e-mail: r.trotta@imperial.ac.uk

Bayesian methods

Roberto Trotta

Abstract These notes aim at presenting an overview of Bayesian statistics, the underlying concepts and application methodology that will be useful to astronomers seeking to analyse and interpret a wide variety of data about the Universe. The level starts from elementary notions, without assuming any previous knowledge of statistical methods, and then progresses to more advanced, research-level topics. After an introduction to the importance of statistical inference for the physical sciences, elementary notions of probability theory and inference are introduced and explained. Bayesian methods are then presented, starting from the meaning of Bayes Theorem and its use as inferential engine, including a discussion on priors and posterior distributions. Numerical methods for generating samples from arbitrary posteriors (including Markov Chain Monte Carlo and Nested Sampling) are then covered. The last section deals with the topic of Bayesian model selection and how it is used to assess the performance of models, and contrasts it with the classical p-value approach. A series of exercises of various levels of difficulty are designed to further the understanding of the theoretical material.

Roberto Trotta
Imperial College London, Imperial Centre for Inference and Cosmology, Blackett Laboratory, Prince Consort Road, London SW7 2AZ

Contents

1 Introduction

The purpose of physics is to learn about regularities in the natural phenomena in the world, which we call "Laws of Physics". Theoretical models expressed in mathematical form (e.g., Newton's theory of gravitation) have to be validated through experiments or observations of the phenomena they aim to describe (e.g., measurement of the time it takes for an apple to fall). Thus an essential part of physics is the quantitative comparison of its theories (i.e., models, equations, predictions) with observations (i.e., data, measurements). This leads to confirm theories or to refute them.

Measurements often have uncertainties associated with them. Those could originate in the noise of the measurement instrument, or in the random nature of the process being observed, or in selection effects. Statistics is the tool by which we can extract information about physical quantities from noisy, uncertain and/or incomplete data. Uncertainties however are more general than that. There might be uncertainty in the relationship between quantities in a model (as a consequence of limited information or true intrinsic variability of the objects being studied); uncertainty in the completeness of the model itself; and uncertainty due to unmodelled systematics (to name but a few).

The purpose of these lectures is to provide an appreciation of the fundamental principles underpinning statistical inference, i.e., the process by which we reconstruct quantities of interest from data, subject to the various sources of uncertainty above. The lectures will also endeavour to provide the conceptual, analytical and numerical tools required to approach and solve some of the most common inference problems in the physical sciences, and in particular in cosmology. References are provided so that the reader can further their understanding of the more advanced topics, at research level and beyond.

Probability theory, as a branch of mathematics, is concerned with studying the properties of sampling distributions, i.e., probability distributions that describe the relative frequency of occurrence of random phenomena. In this sense, probability theory is "forward statistics": given the properties of the underlying distributions, it predicts the outcome of data drawn from such distributions.

Statistical inference, by contrast, asks the question of what can be learnt about the underlying distributions from the observed data. It therefore is sometimes called "inverse probability", in that it seeks to reconstruct the parameters of the distributions out of which the data are believed to have been generated.

Statistics addresses several relevant questions for physicists:

(i) How can we learn about regularities in the physical world given that any measurement is subject to a degree of randomness?
(ii) How do we quantify our uncertainty about observed properties in the world?
(iii) How can we make predictions about the future from past experience and theoretical models?.

Inference and statistics are today at the heart of the scientific process, not merely an optional nuisance. Ernest Rutherford is reported to have said, over a century ago: "If you need statistics, you did the wrong experiment". While this might have had some merit at the time, it completely misses the point of what science has become today. All scientific questions at the forefront of research involve increasingly complicated models that try to explain subtle effects in complex, multidimensional data sets. The sheer amount of data available to astrophysicists and cosmologists has increased by orders of magnitudes in the last 20 years. Correspondingly, the sophistication of our statistical analysis tools has to keep up: increasingly, the limiting factor of our knowledge about the Universe is not the amount of data we have, but rather our ability of analyse, interpret and make sense of them.

To paraphrase Rutherford, in 21st Century astrophysics f you do *not* need statistics, it's because you are doing the wrong kind of physics! There are (at least) five good reasons why every professional astrophysicist and cosmologist ought to have a solid training in advanced statistical methods:

(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 2–3σ level, i.e., with a nominal statistical significance somewhere in the region of 95% to 99.7%. This is when tantalizing suggestions for an effect start 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 we don't have the luxury to be able to gather better or further 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 floor.

2 Elementary notions

2.1 The notion of probability

There are two different ways of understanding what probability is. The classical (so-called "frequentist") notion of probability is that probabilities are tied to the frequency of outcomes over a long series of trials. Repeatability of an experiment is the key concept.

The Bayesian outlook[1] is that probability expresses a degree of belief in a proposition, based on the available knowledge of the experimenter. Information is the key concept. Bayesian probability theory is more general than frequentist theory, as the former can deal with unique situations that the latter cannot handle (e.g., "what is the probability that it will rain tomorrow?).

Let $A, B, C, \ldots$ denote propositions (e.g., that a coin toss gives tails). Let Ω describe the sample space (or state space) of the experiment, i.e., Ω is a list of all the possible outcomes of the experiment.

Example 1. If we are tossing a coin, $\Omega = \{T, H\}$, where T denotes "tails" and H denotes "head". If we are rolling a regular die, $\Omega = \{1, 2, 3, 4, 5, 6\}$. If we are drawing one ball from an urn containing white and black balls, $\Omega = \{W, B\}$, where W denotes a white ball and B a black ball.

Frequentist definition of probability: The number of times an event occurs divided by the total number of events in the limit of an infinite series of equiprobable trials.

Definition 1. The joint probability of A and B is the probability of A and B happening together, and is denoted by $P(A, B)$. The conditional probability of A given B is the probability of A happening given that B has happened, and is denoted by $P(A|B)$.

The sum rule:

$$P(A) + P(\overline{A}) = 1, \tag{1}$$

where $\overline{A}$ denotes the proposition "not A".

The product rule:

$$P(A, B) = P(A|B)P(B). \tag{2}$$

By inverting the order of A and B we obtain that

$$P(B, A) = P(B|A)P(A) \tag{3}$$

[1] So-called after Rev. Thomas Bayes (1701(?)–1761), who was the first to introduce this idea in a paper published posthumously in 1763, "An essay towards solving a problem in the doctrine of chances" [3].

and because $P(A,B) = P(B,A)$, we obtain Bayes theorem by equating Eqs. (2) and (3):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4}$$

The marginalisation rule follows from the two rules above and it reads:

$$P(A) = P(A,B_1) + P(A,B_2) + \cdots = \sum_i P(A,B_i) = \sum_i P(A|B_i)P(B_i), \tag{5}$$

where the sum is over all possible outcomes for proposition B.

Definition 2. Two propositions (or events) are said to be independent if and only if

$$P(A,B) = P(A)P(B). \tag{6}$$

2.2 Random variables, parent distributions and samples

Definition 3. A random variable (RV) is a function mapping the sample space Ω of possible outcomes of a random process to the space of real numbers.

Example 2. When tossing a coin once, the RV X can be defined as

$$X = \begin{cases} 0, & \text{if coin lands T} \\ 1, & \text{if coin lands H.} \end{cases} \tag{7}$$

When rolling a regular, 6-sided die, the RV X can be defined as

$$X = \begin{cases} 1, & \text{if a 1 is rolled} \\ 2, & \text{if a 2 is rolled} \\ 3, & \text{if a 3 is rolled} \\ 4, & \text{if a 4 is rolled} \\ 5, & \text{if a 5 is rolled} \\ 6, & \text{if a 6 is rolled.} \end{cases} \tag{8}$$

When drawing one ball from an urn containing black and white balls, the RV X can be defined as

$$X = \begin{cases} 0, & \text{if the ball drawn is white} \\ 1, & \text{if the ball drawn is black.} \end{cases} \tag{9}$$

A RV can be discrete (only a countable number of outcomes is possible, such as in coin tossing) or continuous (an uncountable number of outcomes is possible, such as in a temperature measurement). It is mathematically subtle to carry out the passage from a discrete to a continuous RV, although as physicists we won't

bother too much with mathematical rigour here. Heuristically, we simply replace summation sums over discrete variables with integrals over continuous variables.

Definition 4. Each RV has an associated probability distribution to it. The probability distribution of a discrete RV is called probability mass function (pmf), which gives the probability of each outcome: $P(X = x_i) = P_i$ gives the probability of the RV X assuming the value x_i. In the following we shall use the shorthand notation $P(x_i)$ to mean $P(X = x_i)$.

Example 3. If X is the RV of Eq. (8), and the die being tossed is fair, then $P_i = 1/6$ for $i = 1, \ldots, 6$, where x_i is the outcome "a the face with i pips comes up".

The probability distribution associated with a continuous RV is called the probability density function (pdf), denoted by $p(X)$. The quantity $p(x)dx$ gives the probabilty that the RV X assumes the value between x and $x + dx$.

The choice of probability distribution to associate to a given random process is dictated by the nature of the random process one is investigating (a few examples are given below).

For a discrete pmf, the cumulative probability distribution function (cdf) is given by

$$C(x_i) = \sum_{j=1}^{i} P(x_j). \tag{10}$$

The cdf gives the probabilty that the RV X takes on a value less than or equal to x_i, i.e. $C(x_i) = P(X \leq x_i)$.

For a continuous pdf, the cdf is given by

$$P(x) = \int_{-\infty}^{x} p(y)dy, \tag{11}$$

with the same interpretation as above, i.e. it is the probability that the RV X takes a value smaller than x.

When we make a measurement, (e.g., the temperature of an object, or we toss a coin and observe which face comes up), nature selects an outcome from the sample space with probability given by the associated pmf or pdf. The selection of the outcome is such that if the measurement was repeated an infinite number of times the relative frequency of each outcome is the same as the the probability associated with each outcome under the pmf or pdf. This is another formulation of the frequentist definition of probability given above.

Outcomes of measurements realized by nature are called samples[2]. They are a series of real (or integer) numbers, $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$. In this notes, I will denote samples (i.e., measured values) with a hat symbol, $\hat{}$.

Definitions and background material on some of the most important and most commonly-encountered sampling distributions (the uniform, Poisson, Binomial, exponential and Gaussian distributions) are given in Appendix 4.6.

[2] The probability theory notion of sample encountered here is not to be confused with the idea of MCMC (posterior) samples, which we will introduce later in section 3.6.

2.3 The Central Limit Theorem

The Central Limit Theorem (CLT) is a very important result justifying why the Gaussian distribution is ubiquitous.

Theorem 1. *Simple formulation of the CLT: Let $X_1, X_2, \ldots, X_N$ be a collection of independent RV with finite expectation value μ and finite variance σ^2. Then, for $N \to \infty$, thir sum is Gaussian distributed with mean $N\mu$ and variance $N\sigma^2$.*

Note: it does not matter what the detailed shape of the underlying pdf for the individual RVs is!

Consequence: whenever a RV arises as the sum of several independent effects (e.g., noise in a temperature measurement), we can be confident that it will be very nearly Gaussian distributed.

Theorem 2. *More rigorous (and more general) formulation of the CLT: Let $X_1, X_2, \ldots, X_N$ be a collection of independent RV, each with finite expectation value μ_i and finite variance σ_i^2. Then the variable*

$$Y = \frac{\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \mu_i}{\sum_{i=1}^{N} \sigma_i^2} \tag{12}$$

is distributed as a Gaussian with expectation value 0 and unit variance.

2.4 The likelihood function

The problem of inference can be stated as follows: given a collection of samples, $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$, and a generating random process, what can be said about the properties of the underlying probability distribution?

Example 4. You toss a coin 5 times and obtain 1 head. What can be said about the fairness of the coin?

Example 5. With a photon counter you observe 10 photons in a minute. What can be said about the average photon rate from the source?

Example 6. You measure the temperature of an object twice with two different instruments, yielding the following measurements: $T = 256 \pm 10$ K and $T = 260 \pm 5$ K. What can be said about the temperature of the object?
Schematically, we have that:

$$\begin{array}{c} \text{pdf - e.g., Gaussian with a given } (\mu, \sigma) \to \text{Probability of observation} \\ \text{Underlying } (\mu, \sigma) \leftarrow \text{Observed events} \end{array} \tag{13}$$

The connection between the two domains is given by the likelihood function.

Definition 5. Given a pdf or a pmf $p(X|\theta)$, where X represents a random variable and θ a collection of parameters describing the shape of the pdf[3] and the observed data $\hat{x} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$, the likelihood function $\mathscr{L}$ (or "likelihood" for short) is defined as

$$\mathscr{L}(\theta) = p(X = \hat{x}|\theta). \tag{14}$$

On the right-hand side of the above equation, the probability (density) of observing the data that have been obtained ($X = \hat{x}$) is considered *as a function of the parameters* θ. A very important – and often misunderstood! – point is that the likelihood is *not* a pdf in θ. This is why it's called likelihood *function*! It is normalised over X, but not over θ.

Example 7. In tossing a coin, let θ be the probability of obtaining heads in one throw. Suppose we make $N = 5$ flips and obtain the sequence $\hat{x} = \{H, T, T, T, T\}$. The likelihood is obtained by taking the binomial, Eq. (134), and replacing for r the number of heads obtained ($r = 1$) in $N = 5$ trials, and looking at it as a function of the parameter we are interested in determining, here θ. Thus

$$\mathscr{L}(\theta) = \binom{5}{1} \theta^1 (1-\theta)^4 = 5\theta(1-\theta)^4, \tag{15}$$

which is plotted as a function of θ in Fig. 1.

If instead of $r = 1$ heads we had obtained a different number of heads in our $N = 5$ trials, the likelihood function would have looked as shown in Fig. 2 for a few different choices for r.

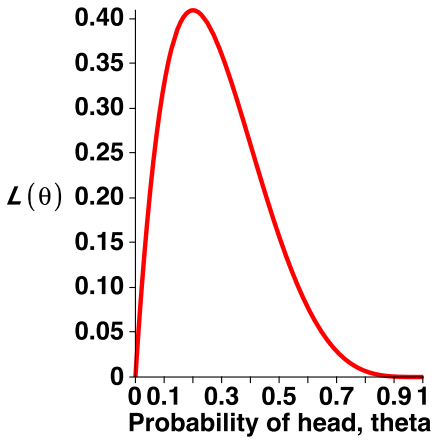

Fig. 1 The likelihood function for the probability of heads (θ) for the coin tossing example, with $N = 5, r = 1$.

[3] For example, for a Gaussian $\theta = \{\mu, \sigma\}$, for a Poisson distribution, $\theta = \lambda$ and for a binomial distribution, $\theta = p$, the probability of success in one trial.

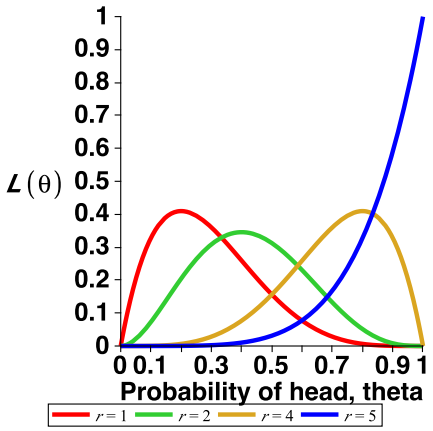

Fig. 2 The likelihood function for the probability of heads (θ) for the coin tossing example, with $n = 5$ trials and different values of r.

This example leads to the formulation of the Maximum Likelihood Principle: if we are trying to determine the value of θ given what we have observed (e.g., the sequence of H/T in coin tossing), we should choose the value that maximises the likelihood, because this maximises the probability of obtaining the data that we got. Notice that this is *not* necessarily the same as maximising the probability of θ. Doing so requires the use of Bayes theorem, see section 3.

2.5 The Maximum Likelihood Principle

The Maximum Likelihood Principle (MLP): given the likelihood function $\mathscr{L}(\theta)$ and seeking to determine the parameter θ, we should choose the value of θ in such a way that the value of the likelihood is maximised.

Definition 6. The Maximum Likelihood Estimator (MLE) for θ is

$$\theta_{\mathrm{ML}} \equiv \max_{\theta} \mathscr{L}(\theta). \tag{16}$$

It can be shown that the MLE as defined above has the following properties: it is asymptotically unbiased (i.e., $\theta_{\mathrm{ML}} \to \theta$ for $N \to \infty$, i.e., the ML estimate converges to the true value of the parameters for infinitely many data points) and it is asymptotically the minimum variance estimator, i.e. the one with the smallest errors.

To find the MLE, we maximise the likelihood by requiring its first derivative to be zero and the second derivative to be negative:

$$\left.\frac{\partial \mathscr{L}(\theta)}{\partial \theta}\right|_{\theta_{\mathrm{ML}}} = 0, \qquad \text{and} \qquad \left.\frac{\partial^2 \mathscr{L}(\theta)}{\partial \theta^2}\right|_{\theta_{\mathrm{ML}}} < 0. \tag{17}$$

In practice, it is often more convenient to maximise the logarithm of the likelihood (the "log-likelihood") instead. Since log is a monotonic function, maximising the likelihood is the same as maximising the log-likelihood. So one often uses

$$\frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta}\Big|_{\theta_{\rm ML}} = 0, \qquad \text{and} \qquad \frac{\partial^2 \ln \mathscr{L}(\theta)}{\partial \theta^2}\Big|_{\theta_{\rm ML}} < 0. \tag{18}$$

Example 8. MLE of the mean of a Gaussian. Imagine we have N independent measurements of a Gaussian-distributed quantity, and let's denote them by $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$. Here the parameters we are interested in determining are μ (the mean of the distribution) and σ (the standard deviation of the distribution), hence we write $\theta = \{\mu, \sigma\}$.Then the joint likelihood function is given by

$$\mathscr{L}(\mu, \sigma) = p(\hat{x}|\mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\hat{x}_i - \mu)^2}{\sigma^2}\right), \tag{19}$$

Often, the expression above is written as

$$\mathscr{L} = L_0 \exp\left(-\chi^2/2\right) \tag{20}$$

where the so-called "chi-squared" is defined as

$$\chi^2 = \sum_{i=1}^{N} \frac{(\hat{x}_i - \mu)^2}{\sigma^2}. \tag{21}$$

We want to estimate the (true) mean of the Gaussian. The MLE for the mean is obtained by solving

$$\frac{\partial \ln \mathscr{L}}{\partial \mu} = 0 \Rightarrow \mu_{\rm ML} = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_i, \tag{22}$$

i.e., the MLE for the mean is just the sample mean (i.e., the average of the measurements).

Example 9. MLE of the standard deviation of a Gaussian. If we want to estimate the standard deviation σ of the Gaussian, the MLE for σ is:

$$\frac{\partial \ln \mathscr{L}}{\partial \sigma} = 0 \Rightarrow \sigma^2_{\rm ML} = \frac{1}{N}\sum_{i=1}^{N} (\hat{x}_i - \mu)^2. \tag{23}$$

However, the MLE above is "biased", i.e. it can be shown that

$$E(\sigma^2_{\rm ML}) = \left(1 - \frac{1}{N}\right)\sigma^2 \neq \sigma^2, \tag{24}$$

where $E(\dot{})$ denotes the expectation value. I.e., for finite N the expectation value of the ML estimator is not the same as the true value, σ^2. In order to obtain an unbiased

estimator we replace the factor $1/N$ by $1/(N-1)$. Also, because the true μ is usually unknown, we replace it in Eq. (23) by the MLE estimator for the mean, μ_{ML}.

Therefore, the unbiased MLE estimator for the variance is

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\hat{x}_i - \mu_{\mathrm{ML}})^2. \tag{25}$$

In general, you should always use Eq. (25) as the ML estimator for the variance, and not Eq. (23).

Example 10. MLE for the success probability of a binomial distribution. We go back to the coin tossing example, but this time we solve it in all generality. Let's define "success" as "the coin lands heads" (H). Having observed H heads in a number N of trials, the likelihood function of a binomial is given by Eq. (134), where the unknown parameter is θ (the success probability for one trial, i.e., the probability that the coin lands H):

$$\mathscr{L}(\theta) = P(H|\theta,N) = \binom{N}{H}\theta^H(1-\theta)^{N-H}, \tag{26}$$

The Maximum Likelihood Estimator the success probability is found by maximising the log likelihood:

$$\begin{aligned}\frac{\partial \ln \mathscr{L}(\theta)}{\partial\theta} &= \frac{\partial}{\partial\theta}\left(\ln\binom{N}{H} + H\ln\theta + (N-H)\ln(1-\theta)\right) = \frac{H}{\theta} - \frac{N-H}{1-\theta} \stackrel{!}{=} 0 \\ &\Leftrightarrow \theta_{\mathrm{ML}} = \frac{H}{N}.\end{aligned} \tag{27}$$

Thus the MLE is simpy given by the observed fraction of heads, which is intuitively obvious.

