Bayes in the sky:
Bayesian inference and model selection in cosmology

Roberto Trotta*

Oxford University, Astrophysics Department
Denys Wilkinson Building, Keble Rd, Oxford, OX1 3RH, UK
(*March 28, 2008*)

The application of Bayesian methods in cosmology and astrophysics has flourished over the past decade, spurred by data sets of increasing size and complexity. In many respects, Bayesian methods have proven to be vastly superior to more traditional statistical tools, offering the advantage of higher efficiency and of a consistent conceptual basis for dealing with the problem of induction in the presence of uncertainty. This trend is likely to continue in the future, when the way we collect, manipulate and analyse observations and compare them with theoretical models will assume an even more central role in cosmology.

This review is an introduction to Bayesian methods in cosmology and astrophysics and recent results in the field. I first present Bayesian probability theory and its conceptual underpinnings, Bayes' Theorem and the role of priors. I discuss the problem of parameter inference and its general solution, along with numerical techniques such as Monte Carlo Markov Chain methods. I then review the theory and application of Bayesian model comparison, discussing the notions of Bayesian evidence and effective model complexity, and how to compute and interpret those quantities. Recent developments in cosmological parameter extraction and Bayesian cosmological model building are summarized, highlighting the challenges that lie ahead.

Keywords: Bayesian methods; model comparison; cosmology; parameter inference; data analysis; statistical methods.

1 Introduction

At first glance, it might appear surprising that a trivial mathematical result obtained by an obscure minister over 200 hundred years ago ought still to excite so much interest across so many disciplines, from econometrics to biostatistics, from financial risk analysis to cosmology. Published posthumously thanks to Richard Price in 1763, "An essay towards solving a problem in the doctrine of chances" by the rev. Thomas Bayes (1701(?)–1761) [1] had nothing in it that could herald the growing importance and enormous domain of application that the subject of *Bayesian probability theory* would acquire more than two centuries afterwards. However, upon reflection there is a very good reason why Bayesian methods are undoubtedly on the rise in this particular historical epoch: the exponential increase in computational power of the last few decades made massive numerical inference feasible for the first time, thus opening the door to the exploitation of the power and flexibility of a rich set of Bayesian tools. Thanks to fast and cheap computing machines, previously unsolvable inference problems became tractable, and algorithms for numerical simulation flourished almost overnight.

Historically, the connections between physics and Bayesian statistics have always been very strong. Many ideas were developed because of related physical problems, and physicists made several distinguished contributions. One has only to think of people like Laplace, Bernouilli, Gauss, Metropolis, Jeffreys, etc. Cosmology is perhaps among the latest disciplines to have embraced Bayesian methods, a development mainly driven by the data explosion of the last decade, as Figure 1 indicates. However, motivated by difficult and computationally intensive inference problems, cosmologists are increasingly coming up with new solutions that add to the richness of a growing Bayesian literature.

*Email: rxt@astro.ox.ac.uk

Some cosmologists are sceptic regarding the usefulness of employing more advanced statistical methods, perhaps because they think with Mark Twain that there are "lies, damned lies and statistics". One argument that is often heard is that there is no point in bothering too much about refined statistical analyses, as better data will in the future resolve the question one way or another, be it the nature of dark energy or the initial conditions of the Universe. I strongly disagree with this view, and would instead argue that sophisticated statistical tools will be increasingly central for modern cosmology. This opinion is motivated by the following reasons:

(i) The complexity of the modelling of both our theories and observations will always increase, thus requiring correspondingly more refined statistical and data analysis skills. In fact, the scientific return of the next generation of surveys will be limited by the level of sophistication and efficiency of our inference tools.

(ii) The discovery zone for new physics is when a potentially new effect is seen at the 3–4 σ level. This is when tantalizing suggestion for an effect starts to accumulate but there is no firm evidence yet. In this potential discovery region a careful application of statistics can make the difference between claiming or missing a new discovery.

(iii) If you are a theoretician, you do not want to waste your time trying to explain an effect that is not there in the first place. A better appreciation of the interpretation of statistical statements might help in identifying robust claims from spurious ones.

(iv) Limited resources mean that we need to focus our efforts on the most promising avenues. Experiment forecast and optimization will increasingly become prominent as we need to use all of our current knowledge (*and* the associated uncertainty) to identify the observations and strategies that are likely to give the highest scientific return in a given field.

(v) Sometimes there will be no better data! This is the case for the many problems associated with cosmic variance limited measurements on large scales, for example in the cosmic background radiation, where the small number of independent directions on the sky makes it impossible to reduce the error below a certain level.

This review focuses on Bayesian methodologies and related issues, presenting some illustrative results where appropriate and reviewing the current state–of–the art of Bayesian methods in cosmology. The emphasis is on the innovative character of Bayesian tools. The level is introductory, pitched for graduate students who are approaching the field for the first time, aiming at bridging the gap between basic textbook examples and application to current research. In the last sections we present some more advanced material that we hope might be useful for the seasoned practitioner, too. A basic understanding of cosmology and of the interplay between theory and cosmological observations (at the level of the introductory chapters in [2]) is assumed. A full list of references is provided as a comprehensive guidance to relevant literature across disciplines.

This paper is organized in two main parts. The first part, sections 2–4, focuses on probability theory, methodological issues and Bayesian methods generally. In section 2 we present the fundamental distinction between probability as frequency or as degree of belief, we introduce Bayes' Theorem and discuss the meaning and role of priors in Bayesian theory. Section 3 is devoted to Bayesian parameter inference and related issues in parameter extraction. Section 4 deals with the topic of Bayesian model comparison from a conceptual and technical point of view, covering Occam's razor principle, its practical implementation in the form of the Bayesian evidence, the effective number of model parameters and information criteria for approximate model comparison. The second part presents applications to cosmological parameter inference and related topics (section 5) and to Bayesian cosmological model building (section 6), including multi–model inference and model comparison forecasting. Section 7 gives our conclusions.

2 Bayesian probability theory

In this section we introduce the basic concepts and the notation we employ. After a discussion of what probability is, we turn to the central formula for Bayesian inference, namely Bayes theorem. The whole of
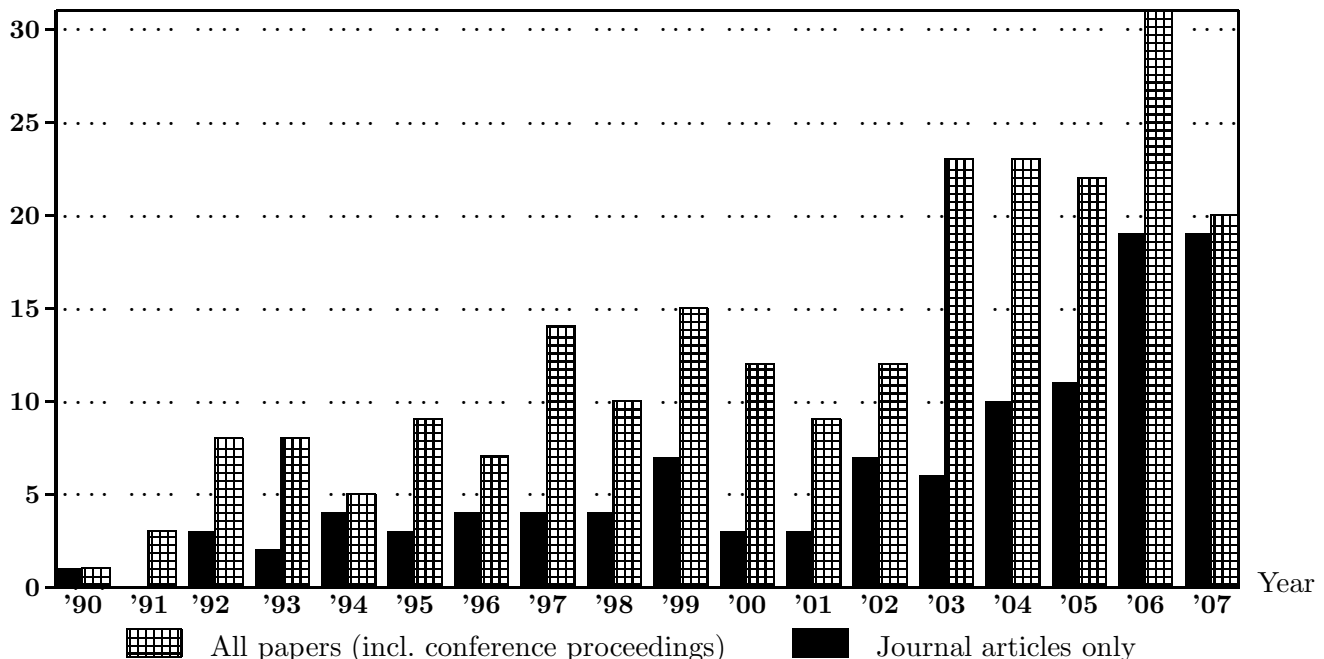
Number of Bayesian papers in cosmology and astrophysics

Figure 1. The evolution of the B–word: number of articles in astronomy and cosmology with "Bayesian" in the title, as a function of publication year. The number of papers employing one form or another of Bayesian methods is of course much larger than that. Up until about 1995, Bayesian papers were concerned mostly with image reconstruction techniques, while in subsequent years the domain of application grew to include signal processing, parameter extraction, object detection, cosmological model building, decision theory and experiment optimization, and much more. It appears that interest in Bayesian statistics began growing around 2002 (source: NASA/ADS).

Bayesian inference follows from this extremely simple cornerstone. We then present some views about the meaning of priors and their role in Bayesian theory, an issue which has always been (wrongly) considered a weak point of Bayesian statistics.

There are many excellent textbooks on Bayesian statistics: the works by Sir Harold Jeffreys [3] and Bruno de Finetti [4] are classics, while an excellent modern introduction with an extensive reading list is given by [5]. A good textbook is [6]. Worth reading as a source of inspiration is the though–provoking monograph by E.T. Jaynes [7]. Computational aspects are treated in [8], while MacKay [9] has a quite unconventional but inspiring choice of topics with many useful exercices. Two very good textbooks on the subject written by physicists are [10, 11]. A nice introductory review aimed at physicists is [12] (see also [13]). Tom Loredo has some masterfully written introductory material, too [14, 15]. A good source expanding on many of the topics covered here is Ref. [16].

2.1 *What is probability?*

2.1.1 *Probability as frequency.* The classical approach to statistics defines the probability of an event as

> "the number of times the event occurs over the total number of trials, in the limit of an infinite series of equiprobable repetitions."

This is the so–called *frequentist* school of thought. This definition of probability is however unsatisfactory in many respects.

(i) Strikingly, this definition of probability in terms of relative frequency of outcomes is *circular*, i.e. it assumes that repeated trials have the same probability of outcomes – but it was the the very notion of *probability* that we were trying to define in the first place!

(ii) It cannot handle with unrepeatable situations, such as the probability that I will be overrun by a

car when crossing the street, or, in the cosmological context, questions concerning the properties of the observable Universe as a whole, of which we have exactly one sample. Indeed, perfectly legitimate questions such as "what is the probability that it was raining in Oxford when William I was crowned?" cannot even be formulated in classical statistics.

(iii) The definition only holds exactly for an infinite sequence of repetitions. In practice we always handle with a finite number of measurements, sometimes with actually only a very small number of them. How can we assess when "how many repetitions" are sufficient? And what shall we do when we have only a handful of repetitions? Frequentist statistics does not say, except sometimes devising complicated *ad-hockeries* to correct for "small sample size" effects. In practice, physicists tend to forget about the "infinite series" requirement and use this definitions and the results that go with it (for example, about asymptotic distributions of test statistics) for whatever number of samples they happen to be working with.

Another, more subtle aspects has to do with the notion of "randomness". Restricting ourselves to classical (non–chaotic) physical systems for now, let us consider the paradigmatic example of a series of coin tosses. From an observed sequence of heads and tails we would like to come up with a statistical statement about the fairness of the coin, which is deemed to be "fair" if the probability of getting heads is $p_H = 0.5$. At first sight, it might appear plausible that the task is to determine whether the coin possesses some physical property (for example, a tensor of intertia symmetric about the plane of the coin) that will ensure that the outcome is indifferent with respect to the interchange of heads and tails. As forcefully argued by Jaynes [7], however, the probability of the outcome of a sequence of tosses has nothing to do with the physical properties of the coin being tested! In fact, a skilled coin–tosser (or a purpose–built machine, see [17]) can influence the outcome quite independently of whether the coin is well–balanced (i.e., symmetric) or heavily loaded. The key to the outcome is in fact the definition of *random* toss. In a loose, intuitive fashion, we sense that a carefully controlled toss, say in which we are able to set quite precisely the spin and speed of the coin, will spoil the "randomness" of the experiment — in fact, we might well call it "cheating". However, lacking a precise operational definition of what a "random toss" means, we cannot meaningfully talk of the probability of getting heads as of a physical property of the coin itself. It appears that the outcome depends on our *state of knowledge* about the initial conditions of the system (angular momentum and velocity of the toss): an lack of precise information about the initial conditions results in a state of knowledge of indifference about the possible outcome with respect to the specification of heads or tails. If however we insist on defining probability in terms of the outcome of random experiments, we immediately get locked up in a circularity when we try to specify what "random" means. For example, one could say that

"*a random toss is one for which the sequence of heads and tails is compatible with assuming the hypothesis $p_H = 0.5$*".

But the latter statement is exactly what we were trying to test in the first place by using a sequence of random tosses! We are back to the problem of circular definition we highlighted above.

2.1.2   *Probability as degree of belief.* Many of the limitations above can be avoided and paradoxes resolved by taking a Bayesian stance about probabilities. The Bayesian viewpoint is based on the simple and intuitive *tenet* that

"*probability is a measure of the degree of belief about a proposition*".

It is immediately clear that this definition of probability applies to any event, regardless whether we are considering repeated experiments (e.g., what is the probability of obtaining 10 heads in as many tosses of a coin?) or one–off situations (e.g., what is the probability that it will rain tomorrow?). Another advantage is that it deals with uncertainty independently of its origin, i.e. there is no distinction between "statistical uncertainty" coming from the finite precision of the measurement apparatus and the associated random noise and "systematic uncertainty", deriving from deterministic effects that are only partially known (e.g., calibration uncertainty of a detector). From the coin tossing example above we learn that it makes good sense to think of probability as a state of knowledge in presence of partial information and that

"randomness" is really a consequence of our lack of information about the exact conditions of the system (if we knew the precise way the coin is flipped we could predict the outcome of any toss with certainty. The case of quantum probabilities is discussed below). The rule for manipulating states of belief is given by Bayes' Theorem, which is introduced in Eq. (5) below.

It seems to us that the above arguments strongly favour the Bayesian view of probability (a more detailed discussion can be found in [7, 14]). Ultimately, as physicists we might as well take the pragmatic view that the approach that yields demonstrably superior results ought to be preferred. In many real–life cases, there are several good reasons to prefer a Bayesian viewpoint:

(i) Classic frequentist methods are often based on asymptotic properties of estimators. Only a handful of cases exist that are simple enough to be amenable to analytic treatment (in physical problems one most often encounters the Normal and the Poisson distribution). Often, methods based on such distributions are employed not because they accurately describe the problem at hand, but because of the lack of better tools. This can lead to serious mistakes. Bayesian inference is not concerned by such problems: it can be shown that *application of Bayes' Theorem recovers frequentist results (in the long run) for cases simple enough where such results exist*, while remaining applicable to questions that cannot even be asked in a frequentist context.

(ii) Bayesian inference deals effortlessly with *nuisance parameters*. Those are parameters that have an influence on the data but are of no interest for us. For example, a problem commonly encountered in astrophysics is the estimation of a signal in the presence of a background rate (see [14, 18, 19]). The particles of interest might be photons, neutrinos or cosmic rays. Measurements of the source s must account for uncertainty in the background, described by a nuisance parameter b. The Bayesian procedure is straightforward: infer the joint probability of s and b and then integrate over the uninteresting nuisance parameter b ("marginalization", see Eq. (16)). Frequentist methods offer no simple way of dealing with nuisance parameters (the very name derives from the difficulty of accounting for them in classical statistics). However neglecting nuisance parameters or fixing them to their best–fit value can result in a very serious underestimation of the uncertainty on the parameters of interest (see [20] for an example involving galaxy evolution models).

(iii) In many situations *prior information* is highly relevant and omitting it would result in seriously wrong inferences. The simplest case is when the parameters of interest have a physical meaning that restricts their possible values: masses, count rates, power and light intensity are examples of quantities that must be positive. Frequentist procedures based only on the likelihood can give best–fit estimates that are negative, and hence meaningless, unless special care is taken (for example, constrained likelihood methods). This often happens in the regime of small counts or low signal to noise. The use of Bayes' Theorem ensures that relevant prior information is accounted for in the final inference and that physically meaningless results are weeded out from the beginning.

(iv) Bayesian statistics only deals with the *data that were actually observed*, while frequentist methods focus on the distribution of possible data that have not been obtained. As a consequence, *frequentist results can depend on what the experimenter thinks about the probability of data that have not been observed.* (this is called the "stopping rule" problem). This state of affairs is obviously absurd. Our inferences should not depend on the probability of what could have happened but should be conditional on whatever has actually occurred. This is built into Bayesian methods from the beginning since inferences are by construction conditional on the observed data.

However one looks at the question, it is fair to say that the debate among statisticians is far from settled (for a discussion geared for physicists, see [21]). Louis Lyons neatly summarized the state of affairs by saying that [22]

> "*Bayesians address the question everyone is interested in by using assumptions no–one believes, while frequentists use impeccable logic to deal with an issue of no interest to anyone*".

2.1.3   *Quantum probability.* Ultimately the quantum nature of the microscopic world ensures that the fundamental meaning of probability is to be found in a theory of quantum measurement. The fundamental problem is how to make sense of the indeterminism brought about when the wavefunction collapses into one or other of the eigenstates being measured. The classic textbook view is that the probability of each outcome is given by the square of the amplitude, the so–called "Born rule". However the question remains — where do quantum probabilities come from?

Recent developments have scrutinized the scenario of the Everettian many–worlds interpretation of quantum mechanics, in which the collapse never happens but rather the world "splits" in disconnected "branches" every time a quantum measurement is performed. Although all outcomes actually occur, uncertainty comes into the picture because an observer is unsure about which outcome will occur in her branch. David Deutsch [23] suggested to consider the problem from a decision–theoretic point of view. He proposed to consider a formal system in which the probabilities ("weights") to be assigned to each quantum branch are expressed in terms of the preferences of a rational agent playing a quantum game. Each outcome of the game (i.e., weights assignment) has an associated utility function that determines the payoff of the rational agent. Being rational, the agent will act to maximise the expectation value of her utility. The claim is that it is possible to find a simple and plausible set or rules for the quantum game that allow to derive unique outcomes (i.e., probability assignments) in agreement with the Born rule, *independently* of the utility function chosen, i.e. the payoffs. In such a way, one obtains a bridge between subjective probabilities (the weight assignment by the rational agent) and quantum chance (the weights attached to the branches according to the Born rule).

This approach to the problem of quantum measurement remains highly controversial. For an introduction and further reading, see e.g. [24, 25] and references therein. An interesting comparison between classical and quantum probabilities can be found in [26].

2.2   *Bayes' Theorem*

Bayes' Theorem is a simple consequence of the axioms of probability theory, giving the rules by which probabilities (understood as degree of belief in propositions) should be manipulated. As a mathematical statement it is not controversial — what is a matter of debate is whether it should be used as a basis for inference and in general for dealing with uncertainty. In the previous section we have given some arguments why we strongly believe this to be the case. An important result is that Bayes' Theorem can be derived from a set of basic consistency requirements for plausible reasoning, known as Cox axioms [27]. Therefore, Bayesian probability theory can be shown to be the unique generalization of boolean logic into a formal system to manipulate propositions in the presence of uncertainty [7]. In other words, Bayesian inference is the unique generalization of logical deduction when the available information is incomplete.

We now turn to the presentation of the actual mathematical framework. We adopt a fairly relaxed notation, as mathematical rigour is not the aim of this review. For a more formal introduction, see e.g. [3,7].

