# Bayesian Inference I - AY 2024-25

Roberto Trotta

# Contents

 

# Illustrations

 

# Tables

# 1

# Symbols and notation

- Random variables are noted with upper case Latin letters, e.g., $X$.
- Observed samples from a random variable are noted with the same letter, but lowercase: e.g., $x \sim f(X)$, where $f$ is the distribution of $X$.
- An estimator for quantity $x$ is noted with a hat symbol: $\hat{x}$. Often, the symbol denotes maximum likelihood estimation.
- Scalars are in normal type, while vectors and matrices are boldface: e.g., $x \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^N$.
- Parameters are denoted by Greek letters, e.g., $\boldsymbol{\theta}$.
- The multivariate Normal distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is noted: $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. When necessary for clarity, the random variable, $\mathbf{x}$, is explicitly indicated as a subscript: $\mathcal{N}_{\mathbf{x}}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$.
- A probability distribution (either continuous or discrete) is noted $\Pr(\cdot)$.

# 2

# Fundamental notions

## 2.1 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:

1 How can we learn about regularities in the physical world given that any measurement is subject to a degree of randomness?
2 How do we quantify our uncertainty about observed properties in the world?
3 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 if 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:

1 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.
2 The discovery zone for new physics is when a potentially new effect is seen at the 2–3$\sigma$ 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.

3 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.

4 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.

5 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.

## 2.2 What is probability?

Let us start with some definitions.

**Definition 2.1** (Frequentist probability) The number of times that an event occurs over the total number of trials, in the limit of an infinite series of equiprobable events.

We note that this definition implicitly assumes that the event under consideration can be repeated an arbitrary number of times; crucially, repetitions must be *equiprobable*, in the sense that the conditions under which the event takes place must not change from one repetition to the next.

The prototypical example, often invoked as a simple illustration, is the tossing of a coin: We can repeat the toss several times and keep track of how often the coin will land, say, heads. The number of heads divided by the total number of tosses, in the limit of infinitely many tosses, is the probability of the event 'the coin lands heads'. If the coin is fair, this probability equals to 1/2. Suppose however that, after a few repetitions, the tosser gets bored and starts to slap the coin on the table instead of throwing it up in the air. This change in the 'tossing routine' entails that the events we are counting are no longer equiprobable and the frequency of heads does not define, in the limit of infinite tosses, the probability of getting heads. This example shows how the 'equiprobable' assumption is cru-

cial, yet it is only vaguely defined: at which point do repeated experiments stop being 'equiprobable', say if the hand of the person tossing the coin gets tired?

Other problematic aspects of this definition are:

1 the definition is circular (what does *equiprobable* mean?);
2 the *infinite* limit does not make sense in practice (when is the number of trials big enough?);
3 *repetitions* do not allow for unique events.

In the *Bayesian*[1] context, instead, probability is defined in more general terms, as follows.

**Definition 2.2** (Bayesian probability) A measure of the subjective degree of belief about a proposition.

Note, in particular, that propositions are not restricted to random variables, and can instead be any statement at all. Furthermore, this definition does not require repeatability of events, nor an infinite number of repetitions. Advantages of the Bayesian definition of probability include:

1 It is more general, since it encompasses cases in which repetitions can be made, as well as unique situations;
2 it covers both *epistemic uncertainty* (the kind due to incomplete knowledge about the system under study — referred to as *systematics* in physics) and *stochastic uncertainty* (i.e. intrinsic, aleatory variability — *statistical uncertainty* in physics); in fact, in the Bayesian approach there is no need to distinguish the two;
3 it covers cases in which different observers have different states of knowledge.

These advantages come at a price, though: that of introducing *prior probabilities*, as we shall see later. Furthermore, since probabilities are no longer just frequencies, how do we justify, form a Bayesian point of view, the success of this approach and the fact that probabilities do follow frequencies? In other words, how do we make contact with the Frequentist approach in the cases where the latter works well?

## 2.3 Probabilities: the Kolmogorov view

In 1933, the Russian mathematician Andrey Kolmogorov published his groundbreaking work "Foundations of the Theory of Probability,? in which he intro-

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

duced a rigorous axiomatic framework for probability theory. This work established probability as a mathematical discipline, defining it as a measure and providing a coherent set of axioms that underpin modern probability theory. Below, we introduce Frequentist probabilities following Kolmogorov now-classic definition.

Let $A, B, C, \ldots$ denote disjoint propositions. Let $\Omega$ describe the sample space (or state space) of the experiment, i.e., $\Omega$ is a list of all the possible outcomes of the experiment.

**Definition 2.3** The **joint probability** of $A$ and $B$ is the probability of $A$ and $B$ happening together, and is denoted by $\Pr(A, B)$.

**Definition 2.4** The **conditional probability** of $A$ given $B$ and denoted by $\Pr(\mid AB)$, is the probability of $A$ happening given that $B$ has happened. It is defined as

$$\Pr(\mid AB) \equiv \frac{\Pr(A, B)}{\Pr(B)}. \tag{2.1}$$

**Definition 2.5** Two propositions are said to be **independent** if and only if

$$\Pr(A, B) = \Pr(A)\Pr(B)\ . \tag{2.2}$$

**Definition 2.6** Probabilities obey the following three Kolmogorov axioms:

1 **Non-negativity**: for any $A$, $\Pr(A) \geq 0$.
2 **Normalization**: the probability for the entire sample space is unity, i.e, $\Pr(\Omega) = 1$
3 **Countable additivity**: For any countable sequence of disjoint events $A_1, A_2, \ldots$, the probability of their union is the sum of their probabilities:

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i). \tag{2.3}$$

From countable additivity, follows the **sum rule**:

$$\Pr(A) + \Pr(\neg A) = 1\ , \tag{2.4}$$

where $\neg A$ denotes the proposition "not $A$". From the definition of conditional probability, follows the **product rule**:

$$\Pr(A, B) = \Pr(\mid AB)\Pr(B)\ . \tag{2.5}$$

By combining sum and product rule, we obtain the **marginalisation rule**:

$$\Pr(A) = \Pr(A, B_1) + \Pr(A, B_2) + \cdots = \sum_i \Pr(A, B_i) = \sum_i \Pr(\mid AB_i)\Pr(B_i)\ , \tag{2.6}$$

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

By inverting the order of $A$ and $B$ in Eq. (2.5) we obtain that

$$\Pr(B, A) = \Pr(\mid B\,A)\Pr(A)\,. \tag{2.7}$$

and because $\Pr(A, B) = \Pr(B, A)$, we obtain **Bayes theorem** by equating Eqs. (2.5) and (2.7):

$$\Pr(\mid AB) = \frac{\Pr(\mid B\,A)\Pr(A)}{\Pr(B)}\,. \tag{2.8}$$

**Definition 2.7** A **random variable** (RV) is a function mapping the sample space $\Omega$ of possible outcomes of a random process to the space of real numbers.

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 2.8** 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: $\Pr(X = x_i) = P_i$ gives the probability of the RV $X$ assuming the value $x_i$. In the following we shall use the shorthand notation $\Pr(x_i)$ to mean $\Pr(X = x_i)$.

**Definition 2.9** The probability distribution associated with a continuous RV is called the **probability density function** (pdf), denoted by $\Pr(X)$. The quantity $\Pr(x)\,\mathrm{d}x$ gives the probability that the RV $X$ assumes the value between $x$ and $x + \mathrm{d}x$.

The choice of probability distribution to associate to a given random process is dictated by the nature of the random process one is investigating.

**Definition 2.10** For a discrete pmf, the **cumulative probability distribution function** (CDF) is given by

$$\mathrm{CDF}(x_k) = \sum_{i=1}^{k} \Pr(x_i)\,. \tag{2.9}$$

The $\mathrm{CDF}(x_k)$ gives the probability that the RV $X$ takes on a value less than or equal to $x_k$, i.e. $\mathrm{CDF}(x_i) = \Pr(X \leq x_i)$. For a continuous pdf, the CDF is given by

$$\mathrm{CDF}(x) = \int_{-\infty}^{x} \Pr\left(x'\right)\mathrm{d}x', \tag{2.10}$$

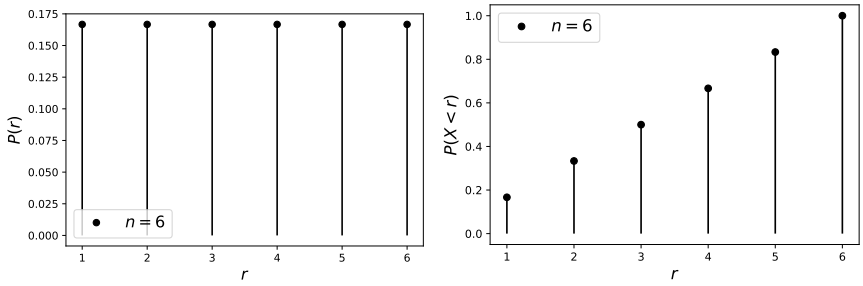


Figure 2.1 Left panel: uniform discrete distribution for $n = 6$. Right panel: the corresponding cdf.

