

Before we make a Bayesian estimation, it is necessary to specify the priors. As we have seen, a good prior is a non-informative one. Suppose we only know some limits for a and b, as can be seen when plotting the data. Then, we consider the flat priors

$$a \propto U[0,5] \quad \text{and} \quad b \propto U[0,3],$$ (69)

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

From Equation (30), we can write our likelihood as

$$L(D; line) \propto \exp\left[-\sum_d \frac{(y_d - y)^2}{2\sigma_d^2}\right],$$ (70)

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 the Markov chains. In our analysis, we ran 5 chains with 10,000 steps for each one. We ran each chain with temperature $T = 2$ and we thinned them every 50 steps. The result corresponds to $a = 2.982 \pm 0.047$ and $b = 1.994 \pm 0.013$, and their posterior distributions are plotted in Figure 9. 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 more likely to 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 achieved. The bottom panel of Figure 9 displays the $1 - 4\,\sigma$ confidence regions. We also added a point in red to show the real value for our parameters. The real value for a and b are within the curve corresponding to one standard deviation of our estimations in the inferential method.

As we mentioned, we need the autocorrelation to be small as k increases in order to consider that our analysis is converging. We see in Figure 10 such plots and notice that our convergence criterion is fulfilled. Then, in Case 1, we can see that the model H_0 seems to be a very good estimation procedure.
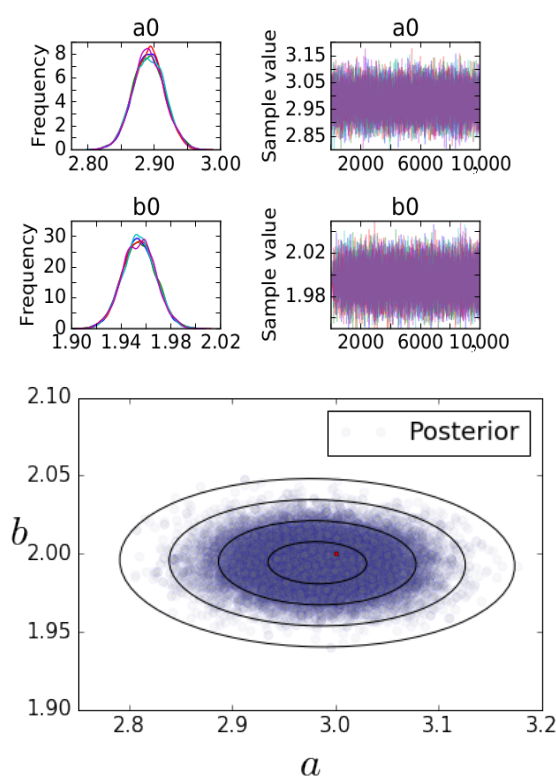

Figure 9. Top panel: One-dimensional marginalized posterior distributions for our samples and the Markov chains for model H_0. Bottom panel: Two-dimensional marginalized posterior distributions, along with 1–4 confidence regions for our parameters for model H_0. The red point corresponds to the true value.
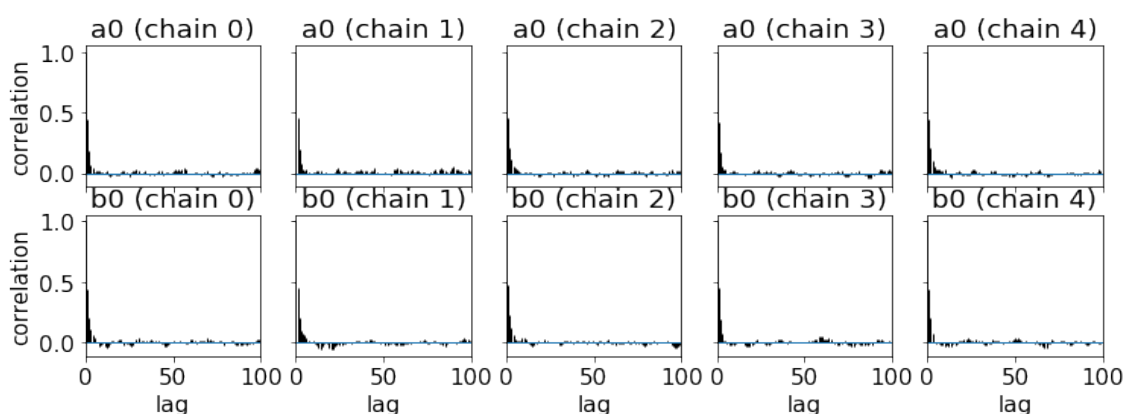

Figure 10. Autocorrelation plots for model H_0.

Model H_1: with hyperparameters.

Now, let us consider the Hyperparameter method. In this case, our likelihood can be written as Equation (50). Similar to the last procedure, we compute the posterior with flat priors, using 5 chains with 10,000 steps for each one, and check for autocorrelations. Our results are as follows: $a = 2.97 \pm 0.038$ 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, Figure 9 and the top panel of Figure 11 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 Equation (51)

$$K = 3. \tag{71}$$

Then, comparing with Table 2, we can say that the evidence for H_1 to be better than H_0 is weak. In such a case, it should be equally better to work with H_0 as to H_1, as explained before.

Finally, in order to exemplify our results, let us plot in the bottom panel of Figure 11 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 lower panel of Figure 8. These correspond to our dataset D_1 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 the parameter estimation for the parameters a and b and look for the differences in both procedures.
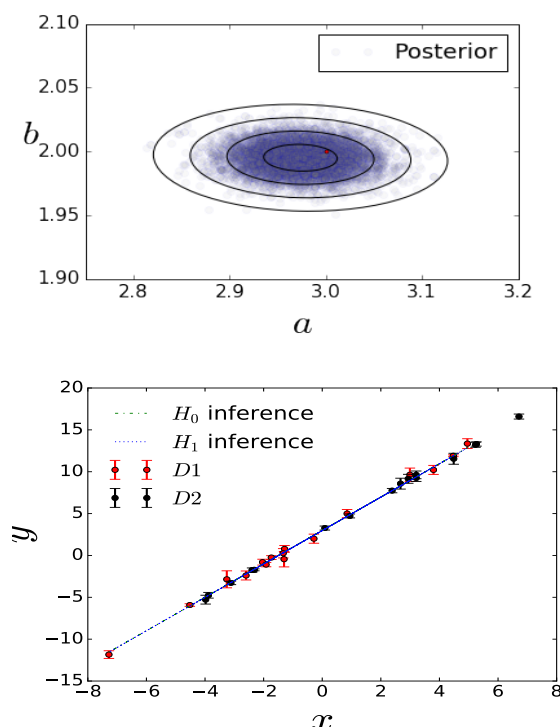

Figure 11. Upper panel: Confidence regions for the parameters in model H_1. Lower panel: The best straight-lines inferred, along with the data.

Model H_0: without hyperparameters.

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 autocorre-

lation plots. Our results are the following: $a = 3.528 \pm 0.056$ and $b = 1.795 \pm 0.014$. Then, we plotted $1 - 4\sigma$ confidence regions in the upper panel of Figure 12. It is easy to see that our estimation differs so much from the real parameters of the datasets (red points). This is because we are trying to fit a model with non-auto-consistent datasets; therefore, we arrive at incorrect results. Now, let us see what happens in the hyperparameters procedure.

Model H_1: With hyperparameters.

In the bottom-left panel of Figure 12, we plotted the 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).

Given the fact that we know a priori the real values of the parameters, we could immediately say that the method with hyperparameters is a better approximation than the case without them. However, we confirm this assumption by computing the ratio K between both models. We then obtain

$$K = 37, \tag{72}$$

which means that we have very strong evidence that H_1 is better that H_0. Finally, we can plot the straight-line inferred by model H_0 and the two lines inferred by model H_1. Considering the parameters inside the two regions in the bottom-left of Figure 12, we obtain the bottom-right panel of the same Figure.
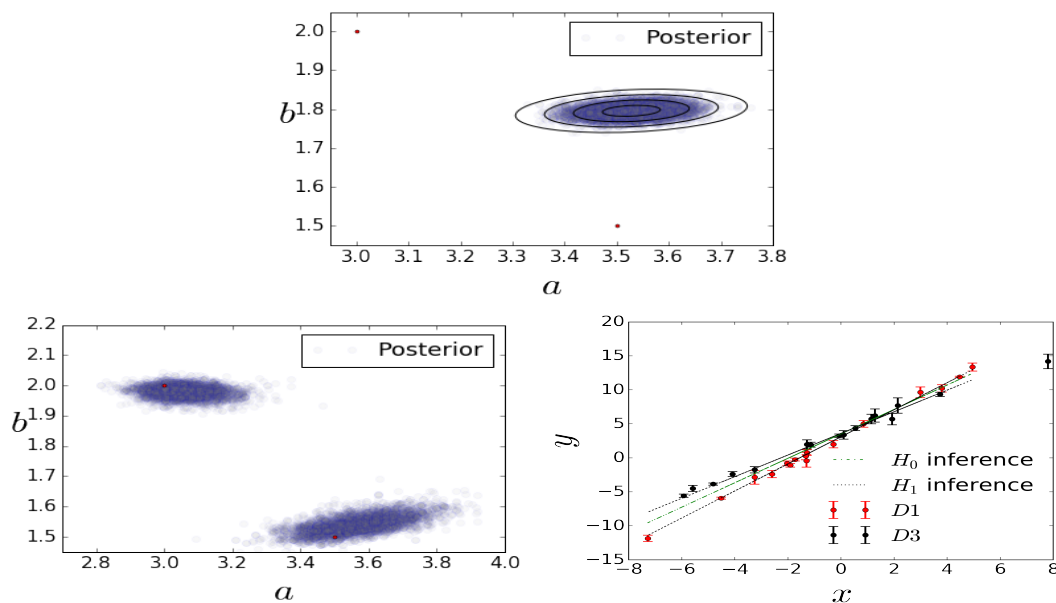