Let us consider a proposition A, which could be a random variable (e.g., the probability of obtaining 12 when rolling two dices) or a one–off proposition (the probability that Prince Charles will become king in 2015), and its negation, $\overline{A}$. The *sum rule* reads

$$p(A|I) + p(\overline{A}|I) = 1, \tag{1}$$

where the vertical bar means that the probability assignment is conditional on assuming whatever information is given on its right. Above, I represents any relevant information that is assumed to be true. The *product rule* is written as

$$p(A, B|I) = p(A|B, I)p(B|I), \tag{2}$$

which says that the *joint probability* of A and B equals to the probability of A given that B occurs times the probability of B occurring on its own (both conditional on information I). If we are interested in the

probability of B alone, irrespective of A, the sum and product rules together imply that

$$p(B|I) = \sum_A p(A, B|I), \qquad (3)$$

where the sum runs over the possible outcomes for proposition A. The quantity on the left–hand–side is called *marginal probability* of B. Since obviously $p(A, B|I) = p(B, A|I)$, the product rule can be rewritten to give Bayes' Theorem:

$$p(B|A, I) = \frac{p(A|B, I)p(B|I)}{p(A|I)} \quad \text{(Bayes theorem)}. \qquad (4)$$

The interpretation of this simple result is more illuminating if one replaces for A the observed data d and for B the hypothesis H we want to assess, obtaining

$$p(H|d, I) = \frac{p(d|H, I)p(H|I)}{p(d|I)}. \qquad (5)$$

On the left–hand side, $p(H|d, I)$ is the *posterior probability* of the hypothesis taking the data into account. This is proportional to the *sampling distribution* of the data $p(d|H, I)$ assuming the hypothesis is true, times the *prior probability* for the hypothesis, $p(H|I)$ ("the prior", conditional on whatever external information we have, I), which represents our state of knowledge before seeing the data. The sampling distribution encodes how the degree of plausibility of the hypothesis changes when we acquire new data. Considered as a function of the hypothesis, for fixed data (the ones that have been observed), it is called *the likelihood function* and we will often employ the shortcut notation $\mathcal{L}(H) \equiv p(d|H, I)$. Notice that as a function of the hypothesis the likelihood is *not* a probability distribution. The normalization constant on the right–hand–side in the denominator is the *marginal likelihood* (in cosmology often called the "Bayesian evidence") given by

$$p(d|I) \equiv \sum_H p(d|H, I)p(H|I) \quad \text{(Bayesian evidence)}. \qquad (6)$$

where the sum runs over all the possible outcomes for the hypothesis H. This is the central quantity for model comparison purposes, and it is further discussed in section 4. The posterior is the relevant quantity for Bayesian inference as it represents our state of belief about the hypothesis after we have considered the information in the data (hence the name). Notice that there is a *logical* sequence in going from the prior to the posterior, not necessarily a *temporal* one, i.e. a scientist might well specify the prior after the data have been gathered provided that the prior reflects her state of knowledge irrespective of the data. Bayes' Theorem is therefore a prescription as to how one learns from experience. It gives a unique rule to update one's beliefs in the light of the observed data.

The need to specify a prior describing a "subjective" state of knowledge has exposed Bayesian inference to the criticism that it is not objective, and hence unfit for scientific reasoning. Exactly the contrary is true — a thorny issue to which we now briefly turn out attention.

2.3 *Subjectivity, priors and all that*

The prior choice is a fundamental ingredient of Bayesian statistics. Historically, it has been regarded as problematic, since the theory does not give guidance about how the prior should be selected. Here we argue that this issue has been given undue emphasis and that prior specification should be regarded as a *feature* of Bayesian statistics, rather than a limitation.

The guiding principle of Bayesian probability theory is that there can be no inference without assumptions, and thus the prior choice ought to reflect as accurately as possible one's assumptions and state of knowledge about the problem in question before the data come along. Far from undermining objectivity,

this is obviously a positive feature, because Bayes' Theorem gives a *univoque* procedure to update different degrees of beliefs that different scientist might have held before seeing the data. Furthermore, there are many cases where prior (i.e., external) information is relevant and it is sensible to include it in the inference procedure[1].

It is only natural that two scientists might have different priors as a consequence of their past scientific experiences, theoretical outlook and based on the outcome of previous observations they might have performed. As long as the prior $p(H|I)$ (where the extra conditioning on I denotes the external information of the kind listed above) has a support that is non–zero in regions where the likelihood is large, repeated application of Bayes theorem, Eq. (5), will lead to a posterior pdf that converges to a common, i.e. objective inference on the hypothesis. As an example, consider the case where the inference concerns the value of some physical quantity θ, in which case $p(\theta|d, I)$ has to be interpreted as the posterior probability *density* and $p(\theta|d, I)\mathrm{d}\theta$ is the probability of θ to take on a value between θ and $\theta + \mathrm{d}\theta$. Alice and Bob have different prior beliefs regarding the possible value of θ, perhaps based on previous, independent measurements of that quantity that they have performed. Let us denote their priors by $p(\theta|I_i)$ $(i = A, B)$ and let us assume that they are described by two Gaussian distributions of mean μ_i and variance Σ_i^2, $i = A, B$ representing the state of knowledge of Alice and Bob, respectively. Alice and Bob go together in the lab and perform a measurement of θ with an apparatus subject to Gaussian noise of known variance σ^2. They obtain a value m_1, hence their likelihood is

$$\mathcal{L}(\theta) \equiv p(m_1|\theta) = \mathcal{L}_0 \exp\left(-\frac{1}{2}\frac{(\theta - m_1)^2}{\sigma^2}\right). \tag{7}$$

Replacing the hypothesis H by the continuous variable θ in Bayes' Theorem[1], we obtain for their respective posterior pdf's after the new datum

$$p(\theta|m_1, I_i) = \frac{\mathcal{L}(\theta)p(\theta|I_i)}{p(m_1|I_i)} \quad (i = A, B). \tag{8}$$

It is easy to see that the posterior pdf's of Alice and Bob are again Gaussians with means

$$\overline{\mu}_i = \frac{m_1 + (\sigma/\Sigma_i)^2 \mu_i}{1 + (\sigma/\Sigma_i)^2} \tag{9}$$

and variance

$$\tau_i^2 = \frac{\sigma^2}{1 + (\sigma/\Sigma_i)^2} \quad (i = A, B). \tag{10}$$

Thus if the likelihood is more informative than the prior, i.e. for $(\sigma/\Sigma_i) \ll 1$ the posterior means of Bob and Alice will converge towards the measured value, m_1. As more and more data points are gathered, one can simply replace m_1 in the above equations by the mean $\overline{m}$ of the observations and σ^2 by σ^2/N, with N the number of data points. Thus we can see that the initial prior means μ_i of Alice and Bob will progressively be overridden by the data. This process is illustrated in Figure 2.

Finally, objectivity is ensured by the fact that two scientists in the same state of knowledge should assign the same prior, hence their posterior are identical if they observe the same data. The fact that the prior assignment eventually becomes irrelevant as better and better data make the posterior likelihood–dominated in uncontroversial in principle but may be problematic in practice. Often the data are not strong enough to override the prior, in which case great care must be given in assessing how much of

---

[1]As argued above, often it would be a mistake not to do so, for example when trying to estimate a mass m from some data one should enforce it to be a positive quantity by requiring that $p(m) = 0$ for $m < 0$.

[1]Strictly speaking, Bayes' Theorem holds for discrete probabilities and the passage to hypotheses represented by continuous variables ought to be performed with some mathematical care. Here we simply appeal to the intuition of physicists without being too much concerned by mathematical rigour.
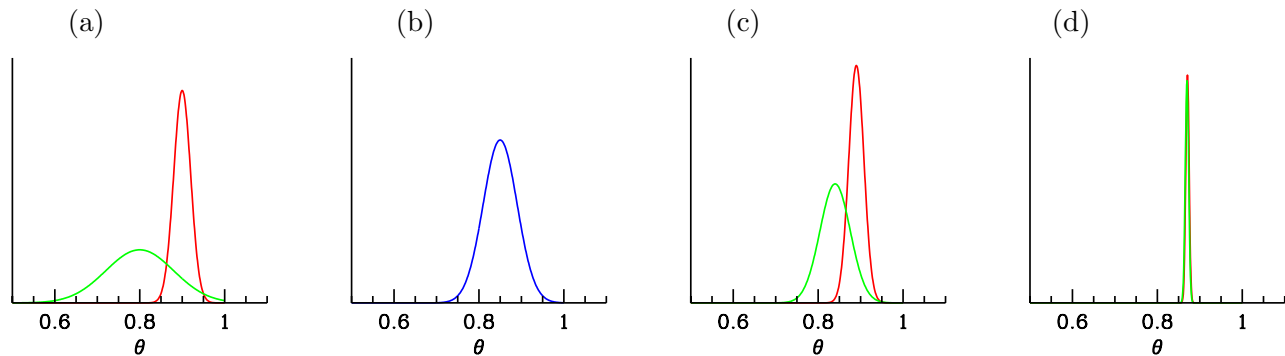
(a)   (b)   (c)   (d)

Figure 2. Converging views in Bayesian inference. Two scientists having different prior believes $p(\theta|I_i)$ about the value of a quantity θ (panel (a), red and green pdf's) observe one datum with likelihood $\mathcal{L}(\theta)$ (panel (b)), after which their posteriors $p(\theta|m_1)$ (panel (c), obtained via Bayes Theorem, Eq. (8)) represent their updated states of knowledge on the parameter. After observing 100 data points, the two posteriors have become essentially indistinguishable (d).

the final inference depends on the prior choice. This might occur for small sample sizes, or for problems where the dimensionality of the hypothesis space is larger than the number of observations (for example, in image reconstruction). Even in such a case, if different prior choices lead to different posteriors we can still conclude that the data are not informative enough to completely override our prior state of knowledge and hence we have learned something useful about the constraining power (or lack thereof) of the data.

The situation is somewhat different for model comparison questions. In this case, it is precisely the available prior volume that is important in determining the penalty that more complex models with more free parameters should incur into (this is discussed in detail in section 4). Hence the impact of the prior choice is much stronger when dealing with model selection issues, and care should be exercised in assessing how much the outcome would change for physically reasonable changes in the prior.

There is a vast literature on priors that we cannot begin to summarize here. Important issues concern the determination of "ignorance priors", i.e. priors reflecting a state of indifference with respect to symmetries of the problem considered. "Reference priors" exploit the idea of using characteristics of what the experiment is expected to provide to construct priors representing the least informative state of knowledge. In order to be probability distributions, priors must be *proper*, i.e. normalizable to unity probability content. "Flat priors" are often a standard choice, in which the prior is taken to be constant within some minimum and maximum value of the parameters, i.e. for a 1–dimensional case $p(\theta) = (\theta_{\max} - \theta_{\min})^{-1}$. The rationale is that we should assign equal probability to equal states of knowledge. However, flat priors are not always as harmless as they appear. One reason is that a flat prior on a parameter θ does not correspond to a flat prior on a non–linear function of that parameter, $\psi(\theta)$. The two priors are related by

$$p(\psi) = p(\theta) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\psi} \right|, \tag{11}$$

so for a non–linear dependence $\psi(\theta)$ the term $|\mathrm{d}\theta/\mathrm{d}\psi|$ means that an uninformative (flat) prior on θ might be strongly informative about ψ (in the multi–dimensional case, the derivative term is replaced by the determinant of the Jacobian for the transformation). Furthermore, if we are ignorant about the *scale* of a quantity θ, it can be shown (see e.g. [7]) that the appropriate prior is flat on $\ln \theta$, which gives equal weight to all orders of magnitude. This prior corresponds to $p(\theta) \propto \theta^{-1}$ and is called "Jeffreys' prior". It is appropriate for example for the rate of a Poisson distribution. For further details on prior choice, and especially so–called "objective priors", see e.g. chapter 5 in [28] and references therein.

3 Bayesian parameter inference

3.1 *The general problem and its solution*

The general problem of Bayesian parameter inference can be specified as follows. We first choose a model containing a set of hypotheses in the form of a vector of parameters, θ. The parameters might describe

any aspect of the model, but usually they will represent some physically meaningful quantity, such as for example the mass of an extra–solar planet or the abundance of dark matter in the Universe. Together with the model we must specify the priors for the parameters. Priors should summarize our state of knowledge about the parameters before we consider the new data, and for the parameter inference step the prior for a new observation might be taken to be the posterior from a previous measurement (for model comparison issues the prior is better understood in a different way, see section 4). The *caveats* about priors and prior specifications presented in the previous section will apply at this stage.

The central step is to construct the likelihood function for the measurement, which usually reflects the way the data are obtained. For example, a measurement with Gaussian noise will be represented by a Normal distribution, while γ–ray counts on a detector will have a Poisson distribution for a likelihood. Nuisance parameters related to the measurement process might be present in the likelihood, e.g. the variance of the Gaussian might be unknown or the background rate in the absence of the source might be subject to uncertainty. This is no matter of concern for a Bayesian, as the general strategy is always to work out the joint posterior for all of the parameters in the problem and then marginalize over the ones we are not interested in. Assuming that we have a set of physically interesting parameters ϕ and a set of nuisance parameters ψ, the joint posterior for $\theta = (\phi, \psi)$ is obtained through Bayes' Theorem:

$$p(\theta|d, \mathcal{M}) = \mathcal{L}(\theta)\frac{p(\theta|\mathcal{M})}{p(d|\mathcal{M})} \tag{12}$$

where we have made explicit the choice of a model $\mathcal{M}$ by writing it on the righ–hand–side of the conditioning symbol. Recall that $\mathcal{L}(\theta) \equiv p(d|\theta, \mathcal{M})$ denotes the likelihood and $p(\theta|\mathcal{M})$ the prior. The normalizing constant $p(d|\mathcal{M})$ ("the Bayesian evidence") is irrelevant for parameter inference (but central to model comparison, see section 4), so we can write the marginal posterior on the parameter of interest as (marginalizing over the nuisance parameters)

$$p(\phi|d, \mathcal{M}) \propto \int \mathcal{L}(\phi, \psi)p(\phi, \psi|\mathcal{M})\mathrm{d}\psi. \tag{13}$$

The final inference on ϕ from the posterior can then be communicated either by some summary statistics (such as the mean, the median or the mode of the distribution, its standard deviation and the correlation matrix among the components) or more usefully (especially for cases where the posterior presents multiple peaks or heavy tails) by plotting one or two dimensional subsets of ϕ, with the other components marginalized over.

In real life there are only a few cases of interest for which the above procedure can be carried out analytically. Quite often, however, the simple case of a Gaussian prior and a Gaussian likelihood can offer useful guidance regarding the behaviour of more complex problems. An analytical model of a Poisson–distributed likelihood for estimating source counts in the presence of a background signal is worked out in [14]. In general, however, actual problems in cosmology and astrophysics are not analytically tractable and one must resort to numerical techniques to evaluate the likelihood and to draw samples from the posterior. Fortunately this is not a major hurdle thanks to the recent increase of cheap computational power. In particular, numerical inference often employs a technique called *Markov Chain Monte Carlo*, which allows to map out numerically the posterior distribution of Eq. (12) even in the most complicated situations, where the likelihood can only be obtained by numerical simulation, the parameter space can have hundreds of dimensions and the posterior has multiple peaks and a complicated structure.

For further reading about Bayesian parameter inference, see [10, 11, 29, 30]. For more advanced applications to problems in astrophysics and cosmology, see [16, 31].

3.2 *Markov Chain Monte Carlo techniques for parameter inference*

The general solution to any inference problem has been outlined in the section above: it remains to find a way to evaluate the posterior of Eq. (12) for the usual case where analytical solutions do not exist or are

insufficiently accurate. Nowadays, Bayesian inference heavily relies on numerical simulation, in particular in the form of Markov Chain Monte Carlo (MCMC) techniques, which are discussed in this section.

The purpose of the Markov chain Monte Carlo algorithm is to construct a sequence of points in parameter space (called "a chain"), whose density is proportional to the posterior pdf of Eq. (12). Developing a full theory of Markov chains is beyond the scope of the present article (see e.g. [32,33] instead). For our purposes it suffices to say that a Markov chain is defined as a sequence of random variables $\{X^{(0)}, X^{(1)}, \ldots, X^{(M-1)}\}$ such that the probability of the $(t + 1)$–th element in the chain only depends on the value of the t–th element. The crucial property of Markov chains is that they can be shown to converge to a stationary state (i.e., which does not change with t) where successive elements of the chain are samples from the *target distribution*, in our case the posterior $p(\theta|d)$. The generation of the elements of the chain is probabilistic in nature, and several algorithms are available to construct Markov chains. The choice of algorithm is highly dependent on the characteristics of the problem at hand, and "tayloring" the MCMC to the posterior one wants to explore often takes a lot of effort. Popular and effective algorithms include the Metropolis–Hastings algorithm [34, 35], Gibbs sampling (see e.g. [36]), Hamiltonian Monte Carlo (see e.g. [37] and importance sampling.

Once a Markov chain has been constructed, obtaining Monte Carlo estimates of expectations for any function of the parameters becomes a trivial task. For example, the posterior mean is given by (where $\langle \cdot \rangle$ denotes the expectation value with respect to the posterior)

$$\langle \theta \rangle \approx \int p(\theta|d)\theta \mathrm{d}\theta = \frac{1}{M} \sum_{t=0}^{M-1} \theta^{(t)}, \tag{14}$$

where the equality with the mean of the samples from the MCMC follows because the samples $\theta^{(t)}$ are generated from the posterior by construction. In general, one can easily obtain the expectation value of any function of the parameters $f(\theta)$ as

$$\langle f(\theta) \rangle \approx \frac{1}{M} \sum_{t=0}^{M-1} f(\theta^{(t)}). \tag{15}$$

It is usually interesting to summarize the results of the inference by giving the 1–dimensional *marginal probability* for the j–th element of θ, θ_j. Taking without loss of generality $j = 1$ and a parameter space of dimensionality n, the equivalent expression to Eq. (3) for the case of continuous variables is

$$p(\theta_1|d) = \int p(\theta|d)\mathrm{d}\theta_2 \ldots \mathrm{d}\theta_n, \tag{16}$$

where $p(\theta_1|d)$ is the *marginal posterior* for the parameter θ_1. From the Markov chain it is trivial to obtain and plot the marginal posterior on the left–hand–side of Eq. (16): since the elements of the Markov chains are samples from the full posterior, $p(\theta|d)$, their density reflects the value of the full posterior pdf. It is then sufficient to divide the range of θ_1 in a series of bins and *count the number of samples falling within each bin*, simply ignoring the coordinates values $\theta_2, \ldots, \theta_n$. A 2–dimensional posterior is defined in an analogous fashion.

There are several important practical issues in working with MCMC methods (for details see e.g. [32]). Especially for high–dimensional parameter spaces with multi–modal posteriors it is important *not* to use MCMC techniques as a black box, since poor exploration of the posterior can lead to serious mistakes in the final inference if it remains undetected. Considerable care is required to ensure as much as possible that the MCMC exploration has covered the relevant parameter space.

4 Bayesian model comparison

4.1 *Shaving theories with Occam's razor*

When there are several competing theoretical models, Bayesian model comparison provides a formal way of evaluating their relative probabilities in light of the data and any prior information available. The "best" model is then the one which strikes an optimum balance between quality of fit and predictivity. In fact, it is obvious that a model with more free parameters will always fit the data better (or at least as good as) a model with less parameters. However, more free parameters also mean a more "complex" model, in a sense that we will quantify below in section 4.6. Such an added complexity ought to be avoided whenever a simpler model provides an adequate description of the observations. This guiding principle of simplicity and economy of an explanation is known as *Occam's razor* — the simplest theory compatible with the available evidence ought to be preferred[1]. Bayesian model comparison offers a formal way to evaluate whether the extra complexity of a model is required by the data, thus putting on a firmer statistical grounds the evaluation and selection process of scientific theories that scientists often carry out at a more intuitive level. For example, a Bayesian model comparison of the Ptolemaic model of epicycles versus the heliocentric model based on Newtonian gravity would favour the latter because of its simplicity and ability to explain planetary motions in a more economic fashion than the baroque construction of epicycles.

An important feature is that an alternative model must be specified against which the comparison is made. In contrast with frequentist goodness–of–fit tests (such as chi–square tests), Bayesian model comparison maintains that it is pointless to reject a theory unless an alternative explanation is available that fits the observed facts better (for more details about the difference in approach with frequentist hypothesis testing, see [14]). In other words, unless the observations are totally impossible within a model, finding that the data are improbable given a theory does not say anything about the probability of the theory itself *unless we can compare it with an alternative.* A consequence of this is that the probability of a theory that makes a correct prediction can increase if the prediction is confirmed by observations, provided competitor theories do not make the same prediction. This agrees with our intuition that a verified prediction lends support to the theory that made it, in contrast with the limited concept of falsifiability advocated by Popper (i.e., that scientific theories can only be tested by proving them wrong). So for example, perturbations to the motion of Uranus led the French astronomer U.J.J Leverrier and the English scholar J.C. Adams to formulate the prediction, based on Newtonian theory, that a further planet ought to exist beyond the orbit of Uranus. The discovery of Neptune in 1846 within 1 degree of the predicted position thus should strengthen our belief in the correctness of Newtonian gravity. However, as discussed in detail in chapter 5 of Ref [7], the change in the plausibility of Newton's theory following the discovery of Uranus crucially depends on the alternative we are considering. If the alternative theory is Einstein gravity, then obviously the two theories make the same predictions as far as the orbit of Uranus is concerned, hence their relative plausibility is unchanged by the discovery. The alternative "Newton theory is false" is not useful in Bayesian model comparison, and we are forced to put on the table a more specific model than that before we can assess how much the new observation changes our relative degree of belief between an alternative theory and Newtonian gravity.

In the context of model comparison it is appropriate to think of a model as a specification of a set of parameters θ *and* of their prior distribution, $p(\theta|\mathcal{M})$. As shown below, it is the number of free parameters *and* their prior range that control the strength of the Occam's razor effect in Bayesian model comparison: models that have many parameters that can take on a wide range of values but that are not needed in the light of the data are penalized for their unwarranted complexity. Therefore, *the prior choice ought to reflect the available parameter space under the model $\mathcal{M}$, independently of experimental constraints we might already be aware of.* This is because we are trying to assess the economy (or simplicity) of the model itself, and hence the prior should be based on theoretical or physical constraints on the model under consideration. Often these will take the form of a range of values that are deemed "intuitively" plausible, or "natural". Thus the prior specification is inherent in the model comparison approach.