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

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, $\{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$. Samples (i.e., measured values) are denoted with a hat ( $\hat{}$ ) symbol.

## 2.4 Some important distributions and their properties

### *2.4.1 The uniform, binomial and Poisson distributions*

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

$$P(r) = \begin{cases} 1/n & \text{for } 1 \leq r \leq n \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

It is plotted in Fig. 2.1 alongside with its cdf for the case of the tossing of a fair die ($n = 6$).

**The binomial distribution:** the binomial describes the probability of obtaining $r$ "successes" in a sequence of $n$ independent trials, each of which has prob-

[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 **??**.

ability $p$ 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:

$$P(r \mid n,p) \equiv B(n,p) = \binom{n}{r} p^r (1-p)^{n-r}, \tag{2.12}$$

where the "choose" symbol is defined as

$$\binom{n}{r} \equiv \frac{n!}{(n-r)!r!} \tag{2.13}$$

for $0 \leq r \leq n$ (remember, $0! = 1$). Some examples of the binomial for different choices of $n, p$ are plotted in Fig. 2.2.

**Example 2.11** A bent coin has a probability of landing heads $p = 0.7$. You toss it $n = 10$ times. What is the probability of getting 6 heads?
**Answer:**

$$P(6 \mid n = 10, p = 0.7) = \binom{10}{6} 0.7^6 0.3^4 = 0.2. \tag{2.14}$$

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 $\lambda$ 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 2.12** 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

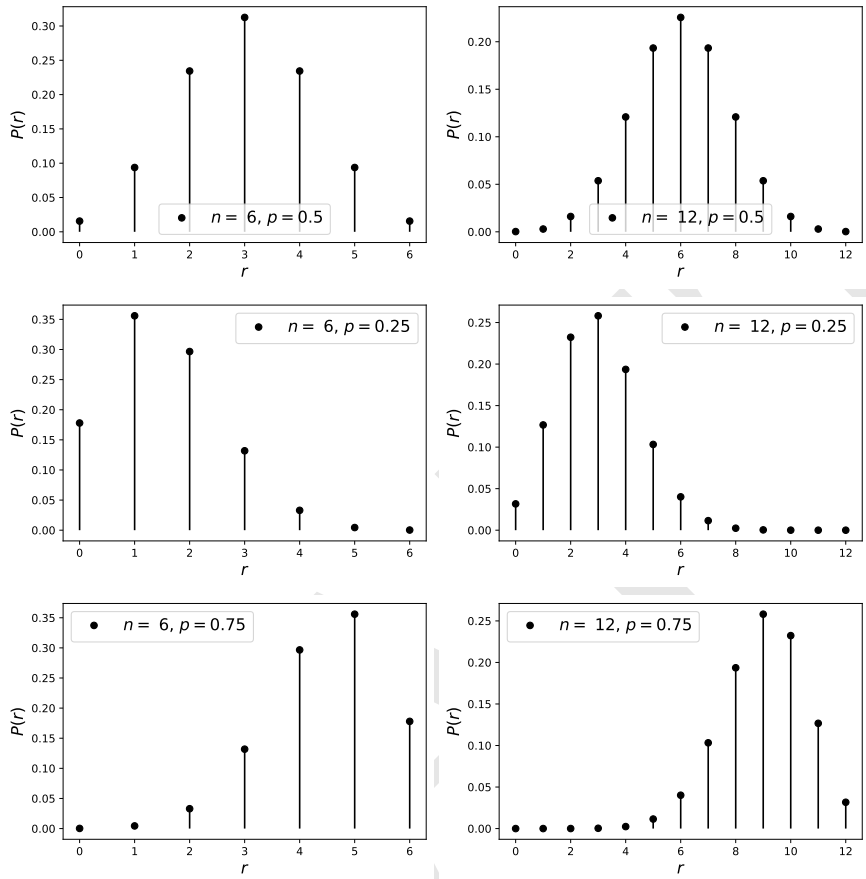


Figure 2.2 Some examples of the binomial distribution, Eq. (2.12), for different choices of $n, p$.

(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**:

$$P(r \mid \lambda, t) \equiv \text{Poisson}(\lambda) = \frac{(\lambda t)^r}{r!} e^{-\lambda t}. \tag{2.15}$$

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:

$$P(r \mid \lambda, A) \equiv \text{Poisson}(\lambda) = \frac{(\lambda A)^r}{r!} e^{-\lambda A}. \tag{2.16}$$

Notice that this is a discrete pmf in the number of events $r$, and **not** a con-

 

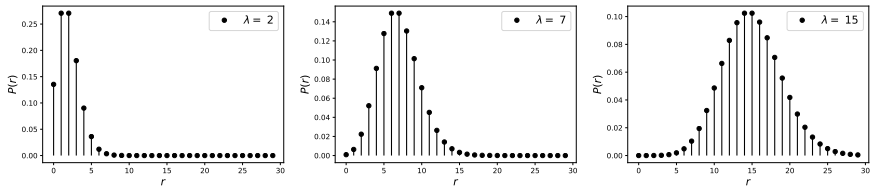


Figure 2.3 Some examples of the Poisson distribution, Eq. (2.15), for different choices of $\lambda = 2, 7, 15$ (from left to right).

tinuous pdf in $t$ or $A$. The probability of getting $r$ events in a unit time interval is obtained by setting $t = 1$ in Eq. (2.15); similarly, the probability of getting $r$ events in a unit area is obtained by setting $A = 1$ in Eq. (2.16)

**Example 2.13** A particle detector measures protons which are emitted with an average rate $\lambda = 4.5/\text{s}$. What is the probability of measuring 6 protons in 2 seconds?

**Answer:**

$$P(6 \mid \lambda = 4.5\text{s}^{-1}, t = 2\text{s}) = \frac{(4.5 \cdot 2)^6}{6!} e^{-4.5 \cdot 2} = 0.09. \tag{2.17}$$

So the probability is about 9%.

The Poisson distribution of Eq. (2.15) is plotted in Fig. 2.3 as a function of $r$ for a few choices of $\lambda$ (notice that in the figure $t = 1$ has been assumed, in the appropriate units).

The derivation of the Poisson distribution goes as follows. Consider a time interval between 0 and $t$, and divide it in $N$ small time segments, each of duration $\Delta t$. We assume that $\lambda$ is the (constant) average rate at which events happen in a unit time interval. Both $\lambda$ and $\Delta t$ need to be sufficiently small, so that in any given time interval there can only be either 0 or 1 events.

Let $P_n(t)$ denote the probability that after a time $t$, $n$ events have occurred. Since the probability of 1 event happening in a time interval $\Delta t$ is $\lambda \Delta t$ (and $\lambda \Delta t \ll 1$, since both are small), the probability of no events in the time interval $\Delta t$ is $1 - \lambda \Delta t$. Therefore the probability of no events in the total time $t$ is

$$P_0(t) = (1 - \lambda \Delta t)^N = (1 - \lambda \Delta t)^{t/\Delta t}. \tag{2.18}$$

Now we take the infinitesimal limit, making $\Delta t \to dt \to 0$ and we obtain

$$P_0(t) = (1 - \lambda \Delta t)^{t/dt} \to e^{-\lambda t} \quad (\text{for } dt \to 0). \tag{2.19}$$

We now consider the probability of $n$ events in a time $t + dt$: this is given by the probability of $n$ events in the time interval $[0, t]$ and none in the interval

$[t, t+dt]$; plus the probability of $n-1$ events in the interval $[0, t]$ and one event in the interval $[t, t+dt]$, i.e.

$$P_n(t+dt) = P_n(t)P_0(dt) + P_{n-1}(t)P_1(dt). \tag{2.20}$$

Since $dt$ is small, we can write $P_0(dt) = 1-\lambda dt$ and $P_1(dt) = \lambda dt$, thus obtaining

$$\frac{P_n(t+dt) - P_n(t)}{dt} = P_{n-1}(t)\lambda - P_n(t)\lambda. \tag{2.21}$$

Now we take the limit $dt \to 0$ and this gives a differential equation for $P_n$:

$$\frac{dP_n(t)}{dt} = \lambda[P_{n-1}(t) - P_n(t)] \tag{2.22}$$

whose solution by inspection is found to be the Poisson distribution of Eq. (2.15).