Example 11. MLE for the rate of a Poisson distribution. The likelihood function is given by Eq. (136), using the notation $\theta = \lambda$ (i.e., the parameter θ we are interested in is here the rate λ):

$$\mathscr{L}(\lambda) = P(n|\lambda) = \frac{(\lambda t)^n}{n!}\exp(-\lambda t), \tag{28}$$

The unknown parameter is the rate λ, while the data are the observed counts, n, in the amount of time t. The Maximum Likelihood Estimate for λ is obtained by finding the maximum of the log likelihood as a function of the parameter (here, the rate λ). Hence we need to find the value of λ such that:

$$\frac{\partial \ln P(n|\lambda)}{\partial\lambda} = 0. \tag{29}$$

The derivative gives

$$\frac{\partial \ln P(n|\lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda}(n\ln(\lambda t) - \ln n! - \lambda t) = n\frac{t}{\lambda t} - t = 0 \Leftrightarrow \lambda_{MLE} = \frac{n}{t}. \quad (30)$$

So the maximum likelihood estimator for the rate is the observed average number of counts.

We can thus summarise the MLE recipe:

(i) Write down the likelihood. This depends on the kind of random process you are considering. Identify what is the parameter that you are interested in, θ.
(ii) Find the "best fit" value of the parameter of interest by maximising the likelihood $\mathscr{L}$ as a function of θ. This is your MLE, $\theta_{\rm ML}$.
(iii) Evaluate the uncertainty on $\theta_{\rm ML}$, i.e. compute the confidence interval (see next section).

2.6 Confidence intervals (frequentist)

Consider a general likelihood function, $\mathscr{L}(\theta)$ and let us do a Taylor expansion of the log-likelihood $\ln\mathscr{L}$ around its maximum, given by $\theta_{\rm ML}$:

$$\ln\mathscr{L}(\theta) = \ln\mathscr{L}(\theta_{\rm ML}) + \frac{\partial \ln\mathscr{L}(\theta)}{\partial\theta}\Big|_{\theta_{\rm ML}}(\theta-\theta_{\rm ML}) + \frac{1}{2}\frac{\partial^2 \ln\mathscr{L}(\theta)}{\partial\theta^2}\Big|_{\theta_{\rm ML}}(\theta-\theta_{\rm ML})^2 + \ldots \quad (31)$$

The second term on the RHS vanishes (by definition of the Maximum Likelihood value), hence we can approximate the likelihood as

$$\mathscr{L}(\theta) \approx \mathscr{L}(\theta_{\rm ML})\exp\left(-\frac{1}{2}\frac{(\theta-\theta_{\rm ML})^2}{\Sigma_\theta^2}\right) + \ldots, \quad (32)$$

with

$$\frac{1}{{\Sigma_\theta}^2} = -\frac{\partial^2 \ln\mathscr{L}(\theta)}{\partial\theta^2}\Big|_{\theta_{\rm ML}}. \quad (33)$$

A general likelihood function can be approximated to second order as a Gaussian around the ML value, as shown by Eq. (32). Therefore, to the extent that this second order Taylor expansion is sufficiently accurate, the uncertainty around the ML value, Σ_θ, is approximately given by Eq. (33).

Example 12. Let's go back to the Gaussian problem of Eq. (19). We have seen in Eq. (22) that the sample mean is the MLE for the mean of the Gaussian. We now want to compute the uncertainty on this value. Applying Eq. (33) to the likelihood of Eq. (19) we obtain

$$\Sigma_\mu^2 = \sigma^2/N. \quad (34)$$

This means that the the uncertainty on our ML estimate for μ (as expressed by the standard deviation Σ_μ) is proportional to $1/\sqrt{N}$, with N being the number of measurements.

As the likelihood function can be approximated as a Gaussian (at least around the peak), we can use the results for a Gaussian distribution to approximate the probability content of an interval around the ML estimate for the mean. The interval $[\mu_{\min}, \mu_{\max}]$ is called a $100\alpha\%$ confidence interval for the mean μ if $P(\mu_{\min} < \mu < \mu_{\max}) = \alpha$.

Example 13. For example, the interval $[\mu_{\rm ML} - \Sigma_\mu < \mu < \mu_{\rm ML} + \Sigma_\mu]$ is a 68.3% confidence interval for the mean (a so-called "1σ interval"), while $[\mu_{\rm ML} - 2\Sigma_\mu < \mu < \mu_{\rm ML} + 2\Sigma_\mu]$ is a 95.4% confidence interval (a "2σ interval").

Example 14. In the temperature measurement example of Eq. (42), the 68.3% confidence interval for the mean is $198.0\text{K} < \mu < 201.2\text{K}$. The 95.4% confidence interval is $196.4\text{K} < \mu < 202.8\text{K}$.

Generally, the value after the "$\pm$" sign will usually give the 1σ (i.e., 68.3%) region. Sometimes you might find a notation like 50 ± 1 (95% CL), where "CL" stands for "Confidence Level". In this case, ± 1 encompasses a region of 95% confidence (rather than 68.3%), which corresponds to 1.96 σ (see Table 4).

In the multi-dimensional case, additional parameters are eliminated from the likelihood by profiling over them, i.e., maximising over their value.

Definition 7. The profile likelihood for the parameter θ_1 (without loss of generality) is defined as

$$\mathscr{L}(\theta_1) \equiv \max_{\theta_2,\ldots,\theta_N} \mathscr{L}(\theta), \tag{35}$$

where in our case $\mathscr{L}(\theta)$ is the full likelihood function.

Thus in the profile likelihood one maximises the value of the likelihood along the hidden dimensions, rather than integrating it out as in the marginal posterior (see Eq. (61) below).

The profile likelihood can be directly interpreted as a if it were a genuine likelihood function, except that it does account for the effect of the hidden parameters.

Confidence intervals from the profile likelihood can be obtained via the likelihood ratio test as follows.

Classical confidence intervals based on the Neyman construction are defined as the set of parameter points in which some real-valued function, or *test statistic*, t evaluated on the data falls in an acceptance region $W_\theta = [t_-, t_+]$. Likelihood ratios are often chosen as the test statistic on which frequentist intervals are based. When θ is composed of parameters of interest, θ, and nuisance parameters, ψ, a common choice of test statistic is the profile likelihood ratio

$$\lambda(\theta) \equiv \frac{\mathscr{L}(\theta, \hat{\hat{\psi}})}{\mathscr{L}(\hat{\theta}, \hat{\psi})}. \tag{36}$$

where $\hat{\hat{\psi}}$ is the conditional maximum likelihood estimate (MLE) of ψ with θ fixed and $\hat{\theta}, \hat{\psi}$ are the unconditional MLEs. Under certain regularity conditions[4], Wilks

[4] One important and often-overlooked condition for the validity of Wilks' theorem is that the parameter it is being applied to cannot lie at the boundary of the allowed parameter space. In this

showed [59] that the distribution of $-2\ln\lambda(\theta)$ converges to a chi-square distribution with a number of degrees of freedom given by the dimensionality of θ.

This leads to the following prescription. Starting from the best-fit value in parameter space, an $\alpha\%$ confidence interval encloses all parameter values for which minus twice the log–likelihood increases less than $\Delta\chi^2(\alpha,n)$ from the best fit value. The threshold value depends on α and on the number n of parameters one is simultaneously considering (usually $n=1$ or $n=2$), and it is obtained by solving

$$\alpha = \int_0^{\Delta\chi^2} \chi_n^2(x)dx, \tag{37}$$

where $\chi_n^2(x)$ is the chi–square distribution for n degrees of freedom, Eq. (170).

One has to be careful with the interpretation of confidence intervals as this is often misunderstood!

Interpretation: if we were to repeat an experiment many times, and each time report the observed $100\alpha\%$ confidence interval, we would be correct $100\alpha\%$ of the time. This means that (ideally) a $100\alpha\%$ confidence intervals contains the true value of the parameter $100\alpha\%$ of the time.

In a frequentist sense, it does not make sense to talk about "the probability of θ". This is because every time the experiment is performed we get a different realization (different samples), hence a different numerical value for the confidence interval. Each time, either the true value of θ is inside the reported confidence interval (in which case, the probability of θ being inside is 1) or the true value is outside (in which case its probability of being inside is 0). Confidence intervals do not give the probability of the parameter! In order to do that, you need Bayes theorem.

2.7 Exercices

Those exercises are designed to help you put into practice the above introductory concepts. Please make sure you are familiar with these notions before moving on to the next section.

(i) Gaussian 1D problem. The surface temperature on Mars is measured by a probe 10 times, yielding the following data (units of K):

$$191.9, 201.6, 206.1, 200.4, 203.2, 201.6, 196.5, 199.5, 194.1, 202.4 \tag{38}$$

---

case, one ought to employ Chernoff's theorem instead [11]. A modern discussion of the regularity conditions necessary for the asymptotic distribution of the likelihood ratio test statistics to be valid can be found in [47].

a. Assume that each measurement is independently Normally distributed with known variancee $\sigma^2 = 25$ K^2. What is the likelihood function for the whole data set?
Answer: The measurements are independent, hence the total likelihood is the product of the likelihoods for each measurement:

$$\mathscr{L}_{\rm tot}(T) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\hat{T}_i - T)^2}{\sigma^2}\right) \tag{39}$$

where $\hat{T}_i$ are the data given, T is the temperature we are trying to determine (unknown parameter) and $\sigma = 5$ K.

b. Find the Maximum Likelihood Estimate (MLE) for the surface temperature, $T_{\rm ML}$, and express your result to 4 significant figures accuracy.
Answer: The MLE for the mean of a Gaussian is given by the mean of the sample, see Eq. (22), hence

$$T_{\rm ML} = \frac{1}{10}\sum_{i=1}^{10} T_i = 199.7\text{K}. \tag{40}$$

c. Determine symmetric confidence intervals at 68.3%, 95.4% and 99% around $T_{\rm ML}$ (4 significant figures accuracy).
Answer: The variance of the mean is given by σ^2/N, see Eq. (34). Therefore the standard deviation of our estimate $T_{\rm ML}$ is given by $\Sigma_T = \sigma/\sqrt{N} = 5/\sqrt{10}$ = 1.58 K, which corresponds to the 68.3% interval: 199.7 ± 1.6 K, i.e. the range $[198.1, 201.3]$ K (4 s.f. accuracy). Confidence intervals at 95.4% and 99% corresponds to symmetric intervals around the mean of length 2.0 and 2.57 times the standard deviation Σ_T. Hence the required confidence intervals are $[196.5, 202.9]$ K (95.4%) and $[195.6, 203.8]$ K (99%).

d. How many measurements would you need to make if you wanted to have a 1σ confidence interval around the mean of length less than 1 K (on each side)?
Answer: A 1σ confidence interval lenght 1 K means that the value of Σ_T should be 1 K. Using that the standard deviation scales as $1/\sqrt{N}$, we have

$$1 = 5/\sqrt{N} \Rightarrow N = 25. \tag{41}$$

You would need $N = 25$ measurements to achieve the desired accuracy.

(ii) The surface temperature on Mars is measured by a probe 10 times, yielding the following data (units of K):

$$197.2, 202.4, 201.8, 198.8, 207.6, 191.4, 201.4, 198.2, 195.7, 201.2. \tag{42}$$

Assuming that each measurement is independently Gaussian distributed with known variance $\sigma^2 = 5$ K^2, what is the likelihood function for the whole data set?
Answer: the measurements are independent, hence the total likelihood is the

product of the likelihoods for each measurement, see Eq. (19):

$$\mathscr{L}(T) = \prod_{i=1}^{10} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\hat{T}_i - T)^2}{\sigma^2}\right) \tag{43}$$

What is the MLE of the mean, T_{ML}?
Answer: the MLE for the mean of a Gaussian is given by the mean of the sample, see Eq. (22), hence

$$T_{\rm ML} = \frac{1}{10}\sum_{i=1}^{10} \hat{T}_i = 199.6K. \tag{44}$$

What is the uncertainty on our MLE for the mean?
Answer: The variance of the mean is given by $\Sigma_\mu^2 = \sigma^2/N$, where $\sigma^2 = 5$ K^2 and $N = 10$. Therefore the standard deviation of our temperature estimate $T_{\rm ML}$ is given by $\Sigma_T = 5/\sqrt{10}$ = 1.6 K. The measurement can thus be summarized as $T = 199.6 \pm 1.6$ K, where the $\pm$1.6 K gives the range of the 1σ (or 68.3%) confidence interval.

(iii) A laser beam is used to measure the deviation of the distance between the Earth and the Moon from its average value, giving the following data, in units of cm:

$$119, \quad 119, \quad 122, \quad 121, \quad 116. \tag{45}$$

a. Assuming that each measurement above follows an independent Gaussian distribution of known standard deviation $\sigma = 3$ cm, write down the joint likelihood function for Δ, the deviation of the Earth-Moon distance from its average value.
Answer: The joint Gaussian likelihood function for Δ is given by

$$P(\Delta|d) \equiv \mathscr{L}(\Delta) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\Delta - d_i)^2}{\sigma^2}\right), \tag{46}$$

where $\sigma = 3$ cm and d_i are the measurements given in the question.

b. Compute the maximum likelihood estimate for Δ and its uncertainty, both to 3 significant figures.
Answer: The maximum likelihood estimate for Δ is found by maximising the log-likelihood function wrt Δ:

$$\frac{\partial \ln \mathscr{L}}{\partial \Delta} = -\sum_{i=1}^{5} \frac{\Delta - d_i}{\sigma^2} = 0 \rightarrow \Delta_{\rm MLE} = \frac{1}{N}\sum_{i=1}^{5} d_i \tag{47}$$

The numerical value is $\Delta_{\rm MLE} = 119.4cm \approx 119$ (cm, 3 s.f.).
The uncertainty Σ on Δ is estimated from the inverse curvature of the log likelihood function at the MLE point:

$$-\frac{\partial^2 \ln \mathscr{L}}{\partial \Delta^2} = \frac{N}{\sigma^2} \rightarrow \Sigma = \left(-\frac{\partial^2 \ln \mathscr{L}}{\partial \Delta^2}\right)^{-1/2} = \frac{\sigma}{\sqrt{N}} \tag{48}$$

Numerically this gives $\Sigma = 3/\sqrt{5} = 1.34 \approx 1$ cm.

c. How would you report the measurement of Δ?
Answer: The measurement of Δ would thus be reported as $\Delta = (119 \pm 1)$ cm.

(iv) An experiment counting particles emitted by a radioactive decay measures r particles per unit time interval. The counts are Poisson distributed.

a. If λ is the average number of counts per per unit time interval, write down the appropriate probability distribution function for r.
b. Now we seek to determine λ by repeatedly measuring for M times the number of counts per unit time interval. This series of measurements yields a sequence of counts $\hat{r} = \{\hat{r}_1, \hat{r}_2, \hat{r}_3, ..., \hat{r}_M\}$. Each measurement is assumed to be independent. Derive the combined likelihood function for λ, $\mathscr{L}(\lambda) = P(\hat{r}|\lambda)$, given the measured sequence of counts $\hat{r}$.
c. Use the Maximum Likelihood Principle applied to the the log likelihood $\ln \mathscr{L}(\lambda)$ to show that the Maximum Likelihood estimator for the average rate λ is just the average of the measured counts, $\hat{r}$, i.e.

$$\lambda_{\mathrm{ML}} = \frac{1}{M} \sum_{i=1}^{M} \hat{r}_i \, .$$

d. By considering the Taylor expansion of $\ln \mathscr{L}(\lambda)$ to second order around λ_{ML}, derive the Gaussian approximation for the likelihood $\mathscr{L}(\lambda)$ around the Maximum Likelihood point (see Eq. (63) in the handout), and show that it can be written as

$$\mathscr{L}(\lambda) \approx L_0 \exp\left(-\frac{1}{2} \frac{M}{\lambda_{\mathrm{ML}}} (\lambda - \lambda_{\mathrm{ML}})^2 \right) ,$$

where L_0 is a normalization constant.
e. Compare with the equivalent expression for M Gaussian-distributed measurements to show that the variance σ^2 of the Poisson distribution is given by $\sigma^2 = \lambda$.

(v) An astronomer measures the photon flux from a distant star using a very sensitive instrument that counts single photons. After one minute of observation, the instrument has collected $\hat{r}$ photons. One can assume that the photon counts, $\hat{r}$, are distributed according to the Poisson distribution. The astronomer wishes to determine λ, the emission rate of the source.

a. What is the likelihood function for the measurement? Identify explicitly what is the unknown parameter and what are the data in the problem.
b. If the true rate is $\lambda = 10$ photons/minute, what is the probability of observing $\hat{r} = 15$ photons in one minute?
c. Find the Maximum Likelihood Estimate for the rate λ (i.e., the number of photons per minute). What is the maximum likelihood estimate if the observed number of photons is $\hat{r} = 10$?

d. Upon reflection, the astronomer realizes that the photon flux is the superposition of photons coming from the star plus "background" photons coming from other faint sources within the field of view of the instrument. The background rate is supposed to be known, and it is given by λ_b photons per minute (this can be estimated e.g. by pointing the telescope away from the source and measuring the photon counts there, when the telescope is only picking up background photos). She then points to the star again, measuring $\hat{r}_t$ photons in a time t_t. What is her maximum likelihood estimate of the rate λ_s from the star in this case?
Hint: The total number of photons $\hat{r}_t$ is Poisson distributed with rate $\lambda = \lambda_s + \lambda_b$, where λ_s is the rate for the star.

e. What is the source rate (i.e., the rate for the star) if $\hat{r}_t = 30$, $t_t = 2$ mins, and $\lambda_b = 12$ photons per minute? Is it possible that the measured average rate from the source (i.e., $\hat{r}_t/t_t$) is less than λ_b? Discuss what happens in this case and comment on the physicality of this result.

(vi) This problem generalizes the Gaussian measurement case to the case where the measurements have different uncertainties among them.
You measure the flux F of photons from a laser source using 4 different instruments and you obtain the following results (units of 10^4 photons/cm^2):

$$34.7 \pm 5.0, \quad 28.9 \pm 2.0, \quad 27.1 \pm 3.0, \quad 30.6 \pm 4.0. \tag{49}$$

a. Write down the likelihood for each measurement, and explain why a Gaussian approximation is justified in this case.

b. Write down the total likelihood for the combination of the 4 measurements.

c. Find the MLE of the photon flux, $F_{\rm ML}$, and show that it is given by:

$$F_{\rm ML} = \sum_i \frac{\hat{n}_i}{\hat{\sigma}_i^2/\bar{\sigma}^2}, \tag{50}$$

where

$$\frac{1}{\bar{\sigma}^2} \equiv \sum_i \frac{1}{\hat{\sigma}_i^2}. \tag{51}$$

d. Compute $F_{\rm ML}$ from the data above and compare it with the sample mean.

e. Find the 1σ confidence interval for your MLE for the mean, and show that it is given by:

$$\left(\sum_i \frac{1}{\hat{\sigma}_i^2} \right)^{-1/2}. \tag{52}$$

Evaluate the confidence interval for the above data. How would you summarize your measurement of the flux F?

3 Bayesian parameter inference

In this section we introduce the meaning and practical application of Bayes Theorem, Eq. (4), which encapsulates the notion of probability as degree of belief.

3.1 Bayes theorem as an inference device

As a mathematical result, Bayes Theorem is elementary and uncontroversial. It becomes interesting for the purpose of inference when we replace in Bayes theorem, Eq. (4), $A \to \theta$ (the parameters) and $B \to d$ (the observed data, or samples), obtaining

$$P(\theta|d) = \frac{P(d|\theta)P(\theta)}{P(d)}. \tag{53}$$

On the LHS, $P(\theta|d)$ is the posterior probability for θ (or "posterior" for short), and it represents our degree of belief about the value of θ after we have seen the data d.