Figure 12. Upper panel: Confidence regions for the parameters in model H_0. Bottom-left panel: Confidence regions for the parameters in model H_1. Bottom-right panel: Best-fit values for the straight-lines for Case 2 inferred with the datasets shown.

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 and apply the Bayesian statistics to perform the parameter inference. In our examples, we will focus on the background Universe—and avoid perturbations—since the main purpose of this article is the application 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—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 with 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

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{73}$$

where $G_{\mu\nu}$ is the Einstein tensor (geometry of the spacetime), and G is the gravitational Newton constant [45–48]. Throughout this review, we use natural units $c = \hbar = 1$.

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

$$ds^2 = g_{\mu\nu}dx^\mu dx^\nu, \tag{74}$$

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 stated otherwise, greek letters μ, ν, ... denote spacetime indices ranging from 0 to 3, while latin letters i, j, ... 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

$$ds^2 = dt^2 - a^2(t)\gamma_{ij}dx^i dx^j, \tag{75}$$

where

$$\gamma_{ij} \equiv \delta_{ij} + \kappa \frac{x_i x_j}{1 - \kappa\left(x_k x^k\right)}. \tag{76}$$

In Equation (75), *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 expression (76), x^i labeled 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; hence, here it is described by the energy-momentum tensor of a perfect fluid

$$T_{\mu\nu} = (\rho + P)U_\mu U_\nu - P g_{\mu\nu}, \tag{77}$$

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

$$T^\mu_\nu = g^{\mu\lambda} T_{\lambda\nu} = \begin{pmatrix} \rho & 0 & 0 & 0 \\ 0 & -P & 0 & 0 \\ 0 & 0 & -P & 0 \\ 0 & 0 & 0 & -P \end{pmatrix}. \tag{78}$$

Using Equations (73) and (78), with the FLRW metric, we can obtain the Friedmann equations

$$H^2 \equiv \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3} \sum_i \rho_i - \frac{\kappa}{a^2}, \tag{79}$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \sum_i (\rho_i + 3P_i). \tag{80}$$

In these expressions, H accounts for the rate of expansion/contraction of 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 (79) and (80), we can obtain the continuity equation given by

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + P) = 0. \tag{81}$$

The meaning of (81) 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 its content. The standard cosmological model, also known as Λ Cold Dark Matter (ΛCDM), is one of the most accepted models to describe the Universe, with its content being:

- **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 [49,50]. The ΛCDM model assumes the dark matter only interacts gravitationally (and possibly by weak interaction) with the rest of the Universe, hence its name, Cold Dark Matter (CDM). Since it is proposed as interacting only via gravitational force, there can be several candidates to fulfill this requirement: it could be conformed by weakly interacting massive particles (WIMPs), by gravitationally-interacting massive particles (GIMPs); by axions (hypothetical elementary particles); or by sterile neutrinos, just to mention a few. For a short review about Dark Matter and possible candidates, see Reference [51].
- **Radiation:** This corresponds to relativistic particles that follow the relation $P = \frac{1}{3}\rho$, which implies a density with a behavior $\rho \propto a^{-4}$. We consider photons ρ_γ and massless neutrinos ρ_ν as radiation, so the total radiation energy density in the Universe is given by

$$\rho_r = \rho_\gamma + \rho_\nu. \tag{82}$$

The relation between these quantities is

$$\rho_\nu = N_{\text{eff}} \times \frac{7}{8} \times \left(\frac{4}{11}\right)^{4/3} \rho_\gamma, \tag{83}$$

where N_{eff} is the effective number of relativistic degrees of freedom, with standard value $N_{\text{eff}} = 3.046$ [13].

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

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

$$\Omega_i \equiv \frac{\rho_i}{\rho_{\text{crit}}}, \quad \text{with} \quad \rho_{\text{crit}} = \frac{3H^2}{8\pi G}, \tag{84}$$

with ρ_{crit} being the condition to have a flat Universe or equivalently zero curvature, we can rewrite (79) as

$$\frac{H^2}{H_0^2} = \Omega_{r,0}a^{-4} + \Omega_{m,0}a^{-3} + \Omega_{k,0}a^{-2} + \Omega_{\Lambda,0}, \tag{85}$$

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 with the Cosmological Constant, and the subscript zero indicates they are evaluated by today's values.

Table 4. Equation of state associated with each component of the Universe.

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

6.1.4. Alternatives to the ΛCDM Model

The ΛCDM 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, Reference [50,52]). In addition, all attempts to detect WIMPs either directly in the laboratory, or indirectly by astronomical signals of distant objects have failed so far. For some of these reasons, it seems necessary to explore alternatives to the standard ΛCDM model. With this in mind, several alternatives have been suggested. For instance, the Scalar Field Dark Matter (SFDM) model proposes the dark matter as a spin 0 boson particle [53–58]; or the Self Interacting Dark Matter, as its name states, relies on the cold dark matter to be made of self interacting particles [59–61]. 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 [62], braneworld models [63,64]. There are also some alternatives to the cosmological constant as Dark Energy, i.e., scalar fields (quintessence, K-essence, phantom, quintom, non-minimally coupled scalar fields [65–69]); or many more alternatives, i.e., anisotropic Universes [70–72]. Finally, if the dark energy is assumed to be a perfect fluid, one of the most popular time-evolving parameterizations 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 [73,74]. It may also be expanded into Fourier series [75], or many more Bayesian approaches, as have been suggested to account for a dynamical dark energy [76–78].

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 fitted 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 (85). In particular, the radiation contribution is measured with great precision, so that Ω_γ is pinned down very accurately, and, hence, there is no need to fit this parameter. Similarly, neutrinos, as long as they maintain a relativistic behavior, can be related to the density of the photons through (83).

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 others 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 of matter-radiation, or the epoch of reionization. In the standard cosmological model (ΛCDM), 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 [13].

6.3. Cosmological Observations

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

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 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; 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 traveled through the primordial plasma before recombination. The sound horizon r_d is given by

$$r_d = \int_{z_d}^{\infty} \frac{c_s(z)}{H(z)} dz,\tag{86}$$

where the sound speed (in terms of redshift z) in the photon-baryon fluid is $c_s(z) = 3^{-1/2}c\left[1 + \frac{3}{4}\rho_b(z)/\rho_\gamma(z)\right]^{-1/2}$, 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 $\alpha_\perp$ and one parallel $\alpha_\parallel$ to the line of sight. A measurement of $\alpha_\perp$ constrains the ratio of the comoving angular diameter distance to the sound horizon [79]:

$$\frac{D_M(z)}{r_d} = \alpha_\perp \frac{D_{M,fid}(z)}{r_{d,fid}},\tag{87}$$

where the comoving angular diameter distance is given by

$$D_M(z) = \frac{c}{H_0} S_k\left(\frac{D_c(z)}{c/H_0}\right). \tag{88}$$

The line-of-sight comoving distance is defined as

$$D_c(z) = \frac{c}{H_0} \int_0^z dz' \frac{H_0}{H(z')}, \tag{89}$$

and $S_k(z)$

$$S_k(x) = \begin{cases} \sinh\left(\sqrt{\Omega_k}x\right)/\sqrt{\Omega_k} & \Omega_k > 0, \\ x & \Omega_k = 0, \\ \sin\left(\sqrt{-\Omega_k}x\right)/\sqrt{-\Omega_k} & \Omega_k < 0. \end{cases} \tag{90}$$

The Hubble parameter can be constrained by measuring $\alpha_\parallel$ using an analogous quantity

$$\frac{D_H(z)}{r_d} = \alpha_\parallel \frac{D_{H,fid}(z)}{r_{d,fid}}, \tag{91}$$

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

If redshift-space distortions are weak (which is valid for luminous galaxy surveys but not for the Ly-α), an isotropic analysis measures an effective combination of (87) and (91), and the volume averaged distance $D_V(z)$ [79]

$$\frac{D_V(z)}{r_d} = \alpha \frac{D_{V,fid}(z)}{r_{d,fid}}, \tag{92}$$

with $D_V(z) = \left[z D_H(z) D_M^2(z)\right]^{1/3}$.