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.

**Example 2.14** 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

$$P(n = 6 \mid \lambda, t = 1\,\text{min}) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \approx 0.05 \tag{2.23}$$

So the probability is about 5%.

The discrete distributions above depend on parameters (such as $p$ for the binomial, $\lambda$ 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.

### *2.4.2 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. The **expectation value** $\mathrm{E}[X]$ (often called "mean", or "expected value"[3]) of the discrete RV $X$ is defined as

$$\mathrm{E}[X] = \langle X \rangle \equiv \sum_i x_i P_i. \tag{2.24}$$

**Example 2.15** You toss a fair die, which follows the uniform discrete distribution, Eq. (2.11). What is the expectation value of the outcome?
**Answer:** the expectation value is given by $\mathrm{E}[X] = \sum_i i \cdot \frac{1}{6} = 21/6$.

The **variance or dispersion** $\mathrm{Var}[(] X)$ of the discrete RV $X$ is defined as

$$\mathrm{Var}[X] \equiv \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2. \tag{2.25}$$

Since the expectation value is a linear operation, we can rewrite this as:

$$\mathrm{Var}[X] = \mathrm{E}\left[(X^2) - 2\mathrm{E}[X]^2 + \mathrm{E}[X]^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2. \tag{2.26}$$

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

**Example 2.16** For the case of tossing a fair die once, the variance is given by

$$\mathrm{Var}[X] = \sum_i (x_i - \langle X \rangle)^2 P_i = \sum_i x_i^2 P_i - \left(\sum_i x_i P_i\right)^2 = \sum_i i^2 \frac{1}{6} - \left(\frac{21}{6}\right)^2 = \frac{105}{36}. \tag{2.27}$$

For the binomial distribution of Eq. (2.12), the expectation value and variance are given by:

$$\mathrm{E}[X] = np, \quad \mathrm{Var}[X] = np(1-p). \tag{2.28}$$

**Example 2.17** 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. (2.28), with $p = 1/2$ (as the coin is fair), thus

$$E(H) = np = N/2 \quad \text{and} \quad \mathrm{Var}(H) = np(1-p) = n/4. \tag{2.29}$$

The probability of obtaining 8 or more heads for a fair coin ($p = 1/2$) is given by

$$P(H \geq 8 \mid p = 1/2) = \sum_{H=8}^{10} P(H \text{heads} \mid p = 1/2) = \frac{1}{2^{10}} \sum_{H=8}^{10} \binom{10}{H} = \frac{56}{1024} \approx 0.055. \tag{2.30}$$

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

As we did above for the discrete distribution, we now define the following

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

properties for continuous distributions. The **expectation value** $\mathrm{E}[X]$ of the continuous RV $X$ with pdf $p(X)$ is defined as

$$\mathrm{E}[X] = \langle X \rangle \equiv \int x \Pr(x)\, \mathrm{d}\theta. \tag{2.31}$$

The **variance** (which physicists often call "dispersion") $\mathrm{Var}[(]X)$ of the continuous RV $X$ is defined as

$$\mathrm{Var}[X] \equiv \mathrm{E}[(X - \mathrm{E}[X])]^2 = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 = \int x^2 \Pr(x)\, \mathrm{d}\theta - \left(\int x \Pr(x)\, \mathrm{d}\theta\right)^2. \tag{2.32}$$

For the Poisson distribution of Eq. (2.15), the expectation value and variance are given by:

$$\mathrm{E}[X] = \lambda t, \quad \mathrm{Var}[X] = \lambda t, \tag{2.33}$$

while for the spatial version of the Poisson distribution, Eq. (2.16), they are given by:

$$\mathrm{E}[X] = \lambda A, \quad \mathrm{Var}[X] = \lambda A. \tag{2.34}$$

A proof of Eqs. (2.28) and (2.34) is given in Appendix **??**.

### *2.4.3 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 $\lambda$ (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 $P(0 \mid \lambda, t) = \exp(-\lambda t)$ from Eq. (2.15), 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

$$\Pr\left(\text{1st event happens at time } t \mid \lambda\right) = \lambda e^{-\lambda t}, \tag{2.35}$$

where $\lambda$ is the mean number of events per unit time. This is the exponential distribution, shown in Fig. 2.4.

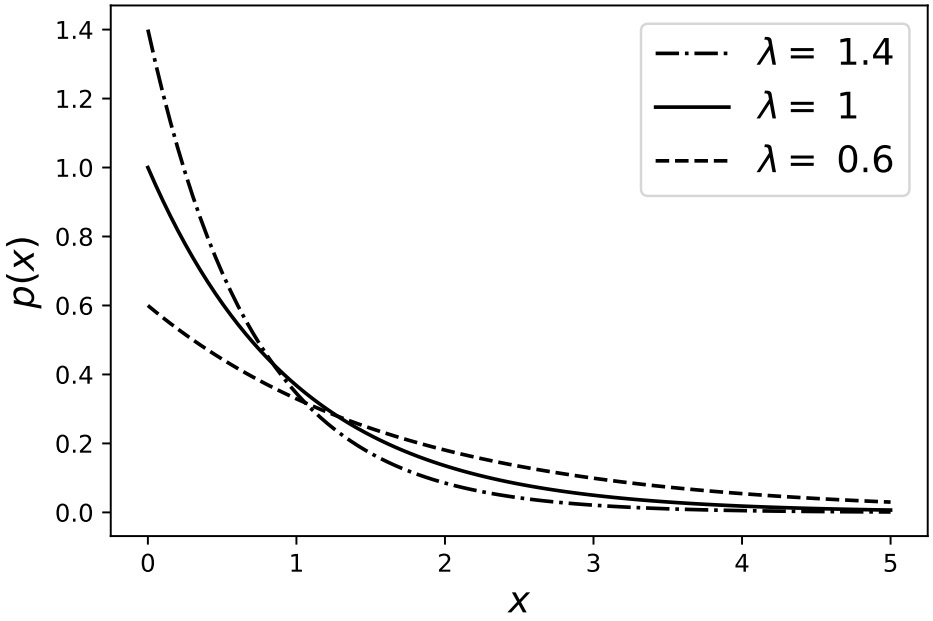


Figure 2.4 Three examples of the exponential distribution, Eq. (2.35), for different choices of $\lambda$.

**Example 2.18** 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

$$\int_{t_0}^{\infty} \Pr\big(\text{1st event happens at time } t \mid \lambda\big)\, dt = \int_{t_0}^{\infty} \lambda e^{-\lambda t} dt = e^{-\lambda t_0} = 0.29, \quad (2.36)$$

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

$$\Pr(T > t + s \mid T > s) = \Pr(T > t)\,. \quad (2.37)$$

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 (this is proved in Appendix **??**).

For the exponential distribution of Eq. (2.35), the expectation value and variance for the time $t$ are given by

$$\mathrm{E}[t] = 1/\lambda, \quad \mathrm{Var}[t] = 1/\lambda^2. \tag{2.38}$$

This is proved in the Appendix.

### *2.4.4 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, as it is the limiting distribution for the sum of many independent RV, see theorem 2.19). A heuristic derivation of how the Gaussian arises follows from the example of darts throwing (see Appendix **??**).

The Gaussian pdf is a continuous distribution with mean $\mu$ and standard deviation $\sigma$ is noted $\mathcal{N}(\mu,\sigma^2)$ and it is given by

$$\Pr(x \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \tag{2.39}$$

and it is plotted in Fig. 2.5 for two different choices of $\{\mu,\sigma\}$. The Gaussian is the famous bell-shaped curve.

For the Gaussian distribution of Eq. (2.39), the expectation value and variance are given by:

$$\mathrm{E}[X] = \mu, \quad \mathrm{Var}[X] = \sigma^2. \tag{2.40}$$

This is proven in Appendix **??**.