On the RHS, $P(d|\theta) = \mathscr{L}(\theta)$ is the likelihood we already encountered. It is the probability of the data given a certain value of the parameters.

The quantity $P(\theta)$ is the prior probability distribution (or "prior" for short). It represents our degree of belief in the value of θ before we see the data (hence the name). This is an essential ingredient of Bayesian statistics. The Bayesian school is divided between "subjectivists" (who maintain that the prior is a reflection of the subject state of knowledge of the individual researcher adopting it) and "objectivists" (who argue for the use of "standard" priors to enforce inter-subjectivity between different researchers). However formulated, the posterior distribution usually converges to a prior-independent regime for sufficiently large data sets.

In the denominator, $P(d)$ is a normalizing constant (often called "the evidence" or "marginal likelihood"), than ensures that the posterior is normalized to unity:

$$P(d) = \int d\theta P(d|\theta)P(\theta). \tag{54}$$

The evidence is important for Bayesian model selection (see section 4).

Interpretation: Bayes theorem relates the posterior probability for θ (i.e., what we know about the parameter after seeing the data) to the likelihood and the prior (i.e., what we knew about the parameter before we saw the data). It can be thought of as a general rule to update our knowledge about a quantity (here, θ) from the prior to the posterior.

Remember that in general $P(\theta|d) \neq P(d|\theta)$, i.e. the posterior $P(\theta|d)$ and the likelihood $P(d|\theta)$ are two different quantities with different meaning!

Example 15. We want to determine if a randomly-chosen person is male (M) or female (F)[5]. We make one measurement, giving us information on whether the person is pregnant (Y) or not (N). Let's assume we have observed that the person is pregnant, so $d = Y$.

The likelihood is $P(d = Y|\theta = F) = 0.03$ (i.e., there is a 3% probability that a randomly selected female is pregnant), but the posterior probability $P(\theta = F|d = Y) = 1.0$, i.e., if we have observed that the person is pregnant, we are sure she is a woman. This shows that the likelihood and the posterior probability are in general different!

This is because they mean two different things: the likelihood is the probability of making the observation if we know what the parameter is (in this example, if we know that the person is female); the posterior is the probability of the parameter given that we have made a certain observation (in this case, the probability of a person being female if we know she is pregnant). The two quantities are related by Bayes theorem (prove this in the example given here).

Bayesian inference works by updating our state of knowledge about a parameter (or hypothesis) as new data flow in. The posterior from a previous cycle of observations becomes the prior for the next.

3.2 Advantages of the Bayesian approach

Irrespectively of the philosophical and epistemological views about probability, 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 The particles of interest might be photons, neutrinos or cosmic rays. Measurements of the source s must account

[5] This example is due to Louis Lyons.

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. (76)). 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.

(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.

The cosmology and astrophysics communities have been embracing Bayesian mehods since the turning of the Millennium, spurred by the availability of cheap computational power that has ushered in an era of high-performance computing, thus allowing for the first time to deploy the power of Bayesian statistics thanks to numerical implementations (in particular, MCMC and related techniques). The steep increase in the number of Bayesian papers in the astrophysics literature is shown in Fig. 3.

3.3 Considerations and caveats on priors

Bayesian inference works by updating our state of knowledge about a parameter (or hypothesis) as new data flow in. The posterior from a previous cycle of observations becomes the prior for the next. The price we have to pay is that we have to start somewhere by specifying an initial prior, which is not determined by the theory, but it needs to be given by the user. The prior should represent fairly the state of knowledge of the user about the quantity of interest. Eventually, the posterior will

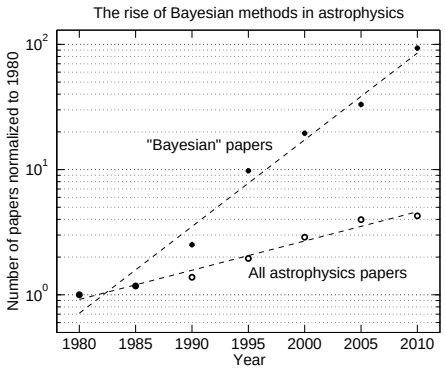

Fig. 3 Number of articles in astronomy and cosmology with "Bayesian" in the title, as a function of publication year (upper data points) and total number of articles (lower data points) as a function of publication year. Numbers are normalized to 1980 levels for each data series. The number of Bayesian papers doubles every 4.3 years, while the total number of papers doubles "only" every 12.6 years. At the present rate, by 2060 all papers on the archive will be Bayesian. (source: NASA/ADS).

converge to a unique (objective) result even if different scientists start from different priors (provided their priors are non-zero in regions of parameter space where the likelihood is large). See Fig. 4 for an illustration.

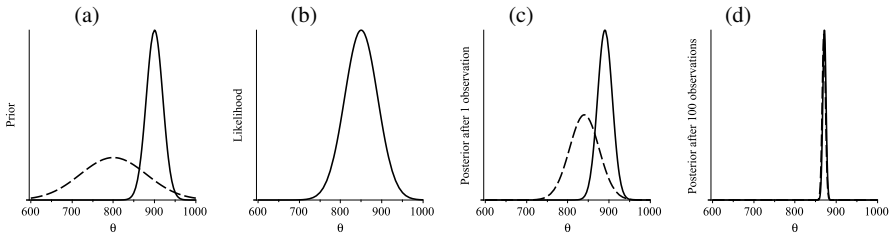

Fig. 4 Converging views in Bayesian inference. Two scientists having different prior believes $p(\theta)$ about the value of a quantity θ (panel (a), the two curves representing two different priors) observe one datum with likelihood $\mathscr{L}(\theta)$ (panel (b)), after which their posteriors $p(\theta|d)$ (panel (c), obtained via Bayes Theorem, Eq. (4)) represent their updated states of knowledge on the parameter. This posterior then becomes the prior for the next observation. After observing 100 data points, the two posteriors have become essentially indistinguishable (d).

There is a vast literature about how to select a prior in an appropriate way. Some aspects are fairly obvious: if your parameter θ describes a quantity that has e.g. to be strictly positive (such as the number of photons in a detector, or an amplitude), then the prior will be 0 for values $\theta < 0$.

A standard (but by no means harmless, see below) choice is to take a *uniform prior* (also called "flat prior") on θ, defined as:

$$P(\theta) = \begin{cases} \frac{1}{(\theta_{\max} - \theta_{\min})} & \text{for } \theta_{\min} \leq \theta \leq \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{55}$$

With this choice of prior in Bayes theorem, Eq. (53), the posterior becomes functionally identical to the likelihood up to a proportionality constant:

$$P(\theta|d) \propto P(d|\theta) = \mathscr{L}(\theta). \tag{56}$$

In this case, all of our previous results about the likelihood carry over (but with a different interpretation). In particular, the probability content of an interval around the mean for the posterior should be interpreted as a statement about our degree of belief in the value of θ (differently from confidence intervals for the likelihood).

Example 16. Let's look once more to the temperature estimation problem of Eq. (42). The Bayesian estimation of the temperature proceeds as follows. We first need to specify the likelihood function – this is the same as before, and it is given by Eq. (42). If we want to estimate the temperature, we need to compute the posterior probability for T, given by (up to a normalization constant)

$$P(T|d) \propto \mathscr{L}(T)P(T) \tag{57}$$

where the likelihood $\mathscr{L}(T)$ is given by Eq. (42). We also need to specify the prior, $P(T)$. For this particular case, we know that $T > 0$ (the temperature in K of an object needs to be positive) and let's assume we know that the temperature cannot exceed 300 K. Therefore we can pick a flat prior of the form

$$P(T) = \begin{cases} \frac{1}{300} & \text{for } 0\text{K} \leq T \leq 300\text{K} \\ 0 & \text{otherwise.} \end{cases} \tag{58}$$

The posterior distribution for T then becomes

$$P(T|d) \propto \begin{cases} \frac{\mathscr{L}(T)}{300} & \text{for } 0\text{K} \leq T \leq 300\text{K} \\ 0 & \text{otherwise.} \end{cases} \tag{59}$$

So the posterior is identical to the likelihood (up to a proportionality constant), at least within the range of the flat prior. Hence we can conclude that the posterior is going to be a Gaussian (just like the likelihood) and we can immediately write the 68.3% posterior range of T as $198.0\text{K} < \mu < 201.2\text{K}$. This is numerically identical to our results obtained via the MLE. However, in this case the interpretation of this interval is that "after seeing the data, and given our prior as specified in Eq. (58), there is 68.3% probability that the true value of the temperature lies within the range $198.0\text{K} < \mu < 201.2\text{K}$".
Under a change of variable, $\Psi = \Psi(\theta)$, the prior transforms according to:

$$P(\Psi) = P(\theta)\left|\det\left(\frac{\partial\theta}{\partial\Psi}\right)\right|. \tag{60}$$

In particular, a flat prior on θ is no longer flat in Ψ if the variable transformation is non-linear.

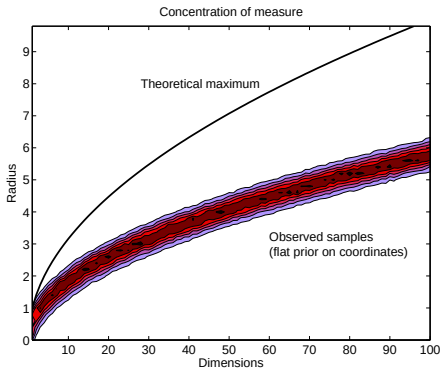

Fig. 5 Illustration of the phenomenon of the concentration of measure in parameter spaces with a large number of dimensions. The coloured band represents the density of samples (as a function of the number of dimensions sampled) obtained with a flat prior on the axis coordinates of a D-dimensional hypercube. It can be seen that samples from the prior concentrate in a thin shell of constant variance, leaving most of the parameter space unexplored. The radius of the shell is given in the vertical axis.

It is important to realize that a flat prior is far from harmless, especially in parameter spaces of high dimensionality. This is the so-called "concentration of measure" phenomenon. Sampling uniformly (i.e., with a uniform prior) along each dimension $x_i \in [0,1]$ of a D-dimensional hypercube leads to the radius $r = \left(\sum_{i=1}^{D} x_i^2\right)^{1/2}$ of the samples to concentrate around the value $\langle r \rangle = (D/3)^{1/2}$ with constant variance. As a consequence, all of the samples are found on a thin shell (see Fig. 5 for an illustration). Even worse, in D dimensions the volume of the hypercube is much larger than the volume of the hypersphere, hence most of the volume is in the corners of the hypercube which are not sampled. This means that an MCMC in D dimensions (where D is large) has a prior distribution that is far from being uniformly distributed in the volume of the hypercube – although any 2-dimensional projection will apparently belie this.

A sensitivity analysis should always be performed, i.e., change the prior in a reasonable way and assess how robust the ensuing posterior is. Unfortunately, this is seldom done in the astrophysics and cosmology literature.

There is a vast body of literature on different types of priors, when to use them and what they are good for. It is a good idea to browse the literature when faced with a new problem, as there is no point in re-inventing the wheel every time. There are essentially two schools of thought: one maintains that priors should be chosen according to subjective degree of belief; the other, that they should be selected according to some formal rule, i.e. priors should be chosen by convention. None of the two approaches is free from difficulties. To give but some relevant examples:

- *reference priors:* the idea is to define a prior so that the contribution of the data to the posterior is maximised. This is achieved by choosing a prior with maximum entropy. For example, in the case of a Gaussian likelihood this leads to the conclusion that the proper prior for the mean μ is flat on μ, while for the standard deviation σ is should be flat in $\log\sigma$ (with appropriate cutoffs of course).
- *ignorance priors:* in 1812 Laplace set forth the principle that when nothing else is known priors should be chosen so as to give equal probability to all alternatives ("the principle of indifference"). Unfortunately this is very difficult to do in the case of continuous parameters: part of the reason is that the notion of "indifference" is not invariant under non-linear reparameterizations. In some relatively simple cases, ignorance priors can be derived using symmetry or invariance arguments, see for examples [26].
- *conjugate priors:* a prior is said to be conjugate to the likelihood if the resulting posterior is of the same family as the likelihood. The convenience of having conjugate priors is that the likelihood updates the prior to a posterior which is of the same type (i.e., same distributional family). For example, Gaussian distributions are self-conjugate, i.e., a Gaussian prior with a Gaussian likelihood leads to a Gaussian posterior; the conjugate prior to both the Poisson and the exponential likelihood is the Gamma distribution; the conjugate prior to a Binomial likelihood is the Beta distribution.

3.4 A general Bayesian solution to inference problems

The general Bayesian recipe to inferential problems can be summarised as follows:

(i) Choose a model containing a set of hypotheses in the form of a vector of parameters, θ (e.g., the mass of an extra–solar planet or the abundance of dark matter in the Universe).

(ii) Specify the priors for the parameters. Priors should summarize your state of knowledge about the parameters before you consider the new data, including an relevant external source of information.

(iii) Construct the likelihood function for the measurement, which usually reflects the way the data are obtained (e.g., 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. Such nuisance parameters are included in the likelihood (with appropriate prior). If external measurements are available for the nuisance parameters, they can be incorporated either as an informative prior on them, or else as additional likelihood terms.

(iv) Obtain the posterior distribution (usually, up to an overall normalisation constant) either by analytical means or, more often, by numerical methods (see below for MCMC and nested sampling algorithms to this effect).

The posterior pdf for one parameter at the time is obtained by *marginalization*, i.e., by integrating over the uninteresting parameters. E.g., assume the the vector of parameters is given by $\theta = \{\phi, \psi\}$, then the 1D posterior pdf for ϕ alone is given by

$$p(\phi|d) \propto \int \mathscr{L}(\phi, \psi) p(\phi, \psi) \mathrm{d}\psi. \tag{61}$$

The final inference on ϕ from the posterior can then be communicated by plotting $p(\phi|d)$, with the other components marginalized over.

From an MCMC chain, one can also obtain the profile likelihood, Eq. (35), by maximising the value of the likelihood in each bin. The profile likelihood is expected to be prior-independent, as long as the scan has gathered a sufficient number of samples in the favoured region, which is in general a difficult task for multi-dimensional parameter spaces. It is also typically much more expensive to compute as it requires a much larger number of samples than the posterior.

The profile likelihood and the Bayesian posterior ask two different statistical questions of the data: the latter evaluates which regions of parameter space are most plausible in the light of the measure implied by the prior; the former singles out regions of high quality of fit, independently of their extent in parameter space, thus disregarding the possibility of them being highly fine tuned. The information contained in both is relevant and interesting, and for non-trivial parameter spaces the two different approaches do not necessarily lead to the same conclusions[6].

3.5 The Gaussian linear model

As idealised a case as it is, the Gaussian linear model is a great tool to hone your computational skills and intuition. This is because it can be solved analytically, and any numerical solution can be compared with the exact one. Furthermore, it applies in an approximate way to many cases of interest. Here we solve analytically the general problem in n dimensions. An application to the 2-dimensional case is then given in the Exercises, section 3.8.2. For a more complete discussion, see [30], where the general case is treated (including errors on the independent variable, general correlations, missing data, upper limits, selection effects and the important subject of Bayesian hierarchical modelling).

We consider the following *linear model*

$$y = F\theta + \varepsilon \tag{62}$$

where the dependent variable y is a d-dimensional vector of observations (the *data*), $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ is a vector of dimension n of unknown parameters that we wish to determine and F is a $d \times n$ matrix of known constants which specify the rela-

[6] In the archetypal case of a Gaussian likelihood and uniform prior, the posterior pdf and the profile likelihood are identical (up to a normalisation constant) and thus the question of which to choose does not arise.

tion between the input variables θ and the dependent variables y (so-called "design matrix").

In the following, we will specialize to the case where observations $y_i(x)$ are fitted with a linear model of the form $f(x) = \sum_{j=1}^{n} \theta_j X^j(x)$. Then the matrix F is given by the basis functions X^j evaluated at the locations x_i of the observations, $F_{ij} = X^j(x_i)$. Notice that the model is linear in θ_j, not necessarily in X^j, i.e. X^j can very well be a non-linear function of x.

Furthermore, ε is a d-dimensional vector of random variables with zero mean (the *noise*). We assume for simplicity that ε follows a multivariate Gaussian distribution with uncorrelated covariance matrix $C \equiv \mathrm{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_d^2)$. The likelihood function takes the form

$$p(y|\theta) = \frac{1}{(2\pi)^{d/2} \prod_j \tau_j} \exp\left[-\frac{1}{2}(b - A\theta)^t (b - A\theta)\right], \tag{63}$$

where we have defined $A_{ij} = F_{ij}/\tau_i$ and $b_i = y_i/\tau_i$ where A is a $d \times n$ matrix and b is a d-dimensional vector. This can be re-cast with some simple algebra as

$$p(y|\theta) = \mathscr{L}_0 \exp\left[-\frac{1}{2}(\theta - \theta_0)^t L (\theta - \theta_0)\right], \tag{64}$$

with the likelihood Fisher matrix L (a $n \times n$ matrix) given by

$$L \equiv A^t A \tag{65}$$

and a normalization constant

$$\mathscr{L}_0 \equiv \frac{1}{(2\pi)^{d/2} \prod_j \tau_j} \exp\left[-\frac{1}{2}(b - A\theta_0)^t (b - A\theta_0)\right]. \tag{66}$$

Here θ_0 denotes the parameter value which maximises the likelihood (i.e., the maximum likelihood value for θ), given by

$$\theta_0 = L^{-1} A^t b. \tag{67}$$

We assume as a prior pdf a multinormal Gaussian distribution with zero mean and the $n \times n$ dimensional prior Fisher information matrix P (recall that that the Fisher information matrix is the inverse of the covariance matrix), i.e.

$$p(\theta) = \frac{|P|^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\theta^t P \theta\right], \tag{68}$$

where $|P|$ denotes the determinant of the matrix P.

It can be shown that the posterior distribution for θ is given by multinormal Gaussian with Fisher information matrix $\mathscr{F}$

$$\mathscr{F} = L + P \tag{69}$$

and mean $\bar{\theta}$ given by

$$\bar{\theta} = \mathscr{F}^{-1} L \theta_0. \tag{70}$$

Finally, the model likelihood (or "Bayesian evidence", i.e., the normalizing constant in Bayes theorem) is given by

$$\begin{aligned} p(y) &= \mathscr{L}_0 \frac{|\mathscr{F}|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}\theta_0^t (L - L\mathscr{F}^{-1}L)\theta_0\right] \\ &= \mathscr{L}_0 \frac{|\mathscr{F}|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}(\theta_0^t L\theta_0 - \bar{\theta}^t \mathscr{F}\bar{\theta})\right]. \end{aligned} \tag{71}$$

3.6 Markov Chain Monte Carlo methods

3.6.1 General theory

The purpose of a Markov chain Monte Carlo algorithm is to construct a sequence of points (or "samples") in parameter space (called "a chain"). The crucial property of the chain is that the density of samples is proportional to the posterior pdf. This allows to construct a map of the posterior distribution.

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 is described by a *transition probability* $T(\theta^{(t)}, \theta^{(t+1)})$, giving the probability of moving from point $\theta^{(t)}$ to point $\theta^{(t+1)}$ in parameter space. A sufficient condition to obtain a Markov Chain is that the transition probability satisfy the *detailed balance condition*

$$p(\theta^{(t)}|d)T(\theta^{(t)}, \theta^{(t+1)}) = p(\theta^{(t+1)}|d)T(\theta^{(t+1)}, \theta^{(t)}). \tag{72}$$

This is perhaps clearer when recast as follows:

$$\frac{T(\theta^{(t)}, \theta^{(t+1)})}{T(\theta^{(t+1)}, \theta^{(t)})} = \frac{p(\theta^{(t+1)}|d)}{p(\theta^{(t)}|d)}, \tag{73}$$

i.e.ratio of the transition probabilities is inversely proportional to the ratio of the posterior probabilities at the two points.

Once samples from the posterior pdf have been gathered, obtaining Monte Carlo estimates of expectations for any function of the parameters becomes a trivial task. The posterior mean is given by

$$E[\theta] = \int P(\theta|d)\theta d\theta \approx \frac{1}{M}\sum_{t=0}^{M-1}\theta^{(t)}, \tag{74}$$

where the (approximate) equality with the mean of the samples from the MCMC follows because the samples $\theta^{(t)}$ are generated from the posterior by construction.