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 13. The data used in this work to constrain D_V/r_d is obtained from the 6dF Galaxy Survey (6dFGS [80]) from UK Schmidt Telescope, the Main Galaxy Sample (MGS [81]), and the BOSS LOWZ Sample [82] from SDSS. On the other hand, the data from BOSS CMASS Sample [82], BOSS Lyman α auto-correlation (Lyα auto, [83]), and BOSS Lyman-α cross correlation (Lyα cross [84]) are used to constrain D_M/r_d, D_H/r_d.
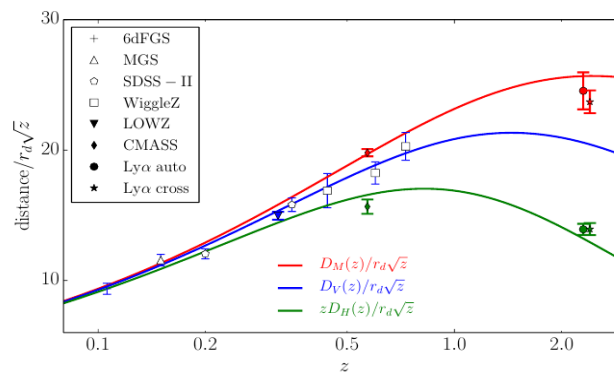

Figure 13. BAO Hubble diagram. BAO measurements of D_V/r_d, D_M/r_d, and zD_H/r_d from the sources indicated in the legend. The scaling factor $\sqrt{z}$ is included for a better display of the error bars. Solid lines are plotted by using the best-fit values obtained by the Planck satellite [85]. The Lyα cross-correlation points have been shifted in redshift; auto-correlation points are plotted at the correct effective redshift.

Supernovae type Ia (SNIa): This type of supernova occurs in binary star systems, one of which is a white dwarf that accretes matter from the star that accompanies it. When the white dwarf accumulates sufficient mass ($\approx$1.4 solar masses), its core will start the ignition temperature for carbon fusion, and, within a few seconds, it releases enough energy to produce the supernova [86]. Since type Ia supernovae (SNIa) are hypothesized to occur near the same mass limit of 1.4 solar masses, commonly referred to as Chandrasekhar mass, their luminosity peaks are fairly consistent and can be standardized and, thus, be used as standard candles [13]. From several analyses of SNIa, the Supernova Cosmology Project and High-z Supernova Search Team both found evidence that the Universe is currently expanding at an accelerated rate [87–89].

These stars allow us to measure relative distances using the luminosity distance given by

$$D_L \equiv \sqrt{\frac{L}{4\pi S}}, \tag{93}$$

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 [49]. 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. Even if the luminosity is unknown, it will appear as a scaling factor [49]. The relation between D_L and the cosmological parameters is given by

$$D_L = D_M(1+z), \tag{94}$$

where D_M is provided by Equation (88). Another important quantity in the observation of supernovae is the standardized distance modulus

$$\mu = m_B^* - M_B + \alpha X_1 - \beta C, \tag{95}$$

where m_B^* is the observed peak magnitude in the rest frame of blue band (B), α, β, and M_B are parameters that depend on host galaxy properties [90]. X_1 is the time stretching of the light curve, and C is the supernova color at maximum brightness. The relation between the standardized distance modulus and the luminosity distance is

$$\mu = 5log_{10}\left(\frac{D_L}{10pc}\right). \tag{96}$$

The data used in this paper (shown in Figure 14) 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 [91]) 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 the Hubble Space Telescope (HST) [90]. 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 [79].

Although SNIa have been widely used to restrict cosmological models, there are still discussions about the way it is done. That is, in order to extract information from them, an a priori cosmological model has to be assumed, which may biased the outcomes [92].
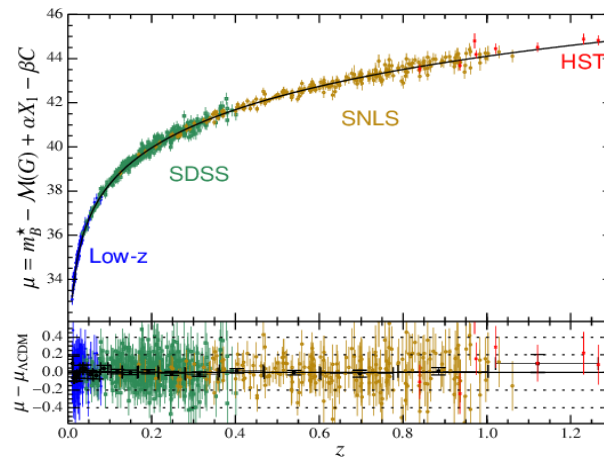
Figure 14. Joint Light-curve Analysis data (JLA). Vertical axis is the standardized distance modulus μ (luminosity distance function), and the horizontal axis is the redshift z. Source: Reference [93].

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 traveled 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 recent collaborations that studies the CMB corresponds to the Planck satellite, and previous probes include COBE [94] and WMAP [95]. It is a European Space Agency mission, in which the 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 and the nature of dark matter and dark energy, as well as to test inflationary theories, determine whether the Universe is homogeneous or not, and obtain maps of galaxies in the microwave [96–103].

In this work, we use the CMB information as a BAO located at redshift $z = 1090$, measuring the angular scale of the sound horizon $D_M(1090)/r_d$, in addition, to calibrating the absolute length of the BAO ruler through the determination of $\Omega_b h^2$ and $\Omega_{cb} h^2$ (the density parameters of baryonic and dark matter, respectively; more details about these parameters can be found in Reference [79]).

Large Red Galaxies: (Cosmic Chronometers) These are the most massive galaxies for each redshift z and contain the oldest star population. These kinds of galaxies are used to estimate the Hubble factor because they contain little stardust, which makes it easier to get their spectra. The way these chronometers work is by selecting two galaxies at different redshifts 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 [104,105]

$$H(z) = \frac{-1}{1+z}\frac{dz}{dt} \simeq \frac{-1}{1+z}\frac{\Delta z}{\Delta t}. \tag{97}$$

In this work, the data used to constrain $H(z)$ was obtained from References [106–109]. In Figure 15, a compilation of these data is shown.

Another important probe in the foreseen future corresponds to the Gravitational waves from merging black holes (standard sirens), as they directly measure the luminosity distance to the merger. Therefore, it is possible to constrain the distance-redshift relation and, hence, the cosmological parameters [110,111].

Even though this paper is focused on cosmological constraints on the background level, there exist several observations to pin down the parameters at the perturbation level, and some of them include gravitational lensing [112–116], clustering [117–120], the matter

power spectrum [121,122], and primordial gravitational waves and the primordial power spectrum, both generated during inflation [123–125], among many others.

7. Parameters Inference in Cosmology

Here, we estimate the Hubble parameter H_0 at the present time and the matter density of the Universe Ω_m, assuming a ΛCDM standard model. We use our own Python code, which can be found in Reference [44].

The data—For this particular example, we focus only on the Cosmic Chronometers, as shown in Figure 15.

The model—Our interest 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 (85), and, for simplicity, we assume a flat Universe with a well measured radiation content. Then, the model is given by

$$H^2 = H_0^2[\Omega_{m,0}(1+z)^3 + \Omega_{\Lambda,0}], \tag{98}$$

constrained by the relation

$$\Omega_{m,0} + \Omega_{\Lambda,0} = 1. \tag{99}$$

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

$$L \propto \exp\left[-\sum_i \frac{(H_i - H(z_i))^2}{2\sigma_i^2}\right], \tag{100}$$

where $H(z_i)$ is described by Equation (98) evaluated at each redshift z_i, and H_i is the value of the Hubble parameter measured at z_i and σ_i the error of the i-th measurement.