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. 2.6, the Gaussian approximation to either the binomial or the Poisson distribution is very good even for fairly moderate values of $n$ and $\lambda$.

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

$$\Pr(\mu-\kappa\sigma < x < \mu+\kappa\sigma) = \int_{\mu-\kappa\sigma}^{\mu+\kappa\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \mathrm{d}\theta \tag{2.41}$$

$$= \frac{2}{\sqrt{\pi}} \int_0^{\kappa/\sqrt{2}} \exp\left(-y^2\right) dy \tag{2.42}$$

$$= \mathrm{erf}(\kappa/\sqrt{2}), \tag{2.43}$$

 

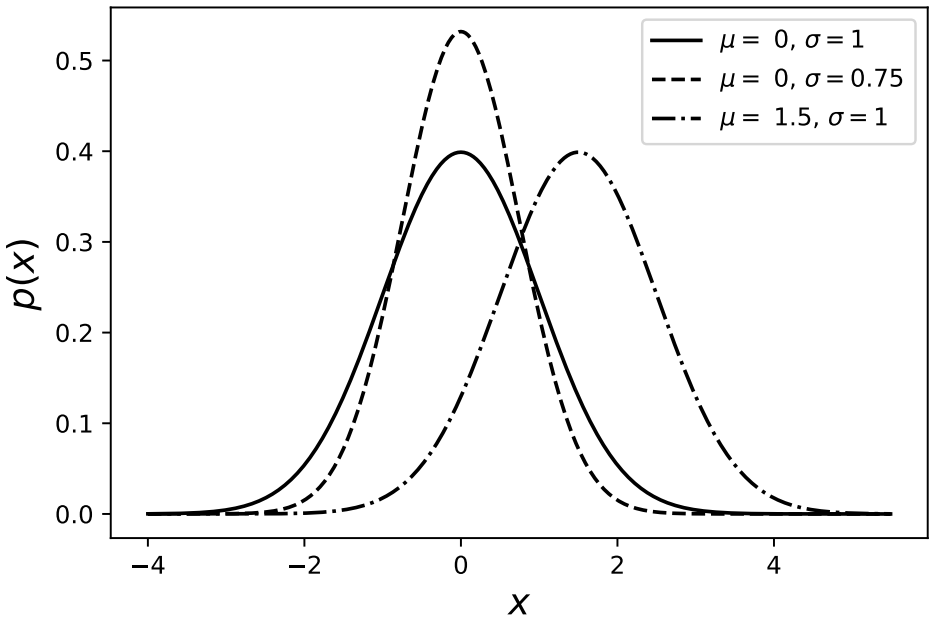


Figure 2.5 Three examples of the Gaussian distribution, Eq. (2.39), for different choices of $\mu, \sigma$. The expectation value $\mu$ controls the location of the pdf (i.e., when changing $\mu$ the peak moves horizontally, without changing its shape), while the standard deviation $\sigma$ controls its width (i.e., when changing $\sigma$ the spread of the peak changes but not its location).

where the **error function** erf is defined as

$$\mathrm{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-y^2\right) dy, \tag{2.44}$$

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:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \mathrm{d}\theta = \sqrt{2\pi}\sigma. \tag{2.45}$$

Eq. (2.41) 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 2.1.

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

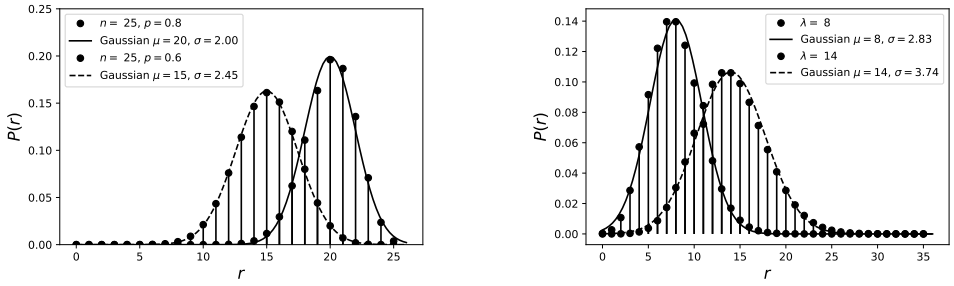


Figure 2.6 Gaussian approximation to the binomial (left panel) and the Poisson distribution (right panel). The solid curve gives in each case the Gaussian approximation to each pmf.

**Theorem 2.19** (Central Limit Theorem) *Let $X_1, X_2, \ldots, X_N$ be a collection of independent RV, each with finite expectation value $\mu_i$ and finite variance $\sigma_i^2$. Then, for $N \to \infty$, the RV*

$$Y \equiv \frac{\sum_{i=1}^{N} X_i - \sum_{i=1}^{N} \mu_i}{\sum_{i=1}^{N} \sigma_i^2} \tag{2.46}$$

*is distributed as $\mathcal{N}(0,1)$.*

The theorem is important as it says that whenever a RV arises as the sum of a large number of independent effects (e.g., noise in a temperature measurement), we can be confident that it will be very nearly Gaussian distributed. Notice that the exact shape of each distribution does not matter, as long as its variance is finite.

## 2.5 The likelihood function and the maximum likelihood principle

The problem of inference can be stated as follows: given a collection of samples, $\{x_1, x_2, \ldots, x_N\}$, and a generating random process, what can be said about the properties of the underlying probability distribution? The connection between the two domains is given by the likelihood function.

**Definition 2.20** Given a pdf or a pmf $\Pr(X|\boldsymbol{\theta})$, where $X$ represents a random variable and $\boldsymbol{\theta}$ a vector of parameters describing the shape of the pdf[4] and the observed data $\boldsymbol{d} = \{x_1, x_2, \ldots, x_N\}$, the **likelihood function** $\mathscr{L}$ (or "likelihood" for

[4] For example, for a Gaussian $\boldsymbol{\theta} = \{\mu, \sigma^2\}$ (the mean and variance of the Gaussian, respectively), for a Poisson distribution, $\theta = \lambda$ (the rate of the Poisson) and for a binomial distribution, $\theta = p$ (the probability of success in one trial).

short) is defined as

$$\mathscr{L}(\boldsymbol{\theta}) = \Pr(X = \boldsymbol{d} \,|\, \boldsymbol{\theta}). \tag{2.47}$$

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

If $\boldsymbol{d} = \{x_1, \ldots, x_N\}$ is a collection of $N$ i.i.d. observations, each of which is distributed according to the same distribution $f(x_i \,|\, \boldsymbol{\theta})$, then their joint distribution is simply the product of the individual pdfs and the joint likelihood is:

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i \,|\, \boldsymbol{\theta}). \tag{2.48}$$

**Example 2.21** As an illustration, consider the following example: in tossing a coin, let $\theta$ be the probability of obtaining heads in one throw. Suppose we make $N = 5$ flips and obtain the sequence $\boldsymbol{d} = \{H, T, T, T, T\}$, where $H(T)$ denotes the coin landing heads (tails) . The likelihood is obtained by taking the binomial distribution for obtaining $r$ successes (in this case, heads) over $N$ trials:

$$\Pr(|\, rN, \theta) = \binom{N}{r} \theta^r (1-\theta)^{N-r}, \tag{2.49}$$

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, $\theta$. Thus

$$\mathscr{L}(\theta) = \binom{5}{1} \theta^1 (1-\theta)^4 = 5\theta(1-\theta)^4, \tag{2.50}$$

which is plotted as a function of $\theta$ in the left panel of Fig. 2.7. 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 the right panel of Fig. 2.7 for a few different choices for $r$.

The above example above leads to the formulation of the Maximum Likelihood Principle: if we are trying to determine the value of $\theta$ given what we have observed (e.g., the sequence of H/T in coin tossing), the maximum likelihood principle states that we should choose the value that maximises the likelihood, because this maximises the probability of obtaining the data that we did observe. Notice that this is *not* the same as maximising the probability of $\theta$ (although in special cases the two procedures will give the same numerical result, as we shall see below). Doing so requires the use of Bayes theorem and the notion of posterior distribution.