One can easily obtain the expectation value of any function of the parameters $f(\theta)$ as

$$E[f(\theta)] \approx \frac{1}{M}\sum_{t=0}^{M-1} f(\theta^{(t)}). \tag{75}$$

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, obtained by integrating out all other parameters from the posterior:

$$P(\theta_1|d) = \int P(\theta|d)d\theta_2 \ldots d\theta_n, \tag{76}$$

where $P(\theta_1|d)$ is the *marginal posterior* for the parameter θ_1. While this would usually require an $n-1$-dimensional integration (which can be numerically difficult), it is easily obtained from the Markov chain. 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.

A 1D 2–tail symmetric $\alpha\%$ credible region is given by the interval (for the parameter of interest) within which fall $\alpha\%$ of the samples, obtained in such a way that a fraction $(1-\alpha)/2$ of the samples lie outside the interval on either side. In the case of a 1–tail upper (lower) limit, we report the value of the quantity below (above) which $\alpha\%$ of the sample are to be found.

Credible regions for a given probability content α can be defined in an infinite number of ways. Two definitions are commonly used. The first is "symmetric credible interval" (in 1D) given above. The second definition is that of Highest Posterior Density (HPD) regions. They are obtained by starting from the maximum of the posterior and reducing the level until the desired fraction α of the posterior probability mass is included. Such a definition delimits a region so that every point inside it has by construction a higher posterior density than any point outside it. For a given probability content α, the HPD region is also the shortest interval. For a Normal 1D posterior, the HPD is identical to the symmetric credible region.

3.6.2 The Metropolis-Hastings algorithm

The simplest (and widely used) MCMC algorithm is the Metropolis-Hastings algorithm [42, 23]:

(i) Start from a random point $\theta^{(0)}$, with associated posterior probability $p_0 \equiv p(\theta^{(0)}|d)$.

(ii) Propose a candidate point $\theta^{(c)}$ by drawing from the *proposal distribution* $q(\theta^{(0)}, \theta^{(c)})$. The proposal distribution might be for example a Gaussian of fixed width σ centered around the current point. For the Metropolis algorithm (as opposed to the more general form due to Hastings), the distribution q satisfies the symmetry condition, $q(x,y) = q(y,x)$.

(iii) Evaluate the posterior at the candidate point, $p_c = p(\theta^{(c)}|d)$. Accept the candidate point with probability

$$\alpha = \min\left(\frac{p_c q(\theta^{(c)}, \theta^{(0)})}{p_0 q(\theta^{(0)}, \theta^{(c)})}, 1\right). \tag{77}$$

For the Metropolis algorithm (where q is symmetric), this simplifies to

$$\alpha = \min\left(\frac{p_c}{p_0}, 1\right). \tag{78}$$

This accept/reject step can be performed by generating a random number u from the uniform distribution $[0,1)$ and accepting the candidate sample if $u < \alpha$, and rejecting it otherwise.

(iv) If the candidate point is accepted, add it to the chain and move there. Otherwise stay at the old point (which is thus counted twice in the chain). Go back to (ii).

Notice from Eq. (78) that whenever the candidate sample has a larger posterior than the previous one (i.e., $p_c > p_0$) the candidate is always accepted. Also, in order to evaluate the acceptance function (78) only the unnormalized posterior is required, as the normalization constant drops out of the ratio. It is easy to show that the Metropolis algorithm satisfies the detailed balance condition, Eq. (72), with the transition probability given by $T(\theta^{(t)}, \theta^{(t+1)}) = q(\theta^{(t)}, \theta^{(t+1)})\alpha(\theta^{(t)}, \theta^{(t+1)})$.

Ref [19] shows that an optimal choice of the proposal distribution is such that it leads to an acceptance rate of approximately 25% (where acceptance rate is the ratio of the number of accepted jumps to the total number of likelihood evaluations). The optimal scale of the proposal distribution is approximately $2.4/\sqrt{d}$ times the scale of the target distribution, where d is the number of dimensions of the parameter space.

The choice of proposal distribution q is crucial for the efficient exploration of the posterior. If the scale of q is too small compared to the scale of the target distribution, exploration will be poor as the algorithm spends too much time locally. If instead the scale of q is too large, the chain gets stuck as it does not jump very frequently.

To improve the exploration of the target, it is advisable to run an exploratory MCMC, compute the covariance matrix from the samples, and then re-run with this covariance matrix (perhaps rescaled by a factor $2.4/\sqrt{d}$ as recommended by [19]) as the covariance of a multivariate Gaussian proposal distribution. This process can be iterated a couple of times. The affine invariant ensemble sampler proposed by [21] evolves a series of "walkers" rather than just one sampler at the time, and uses the

position of the other points in the ensemble to generate a move with greatly reduced auto-correlation length. This also largely dispenses with the need to fine-tune the proposal distribution to match the target density. An algorithm that includes a suitable parallelization of this sampling scheme is described in [18], and is implemented in a publicly available Python package, `emcee`[7].

3.6.3 Gibbs sampling

The Gibbs sampler is a particularly good choice when it is simple (and computationally non-expensive) to sample from the conditional distribution of one of the parameters at the time. It has been shown to work well in a large ($\sim 10^5$) number of dimensions.

In Gibbs sampling, each of the parameters is updated in turn by drawing the proposal distribution from the univariate conditional distribution of that variable (conditional on all the others). This is best explained in a simple example, where the parameter space is 2-dimensional and $\theta = \{x, y\}$. In order to obtain the t-th sample, one draws

$$x^{(t)} \sim p(x|y = y^{(t-1)}) \tag{79}$$

$$y^{(t)} \sim p(y|x = x^{(t)}). \tag{80}$$

$$\tag{81}$$

Notice that in the second step, when drawing y we condition on a value of x that has been updated to the latest draw of x, namely $x^{(t)}$. In the above, p denotes the target distribution, i.e. the posterior density (where we have omitted explicit conditioning on the data for ease of notation).

In a higher number of dimensions of parameter space, one always draws the k-th variable from the conditional distribution $p(\theta_k|\theta_{(-k)})$, where $\theta_{(-k)}$ denotes the vector of variables without the k-th variable.

It is perhaps slightly baffling that one can obtain samples from the joint posterior merely from knowledge of the conditional distributions (although this is not generally true). An explanation of why this is the case (under only very mild conditions) can be found in [10].

The Gibbs sampler can thus be seen as a special case of Metropolis-Hastings, with one-dimensional proposal distributions and an acceptance rate of 1.

The above can also be generalised to blocks of variables, that are all updated simultaneously conditional on all the others. In the so-called "blocked Gibbs sampler" one draws two (or more) variables simultaneously from $p(\theta_{k,j}|\theta_{(-k,j)})$. This can be useful in improving the convergence if the two variables k, j are strongly correlated. A collapsed Gibbs sampler refers to the case when one of the variables has been marginalised out in one of the sampling steps, i.e. one draws from $p(\theta_k|\theta_{(-k,j)})$, where the j-th variable has been marginalised from the joint. More sophisticated

[7] Available from: `http://dan.iel.fm/emcee`.

sampling strategies can also be employed to reduced auto-correlation and improve sampling, see [45, 45] for the Partially Collapsed Gibbs sampler, and [60] for the ancillarity-sufficiency interweaving strategy.

3.6.4 Hamiltonian Monte-Carlo

Hamiltonian Monte Carlo is particularly appealing for physicists, as it is built on the formalism of Hamiltonian dynamics (as the name implies). Only a very sketchy introduction is possible here. Refer to [44] for further details. A Python implementation of HMC can be found at: `mc-stan.org`.

The idea is to augment the vector containing the variable of interest, q (representing position), by another vector of the same dimensionality, p (representing momentum). We then define the potential energy $U(q)$ as the negative log of the unnormalized posterior we wish to sample from,

$$U(q) = -\log(\pi(q)\mathscr{L}(q)), \tag{82}$$

where $\pi(q)$ is the prior and $\mathscr{L}(q)$ the likelihood function. The Hamiltonian of this fictitious system is then given by

$$H(q,p) = K(p) + U(q) \tag{83}$$

where $K(p)$ represents kinetic energy,

$$K(p) = \sum_i \frac{p_i^2}{2m_i}. \tag{84}$$

Here, the sum runs over the dimensionality of the parameter space, and m_i are "mass values" that are chosen for convenience. If we look at the kinetic energy term as the negative log of a probability distribution, then it defines a multivariate Gaussian of 0 mean with variance along each direction given by m_i^2.

From analytical mechanics, we know that physical solutions are obtained by solving the Hamiltonian equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{85}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}. \tag{86}$$

Such solutions have the useful properties of preserving energy (i.e., $dH/dt = 0$) and conserving the phase space volume (in virtue of Liouville's theorem). Those properties are crucial in ensuring that the Hamiltonian MC (HMC) algorithm leaves the desired distribution invariant.

In order to obtain a Markov Chain from the target distribution, the Hamiltonian MC algorithm performs the following steps in each iteration:

(i) resample the momentum variables, $p_i \sim \mathcal{N}(0, m_i^2)$;
(ii) obtain a new candidate location (q_c, p_c) in phase space by evolving the system via approximate Hamiltonian dynamics (e.g. via the leapfrog method);
(iii) take a Metropolis accept/reject step at the candidate location (this is necessary as in practice numerical approximation schemes mean that the energy of the system is only approximately conserved).

The Hamiltonian dynamics preserves energy, but it changes the value of both the momentum (in step (1)) and position variables (in step (2)), thus accomplishing a large jump in the parameters of interest, namely q.

The key advantages of HMC is that it produces samples that are much less correlated than ordinary Metropolis-Hastings (in virtue of the large distance travelled via the Hamiltonian dynamics step), and that it scales well with the number of dimensions of the parameter space.

3.6.5 Importance sampling

Importance sampling is a useful technique when we want to sample from a target distribution $p(x)$ (usually the posterior), but we have samples from another distribution $q(x)$ (perhaps because the latter is simpler to sample from). In some applications, $q(x)$ could be the posterior from a certain data set, and we then want to add another data set on top of it, thus obtaining $p(x)$. As long as $p(x)$ is not too dissimilar from $q(x)$, it can be obtained by importance sampling.

The expectation value under p of any function $f(x)$ of the RV x can be written as

$$E_p[f(x)] = \int f(x)p(x)dx = \int f(x)q(x)\frac{p(x)}{q(x)}dx = E_q[\frac{p(x)}{q(x)}f(x)]. \tag{87}$$

This shows that we can obtain the expectation value under p by computing the expectation value under q but re-weighting the function of interest by the factor $p(x)/q(x)$.

In terms of the sampling estimate, we can write

$$\mu_f \approx \frac{1}{M}\frac{\sum_{i=0}^{M} w_i f(x_i)}{\sum_{i=0}^{M} w_i} \tag{88}$$

where $w_i = p(x_i)/q(x_i)$ are the importance sampling weights and $x_i \sim q(x)$. Notice that only the unnormalized values of p and q are necessary in Eq. (88), since the normalisation cancels in the ratio.

3.7 Practical and numerical issues

It is worth mentioning several important practical issues in working with MCMC methods. Poor exploration of the posterior can lead to serious mistakes in the final inference if it remains undetected – especially in high–dimensional parameter spaces with multi–modal posteriors. It is therefore important *not* to use MCMC techniques as a black box, but to run adequate tests to ensure insofar as possible that the MCMC sampling has converged to a fair representation of the posterior.

Some of the most relevant aspects are:

(i) Initial samples in the chain must be discarded, since the Markov process is not yet sampling from the equilibrium distribution (so–called *burn–in period*). The length of the burn–in period can be assessed by looking at the evolution of the posterior density as a function of the number of steps in the chain. When the chain is started at a random point in parameter space, the posterior probability will typically be small and becomes larger at every step as the chain approaches the region where the fit to the data is better. Only when the chain has moved in the neighborhood of the posterior peak the curve of the log posterior as a function of the step number flattens and the chain begins sampling from its equilibrium distribution. Samples obtained before reaching this point must be discarded, see Fig. 6

(ii) A difficult problem is presented by the assessment of *chain convergence*, which aims at establishing when the MCMC process has gathered enough samples so that the Monte Carlo estimate (75) is sufficiently accurate. Useful diagnostic tools include the Raftery and Lewis statistics [48] and the Gelman and Rubin criterion [20].

(iii) One has to bear in mind that MCMC is a *local algorithm*, which can be trapped around local maxima of the posterior density, thus missing regions of even higher posterior altogether. Considerable experimentation is sometimes required to find an implementation of the MCMC algorithm that is well suited to the exploration of the parameter space of interest. Experimenting with different algorithms (each of which has its own strength and weaknesses) is highly recommended.

(iv) Successive samples in a chain are in general correlated. Although this is not prejudicial for a correct statistical inference, it is often interesting to obtain *independent samples* from the posterior. This can be achieved by "thinning" the chain by an appropriate factor, i.e. by selecting only one sample every K. The autocorrelation is a good measure of the number of steps required before the chain has "forgotten" its previous state. It can be estimated from the MCMC samples as

$$\hat{\gamma}(k) = \frac{\sum_{i=0}^{M-k}(\theta_i - \bar{\theta})(\theta_{i+k} - \bar{\theta})}{\sum_{i=0}^{M-k}(\theta_i - \bar{\theta})^2}, \tag{89}$$

where k is called the lag and $\bar{x}$ is the sample mean (the above equation should be understood component by component if the parameter vector θ is multi-dimensional). A plot of $\hat{\gamma}$ versus lag k is called "autocorrelation function" (ACF)

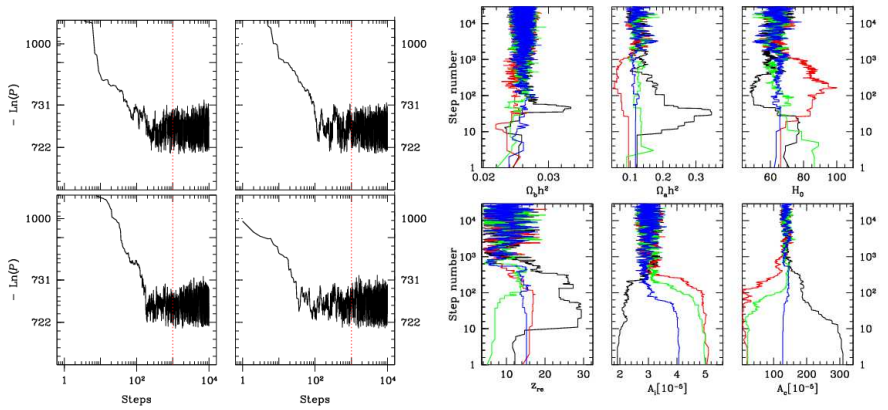

Fig. 6 Illustration of the burn-in period. Left panel: the logarithm of the log-likelihood, $-\ln P(d|\theta)$, as a function of the step number for four Monte Carlo chains. After the burn-in period (dotted, vertical lines), the value flattens and the chains are sampling from the target distribution. Right panel: the four chains (in different colors) are started in different points of a 6-dimensional parameter space and all converge to the same region after the burn-in. The vertical axis gives the number of steps.

and the value of the lag after which it drops close to 0 provides an estimate of the thinning factor K required to obtain approximate independent samples from the chain.

A discussion of samples independence and how to assess it can be found in [15], along with a convergence test based on the samples' power spectrum.

3.8 Exercices

3.8.1 Bayesian reasoning

(i) A batch of chemistry undergraduates are screened for a dangerous medical condition called *Bacillum Bayesianum* (BB). The incidence of the condition in the population (i.e., the probability that a randomly selected person has the disease) is estimated at about 1%. If the person has BB, the test returns positive 95% of the time. There is also a known 5% rate of false positives, i.e. the test returning positive even if the person is free from BB. One of your friends takes the test and it comes back positive. Here we examine whether your friend should be worried about her health.

a. Translate the information above in suitably defined conditional probabilities. The two relevant propositions here are whether the test returns positive (denote this with a + symbol) and whether the person is actually sick (denote

this with the symbol $BB = 1$. Denote the case when the person is healthy as $BB = 0$).

b. Compute the conditional probability that your friend is sick, knowing that she has tested positive, i.e., find $P(BB = 1|+)$.

c. Imagine screening the general population for a very rare desease, whose incidence in the population is 10^{-6} (i.e., one person in a million has the disease on average, i.e. $P(BB = 1) = 10^{-6}$). What should the reliability of the test (i.e., $P(+|BB = 1)$) be if we want to make sure that the probability of actually having the disease after testing positive is at least 99%? Assume first that the false positive rate $P(+|BB = 0)$ (i.e, the probability of testing positive while healthy), is 5% as in part (a). What can you conclude about the feasibility of such a test?

d. Now we write the false positive rate as $P(+|BB = 0) = 1 - P(-|BB = 0)$. It is reasonable to assume (although this is not true in general) that $P(-|BB = 0) = P(+|BB = 1)$, i.e. the probability of getting a positive result if you have the disease is the same as the probability of getting a negative result if you don't have it. Find the requested reliability of the test (i.e., $P(+|BB = 1)$) so that the probability of actually having the disease after testing positive is at least 99% in this case. Comment on whether you think a test with this reliability is practically feasible.

(ii) In a game, you can pick one of three doors, labelled A, B and C. Behind one of the three doors lies a highly desirable price, such as for example a cricket bat. After you have picked one door (e.g., door A) the person who is presenting the game opens one of the remaining 2 doors so as to reveal that there is no prize behind it (e.g., door C might be opened). Notice that the gameshow presenter *knows* that the door he opens has no prize behind it. At this point you can either stick with your original choice (door A) or switch to the door which remains closed (door B). At the end, all doors are opened, at which point you will only win if the prize is behind your chosen door.

a. Given the above rules (and your full knowledge of them), should you stick with your choice or is it better to switch?

b. In a variation, you are given the choice to randomly pick one of doors B or C and to open it, after you have chosen door A. You pick door C, and upon opening it you discover there is nothing behind it. At this point you are again free to either stick with door A or to switch to door B. Are the probabilities different from the previous scenario? Justify your answers.

(iii) In a TV debate, politician A affirms that a certain proposition S is true. You trust politician A to tell the truth with probability 4/5. Politician B then agrees that what politician A has said is indeed true. Your trust in politician B is much weaker, and you estimate that he lies with probability 3/4.
After you have heard politician B, what is the probability that statement S is indeed true?

(You may assume that you have no other information on the truth of proposition S other than what you heard from politicians A and B)
Hint: Start by denoting by A_T the statement "politician A tells the truth", and by B_T the statement "politician B tells the truth". What you are after is the probability of the statement "proposition S is true" after you have heard politician B say so.

(iv) A body has been found on the Baltimore West Side, with no apparent wounds, although it transpires that the deceased, a Mr Fuzzy Dunlop, was a heavy drug user. The detective in charge suggests to close the case and to attribute the death to drugs overdose, rather than murder.
Knowing that, of all murders in Baltimore, about 30% of the victims were drug addicts, and that the probability of a dead person having died of overdose is 50% (without further evidence apart from the body) estimate the probability that the detective's hunch is correct. (For this problem, you may assume that the possible only causes of death are overdose or murder).

3.8.2 Bayesian parameter inference

(v) This problem takes you through the steps to derive the posterior distribution for a quantity of interest θ, in the case of a Gaussian prior and Gaussian likelihood, for the 1-dimensional case.
Let us assume that we have made N independent measurements, $\hat{x} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$ of a quantity of interest θ (this could be the temperature of an object, the distance of a galaxy, the mass of a planet, etc). We assume that each of the measurements in independently Gaussian distributed with known experimental standard deviation σ. Let us denote the sample mean by $\bar{x}$, i.e.