---

[1]In its formulation by the medieval English philosopher and Franciscan monk William of Ockham (ca. 1285-1349): "Pluralitas non est ponenda sine neccesitate".

The prime tool for model selection is the *Bayesian evidence*, discussed in the next three sections. A quantitative measure of the *effective model complexity* is introduced in section 4.6. We then present some popular approximations to the full Bayesian evidence that attempt to avoid the difficulty of priors choice, the *information criteria*, and discuss the limits of their applicability in section 4.7.

For reviews on model selection see e.g. [3, 9, 38] and [39] for cosmological applications. Good starting points on Bayes factors are [40, 41]. A discussion of the spirit of model selection can be found in the first part of [42].

4.2 *The Bayesian evidence*

The evaluation of a model's performance in the light of the data is based on the *Bayesian evidence*, which in the statistical literature is often called *marginal likelihood* or *model likelihood*. Here we follow the practice of the cosmology and astrophysics community and will use the term "evidence" instead. The evidence is the normalization integral on the right–hand–side of Bayes' theorem, Eq. (6), which we rewrite here for a continuous parameter space $\Omega_{\mathcal{M}}$ and conditioning explicitly on the model under consideration, $\mathcal{M}$:

$$p(d|\mathcal{M}) \equiv \int_{\Omega_{\mathcal{M}}} p(d|\theta, \mathcal{M})p(\theta|\mathcal{M})\mathrm{d}\theta \quad \text{(Bayesian evidence)}. \tag{17}$$

Thus the Bayesian evidence is the average of the likelihood under the prior for a specific model choice. From the evidence, the model posterior probability given the data is obtained by using Bayes' Theorem to invert the order of conditioning:

$$p(\mathcal{M}|d) \propto p(\mathcal{M})p(d|\mathcal{M}), \tag{18}$$

where we have dropped an irrelevant normalization constant that depends only on the data and $p(\mathcal{M})$ is the prior probability assigned to the model itself. Usually this is taken to be non–committal and equal to $1/N_m$ if one considers N_m different models. When comparing two models, $\mathcal{M}_0$ versus $\mathcal{M}_1$, one is interested in the ratio of the posterior probabilities, or *posterior odds*, given by

$$\frac{p(\mathcal{M}_0|d)}{p(\mathcal{M}_1|d)} = B_{01}\frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)} \tag{19}$$

and the *Bayes factor* B_{01} is the ratio of the models' evidences:

$$B_{01} \equiv \frac{p(d|\mathcal{M}_0)}{p(d|\mathcal{M}_1)} \quad \text{(Bayes factor)}. \tag{20}$$

A value $B_{01} > (<) 1$ represents an increase (decrease) of the support in favour of model 0 versus model 1 given the observed data. From Eq. (19) it follows that the Bayes factor gives the factor by which the relative odds between the two models have changed after the arrival of the data, regardless of what we thought of the relative plausibility of the models before the data, given by the ratio of the prior models' probabilities. Therefore the relevant quantity to update our state of belief in two competing models is the Bayes factor.

To gain some intuition about how the Bayes factor works, consider two competing models: $\mathcal{M}_0$ predicting that a quantity $\theta = 0$ with no free parameters, and $\mathcal{M}_1$ which assigns θ a Gaussian prior distribution with 0 mean and variance Σ^2. Assume we perform a measurement of θ described by a normal likelihood of standard deviation σ, and with the maximum likelihood value lying λ standard deviations away from 0, i.e. $|\theta_{\max}/\sigma| = \lambda$. Then the Bayes factor between the two models is given by, from Eq. (20)

$$B_{01} = \sqrt{1 + (\sigma/\Sigma)^{-2}} \exp\left(-\frac{\lambda^2}{2(1 + (\sigma/\Sigma)^2)}\right). \tag{21}$$

Table 1. Empirical scale for evaluating the strength of evidence when comparing two models, $\mathcal{M}_0$ versus $\mathcal{M}_1$ (so–called "Jeffreys' scale"). Threshold values are empirically set, and they occur for values of the logarithm of the Bayes factor of $|\ln B_{01}| = 1.0$, 2.5 and 5.0. The right–most column gives our convention for denoting the different levels of evidence above these thresholds. The probability column refers to the posterior probability of the favoured model, assuming non–committal priors on the two competing models, i.e. $p(\mathcal{M}_0) = p(\mathcal{M}_1) = 1/2$ and that the two models exhaust the model space, $p(\mathcal{M}_0|d) + p(\mathcal{M}_1|d) = 1$.

$|\ln B_{01}|$	Odds	Probability	Strength of evidence
< 1.0	$\lesssim 3 : 1$	< 0.750	Inconclusive
1.0	$\sim 3 : 1$	0.750	Weak evidence
2.5	$\sim 12 : 1$	0.923	Moderate evidence
5.0	$\sim 150 : 1$	0.993	Strong evidence

For $\lambda \gg 1$, corresponding to a detection of the new parameter at many sigma, the exponential term dominates and $B_{01} \ll 1$, favouring the more complex model with a non–zero extra parameter, in agreement with the usual conclusion. But if $\lambda \lesssim 1$ and $\sigma/\Sigma \ll 1$ (i.e., the likelihood is much more sharply peaked than the prior and in the vicinity of 0), then the prediction of the simpler model that $\theta = 0$ has been confirmed. This leads to the Bayes factor being dominated by the Occam's razor term, and $B_{01} \approx \Sigma/\sigma$, i.e. evidence accumulates in favour of the simpler model proportionally to the volume of "wasted" parameter space. If however $\sigma/\Sigma \gg 1$ then the likelihood is less informative than the prior and $B_{01} \to 1$, i.e. the data have not changed our relative belief in the two models.

Bayes factors are usually interpreted against the Jeffreys' scale [3] for the strength of evidence, given in Table 1. This is an empirically calibrated scale, with thresholds at values of the odds of about $3 : 1$, $12 : 1$ and $150 : 1$, representing weak, moderate and strong evidence, respectively. A useful way of thinking of the Jeffreys' scale is in terms of betting odds — many of us would feel that odds of $150 : 1$ are a fairly strong disincentive towards betting a large sum of money on the outcome. Also notice from Table 1 that the relevant quantity in the scale is the logarithm of the Bayes factor, which tells us that evidence only accumulates slowly and that indeed moving up a level in the evidence strength scale requires about an order of magnitude more support than the level before.

Bayesian model comparison *does not* replace the parameter inference step (which is performed within each of the models separately). Instead, model comparison *extends* the assessment of hypotheses in the light of the available data to the space of theoretical models, as evident from Eq. (19), which is the equivalent expression for models to Eq. (12), representing inference about the parameters value within each model (for multi–model inference, merging the two levels, see section 6.2).

4.3 *Computation and interpretation of the evidence*

The computation of the Bayesian evidence (17) is in general a numerically challenging task, as it involves a multi–dimensional integration over the whole of parameter space. An added difficulty is that the likelihood is often sharply peaked within the prior range, but possibly with long tails that do contribute significantly to the integral and which cannot be neglected. Other problematic situations arise when the likelihood is multi–modal, or when it has strong degeneracies that confine the posterior to thin sheets in parameter space. Until recently, the application of Bayesian model comparison has been hampered by the difficulty of reliably estimating the evidence. Fortunately, several methods are now available, each with its own strengths and domains of applicability.

(i) The numerical method of choice until recently has been thermodynamic integration, also called *simulated annealing* (see e.g. [11,43,44] and references therein for details). Its computational cost can become fairly large, as it depends heavily on the dimensionality of the parameter space and on the characteristic of the likelihood function. In typical cosmological applications [45–47], thermodynamic integration can require up to 10^7 likelihood evaluations, two orders of magnitude more than MCMC–based parameter estimation.

(ii) Skilling [48,49] has put forward an elegant algorithm called "nested sampling", which has been implemented in the cosmological context by [50–54] (for a theoretical discussion of the algorithmic properties, see [55]). The gist of nested sampling is that the multi–dimensional evidence integral is recast into a one–dimensional integral that is easy to evaluate numerically. This technique allows to reduce the computational burden to about 10^5 likelihood evaluations[1]. Recently, the development of what is called "multi–modal nested sampling" has allowed to increase significantly the efficiency of the method [53], reducing the number of likelihood evaluations by another order of magnitude.

(iii) Approximations to the Bayes factor, Eq. (20), are available for situations in which the models being compared are *nested* into each other, i.e. the more complex model ($\mathcal{M}_1$) reduces to the original model ($\mathcal{M}_0$) for specific values of the new parameters. This is a fairly common scenario in cosmology, where one wishes to evaluate whether the inclusion of the new parameters is supported by the data. For example, we might want to assess whether we need isocurvature contributions to the initial conditions for cosmological perturbations, or whether a curvature term in Einstein's equation is needed, or whether a non–scale invariant distribution of the primordial fluctuation is preferred (see Table 4 for actual results). Writing for the extended model parameters $\theta = (\phi, \psi)$, where the simpler model $\mathcal{M}_0$ is obtained by setting $\psi = 0$, and assuming further that the prior is separable (which is usually the case in cosmology), i.e. that

$$p(\phi, \psi | \mathcal{M}_1) = p(\psi | \mathcal{M}_1) p(\phi | \mathcal{M}_0), \tag{22}$$

the Bayes factor can be written in all generality as

$$B_{01} = \left. \frac{p(\psi | d, \mathcal{M}_1)}{p(\psi | \mathcal{M}_1)} \right|_{\psi = 0}. \tag{23}$$

This expression is known as the Savage–Dickey density ratio (SDDR, see [56] and references therein). The numerator is simply the marginal posterior under the more complex model evaluated at the simpler model's parameter value, while the denominator is the prior density of the more complex model evaluated at the same point. This technique is particularly useful when testing for one extra parameter at the time, because then the marginal posterior $p(\psi | d, \mathcal{M}_1)$ is a 1–dimensional function and normalizing it to unity probability content only requires a 1–dimensional integral, which is simple to do using for example the trapezoidal rule.

(iv) An instructive approximation to the Bayesian evidence can be obtained when the likelihood function is unimodal and approximately Gaussian in the parameters. Expanding the likelihood around its peak to second order one obtains the Laplace approximation

$$p(d | \theta, \mathcal{M}) \approx \mathcal{L}_{\max} \exp \left[-\frac{1}{2} (\theta - \theta_{\max})^t L (\theta - \theta_{\max}) \right], \tag{24}$$

where $\theta_{\max}$ is the maximum–likelihood point, $\mathcal{L}_{\max}$ the maximum likelihood value and L the likelihood Fisher matrix (which is the inverse of the covariance matrix for the parameters). Assuming as a prior a multinormal Gaussian distribution with zero mean and Fisher information matrix P one obtains for the evidence, Eq. (17)

$$p(d | \mathcal{M}) = \mathcal{L}_{\max} \frac{|F|^{-1/2}}{|P|^{-1/2}} \exp \left[-\frac{1}{2} (\theta_{\max}{}^t L \theta_{\max} - \overline{\theta}^t F \overline{\theta}) \right], \tag{25}$$

where the posterior Fisher matrix is $F = L + P$ and the posterior mean is given by $\overline{\theta} = F^{-1} L \theta_{\max}$.

---

[1]Publicly available modules implementing nested sampling can be found at `cosmonest.org` [51] and `http://www.mrao.cam.ac.uk/software/cosmoclust/` [52] (accessed Feb 2008).
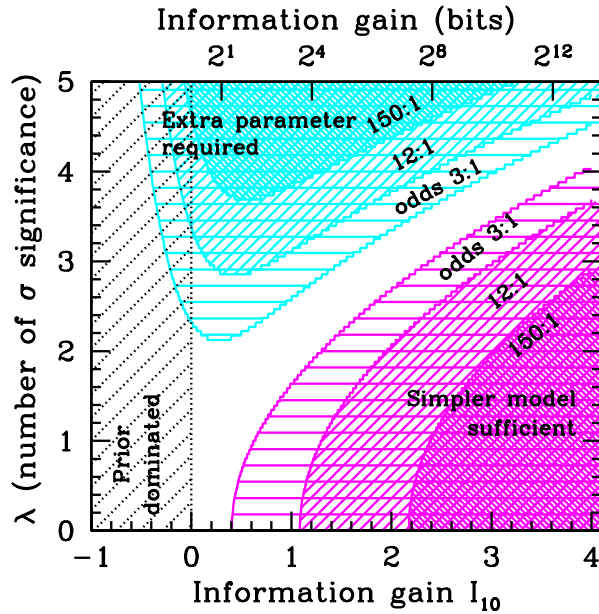
Figure 3. Illustration of Bayesian model comparison for two nested models, where the more complex model has one extra parameter. The outcome of the model comparison depends both on the information content of the data with respect to the *a priori* available parameter space, I_{10} (horizontal axis) and on the quality of fit (vertical axis, λ, which gives the number of sigma significance of the measurement for the extra parameter). The contours are computed from Eq. (23), assuming a Gaussian likelihood and prior (adapted from [58]).

From Eq. (25) we can deduce a few qualitatively relevant properties of the evidence. First, the quality of fit of the model is expressed by $\mathcal{L}_{\max}$, the best–fit likelihood. Thus a model which fits the data better will be favoured by this term. The term involving the determinants of P and F is a volume factor, encoding the Occam's razor effect. As $|P| \leq |F|$, it penalizes models with a large volume of wasted parameter space, i.e. those for which the parameter space volume $|F|^{-1/2}$ which survives after arrival of the data is much smaller than the initially available parameter space under the model prior, $|P|^{-1/2}$. Finally, the exponential term suppresses the likelihood of models for which the parameters values which maximise the likelihood, $\theta_{\max}$, differ appreciably from the expectation value under the posterior, $\overline{\theta}$. Therefore when we consider a model with an increased number of parameters we see that *its evidence will be larger only if the quality–of–fit increases enough to offset the penalizing effect of the Occam's factor* (see also the discussion in [57]).

On the other hand, it is important to notice that the Bayesian evidence does *not* penalizes models with parameters that are unconstrained by the data. It is easy to see that unmeasured parameters (i.e., parameters whose posterior is equal to the prior) do not contribute to the evidence integral, and hence model comparison does not act against them, awaiting better data.

4.4 *The rough guide to model comparison*

The gist of Bayesian model comparison can be summarized by the following, back–of–the–envelope Bayes factor computation for nested models. The result is surprisingly close to what one would obtain from the more elaborate, fully–fledged evidence evaluation, and can serve as a rough guide for the Bayes factor determination.

Returning to the example of Eq. (21), if the data are informative with respect to the prior on the extra parameter (i.e., for $\sigma/\Sigma \ll 1$) the logarithm of the Bayes factor is given approximately by

$$\ln B_{01} \approx \ln\left(\Sigma/\sigma\right) - \lambda^2/2, \tag{26}$$

where as before λ gives the number of sigma away from a null result (the "significance" of the measurement). The first term on the right–hand–side is approximately the logarithm of the ratio of the prior to posterior

volume. We can interpret it as the information content of the data, as it gives the factor by which the parameter space has been reduced in going from the prior to the posterior. This term is positive for informative data, i.e. if the likelihood is more sharply peaked than the prior. The second term is always negative, and it favours the more complex model if the measurement gives a result many sigma away from the prediction of the simpler model (i.e., for $\lambda \gg 0$). We are free to measure the information content in base–10 logarithm (as this quantity is closer to our intuition, being the order of magnitude of our information increase), and we define the quantity $I_{10} \equiv \log_{10}(\Sigma/\sigma)$. Figure 3 shows contours of $|\ln B_{01}| = \text{const}$ for const $= 1.0, 2.5, 5.0$ in the (I_{10}, λ) plane, as computed from Eq. (26). The contours delimit significative levels for the strength of evidence, according to the Jeffreys' scale (Table 1). For moderately informative data ($I_{10} \approx 1-2$) the measured mean has to lie at least about 4σ away from 0 in order to robustly disfavor the simpler model (i.e., $\lambda \gtrsim 4$). Conversely, for $\lambda \lesssim 3$ highly informative data ($I_{10} \gtrsim 2$) do favor the conclusion that the extra parameter is indeed 0. In general, a large information content favors the simpler model, because Occam's razor penalizes the large volume of "wasted" parameter space of the extended model.

An useful properties of Figure 3 is that the impact of a change of prior can be easily quantified. A different choice of prior width (i.e., Σ) amounts to a *horizontal shift* across Figure 3, at least as long as $I_{10} > 0$ (i.e., the posterior is dominated by the likelihood). Picking more restrictive priors (reflecting more predictive theoretical models) corresponds to shifting the result of the model comparison to the left of Figure 3, returning an inconclusive result (white region) or a prior–dominated outcome (hatched region). Notice that results in the 2–3 sigma range, which are fairly typical in cosmology, can only support the more complex model in a very mild way at best (odds of 3 : 1 at best), while actually being most of the time either inconclusive or in favour of the simpler hypothesis (pink shaded region in the bottom right corner).

Bayesian model comparison is usually *conservative* when it comes to admitting a new quantity in our model, even in the case when the prior width is chosen incorrectly. Consider the following two possibilities:

- If the prior range is too small, the model comparison result will be non–committal (white region in Figure 3), or even prior dominated (hatched region, where the posterior is dominated by the prior). Hence in this case we have to hold judgement until better data come along.
- Too wide a prior will instead unduly favour the simpler model (pink, shaded regions). However, as new, better data come along the result will move to the right (for a fixed prior width, as the likelihood becomes narrower) but eventually also upwards, towards a larger number of sigma significance, if the true model really has a non–zero extra parameter. Eventually, our initial "poor" prior choice will be overridden as the number of sigma becomes large enough to take the result into the blue, shaded region.

In both cases the result of the model comparison will eventually override the "wrong" prior choice (although it might take a long time to do so), exactly as it happens for parameter inference.

4.5 *Getting around the prior – The maximal evidence for a new parameter*

For nested models, Eq. (23) shows that the relative probability of the more complex model can be made arbitrarily small by increasing the broadness of the prior for the extra parameters, $p(\psi|\mathcal{M}_1)$ (as the prior is a pdf, it must integrate to unit probability. Hence a broader prior corresponds to a smaller value of $p(\psi|\mathcal{M}_1)_{\psi=0}$ in the denominator). Often, this is not problematical as prior ranges for the new parameters can (and should) be motivated from the underlying theory. For example, in assessing whether the scalar spectral index (n_s) of the primordial perturbations differs from the scale–invariant value $n_s = 1$, the prior range of the index can be constrained to be $0.8 \lesssim n_s \lesssim 1.2$ within the theoretical framework of slow roll inflation (more on this in section 6). The sensitivity of the model comparison result can also be investigated for other plausible, physically motivated choices of prior ranges, see e.g. [59,60]. If the model comparison outcome is qualitatively the same for a broad choice of plausible priors, then we can be confident that the result is robust.

Although the Bayesian evidence offers a well–defined framework for model comparison, there are cases where there is not a specific enough model available to place meaningful limits on the prior ranges of

Table 2. Translation table (using Eq. (27)) between frequentist significance values (p–values) and the upper bounds on the odds ($\overline{B}_{10}$) in favour of the more complex model. No other choice of prior (within the family considered in the text) will give higher evidence in favour of the extra parameters. The "sigma" column is the corresponding number of standard deviations away from the mean for a normal distribution. The "category" column gives the Jeffreys' scale of Table 1 (from [65]).

p–value	$\overline{B}_{10}$	$\ln \overline{B}_{10}$	sigma	category
0.05	2.5	0.9	2.0	
0.04	2.9	1.0	2.1	'weak' at best
0.01	8.0	2.1	2.6	
0.006	12	2.5	2.7	'moderate' at best
0.003	21	3.0	3.0	
0.001	53	4.0	3.3	
0.0003	150	5.0	3.6	'strong' at best
6×10^{-7}	43000	11	5.0	

new parameters in a model. This hurdle arises frequently in cases when the new parameters are a phenomenological description of a new effect, only loosely tied to the underlying physics, such as for example expansion coefficients of some series. An interesting possibility in such a case is to choose the prior on the new parameters in such a way as to *maximise the probability of the new model*, given the data. If, even under this best case scenario, the more complex model is not significantly more probable than the simpler model, then one can confidently say that the data does not support the addition of the new parameters, without worrying that some other choice of prior will make the new model more probable [61–63].

Consider the Bayes factor in favour of the more complex model, $B_{10} \equiv 1/B_{01}$, with B_{01} given by Eq. (20). The simpler model, $\mathcal{M}_0$, is obtained from $\mathcal{M}_0$ by setting $\theta = \theta^*$. An absolute upper bound to the evidence in favour of the more complex model is obtained by choosing $p(\theta|\mathcal{M}_1)$ to be a delta function centered at the maximum likelihood value under $\mathcal{M}_1$, $\theta_{\max}$. It can be shown that in this case the upper bound $\overline{B}_{10}$ corresponds to the likelihood ratio between $\theta_{\max}$ and θ^*. However, it can be argued that such a choice for the prior is unjustified, as it can only be made *ex post facto* after one has seen the data and obtained the maximum likelihood estimate. It is more natural to appeal to a mild principle of indifference as to the value of the parameter under the more complex model, and thus to maximize the evidence over priors that are symmetric about θ^* and unimodal. This can be shown to be equivalent to maximizing over all priors that are uniform and symmetric about θ^*. This procedure leads to a very simple expression for the lower bound on the Bayes factor [61]

$$B_{10} \leq \overline{B}_{10} = \frac{-1}{\mathrm{e}\wp \ln \wp} \tag{27}$$

for $\wp \leq \mathrm{e}^{-1}$, where e is the exponential of one. Here, $\wp$ is the *p–value*, the probability that the the value of some test statistics be as large as or larger than the observed value *assuming the null hypothesis (i.e., the simpler model) is true* (see [61, 62] for a detailed discussion). A more precise definition of p–values is given in any standard statistical textbook, e.g. [64].