The only a priori information we have about the free parameters is that each component of the Universe $\Omega_{i,0}$ must satisfy the relation $0 \leq \Omega_{i,0} \leq 1$, 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 a uniform distribution (flat prior) with limits $\Omega_m \in [0,1]$ and $H_0 \in [10,100]$. Hence, the priors

$$\Omega_{m,0} \sim U[0,1], \qquad H_0 \sim U[10,100]. \tag{101}$$

Numerical estimation—We follow the same procedure as in the straight-line example. In the left panel of Figure 16, we have plotted the chains obtained for our estimations and its corresponding 1D posterior distribution. Similar to the above examples, we have also plotted the 2D posterior distributions with 1–4σ 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 to the values reported in the literature [85].
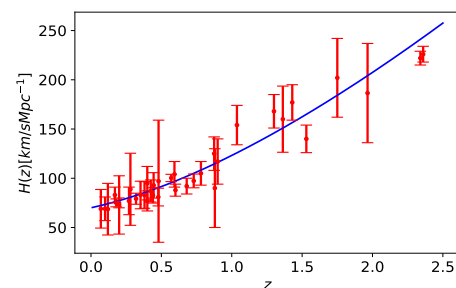

Figure 15. Hubble data as a function of redshift z for Cosmic Chronometers. The solid line corresponds to the best-fit by using ΛCDM model.
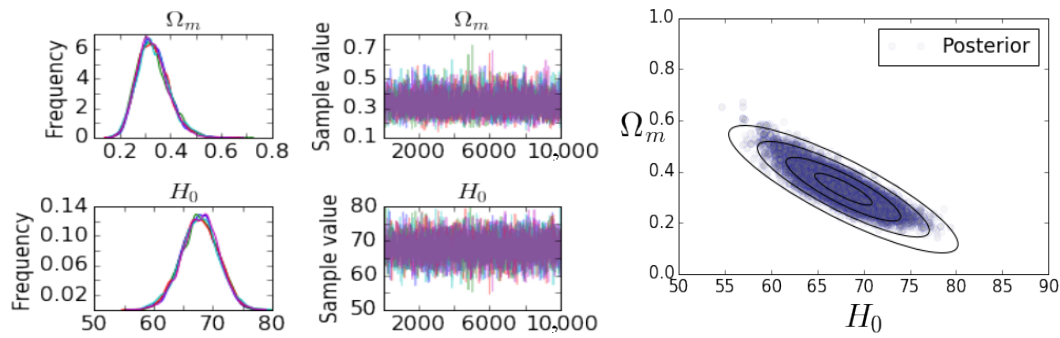
Figure 16. Results for the ΛCDM model. Left panel: One-dimensional posterior distributions for the parameters H_0 and Ω_m, along with its chains. Right panel: Joint 2D posterior distributions with 1–4 confidence levels.

7.1. Cosmological and Statistical Codes

As we have shown, until now, during the process of parameter inference for 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 (98) and (100). Finally, it is necessary to program the numerical tools in order to obtain the parameter inference. Such programming can be done, for example, in PyMC3, as we saw before.

It is notorious that the above process can be a tremendous task for programming, for instance, when the theories involve numerous parameters, or some of them must be marginalized in 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: the first is by accepting the challenge and creating our own code. Developing a new code may be a good option rather than using others, as it would require a long time to learn the implementation and modification (specially if the theory we are dealing with 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 to test cosmological models.

7.1.1. Cosmological Codes

Today, there are several cosmological Boltzmann codes available. Some of them include: CMBFAST (written in FORTRAN 77 [126–128]), CMBEASY (C++ [129]), CAMB (FORTRAN 90 [130,131]), CLASS (C [132,133]), and COSMOSIS [134] (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. There are several MCMC codes that can make this task easy to handle. Some of them are:

Monte Python [135–137]—It is a Monte Carlo code for Cosmological Parameter extraction that 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.

CosmoMC [138]—This is a Fortran MCMC engine for exploring cosmological parameter space. 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 an MCMC code for cosmological parameter estimation where only the expansion history of the Universe matters. It was written by Anže Slosar and José A. Vázquez, initially released in Reference [79], and can be downloaded in Reference [139]. This code solves the cosmological equations for the background parameters in the same way as CLASS or CAMB, and it contains the statistical parameter inference of CosmoMC/MontePython. An advantage of this code is that it is completely written in Python with an interface to machine learning tools, such as artificial neural networks, genetic algorithms, as well as algorithms to compute the Bayesian evidence, i.e., Dynesty [140] or MCEvidence [141].

The main idea of MCEvidence is that, asymptotically, the number density n of the MCMC is proportional to the density of the likelihood multiplied by the prior, that is, the non-normalized posterior

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

where $n(D|\theta, H) = NP(D|\theta, H)$ with N the length of the chain. The code uses Bayesian inference to find a, the constant of proportionality (see Reference [142] for details). Once this constant has been found, the Bayesian evidence is given by $P(D|H) = aN$.

8. Examples with SimpleMC

The main interest of this section is to test several cosmological models through SimpleMC. Even though we selected this code, the results we present here can also be obtained in any of the aforementioned codes.

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

$$L \propto \exp\left[-\sum_i \frac{(T(z_i) - T_i)^2}{2\sigma_i}\right], \tag{102}$$

where $T(z_i)$ is the theoretical value related to the observation T_i; and σ_i are the corresponding errors associated to each measurement. In our estimation, $T(z_i)$ is given by (94) for Supernovae; (87) for CMB; (87), (91), and (92) for BAO; and (97) for Cosmic Chronometers.

We use the BAO data mentioned in Section 6.3 (labeled as BBAO); for Supernovae, 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).

Models of the Universe

The base example corresponds to the flat ΛCDM model, already explored in Section 7, but now including the full observations presented in Section 6.3 and using the SimpleMC code. Here, in order to test the Friedmann Equation (85) 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 ΛCDM model, we let the curvature of the Universe be a free parameter (model oΛCDM), 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 labeled it as model ωCDM. We may combine the addition of curvature and constant ω to define the oωCDM 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 labeled it as $\omega_0\omega_a$CDM model. Again, we can incorporate the curvature of the Universe to the CPL parameterization, named as o$\omega_0\omega_a$CDM.

By using the combined dataset BBAO+Planck+SN+HD, Table 5 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 ΛCDM 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 16 is updated with new data in order to get Figure 17. Here, the upper panel displays the chain for the parameter $H_0 = 100 \cdot h$, with 9000 steps. In the lower panel of the same figure, we plot the 2D posterior distribution, along with 1 and 2σ confidence regions obtained from our estimations.

From Table 5, we also observe the best fit of most of the new parameters—additional to the base ΛCDM model—remained well inside the 1σ confidence level. However, the exception is for the $o\omega_0\omega_a$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 18, where the 2D posterior distributions, along with 1 and 2σ confidence regions, are shown. Here, the standard ΛCDM 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}, before the last row of Table 5. However, it also carries out a penalization factor that affects directly the model selection, as seen in Reference [75,77,143].

Table 5. 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 ($\Omega_k = 0$, $\omega = \omega_0 = -1$, $\omega_a = 0$).

Parameter	ΛCDM	oΛCDM	ωCDM	oωCDM	$\omega_0\omega_a$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	$\cdots$	-0.004 ± 0.002	$\cdots$	-0.003 ± 0.003	$\cdots$	-0.006 ± 0.003
ω_0	$\cdots$	$\cdots$	-0.96 ± 0.05	-0.97 ± 0.05	-0.91 ± 0.10	-0.83 ± 0.11
ω_a	$\cdots$	$\cdots$	$\cdots$	$\cdots$	-0.17 ± 0.41	-0.52 ± 0.51
χ^2_{min}	73.57	71.59	73.13	71.5	72.8	69.7
$|\mathcal{B}_{\Lambda\text{CDM},i}|$	0	3.61	1.48	5.59	1.23	4.3

In the last row of the same table, we show the Bayes factor for the extended models using the LCDM model as reference, where the Bayesian evidences were obtained using the MCEvidence code. We found significant evidence in favor of ΛCDM compared to ωCDM and $\omega_0\omega_a$CDM; strong evidence with respect to oΛCDM and o$\omega_0\omega_a$CDM; and decisive evidence against oωCDM. This is in agreement with the results presented in Reference [144], where Planck 2013 found no evidence against ΛCDM.
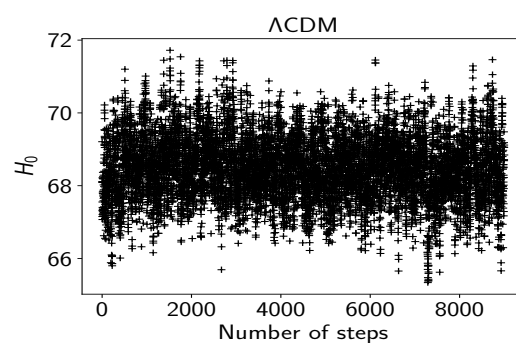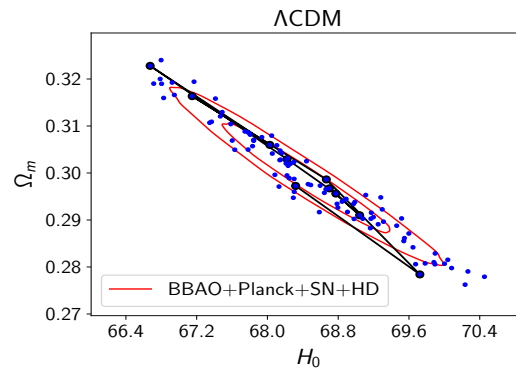