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_i. \tag{90}$$

Before we do the experiment, our state of knowledge about the quantity of interest θ is described by a Gaussian distribution on θ (i.e., the prior entering Eq. (81) in the handout), centered around 0 (we can always choose the units in such a way that this is the case). Such a prior might come e.g. from a previous experiment we have performed. The new experiment is however much more precise, i.e. $\Sigma \gg \sigma$. Our prior state of knowledge be written in mathematical form as the following Gaussian pdf:

$$p(\theta) \sim \mathcal{N}(0, \Sigma^2). \tag{91}$$

a. Write down the likelihood function for the measurements and show that it can be recast in the form:

$$\mathcal{L}(\theta) = L_0 \exp\left(-\frac{1}{2}\frac{(\theta - \bar{x})^2}{\sigma^2/N}\right), \tag{92}$$

where L_0 is a constant that does not depend on θ.

b. By using Bayes theorem, compute the posterior probability for θ after the data have been taken into account, i.e. compute $p(\theta|\hat{x})$. Show that it is given by a Gaussian of mean $\bar{x}\frac{\Sigma^2}{\Sigma^2+\sigma^2/N}$ and variance $\left[\frac{1}{\Sigma^2}+\frac{N}{\sigma^2}\right]^{-1}$.
Hint: you may drop the normalization constant from Bayes theorem, as it does not depend on θ

c. Show that as $N \to \infty$ the posterior distribution becomes independent of the prior.

d. Show that as $N \to \infty$ the mean of the posterior distribution converges to the MLE of the mean for θ. This means that for a large number of measurements, the Bayesian result matches the frequentist MLE result.

(vi) We already encountered the coin tossing problem, but this time you'll do it in the Bayesian way.
A coin is tossed N times and heads come up H times.

a. What is the likelihood function? Identify clearly the parameter, θ, and the data.

b. What is a reasonable, non-informative prior on θ?

c. Compute the posterior probability for θ. Recall that θ is the probability that a single flip will give heads. This integral will prove useful:

$$\int_0^1 d\theta\,\theta^N(1-\theta)^M = \frac{\Gamma(N+1)\Gamma(M+1)}{\Gamma(N+M+2)}. \tag{93}$$

d. Determine the posterior mean and standard deviation of θ.

e. Plot your results as a function of H for $N = 10, 100, 1000$.

f. † Generalize your prior to the Beta distribution,

$$p(\theta|\nu_1,\nu_2) = \frac{1}{B(\nu_1,\nu_2)}\theta^{\nu_1-1}(1-\theta)^{\nu_2-1} \tag{94}$$

where $B(\nu_1,\nu_2) = \Gamma(\nu_1)\Gamma(\nu_2)/\Gamma(\nu_1+\nu_2)$ is the beta function and the "hyperparameters" $\nu_1,\nu_2 > 0$. Clearly, a uniform prior is given by the choice $(\nu_1,\nu_2) = (1,1)$. Evaluate the dependency of your result to the choice of hyperparameters.

g. † What is the probability that the $(N+1)$-th flip will give heads?

(vii) Prove Eqs. (64), (69) and (71) in the notes for the Gaussian linear model given by

$$y = F\theta + \varepsilon. \tag{95}$$

Hint: recall this standard result for Gaussian integrals:

$$\int \exp\left[-\frac{1}{2}(x-m)^t\Sigma^{-1}(x-m)\right]\mathrm{d}x = \sqrt{\det(2\pi\Sigma)} \tag{96}$$

(viii) Now we specialize to the case $n = 2$, i.e. we have two parameters of interest, $\theta = \{\theta_1, \theta_2\}$ and the linear function we want to fit is given by

$$y = \theta_1 + \theta_2 x. \tag{97}$$

(In the formalism above, the basis vectors are $X^1 = 1, X^2 = x$).
Table 1 gives an array of $d = 10$ measurements $y = \{y_1, y_2, \ldots, y_{10}\}$, together with the values of the independent variable x_i. Assume that the uncertainty in the same for all measurements, i.e. $\tau_i = 0.1$ ($i = 1, \ldots, 10$). You may further assume that measurements are uncorrelated. The data set is shown in the left panel of Fig. 7

Table 1 Data sets for the Gaussian linear model exercise. You may assume that all data points are independently and identically distributed with standard deviation of the noise $\sigma = 0.1$.

x	y
0.8308	0.9160
0.5853	0.7958
0.5497	0.8219
0.9172	1.3757
0.2858	0.4191
0.7572	0.9759
0.7537	0.9455
0.3804	0.3871
0.5678	0.7239
0.0759	0.0964

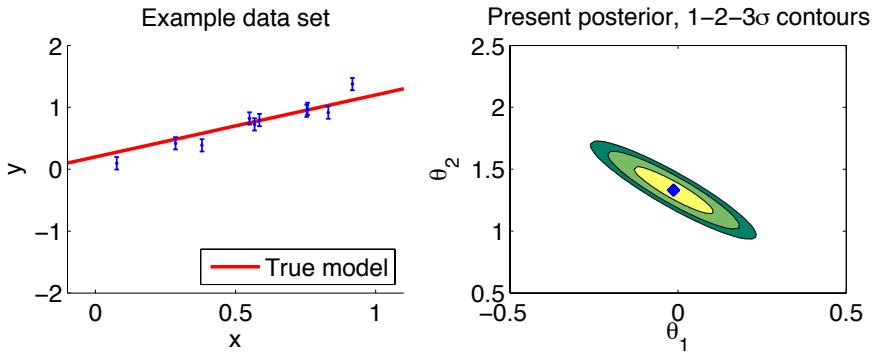

Fig. 7 Left panel: data set for the Gaussian linear problem. The solid line shows the true value of the linear model from which the data have been generated, subject to Gaussian noise. Right panel: 2D credible intervals from the posterior distribution for the parameters. The the blue diamond is the Maximum Likelihood Estimator, from Eq. (67), whose value for this data set is $x = -0.0136, y = 1.3312$.

a. Assume a Gaussian prior with Fisher matrix $P = \mathrm{diag}\left(10^{-2}, 10^{-2}\right)$ for θ.
Find the posterior distribution for θ given the data, and plot it in 2 dimensions in the (θ_1, θ_2) plane (see right panel of Fig. 7).
Use the appropriate contour levels to demarcate 1, 2 and 3 sigma joint credible intervals of the posterior.
b. In a language of your choice, write an implementation of the Metropolis-Hastings Markov Chain Monte Carlo algorithm, and use it to obtain samples from the posterior distribution.
Plot *equal weight* samples in the (θ_1, θ_2) space, as well as marginalized 1-dimensional posterior distributions for each parameter.
c. Compare the credible intervals that you obtained from the MCMC with the analytical solution.

(ix) Supernovae type Ia can be used as standardizable candles to measure distances in the Universe. This series of problems explores the extraction of cosmological information from a simplified SNIa toy model.
The cosmological parameters we are interested in constraining are

$$\mathscr{C} = \{\Omega_m, \Omega_\Lambda, h\} \tag{98}$$

where Ω_m is the matter density (in units of the critical energy density) and Ω_Λ is the dark energy density, assumed here to be in the form of a cosmological constant, i.e. $w = -1$ at all redshifts. In the following, we will fix $h = 0.72$ for simplicity, where the Hubble constant today is given by $H_0 = 100h$km/s/Mpc.
In an FRW cosmology defined by the parameters $\mathscr{C}$, the distance modulus μ (i.e., the difference between the apparent and absolute magnitudes, $\mu = m - M$) to a SN at redshift z is given by

$$\mu(z, \mathscr{C}) = 5\log\left[\frac{D_L(z, \Omega_m, \Omega_\Lambda, h)}{\mathrm{Mpc}}\right] + 25, \tag{99}$$

where D_L denotes the luminosity distance to the SN. Recalling that $D_L = cd_L/H_0$, We can rewrite this as

$$\mu(z, \mathscr{C}) = \eta + 5\log d_L(z, \Omega_m, \Omega_\Lambda), \tag{100}$$

where

$$\eta = -5\log\frac{100h}{c} + 25 \tag{101}$$

and c is the speed of light in km/s. We have defined the dimensionless luminosity distance

$$d_L(z, \Omega_m, \Omega_\Lambda) = \frac{(1+z)}{\sqrt{|\Omega_\kappa|}}\mathrm{sinn}\{\sqrt{|\Omega_\kappa|}\int_0^z \mathrm{d}z'[(1+z')^3\Omega_m + \Omega_\Lambda + (1+z')^2\Omega_\kappa]^{-1/2}\}. \tag{102}$$

The curvature parameter is given by the constraint equation

$$\Omega_\kappa = 1 - \Omega_m - \Omega_\Lambda \tag{103}$$

and the function

$$\mathrm{sinn}(x) = \begin{cases} x & \text{for a flat Universe } (\Omega_\kappa = 0); \\ \sin(x) & \text{for a closed Universe } (\Omega_\kappa < 0); \\ \sinh(x) & \text{for an open Universe } (\Omega_\kappa > 0). \end{cases} \tag{104}$$

We now assume that from each SNIa in our sample we get a measurement of the distance modulus with Gaussian noise[8], i.e., that the likelihood function for each SN i ($i = 1, \ldots, N$) is of the form

$$\mathscr{L}_i(z_i, \mathscr{C}, M) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2} \frac{(\hat{\mu}_i - \mu(z_i, \mathscr{C}))^2}{\sigma_i^2} \right). \tag{105}$$

The observed distance modulus is given by $\hat{\mu}_i = \hat{m}_i - M$, where $\hat{m}_i$ is the observed apparent magnitude and M is the intrinsic magnitude of the SNIa. We assume that each SN observation is independent of all the others.
The provided data file[9] (`SNe_simulated.dat`) contains simulated observations from the above simplified model of $N = 300$ SNIa. The two columns give the redshift z_i and the observed apparent magnitude $\hat{m}_i$. The observational error is the same for all SNe, $\sigma_i = \sigma = 0.4$ mag for $i = 1, \ldots, N$.
A plot of the data set is shown in the left panel of Fig. 8. The characteristics of the simulated SNe are designed to mimic currently available datasets (see [32, 1, 31, 49, 6]).

a. We assume that the intrinsic magnitude[10] is known and fix $M = M_0 = -19.3$ and that $h = 0.72$. We also assume that the observational error is known, given by the value above.
   Using a language of your choice, write a code to carry out an MCMC sampling of the posterior probability for $(\Omega_m, \Omega_\Lambda)$ and plot the resulting 68% and 95% posterior regions, both in 2D and marginalized to 1D, using uniform priors on $(\Omega_m, \Omega_\Lambda)$ (be careful to define them explicitly).
   You should obtain a result similar to the 2D plot shown in the right panel of Fig. 8.
b. † Add the quantity σ (the observational error) to the set of unknown parameters and estimate it from the data along with $\mathscr{C}$. Notice that since σ is a "scale parameter", the appropriate (improper) prior is $p(\sigma) \propto 1/\sigma$ (see [7] for a justification).

[8] We neglect the important issue of applying the empirical corrections known as Phillip's relations to the observed light curve. This is of fundamental important in order to reduce the scatter of SNIa within useful limits for cosmological distance measurements, but it would introduce a technical complication here without adding to the fundamental scope of this exercise.

[9] Thanks to Marisa March for help with the simulation.

[10] In reality the SNe intrinsic magnitude is not fixed, but there is an "intrinsic dispersion" (even after Phillips' corrections) reflecting perhaps intrinsic variability in the explosion mechanism, or environmental parameters which are currently poorly understood.

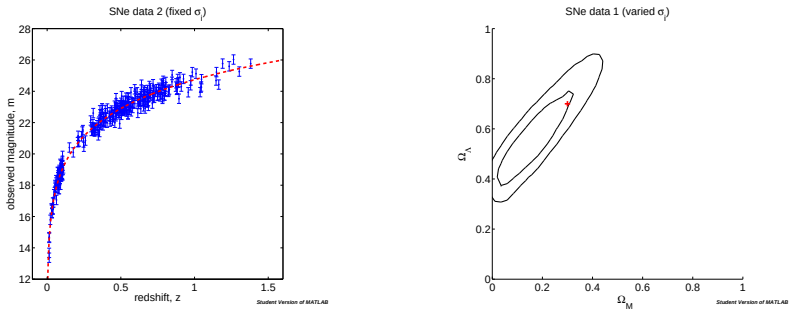

Fig. 8 Left: Simulated SNIa dataset, `SNe_simulated.dat`. The solid line is the true underlying cosmology. Right: constraints on Ω_m, Ω_Λ from this dataset, with contours delimiting 2D joint 68% and 95% credible regions (uniform priors on the variables Ω_m, Ω_Λ, assuming $M = M_0$ fixed and $h = 0.72$). The red cross denotes the true value.

c. The location of the peaks in the CMB power spectrum gives a precise measurement of the angular diameter distance to the last scattering surface, divided by the sound horizon at decoupling. This approximately translates into an effective constraint (see [52], Fig. 20) on the following degenerate combination of Ω_m and Ω_Λ:

$$1.41\Omega_\Lambda + \Omega_m = 1.30 \pm 0.04. \tag{106}$$

Add this constraint (assuming a Gaussian likelihood, with the above mean and standard deviation) to the SNIa likelihood and plot the ensuing combined 2D and 1D limits on $(\Omega_m, \Omega_\Lambda)$.

d. The measurement of the baryonic acoustic oscillation scale in the galaxy power spectrum at small redshift gives an effective constraint on the angular diameter distance D_A out to $z \sim 0.3$. This measurement can be summarized as [2]:

$$D_A(z = 0.57) = (1408 \pm 45)\text{Mpc}. \tag{107}$$

Add this constraints (again assuming a Gaussian likelihood) to the above CMB+SNIa limits and plot the resulting combined 2D and 1D limits on $(\Omega_m, \Omega_\Lambda)$.
Hint: recall that $D_L(z) = (1+z)^2 D_A(z)$.

4 Bayesian model selection

4.1 The three levels of inference

For the purpose of this discussion, it is convenient to divide Bayesian inference in three different levels:

(i) Level 1: We have chosen a model $\mathscr{M}_0$, assumed true, and we want to learn about its parameters, θ_0. E.g.: we assume ΛCDM to be the true model for the Universe and try to constrain its parameters. This is the usual parameter inference step.
(ii) Level 2: We have a series of alternative models being considered ($\mathscr{M}_1, \mathscr{M}_2, \ldots$) and we want to determine which of those is in best agreement with the data. This is a problem of model selection, or model criticism. For example, we might want to decide whether a dark energy equation of state $w = -1$ is a sufficient description of the available observations or whether we need an evolving dark energy model, $w = w(z)$.
(iii) Level 3: Of the N models considered in Level 2, there is no clear "best" model. We want to report inferences on parameters that account for this model uncertainty. This is the subject of Bayesian model averaging. For example, we want to determine Ω_m independently of the assumed dark energy model.

The Frequentist approach to model criticism is in the form of hypothesis testing (e.g., "chi-squared-per-degree-of-freedom" type of tests). One ends up rejecting (or not) a null hypothesis H_0 based on the p-value, i.e., the probability of getting data as extreme or more extreme than what has been observed if one assumes that H_0 is true. Notice that this is *not* the probability for the hypothesis! Classical hypothesis testing assumes the hypothesis to be true and determines how unlikely are our observations given this assumption. This is arguably not the quantity we are actually interested in, namely, the probability of the hypothesis itself given the observations in hand. Ref. [50] is a highly recommended read on this topic.

The Bayesian approach takes the view that there is no point in rejecting a model unless there are specific alternatives available: it takes therefore the form of model *comparison*. The key quantity for model comparison is the Bayesian evidence. Bayesian model comparison automatically implements a quantitative version of Occam's razor, i.e., the notion that simpler models ought to be preferred if they can explain the data sufficiently well.

4.2 The Bayesian evidence

4.2.1 Definition

The evaluation of a model's performance in the light of the data is based on the *Bayesian evidence*. This is the normalization integral on the right–hand–side of

Bayes' theorem, Eq. (54), which we rewrite here conditioning explicitly on the model under consideration, $\mathscr{M}$, with parameter space $\Omega_{\mathscr{M}}$:

$$p(d|\mathscr{M}) \equiv \int_{\Omega_{\mathscr{M}}} p(d|\boldsymbol{\theta},\mathscr{M})p(\boldsymbol{\theta}|\mathscr{M})\mathrm{d}\boldsymbol{\theta} \quad \text{(Bayesian evidence)}. \tag{108}$$

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:

$$p(\mathscr{M}|d) \propto p(\mathscr{M})p(d|\mathscr{M}), \tag{109}$$

where we have dropped an irrelevant normalization constant that depends only on the data and $p(\mathscr{M})$ 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, $\mathscr{M}_0$ versus $\mathscr{M}_1$, one is interested in the ratio of the posterior probabilities, or *posterior odds*, given by

$$\frac{p(\mathscr{M}_0|d)}{p(\mathscr{M}_1|d)} = B_{01}\frac{p(\mathscr{M}_0)}{p(\mathscr{M}_1)}. \tag{110}$$

Definition 8. The *Bayes factor* B_{01} is the ratio of the models' evidences:

$$B_{01} \equiv \frac{p(d|\mathscr{M}_0)}{p(d|\mathscr{M}_1)} \quad \text{(Bayes factor)}. \tag{111}$$

A value $B_{01} > (<)\ 1$ represents an increase (decrease) of the support in favour of model 0 versus model 1 given the observed data (see [29] for more details on Bayes factors).

Bayes factors are usually interpreted against the Jeffreys' scale [27] for the strength of evidence, given in Table 2. 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.

$\lvert\ln B_{01}\rvert$	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

Table 2 Empirical scale for evaluating the strength of evidence when comparing two models, $\mathscr{M}_0$ versus $\mathscr{M}_1$ (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., $p(\mathscr{M}_0) = p(\mathscr{M}_1) = 1/2$ and that the two models exhaust the model space, $p(\mathscr{M}_0|d) + p(\mathscr{M}_1|d) = 1$.

4.2.2 The Occam's razor effect

We begin by considering the example of two nested models. Consider two competing models: $\mathscr{M}_0$ predicting that a parameter $\theta = 0$ with no free parameters, and $\mathscr{M}_1$ which assigns to it 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. (111)

$$B_{01} = \sqrt{1+(\sigma/\Sigma)^{-2}}\exp\left(-\frac{\lambda^2}{2(1+(\sigma/\Sigma)^2)}\right). \tag{112}$$

For $\lambda \gg 1$, corresponding to a detection of the new parameter with high significance, the exponential term dominates and $B_{01} \ll 1$, favouring the more complex model with a non–zero extra parameter, in agreement with what one would get using Frequentist hypothesis testing. 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.

In the above example, 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

$$\ln B_{01} \approx \ln(\Sigma/\sigma) - \lambda^2/2, \tag{113}$$

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 9 shows contours of $|\ln B_{01}| = \text{const}$ for $\text{const} = 1.0, 2.5, 5.0$ in the (I_{10}, λ) plane, as computed from Eq. (113). The contours delimit significative levels for the strength of evidence, according to the Jeffreys' scale (Table 2). 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 9 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 9, 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 9, 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).

Notice that 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" (whatever that means!). In general 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.

Bayesian model selection does not penalize parameters which are unconstrained by the data. This is easily seen from Eq. (113): if a parameter is unconstrained, its posterior width σ is approximately equal to the prior width, Σ, and the Occam's razor penalty term goes to zero. In such a case, consideration of the Bayesian model complexity might help in judging model performance, see [33] for details.

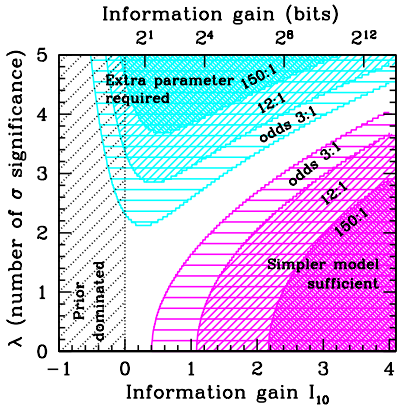

Fig. 9 Illustration of Bayesian model comparison for two nested models, where the more complex model has one extra parameter. The outcome of the model comparison depends both on the information content of the data with respect to the *a priori* available parameter space, I_{10} (horizontal axis) and on the quality of fit (vertical axis, λ, which gives the number of sigma significance of the measurement for the extra parameter). Adapted from [54].

4.2.3 Comparison with p-values

Classical hypothesis testing relies on comparing the observed value of some test statistics, $T(X)$ (where X is a RV with density $p(X|\theta)$) with its the expected distribution under a null hypothesis (usually denoted by H_0). The hypothesis test is to compare $H_0 : \theta = \theta_0$ vs an alternative $H_1 : \theta \neq \theta_0$. The test statistics is so chosen that more extreme values denote a stronger disagreement with the null.