Eq. (27) offers a useful calibration of frequentist significance values (p–values) in terms of upper bounds on the Bayesian evidence in favour of the extra parameters. The advantage is that the quantity on the left–hand side of Eq. (27) can be straightforwardly interpreted as an upper bound on the odds for the more complex model, whereas the p–values cannot. This point is illustrated very clearly with an astronomical example in [66]. In fact, a word of caution is in place regarding the meaning of the p–value, which is often *misinterpreted* as an error probability, i.e. as giving the fraction of wrongly rejected nulls in the long run. For example, when a frequentist test rejects the null hypothesis (in our example, that $\theta = \theta^*$) at the 5% level, this *does not* mean that one will make a mistake roughly 5% of the time if one were to repeat the test many times. The actual error probability is much larger than that, and can be shown to be *at least* 29% (for unimodal, symmetric priors, see Table 6 in [62]). This important conceptual point is discussed in in greater detail in [62, 63, 67]. The fundamental reason for this discrepancy with intuition is that frequentist significance tests give the probability of observing data as extreme or more extreme than

what has actually been measured, *assuming* the null hypothesis $\mathcal{H}_0$ to be true (which in Bayesian terms amounts to the choice of a model, $\mathcal{M}_0$). But the quantity one is interested in is actually the probability of the model $\mathcal{M}_0$ given the observations, which can only be obtained by using Bayes' Theorem to invert the order of conditioning[1]. Indeed, in frequentist statistics, a hypothesis is either true or false (although we do not know which case it is) and it is meaningless to attach to it a probability statement.

Table 2 lists $\overline{B}_{10}$ for some common thresholds for significance values and the strength of evidence scale, thus giving a conversion table between significance values and upper bounds on the Bayesian evidence, *independent of the choice of prior for the extra parameter* (within the class of unimodal and symmetric priors). It is apparent that in general the upper bound on the Bayesian evidence is much more conservative than the p–value, e.g. a 99% result (corresponding to $\wp = 0.01$) corresponds to odds of $8 : 1$ at best in favour of the extra parameters, which fall short of even the "moderate evidence" threshold. Strong evidence at best requires at least a 3.6 sigma result. A useful rule of thumb is thus to think of a s sigma result as a $s - 1$ sigma result, e.g. a 99.7% result (3 sigma) really corresponds to odds of $21 : 1$, i.e. about 95% probability for the more complex model. Thus when considering the detection of a new parameter, instead of reporting frequentist significance values it is more appropriate to present the upper bound on the Bayes factor, as this represents the maximum probability that the extra parameter is different from its value under the simpler model.

This approach has been applied to the cosmological context in [65], who analysed the evidence in favour of a non–scale invariant spectral index and of asymmetry in the cosmic microwave background maps. For further details on the comparison between frequentist hypothesis testing and Bayesian model selection, see [28].

4.6 *The effective number of parameters – Bayesian model complexity*

The usefulness of a Bayesian model selection approach based on the Bayesian evidence is that it tells us whether the increased "complexity" of a model with more parameters is justified by the data. However, it is desirable to have a more refined definition of "model complexity", as the number of free parameters is not always an adequate description of this concept. For example, if we are trying to measure a periodic signal in a time series, we might have a model of the data that looks like

$$f(t) = A(1 + \theta \cos(t + \delta)), \tag{28}$$

where A, θ, δ are free parameters we wish to constrain. But if θ is very small compared to 1 and the noise is large compared to θ, then the oscillatory term remains unconstrained by the data and effectively we can only measure the normalization A. Thus the parameters θ, δ should not count as free parameters as they cannot be constrained given the data we have, and the effective model complexity is closer to 1 than to 3. From this example it follows that the very notion of "free parameter" is not absolute, but it depends on both what our expectations are under the model, i.e. on the prior, and on the constraining power of the data at hand.

In order to define a more appropriate measure of complexity, in [68] the notion of *Bayesian complexity* was introduced, which measures the number of parameters that the data can support. Consider the information gain obtained when upgrading the prior to the posterior, as measured by the the Kullback–Leibler (KL) divergence [69] between the posterior, p and the prior, denoted here by π:

$$D_{\mathrm{KL}}(p, \pi) \equiv \int p(\theta|d, \mathcal{M}) \ln \frac{p(\theta|d, \mathcal{M})}{\pi(\theta|\mathcal{M})} \mathrm{d}\theta. \tag{29}$$

In virtue of Bayes' theorem, $p(\theta|d, \mathcal{M}) = \mathcal{L}(\theta)\pi(\theta|\mathcal{M})/p(d|\mathcal{M})$ hence the KL divergence becomes the sum

---

[1]To convince oneself of the difference between the two quantities, consider the following example [22]. Imagine selecting a person at random — the person can either be male or female (our hypothesis). If the person is female, her probability of being pregnant (our data) is about 3%, i.e. $p(\text{pregnant}|\text{female}) = 0.03$. However, if the person is pregnant, her probability of being female is much larger than that, i.e. $(\text{female}|\text{pregnant}) \gg 0.03$. The two conditional probabilities are related by Bayes' theorem.

of the negative log evidence and the expectation value of the log–likelihood under the posterior:

$$D_{\mathrm{KL}}(p, \pi) = -\ln p(d|\mathcal{M}) + \int p(\theta|d, \mathcal{M}) \ln \mathcal{L}(\theta) \mathrm{d}\theta. \tag{30}$$

To gain a feeling for what the KL divergence expresses, let us compute it for a 1–dimensional case, with a Gaussian prior around 0 of variance Σ^2 and a Gaussian likelihood centered on $\theta_{\max}$ and variance σ^2. We obtain after a short calculation

$$D_{\mathrm{KL}}(p, \pi) = -\frac{1}{2} - \ln \frac{\sigma}{\Sigma} + \frac{1}{2} \left[\left(\frac{\sigma}{\Sigma} \right)^2 \left(\frac{\theta_{\max}^2}{\sigma^2} - 1 \right) \right]. \tag{31}$$

The second term on the right–hand side gives the reduction in parameter space volume in going from the prior to the posterior. For informative data, $\sigma/\Sigma \ll 1$, this terms is positive and grows as the logarithm of the volume ratio. On the other hand, in the same regime the third term is small unless the maximum likelihood estimate is many standard deviations away from what we expected under the prior, i.e. for $\theta_{\max}/\sigma \gg 1$. This means that the maximum likelihood value is "surprising", in that it is far from what our prior led us to expect. Therefore we can see that the KL divergence is a summary of the amount of information, or "surprise", contained in the data.

Let us now define an effective χ^2 through the likelihood as $\mathcal{L}(\theta) = \exp(-\chi^2/2)$. Then Eq. (30) gives

$$D_{\mathrm{KL}}(p, \pi) = -\frac{1}{2}\overline{\chi^2(\theta)} + \ln p(d|\mathcal{M}), \tag{32}$$

where the bar indicates a mean taken over the posterior distribution. The posterior average of the effective chi–square is a quantity can be easily obtained by Markov chain Monte Carlo techniques (see section 3.2). We then subtract from the "expected surprise" the estimated surprise in the data after we have actually fitted the model parameters, denoted by

$$\widehat{D_{\mathrm{KL}}} \equiv -\frac{1}{2}\chi^2(\hat{\theta}) + \ln p(d|\mathcal{M}), \tag{33}$$

where the first term on the right–hand–side is the effective chi–square at the estimated value of the parameters, indicated by a hat. This will usually be the posterior mean of the parameters, but other possible estimators are the maximum likelihood point or the posterior median, depending on the details of the problem. We then define the quantity

$$\mathcal{C}_b \equiv -2\left(D_{\mathrm{KL}}(p, \pi) - \widehat{D_{\mathrm{KL}}} \right) = \overline{\chi^2(\theta)} - \chi^2(\hat{\theta}) \quad \text{(Bayesian complexity)}, \tag{34}$$

(notice that the evidence term is the same in Eqs. (32) and (33) as it does not depend on the parameters and therefore it disappears from the complexity). The Bayesian complexity gives the effective number of parameters as *a measure of the constraining power of the data as compared to the predictivity of the model*, i.e. the prior. Hence $\mathcal{C}_b$ depends both on the data and on the prior available parameter space. This can be understood by considering further the toy example of a Gaussian likelihood of variance σ^2 around $\theta_{\max}$ and a Gaussian prior around 0 of variance Σ^2. Then a short calculation shows that the Bayesian complexity is given by (see [70] for details)

$$\mathcal{C}_b = \frac{1}{1 + (\sigma/\Sigma)^2}. \tag{35}$$

So for $\sigma/\Sigma \ll 1$, $\mathcal{C}_b \approx 1$ and the model has one effective, well constrained parameter. But if the likelihood width is large compared to the prior, $\sigma/\Sigma \gg 1$, then the experiment is not informative with respect to our prior beliefs and $\mathcal{C}_b \to 0$.

The Bayesian complexity can be a useful diagnostic tool in the tricky situation where the evidence for two competing models is about the same. Since the evidence does not penalize parameters that are unmeasured, from the evidence alone we cannot know if we are in the situation where the extra parameters are simply unconstrained and hence irrelevant ($\theta \ll 1$ in the example of Eq. (28)) or if they improve the quality–of–fit just enough to offset the Occam's razor penalty term, hence giving the same evidence as the simpler model. The Bayesian complexity breaks this degeneracy in the evidence allowing to distinguish between the two cases:

(i) $p(d|\mathcal{M}_0) \approx p(d|\mathcal{M}_1)$ and $\mathcal{C}_b(\mathcal{M}_1) > \mathcal{C}_b(\mathcal{M}_0)$: the quality of the data is sufficient to measure the additional parameters of the more complicated model ($\mathcal{M}_1$), but they do not improve its evidence by much. We should prefer model $\mathcal{M}_0$, with less parameters.

(ii) $p(d|\mathcal{M}_0) \approx p(d|\mathcal{M}_1)$ and $\mathcal{C}_b(\mathcal{M}_1) \approx \mathcal{C}_b(\mathcal{M}_0)$: both models have a comparable evidence and the effective number of parameters is about the same. In this case the data is not good enough to measure the additional parameters of the more complicated model (given the choice of prior) and we cannot draw any conclusions as to whether the extra parameter is needed.

The first application of this technique in the cosmological context is in [70], where it is applied to the number of effective parameters from cosmic microwave background data, while in [71] it was used to determine the number of effective dark energy parameters.

4.7 *Information criteria for approximate model comparison*

Sometimes it might be useful to employ methods that aim at an approximate model selection under some simplifying assumptions that give a default penalty term for more complex models, which replaces the Occam's razor term coming from the different prior volumes in the Bayesian evidence. While this is an obviously appealing feature, on closer examination it has the drawback of being meaningful only in fairly specific cases, which are not always met in astrophysical or cosmological applications. In particular, it can be argued that the Bayesian evidence (ideally coupled with an analysis of the Bayesian complexity) ought to be preferred for model building since it is precisely the lack of predictivity of more complicated models, as embodied in the physically motivated range of the prior, that ought to penalize them.

With this *caveat* in mind, we list below three types of information criteria that have been widely used in several astrophysical and cosmological contexts. An introduction to the information criteria geared for astrophysicists is given by Ref. [72]. A discussion of the differences between the different information criteria as applied to astrophysics can be found in [73], which also presents a few other information criteria not discussed here.

- **Akaike Information Criterion (AIC):** Introduced by Akaike [74], the AIC is an essentially frequentist criterion that sets the penalty term equal to twice the number of free parameters in the model, k:

$$\text{AIC} \equiv -2 \ln \mathcal{L}_{\max} + 2k \qquad (36)$$

where $\mathcal{L}_{\max} \equiv p(d|\theta_{\max}, \mathcal{M})$ is the maximum likelihood value. The derivation of the AIC follows from an approximate minimization of the KL divergence between the true model distribution and the distribution being fitted to the data.

- **Bayesian Information Criterion (BIC):** Sometimes called "Schwarz Information Criterion" (from the name of its proposer [75]), the BIC follows from a Gaussian approximation to the Bayesian evidence in the limit of large sample size:

$$\text{BIC} \equiv -2 \ln \mathcal{L}_{\max} + k \ln N \qquad (37)$$

where k is the number of fitted parameters as before and N is the number of data points. The best model is again the one that minimizes the BIC.

- **Deviance Information Criterion (DIC):** Introduced by [68], the DIC can be written as

$$\mathrm{DIC} \equiv -2\widehat{D_{\mathrm{KL}}} + 2\mathcal{C}_b. \tag{38}$$

In this form, the DIC is reminiscent of the AIC, with the $\ln\mathcal{L}_{\max}$ term replaced by the estimated KL divergence and the number of free parameters by the effective number of parameters, $\mathcal{C}_b$, from Eq. (34). Indeed, in the limit of well–constrained parameters, the AIC is recovered from (38), but the DIC has the advantage of accounting for unconstrained directions in parameters space.

The information criteria ought to be interpreted with care when applied to real situations. Comparison of Eq. (37) with Eq. (36) shows that for $N > 7$ the BIC penalizes models with more free parameters more harshly than the AIC. Furthermore, both criteria penalize extra parameters regardless of whether they are constrained by the data or not, unlike the Bayesian evidence. This comes about because implicitly both criteria assume a "data dominated" regime, where all free parameters are well constrained. But in general the number of free parameters might not be a good representation of the actual number of effective parameters, as discussed in section 4.6. In a Bayesian sense it therefore appears desirable to replace the number of parameters k by the effective number of parameters as measured by the Bayesian complexity, as in the DIC.

It is instructive to inspect briefly the derivation of the BIC. The unnormalized posterior $g(\theta) \equiv \mathcal{L}(\theta)p(\theta|\mathcal{M})$ can be approximated by a multi–variate Gaussian around its mode $\overline{\theta}$, i.e. $g(\theta) \approx g(\overline{\theta}) - 1/2(\theta - \overline{\theta})^t F(\theta - \overline{\theta})$, where F is minus the Hessian of the posterior evaluated at the posterior mode. Then the evidence integral can be computed analytically, giving

$$p(d|\mathcal{M}) \approx \exp(g(\overline{\theta}))(2\pi)^{k/2}|F|^{-1/2}. \tag{39}$$

For large samples, $N \to \infty$, the posterior mode tends to the maximum likelihood point, $\overline{\theta} \to \theta_{\max}$, hence $g(\overline{\theta}) \to \mathcal{L}_{\max}p(\theta_{\max})$ and the log–evidence becomes

$$\ln p(d|\mathcal{M}) \to \ln\mathcal{L}_{\max} + \ln p(\theta_{\max}) + \frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln|F| \quad (N \to \infty), \tag{40}$$

where the error introduced by the various approximations scales to leading order as $\mathcal{O}(N^{-1/2})$. On the right–hand–side, the first term scales as $\mathcal{O}(N)$, the second and third terms are constant in N, while the last term is given by $\ln|F| \approx k\ln N$, since the variance scales as the number of data points, N. Dropping terms of order $\mathcal{O}(1)$ or below[1], we obtain

$$\ln p(d|\mathcal{M}) \to \ln\mathcal{L}_{\max} - \frac{k}{2}\ln N \quad (N \to \infty). \tag{41}$$

Thus maximising the evidence of Eq. (41) is equivalent to minimizing the BIC in Eq. (37). We see that dropping the term $\ln p(\theta_{\max})$ in (40) means that effectively we expect to be in a regime where the model comparison is dominated by the likelihood, and that the prior Occam's razor effect becomes negligible. This is often not the case in cosmology. Furthermore, this "weak" prior choice is intrinsic (even though hidden from the user) in the form of the BIC, and often it is not justified. In conclusion, it appears that what makes the information criteria attractive, namely the absence of an explicit prior specification, represents also their intrinsic limitation.

---

[1]See [76] for a more careful treatment, where a better approximation is obtained by assuming a weak prior which contains the same information as a single datum.

5 Cosmological parameter inference

Driven by the emergence of inexpensive sensors and computing capabilities, the amount of cosmological data has been increasing exponentially over the last 15 years or so. For example, the first map of cosmic microwave background (CMB) anisotropies obtained in 1992 by COBE [77] contained $\sim 10^3$ pixels, which became $\sim 5 \times 10^4$ by 2002 with CBI [78,79]. Current state–of–the art maps (from the WMAP satellite [80]) involve $\sim 10^6$ pixels, which are set to grow to $\sim 10^7$ with Planck in the next couple of years. Similarly, angular galaxy surveys contained $\sim 10^6$ objects in the 1970's, while by 2005 the Sloan Digital Sky Survey [81] had measured $\sim 2 \times 10^8$ objects, which will increase to $\sim 3 \times 10^9$ by 2012 when the Large Synoptic Survey Telescope [82] comes online[2].

This data explosion drove the adoption of more efficient map making tools, faster component separation algorithms and parameter inference methods that would scale more favourably with the number of dimensions of the problem. As data sets have become larger and more precise, so has grown the complexity of the models being used to describe them. For example, if only 2 parameters could meaningfully be extracted from the COBE measurement of the large–scale CMB temperature power spectrum (namely the normalization and the spectral tilt [77, 83]), the number of model parameters had grown to 11 by 2002, when smaller–scale measurements of the acoustic peaks had become available. Nowadays, parameter spaces of up to 20 dimensions are routinely considered.

This section gives an introduction to the broad problem of cosmological parameter inference and highlights some of the tools that have been introduced to tackle it, with particular emphasis on innovative techniques. This is a vast field and any summary is bound to be only sketchy. We give throughout references to selected papers covering both historically important milestones and recent major developments.

5.1 *The "vanilla" ΛCDM cosmological model*

Before discussing the quantities we are interested in measuring in cosmology (the "cosmological parameters") and some of the observational probes available to do so, we briefly sketch the general framework which goes under the name of "cosmological concordance model". Because it is a relatively simple scenario containing both a cosmological constant (Λ) and cold dark matter (CDM) (more about them below), it is also known as the "vanilla" ΛCDM model.

Our current cosmological picture is based on the scenario of an expanding Universe, as implied by the observed redshift of the spectra of distant galaxies (Hubble's law). This in turn means that the Universe began from a hot and dense state, the initial singularity of the Big Bang. The existence of the cosmic microwave background lends strong support to this idea, as it is interpreted as the relic radiation from the hotter and denser primordial era. The expanding spacetime is described by Einstein's general relativity. The cosmological principle states that the Universe is isotropic (i.e., the same in all directions) and homogeneous (the same everywhere). If follows that an isotropically expanding Universe is described by the so–called Friedmann–Robertson–Walker metric,

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \frac{a^2(t)}{c^2}\left[\frac{\mathrm{d}r^2}{1+\kappa r^2} + r^2(\mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2)\right] \tag{42}$$

where κ defines the geometry of spatial sections (if κ is positive, the geometry is spherical; if it is zero, the geometry is flat; if it is negative, the geometry is hyperbolic). The *scale factor* $a(t)$ describes the expansion of the Universe, and it is related to redshift z by

$$1 + z = \frac{a(t_0)}{a(t)}, \tag{43}$$

---

[2]Alex Szalay, talk at the specialist discussion meeting "Statistical challenges in cosmology and astroparticle physics", held at the Royal Astronomical Society, London, Oct 2007.

where $a(t_0)$ is the scale factor today and $a(t)$ the scale factor at redshift z. The relation between redshift and comoving distance r is obtained from the above metric via the Friedmann equation, and is given by

$$a_0 \mathrm{d}r = \frac{c}{H_0} \left[\Omega_\kappa (1+z)^2 + \Omega_\Lambda + (\Omega_b + \Omega_{\mathrm{cdm}})(1+z)^3 + (\Omega_\gamma + \Omega_\nu)(1+z)^4 \right]^{-1/2} \mathrm{d}z, \qquad (44)$$

where H_0 is the *Hubble constant*, and Ω_x are the *cosmological parameters* describing the matter–energy content of the Universe. Standard parameters included in the vanilla model are neutrinos (Ω_ν, with a mass $\lesssim 1$ eV), photons (Ω_γ), baryons (Ω_b), cold dark matter (Ω_{cdm}) and a cosmological constant (Ω_Λ). The curvature term Ω_κ is included for completeness but is currently not required by the standard cosmological model (see section 6). The comoving distance determines the apparent brightness of objects, their apparent size on the sky and the number density of objects in a comoving volume. Hence measurements of the brightness of standard candles, of the length of standard rulers or of the number density of objects at a given redshift leads to the determination of the cosmological parameters in Eq. (44) (see next section).