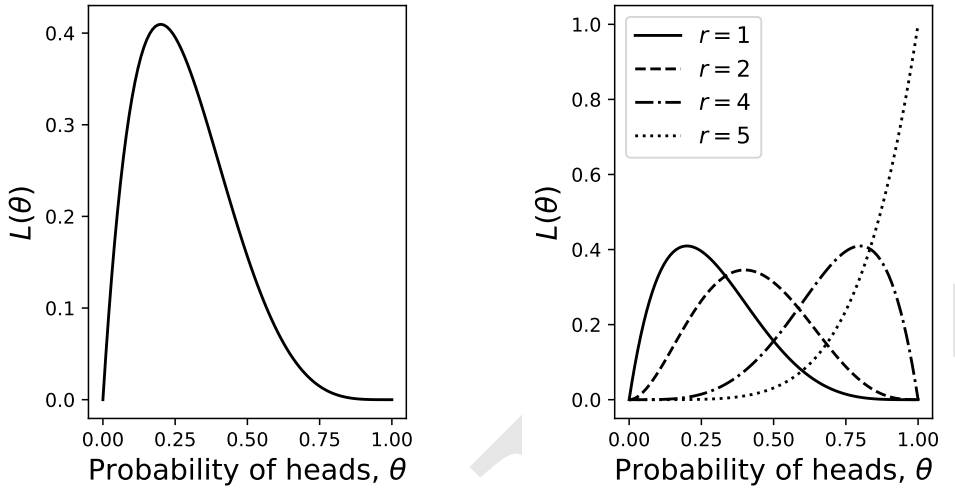


Figure 2.7 The likelihood function for the probability of heads ($\theta$) for the coin tossing example, for $N = 5, r = 1$ (left) and for $N = 5$ trials and different values of $r$ (right).

**Definition 2.22 The Maximum Likelihood Principle**: given the likelihood function $\mathscr{L}(\boldsymbol{\theta})$ and seeking to determine the parameters $\boldsymbol{\theta}$, we should choose the value of $\boldsymbol{\theta}$ in such a way that the value of the likelihood is maximised.

**Definition 2.23 The Maximum Likelihood Estimator (MLE)** for $\boldsymbol{\theta}$ is obtained from the likelihood function as:

$$\boldsymbol{\theta}_{\text{MLE}} \equiv \text{argmax}_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}) \,. \tag{2.51}$$

Under some regularity conditions[5], the MLE as defined above has the following properties, which make it useful for inference:

1 it is **consistent**, which means that, as the sample size grows it converges in probability to the true parameter value, $\boldsymbol{\theta}_0$:

$$\boldsymbol{\theta}_{\text{MLE}} \xrightarrow{\text{in P.}} \boldsymbol{\theta}_0 \quad \text{for } N \to \infty .$$

2 **asymptotically unbiased**, i.e.

$$\lim_{N\to\infty} \text{E}\left[\boldsymbol{\theta}_{\text{MLE}}\right] = \boldsymbol{\theta}_0 \,,$$

although it may be biased for finite (small) $N$, see e.g. Eq. (2.55).

[5] One important regularity condition is that the true value of the parameter must lie inside the parameter space, i.e., not on the boundary. This may be violated in situations where one is using the MLE to estimate a signal strength, bounded from below by 0, in data where none is present.

3 it tends to **a Gaussian random variable** for large sample size; more precisely,

$$\frac{\boldsymbol{\theta}_{\mathrm{MLE}} - \boldsymbol{\theta}_0}{\sqrt{\mathrm{Var}\left[\boldsymbol{\theta}_{\mathrm{MLE}}\right]}} \xrightarrow{\text{in d.}} \mathcal{N}(0,1)$$

4 it is **asymptotically efficient**, meaning that no other estimator has smaller variance than the MLE asymptotically (i.e., it saturates the Cramér-Rao bound, introduced in Eq. (3.68)).

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

**Example 2.1: MLE of the mean of a Gaussian**

We have $N$ i.i.d. measurements of a Gaussian-distributed quantity, and let's denote them by $\{x_1, x_2, \dots, x_N\}$. Here the parameters we are interested in determining are $\mu$ (the mean of the distribution) and $\sigma$ (the standard deviation of the distribution), hence we write $\theta = \{\mu, \sigma\}$.Then the joint likelihood function is given by

$$\mathscr{L}(\mu,\sigma) = p(\mathbf{x} \mid \mu,\sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right), \tag{2.52}$$

The MLE for the mean is obtained by solving

$$\frac{\partial \ln \mathscr{L}}{\partial \mu} = 0 \Rightarrow \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{2.53}$$

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

Therefore, the unbiased MLE estimator for the variance is

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_{\mathrm{ML}})^2. \tag{2.57}$$

In general, you should always use Eq. (2.57) as the ML estimator for the variance, and not Eq. (2.54).

**Example 2.2: MLE of the standard deviation of a Gaussian**

If we want to estimate the standard deviation $\sigma$ of the Gaussian, the MLE for $\sigma$ is:

$$\frac{\partial \ln \mathscr{L}}{\partial \sigma} = 0 \Rightarrow \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2. \tag{2.54}$$

However, the MLE above is biased for finite $N$, i.e. it can be shown that

$$\mathrm{E}\left[\sigma^2_{\mathrm{ML}}\right] = \sigma^2 - \mathrm{Var}\left[\mu_{\mathrm{ML}}\right] < \sigma^2, \tag{2.55}$$

I.e., for finite $N$ the expectation value of the ML estimator is not the same as the true value, $\sigma^2$. Since the variance of the mean MLE is given by:

$$\mathrm{Var}\left[\mu_{\mathrm{ML}}\right] = \mathrm{E}\left[(\mu_{\mathrm{ML}} - \mu)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}\left[x_i\right] = \frac{\sigma^2}{N}, \tag{2.56}$$

we can obtain an unbiased estimator for the variance by replacing the factor $1/N$ by $1/(N-1)$. Also, because the true $\mu$ is usually unknown, we replace it in Eq. (2.54) by the MLE estimator for the mean, $\mu_{\mathrm{ML}}$.

## 2.6 Confidence intervals

### *2.6.1 The Neyman belt construction*

The MLE gives a point estimate of what we consider to be the "best" value for the parameters, given data $\boldsymbol{d}$: namely, the value of $\boldsymbol{\theta}$ that maximises the probability of obtaining the observed data. But a point estimate is insufficient for inference. We also require a method to determine a range of values for the parameters that we consider "compatible" with the observed data. In a Frequentist context, the parameters $\boldsymbol{\theta}$ have an unknown but fixed value, so it makes no sense to speak of "the probability for the parameters" – this requires a Bayesian approach. Instead, we must consider the properties of the sampling distribution $\Pr(|\, \boldsymbol{d}\boldsymbol{\theta})$, which is a distribution over the *data*, for any given value of $\boldsymbol{\theta}$.

For simplicity, we consider a one-dimensional case, where both the parameter $\theta$ and the data $x$ are real scalar. Let $x$ be an observable random variable with a probability density (or mass) function $f(x \mid \theta)$, where $\theta$ is an unknown parameter. Our goal is to construct a confidence interval for $\theta$ based on the observed value of $x$, denoted $\boldsymbol{x}^{\mathrm{obs}}$. For a chosen value of $0 < \alpha < 1$, the **Neyman belt construction** produces a **confidence interval** at the $1-\alpha$ **confidence level**, denoted by $[\theta_{\min}(\boldsymbol{x}^{\mathrm{obs}}), \theta_{\max}(\boldsymbol{x}^{\mathrm{obs}})]$ with the property that the true (unknown) value of $\theta$ is contained within the collection of such intervals (over repeated experiments)

**Example 2.3: MLE for the 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. (2.12), where the unknown parameter is $\theta$ (the success probability for one trial, i.e., the probability that the coin lands H):

$$\mathscr{L}(\theta) = P(H \mid \theta, N) = \binom{N}{H} \theta^H (1-\theta)^{N-H}, \tag{2.58}$$

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

$$\begin{aligned} \frac{\partial \ln \mathscr{L}(\theta)}{\partial \theta} &= \frac{\partial}{\partial \theta}\left(\ln \binom{N}{H} + H \ln \theta + (N-H)\ln(1-\theta)\right) = \frac{H}{\theta} - \frac{N-H}{1-\theta} \stackrel{!}{=} 0 \\ &\Leftrightarrow \theta_{\text{MLE}} = \frac{H}{N}. \end{aligned} \tag{2.59}$$

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

with probability $1-\alpha$. In other words, the value of $\alpha$ represents the proportion of times, over repeated experiments, that the true value of $\theta$ is not contained in (i.e., not "covered" by) the confidence interval. This is the concept of "**coverage**", which defines the classical confidence interval.