Definition 9. The p-value (or observed significance level) is given by the probability under the null that T achieves values as extremes or more extremes that have been observed (assuming here that the larger the value of T, the stronger the disagreement):

$$\wp = p(T(X) \geq T^{\text{obs}}|H_0). \tag{114}$$

As an example, consider the case where under H_0, $x_i \sim \mathcal{N}(\theta_0, \sigma)$ for fixed θ_0 (the null hypothesis), while under the alternative H_1, $x \sim \mathcal{N}(\theta, \sigma)$ and n data samples are available (with σ known). The usual test statistics is then given by

$$T(X) = \sqrt{n}\frac{|\bar{X} - \theta_0|}{\sigma}. \tag{115}$$

The p-value is then given by

$$\wp = 2(1 - \text{erf}(T^{\text{obs}})) \tag{116}$$

where the observed value of the test statistics is

$$T^{\text{obs}} = \sqrt{n}\frac{|\bar{x} - \theta_0|}{\sigma} \tag{117}$$

and $\bar{x}$ is the sample mean.

The classical procedure of reporting the observed $\wp$ leads to a gross misrepresentation of the evidence against the null (this is in contrast with the Neyman-Person procedure of setting a threshold p-value before the experiment is performed, and then only reporting whether or not that threshold has been exceeded). This is because it *does not* obey the frequentist principle: in repeated use of a statistical procedure, the long–run average actual error should not be greater than the long–run average reported error [4]. This means that, for example, of all reported 95% confidence results, on average many more than 5% turn out to be wrong, and typically more than 50% are wrong.

Jeffreys famously criticised the use of p-values thus ([28] cited in [5]):

> I have always considered the arguments for the use of [p-values] absurd. They amount to saying that a hypothesis that may or may not be true is rejected because a greater departure from the trial value was improbable; that is, that it has not predicted something that has not happened.

An interesting illustration is given in [5]. Consider the case described above, and let us generate data from a random sequence of null hypothesis (H_0) and alternatives

(H_1), with $\theta_0 = 0$, $\sigma = 1$ and $\theta \sim \mathcal{N}(0,1)$. Suppose that the proportion of nulls and alternatives is equal. We then compute the p-value using Eq. (116) and we select all the tests that give $\wp \in [\alpha - \varepsilon, \alpha + \varepsilon]$, for a certain value of α and $\varepsilon \ll \alpha$ (the exact value of ε is unimportant). Among such results, which rejected the null hypothesis at the $1 - \alpha$ level, we then determine the proportion that actually came from the null, i.e. the percentage of wrongly rejected nulls. The results are shown in Table 3. We notice that among all the "significant" effects at the 95% level about 50% are wrong, and in general when there is only a single alternative at least 29% of the 95% confidence level results will be wrong.

p-value	sigma	fraction of true nulls	lower bound
0.05	1.96	0.51	0.29
0.01	2.58	0.20	0.11
0.001	3.29	0.024	0.018

Table 3 Proportion of wrongly rejected nulls among all results reporting a certain p-value (simulation results). The "lower bound" column gives the minimum fraction of true nulls (derived in [5]). This illustrates that the reported p-value is not equal to the fraction of wrongly rejected true nulls, which can be considerably worse.

The root of this striking disagreement with a common misinterpretation of the p-value (namely, that the p-value gives the fraction of wrongly rejected nulls in the long run) is twofold. While the p-value gives the probability of obtaining data that are as extreme or more extreme than what has actually been observed *assuming the null hypothesis is true*, one is not allowed to interpret this as the probability of the null hypothesis to be true, which is actually the quantity one is interested in assessing. The latter step requires using Bayes theorem and is therefore not defined for a frequentist. Also, quantifying how rare the observed data are under the null is not meaningful unless we can compare this number with their rareness under an alternative hypothesis.

A useful rule of thumb is obtained by [5]: it is recommended to think of a $n\sigma$ result as of a $(n-1)\sigma$ result. Reducing the number of sigma significance brings the naive p-value interpretation in better alignment with the above results. All these points are discussed in greater detail in [5, 50, 4, 38, 14].

4.3 Computation of the evidence

4.3.1 Nested sampling

A powerful and efficient alternative to classical MCMC methods has emerged in the last few years in the form of the so–called "nested sampling" algorithm, out forward by John Skilling [51]. Although the original motivation for nested sampling was to compute the evidence integral of Eq. (108), the development of the multi–modal nested sampling technique [17] (and more recently the PolyChord algorithm [22]) provides a powerful and versatile algorithm that can sample efficiently from complex, multi-modal likelihood surfaces, see Fig. 10.

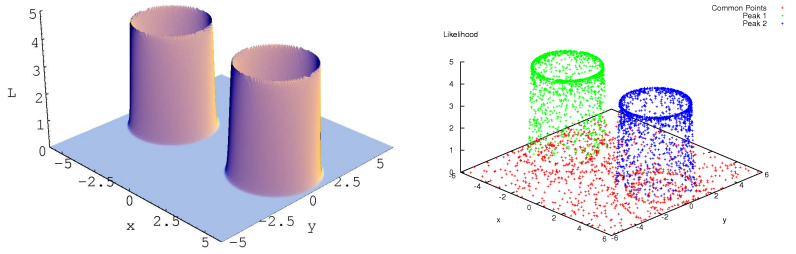

Fig. 10 Example of posterior reconstruction using Nested Sampling. Left panel: target likelihood in a 2D parameter space (x,y). Right panel: reconstructed posterior (with flat priors) using Nested Sampling. From Ref. [17].

The gist of nested sampling is that the multi–dimensional evidence integral is recast into a one–dimensional integral, by defining the prior volume X as $\mathrm{d}X \equiv p(\theta|\mathscr{M})\mathrm{d}\theta$ so that

$$X(\lambda) = \int_{\mathscr{L}(\theta)>\lambda} p(\theta|\mathscr{M})\mathrm{d}\theta \tag{118}$$

where $\mathscr{L}(\theta) \equiv p(d|\theta,\mathscr{M})$ is the likelihood function and the integral is over the parameter space enclosed by the iso–likelihood contour $\mathscr{L}(\theta) = \lambda$. So $X(\lambda)$ gives the volume of parameter space above a certain level λ of the likelihood.

The Bayesian evidence, Eq. (108), can be written as

$$p(d|\mathscr{M}) = \int_0^1 \mathscr{L}(X)\mathrm{d}X, \tag{119}$$

where $\mathscr{L}(X)$ is the inverse of Eq. (118). Samples from $\mathscr{L}(X)$ can be obtained by drawing uniformly samples from the likelihood volume within the iso–contour surface defined by λ. This is the difficult part of the algorithm.

Finally, the 1–dimensional integral of Eq. (119) can be obtained by simple quadrature, thus

$$p(d|\mathscr{M}) \approx \sum_i \mathscr{L}(X_i) W_i, \tag{120}$$

where the weights are $W_i = \frac{1}{2}(X_{i-1} - X_{i+1})$, see [51, 43] for details[11].

4.3.2 Thermodynamic integration

Thermodynamic integration computes the evidence integral by defining

$$E(\mu) \equiv \int_{\Omega_{\mathscr{M}}} \mathscr{L}(\theta)^{\mu} p(\theta|\mathscr{M}) \mathrm{d}\theta, \tag{121}$$

where μ is an annealing parameter and $\mathscr{L}(\theta) \equiv p(d|\theta, \mathscr{M})$. Obviously the desired evidence corresponds to $E(1)$. One starts by performing a standard MCMC sampling with $\mu = 0$ (i.e., sampling from the prior), then gradually increases μ to 1 according to some annealing schedule. The log of the evidence is then given by

$$\ln E(1) = \ln E(0) + \int_0^1 \frac{d \ln E}{d\mu} \mathrm{d}\mu = \int_0^1 \langle \ln \mathscr{L} \rangle_{\mu} \mathrm{d}\mu, \tag{122}$$

where the average log-likelihood is taken over the posterior with annealing parameter μ, i.e.

$$\langle \ln \mathscr{L} \rangle_{\mu} = \frac{\int_{\Omega_{\mathscr{M}}} (\ln \mathscr{L}) \mathscr{L}(\theta)^{\mu} p(\theta|\mathscr{M}) \mathrm{d}\theta}{\int_{\Omega_{\mathscr{M}}} \mathscr{L}(\theta)^{\mu} p(\theta|\mathscr{M}) \mathrm{d}\theta}. \tag{123}$$

The drawback is that the end result might depend on the annealing schedule used and that typically this methods takes 10 times as many likelihood evaluations as parameter estimation. For an overview of so-called "population Monte Carlo" algorithms and annealed importance sampling, see [25, 8].

4.3.3 Laplace approximation

An 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

$$p(d|\theta, \mathscr{M}) \approx \mathscr{L}_{\max} \exp\left[-\frac{1}{2} (\theta - \theta_{\max})^t L (\theta - \theta_{\max}) \right], \tag{124}$$

where $\theta_{\max}$ is the maximum–likelihood point, $\mathscr{L}_{\max}$ the maximum likelihood value and L the likelihood Fisher matrix (which is the inverse of the covariance matrix for

[11] Publicly available software implementing nested sampling can be found at `http://www.mrao.cam.ac.uk/software/cosmoclust/`.

the parameters). Assuming as a prior a multinormal Gaussian distribution with zero mean and Fisher information matrix P one obtains for the evidence, Eq. (108)

$$p(d|\mathscr{M}) = \mathscr{L}_{\max} \frac{|F|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}({\theta_{\max}}^t L \theta_{\max} - \overline{\theta}^t F \overline{\theta})\right], \tag{125}$$

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

4.3.4 The Savage-Dickey density ratio

A useful approximation to the Bayes factor, Eq. (111), is available for situations in which the models being compared are *nested* into each other, i.e. the more complex model ($\mathscr{M}_1$) reduces to the original model ($\mathscr{M}_0$) for specific values of the new parameters. This is a fairly common scenario when one wishes to evaluate whether the inclusion of the new parameters is supported by the data (e.g., do 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).

Writing for the extended model parameters $\theta = (\phi, \psi)$, where the simpler model $\mathscr{M}_0$ is obtained by setting $\psi = 0$, and assuming further that the prior is separable (which is usually the case), i.e. that

$$p(\phi, \psi|\mathscr{M}_1) = p(\psi|\mathscr{M}_1) p(\phi|\mathscr{M}_0), \tag{126}$$

the Bayes factor can be written in all generality as

$$B_{01} = \left.\frac{p(\psi|d, \mathscr{M}_1)}{p(\psi|\mathscr{M}_1)}\right|_{\psi=0}. \tag{127}$$

This expression is known as the Savage–Dickey density ratio (see [54] 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 $p(\psi|d, \mathscr{M}_1)$ 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.

4.3.5 Information criteria for approximate model selection

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 [34].

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

$$\mathrm{AIC} \equiv -2\ln \mathscr{L}_{\max} + 2k \tag{128}$$

where $\mathscr{L}_{\max} \equiv p(d|\theta_{\max}, \mathscr{M})$ is the maximum likelihood value.

Bayesian Information Criterion (BIC): the BIC follows from a Gaussian approximation to the Bayesian evidence in the limit of large sample size:

$$\mathrm{BIC} \equiv -2\ln \mathscr{L}_{\max} + k\ln N \tag{129}$$

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): the DIC can be written as

$$\mathrm{DIC} \equiv -2D_{\mathrm{KL}} + 2\mathscr{C}_b. \tag{130}$$

In this form, the DIC is reminiscent of the AIC, with the $\ln \mathscr{L}_{\max}$ term replaced by the estimated KL divergence D_{KL} and the number of free parameters by the effective number of parameters, $\mathscr{C}_b$ (see [55] for definitions).

The information criteria ought to be interpreted with care when applied to real situations. Comparison of Eq. (129) with Eq. (128) 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. In conclusion, what makes the information criteria attractive, namely the absence of an explicit prior specification, represents also their intrinsic limitation.

4.4 Example: model selection for the inflationary landscape

The inflationary model comparison carried out in Ref. [41, 40] is an example of the application of the above formalism to the problem of deciding which theoretical model is the best description of the available observations. Although the technical details are fairly involved, the underlying idea can be sketched as follows.

The term "inflation" describes a period of exponential expansion of the Universe in the very first instants of its life, some 10^{-32} seconds after the Big Bang, during which the size of the Universe increased by at least 25 orders of magnitude. This huge and extremely fast expansion is required to explain the observed isotropy of the cosmic microwave background on large scales. It is believed that inflation was powered by one or more scalar fields. The behaviour of the scalar field during inflation is determined by the shape of its potential, which is a real-valued function $V(\phi)$

(where ϕ denotes the value of the scalar field). The detailed shape of $V(\phi)$ controls the duration of inflation, but also the spatial distribution of inhomogeneities (perturbations) in the distribution of matter and radiation emerging from inflation. It is from those perturbations that galaxies and cluster form out of gravitational collapse. Hence the shape of the scalar field can be constrained by observations of the large scale structures of the Universe and of the CMB anisotropies.

Theories of physics beyond the Standard Model motivate certain functional forms of $V(\phi)$, which however typically have a number of free parameters, θ. The fundamental model selection question is to use cosmological observations to discriminate between alternative models for $V(\phi)$ (and hence alternative fundamental theories). The major obstacle to this programme is that very little if anything at all is known *a priori* about the free parameters θ describing the inflationary potential. What is worse, such parameters can assume values across several orders of magnitude, according to the theory. Hence the Occam's razor effect of Bayesian model comparison can vary in a very significant way depending on the prior choices for Ψ. Furthermore, a non-linear reparameterization of the problem (which leaves the physics invariant) does in general change the Occam's razor factor, and hence the model comparison result.

In Ref. [41] inflationary model selection was considered from a principled point of view. The Bayesian evidence and complexity of 198 slow-roll single-field models of inflation was computed, using the Planck 2013 Cosmic Microwave Background data. The models considered represented an almost complete and systematic scan of the entire landscape of inflationary scenarios proposed so far (More recently, this works has been extended to more complex scenarios with more than one scalar field [58]). The analysis singled out the most probable models (from an Occam's razor point of view) that are compatible with Planck data. The resulting Bayes factors (normalised to the case of Higgs Inflation) are displayed in Fig. 11.

4.5 Open challenges

I conclude by listing what I think are some of the open questions and outstanding challenges in the application of Bayesian model selection to cosmological model building.

- Is Bayesian model selection always applicable? The Bayesian model comparison approach as applied to cosmological and particle physics problems has been strongly criticized by some authors. E.g., George Efstathiou [16] and Bob Cousins [12, 13] pointed out (in different contexts) that often insufficient attention is given to the selection of models and of priors, and that this might lead to posterior model probabilities which are largely a function of one's unjustified assumptions. This draws attention to the difficult question of how to choose priors on phenomenological parameters, for which theoretical reasoning offers poor or no guidance (as in the inflationary model comparison example above).

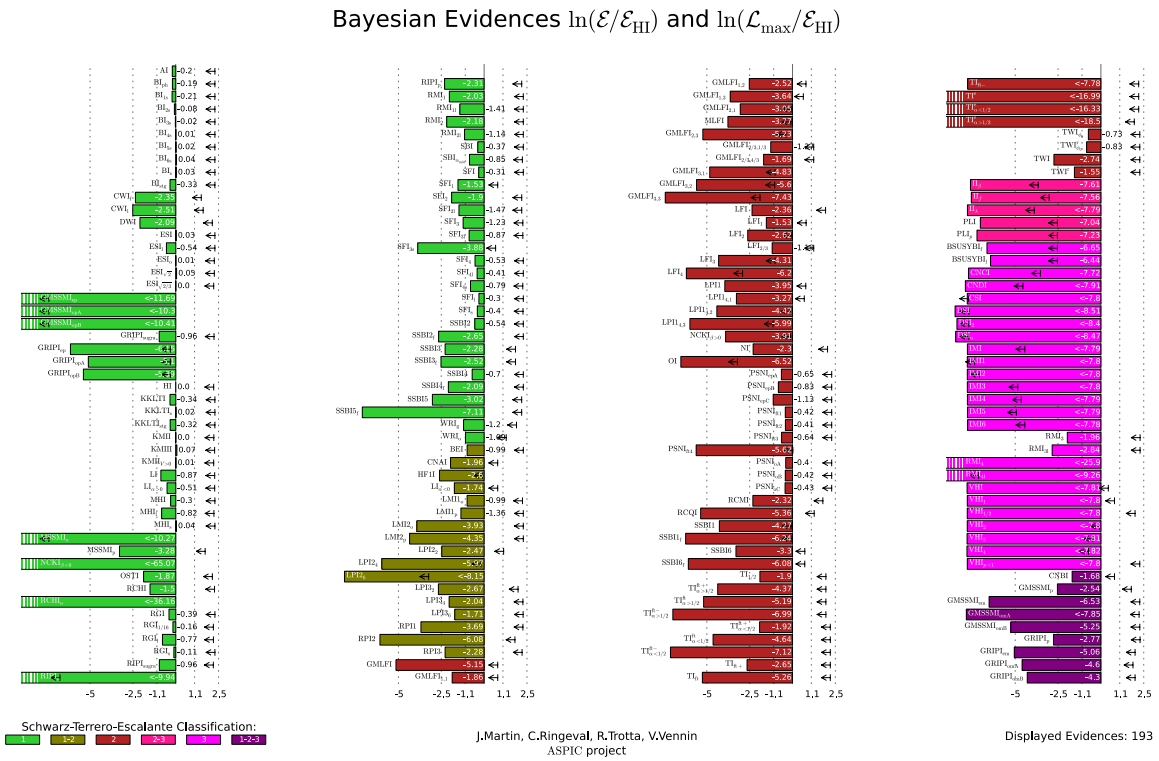

Fig. 11 Bayes factors (bars) and absolute upper bound to the Bayes factors (arrows) for inflationary scenarios, with Higgs inflation as the reference model (see [41] for further details).

- How do we deal with Lindley's paradox? It is simple to construct examples of situations where Bayesian model comparison and classical hypothesis testing disagree (Lindley's paradox [36]). This is not surprising, as frequentist hypothesis testing and Bayesian model selection really ask different questions of the data [50]. Furthermore, as the scaling with the number of data points is different, there isn't even a guarantee that the two approaches will agree in the asymptotic regime of large data sample size. As Louis Lyons aptly put it:

  > 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 [37].

However, such a disagreement is more likely to occur in situations where the signal is weak, which are precisely the kind of "frontier science" cases which are the most interesting ones (e.g., discovery claims). Is there a way to evaluate e.g. the loss function from making the "wrong" decision about rejecting/accepting a model? In this context, perhaps a decision theoretical approach would be beneficial: the loss function of making the wrong decision has to be explicitly formulated, thus helping putting the user's subjective biases and values in the open.

- How do we assess the completeness of the set of known models? Bayesian model selection always returns a best model among the ones being compared, even though that model might be a poor explanation for the available data. Is there a principled way of constructing an *absolute* scale for model performance in a Bayesian context? (for example, along the lines of the notion of Bayesian doubt, introduced in [39]).
- Is Bayesian model averaging useful? Bayesian model averaging can be used to obtain final inferences on parameters which take into account the residual model uncertainty (examples of applications in cosmology can be found in [35, 46, 57, 24]). However, it also propagates the prior sensitivity of model selection to the level of model-averaged parameter constraints. Is it useful to produce model-averaged parameter constraints, or should this task be left to the user, by providing model-specific posteriors and Bayes factors instead?
- Is there such a thing as a "correct" prior? In fundamental physics, models and parameters (and their priors) are supposed to represent (albeit in an idealized way) the real world, i.e., they are not simply useful representation of the data (as they are in other statistical problems, e.g. as applied to social sciences). In this sense, one could imagine that there exist a "correct" prior for e.g. the parameters θ of our cosmological model, which could in principle be derived from fundamental theories such as string theory (e.g., the distribution of values of cosmological parameters across the landscape of string theory [53]). This raises interesting statistical questions about the relationship between physics, reality and probability.

4.6 Exercices

(i) A coin is tossed $N = 250$ times and it returns $H = 140$ heads. Evaluate the evidence that the coin is biased using Bayesian model comparison and contrast your findings with the usual (frequentist) hypothesis testing procedure (i.e. testing the null hypothesis that $p_H = 0.5$). Discuss the dependency on the choice of priors.