Figure 17. *Cont.*

Figure 17. Top panel: Markov chain for the parameter H_0, with 9000 steps. Bottom panel: Two-dimensional posterior distribution with confidence regions 1 and 2 σ for the joint parameters H_0 and Ω_m.
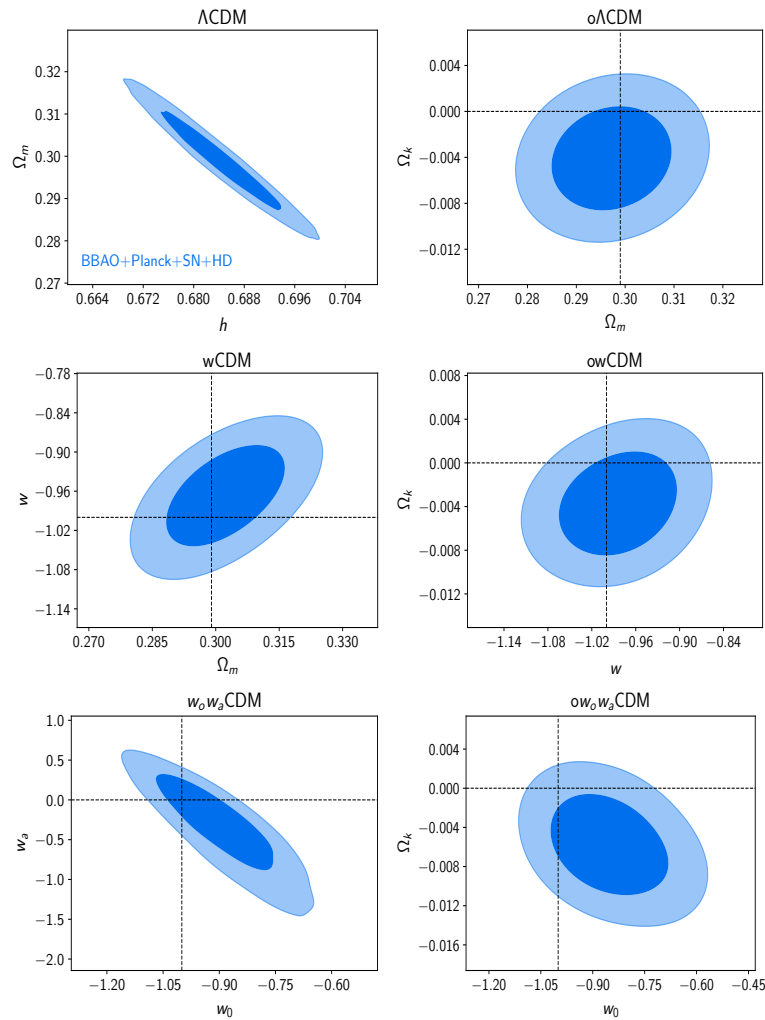

Figure 18. Two-dimensional posterior distributions with 1,2-σ confidence regions for different models. oΛCDM refers to ΛCDM with curvature, ωCDM is a flat Universe with the dark energy equation of state as a free parameter, oωCDM generalizes to non-zero curvature, $w_o w_a$CDM uses the CPL parameterization, and o$w_o w_a$CDM generalizes to non-zero curvature. The dashed lines show the standard ΛCDM values.

9. Conclusions and Discussion

The fact that the number of cosmological observations have increased impressively over the last decade allowed to obtain a better description of the Universe. However, since we still have the limitation of a unique Universe, a Frequentist approach may not be the best to rely on; hence, the Bayesian statistics came into consideration. 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. Bayes' theorem states that

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

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 the prior probability is constant, we can identify the posterior probability with the likelihood $P(\theta|D, H) \propto L(D|\theta, H)$; 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. In addition, for this particular case, in which likelihoods are Gaussian, Fisher's matrix can be computed 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 predicting 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 is may be by introducing the hyperparameter method. These hyperparameters act as weights for every dataset in order to take away data that does not seem to be consistent with all 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 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, through

$\chi^2_{\min}$, corresponds to a curved Universe with a dynamical dark energy, namely o$\omega_o\omega_a$CDM. It is important to clarify that this does not mean that the o$\omega_o\omega_a$CDM model is the final model; it merely shows that, for these particular datasets, there is an improvement in the fit when compared to ΛCDM, but further analysis and more data is necessary to give a verdict. However, adding extra parameters brings up a penalized factor seen through the Bayesian evidence; hence, the favored model becomes ΛCDM.

Author Contributions: All authors contributed equally. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Note

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

References

1.   Trimble, V. The 1920 shapley-curtis discussion: background, issues, and aftermath. *Publ. Astron. Soc. Pac.* **1995**, *107*, 1133. [CrossRef]
2.   Turner, M.S. David Norman Schramm, a Biographical Memoir http://www.nasonline.org/publications/biographical-memoirs/memoir-pdfs/schramm-david.pdf (accessed on 18 May 2021).
3.   Kleijn, B. *Bayesian Statistics*; University of Amsterdam Lecture Notes: Amsterdam, The Netherlands, 2013.
4.   Heavens, A. Statistical techniques in cosmology. *arXiv* **2009**, arxiv:0906.0664.
5.   Trotta, R. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemp. Phys.* **2008**, *49*, 71. [CrossRef]
6.   Verde, L. Statistical methods in cosmology. In *Lectures on Cosmology*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 800, pp. 147–177.
7.   Trotta, R. Bayesian Methods in Cosmology. *arXiv* **2017**, arXiv:1701.01467.
8.   Jaffe, A.H. H0 and odds on cosmology. *Astrophys. J.* **1996**, *24*, 471. [CrossRef]
9.   WMAP Collaboration. First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Parameter estimation methodology. *Astrophys. J. Suppl.* **2003**, *148*, 195. [CrossRef]
10.   D'Agostini, G. Probability and measurement uncertainty in physics: A Bayesian primer. *arXiv* **1995**, arXiv:hep-ph/9512295.
11.   Sharma, S. Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy. *Annu. Rev. Astron. Astrophys.* **2017**, *55*, 213–259.
12.   Liddle, A.R. How many cosmological parameters? *Mon. Not. Roy. Astron. Soc.* **2004**, *351*, L49. [CrossRef]
13.   Lahav, O.; Liddle, A.R. The Cosmological Parameters 2014. *arxiv* **2010**, arxiv:1401.1389.
14.   Mohammad-Djafari, A. Bayesian inference for inverse problems. Bayesian Inference and Maximum Entropy Methods in Science and Engineering. *AIP Conf. Proc. Am. Inst. Phys.* **2002**, *617*, 477–496.
15.   Vazquez, J.A.; Lasenby, A.N.; Bridges, M.; Hobson, M.P. A Bayesian study of the primordial power spectrum from a novel closed universe model. *Mon. Not. Roy. Astron. Soc.* **2012**, *422*, 1948. [CrossRef]
16.   Press, W.H.; Teukolsky, S.A.; Vetterling, W.T.; Flannery, B.P. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*; Cambridge University Press: Cambridge, UK, 2007.
17.   Fisher, R.A. The logic of inductive inference. *J. R. Stat. Soc.* **1935**, *98*, 39. [CrossRef]
18.   Albrecht, A.; Bernstein, G.; Cahn, R.; Freedman, W.L.; Hewitt, J.; Hu, W.; Huth, J.; Kamionkowski, M.; Kolb, E.W.; Knox, L.; et al. Report of the Dark Energy Task Force. *arXiv* **2006**, arXiv:astro-ph/0609591.
19.   Tokdar, S.T.; Kass, R.E. Importance sampling: A Review. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 54. [CrossRef]
20.   Lahav, O.; Bridle, S.L.; Hobson, M.P.; Lasenby, A.N.; Sodre, L., Jr. Bayesian 'hyper-parameters' approach to joint estimation: the hubble constant from cmb measurements. *Mon. Not. Roy. Astron. Soc.* **2000**, *315*, L45. [CrossRef]
21.   Hobson, M.P.; Bridle, S.L.; Lahav, O. Combining cosmological datasets: Hyperparameters and bayesian evidence. *Mon. Not. Roy. Astron. Soc.* **2002**, *335*, 377. [CrossRef]
22.   Medel, R.; Gomez, I.; Vazquez, J.A.; Garcia, R. An introduction to markov chain monte carlo. *Bol. Estad. Investig. Oper.* **2021**, *37*, 47.
23.   Tanner, M.A. *Tools for Statistical Inference*; Springer: Berlin/Heidelberg, Germany, 2012.