The currently accepted paradigm for the generation of density fluctuations in the early Universe is inflation. The idea is that quantum fluctuations in the primordial era were stretched to cosmological scales by an initial period of exponential expansion, called "inflation", possibly driven by a yet unknown scalar field. This increased the scale factor by about 26 orders of magnitude within about 10^{-32}s after the Big Bang. Although presently we have only indirect evidence for inflation, it is commonly accepted that such a short burst of exponential growth in the scale factor is required to solve the horizon problem, i.e. to explain why the CMB is so highly homogeneous across the whole sky. The quantum fluctuations also originated temperature anisotropies in the CMB, whose study has proved to be one of the powerhouses of precision cosmology. From the initial state with small perturbations imprinted on a broadly uniform background, gravitational attraction generated the complex structures we see in the modern Universe, as indicated both by observational evidence and highly sophisticated computer modelling.

Of course it is possible to consider completely different models, based for example on alternative theories of gravity (such as Bekenstein's theory [84, 85] or Jordan–Brans–Dicke theory [86]), or on a different way of comparing model predictions with observations [87–89]. Discriminating among models and determining which model is in best agreement with the data is a task for model comparison techniques, whose application to cosmology is discussed section 6. Here we will take the vanilla ΛCDM model as our starting point for the following considerations on cosmological parameters and how they are measured.

5.2 *Cosmological observations*

The discovery of temperature fluctuations in the Cosmic Microwave Background (CMB) in 1992 by the COBE satellite [77] marked the beginning of the era of precision cosmology. Many other observations have contributed to the impressive development of the field in less than 20 years. For example, around 1990 the picture of flat Universe with both cold dark matter and a positive cosmological constant was only beginning to emerge, and only thanks to the painstakingly difficult work of gluing together several fairly indirect clues [90]. At the time of writing (January 2008), the total density is known with an error of order 1%, and it is likely that this uncertainty will be reduced by another two orders of magnitude in the mid–term [91]. The high accuracy of modern precision cosmology rests on the combination of several different probes, that constrain the physical properties of the Universe at many different redshifts.

(i) **Cosmic microwave background (CMB):** CMB anisotropies offer a snapshot of the Universe at the time of recombination, about 380,000 years after the Big Bang, at a redshift $z \approx 1100$. As described above, the temperature differences measured in the CMB arise from quantum fluctuations during the inflationary phase. Their usefulness lies in the fact that they are small ($\Delta T/T \sim 10^{-5}$) and hence linear perturbation theory is mostly sufficient to predict very accurately their statistical distribution. The 2–point correlation function of the anisotropies is usually described via its Fourier transform, the *angular power spectrum*, which presents a series of characteristic peaks called *acoustic oscillations*, see e.g. [92, 93]. Their structure depends in a rich way on the cosmological parameters introduced in Eq. (44), as well as on the initial conditions for the perturbations emerging from the inflationary era (see e.g. [94–96]
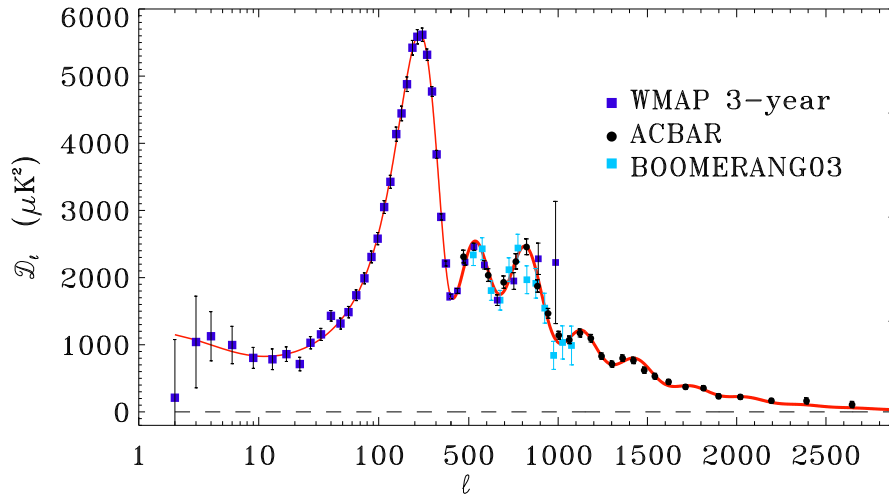
Figure 4. State–of–the-art cosmic microwave background temperature power spectrum measurements along with the best–fit
ΛCDM model (solid line), showing data from WMAP 3–yr [80], the Boomerang 2003 flight [101] and ACBAR [97] (from [97]).

for further details). The anisotropies are polarized at the level of 1%, and measuring accurately the
information encoded by the weak polarization signal is the goal of many ongoing observations. State–
of–the art measurements are described in e.g. [80, 97–101]. An example of recent measurements of the
temperature power spectrum is shown in Figure 4. Later this year, the Planck satellite is expected to
start full–sky, high–resolution observations of both temperature and polarization.

(ii) **Large scale structures (LSS):** the correlation function among galaxies gives an estimate of the
correlation properties of the underlying dark matter distribution, up to a bias factor relating the dark
matter to the baryon distribution. Current data typically extend out to $z \sim 0.7$. The resulting power
spectrum (recent data are shown in Figure 5) depends mainly on the ratio of the radiation to matter
energy density, on the initial distribution of the perturbations with scale (spectral index) and on the
overall normalization (which can be extracted once a bias model is specified). Heavy numerical simula-
tion is nowadays used to model accurately small scales, where non–linear effects become dominant. The
tool of choice to measure the power spectrum on small scales is becoming the observations of absorp-
tion lines from neutral hydrogen clouds, the so-called *Lyman-α forest* [102], although concerns remain
about the reliability of the theoretical modelling of non–linear effects. Recently, both the Sloan Digital
Sky Survey [103] and the 2dF Galaxy Redshift Survey [104] have detected the presence of *baryonic
acoustic oscillations*, which appear as a bump in the galaxy–galaxy correlation function corresponding
to the scale of the acoustic oscillations in the CMB. The physical meaning is that galaxies tend to form
preferentially at a separation corresponding to the characteristic scale of inhomogeneities in the CMB.
Baryonic oscillations can be used as rulers of known length (measured via the CMB acoustic peaks)
located at a much smaller redshift than the CMB (currently, $z \sim 0.3$), and hence they are powerful
probes of the recent expansion history of the Universe, with particular focus on dark energy properties.
The distribution of clusters with redshift can also be employed to probe the growth of perturbations
and hence to constrain cosmology. Current galaxy redshift surveys have catalogued about half a million
objects, but a new generation of surveys aims at taking this number to a over a billion.

(iii) **Supernovae type Ia:** the commonly accepted scenario for the formation of supernovae type Ia is
a white dwarf accreting material from a binary companion. The core heats up as the gravitational
pressure increases, eventually leading to carbon fusion ignition, followed by oxygen burning. This
runaway reaction releases within seconds a huge amount of energy, resulting in a violent explosion which
is accompanied by a massive surge in luminosity. Observationally, type Ia supernovae are characterized
by the absence of hydrogen lines in their spectrum and they are considered almost standard candles, in
the sense that there is a strong empirical correlation between their duration and their peak luminosity.
From measurements of their brightness as a function of time (the light curve), their intrinsic luminosity
can be reconstructed. The data are then used to reconstruct the redshift–distance relationship, i.e. the
Hubble diagram, which in turn depends on the cosmological parameters [106–109]. Current data extend
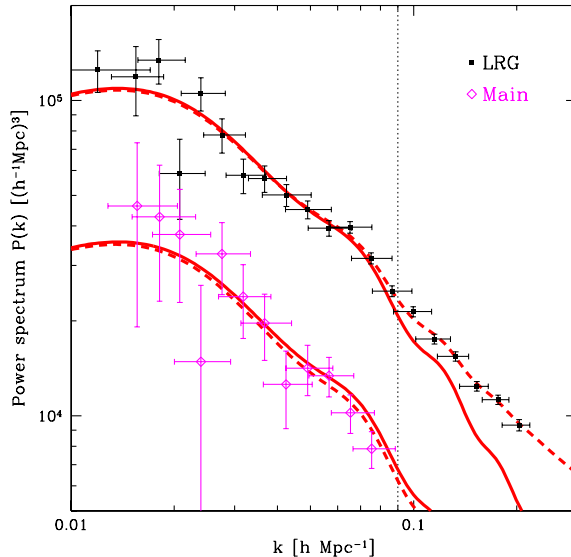
Figure 5. Current matter power spectrum measurements from the Sloan Digital Sky Survey and best–fit ΛCDM power spectrum (solid/dashed lines, without/with non–linear corrections), corresponding to the cosmological parameters extracted from WMAP 3–yr CMB data (the top and bottom curves are for two different samples). This shows that even without fitting to matter power spectrum data, the best–fit CMB model is in good agreement with the galaxy distribution data (from [105]).

out to about $z \sim 1.4$ and encompass a few hundreds of supernovae (a number which will increase by a factor of 10 thanks to planned searches). Supernovae data were the first line of evidence in 1998 [110,111] that the expansion of the Universe is accelerating – an effect attributed to the existence of dark energy (see e.g. [112]).

(iv) **Weak gravitational lensing:** the presence of inhomogeneities in the distribution of matter along the line of sight distorts the shape of background galaxies due to the deflection of light rays. This is most spectacularly displayed in observations of strong lensing, showing the characteristic arc–shaped multiple images of background galaxies. However, the same physics affects to a much smaller degree the shape of any background galaxy, distorting it by about 0.1 to 1%. This is called *weak gravitational lensing*. Although it is impossible to measure such small distortions for any single galaxy, the effect can still be detected statistically, by correlating the shear pattern (distortion due to gravitational lensing) of several thousand galaxies. The resulting weak lensing power spectrum probes a combination of the geometrical setup (distance to the background sources and to the lens) and of the parameters controlling the growth of structures, in particular the amount of matter (both visible and dark) and the strength of the perturbations (for a recent review, see [113] Some recent observations are reported in [114–117]). By dividing the source galaxies into slices of different redshift, it is possible to carry out a sort of "cosmic tomography", reconstructing the dark matter distribution between us and the sources [118]. Although it has not yet reached the same level of precision of the CMB, weak lensing shows great promise for the future in constraining cosmological parameters and in particular dark energy.

5.3 *Constraining cosmological parameters*

As outlined is section 3.1, our inference problem is fully specified once we give the model (which parameters are allowed to vary and their prior distribution) and the likelihood function for the data sets under consideration. For the cosmological observations described above, relevant cosmological parameters can be broadly classified in four categories.

(i) *Parameters describing the dynamics of the background evolution:* they represent the matter–energy content of the Universe and its expansion history, relating redshift with comoving distance, see Eq. (44). The Hubble constant today is written as $H_0 = 100h$ km/s/Mpc, and is used to define the *critical energy density*, i.e. the energy density needed to make the Universe spatially flat: $\rho_{\mathrm{crit}} = 1.88 \times 10^{-29} h^2$

g/cm^3. The remaining density parameters (Ω_x in Eq. (44)) are then written in units of the critical energy density, so that for example the energy density in baryons is given by $\rho_b = \Omega_b \rho_{\mathrm{crit}}$. Standard parameters include the energy density in photons (Ω_γ), neutrinos (Ω_ν), baryons (Ω_b), cold dark matter (Ω_{cdm}), cosmological constant (Ω_Λ) or, more generally, a possibly time–dependent dark energy (Ω_{de}).

(ii) *Parameters describing the initial conditions for the fluctuations:* they give the type of initial conditions, adiabatic (where the spatial distribution of fluctuations is the same for all fluids emerging from inflation, up to multiplicative factors) or isocurvature (where there is a mismatch between perturbations among two components). The most general type of initial conditions is described by a correlation matrix that contains 10 free parameters representing the excitation amplitude of each mode [119–121]. The simplest parameterization of the distribution of perturbations with scale is then given for each mode in terms of a spectral index.

(iii) *Nuisance parameters:* these often relate to insufficiently constrained aspects of the physics of the observed objects, or to uncertainties in the measuring process. We are not interested in determining them, but accounting for their uncertainty is important in order to obtain a correct estimate of the error on the physical parameters we are seeking to determine. If the observable quantity has a strong dependence on poorly determined nuisance parameters, then simply fixing the nuisance parameters instead of marginalizing over them will lead to serious underestimation of the uncertainty for the remaining parameters. Example of nuisance parameters are the bias factor in galaxy surveys, residual beam calibration uncertainty for CMB data, supernovae intrinsic evolution parameters, intrinsic ellipticity of background galaxies in weak lensing and others.

(iv) *Parameters describing new physics:* this is where the exciting frontier of data analysis lies, and we are trying to constrain or detect effects arising from new physics in the model, such as time–variation of the fine structure constant, the presence of cosmic strings, the mass of neutrinos, non–trivial topology of the Universe, extra dimensions, time–variations of dark energy properties, and much more. Although often framed as a parameter inference problem, this is actually a model comparison question, and is therefore best tackled with the methods described in section 4. Therefore, in this case the parameter inference step is only the first level towards working out the outcome of the higher level model comparison step.

5.3.1 *The joint likelihood function.* When the observations are independent, the log–likelihoods for each observation simply add[1]. Defining $\chi^2 \equiv -2 \ln \mathcal{L}$, we have that

$$\chi^2_{\mathrm{tot}} = \chi^2_{\mathrm{CMB}} + \chi^2_{\mathrm{SN}} + \chi^2_{\mathrm{lens}} + \chi^2_{\mathrm{LSS}} + \ldots \tag{45}$$

One important advantage of combining different observations as in Eq. (45) is that each observable has different *degenerate directions*, i.e. directions in parameter space that are poorly constrained by the data. By combining two or more types of observables, it is often the case that the two data sets together have a much stronger constraining power than each one of them separately, because they mutually break parameters degeneracies. Combination of data sets should never be carried out blindly, however. The danger is that the data sets might be mutually inconsistent, in which case combining them singles out in the posterior a region that is not favoured by any of the two data sets separately, which is obviously unsatisfactory. Such discrepancies might arise because of undetected systematics, or insufficient modelling of the observations.

In order to account for possible discrepancies of this kind, Ref. [122] suggested to replace Eq. (45) by

$$\chi^2_{\mathrm{tot}} = \sum_i \alpha_i \chi^2_i \tag{46}$$

where α_i are (unknown) weight factors ("hyperparameters") for the various data sets, which determine the relative importance of the observations. A non–informative prior is specified for the hyperparameters,

---

[1]This is of course not the case when one is carrying out correlation studies, where the aim is precisely to exploit correlation among *different* observables (for example, the late Integrated Sachs–Wolfe effect).

which are then integrated out in a Bayesian way, obtaining an effective chi–square

$$\chi^2_{\text{eff}} = \sum_i N_i \ln \chi^2_i, \qquad (47)$$

where N_i is the number of data points in data set i. This method has been applied to combine different CMB observations in the pre–WMAP era [122,123]. A technique based on the comparison of the Bayesian evidence for different data sets has been employed in [124], while Ref. [125] uses a technique similar in spirit to the hyperparameter approach outlined above to perform a binning of mutually inconsistent observations suffering from undetected systematics, as explained in [126].

After the likelihood has been specified, it remains to work out the posterior pdf, usually obtained numerically via MCMC technology, and report posterior constraints on the model parameters, e.g. by plotting 1 or 2–dimensional posterior contours. We now sketch the way this program has been carried out as far as cosmological parameter estimation is concerned.

5.3.2   *Likelihood–based parameter determination.* Up until around 2002, the method of choice for cosmological parameter estimation was either direct numerical maximum likelihood search [127, 128] or evaluation of the likelihood on a grid in parameter space. Once the likelihood has been mapped out, (frequentist) confidence intervals for the parameters are obtained by finding the maximum–likelihood point (or, equivalently, the minimum chi–square) and by delimiting the region of parameter space where the log–likelihood drops by a specified amount (details can be found in any standard statistics textbook). If the likelihood is a multi–normal Gaussian, then this procedure leads to the familiar "delta chi–square" rule–of–thumb, i.e. that e.g. a 95.4% (2σ) confidence interval for 1 parameter is delimited by the region where the $\chi^2 \equiv -2 \log \mathcal{L}$ increases by $\Delta\chi^2 = 4.00$ from its minimum value (see e.g. [43]). Of course the value of $\Delta\chi^2$ depends both on the number of parameters being constrained and on the desired confidence interval[1]

Approximate confidence intervals for each parameter were then usually obtained from the above procedure, after maximising the likelihood across the hidden dimensions rather than marginalising [130,131], since the latter procedure required a computationally expensive multi–dimensional integration. The rationale was that maximisation is approximately equivalent to marginalisation for close–to–Gaussian problems (a simple proof can be found in Appendix A of [132]), although it was early recognized that this is not always a good approximation for real–life situations [133]. Marginalisation methods based on multi–dimensional interpolation were devised and applied in order to improve on this respect [132,134]. Many studies adopted this methodology, which could not quite be described as fully Bayesian yet since it was likelihood–based and the the notion of posterior was not explicitly introduced. Often, the choice of particular theoretical scenarios (for example, a flat Universe or adiabatic initial conditions) or the inclusion of external constraints (such as bounds on the baryonic density coming from Big Bang nucleosyhntesis) were described as "priors". A more rigorous terminology would call the former a model choice, while the latter amounts to inclusion (in the likelihood) of external information. Since the likelihood could be well approximated by a simple log–normal distribution, its computation cost was fairly low. With the advent of `CMBFAST` [135], the availability of a fast numerical code for the computation of CMB and matter power spectra meant that grids of up to 30 million points and parameter spaces of dimensionality up to order 10 could be handled in this way [134].

---

[1]An important technical point is that frequentist confidence intervals are considered random variables — they give the range within which our estimate of the parameter value will fall e.g. 95.4% of the time if we repeat our measurement $N \rightarrow \infty$ times. The true value of the parameter is given (although unknown to us) and has no probability statement attached to it. On the contrary, Bayesian *credible intervals* containing for example 95.4% of the posterior probability mass represent our degree of belief about the value of the parameter itself. Often, Bayesian credible intervals are imprecisely called "confidence intervals" (a term that should be reserved for the frequentist quantity), thus fostering confusion between the two. Perhaps this happens because for Gaussian cases the two results are formally identical, although their interpretation is profoundly different. This can have important consequences if the true value is near the boundary of the parameter space, in which case the results from the frequentist and Bayesian procedure may differ substantially – see [129] for an interesting example involving the determination of neutrino masses.

Table 3. State–of–the art cosmological parameter inference from WMAP 3–year CMB data [80] and Sloan Digital Sky Survey data [105]. Posterior median and 68% posterior region, obtained for flat priors on the parameter set in the top section, with the exception of the reionization optical depth τ, for which a flat prior has been adopted on $\exp(-2\tau)$ instead (adapted from [105]).

Parameter	Value	Meaning	Definition
Matter budget parameters			
Θ_s	$0.5918^{+0.0020}_{-0.0020}$	CMB acoustic angular scale fit (degrees)	$\Theta_s = r_s(z_{\rm rec})/d_A(z_{\rm rec}) \times 180/\pi$
ω_b	$0.0222^{+0.0007}_{-0.0007}$	Baryon density	$\omega_b = \Omega_b h^2 \approx \rho_b/(1.88 \times 10^{-26}{\rm kg/m}^3)$
ω_c	$0.1050^{+0.0041}_{-0.0040}$	Cold dark matter density	$\omega_c = \Omega_{\rm cdm} h^2 \approx \rho_c/(1.88 \times 10^{-26}{\rm kg/m}^3)$
Initial conditions parameters			
A_s	$0.690^{+0.045}_{-0.044}$	Scalar fluctuation amplitude	Primordial scalar power at $k = 0.05/{\rm Mpc}$
n_s	$0.953^{+0.016}_{-0.016}$	Scalar spectral index	Primordial spectral index at $k = 0.05/{\rm Mpc}$
Reionization history (abrupt reionization)			
τ	$0.087^{+0.028}_{-0.030}$	Reionization optical depth	
Nuisance parameters (for galaxy power spectrum)			
b	$1.896^{+0.074}_{-0.069}$	Galaxy bias factor	See [105] for details.
$Q_{\rm nl}$	$30.3^{+4.4}_{-4.1}$	Nonlinear correction parameter	See [105] for details.
Derived parameters (functions of those above)			
$\Omega_{\rm tot}$	1.00 (flat Universe assumed)	Total density/critical density	$\Omega_{\rm tot} = \Omega_m + \Omega_\Lambda = 1 - \Omega_\kappa$
h	$0.730^{+0.019}_{-0.019}$	Hubble parameter	$h = \sqrt{(\omega_b + \omega_c)/(\Omega_{\rm tot} - \Omega_\Lambda)}$
Ω_b	$0.0416^{+0.0019}_{-0.0018}$	Baryon density/critical density	$\Omega_b = \omega_b/h^2$
Ω_c	$0.197^{+0.016}_{-0.015}$	CDM density/critical density	$\Omega_{\rm cdm} = \omega_c/h^2$
Ω_m	$0.239^{+0.018}_{-0.017}$	Matter density/critical density	$\Omega_m = \Omega_b + \Omega_{\rm cdm}$
Ω_Λ	$0.761^{+0.017}_{-0.018}$	Cosmological constant density/critical density	$\Omega_\Lambda \approx h^{-2}\rho_\Lambda(1.88 \times 10^{-26}{\rm kg/m}^3)$
σ_8	$0.756^{+0.035}_{-0.035}$	Density fluctuation amplitude	See [105] for details.