The confidence level $1-\alpha$ represents the proportion of times, in repeated experiments, that the constructed confidence intervals will contain the true value of $\theta$. Common choices for $\alpha$ are 0.05 (95% confidence level) or 0.01 (99% confidence level). This level is fixed at the beginning of the procedure. In the next step, for each fixed value of $\theta$, we construct a region in $x$-space that contains the central $1-\alpha$ probability under the distribution $f(x \mid \theta)$. A common choice is to use the highest probability density (HPD) interval (see Eq. (3.33)), which captures the region where $f(x \mid \theta)$ is largest, but a symmetric interval can also be adopted.

Let $F(x \mid \theta)$ be the CDF of $f(x \mid \theta)$ (as defined in Eq. (2.10)). We find $x_{\min}(\theta)$ and $x_{\max}(\theta)$ such that:

$$F(x_{\max}(\theta) \mid \theta) - F(x_{\min}(\theta) \mid \theta) = 1 - \alpha.$$

**Example 2.4: MLE for the rate of a Poisson distribution**

The likelihood function is given by Eq. (2.15), using the notation $\theta = \lambda$ (i.e., the parameter $\theta$ we are interested in is here the rate $\lambda$):

$$\mathscr{L}(\lambda) = P(n \mid \lambda) = \frac{(\lambda t)^n}{n!} \exp(-\lambda t), \tag{2.60}$$

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

$$\frac{\partial \ln P(n \mid \lambda)}{\partial \lambda} = 0. \tag{2.61}$$

The derivative gives

$$\frac{\partial \ln P(n \mid \lambda)}{\partial \lambda} = \frac{\partial}{\partial \lambda} (n \ln(\lambda t) - \ln n! - \lambda t) = n \frac{t}{\lambda t} - t = 0 \Leftrightarrow \lambda_{ML} = \frac{n}{t}. \tag{2.62}$$

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

The region $[x_{\min}(\theta), x_{\max}(\theta)]$ contains $1-\alpha$ of the probability mass of $x$ for the fixed value of $\theta$. These intervals define a band or "belt" in the $x$-space, as illustrated in Fig. 2.8. Once the confidence belt is constructed, it needs to be inverted in order to find the confidence interval for $\theta$ given an observed value of $x = \boldsymbol{x}^{\text{obs}}$. The confidence interval for $\theta$ is given by the set of all $\theta$ values such that the interval $[x_{\min}(\theta), x_{\max}(\theta)]$ contains $\boldsymbol{x}^{\text{obs}}$:

$$\Theta(\boldsymbol{x}^{\text{obs}}) = \left\{ \theta \mid x_{\min}(\theta) \leq \boldsymbol{x}^{\text{obs}} \leq x_{\max}(\theta) \right\}.$$

This set of $\theta$ values forms the confidence interval for the parameter $\theta$ at the confidence level $1-\alpha$.

**Example 2.24** Neyman belt for a Gaussian distribution.

Consider an example where the observable $x$ follows a Gaussian distribution with mean $\theta$ and known standard deviation $\sigma$:

$$f(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(x-\theta)^2}{2\sigma^2} \right).$$

For this distribution, the confidence belt can be constructed as follows:

1 Fix the confidence level $1-\alpha$. For instance, if $\alpha = 0.05$, the confidence level is 95%.

2 For each $\theta$, the interval $[x_{\min}(\theta), x_{\max}(\theta)]$ corresponding to the central $1-\alpha$ region of the Gaussian distribution is obtained from the CDF of the standard normal distribution:

$$x_{\min}(\theta) = \theta - z_{\alpha/2}\sigma, \quad x_{\max}(\theta) = \theta + z_{\alpha/2}\sigma,$$

where $z_{\alpha/2}$ is the critical value of the standard normal distribution corresponding to the tail probability $\alpha/2$, i.e.

$$\int_{-\infty}^{z_{\alpha/2}} \mathcal{N}(x', 1)\mathrm{d}x' = \alpha/2\,. \tag{2.63}$$

3 Given an observed value $\boldsymbol{x}^{\text{obs}}$, the confidence interval for $\theta$ is obtained from inverting the interval containing probability $(1-\alpha)$:

$$\begin{aligned}
&p\left[\theta - z_{\alpha/2}\sigma \le \boldsymbol{x}^{\text{obs}} \le \theta + z_{\alpha/2}\sigma\right] = 1-\alpha \\
&p\left[-z_{\alpha/2}\sigma \le \boldsymbol{x}^{\text{obs}} - \theta \le z_{\alpha/2}\sigma\right] = 1-\alpha \\
&p\left[-z_{\alpha/2}\sigma - \boldsymbol{x}^{\text{obs}} \le -\theta \le z_{\alpha/2}\sigma - \boldsymbol{x}^{\text{obs}}\right] = 1-\alpha \\
&p\left[\boldsymbol{x}^{\text{obs}} - z_{\alpha/2}\sigma \le \theta \le \boldsymbol{x}^{\text{obs}} + z_{\alpha/2}\sigma\right] = 1-\alpha.
\end{aligned}$$

**Example 2.25** Consider again the Gaussian case, but now the data $\boldsymbol{d} = \{x_1, \ldots, x_n\}$ are i.i.d. distributed according to:

$$x_i \sim \mathcal{N}\left(\mu, \sigma^2\right) \quad (i = 1, \ldots, n)\,, \tag{2.64}$$

and we would like to determine $\mu$, with known $\sigma$. As shown in Eq. (2.53), the MLE for the mean $\mu$ is the sample average, $\bar{x} = 1/n \sum_i x_i$, which is also a sufficient statistics[6] for $\mu$. The sample mean is normally distributed according to (see Exercise **??**):

$$\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n^2}\right). \tag{2.65}$$

Using a similar reasoning as above, the Neyman belt construction gives for $\mu$, at the confidence level $1-\alpha$, the interval:

$$p[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}] = 1-\alpha\,. \tag{2.66}$$

Notice that, here too, $\mu$ is fixed (and unknown), and at each new hypothetical

[6] A function $g(\mathbf{x})$ of data $\mathbf{x}$ is said to be sufficient w.r.t. to the parameter $\theta$ of the data generating distribution if the conditional distribution of $\mathbf{x}$ given $g(\mathbf{x})$ does not depend on $\theta$. This implies that no other function of $\mathbf{x}$ can give additional information on $\theta$.

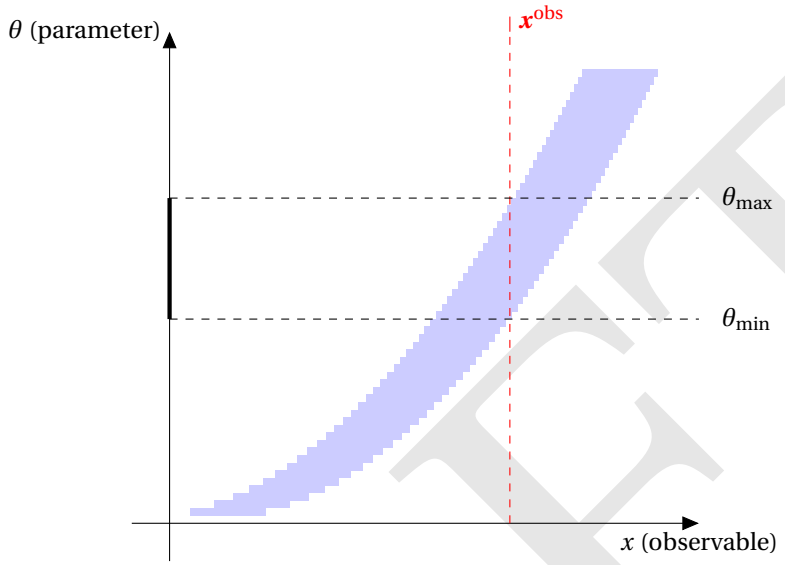


Figure 2.8 Illustration of the Neyman belt construction. For each value of the parameter $\theta$ and for a fixed confidence level $\alpha$, the horizontal blue line represents the $1-\alpha$ HPD interval in the distribution of the observable $x$. Once the data are obtained (vertical, red line) the intersection of the belt with the observed value $\boldsymbol{x}^{\text{obs}}$ gives the $1-\alpha$ confidence interval for the parameter $\theta$ (bold range on the vertical axis) as the interval $[\theta_{\min}(\boldsymbol{x}^{\text{obs}}); \theta_{\max}(\boldsymbol{x}^{\text{obs}})]$.

repetition of the experiment the value of $\bar{x}$ changes, and therefore the random interval $[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$ varies, as well. Over many such experiments, the collection of random intervals will contain (i.e., "cover") the true value of $\mu$ a fraction $1-\alpha$ of the time.