24. Gilks, W.R.; Richardson, S.; Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*; Chapman and Hall/CRC: New York, NY, USA, 1995.
25. Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.B. *Bayesian Data Analysis*; Chapman and Hall/CRC: New York, NY, USA, 2013.
26. Ross, S.M. *Introduction to Probability Models*; Academic Press: Cambridge, MA, USA, 2014.
27. Metropolis, N.; Rosenbluth, A.W.; Rosenbluth, M.N.; Teller, A.H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087. [CrossRef]
28. Hastings, W.K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97. [CrossRef]
29. Salvatier, J.; Wiecki, T.V.; Fonnesbeck, C. Probabilistic programming in python using pymc3. *PeerJ Comput. Sci.* **2016**, *2*, e55. [CrossRef]
30. Gelman, A.; Rubin, D.B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **1992**, *7*, 457. [CrossRef]
31. Brooks, S.P.; Gelman, A. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **1998**, *7*, 434.
32. Link, W.; Eaton, M. On thinning of chains in mcmc. methods in ecology and evolution. *Methods Ecol. Evol.* **2012**, *3*, 112–115. [CrossRef]
33. Geweke, J. *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*; Federal Reserve Bank of Minneapolis, Research Department Minneapolis: Minneapolis, MN, USA, 1991; Volume 196.
34. Model Checking and Diagnostics. 2021. Available online: https://pymc-devs.github.io/pymc/modelchecking.html (accessed on 18 May 2021).
35. Geman, S.; Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, 721–741.. [CrossRef] [PubMed]
36. Vousden, W.; Farr, W.M.; Mandel, I. Dynamic temperature selection for parallel tempering in markov chain monte carlo simulations. *Mon. Not. R. Astron. Soc.* **2015**, *455*, 1919. [CrossRef]
37. Cappé, O.; Guillin, A.; Marin, J.M.; Robert, C.P. Population Monte Carlo. *J. Comput. Graph. Stat.* **2004**, *13*, 907. [CrossRef]
38. Kilbinger, M.; Benabed, K.; Cappe, O.; Cardoso, J.-F.; Coupon, J.; Fort, G.; McCracken, H.J.; Prunet, S.; Robert, C.P.; Wraith, D CosmoPMC: Cosmology Population Monte Carlo. *arXiv* **2012**, arXiv:1101.0950.
39. Foreman-Mackey, D.; Hogg, D.W.; Lang, D.; Goodman, J. emcee: The MCMC Hammer. *Publ. Astron. Soc. Pac.* **2013**, *125*, 306. [CrossRef]
40. emcee. 2021. Available online: http://dfm.io/emcee/current/ (accessed on 18 May 2021).
41. Goodman, J.; Weare, J. Ensemble samplers with affine invariance. *Commun. Appl. Math. Comput. Sci.* **2010**, *5*, 65. [CrossRef]
42. Hanson, K.M. Markov chain monte carlo posterior sampling with the hamiltonian method. In Proceedings of the Medical Imaging 2001: Image Processing. International Society for Optics and Photonics, Los Alamos, NM, USA, 21 February 2001; Volume 4322, pp. 456–467.
43. Neal, R.M. Mcmc using hamiltonian dynamics. *Handb. Markov Chain. Monte Carlo* **2011**, *2*, 2.
44. Available online: https://github.com/ja-vazquez/Cosmologia_observacional.git (accessed on 18 May 2021).
45. Baumann, D. *Cosmology, Part III Mathematical Tripos*; University Lecture Notes: Cambridge, UK, 2014.
46. Wald, R. General relativity. the university of chicago. *Chicago Sect.* **1984**, *6*, 72–73.
47. Iorio, L. Editorial for the Special Issue 100 Years of Chronogeometrodynamics: The Status of the Einstein's Theory of Gravitation in Its Centennial Year. *Universe* **2015**, *1*, 38. [CrossRef]
48. Debono, I.; Smoot, G.F. General Relativity and Cosmology: Unsolved Questions and Future Directions. *Universe* **2016**, *2*, 23. [CrossRef]
49. Liddle, A. *An Introduction to Modern Cosmology*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
50. Vázquez-González, A.; Matos, T. La materia oscura del universo: retos y perspectivas. *Rev. Mex. FíSica* **2008**, *54*, 193.
51. Arun, K.; Gudennavar, S.; Sivaram, C. Dark matter, dark energy, and alternate models: A review. *Adv. Space Res.* **2017**, *60*, 166. [CrossRef]
52. Bull, P.; Akrami, Y.; Adamek, J.; Baker, T.; Bellini, E.; Jiménez J.B.; Bentivegna, E.; Camera, S.; Clesse, S.; Davis, J.H.; et al. Beyond ΛCDM: Problems, solutions, and the road ahead. *Phys. Dark Univ.* **2016**, *12*, 56. [CrossRef]
53. Matos, T.; Luevano, J.-R.; Quiros, I.; Urena-Lopez, L.A.; Vazquez, J.A. Dynamics of Scalar Field Dark Matter With a Cosh-like Potential. *Phys. Rev. D* **2009**, *80*, 123521. [CrossRef]
54. Sin, S.-J. Late-time phase transition and the galactic halo as a bose liquid. *Phys. Rev.* **1994**, *50*, 3650. [CrossRef]
55. FGuzman, S.; Matos, T.; Villegas, H. Scalar fields as dark matter in spiral galaxies: comparison with experiments. *Astron. Nachrichten News Astron. Astrophys.* **1999**, *320*, 97. [CrossRef]
56. Matos, T.; Guzman, F.S. Scalar fields as dark matter in spiral galaxies. *Class. Quant. Grav.* **2000**, *17*, L9. [CrossRef]
57. Lee, J.-W.; Koh, I.-G. Galactic halos as boson stars. *Phys. Rev. D* **1996**, *53*, 2236. [CrossRef] [PubMed]
58. Matos, T.; Vazquez-Gonzalez, A.; Magana, J. ϕ^2 as Dark Matter. *Mon. Not. Roy. Astron. Soc.* **2009**, *393*, 1359. [CrossRef]
59. Spergel, D.N.; Steinhardt, P.J. Observational evidence for selfinteracting cold dark matter. *Phys. Rev. Lett.* **2000**, *84* 3760. [CrossRef] [PubMed]
60. Gonzalez, T.; Matos, T.; Quiros, I.; Vazquez-Gonzalez, A. Self-interacting Scalar Field Trapped in a Randall-Sundrum Braneworld: The Dynamical Systems Perspective. *Phys. Lett. B* **2009**, *676*, 161. [CrossRef]