5.3.3 *Bayes in the sky — The rise of MCMC.* The watershed moment after which methods based on likelihood evaluation on a grid where definitely overcome by Bayesian MCMC methods can perhaps be indicated in Ref. [136], which marked one of the last major studies performed using essentially frequentist techniques. Pioneering works in using MCMC technology for cosmological parameter extraction include the application to VSA data [45, 137], the use of simulated annealing [138] and the study of Ref. [139]. But it was the release of the `CosmoMC` code[1] in 2002 [140] that made a huge impact on the cosmological community, as `CosmoMC` quickly became a standard and user–friendly tool for Bayesian parameter inference from CMB, large scale structure and other data. The favourable scalability of MCMC methods with dimensionality of the parameter space and the easiness of marginalization were immediately recognized as major advantages of the method. `CosmoMC` employs the `CAMB` code [141] to compute the matter and CMB power spectra from the physical model parameters. It then employs various MCMC algorithms to sample from the posterior distribution given current CMB, matter power spectrum (galaxy power spectrum, baryonic acoustic wiggles and Lyman–α observations) and supernovae data.

State–of–the–art applications of cosmological parameter inference can be found in papers such as [97, 105, 142, 143]. Table 3 summarizes recent posterior credible intervals on the parameters of the vanilla ΛCDM model introduced above while Figure 6 shows the full 1–D posterior pdf for 6 relevant parameters (both from Ref. [105]). The initial conditions emerging from inflation are well described by one adiabatic degree of freedom and a distribution of fluctuations that deviates slightly from scale invariance, but which is otherwise fairly featureless.

Addition of extra parameters to this basic description (for example, a curvature term, or a time–evolution of the cosmological constant, in which case it is generically called "dark energy") is best discussed in terms of model comparison rather than parameter inference (see next section). The breath and range of studies aiming at constraining extra parameters is such that it would be impossible to give even a rough sketch here. However we can say that the simple model described by the 6 cosmological parameters given in Table 3 appears at the moment appropriate and sufficient to explain most of the presently available data. MCMC is nowadays almost universally employed in one form or another in the cosmology community.

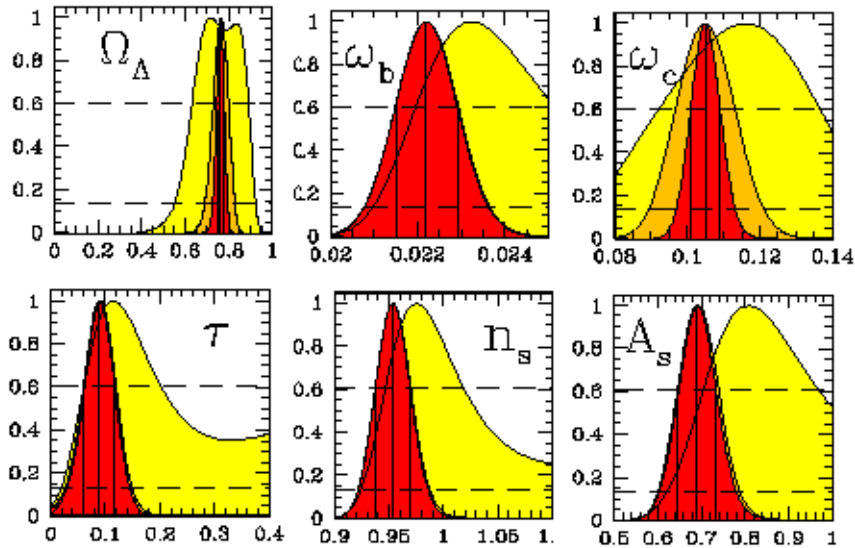[1]Available from: `cosmologist.info/cosmomc` (accessed Jan 20th 2008).

Figure 6. Posterior constraints on key cosmological parameters from recent CMB and large scale structure data, compare Table 3. Top row, from left to right, posterior pdf (normalized to the peak) for the cosmological constant density in units of the critical density, the (physical) baryons and cold dark matter densities. Bottom row, from left to right: optical depth to reionization, scalar tilt and scalar fluctuations amplitude. Yellow using WMAP 1–yr data, orange WMAP 3–yr data and red adding Sloan Digital Sky Survey galaxy distribution data. Spatial flatness and adiabatic initial conditions have been assumed. This set of only 6 parameters (plus 2 other nuisance parameters not shown here) appear currently sufficient to describe most cosmological observations (adapted from [105]).

5.3.4   *Recent developments in parameter inference.* Nowadays, the likelihood evaluation step is becoming the bottleneck of cosmological parameter inference, as the WMAP likelihood code [80] requires the evaluation of fairly complex correlation terms, while the computational time for the actual model prediction in terms of power spectra is subdominant (except for spatially curved models or non–standard scenarios containing active seeds, such as cosmic strings). This trend is likely to become stronger with future CMB datasets, such as Planck. Currently, for relatively straightforward extensions of the concordance model presented in Table 3, a Markov chain with enough samples to give good inference can be obtained in a couple of days of computing time on a "off–the–shelf" machine with 4 CPUs.

Massive savings in computational effort can be obtained by employing neural networks techniques, a computational methodology consistent of a network of processors called "neurons". Neural networks learn in an unsupervised fashion the structure of the computation to be performed (for example, likelihood evaluation across the cosmological parameter space) from a training set provided by the user. Once trained, the network becomes a fast and efficient interpolation algorithm for new points in the parameter space, or for classification purposes (for example to determine the redshift of galaxies from photometric data [144,145]). When applied to the problem of cosmological parameter inference, neural networks can teach themselves the structure of the parameter space (for models up to about 10 dimensions) by employing as little as a few thousands training points, for which the likelihood has to be computed numerically as usual. Once trained, the network can then interpolate extremely fast between samples to deliver a complete Markov chain within a few minutes. The speed–up can thus reach a factor of 30 or more [146,147]. Another promising tool is a machine–learning based algorithm called PICO [148, 149][1], requiring a training set of order $\sim 10^4$ samples, which are then used to perform an interpolation of the likelihood across parameter space. This procedure can achieve a speed–up of over a factor of 1000 with respect to conventional MCMC.

The forefront of Bayesian parameter extraction is quickly moving on to tackle even more ambitious problems, involving thousands of parameters. This is the case for the Gibbs sampling technique to extract the full posterior probability distribution for the power spectrum C_ℓ's directly from CMB maps, performing component separations (i.e., multi–frequency foregrounds removal) at the same time and fully propagating uncertainties to the level of cosmological parameters [150–153]. This method has been tested up to $\ell \lesssim 50$ on WMAP temperature data and is expected to perform well up to $\ell \lesssim 100 - 200$ for Planck–quality data.

---

[1] Available from: http://cosmos.astro.uiuc.edu/pico/ (accessed Jan 14th 2008).

Equally impressive is the Hamiltonian sampling approach [154], which returns the C_ℓ's posterior pdf from a (previously foreground subtracted) CMB map. At WMAP resolution, this involves working with $\sim 10^5$ parameters, but the efficiency is such that the 800 C_ℓ distributions (for the temperature signal) can be obtained in about a day of computation on a high–end desktop computer.

Another frontier of Bayesian methods is represented by high energy particle physics, which for historical and methodological reasons has been so far mostly dominated by frequentist techniques. However, the Bayesian approach to parameter inference for supersymmetric theories is rapidly gathering momentum, due to its superior handling of nuisance parameters, marginalization and inclusion of all sources of uncertainty. The use of Bayesian MCMC for supersymmetry parameter extraction has been first advocated in [155], and has then been rigorously applied to a detailed analysis of the Constrained Minimal Supersymmetric Standard Model [156–161], a problem that involves order 10 free parameters. A public code called `SuperBayeS` is available to perform an MCMC Bayesian analysis of supersymmetric models[1]. Presently, there are hints that the constraining power of the data is insufficient to override the prior choice in this context [162], but future observations, most notably by the Tevatron or LHC [163, 164] and tighter limits (or a detection) on the neutralino scattering cross section [159], should considerably improve the situation in this respect.

5.4 *Caveats and common pitfalls*

Although Bayesian inference is quickly maturing to become an almost automated procedure, we should not forget that a "black box" approach to the problem always hides dangers and pitfalls. Every real world problem presents its own peculiarities that demand careful consideration and statistical inference remains very much a craft as much as a science. While Bayes' Theorem is never wrong, incorrect specification of the prior (for example, making unwarranted assumptions of failing to specify relevant external information) or inappropriate construction of the likelihood (e.g., not reflecting the experiment or neglecting relevant sources of uncertainty) will easily lead to wrong inferences. Below we list a few common pitfalls that need to be considered in the cosmological context.

(i) **Hidden prior information.** Sometimes the choice of flat priors on the parameters is uncritically taken to be uninformative. This is often not the case. For instance, a flat prior is indeed non–informative if we are estimating the mean of a Gaussian, but if we are interested in its standard deviation σ a prior which is flat in $\ln \sigma$ ("Jeffreys' prior") is instead the appropriate choice (see e.g. [6]). When considering extensions of the ΛCDM model, for example, it is common practice to parameterize the new physics with a set of quantities about which very little (if anything at all) is known *a priori*. Assuming flat priors on those parameters is often an unwarranted choice. Flat priors in "fundamental" parameters will in general not correspond to a flat distribution for the observable quantities that are derived from those parameters, nor for other quantities we might be interested in constraining. Hence the apparently non–informative choice for the fundamental parameters is actually highly informative for other quantities that appear directly in the likelihood. It is extremely important that this "hidden prior information" be brought to light, or one could mistake the effect of the prior for constraining power of the data. An effective way to do that is to plot pdf's for the quantities of interest without including the data, i.e. run an MCMC sampling from the prior only (see [45] for an instructive example).

(ii) **No well–defined prior measure.** Another difficulty is that in many cases several physically equally plausible parameterizations exist, in particular for problems involving unknown amplitudes, such as for example isocurvature modes in the initial conditions. Since a flat prior on parameterization A is not flat in parameterization B if the two are related by a non–linear transformation, see Eq. (11), two physically equivalent setups might lead to widely different inferences. In the absence of theoretical or physical reasons to prefer parameterization A or B this leads to the unsatisfactory dependence of the posterior on the volume enclosed by the prior, as discussed for the problem of general initial conditions in [120]. For an example of such a "prior volume effect" applied to inflationary models parameters, see [165]. Other

---

[1]Code available from: `superbayes.org` (accessed Feb 13th 2008).

obvious domains where this effect might be problematic are dark energy equation of state reconstruction (e.g., [166, 167]), initial power spectrum reconstruction [168–170] and determination of inflationary potential parameters [171–173]. In some special cases, fundamental principles can be invoked to define the appropriate prior measure, which exploits either symmetries or invariance properties of the problem. An important example is image reconstruction, where the Maximum Entropy principle is employed to define an informative prior on the image space (see e.g. [174–176] for an astronomical application).

(iii) **Lindley's paradox.** A methodological issue is the widespread use of inappropriate tools to answer what is actually a model selection question. Often we want to assess the "significance" of an effect, when a deviation is observed in the posterior from the parameter value that would correspond to the absence of that effect. This situation is extremely common across very different domains, from estimation of photon number counts in presence of a background (see [18, 19, 177] for a proper Bayesian treatment), to the assessment of the anomalies in the large–scale CMB power spectrum ([178] compare frequentist methods with the Bayesian evidence), the detection of gravitational waves ([179] introduce a Bayesian technique that is automatically optimal) or the discovery of extra–solar planets [180]. In general, the "number of sigma" away from the expected value in the absence of a signal is *not* a good indicator of the significance of the effect, a result that goes under the name of "Lindley's paradox" [181]. For further details, see Appendix A in [58].

(iv) **Using the data twice.** Often the data are used to guide in a qualitative fashion the model building or the choice of priors. If the same data are then used for a quantitative comparison the significance of the effect can be drastically overestimated. This is a well–known problem in frequentist statistics, which therefore insists that one should design one's statistical tests before seeing the data at all. In cosmology this might often be problematic or impossible to achieve, as new observations will uncover completely unexpected phenomena (for example, the large–scale anomalies in the cosmic microwave background). It is however important to keep in mind that the significance of such effects is difficult to assess.

6 Cosmological Bayesian model building

The Bayesian model comparison approach based on the evaluation of the evidence is being increasingly applied to model building questions such as: are isocurvature contributions to the initial conditions required by the data [46, 58, 60, 182]? Is the Universe flat [58, 70, 183]? What is the best description of the primordial power spectrum for density perturbations [47, 51, 58, 70, 165, 184, 185]? Is dark energy best described as a cosmological constant [50, 51, 186–191]? In this section we review the status of the field.

6.1 *Evidence for the cosmological concordance model*

Table 4 is a fairly extensive compilation of recent results regarding possible extensions to (or reduction of) the vanilla ΛCDM concordance cosmological model introduced in section 5.1. We have chosen to compile only results obtained using the full Bayesian evidence, rather than approximate model comparisons obtained via the information criteria because the latter are often *not* adequate approximations, for the reasons explained in section 4.7. Of course, the outcome depends on the Occam's razor effect brought about by the prior volume (and sometimes, by the choice of parameterization). Where applicable, we have show the sensitivity of the result on the prior assumptions by giving a ballpark range of values for the Bayes factor, as presented in the original studies. The reader ought to refer to the original works for the precise prior and parameter choices and for the justification of the assumed prior ranges.

As anticipated, the 6 parameters ΛCDM concordance model is currently well supported by the data, as the inclusion of extra parameters is not required by the Bayesian evidence. This is shown by the fact that most model comparisons return either an undecided result or they support the ΛCDM model (negative values for $\ln B$ in Table 4). The only exception is the support for a cut–off on large scales in the power spectrum reported by [185]. This is clearly driven by the anomalies in the large scale CMB power spectrum, which in this case are interpreted as being a reflection of a lack of power in the primordial power

Table 4. Summary of model comparison results against the ΛCDM concordance model (see Table 3) using Bayesian model comparison for nested models. A negative (positive) value for $\ln B$ indicates that the competing model is disfavoured (supported) with respect to the ΛCDM model. The column ΔN_{par} gives the difference in the number of free parameters with respect to the ΛCDM concordance model. A negative value means that one of the parameters has been fixed. See references for full details and in particular for the choice of priors on the model parameters, which control the strengths of the Occam's razor effect.

Competing model	ΔN_{par}	$\ln B$	Ref	Data	Outcome
Initial conditions					
Isocurvature modes					
CDM isocurvature	+1	−7.6	[58]	WMAP3+, LSS	Strong evidence for adiabaticity
+ arbitrary correlations	+4	−1.0	[46]	WMAP1+, LSS, SN Ia	Undecided
Neutrino entropy	+1	$[-2.5, -6.5]^p$	[60]	WMAP3+, LSS	Moderate to strong evidence for adiabaticity
+ arbitrary correlations	+4	−1.0	[46]	WMAP1+, LSS, SN Ia	Undecided
Neutrino velocity	+1	$[-2.5, -6.5]^p$	[60]	WMAP3+, LSS	Moderate to strong evidence for adiabaticity
+ arbitrary correlations	+4	−1.0	[46]	WMAP1+, LSS, SN Ia	Undecided
Primordial power spectrum					
No tilt ($n_s = 1$)	−1	+0.4	[47]	WMAP1+, LSS	Undecided
		$[-1.1, -0.6]^p$	[51]	WMAP1+, LSS	Undecided
		−0.7	[58]	WMAP1+, LSS	Undecided
		−0.9	[70]	WMAP1+	Undecided
		$[-0.7, -1.7]^{p,d}$	[185]	WMAP3+	$n_s = 1$ weakly disfavoured
		−2.0	[184]	WMAP3+, LSS	$n_s = 1$ weakly disfavoured
		−2.6	[70]	WMAP3+	$n_s = 1$ moderately disfavoured
		−2.9	[58]	WMAP3+, LSS	$n_s = 1$ moderately disfavoured
		$< -3.9^c$	[65]	WMAP3+, LSS	Moderate evidence at best against $n_s \neq 1$
Running	+1	$[-0.6, 1.0]^{p,d}$	[185]	WMAP3+, LSS	No evidence for running
		$< 0.2^c$	[165]	WMAP3+, LSS	Running not required
Running of running	+2	$< 0.4^c$	[165]	WMAP3+, LSS	Not required
Large scales cut–off	+2	$[1.3, 2.2]^{p,d}$	[185]	WMAP3+, LSS	Weak support for a cut–off
Matter–energy content					
Non–flat Universe	+1	−3.8	[70]	WMAP3+, HST	Flat Universe moderately favoured
		−3.4	[58]	WMAP3+, LSS, HST	Flat Universe moderately favoured
Coupled neutrinos	+1	−0.7	[192]	WMAP3+, LSS	No evidence for non–SM neutrinos
Dark energy sector					
$w(z) = w_{\mathrm{eff}} \neq -1$	+1	$[-1.3, -2.7]^p$	[186]	SN Ia	Weak to moderate support for Λ
		−3.0	[50]	SN Ia	Moderate support for Λ
		−1.1	[51]	WMAP1+, LSS, SN Ia	Weak support for Λ
		$[-0.2, -1]^p$	[187]	SN Ia, BAO, WMAP3	Undecided
		$[-1.6, -2.3]^d$	[188]	SN Ia, GRB	Weak support for Λ
$w(z) = w_0 + w_1 z$	+2	$[-1.5, -3.4]^p$	[186]	SN Ia	Weak to moderate support for Λ
		−6.0	[50]	SN Ia	Strong support for Λ
		−1.8	[187]	SN Ia, BAO, WMAP3	Weak support for Λ
$w(z) = w_0 + w_a(1 - a)$	+2	−1.1	[187]	SN Ia, BAO, WMAP3	Weak support for Λ
		$[-1.2, -2.6]^d$	[188]	SN Ia, GRB	Weak to moderate support for Λ
Reionization history					
No reionization ($\tau = 0$)	−1	−2.6	[70]	WMAP3+, HST	$\tau \neq 0$ moderately favoured
No reionization and no tilt	−2	−10.3	[70]	WMAP3+, HST	Strongly disfavoured

^{d} Depending on the choice of datasets.

^{p} Depending on the choice of priors.

^{c} Upper bound using Bayesian calibrated p–values, see section 4.5.

Data sets: WMAP1+ (WMAP3+): WMAP 1st year (3–yr) data and other CMB measurements. LSS: Large scale structures data. SN Ia: supernovae type Ia. BAO: baryonic acoustic oscillations. GRB: gamma ray bursts.

spectrum. Whether such anomalies are of cosmological origin remains however an open question [193, 194]. If extensions of the model are not supported, reduction of ΛCDM to simpler models is not viable, either: recent studies employing WMAP 3–yr data find that a scale invariant spectrum with no spectral tilt is now weakly to moderately disfavoured [58, 65, 70, 184]. Also, a Universe with no reionization is no longer a good description of CMB data, and a non–zero optical depth τ is indeed required [70].

A few further comments about the results reported in Table 4 are in place:

(i) Regarding the type of initial conditions for cosmological perturbations, all parameter extraction studies to date (with the exception of [195]) find that a purely adiabatic mode is in agreement with observations, and constrain the isocurvature fraction to be below about 10% for one single isocurvature mode at

the time [60] and below about 50% for a general mixture of modes [196, 197]. From a model selection perspective, this means that we expect the purely adiabatic model to be preferred over a more complex model with a mixture of isocurvature modes. This is indeed the case, but the result is strongly dependent on the parameterization adopted for the isocurvature sector, which determines the strength of the Occam's razor effect. This is a consequence of the difficulty of coming up with a well motivated phenomenological parameterization of the isocurvature amplitudes, see the discussion in [58, 182].

(ii) The shape of the primordial power spectrum has attracted considerable attention from a model comparison perspective [47, 51, 58, 65, 70, 165, 184, 185]. With the exception of the large scale cut–off mentioned above, the current consensus appears to be that a power–law distribution of fluctuations, with power spectrum $P(k) = P_0(k/k_0)^{n_s-1}$ with $n_s < 1$ is currently the best description. This is usually interpreted as evidence for inflation. However, a proper model comparison of inflationary predictions involves including the presence of tensor modes generated by gravitational waves, parameterized in terms of their amplitude parameter r. Including this extra parameter runs into the difficulty of specifying its prior volume, as the two obvious choices of priors flat in r or $\log r$ lead to very different model comparison results [184]. Another problem is that the comparison might be ill–defined, as the simpler model with $n_s = 1$ and $r = 0$ is presumably some sort of alternative, unspecified model without inflation that would not solve the horizon problem. On this ground alone, unless an alternative solution to the horizon problem is put on the table, such an alternative model would be immediately thrown out (see the discussion in [165]). Finally, higher–order terms in the Taylor expansion of the power spectrum, such as a running of the spectral index or a running of the running, are currently not required. This appears a robust result with respect to a wide choice of priors and data sets.