(ii) In 1919 two expeditions sailed from Britain to measure the light deflection from stars behind the Sun's rim during the solar eclipse of May 29th. Einstein's General Relativity predicts a deflection angle

$$\alpha = \frac{4GM}{c^2 R},$$

where G is Newton's constant, c is the speed of light, M is the mass of the gravitational lens and R is the impact parameter. It is well known that this result it exaclty twice the value obtained using Newtonian gravity. For $M = M_\odot$ and $R = R_\odot$ one gets from Einstein's theory that $\alpha = 1.74$ arc seconds.
The team led by Eddington reported 1.61 ± 0.40 arc seconds (based on the position of 5 stars), while the team headed by Crommelin reported 1.98 ± 0.16 arc seconds (based on 7 stars).
What is the Bayes factor between Einstein and Newton gravity from those data? Comment on the strength of evidence.

(iii) Assume that the combined constraints from CMB, BAO and SNIa on the density parameter for the cosmological constant can be expressed as a Gaussian posterior distribution on Ω_Λ with mean 0.7 and standard deviation 0.05. Use the Savage-Dickey density ratio to estimate the Bayes factor between a model with $\Omega_\Lambda = 0$ (i.e., no cosmological constant) and the ΛCDM model, with a flat prior on Ω_Λ in the range $0 \leq \Omega_\Lambda \leq 2$. Comment on the strength of evidence in favour of ΛCDM.

(iv) If the cosmological constant is a manifestation of quantum fluctuations of the vacuum, QFT arguments lead to the result that the vacuum energy density ρ_Λ scales as

$$\rho_\Lambda \sim \frac{c\hbar}{16\pi} k_{\rm max}^4 \tag{131}$$

where $k_{\rm max}$ is a cutoff scale for the maximum wavenumber contributing to the energy density (see e.g. [9]). Adopting the Planck mass as a plausible cutoff scale (i.e., $k_{\rm max} = c/\hbar M_{\rm Pl}$) leads to "the cosmological constant problem", i.e., the fact that the predicted energy density

$$\rho_\Lambda \sim 10^{76} {\rm GeV}^4 \tag{132}$$

is about 120 orders of magnitude larger than the observed value, $\rho_{\rm obs} \sim 10^{-48} {\rm GeV}^4$.

a. Repeat the above estimation of the evidence in favour of a non-zero cosmological constant, adopting this time a flat prior in the range $0 \leq \Omega_\Lambda / \Omega_\Lambda^{\rm obs} < 10^{120}$. What is the meaning of this result? What is the required observational accuracy (as measured by the posterior standard deviation) required to override the Occam's razor penalty in this case?
b. It seems that it would be very difficult to create structure in a universe with $\Omega_\Lambda \gg 100$, and so life (at least life like our own) would be unlikely to evolve. How can you translate this "anthropic" argument into a quantitative statement, and how would it affect our estimate of Ω_Λ and the model selection problem?

(v) This problem follows up the cosmological parameter estimation problem from supernovae type Ia (for a more thorough treatment, see [56, 57]).

a. Adopt uniform priors $\Omega_{\rm m} \sim U(0,2)$ and $\Omega_\Lambda \sim U(0,2)$. Produce a 2D marginalised posterior pdf in the $(\Omega_{\rm m}, \Omega_\Lambda)$ plane.
b. Produce a 1D marginalised posterior pdf for the curvature parameter, $\Omega_\kappa = 1 - \Omega_\Lambda - \Omega_{\rm m}$, paying attention to normalising it to unity probability content.

What is the shape of the prior on Ω_κ implied by your choice of a uniform prior on $\Omega_{\rm m}, \Omega_\Lambda$?

c. Use the Savage-Dickey density ratio formula to estimate from the above 1D posterior the evidence in favour of a flat Universe, $\Omega_\kappa = 0$, compared with a non-flat Universe, $\Omega_\kappa \neq 0$, with prior $P(\Omega_\kappa) = U(-1,1)$.
Discuss the dependency of your result on the choice of the above prior range.

Acknowledgements I would like to thank the many colleagues who provided very valuable input and discussions over the years: Bruce Bassett, Jim Berger, Bob Cousins, Farhan Feroz, Alan Heavens, Mike Hobson, Andrew Jaffe, Martin Kunz, Andrew Liddle, Louis Lyons, John Peacock, David van Dyk. Many thanks to the Organizers of this advanced school for inviting me to present these lectures, and to the students for their piercing and stimulating questions.

Appendix

5 Some background material

5.1 The uniform, binomial and Poisson distributions

The uniform distribution: for n equiprobable outcomes between 1 and n, the uniform discrete distribution is given by

$$P(r) = \begin{cases} 1/n \text{ for } 1 \leq r \leq n \\ 0 \quad \text{otherwise} \end{cases} \tag{133}$$

It is plotted in Fig. 12 alongside with its cdf for the case of the tossing of a fair die ($n = 6$).

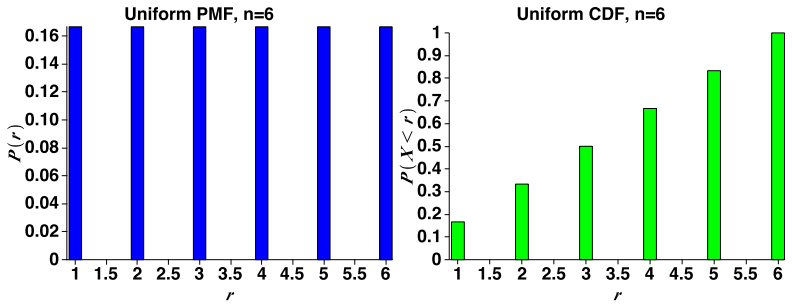

Fig. 12 Left panel: uniform discrete distribution for $n = 6$. Right panel: the corresponding cdf.

The binomial distribution: the binomial describes the probability of obtaining r "successes" in a sequence of n trials, each of which has probability p of success. Here, "success" can be defined as one specific outcome in a binary process (e.g., H/T, blue/red, 1/0, etc). The binomial distribution $B(n,p)$ is given by:

$$P(r|n,p) \equiv B(n,p) = \binom{n}{r} p^r (1-p)^{n-r}, \tag{134}$$

where the "choose" symbol is defined as

$$\binom{n}{r} \equiv \frac{n!}{(n-r)!r!} \tag{135}$$

for $0 \leq r \leq n$ (remember, $0! = 1$). Some examples of the binomial for different choices of n, p are plotted in Fig. 13.

The derivation of the binomial distribution proceeds from considering the probability of obtaining r successes in n trials (p^r), while at the same time obtaining $n-r$ failures ($(1-p)^{n-r}$). The combinatorial factor in front is derived from considerations of the number of permutations that leads to the same total number of successes.

The Poisson distribution: the Poisson distribution describes the probability of obtaining a certain number of events in a process where events occur with a fixed average rate and independently of each other. The process can occur in time (e.g., number of planes landing at Heathrow, number of photons arriving at a photomultiplier, number of murders in London, number of electrons at a detector, etc ... in a certain time interval) or in space (e.g., number of galaxies in a patch on the sky).

Let's assume that λ is the average number of events occuring per unit time or per unit length (depending on the problem being considered). Furthermore, $\lambda =$ constant in time or space.

Example 17. For example, $\lambda = 3.5$ busses/hour is the *average* number of busses passing by a particular bus stop every hour; or $\lambda = 10.3$ droplets/m^2 is the *average* number of drops of water hitting a square meter of the surface of an outdoor swimming pool in a certain day. Notice that of course at every given hour an integer number of busses actually passes by (i.e., we never observe 3 busses and one half passing by in an hour!), but that the average number can be non-integer (for example, you might have counted 7 busses in 2 hours, giving an average of 3.5 busses per hour). The same holds for the droplets of water.

For problems involving the time domain (e.g., busses/hour), the probability of r events happening in a time t is given by the Poisson distribution:

$$P(r|\lambda,t) \equiv \text{Poisson}(\lambda) = \frac{(\lambda t)^r}{r!} e^{-\lambda t}. \tag{136}$$

If the problem is about the spatial domain (e.g., droplets/m^2), the probability of r events happening in an area A is given by:

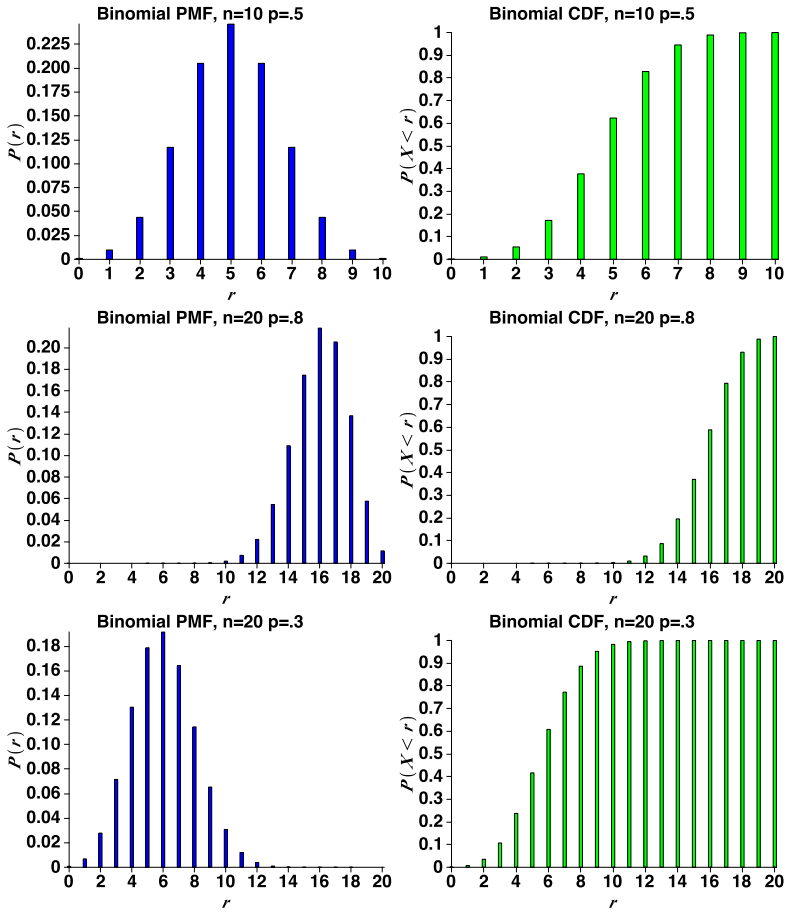

Fig. 13 Some examples of the binomial distribution, Eq. (134), for different choices of n, p, and its corresponding cdf.

$$P(r|\lambda, A) \equiv \text{Poisson}(\lambda) = \frac{(\lambda A)^r}{r!} e^{-\lambda A}. \tag{137}$$

Notice that this is a discrete pmf in the number of events r, and not a continuous pdf in t or A. The probability of getting r events in a unit time interval is obtained by setting $t = 1$ in Eq. (136); similarly, the probability of getting r events in a unit area is obtained by setting $A = 1$ in Eq. (137)

Example 18. A particle detector measures protons which are emitted with an average rate $\lambda = 4.5$/s. What is the probability of measuring 6 protons in 2 seconds? Answer:

$$P(6|\lambda = 4.5\text{s}^{-1}, t = 2\text{s}) = \frac{(4.5 \cdot 2)^6}{6!} e^{-4.5 \cdot 2} = 0.09. \tag{138}$$

So the probability is about 9%.

The Poisson distribution of Eq. (136) is plotted in Fig. 14 as a function of r for a few choices of λ (notice that in the figure $t = 1$ has been assumed, in the appropriate units). The derivation of the Poisson distribution follows from considering the

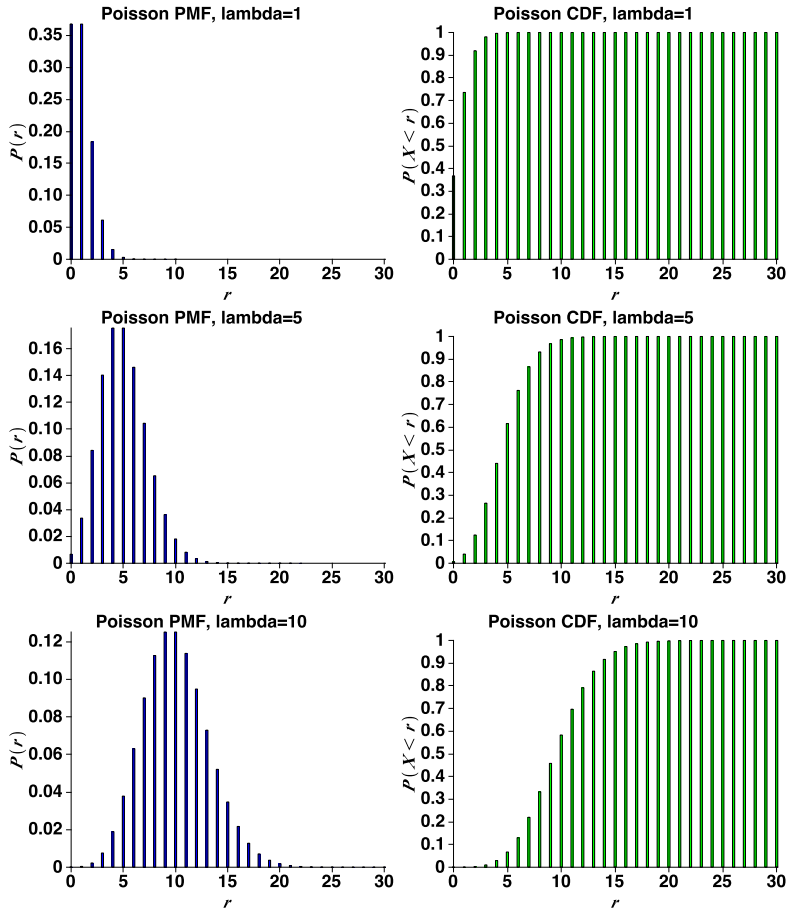

Fig. 14 Some examples of the Poisson distribution, Eq. (136), for different choices of λ, and its corresponding cdf.

probability of 1 event taking place in a small time interval Δt, then taking the limit $\Delta t \to dt \to 0$. It can also be shown that the Poisson distribution arises from the binomial in the limit $pn \to \lambda$ for $n \to \infty$, assuming $t = 1$ in the appropriate units (see lecture).

Example 19. In a post office, people arrive at the counter at an average rate of 3 customers per minute. What is the probability of 6 people arriving in a minute?

Answer: The number of people arriving follows a Poisson distribution with average $\lambda = 3$ (people/min). The probability of 6 people arriving in a minute is given by

$$P(n = 6|\lambda, t = 1\,\text{min}) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \approx 0.015 \tag{139}$$

So the probability is about 1.5%.

The discrete distributions above depend on parameters (such as p for the binomial, λ for Poisson), which control the shape of the distribution. If we know the value of the parameters, we can compute the probability of an observation (as done it the examples above). This is the subject of probability theory, which concerns itself with the theoretical properties of the distributions. The inverse problem of making inferences about the parameters from the observed samples (i.e., learning about the parameters from the observations made) is the subject of statistical inference, addressed later.

5.2 Expectation value and variance

Two important properties of distributions are the expectation value (which controls the location of the distribution) and the variance or dispersion (which controls how much the distribution is spread out). Expectation value and variance are functions of a RV.

Definition 10. The expectation value $E[X]$ (often called "mean", or "expected value"[12]) of the discrete RV X is defined as

$$E[X] = \langle X \rangle \equiv \sum_i x_i P_i. \tag{140}$$

Example 20. You toss a fair die, which follows the uniform discrete distribution, Eq. (133). What is the expectation value of the outcome?
Answer: the expectation value is given by $E[X] = \sum_i i \cdot \frac{1}{6} = 21/6$.

Definition 11. The variance or dispersion $\text{Var}(X)$ of the discrete RV X is defined as

$$\text{Var}(X) \equiv E[(X - E[X])^2] = E(X^2) - E[X]^2. \tag{141}$$

The square root of the variance is often called "standard deviation" and is usually denoted by the symbol σ, so that $\text{Var}(X) = \sigma^2$.

Example 21. For the case of tossing a fair die once, the variance is given by

$$\text{Var}(X) = \sum_i (x_i - \langle X \rangle)^2 P_i = \sum_i x_i^2 P_i - \left(\sum_i x_i P_i \right)^2 = \sum_i i^2 \frac{1}{6} - \left(\frac{21}{6} \right)^2 = \frac{105}{36}. \tag{142}$$

[12] We prefer not to use the term "mean" to avoid confusion with the sample mean.

For the binomial distribution of Eq. (134), the expectation value and variance are given by:

$$E[X] = np, \qquad \mathrm{Var}(X) = np(1-p). \tag{143}$$

Example 22. A fair coin is tossed N times. What is the expectation value for the number of heads, H? What is its variance? For $N = 10$, evaluate the probability of obtaining 8 or more heads.
Answer: The expectation values and variance are given by Eq. (143), with $p = 1/2$ (as the coin is fair), thus

$$E(H) = Np = N/2 \quad \text{and} \quad \mathrm{Var}(H) = Np(1-p) = N/4. \tag{144}$$

The probability of obtaining 8 or more heads is given by

$$P(H = 8 = \sum_{H=8}^{10} P(H\,\text{heads}|N, p = 1/2) = \frac{1}{2^{10}} \sum_{H=8}^{10} \binom{10}{H} = \frac{56}{1024} \approx 0.055. \tag{145}$$

So the probability of obtaining 8 or more heads is about 5.5%.

For the Poisson distribution of Eq. (136), the expectation value and variance are given by:

$$E[X] = \lambda t, \qquad \mathrm{Var}(X) = \lambda t, \tag{146}$$

while for the spatial version of the Poisson distribution, Eq. (137), they are given by:

$$E[X] = \lambda A, \qquad \mathrm{Var}(X) = \lambda A. \tag{147}$$

As we did above for the discrete distribution, we now define the following properties for continuous distributions.

Definition 12. The expectation value $E[X]$ of the continuous RV X with pdf $p(X)$ is defined as

$$E[X] = \langle X \rangle \equiv \int x p(x) dx. \tag{148}$$

Definition 13. The variance or dispersion $\mathrm{Var}(X)$ of the continuous RV X is defined as

$$\mathrm{Var}(X) \equiv E[(X - E[X])^2] = E(X^2) - E[X]^2 = \int x^2 p(x) dx - \left(\int x p(x) dx \right)^2. \tag{149}$$

5.3 The exponential distribution

The exponential distribution describes the time one has to wait between two consecutive events in a Poisson process, e.g. the waiting time between two radioactive particles decays. If the Poisson process happens in the spatial domain, then the exponential distribution describes the distance between two events (e.g., the separation

of galaxies in the sky). In the following, we will look at processes that happen in time (rather than in space).

To derive the exponential distribution, one can consider the arrival time of Poisson distributed events with average rate λ (for example, the arrival time particles in a detector). The probability that the first particle arrives at time t is obtained by considering the probability (which is Poisson distributed) that no particle arrives in the interval $[0,t]$, given by $P(0|\lambda,t) = \exp(-\lambda t)$ from Eq. (136), times the probability that one particle arrives during the interval $[t,t+\Delta t]$, given by $\lambda \Delta t$. Taking the limit $\Delta t \to 0$ it follows that the probability density (denoted by a symbol $p()$) for observing the first event happening at time t is given by

$$p(\text{1st event happens at time } t|\lambda) = \lambda e^{-\lambda t}, \tag{150}$$

where λ is the mean number of events per unit time. This is the exponential distribution.

Example 23. Let's assume that busses in London arrive according to a Poisson distribution, with average rate $\lambda = 5$ busses/hour. You arrive at the bus stop and a bus has just departed. What is the probability that you will have to wait more than 15 minutes?
Answer: the probability that you'll have to wait for $t_0 = 15$ minutes or more is given by

$$\int_{t_0}^{\infty} p(\text{1st event happens at time } t|\lambda)dt = \int_{t_0}^{\infty} \lambda e^{-\lambda t} dt = e^{-\lambda t_0} = 0.29, \tag{151}$$

where we have used $\lambda = 5$busses/hour $= 1/12$ busses/min.

If we have already waited for a time s for the first event to occur (and no event has occurred), then the probability that we have to wait for another time t before the first event happens satisfies

$$p(T > t+s|T > s) = p(T > t). \tag{152}$$

This means that having waited for time s without the event occuring, the time we can expect to have to wait has the same distribution as the time we have to wait from the beginning. The exponential distribution has no "memory" of the fact that a time s has already elapsed.