61. Padilla, L.E.; Vázquez, J.A.; Matos, T.; Germán, G. Scalar Field Dark Matter Spectator During Inflation: The Effect of Self-interaction. *JCAP* **2019**, *5*, 56. [CrossRef]
62. Felice, A.D.; Tsujikawa, S. f(R) theories. *Living Rev. Rel.* **2010**, *13*, 3. [CrossRef]
63. Nojiri, S.; Odintsov, S.; Oikonomou, V. Modified gravity theories on a nutshell: inflation, bounce and late-time evolution. *Phys. Rep.* **2017**, *692*, 1. [CrossRef]
64. Garcia-Aspeitia, M.A.; Hernandez-Almada, A.; Na, J.M.; Amante, M.H.; Motta, V.; Martínez-Robles, C. Brane with variable tension as a possible solution to the problem of the late cosmic acceleration. *Phys. Rev. D* **2018**, *97*. [CrossRef]
65. Vázquez, J.A.; Tamayo, D.; Sen, A.A.; Quiros, I. Bayesian model selection on scalar ϵ-field dark energy. *Phys. Rev. D* **2021**, *103*. [CrossRef]
66. Akarsu, O.; Katirci, N.; Sen, A.A.; Vazquez, J.A. Scalar field emulator via anisotropically deformed vacuum energy: Application to dark energy. *arXiv* **2020**, arXiv:2004.14863.
67. Tsujikawa, S. Quintessence: A Review. *Class. Quant. Grav.* **2013**, *30*. [CrossRef]
68. Yoo, J.; Watanabe, Y. Theoretical Models of Dark Energy. *Int. J. Mod. Phys. D* **2012**, *21*. [CrossRef]
69. Feng, B. The quintom model of dark energy. *arXiv* **2006**, arXiv:astro-ph/0602156.
70. Akarsu, O.; Katırcı, N.; Özdemir, N.; Vázquez, J.A. Anisotropic massive Brans-Dicke gravity extension of the standard ΛCDM model. *Eur. Phys. J. C* **2020**, *80*. [CrossRef]
71. Sharif, M.; Azeem, S. Dark energy models and cosmic acceleration with anisotropic universe in f (t) gravity. *Commun. Theor. Phys.* **2014**, *61*, 482. [CrossRef]
72. Saadeh, D.; Feeney, S.M.; Pontzen, A.; Peiris, H.V.; McEwen, J.D. How isotropic is the universe? *Phys. Rev. Lett.* **2016**, *117*, 131302. [CrossRef]
73. Linden, S.; Virey, J.-M. A test of the CPL parameterization for rapid dark energy equation of state transitions. *Phys. Rev. D* **2008**, *78*, 023526. [CrossRef]
74. Scherrer, R.J. Mapping the Chevallier-Polarski-Linder parametrization onto Physical Dark Energy Models. *Phys. Rev. D* **2015**, *92*, 043001. [CrossRef]
75. Tamayo, D.; Vazquez, J.A. Fourier-series expansion of the dark-energy equation of state. *Mon. Not. Roy. Astron. Soc.* **2019**, *487*. [CrossRef]
76. Vazquez, J.A.; Bridges, M.; Hobson, M.P.; Lasenby, A.N. Reconstruction of the Dark Energy equation of state. *JCAP* **2012**, *9*. [CrossRef]
77. Hee, S.; Vázquez, J.A.; Handley, W.J.; Hobson, M.P.; Lasenby, A.N. Constraining the dark energy equation of state using Bayes theorem and the Kullback–Leibler divergence. *Mon. Not. Roy. Astron. Soc.* **2017**, *466*, 369. [CrossRef]
78. Vazquez, J.A.; Hee, S.; Hobson, M.P.; Lasenby, A.N.; Ibison, M.; Bridges, M. Observational constraints on conformal time symmetry, missing matter and double dark energy. *JCAP* **2018**, *7*, 62. [CrossRef]
79. Aubourg, E.; Bailey, S.; Bautista, J.E.; Beutler, F.; Bhardwaj, V.; Bizyaev, D.; Blanton, M.; Blomqvist, M.; Bolton, A.S.; Bovy, J.; et al. Cosmological implications of baryon acoustic oscillation measurements. *Phys. Rev. D* **2015**, *92*, 123516. [CrossRef]
80. Beutler, F.; Blake, C.; Colless, M.; Jones, D.H.; Staveley-Smith, L.; Campbell, L.; Parker, Q.; Saunders, W.; Watson, F. The 6dF Galaxy Survey: Baryon Acoustic Oscillations and the Local Hubble Constant. *Mon. Not. Roy. Astron. Soc.* **2011**, *416*, 3017. [CrossRef]
81. SDSS Collaboration. Spectroscopic target selection for the Sloan Digital Sky Survey: The Luminous red galaxy sample. *Astron. J.* **2001**, *122*, 2267. [CrossRef]
82. Anderson, L.; Aubourg, E.; Bailey, S.; Beutler, F.; Bhardwaj, V.; Blanton, M.; Bolton, A.S.; Brinkmann, J.; Brownstein, J.R.; Burden, A.; et al. The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: Baryon Acoustic Oscillations in the Data Release 9 Spectroscopic Galaxy Sample. *Mon. Not. Roy. Astron. Soc.* **2013**, *427*, 3435. [CrossRef]
83. Busca, N.G.; Delubac, T.; Rich, J.; Bailey, S.; Font-Ribera, A.; Kirkby, D.; Le Goff, J.-M.; Pieri, M.M.; Slosar, A.; Aubourg, E.; et al. Baryon Acoustic Oscillations in the Ly-α forest of BOSS quasars. *Astron. Astrophys.* **2013**, *552*, A96. [CrossRef]
84. BOSS Collaboration. Quasar-Lyman α Forest Cross-Correlation from BOSS DR11: Baryon Acoustic Oscillations. *JCAP* **2014**, *5*, 27.
85. PLANCK Collaboration. Planck 2013 results. XVI. Cosmological parameters. *Astron. Astrophys.* **2014**, *571*, A16. [CrossRef]
86. What Is a Supernova. 2021. Available online: https://www.nasa.gov/audience/forstudents/5-8/features/nasa-knows/what-is-a-supernova.html (accessed on 18 May 2021).
87. Riess, A.G.; Filippenko, A.V.; Challis, P.; Clocchiatti, A.; Diercks, A.; Garnavich, P.M.; Gilliland, R.L.; Hogan, C.J.; Jha, S.; Kirshner, R.P.; et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron. J.* **1998**, *116*, 1009. [CrossRef]
88. Supernova Search Team Collaboration. Supernova limits on the cosmic equation of state. *Astrophys. J.* **1998**, *509*, 74. [CrossRef]
89. Supernova Cosmology Project Collaboration. Measurements of Ω and Λ from 42 high redshift supernovae. *Astrophys. J.* **1999**, *517*, 565. [CrossRef]
90. SDSS Collaboration. Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples. *Astron. Astrophys.* **2014**, *568*, A22. [CrossRef]
91. SNLS Collaboration. SNLS—The Supernova Legacy Survey. *ASP Conf. Ser.* **2005**, *339*, 60.
92. Vishwakarma, R.G.; Narlikar, J.V. Is it no Longer Necessary to Test Cosmologies with Type Ia Supernovae? *Universe* **2018**, *4*, 73. [CrossRef]

93. SDSS-II/SNLS3 Joint Light-Curve Analysis. 2021. Available online: http://supernovae.in2p3.fr/sdss_snls_jla/ReadMe.html#sec-4-1 (accessed on 18 May 2021).

94. Mather, J.C.; Cheng, E.S.; Eplee, R.E., Jr.; Isaacman, R.B.; Meyer, S.S.; Shafer, R.A.; Weiss, R.; Wright, E.L.; Bennett, C.L.; Boggess, N.W.; et al. A Preliminary measurement of the Cosmic Microwave Background spectrum by the Cosmic Background Explorer (COBE) satellite. *Astrophys. J. Lett.* **1990**, *354*, L37. [CrossRef]

95. Bennett, C.L.; Hill, R.S.; Hinshaw, G.; Larson, D.; Smith, K.M.; Dunkley, J.; Gold, B.; Halpern, M.; Jarosik, N; Kogut, A.; et al. Seven-Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Are There Cosmic Microwave Background Anomalies? *Astrophys. J. Suppl.* **2011**, *192*, 17. [CrossRef]

96. PLANCK Collaboration. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.* **2020**, *641*, A6. [CrossRef]

97. PLANCK Collaboration. Planck 2018 results. VII. Isotropy and Statistics of the CMB. *Astron. Astrophys.* **2020**, *641*, A7. [CrossRef]

98. PLANCK Collaboration. Planck 2018 results. X. Constraints on inflation. *Astron. Astrophys.* **2020**, *641*, A10. [CrossRef]

99. Staggs, S.; Dunkley, J.; Page, L. Recent discoveries from the cosmic microwave background: a review of recent progress. *Rep. Prog. Phys.* **2018**, *81*, 044901. [CrossRef] [PubMed]

100. Samtleben, D.; Staggs, S.; Winstein, B. The cosmic microwave background for pedestrians: A review for particle and nuclear physicists. *Annu. Rev. Nucl. Part. Sci.* **2007**, *57*, 245. [CrossRef]

101. Kamionkowski, M.; Kosowsky, A. The cosmic microwave background and particle physics. *Annu. Rev. Nucl. Part. Sci.* **1999**, *49*, 77. [CrossRef]

102. Erickcek, A.L.; Carroll, S.M.; Kamionkowski, M. Superhorizon perturbations and the cosmic microwave background. *Phys. Rev. D* **2008**, *78*, 083012. [CrossRef]

103. Kamionkowski, M.; Kosowsky, A.; Stebbins, A. Statistics of cosmic microwave background polarization. *Phys. Rev. D* **1997**, *55*, 7368. [CrossRef]

104. Stern, D.; Jimenez, R.; Verde, L.; Stanford, S.A.; Kamionkowski, M. Cosmic Chronometers: Constraining the Equation of State of Dark Energy. II. A Spectroscopic Catalog of Red Galaxies in Galaxy Clusters. *Astrophys. J. Suppl.* **2010**, *188*, 280. [CrossRef]

105. Stern, D.; Jimenez, R.; Verde, L.; Kamionkowski, M.; Stanford, S.A. Cosmic Chronometers: Constraining the Equation of State of Dark Energy. I: H(z) Measurements. *JCAP* **2010**, *2*, 008. [CrossRef]

106. Moresco, M.; Cimatti, A.; Jimenez, R.; Pozzetti, L.; Zamorani, G.; Bolzonella, M.; Dunlop, J.; Lamareille, F.; Mignoli, M.; Pearce, H.; et al. Improved constraints on the expansion rate of the Universe up to z~1.1 from the spectroscopic evolution of cosmic chronometers. *JCAP* **2012**, *8*, 6. [CrossRef]

107. Moresco, M.; Verde, L.; Pozzetti, L.; Jimenez, R.; Cimatti, A. New constraints on cosmological parameters and neutrino properties using the expansion rate of the Universe to z~1.75. *JCAP* **2012**, *7*, 053. [CrossRef]

108. Moresco, M. Raising the bar: new constraints on the hubble parameter with cosmic chronometers at z 2. *Mon. Not. R. Astron. Soc. Lett.* **2015**, *450*, L16. [CrossRef]