**Example 2.26** What if both $\mu$ and $\sigma$ are unknown? In this case, both the sample mean and the sample variance are RV, and it can be shown that the RV $t \equiv (\hat{\mu} - \mu)/(\hat{\sigma}/\sqrt{N})$ is distributed according to a Student's t-distribution with $N-1$ degrees of freedom[7]. For a given $\alpha$, The Neyman belt construction gives a symmetric interval around $t = 0$ with range $[-t_\alpha, t_\alpha]$, where the critical value $t_\alpha$ is obtained by solving

$$\int_{-t_\alpha}^{t_\alpha} p_{\text{Student-t}(N-1)}(x)\mathrm{d}\theta = 1 - \alpha\,. \tag{2.67}$$

[7] The **Student's t-distribution** with $\nu$ degrees of freedom is a continuous pdf, defined as $f(t \mid \nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$, where $\nu > 0$ is the number of degrees of freedom, $\Gamma(\cdot)$ is the Gamma function. Its expectation value and variance are $\mathrm{E}[t] = 0$ for $\nu > 1$, $\mathrm{Var}[t] = \frac{\nu}{\nu-2}$ for $\nu > 2$ (and undefined for $\nu \leq 2$).

By inverting the interval as above, one obtains the $1-\alpha$ confidence interval:

$$\left[\hat{\mu} - t_\alpha \frac{\hat{\sigma}}{\sqrt{N}}; \hat{\mu} + t_\alpha \frac{\hat{\sigma}}{\sqrt{N}}\right]. \tag{2.68}$$

Notice again that $\mu$ is fixed and the RVs are $\hat{\mu}, \hat{\sigma}$.

It is important to remember that coverage does not mean, however, that, for a given value of $\bar{x}$, the confidence interval contains the true $\mu$ with probability $\alpha$! For each repeated experiment, either the true value of $\theta$ is inside the reported confidence interval (in which case, the probability of $\theta$ being inside is 1) or the true value is outside (in which case its probability of being inside is 0). One has to be careful with the interpretation of confidence intervals as this is often misunderstood! Confidence intervals do not give the probability of the parameter and in Frequentist statistics it does not make sense to speak of "the probability of $\theta$", which is an un-defined concept, since $\theta$ is not a RV. In order to be able to attach a probability statement to the parameters, you need Bayes theorem. The correct interpretation is as follows: if we were to repeat an experiment many times, and each time report the observed $1-\alpha$ confidence interval, we would be correct $1-\alpha$ percent of the time. For further discussion of the many subtleties of confidence intervals definition and estimation, see Zech (2001).

Confidence intervals built from the Neyman construction violate the **likelihood principle**, which states that all inferences should be based exclusively on the likelihood $\Pr(|\boldsymbol{d\theta})$ of the observed data, and nothing else. Any other factor, such as the stopping rule adopted, the experimental design etc should be irrelevant. However, the Neyman belt construction requires knowledge of the distribution of other possible (and unobserved) values (when creating the horizontal ranges over $x$) of the observable, which violates the likelihood principle [8].

### *2.6.2 Profile likelihood*

In general, the likelihood function contains more than one parameter. Parameters that are not of direct scientific interest, but that are needed to model the observations (e.g., parameters describing the background or foreground, or the

[8] The so-called "flip-flopping physicist" problem arises in high-energy physics when researchers make ad hoc decisions about reporting results based on whether or not they fall within certain confidence intervals. This issue commonly happens when a result close to the boundaries of a confidence interval leads the physicist to switch between methods of reporting, creating inconsistent intervals. For instance, a physicist may report an upper limit for a parameter in cases where the estimated value is non-physical (e.g., negative mass or negative cross-section), but report two-sided intervals if the estimated value is positive and well-away from zero. This "flip-flopping" distorts the true confidence level, and in particular break coverage. The Feldman and Cousins (1998) construction offers a solution to this problem by providing a unified approach for constructing confidence intervals that automatically transition between one-sided and two-sided intervals. This is achieved thanks to a likelihood ratio ordering and leads to correct coverage.

| $z_{\alpha/2}$ | Confidence level $(1-\alpha)$ |
|---|---|
| 1 | 0.683 |
| 2 | 0.954 |
| 3 | 0.997 |
| 4 | 0.9993 |
| 5 | $1-5.7\times10^{-7}$ |
| 1.64 | 0.90 |
| 1.96 | 0.95 |
| 2.57 | 0.99 |
| 3.29 | 0.999 |

Table 2.1 *Relationship between the size of the symmetric interval around the mean of a standard Normal distribution and the probability contained in the range* $[-z_{\alpha/2}, z_{\alpha/2}]$.

calibration of the instrument) are called **nuisance parameters**. In a Frequentist context, such parameters are eliminated from the likelihood by profiling over them, i.e., by maximising over their value.

**Definition 2.27** Let $\boldsymbol{\theta} = [\theta, \nu]$ describe the parameters of the likelihood $\mathscr{L}(\boldsymbol{\theta})$, with $\theta$ the parameter of interest and $\boldsymbol{\nu}$ a set of nuisance parameters. The **profile likelihood** for the parameter $\theta$ is defined as:

$$\mathscr{L}_{\rm pr}(\theta) \equiv \sup_{\boldsymbol{\nu}} p(\boldsymbol{d} \mid \theta, \boldsymbol{\nu}) = \mathscr{L}(\theta, \hat{\boldsymbol{\nu}}(\theta)) \,, \tag{2.69}$$

where $\hat{\boldsymbol{\nu}}(\theta)$ is the MLE for $\boldsymbol{\nu}$ for a given value of $\theta$ (i.e., the best-fit value for $\boldsymbol{\nu}$ at a fixed value of $\theta$).

Thus in the profile likelihood one maximises the value of the likelihood along the hidden dimensions, treating the unknown nuisance parameters as if they were known and with value given my their conditional MLE $\hat{\boldsymbol{\nu}}(\theta)$. While a pseudo-likelihood, the profile likelihood is useful because of the asymptotic properties of the **profile likelihood ratio**:

**Definition 2.28** The profile likelihood ratio, $\lambda(\theta)$, for the parameter of interest $\theta$ is defined as

$$\lambda(\theta) = \frac{\mathscr{L}_{\rm pr}(\theta)}{\mathscr{L}\left(\hat{\theta}, \hat{\boldsymbol{\nu}}\right)} \,, \tag{2.70}$$

where $\hat{\theta}, \hat{\boldsymbol{\nu}}$ are the MLE for $\theta, \boldsymbol{\nu}$.

Notice that in the denominator the global maximum likelihood value appears (i.e., both $\theta$ and $\boldsymbol{\nu}$ are fitted simultaneously), while the numerator features the

profile likelihood, where $\boldsymbol{\nu}$ has been profiled over. It is convenient to work with the quantity $0 \leq t_\theta \equiv -2\ln\lambda(\theta) \leq 1$. Higher values of $t_\theta$ signal increased incompatibility of the value of $\theta$ and the observed data.

An important theorem due to Wilks (1938) gives the asymptotic distribution of the likelihood ratio test statistics, which we can use to construct confidence intervals. Another important result, the Neyman-Pearson lemma, says that the likelihood ratio is the optimal test (when comparing two simple hypotheses, i.e., when each is fully specified with no free parameters), in a sense that is explained in section 5.3.1.