For the exponential distribution of Eq. (150), the expectation value and variance for the time t are given by

$$E(t) = 1/\lambda, \qquad \text{Var}(t) = 1/\lambda^2. \tag{153}$$

5.4 The Gaussian (or Normal) distribution

The Gaussian pdf (often called "the Normal distribution") is perhaps the most important distribution. It is used as default in many situations involving continuous RV (the reason becomes clear once we have studied the Central Limit Theorem, section 2.3).

The Gaussian pdf is a continuous distribution with mean μ and standard deviation σ is given by

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \tag{154}$$

and it is plotted in Fig. 15 for two different choices of $\{\mu,\sigma\}$. The Gaussian is the famous bell-shaped curve.

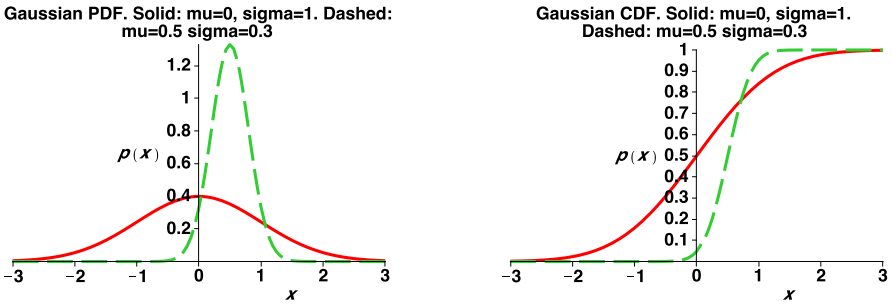

Fig. 15 Two examples of the Gaussian distribution, Eq. (154), for different choices of μ,σ, and its corresponding cdf. The expectation value μ controls the location of the pdf (i.e., when changing μ the peak moves horizontally, without changing its shape), while the standard deviation σ controls its width (i.e., when changing σ the spread of the peak changes but not its location).

For the Gaussian distribution of Eq. (154), the expectation value and variance are given by:

$$E[X] = \mu, \qquad \mathrm{Var}(X) = \sigma^2. \tag{155}$$

It can be shown that the Gaussian arises from the binomial in the limit $n \to \infty$ and from the Poisson distribution in the limit $\lambda \to \infty$. As shown in Fig. 16, the Gaussian approximation to either the binomial or the Poisson distribution is very good even for fairly moderate values of n and λ.

The probability content of a Gaussian of standard deviation σ for a given symmetric interval around the mean of width $\kappa\sigma$ on each side is given by

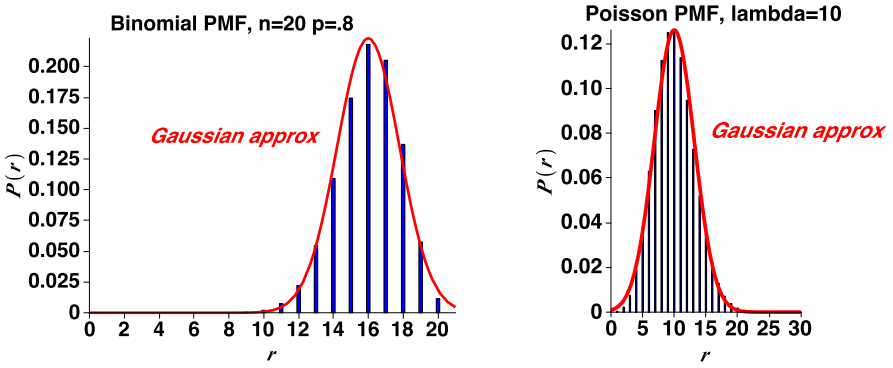

Fig. 16 Gaussian approximation to the binomial (left panel) and the Poisson distribution (right panel). The solid curve gives in each case the Gaussian approximation to each pmf.

$$P(\mu-\kappa\sigma<x<\mu+\kappa\sigma)=\int_{\mu-\kappa\sigma}^{\mu+\kappa\sigma}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)dx \tag{156}$$

$$=\frac{2}{\sqrt{\pi}}\int_0^{\kappa/\sqrt{2}}\exp\left(-y^2\right)dy \tag{157}$$

$$=\mathrm{erf}(\kappa/\sqrt{2}), \tag{158}$$

where the error function erf is defined as

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x\exp\left(-y^2\right)dy, \tag{159}$$

and can be found by numerical integration (also often tabulated and available as a built-in function in most mathematical software). Also recall the useful integral:

$$\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)dx=\sqrt{2\pi}\sigma. \tag{160}$$

Eq. (156) allows to find the probability content of the Gaussian pdf for any symmetric interval around the mean. Some commonly used values are given in Table 4.

Example 24. Measurements are often reported with the notation $T=(100\pm1)$ K (in this case, we assume we have measured a temperature, T). If nothing else is specified, it is usually implied that the error follows a Gaussian distribution. In the example above, ±1 K is the so-called "1σ interval". This means that 68.3% of the probability is contained within the range $[99,101]$ K. A "2σ interval" would have a length of 2 K on either side, so 95.4% of the probability is contained in the interval $[98,102]$ K. If one wanted a 99% interval, one would need a 2.57σ range (see Table 4). Since in this case the 1σ error is 1 K, the 2.57σ error is 2.57 K and the 99% interval is $[97.43,102.57]$ K.

κ "number of sigma"	$P(-\kappa < \frac{x-\mu}{\sigma} < \kappa)$ Probability content	Usually called
1	0.683	1σ
2	0.954	2σ
3	0.997	3σ
4	0.9993	4σ
5	$1-5.7\times10^{-7}$	5σ
1.64	0.90	90% probability interval
1.96	0.95	95% probability interval
2.57	0.99	99% probability interval
3.29	0.999	99.9% probability interval

Table 4 Relationship between the size of the interval around the mean and the probability content for a Gaussian distribution.

A heuristic derivation of how the Gaussian arises follows from this example involving darts throwing. Suppose we are throwing darts towards a target (located at the center of the coordinate system, at the position $x = 0, y = 0$), with the following rules:

(i) Throws are independent.
(ii) Errors in the x and y directions are independent.
(iii) Large errors are less probable than small ones.

The probability of a dart landing in an infinitesimal square located at coordinates (x,y) and of size $(\Delta x, \Delta y)$ (i.e., the dart landing in the interval $[x, x+\Delta x]$ and $[y, y+\Delta y]$) is given by:

$$p(x)\Delta x \cdot p(y)\Delta y = f(r)\Delta x \Delta y, \tag{161}$$

where $p(x)$ is the probability density of landing at position x (and similarly for $p(y)$), which is what we are trying to determine. On the l.h.s. of this equation, we can multiply the probabilities of landing in the x and y direction because of rule number (1) and (2). On the l.h.s., $f(r)$ is a function that only depends on the radial distance from the center, because of rule (2).

We now differentiate the above equation w.r.t. the polar coordinate ϕ:

$$\left(p(x)\frac{dp(x)}{d\phi} + p(y)\frac{dp(y)}{d\phi} \right) \Delta x \Delta y = 0. \tag{162}$$

(Note that the r.h.s. becomes 0 as it does not depend on ϕ). In polar coordinates, $x = r\cos\phi, y = r\sin\phi$, hence

$$\frac{dp(x)}{d\phi} = \frac{\partial p}{\partial x}\frac{\partial x}{\partial \phi} = -\frac{\partial p}{\partial x}y, \tag{163}$$

$$\frac{dp(y)}{d\phi} = \frac{\partial p}{\partial y}\frac{\partial y}{\partial \phi} = \frac{\partial p}{\partial y}x. \tag{164}$$

Eq. (162) becomes

$$\left(-p(x)\frac{\partial p}{\partial x}y+p(y)\frac{\partial p}{\partial y}x\right)\Delta x\Delta y=0, \tag{165}$$

which implies

$$\frac{p(x)}{x}\frac{\partial p}{\partial x}=\frac{p(y)}{y}\frac{\partial p}{\partial y}. \tag{166}$$

Since each side only depends on one of the variables, they must both equal a constant C, and we obtain the differential equation:

$$\frac{\partial p}{\partial x}=Cxp(x) \tag{167}$$

(and similarly for y). Integration gives the solution

$$p(x)=Ae^{\frac{C}{2}x^2} \tag{168}$$

and $C<0$ because of rule (3). We thus define $C=-1/\sigma^2$. Requiring that the distribution is normalized gives $A=\frac{1}{\sqrt{2\pi}\sigma}$, and therefore $p(x)$ has the shape of a Gaussian (similarly for $p(y)$).

5.5 The Chi-Square distribution

We define the RV χ^2 as the sum of the squares of n standardised independent identically distributed Gaussian RV, $x_1,\ldots,x_n$, where $x_i\sim\mathcal{N}(\mu,\sigma)$:

$$\chi^2=\sum_i^n\left(\frac{x_i-\nu}{\sigma}\right)^2 \tag{169}$$

The the RV χ^2 is distributed according to the Chi-Square distribution with n degrees of freedom,

$$p(\chi^2)=\frac{1}{\Gamma(n/2)2^{n/2}}(\chi^2)^{\frac{n}{2}-1}\exp(-\frac{1}{2}\chi^2). \tag{170}$$

For the Chi-Square distribution of Eq. (170), the expectation value and variance are given by:

$$E[X]=n \qquad \mathrm{Var}(X)=2n. \tag{171}$$

References

1. Amanullah, R., Lidman, C., Rubin, D., Aldering, G., Astier, P., et al.: Spectra and Light Curves of Six Type Ia Supernovae at 0.511 ¡ z ¡ 1.12 and the Union2 Compilation. Astrophys.J. **716**,

712–738 (2010). DOI 10.1088/0004-637X/716/1/712
2. Anderson, L., et al.: The clustering of galaxies in the SDSS-III Baryon Oscillation Spectroscopic Survey: measuring D_A and H at z = 0.57 from the baryon acoustic peak in the Data Release 9 spectroscopic Galaxy sample. Mon. Not. Roy. Astron. Soc. **439**(1), 83–101 (2014). DOI 10.1093/mnras/stt2206
3. Bayes, T., Price, R.: An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, communicated by mr. price, in a letter to john canton, m. a. and f. r. s. Phil. Trans. Roy. Soc. **53**(0), 370–418 (1763). Reproduced in: Biometrika, **45**, 293-315 (1958)
4. Berger, J.: Could fisher, jeffreys and neyman have agreed on testing? Statistical Science **18**(1), 1–12 (2003). Rejoinder: *ibid.*, 28-32
5. Berger, J., Sellke, T.: Testing a point null hypothesis: The irreconcilability of p values and evidence. J. Am. Stat. Assoc. **82**(397), 112–122 (1987). Rejoinder: *ibid.*, 135–139
6. Betoule, M., et al.: Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples. Astron. Astrophys. **568**, A22 (2014). DOI 10.1051/0004-6361/201423413
7. Box, G.E.P., Tiao, G.C.: Bayesian Inference in Statistical Analysis. John Wiley & Sons, Chicester, UK (1992)
8. Cappé, O., Guillin, A., M., M., Robert, C.: Population monte carlo. Journal of Computational and Graphical Statistics **13**(4), 907–929 (2004). URL http://www.jstor.org/stable/27594084
9. Carroll, S.M., Press, W.H., Turner, E.L.: The Cosmological constant. Ann.Rev.Astron.Astrophys. **30**, 499–542 (1992). DOI 10.1146/annurev.aa.30.090192.002435
10. Casella, G., Edwards, I.: Explaining the Gibbs sampler. Am. Stat. **46**, 167–174 (1992)
11. Chernoff, H.: On the distribution of the likelihood ratio. The Annals of Mathematical Statistics **25**, 573–578 (1954)
12. Cousins, R.D.: Comment on 'Bayesian Analysis of Pentaquark Signals from CLAS Data', with Response to the Reply by Ireland and Protopopsecu,. Phys. Rev. Lett. . **101**, 029,101 (2008)
13. Cousins, R.D.: The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics. ArXiv e-prints (2013)
14. Demortier, L., Lyons, L.: Testing Hypotheses in Particle Physics: Plots of p_0 Versus p_1. ArXiv e-prints (2014)
15. Dunkley, J., Bucher, M., Ferreira, P.G., Moodley, K., Skordis, C.: Fast and reliable MCMC for cosmological parameter estimation. Mon. Not. Roy. Astron. Soc. **356**, 925–936 (2005)
16. Efstathiou, G.: Limitations of Bayesian Evidence applied to cosmology. Mon. Not. Roy. Astron. Soc. **388**, 1314–1320 (2008). DOI 10.1111/j.1365-2966.2008.13498.x
17. Feroz, F., Hobson, M.P.: Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. Mon. Not. Roy. Astron. Soc. **384**, 449–463 (2008). DOI 10.1111/j.1365-2966.2007.12353.x
18. Foreman-Mackey, D., Hogg, D.W., Lang, D., Goodman, J.: emcee: The MCMC Hammer. Pub. Astron. Soc. Pac. **125**, 306–312 (2013). DOI 10.1086/670067
19. Gelman, A., Roberts, G., Gilks, W.: Efficient Metropolis Jumping Rules. In: J. Bernardo, J. Berger, A. Dawid, A. Smith (eds.) Bayesian statistics 5, vol. 30, pp. 599–607. Oxford University Press (1996)
20. Gelman, A., Rubin, D.: Inference from iterative simulation using multiple sequences (with discussion). Statistical Science **7**, 457–511 (1992)
21. Goodman, J., Weare, J.: Ensemble samplers with affine invariance. Comm. App. Math. Comp. Sci. **5**, 65 (2010)
22. Handley, W.J., Hobson, M.P., Lasenby, A.N.: PolyChord: nested sampling for cosmology. Mon. Not. Roy. Astron. Soc. **450**(1), L61–L65 (2015). DOI 10.1093/mnrasl/slv047
23. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
24. Hee, S., Handley, W., Hobson, M.P., Lasenby, A.N.: Bayesian model selection without evidences: application to the dark energy equation-of-state (2015). DOI 10.1093/mnras/stv2217
25. Iba, Y.: Population Monte Carlo algorithms. Transactions of the Japanese Society for Artificial Intelligence **16**, 279–286 (2001). DOI 10.1527/tjsai.16.279

26. Jaynes, E.T.: Probability Theory. The logic of science. Cambridge University Press, Cambridge, UK (2003)
27. Jeffreys, H.: Theory of probability, 3rd edn edn. Oxford Classics series (reprinted 1998). Oxford University Press, Oxford, UK (1961)
28. Jeffreys, H.: Some general points in probability theory. In: A. Zellner (ed.) Bayesian analysis in econometrics and statistics. North-Hollands, Amsterdam (1980)
29. Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Ass. **90**(430), 773–795 (1995)
30. Kelly, B.C.: Some Aspects of Measurement Error in Linear Regression of Astronomical Data. ApJ **665(2)**, 1489–1506 (2007)
31. Kessler, R., Becker, A., Cinabro, D., Vanderplas, J., Frieman, J.A., et al.: First-year Sloan Digital Sky Survey-II (SDSS-II) Supernova Results: Hubble Diagram and Cosmological Parameters. Astrophys.J.Suppl. **185**, 32–84 (2009). DOI 10.1088/0067-0049/185/1/32
32. Kowalski, M., et al.: Improved Cosmological Constraints from New, Old and Combined Supernova Datasets. Astrophys.J. **686**, 749–778 (2008). DOI 10.1086/589937
33. Kunz, M., Trotta, R., Parkinson, D.: Measuring the effective complexity of cosmological models. Phys. Rev. **D74**, 023,503 (2006)
34. Liddle, A.R.: How many cosmological parameters? Mon. Not. Roy. Astron. Soc. **351**, L49–L53 (2004)
35. Liddle, A.R., Mukherjee, P., Parkinson, D., Wang, Y.: Present and future evidence for evolving dark energy. Phys. Rev. **D74**, 123,506 (2006)
36. Lindley, D.: A statistical paradox. Biometrika **44**, 187–192 (1957)
37. Lyons, L.: A particle physicist's perspective on astrostatistics. In: Statistical Challenges in Modern Astronomy IV Conference, 371, pp. 361–372. Astronomical Society of the Pacific, San Francisco (2007)
38. Lyons, L.: Bayes and Frequentism: a Particle Physicist's perspective. Contemp. Phys. **54**, 1 (2013). DOI 10.1080/00107514.2012.756312
39. March, M., Starkman, G., Trotta, R., Vaudrevange, P.: Should we doubt the cosmological constant? Mon.Not.Roy.Astron.Soc. **410**, 2488–2496 (2011). DOI 10.1111/j.1365-2966.2010.17614.x
40. Martin, J., Ringeval, C., Trotta, R., Vennin, V.: Compatibility of Planck and BICEP2 results in light of inflation. Phys. Rev. D **90**(6), 063501 (2014). DOI 10.1103/PhysRevD.90.063501
41. Martin, J., Ringeval, C., Trotta, R., Vennin, V.: The Best Inflationary Models After Planck. JCAP **1403**, 039 (2014). DOI 10.1088/1475-7516/2014/03/039
42. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)
43. Mukherjee, P., Parkinson, D., Liddle, A.R.: A nested sampling algorithm for cosmological model selection. Astrophys. J. **638**, L51–L54 (2006)
44. Neal, R.: Mcmc using hamiltonian dynamics. In: S. Brooks, A. Gelman, G. Jones, X.L. Meng (eds.) Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC Press (2011)
45. Park, T., van Dyk, D.A.: Partially collapsed Gibbs samplers: Illustrations and applications. Journal of Computational and Graphical Statistics **18**, 283–305 (2009)
46. Parkinson, D., Liddle, A.R.: Application of Bayesian model averaging to measurements of the primordial power spectrum. Phys.Rev. **D82**, 103,533 (2010). DOI 10.1103/PhysRevD.82.103533
47. Protassov, R., van Dyk, D.A., Connors, A., Kashyap, V.L., Siemiginowska, A.: Statistics: handle with care, detecting multiple model components with the likelihood ratio test. The Astrophysical Journal **571**, 545–559 (2002)
48. Raftery, A.: Bayesian model selection in sociological research. Sociological Methodology **25**, 111–163 (1995)
49. Rest, A., et al.: Cosmological Constraints from Measurements of Type Ia Supernovae discovered during the first 1.5 yr of the Pan-STARRS1 Survey. Astrophys. J. **795**(1), 44 (2014). DOI 10.1088/0004-637X/795/1/44
50. Sellke, T., Bayarri, M., Berger, J.O.: Calibration of p values for testing precise null hypotheses. American Statistician **55**(1), 62–71 (2001)

51. Skilling, J.: Nested sampling. In: R. Fischer, R. Preuss, U. von Toussaint (eds.) Bayesian Inference and Maximum Entropy Methods in Science and Engineering, 735, pp. 395–405. Amer. Inst. Phys. conf. proc. (2004)
52. Spergel, D.N., et al.: Wilkinson Microwave Anisotropy Probe (WMAP) three year results: implications for cosmology. Astrophys. J. Suppl. **170**, 377 (2007). DOI 10.1086/513700
53. Tegmark, M.: What does inflation really predict? JCAP **0504**, 001 (2005). DOI 10.1088/1475-7516/2005/04/001
54. Trotta, R.: Applications of bayesian model selection to cosmological parameters. Mon. Not. Roy. Astron. Soc. **378**, 72–82 (2007)
55. Trotta, R.: Bayes in the sky: Bayesian inference and model selection in cosmology. Contemp. Phys. **49**, 71–104 (2008)
56. Vardanyan, M., Trotta, R., Silk, J.: How flat can you get? A model comparison perspective on the curvature of the Universe. Mon.Not.Roy.Astron.Soc. **397**, 431–444 (2009). DOI 10.1111/j.1365-2966.2009.14938.x
57. Vardanyan, M., Trotta, R., Silk, J.: Applications of Bayesian model averaging to the curvature and size of the Universe. Mon.Not.Roy.Astron.Soc. **413**, L91–L95 (2011)
58. Vennin, V., Koyama, K., Wands, D.: Encyclopdia curvatonis. JCAP **1511**(11), 008 (2015). DOI 10.1088/1475-7516/2015/11/008
59. Wilks, S.: The large-sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. pp. 60–62 (1938)
60. Yu, Y., Meng, X.L.: To center or not to center: that is not the question—An ancillarity-sufficiency interweaving strategy (ASIS) for boosting MCMC efficiency (with discussion). Journal of Computational and Graphical Statistics **20**, 531–570 (2011)