109. Moresco, M.; Pozzetti, L.; Cimatti, A.; Jimenez, R.; Maraston, C.; Verde, L.; Thomas, D.; Citro, A.; Tojeiro, R.; Wilkinson, D. A 6 evidence of the epoch of cosmic re-acceleration. *JCAP* **2016**, *5*, 14. [CrossRef]

110. LIGO SCIENTIFIC, VIRGO, 1M2H, DARK ENERGY CAMERA GW-E, DES, DLT40, LAS CUMBRES OBSERVATORY, VINROUGE, MASTER Collaboration. A gravitational-wave standard siren measurement of the Hubble constant. *Nature* **2017**, *551*, 85. [CrossRef]

111. Guidorzi, C.; Margutti, R.; Brout, D.; Scolnic, D.; Fong, W.; Alexander, K.D.; Cowperthwaite, P.S.; Annis, J.; Berger, E.; Blanchard, P.K.; et al. Improved Constraints on H_0 from a Combined Analysis of Gravitational-wave and Electromagnetic Emission from GW170817. *Astrophys. J. Lett.* **2017**, *851*, L36. [CrossRef]

112. Mandelbaum, R. Weak lensing for precision cosmology. *Annu. Rev. Astron. Astrophys.* **2018**, *56*, 393. [CrossRef]

113. Hoekstra, H.; Jain, B. Weak gravitational lensing and its cosmological applications. *Annu. Rev. Nucl. Part. Sci.* **2008**, *58*, 99. [CrossRef]

114. Blandford, R.; Narayan, R. Cosmological applications of gravitational lensing. *Annu. Rev. Astron. Astrophys.* **1992**, *30*, 311. [CrossRef]

115. Smith, K.M.; Zahn, O.; Dore, O. Detection of gravitational lensing in the cosmic microwave background. *Phys. Rev. D* **2007**, *76*, 043510. [CrossRef]

116. Refregier, A. Weak gravitational lensing by large-scale structure. *Annu. Rev. Astron. Astrophys.* 2003, *41*, 645. [CrossRef]

117. Sehgal, N.; Trac, H.; Acquaviva, V.; Ade, P.A.; Aguirre, P.; Amiri, M.; Appel, J.W.; Barrientos, L.F.; Battistelli, E.S.; Bond, J.R.; et al. The atacama cosmology telescope: cosmology from galaxy clusters detected via the sunyaev-zel'dovich effect. *Astrophys. J.* **2011**, *732*, 44. [CrossRef]

118. Allen, S.W.; Evrard, A.E.; Mantz, A.B. Cosmological parameters from observations of galaxy clusters. *Annu. Rev. Astron. Astrophys.* **2011**, *49*, 409. [CrossRef]

119. Hasselfield, M.; Hilton, M.; Marriage, T.A.; Addison, G.E.; Barrientos, L.F.; Battaglia, N.; Battistelli, E.S.; Bond, J.R.; Crichton, D.; Das, S.; et al. The atacama cosmology telescope: Sunyaev-zel'dovich selected galaxy clusters at 148 ghz from three seasons of data. *J. Cosmol. Astropart. Phys.* **2013**, *2013*, 008. [CrossRef]

120. Borgani, S.; Murante, G.; Springel, V.; Diaferio, A.; Dolag, K.; Moscardini, L.; Tormen, G.; Tornatore, L.; Tozzi, P. X-ray properties of galaxy clusters and groups from a cosmological hydrodynamical simulation. *Mon. Not. R. Astron. Soc.* **2004**, *348*, 1078. [CrossRef]

121. Schneider, A.; Teyssier, R.; Potter, D.; Stadel, J.; Onions, J.; Reed, D.S.; Smith, R.E.; Springel, V.; Pearce, F.R.; Scoccimarro, R. Matter power spectrum and the challenge of percent accuracy. *J. Cosmol. Astropart. Phys.* **2016**, *2016*, 47. [CrossRef]
122. Habib, S.; Heitmann, K.; Higdon, D.; Nakhleh, C.; Williams, B. Cosmic calibration: Constraints from the matter power spectrum and the cosmic microwave background. *Phys. Rev. D* **2007**, *76*, 083503. [CrossRef]
123. Rudd, J.; Whelan, K. Modeling inflation dynamics: A critical review of recent research. *J. Money Credit. Bank.* **2007**, *39*, 155. [CrossRef]
124. Agulló, I.; Navarro-Salas, J.; Olmo, G.J.; Parker, L. Revising the observable consequences of slow-roll inflation. *Phys. Rev. D* **2010**, *81*, 043514. [CrossRef]
125. Vázquez, J.A.; Padilla, L.E.; Matos, T. Inflationary Cosmology: From Theory to Observations. *arXiv* **2018**, arXiv:1810.09934.
126. Seljak, U.; Zaldarriaga, M. A Line of sight integration approach to cosmic microwave background anisotropies. *Astrophys. J.* **1996**, *469*, 437. [CrossRef]
127. Zaldarriaga, M.; Seljak, U.; Bertschinger, E. Integral solution for the microwave background anisotropies in nonflat universes. *Astrophys. J.* **1998**, *494*, 491. [CrossRef]
128. Zaldarriaga, M.; Seljak, U. Cmbfast for spatially closed universes. *Astrophys. J. Suppl.* **2000**, *129*, 431. [CrossRef]
129. Doran, M. CMBEASY: An object oriented code for the cosmic microwave background. *JCAP* **2005**, *10*, 11. [CrossRef]
130. Lewis, A.; Challinor, A.; Lasenby, A. Efficient computation of CMB anisotropies in closed FRW models. *Astrophys. J.* **2000**, *538*, 473. [CrossRef]
131. Howlett, C.; Lewis, A.; Hall, A.; Challinor, A. CMB power spectrum parameter degeneracies in the era of precision cosmology. *JCAP* **2012**, *1204*, 27. [CrossRef]
132. Blas, D.; Lesgourgues, J.; Tram, T. The Cosmic Linear Anisotropy Solving System (CLASS) II: Approximation schemes. *JCAP* **2011**, *7*, 34. [CrossRef]
133. Zumalacárregui, M.; Bellini, E.; Sawicki, I.; Lesgourgues, J.; Ferreira, P.G. hi_class: Horndeski in the Cosmic Linear Anisotropy Solving System. *JCAP* **2017**, *8*, 19. [CrossRef]
134. Zuntz, J.; Paterno, M.; Jennings, E.; Rudd, D.; Manzotti, A.; Dodelson, S.; Bridle, S.; Sehrish, S.; Kowalkowski, J.; CosmoSIS: Modular cosmological parameter estimation. *Astron. Comput.* **2015**, *12*, 45. [CrossRef]
135. Brinckmann, T.; Lesgourgues, J. MontePython 3: Boosted MCMC sampler and other features. *Phys. Dark Universe* **2019**, *24*, 100260. [CrossRef]
136. Audren, B.; Lesgourgues, J.; Benabed, K.; Prunet, S. Conservative Constraints on Early Cosmology: an illustration of the Monte Python cosmological parameter inference code. *JCAP* **2013**, *1302*, 1.
137. Audren, B.; Lesgourgues, J.; Benabed, K.; Prunet, S. Monte python: Monte carlo code for class in python. *Astrophys. Source Code Libr.* **2013**, *ascl-1307*.
138. Lewis, A.; Bridle, S. Cosmological parameters from CMB and other data: A Monte Carlo approach. *Phys. Rev. D* **2002**, *66*, 103511. [CrossRef]
139. Available online: https://github.com/slosar/april (accessed on 18 May 2021).
140. Speagle, J.S. Dynesty: A dynamic nested sampling package for estimating bayesian posteriors and evidences. *Mon. Not. R. Astron. Soc.* **2020**, *493*, 3132–3158. [CrossRef]
141. Available online: https://github.com/yabebalFantaye/MCEvidence (accessed on 18 May 2021).
142. Heavens, A.; Fantaye, Y.; Mootoovaloo, A.; Eggers, H.; Hosenie, Z.; Kroon, S.; Sellentin, E. Marginal Likelihoods from Monte Carlo Markov Chains. *arXiv* **2017**, arXiv:1704.03472.
143. Zhao, G.-B.; Raveri, M.; Pogosian, L.; Wang, Y.; Crittenden, R.G.; Handley, W.J.; Percival, W.J.; Beutler, F.; Brinkmann, J.; Chuang, C.-H.; et al. Dynamical dark energy in light of the latest observations. *Nat. Astron.* **2017**, *1*, 627. [CrossRef]
144. Heavens, A.; Fantaye, Y.; Sellentin, E.; Eggers, H.; Hosenie, Z.; Kroon, S.; Mootoovaloo, A. No evidence for extensions to the standard cosmological model. *Phys. Rev. Lett.* **2017**, *119*, 101301. [CrossRef] [PubMed]