(iii) Present–day constraints on the curvature of spatial sections are of order $|\Omega_\kappa| \lesssim 0.01$ (with $\Omega_\kappa = 0$ corresponding to a flat, Euclidean geometry), stemming from a combination of CMB, large scale structures and supernovae data. Choosing a phenomenological prior of width $\Delta\Omega_\kappa = 1$ around 0 delivers a moderate support for a flat Universe versus curved models [58, 70]. However, adopting an inflation–motivated prior instead, $\Delta\Omega_\kappa \sim 10^{-5}$, would lead to an undecided result ($\ln B = 0$) for the model comparison, as the data are not strong enough to discriminate between the two models in this case. This can be formalized by considering the Bayesian model complexity for the two choices of priors, Eq. (35). Noticing that σ/Σ is the ratio between the likelihood and prior widths, for a prior on the curvature parameter of width 1, $\sigma/\Sigma \sim 10^{-2}$ and $\mathcal{C}_b \approx 1$, hence the parameter has been measured. But if we take a prior width $\sim 10^{-5}$, $\sigma/\Sigma \sim 10^3$ hence $\mathcal{C}_b \to 0$. In the latter case, we can see from Eq. (21) that the Bayes factor between the two models $B_{01} \to 1$ and the evidence is inconclusive, awaiting better data.

(iv) Model comparisons regarding the dark energy sector suffer from considerably uncertainty. Clearly, the model to beat is the cosmological constant (with equation of state parameter $w = -1$ at all redshifts), but alternative dark energy scenarios suffer from the fundamental difficulty of motivating physically both the parameterization of the dark energy time dependence and the prior volume for the extra parameters [198] (see [199] for a review of models and [200] for on overview of recent constraints). However, the semi–phenomenological studies shown in Table 4 do agree in deeming a cosmological constant a sufficient description of the data. This is again a consequence of the fact that no time evolution of the equation of state is detected in the data, hence the strength of the support in favour of the cosmological constant becomes a function of the available parameter space under the more complex, alternative models. Given this result, it is interesting to ask what level of accuracy is required before our degree of belief in the cosmological constant is overwhelmingly larger than for an evolving dark energy, assuming of course that future data will not detect any significant departure from $w = -1$. To this end, a simple classification of models has been given in [201] in terms of their effective equation of space parameter, w_{eff}, representing the time–varying equation of state averaged over redshift with the appropriate weighting factor for the observable [202]. The three categories considered are "phantom models" (exhibiting large, negative values for the equation of state, $-11 \leq w_{\mathrm{eff}} \leq -1$), "fluid–like dark energy" ($-1 \leq w_{\mathrm{eff}} \leq -1/3$) and "small–departures from Λ" models ($-1.01 \leq w_{\mathrm{eff}} \leq 0.99$). Assuming a flat prior on these ranges of values for w_{eff}, consideration of the Bayes factor between each of those models and the cosmological constant shows that gathering strong evidence against each of the models

Table 5. Summary of model comparison results against the ΛCDM concordance model for some alternative (i.e., non–nested) cosmological models. A negative (positive) value for $\ln B$ indicates that the competing model is disfavoured (supported) with respect to ΛCDM. The column N_{par} gives the number of free parameters in the alternative model. See references for full details about the models, priors and data used.

Competing model	N_{par}	$\ln B$	Ref	Data	Outcome
Alternatives to FRW					
Bianchi VII$_h$	5 to 8	$[-0.9, 1.2]^{d,p}$	[54]	WMAP1, WMAP3	Weak support (at best) for Bianchi template
	5 to 6	$[-0.1, -1.2]^{p}$	[203]	WMAP3	No evidence after texture correction
LTB models	4	-3.6	[204]	WMAP3, BAO, SN Ia	Moderate evidence against LTB
Fractal bubble model	2	0.3	[89]	SN Ia	Undecided
Asymmetry in the CMB					
Anomalous dipole	3	1.8	[205]	WMAP3	Weak evidence for anomalous dipole
		$< 2.2^{c}$	[65]	WMAP3	Weak evidence at best

^{d} Depending on the choice of datasets.

^{p} Depending on the choice of priors.

^{c} Upper bound using Bayesian calibrated p–values, see section 4.5.

requires an accuracy on w_{eff} of order $\sigma_{\text{eff}} = 0.05$ for phantom models (which are therefore already under pressure from current data, which have an accuracy of order ~ 0.1), $\sigma_{\text{eff}} = 3 \times 10^{-3}$ for fluid–like models (about a factor of 5 better than optimistic constraints from future observations) and $\sigma_{\text{eff}} = 5 \times 10^{-5}$ for small–departure models. Refinements of this approach that employ more fundamentally–motivated priors could lead to an analysis of the expected costs/benefits from future dark energy observations in terms of their likely model selection outcome (we return on this issue in section 6.2).

Let us now turn to models that are not nested within ΛCDM— i.e., alternative theoretical scenarios. Table 5 gives some examples of the outcome of the Bayesian model comparison with the concordance model. As above, we restrict our considerations to studies employing the full Bayesian evidence (there are many other examples in the literature carrying out approximate model comparison using information criteria instead). The model comparison is often more difficult for non–nested models, as priors must be specified for all of the parameters in the alternative model (and in the ΛCDM model, as well), in order to compute the evidence ratio. The usual *caveats* on prior choice apply in this case. From Table 5 it appears that the data do not seem to require fundamental changes in our underlying theoretical model, either in the form of Bianchi templates representing a violation of cosmic isotropy (see also [206]), or as Lemaitre–Tolman–Bondi models or fractal bubble scenarios with dressed cosmological parameters. The anomalous dipole in the CMB temperature maps is a fine example of Lindley's paradox. When fitting a dipolar template to the CMB maps, the effective chi–square improves by 9 to 11 units (depending on the details of the analysis) for 3 extra parameters [205, 207], which would be deemed a "significant" effect using a standard goodness–of–fit test. However, the Bayesian evidence analysis shows that the odds in favour of an anomalous dipole are 9 to 1 *at best* (corresponding to $\ln B < 2.2$), which does not reach the "moderate evidence at best" threshold. Hence Bayesian model comparison is conservative, requiring a stronger evidence before deeming an effect to be favoured.

6.2 *Other uses of the Bayesian evidence*

Beside cosmological model building, the Bayesian evidence can be employed in many other different ways. Here we presents two aspects that are relevant to our topic, namely the applications to the field of *multi–model inference* and *model selection forecasting*.

(i) **Multi–model inference.** Once we realize that there are several possible models for our data, it becomes interesting to present parameter inferences that take into account the model uncertainty associated with this plurality of possibilities. In other words, instead of just constraining parameters within each model, we can take a step further and produce parameter inferences that are *averaged over the models being considered*. Let us suppose that we have a minimal model (in our case, ΛCDM) and a series of augmented models with extra parameters. A typical example from cosmology is dark

energy (first discussed in the context of multi–model averaging in [208]), where the minimal model has $w = -1$ fixed and there are several other candidate models with a time–varying equation of state, parameterized in terms of a number of free parameters and their priors. Let us denote by θ the cosmological parameters common to all models. For the extended models, the redshift–dependence of the dark energy equation of state is described by a vector of parameters ω_i (under model $\mathcal{M}_i$). The ΛCDM model has no free parameters for the equation of state, hence the prior on $\omega_{\Lambda\text{CDM}}$ is a delta function centered on $w(z) = w_0 = -1$. Then a straightforward application of Bayes' theorem leads to the following posterior distribution for the parameters:

$$p(\theta, \omega | d) = \sum_i p(\mathcal{M}_i | d) p(\theta, \omega | d, \mathcal{M}_i), \tag{48}$$

where $p(\theta, \omega | d, \mathcal{M}_i)$ is the posterior within each model $\mathcal{M}_i$, and it is understood that the posterior has non–zero support only along the parameter directions $\omega_i \subset \omega$ that are relevant for the model, and delta–functions along all other directions. Each term is weighted by the corresponding posterior model probability,

$$p(\mathcal{M}_i | d) = \frac{p(d | \mathcal{M}_i) p(\mathcal{M}_i)}{\sum_i p(d | \mathcal{M}_i) p(\mathcal{M}_i)}. \tag{49}$$

The prior model probabilities $p(\mathcal{M}_i)$ are usually set equal, but a model preference can be incorporated here if necessary. The *model averaged* posterior distribution of Eq. (48) then represents the parameter constraints obtained independently of the model choice, which has been marginalized over. Unless one of the models is overwhelmingly more probable than the others (in which case the model averaging essentially disappears, as all of the weights for the other models go to zero), the model–averaged posterior distribution can be significantly different from the model–specific distribution. A counter–intuitive consequence is that in the case of dark energy, the model–averaged posterior shows *tighter constraints around $w = -1$* than any of the evolving dark energy models by itself. This comes about because ΛCDM is the preferred model and hence much of the weight in the model–averaged posterior is shifted to the point $w = -1$ [208]. For further details on multi–model inference, see e.g. [209, 210].

(ii) **Model selection forecasting.** When considering the capabilities of future experiments, it is common stance to predict their performance in terms of constraints on relevant parameters, assuming a fiducial point in parameter space as the true model (often, the current best–fit model). While this is a useful indicator for parameter inference tasks, many questions in cosmology fall rather in the model comparison category. A notable example is again dark energy, where the science driver for many future multi–million–dollar probes is to detect possible departures from a cosmological constant, hence to gather evidence in favour of an evolving dark energy model. It is therefore preferable to assess the capabilities of future experiments by their ability to answer model selection questions.

The procedure is as follows (see [211] for details and the application to dark energy scenarios). At every point in parameter space, mock data from the future observation are generated and the Bayes factor between the competing models is computed, for example between an evolving dark energy and a cosmological constant. Then one delimits in parameter space the region where the future data would *not* be able to deliver a clear model comparison verdict, for example $|\ln B| < 5$ (evidence falling short of the "strong" threshold). The experiment with the smallest "model–confusion" volume in parameter space is to be preferred, since it achieves the highest discriminative power between models. An application of a related technique to the spectral index from the Planck satellite is presented in [212, 213].

Alternatively, we can investigate the full probability distribution for the Bayes factor from a future observation. This allows to make probabilistic statements regarding the outcome of a future model comparison, and in particular to quantify the probability that a new observation will be able to achieve a certain level of evidence for one of the models, given current knowledge. This technique is based on the *predictive distribution* for a future observation, which gives the expected posterior on θ for an observation with experimental capabilities described by e (this might describe sky coverage, noise

levels, target redshift, etc):

$$p(\theta|e, d) = \sum_i p(\mathcal{M}_i|d) \int \mathrm{d}\psi_i^\star p(\theta|\psi_i^\star, e, \mathcal{M}_i) p(\psi_i^\star|d, \mathcal{M}_i). \qquad (50)$$

Here, d are the currently available observations, $p(\mathcal{M}_i|d)$ is the current model posterior, $p(\theta|\psi_i^\star, e, \mathcal{M}_i)$ is the posterior on θ from a future observation e computed assuming $\psi_i^\star$ are the correct model parameters, while each term is weighted by the present probability that $\psi_i^\star$ is the true value of the parameters, $p(\psi_i^\star|d, \mathcal{M}_i)$. The sum over i ensures that the prediction averages over models, as well. From Eq. (50) we can compute the corresponding probability distribution for $\ln B$ from experiment e, for example by employing MCMC techniques (further details are given in [59]). This method is called PPOD, for *predictive posterior odds distribution* and can be useful in the context of experiment design and optimization, when the aim is to determine which choice of e will lead to the best scientific return from the experiment, in this case in terms of model selection capabilities (see [214–216] for a discussion of performance optimization for parameter constraints). For further details on Bayes factor forecasts and experiment design, see [217].

7 Conclusions

Bayesian probability theory offers a consistent framework to deal with uncertainty in several different situations, from parameter inference to model comparison, from prediction to optimization. The notion of probability as a degree of belief is far more general than the restricted view of probability as frequency, and it can be applied equally well both to repeatable experiments and to one–off situations. We have seen how Bayes' theorem is a unique prescription to update our state of knowledge in the light of the available data, and how the basic laws of probability can be used to incorporate all sorts of uncertainty in our inferences, including noise (measurement uncertainty), systematic errors (hyper–parameters), imperfect knowledge of the system (nuisance parameters) and modelling uncertainty (model comparison and model averaging). The same laws can equally well be applied to the problem of prediction, and there is considerable potential for a systematic exploration of experiment optimization and Bayesian decision theory (e.g., given what we know about the Universe and our theoretical models, what are the best observations to achieve a certain scientific goal?).

The exploration of the full potential of Bayesian methods is only just beginning. Thanks to the increasing availability of cheap computational power, it now becomes possible to handle problems that were of intractable complexity until a few years ago. Markov Chain Monte Carlo techniques are nowadays a standard inference tool to derive parameter constraints, and many algorithms are available to explore the posterior pdf in a variety of settings. We have highlighted how the issue of priors — which has traditionally been held against Bayesian methods — is a false problem stemming from a misunderstanding of what Bayes' theorem says. This is not to deny that it can be difficult in practice to choose a prior that is a fair representation of one's degree of belief. But we should not shy away from this task — the fact is, there is no inference without assumptions and a correct application of Bayes' theorem forces us to be absolutely clear about which assumptions we are making. It remains important to quantify as much as possible the extent by which our priors are influencing our results, since in many cases when working at the cutting–edge of research we might not have the luxury of being in a data–dominated regime.

The model comparison approach can formalize in a quantitative manner the intuitive assessment of scientific theories, based on the Occam's razor notion that simpler models ought to be preferred if they offer a satisfactory explanation for the observations. The Bayesian evidence and complexity tell us which models are supported by the data, and what is their effective number of parameters. Multi–model inference delivers model–averaged parameter constraints, thus merging the two levels of parameter inference and model comparison.

The application of Bayesian tools to cosmology and astrophysics is blossoming. As both data sets and models become more complex, our inference tools must acquire a corresponding level of sophistication, as

basic statistical analyses that served us well in the past are no longer up to the task. There is little doubt that the field of cosmostatistics will grow in importance in the future, and Bayesian methods will have a great role to play.

Acknowledgments

References

[1] T. Bayes, Phil. Trans. Roy. Soc. **53** 370–418 (Reproduced in: Biometrika, **45**, 293-315 (1958)) (1763).
[2] S. Dodelson, *Modern Cosmology* (Academic Press, London, 2003).
[3] H. Jeffreys, *Theory of probability*, 3rd edn , Oxford Classics series (reprinted 1998) (Oxford University Press, Oxford, UK, 1961).
[4] B. de Finetti, *Theory of probability* (John Wiley & Sons, Chichester, UK, 1974; reprinted 1995).
[5] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory* (John Wiley & Sons, Chicester, UK, 1994).
[6] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis* (John Wiley & Sons, Chicester, UK, 1992).
[7] E. T. Jaynes, *Probability Theory. The logic of science* (Cambridge University Press, Cambridge, UK, 2003).
[8] J. M. Marin and C. P. Robert, *Bayesian Core. A Practical Approach to Computational Bayesian Statistics* (Springer, New York, 2007).
[9] D. MacKay, *Information theory, inference, and learning algorithms* (Cambridge University Press, Cambridge, UK, 2003).
[10] D. Sivia, *Data Analysis: A Bayesian tutorial* (Oxford University Press, Oxford, UK, 1996).
[11] P. Gregory, *Bayesian logical data analysis for the physical sciences* (Cambridge University Press, Cambridge, UK, 2003).
[12] G. D'Agostini, Probability and Measurement Uncertainty in Physics - a Bayesian Primer, (Available as pre–print: hep-ph:/9512295) (1995).
[13] G. D'Agostini, Rept. Prog. Phys. **66** 1383–1420 (2003).
[14] T. J. Loredo, From Laplace to Supernova SN 1987A: Bayesian Inference in Astrophysics, in T. Fougere (Editor) *Maximum-Entropy and Bayesian Methods*, Available from: http://bayes.wustl.edu/gregory/articles.pdf (accessed Jan 15 2008) (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990), pp. 81–142.
[15] T. J. Loredo, The promise of Bayesian inference for astrophysics, in E. D. Feigelson and G. J. Babu (Eds) *Statistical Challenges in Modern Astronomy*, Available from: http://www.astro.cornell.edu/staff/loredo/bayes/promise.pdf (accessed Jan 15 2008) (Springer, New York, 1992), pp. 275–297.
[16] M. Hobson, A. Jaffe, A. Liddle, P. Mukherjee, *et al.* (Eds) *Bayesian Methods in Cosmology* (Cambridge University Press, Cambridge, UK, to appear, 2008).
[17] P. Diaconis, S. Holmes and R. Montgomery, SIAM Review **49** 211–235 (2007).
[18] R. Protassov, D. A. van Dyk, A. Connors, V. L. Kashyap, *et al.*, Astrophys. J. **571** 545–559 (2002).
[19] G. D'Agostini, Inferring the success parameter p of a binomial model from small samples affected by background, (Available as pre–print: physics/0412069v1) (2004).
[20] S. Andreon, Astron. Astrophys. **416** 865–873 (2004).
[21] R. D. Cousins, Am. J. Phys. **63** 398 (1995).
[22] L. Lyons, A particle physicist's perspective on astrostatistics, in *Proceedings of the Statistical Challenges in Modern Astronomy IV Conference*, 371, Pennsylvania State University, Pennsylvania, USA, 12-15 June 2006 (Astronomical Society of the Pacific, San Francisco, 2007), pp. 361–372.
[23] D. Deutsch, Proceedings of the Royal Society of London **A455** 3129–3137 (1999).
[24] S. Saunders, What is probability?, in E. S. Dolev and N. Kolenda (Eds) *Quo Vadis Quantum Mechanics* (Springer, New York, 2005).
[25] H. Greaves, Probability in the Everett interpretation, to appear. Available from: http://philsci-archive.pitt.edu/archive/00003103/ (accessed Jan 16 2008). (2007).
[26] L. Hardy, Why is nature described by quantum theory?, in J. Barrow, P. Davies and C. J. Harper (Eds) *Science and Ultimate Reality: Quantum Theory, Cosmology, and Complexity* (Cambridge University Press, Cambridge, UK, 2004), pp. 45–71.
[27] R. Cox, J. Am. Stat. Assoc. **14** 1–13 (1946).
[28] J. Ghosh, M. Delampady and T. Samanta, *An Introduction to Bayesian Analysis* (Springer, New York, 2006).
[29] G. Babu and E. Feigelson, *Astrostatistics* (Chapman & Hall/CRC, New York, 1996).
[30] A. Gelman, J. Carlin, H. Stern, *et al.*, *Bayesian data analysis* (Chapman & Hall/CRC, New York, 1995).
[31] E. D. Feigelson and G. J. Babu (Eds) *Statistical Challenges in Astronomy* (Springer, New York, 2003).
[32] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, 2nd edn (Springer, New York, 2004).
[33] D. Gamerman and H. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd edn (Chapman & Hall/CRC, Boca Raton (USA), 2006).
[34] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, *et al.*, J. Chem. Phys. **21** 1087–1092 (1953).
[35] W. K. Hastings, Biometrika **57** 97–109 (1970).
[36] A. Smith and G. Roberts, J. R. Statist. Soc. B **55** 3–23 (1993).
[37] K. M. Hanson, Markov Chain Monte Carlo posterior sampling with the Hamiltonian method, in M. Sonka and K. M. Hanson (Eds) *Medical Imaging: Image Processing* Vol. 4322, Proc. SPIE, pp. 456–467.
[38] K. P. Burnham and D. R. Anderson, *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach*, 2nd edn (Springer, New York, 2002).
[39] P. Mukherjee and D. Parkinson, IJMPA **23** 787–802 (2008).
[40] R. E. Kass and A. E. Raftery, J. Am. Stat. Ass. **90** 773–795 (1995).
[41] C. Han and B. P. Carlin, J. Am. Stat. Assoc. **96** 1122–1132 (2001).

[42] A. Garrett and P. Coles, Comments on Astrophysics **17** 23–47 (1993).
[43] W. H. Press, S. A. Teukolsky, W. T. Vetterling, *et al.*, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge, UK, 2007).
[44] M. Clyde, J. Berger, F. Bullard, E. Ford, W. Jefferys, R. Luo, R. Paulo, *et al.*, Current challenges in Bayesian model choice, in *Proceedings of the Statistical Challenges in Modern Astronomy IV Conference*, 371, Pennsylvania State University, Pennsylvania, USA, 12-15 June 2006 (Astronomical Society of the Pacific, San Francisco, 2007), pp. 224–240.
[45] A. Slosar *et al.*, Mon. Not. Roy. Astron. Soc. **341** L29 (2003).
[46] M. Beltran, J. Garcia-Bellido, J. Lesgourgues, A. R. Liddle, *et al.*, Phys. Rev. **D71** 063532 (2005).
[47] M. Bridges, A. N. Lasenby and M. P. Hobson, Mon. Not. Roy. Astron. Soc. **369** 1123–1130 (2006).
[48] J. Skilling, Nested sampling, in R. Fischer, R. Preuss and U. von Toussaint (Eds) *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, 735 (Amer. Inst. Phys. conf. proc. 2004), pp. 395–405.
[49] J. Skilling, Bayesian Analysis **1** 833–861 (2006).
[50] B. A. Bassett, P. S. Corasaniti and M. Kunz, Astrophys. J. **617** L1–L4 (2004).
[51] P. Mukherjee, D. Parkinson and A. R. Liddle, Astrophys. J. **638** L51–L54 (2006).
[52] R. Shaw, M. Bridges and M. P. Hobson, Mon. Not. Roy. Astron. Soc. **378** 1365–1370 (2007).
[53] F. Feroz and M. P. Hobson (Available as pre-print: arXiv:0704.3704) (2007).
[54] M. Bridges, J. D. McEwen, A. N. Lasenby, *et al.*, Mon. Not. Roy. Astron. Soc. **377** 1473–1480 (2007).
[55] N. Chopin and C. P. Robert, Contemplating Evidence: properties, extensions of, and alternatives to Nested Sampling.. (Available as preprint from: http://www.crest.fr/pageperso/Nicolas.Chopin/Nicolas.Chopin.htm. Accessed Jan 2008.).
[56] I. Verdinelli and L. Wasserman, J. Amer. Stat. Assoc. **90** 614–618 (1995).
[57] A. F. Heavens, T. D. Kitching and L. Verde, Mon. Not. Roy. Astron. Soc. **380** 1029–1035 (2007).
[58] R. Trotta, Mon. Not. Roy. Astron. Soc. **378** 72–82 (2007).
[59] R. Trotta, Mon. Not. Roy. Astron. Soc. **378** 819–824 (2007).
[60] R. Trotta, Mon, Not. R. Astron. Soc. **375** L26–L30 (2007).
[61] T. Sellke, M. Bayarri and J. O. Berger, American Statistician **55** 62–71 (2001).
[62] J. Berger and T. Sellke, J. Am. Stat. Assoc. **82** 112–122 (Rejoinder: *ibid.*, 135–139) (1987).
[63] J. Berger and M. Delampady, Statistical Science **2** 317–335 (Rejoinder: *ibid.*, 348–352) (1987).
[64] M. Kendall and A. Stuart, *The advanced theory of statistics*, 4th (Griffin & Co., London and High Wycombe, 1977).
[65] C. Gordon and R. Trotta, Mon. Not. Roy. Astron. Soc. **382** 1859–1863 (2007).
[66] S. Andreon, A Bayesian approach to galaxy evolution studies, in M. Hobson, A. Jaffe, A. Liddle, P. Mukherjee and D. R. Parkinson (Eds) *Bayesian Methods in Cosmology* (Cambridge University Press, Cambridge, UK, to appear, 2008).
[67] J. Berger, Statistical Science **18** 1–12 (Rejoinder: *ibid.*, 28-32) (2003).
[68] D. Spiegelhalter, N. G. Best, B. P. Carlin, *et al.*, J. Royal. Stat. Soc. B **64** 583–639 (2002).
[69] S. Kullback and R. A. Leibler, Ann. Math. Stat. **22** 79–86 (1951).
[70] M. Kunz, R. Trotta and D. Parkinson, Phys. Rev. **D74** 023503 (2006).
[71] M. Szydlowski and A. Kurek (Available as pre–print: arXiv:0801.0638) (2008).
[72] A. R. Liddle, Mon. Not. Roy. Astron. Soc. **351** L49–L53 (2004).
[73] A. R. Liddle, Mon. Not. Roy. Astron. Soc. Lett. **377** L74–L78 (2007).
[74] H. Akaike, I.E.E.E. Trans. Auto. Control **AC-19** 716–723 (1974).
[75] G. Schwarz, Ann. Statist. **6** 461–464 (1978).
[76] A. Raftery, Sociological Methodology **25** 111–163 (1995).
[77] G. F. Smoot *et al.*, Astrophys. J. **396** L1–L5 (1992).
[78] T. J. Pearson *et al.*, Astrophys. J. **591** 556–574 (2003).
[79] B. S. Mason *et al.*, Astrophys. J. **591** 540–555 (2003).
[80] G. Hinshaw *et al.*, Astrophys. J. Suppl. **170** 288 (2007).
[81] J. K. Adelman-McCarthy *et al.*, Astrophys. J. Suppl. **162** 38–48 (2006).
[82] J. A. Tyson *et al.*, Proc. SPIE Int. Soc. Opt. Eng. **4836** 10 (2002).
[83] K. M. Gorski, G. Hinshaw, A. J. Banday, C. L. Bennett, E. L. Wright, A. Kogut, G. F. Smoot, *et al.*, Astrophys. J. Lett. **430** L89–L92 (1994).
[84] F. Bourliot, P. G. Ferreira, D. F. Mota, *et al.*, Phys. Rev. **D75** 063508 (2007).
[85] C. Skordis, D. F. Mota, P. G. Ferreira, *et al.*, Phys. Rev. Lett. **96** 011301 (2006).
[86] V. Acquaviva, C. Baccigalupi, S. M. Leach, A. R. Liddle, *et al.*, Phys. Rev. **D71** 104025 (2005).
[87] D. L. Wiltshire, New J. Phys. **9** 377 (2007).
[88] D. L. Wiltshire, Phys. Rev. Lett. **99** 251101 (2007).
[89] B. M. Leith, S. C. C. Ng and D. L. Wiltshire, Astrophys. J. **672** L91–L94 (2008).
[90] G. Efstathiou, W. J. Sutherland and S. J. Maddox, Nature **348** 705–707 (1990).
[91] Y. Mao, M. Tegmark, M. McQuinn, M. Zaldarriaga, *et al.* (Available as pre–print: arXiv:0802.1710) (2008).
[92] A. Challinor (Invited review at International Scientific Workshop on Cosmology and Gravitational Physics, Thessaloniki, Greece, 15-16 Dec 2005. Available as pre-print: astro-ph/0606548) (2006).
[93] E. Martinez-Gonzalez (To appear in the proceedings of International Summer School on Data Analysis in Cosmology, Valencia, Spain, 6-10 Sep 2004. Available as pre–print: astro-ph/0610162) (2006).
[94] W. Hu and N. Sugiyama, Astrophys. J. **471** 542–570 (1996).
[95] W. Hu, D. Scott, N. Sugiyama, *et al.*, Phys. Rev. **D52** 5498–5515 (1995).
[96] W. Hu, N. Sugiyama and J. Silk, Nature **386** 37–43 (1997).
[97] C. L. Reichardt *et al.* (Available as pre-print: arXiv:0801.1491) (2008).
[98] A. C. S. Readhead *et al.*, Astrophys. J. **609** 498–512 (2004).
[99] T. E. Montroy *et al.*, Astrophys. J. **647** 813 (2006).
[100] C. L. Kuo *et al.*, Astrophys. J. **664** 687–701 (2007).
[101] W. C. Jones *et al.*, Astrophys. J. **647** 823 (2006).
[102] U. Seljak, A. Slosar and P. McDonald, JCAP **0610** 014 (2006).
[103] D. J. Eisenstein *et al.*, Astrophys. J. **633** 560–574 (2005).
[104] S. Cole *et al.*, Mon. Not. Roy. Astron. Soc. **362** 505–534 (2005).
[105] M. Tegmark *et al.*, Phys. Rev. **D74** 123507 (2006).
[106] A. G. Riess *et al.*, Astrophys. J. **607** 665–687 (2004).
[107] A. G. Riess *et al.*, Astrophys. J. **659** 98–121 (2007).

[108] P. Astier *et al.*, Astron. Astrophys. **447** 31–48 (2006).
[109] W. M. Wood-Vasey *et al.*, Astrophys. J. **666** 694–715 (2007).
[110] S. Perlmutter *et al.*, Astrophys. J. **517** 565–586 (1999).
[111] A. G. Riess *et al.*, Astron. J. **116** 1009–1038 (1998).
[112] N. Straumann, Mod. Phys. Lett. **A21** 1083–1098 (2006).
[113] P. Schneider, Weak Gravitational Lensing, in G. Meylan, P. Jetzer and P. North (Eds) *Gravitational Lensing: Strong, Weak and Micro. Lecture Notes of the 33rd Saas-Fee Advanced Course*, Saas-Fee Advanced Courses (Springer, Berlin, 2006), p. 273.
[114] H. Hoekstra *et al.*, Astrophys. J. **647** 116–127 (2006).
[115] M. Jarvis, B. Jain, G. Bernstein, *et al.*, Astrophys. J. **644** 71–79 (2006).
[116] L. Fu *et al.* (Available as pre–print: arXiv:0712.0884) (2007).
[117] J. Benjamin *et al.*, Mon. Not. Roy. Astron. Soc. **381** 702–712 (2007).
[118] R. Massey *et al.*, Nature **445** 286 (2007).
[119] M. Bucher, K. Moodley and N. Turok, Phys. Rev. **D62** 083508 (2000).
[120] M. Bucher, J. Dunkley, P. G. Ferreira, K. Moodley, *et al.*, Phys. Rev. Lett. **93** 081301 (2004).
[121] R. Trotta, A. Riazuelo and R. Durrer, Phys. Rev. Lett. **87** 231301 (2001).
[122] O. Lahav, S. L. Bridle, M. P. Hobson, A. N. Lasenby, *et al.*, Mon. Not. Roy. Astron. Soc. **315** L45 (2000).
[123] M. P. Hobson, S. L. Bridle and O. Lahav, Mon. Not. Roy. Astron. Soc. **335** 377 (2002).
[124] P. Marshall, N. Rajguru and A. Slosar, Phys. Rev. **D73** 067302 (2006).
[125] M. Kampakoglou, R. Trotta and J. Silk, Mon. Not. Roy. Astron. Soc. **384** 1414–1426 (2008).
[126] W. Press, Understanding data better with Bayesian and global statistical methods, in *Proceedings of the Unsolved Problems in Astrophysics. Conference in Honor of John Bahcall,*J. Ostriker (Editor) (Princeton University Press, Princeton, 1996).
[127] S. Dodelson and L. Knox, Phys. Rev. Lett. **84** 3523 (2000).
[128] A. Melchiorri *et al.*, Astrophys. J. **536** L63–L66 (2000).
[129] O. Host, O. Lahav, F. B. Abdalla, *et al.* (Available as pre–print: arXiv:0709.1317) (2007).
[130] M. Tegmark, Astrophys. J. **514** L69–L72 (1999).
[131] C. H. Lineweaver, Astrophys. J. **505** L69 (1998).
[132] M. Tegmark and M. Zaldarriaga, Astrophys. J. **544** 30–42 (2000).
[133] G. Efstathiou, S. L. Bridle, A. N. Lasenby, M. P. Hobson, *et al.*, Mon. Not. Roy. Astron. Soc. **303** L47–L52 (1999).
[134] M. Tegmark, M. Zaldarriaga and A. J. S. Hamilton, Phys. Rev. **D63** 043007 (2001).
[135] U. Seljak and M. Zaldarriaga, Astrophys. J. **469** 437–444 (1996).
[136] X. Wang, M. Tegmark, B. Jain, *et al.*, Phys. Rev. **D68** 123001 (2003).
[137] J. A. Rubino-Martin *et al.*, Mon. Not. Roy. Astron. Soc. **341** 1084 (2003).
[138] S. Hannestad, Phys. Rev. D **61** 023002 (1999).
[139] N. Christensen, R. Meyer, L. Knox, *et al.*, Class. Quant. Grav. **18** 2677 (2001).
[140] A. Lewis and S. Bridle, Phys. Rev. **D66** 103511 (2002).
[141] A. Lewis, A. Challinor and A. Lasenby, Astrophys. J. **538** 473–476 (2000).
[142] D. N. Spergel *et al.*, Astrophys. J. Suppl. **170** 377 (2007).
[143] L. Verde *et al.*, Astrophys. J. Suppl. **148** 195 (2003).
[144] A. A. Collister and O. Lahav, Publ. Astron. Soc. Pac. **16** 345 (2004).
[145] R. Tagliaferri, G. Longo, S. Andreon, S. Capozziello, C. Donalek, *et al.*, Neural Networks for Photometric Redshifts Evaluation, in *Neural Nets*. Lecture Notes in Computer Science Vol. 2859 (Springer, Berlin, 2003), pp. 226–234.
[146] T. Auld, M. Bridges and M. P. Hobson (Available as pre–print: astro-ph/0703445) (2007).
[147] T. Auld, M. Bridges, M. P. Hobson, *et al.*, Mon. Not. Roy. Astron. Soc. Lett. **376** L11–L15 (2007).
[148] W. A. Fendt and B. D. Wandelt, Astrophys. J. **654** 2–11 (2006).
[149] W. A. Fendt and B. D. Wandelt (Available as pre–print: arXiv:0712.0194) (2007).
[150] B. D. Wandelt, D. L. Larson and A. Lakshminarayanan, Phys. Rev. **D70** 083511 (2004).
[151] H. K. Eriksen *et al.* (Available as pre–print: arXiv:0709.1058) (2007).
[152] H. K. Eriksen *et al.*, Astrophys. J. Suppl. **155** 227–241 (2004).
[153] D. L. Larson *et al.*, Astrophys. J. **656** 653–660 (2007).
[154] J. F. Taylor, M. A. J. Ashdown and M. P. Hobson (Available as pre–print: arXiv:0708.2989) (2007).
[155] E. A. Baltz and P. Gondolo, JHEP **10** 052 (2004).
[156] L. Roszkowski, R. Ruiz de Austri and R. Trotta, JHEP **07** 075 (2007).
[157] B. C. Allanach, C. G. Lester and A. M. Weber, JHEP **12** 065 (2006).
[158] L. Roszkowski, R. Ruiz de Austri, J. Silk, *et al.* (Available as pre–print: arXiv:0707.0622) (2007).
[159] R. Trotta, R. Ruiz de Austri and L. Roszkowski **51** 316–320 (2007).
[160] R. Ruiz de Austri, R. Trotta and L. Roszkowski, JHEP **05** 002 (2006).
[161] B. C. Allanach and C. G. Lester, Phys. Rev. **D73** 015013 (2006).
[162] B. C. Allanach, Phys. Lett. **B635** 123–130 (2006).
[163] L. Roszkowski, R. Ruiz de Austri and R. Trotta, JHEP **04** 084 (2007).
[164] B. C. Allanach, K. Cranmer, C. G. Lester, *et al.*, JHEP **08** 023 (2007).
[165] G. Ballesteros, J. A. Casas, J. R. Espinosa, R. Ruiz de Austri, *et al.* (Available as pre–print: arXiv:0711.3436) (2007).
[166] D. Huterer and A. Cooray, Phys. Rev. **D71** 023506 (2005).
[167] C. Zunckel and R. Trotta, Mon. Not. Roy. Astron. Soc. **380** 865 (2007).
[168] S. L. Bridle, A. M. Lewis, J. Weller, *et al.*, Mon. Not. Roy. Astron. Soc. **342** L72 (2003).
[169] P. Mukherjee and Y. Wang, JCAP **0512** 007 (2005).
[170] S. Leach, Mon. Not. Roy. Astron. Soc. **372** 646–654 (2006).
[171] W. H. Kinney, E. W. Kolb, A. Melchiorri, *et al.*, Phys. Rev. **D74** 023502 (2006).
[172] H. Peiris and R. Easther, JCAP **0610** 017 (2006).
[173] H. Peiris and R. Easther, JCAP **0607** 002 (2006).
[174] S. L. Bridle, M. P. Hobson, A. N. Lasenby, *et al.*, Mon. Not. Roy. Astron. Soc. **299** 895–903 (1998).
[175] P. J. Marshall, M. P. Hobson, S. F. Gull, *et al.*, Mon. Not. Roy. Astron. Soc. **335** 1037 (2002).
[176] S. L. Bridle, M. P. Hobson, R. Saunders, *et al.* (Available as pre–print: astro-ph/0010387) (2000).
[177] T. Park *et al.*, Astrophys. J. **652** 610–628 (2006).
[178] A. Niarchou, A. H. Jaffe and L. Pogosian, Phys. Rev. **D69** 063515 (2004).
[179] B. Allen, J. D. E. Creighton, E. E. Flanagan, *et al.*, Phys. Rev. **D67** 122002 (2003).

[180] E. Ford and P. Gregory, Bayesian model selection and extrasolar planet detection (with discussion), in *Proceedings of the Statistical Challenges in Modern Astronomy IV Conference*, 371, Pennsylvania State University, Pennsylvania, USA, 12-15 June 2006 (Astronomical Society of the Pacific, San Francisco, 2007), pp. 189–205.
[181] D. Lindley, Biometrika **44** 187–192 (1957).
[182] G. Lazarides, R. Ruiz de Austri and R. Trotta, Phys. Rev. **D70** 123527 (2004).
[183] A. H. Jaffe, Astrophys. J. **471** 24 (1996).
[184] D. Parkinson, P. Mukherjee and A. R. Liddle, Phys. Rev. **D73** 123523 (2006).
[185] M. Bridges, A. N. Lasenby and M. P. Hobson, Mon. Not. Roy. Astron. Soc. **381** 68–74 (2007).
[186] T. D. Saini, J. Weller and S. L. Bridle, Mon. Not. Roy. Astron. Soc. **348** 603 (2004).
[187] P. Serra, A. Heavens and A. Melchiorri, Mon. Not. Roy. Astron. Soc. **379** 169–175 (2007).
[188] Y. Gong and X. Chen, Phys. Rev. **D76** 123007 (2007).
[189] O. Elgaroy and T. Multamaki, JCAP **0609** 002 (2006).
[190] M. V. John and J. V. Narlikar, Phys. Rev. **D65** 043506 (2002).
[191] A. Balbi, M. Bruni and C. Quercellini, Phys. Rev. **D76** 103519 (2007).
[192] R. Trotta and A. Melchiorri, Phys. Rev. Lett. **95** 011305 (2005).
[193] C. J. Copi, D. Huterer, D. J. Schwarz, *et al.*, Mon. Not. Roy. Astron. Soc. **367** 79–102 (2006).
[194] D. J. Schwarz, G. D. Starkman, D. Huterer, *et al.*, Phys. Rev. Lett. **93** 221301 (2004).
[195] R. Keskitalo, H. Kurki-Suonio, V. Muhonen, *et al.*, JCAP **0709** 008 (2007).
[196] R. Bean, J. Dunkley and E. Pierpaoli, Phys. Rev. **D74** 063503 (2006).
[197] M. Beltran, J. Garcia-Bellido, J. Lesgourgues, *et al.*, Phys. Rev. **D70** 103530 (2004).
[198] G. Efstathiou (Available as pre–print: arXiv:0802.3185) (2008).
[199] J. P. Uzan, Gen. Rel. Grav. **39** 307–342 (2007).
[200] E. J. Copeland, M. Sami and S. Tsujikawa, Int. J. Mod. Phys. **D15** 1753–1936 (2006).
[201] R. Trotta, New Astron. Rev. **51** 316–320 (2007).
[202] F. Simpson and S. Bridle, Phys. Rev. **D73** 083001 (2006).
[203] M. Bridges *et al.* (Available as pre–print: arXiv:0712.1789) (2007).
[204] J. Garcia-Bellido and T. Haugboelle (Available as pre–print: arXiv:0802.1523) (2008).
[205] H. K. Eriksen, A. J. Banday, K. M. Gorski, F. K. Hansen, *et al.*, Astrophys. J. **660** L81–L84 (2007).
[206] T. R. Jaffe, A. J. Banday, H. K. Eriksen, K. M. Gorski, *et al.*, Astrophys. J. **629** L1–L4 (2005).
[207] C. Gordon, Astrophys. J. **656** 636–640 (2007).
[208] A. R. Liddle, P. Mukherjee, D. Parkinson, *et al.*, Phys. Rev. **D74** 123506 (2006).
[209] J. Hoeting, D. Madigan, A. Raftery, *et al.*, Stat. Sci. **14** 382 (1999).
[210] M. Clyde and E. I. George, Stat. Science **19** 81–94 (2004).
[211] P. Mukherjee, D. Parkinson, P. S. Corasaniti, A. R. Liddle, *et al.*, Mon. Not. Roy. Astron. Soc. **369** 1725–1734 (2006).
[212] C. Pahud, A. R. Liddle, P. Mukherjee, *et al.*, Mon. Not. Roy. Astron. Soc. **381** 489–493 (2007).
[213] C. Pahud, A. R. Liddle, P. Mukherjee, *et al.*, Phys. Rev. **D73** 123524 (2006).
[214] B. A. Bassett, Phys. Rev. **D71** 083517 (2005).
[215] B. A. Bassett, D. Parkinson and R. C. Nichol, Astrophys. J. **626** L1–L4 (2005).
[216] T. J. Loredo, Bayesian Adaptive Exploration, in *Proceedings of the 23rd Annual Conference on Bayesian Methods and Maximum Entropy in Science and Engineering* Vol. 707, AIP Conf. Proc., pp. 330–346.
[217] R. Trotta, M. Kunz, P. Mukherjee, *et al.*, Bayesian experimental design and model selection forecasting, in M. Hobson, A. Jaffe, A. Liddle, P. Mukherjee and D. R. Parkinson (Eds) *Bayesian Methods in Cosmology* (Cambridge University Press, Cambridge, UK, to appear, 2008).

Roberto Trotta is the Lockyer Fellow of the Royal Astronomical Society at the Astrophysics Department of the University of Oxford and a Junior Fellow of St Anne's College, Oxford. He has worked on the problem of constraining exotic physics with cosmic microwave background anisotropies and is interested in several aspects of observational cosmology and astroparticle physics, in particular in relation with dark matter and dark energy. His recent research focuses on statistical issues in cosmology.