**Theorem 2.29** (Wilks (1938)) *Under certain regularity conditions*[9]*, and in the large sample limit (i.e., asymptotically), the test statistics*

$$t_{\theta_0} \equiv -2\ln\frac{\mathscr{L}_{pr}(\theta_0)}{\mathscr{L}\left(\hat{\theta},\hat{\boldsymbol{\nu}}\right)}, \tag{2.71}$$

*where $\theta_0$ is the value of $\theta$ specified under the null hypothesis, is distributed according to a $\chi^2_d$ distribution with a number of degrees of freedom, $d$, equal to the dimensionality of $\theta$, independent of the value of the parameters $\boldsymbol{\nu}$. The $\chi^2_d$ distribution is given by:*

$$\chi^2_d(x) = \frac{1}{\Gamma(d/2)2^{d/2}}(x)^{\frac{d}{2}-1}\exp\left(-\frac{1}{2}x\right). \tag{2.72}$$

Thanks to Wilks' theorem, we can construct a confidence interval for $\theta$ by performing a hypothesis test (see section 5.3.1): the null hypothesis is that $\theta_0 = \hat{\theta}$ (i.e., the maximum likelihood value). A $1-\alpha$ confidence interval for $\theta$ (with $\alpha$ the significance level of the test, e.g., $\alpha = 0.01$) is the set of possible values of $\theta$ for which the probability that the test statistics under the null, $t_{\theta_0}$, is larger than or equal to $t_\theta$ is larger than $\alpha$. This means that such a value of $\theta$ would fail to be rejected with a hypothesis test that $\theta = \hat{\theta}$ at the $\alpha$ significance level. The explicit construction is as follows.

1 For each value of $\theta$ we consider, the $p$-value (tail probability) under the null hypothesis is given by:

$$\wp(\theta) = \int_0^{t_\theta} \chi^2_d(x')\mathrm{d}\theta' = 1 - \mathrm{CDF}_{\chi^2_d}(t_\theta) \tag{2.73}$$

[9] One important (and often-overlooked) condition for the validity of Wilks' theorem is that the parameter value being tested must not lie at the boundary of the allowed parameter space. In this case, one ought to employ Chernoff (1954)'s theorem instead, which says that the asymptotic distribution of $-2\ln\lambda$ is the mixture $\frac{1}{2}\chi^2_d + \frac{1}{2}\delta(0)$, where $\delta(0)$ is the Dirac delta-function centered at 0. Another condition is that the asymptotic regime (which ensures that the MLE is normally distributed) has been achieved – thus Wilks' typically fails for small sample size. 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 Algeri et al. (2020). In case where the asymptotic distribution does not hold, one can replace the analytical distribution with a Monte Carlo study, by simulating the distribution of the test statistics from the null.

where $\mathrm{CDF}_{\chi^2_d}$ is the CDF of the $\chi^2$ distribution with $d$ degrees of freedom (with $d = 1$ for unidimensional $\theta$).

2 The boundary of the $1-\alpha$ confidence region, $t_*$, is obtained by setting the $p$-value equal to the desired tail probability, i.e. $\wp = \alpha$, and solving for the corresponding critical value using the quantile (i.e., inverse CDF) function, denoted by $Q_{\chi^2_d}(\cdot)$:

$$1-\alpha = \mathrm{CDF}_{\chi^2_d}(t_*) \Rightarrow t_* = Q_{\chi^2_d}(1-\alpha). \tag{2.74}$$

3 Recalling that $t_\theta$ is nothing else but the difference in minus twice log-likelihood between $\theta$ and the MLE, we can see that the confidence region is bounded by the parameters values satisfying:

$$\ln \mathscr{L}_{\mathrm{pr}}(\theta) = \ln \mathscr{L}\left(\hat{\theta}, \hat{\boldsymbol{\nu}}\right) - \frac{1}{2} Q_{\chi^2_d}(1-\alpha)\,. \tag{2.75}$$

This leads to the following prescription for constructing approximate profile likelihood-based confidence intervals. Starting from the MLE, a $1-\alpha$ profile likelihood confidence region (or interval for $d = 1$) encloses all parameter values for which minus twice the profile log–likelihood increases less than $Q_{\chi^2_d}(1-\alpha)$ from the MLE value. We can write this as

$$-2\ln \mathscr{L}_{\mathrm{pr}}\left(\boldsymbol{\phi}\right) = -2\ln \mathscr{L}\left(\hat{\theta}, \hat{\boldsymbol{\nu}}\right) + Q^{\alpha}_{\chi^2_d}\,, \tag{2.76}$$

where the threshold value, $Q^{\alpha}_{\chi^2_d} = Q_{\chi^2_d}(1-\alpha)$, depends on the confidence level $1-\alpha$ and on the number of parameters of interest being considered (usually, $d = 1$ or $d = 2$). Values of $Q^{\alpha}_{\chi^2_d}$ for 1- and 2-dimensional regions are tabulated in Table 2.2 for a few common choices for $1-\alpha$. The confidence regions obtained in this manner are therefore the regions in parameter space in which the hypothesis that the data were generated from the MLE value would fail to be rejected with a significance level $\alpha$ using a likelihood ratio test (see section 5.3.1 for more details on hypothesis testing).

We note that profile likelihood-based confidence interval do respect the likelihood principle, in which only the observed value of the likelihood enters the computation, and no distributional assumptions are made beside the asymptotic distribution of the log-likelihood ratio obtained from Wilks' theorem. In that respect, the resulting confidence intervals have to be considered approximate, in that the asymptotic distribution may not be achieved for small sample size.

| $1-\alpha$ | 0.683 | 0.950 | 0.954 | 0.990 | 0.997 |
|---|---|---|---|---|---|
| Normal equivalent | $(1\sigma)$ | $(1.96\sigma)$ | $(2\sigma)$ | $(2.58\sigma)$ | $(3\sigma)$ |
| $d=1$, $Q^{\alpha}_{\chi^2_1}$ | 1.00 | 3.84 | 4.00 | 6.63 | 9.00 |
| $d=2$, $Q^{\alpha}_{\chi^2_2}$ | 2.30 | 5.99 | 6.17 | 9.21 | 11.80 |

Table 2.2 *Values of threshold levels of twice the log-likelihood increase from the MLE for profile likelihood-based confidence intervals in 1 and 2 dimensions, for some common choice of confidence level* $1-\alpha$.

### *2.6.3 Multivariate Wald Confidence Regions*

For a general likelihood function and $\dim(\boldsymbol{\theta}) = d \geq 1$, provided that the $\ln \mathscr{L}(\boldsymbol{\theta})$ is twice differentiable in $\boldsymbol{\theta}$ and that the range of the data is independent of the parameters, the MLE is asymptotically (i.e., for $n \to \infty$) distributed as a multivariate normal around the true value, $\boldsymbol{\theta}_0$, that is:

$$\boldsymbol{\theta}_{\text{MLE}} \sim \mathscr{N}\left(\boldsymbol{\theta}_0, \boldsymbol{\Sigma^2}\right). \tag{2.77}$$

The covariance matrix of this multivariate distribution is given by the expectation value (with respect to the sampling distribution) of the Hessian of the log-likelihood, i.e. the **Fisher information matrix**:

$$\Sigma^{-2}_{ij} = \mathbb{E}\left[-\frac{\partial^2 \ln \mathscr{L}(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right]. \tag{2.78}$$

If $n$ is sufficiently large and there exists a set of $d$ jointly sufficient statistics for the $d$ parameters in $\boldsymbol{\theta}$, then the above covariance can be approximated by the observed (as opposed to the expected) Fisher matrix:

$$\Sigma^{-2}_{ij} \approx -\frac{\partial^2 \ln \mathscr{L}(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\Big|_{\boldsymbol{\theta}_{\text{MLE}}} \equiv \mathbf{I}(\hat{\boldsymbol{\theta}}). \tag{2.79}$$

Since the distribution for the ML estimator can be asymptotically approximated as a multivariate Gaussian, we can use a multivariate generalization of the profile likelihood ratio test above. The Wald statistic for constructing confidence regions is based on the approximate normality of the MLE. Specifically, for a parameter vector $\boldsymbol{\theta}$, the Wald statistic is:

$$W(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\top} \mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

This statistic follows a chi-squared distribution asymptotically with $d$ degrees of freedom

$$W(\boldsymbol{\theta}) \sim \chi^2_d$$

To construct a $(1-\alpha)$ confidence region, we solve for the set of values of $\boldsymbol{\theta}$ such that the Wald statistic $W(\boldsymbol{\theta})$ is less than or equal to the critical value from the chi-squared distribution with $d$ degrees of freedom, $Q^{\alpha}_{\chi^2_d}$:

$$W(\boldsymbol{\theta}) \leq Q^{\alpha}_{\chi^2_d}$$

Explicitly, the confidence region is given by:

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\top} \mathbf{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq Q^{\alpha}_{\chi^2_d}$$

This inequality defines an ellipsoidal region in the parameter space, centered at the MLE $\hat{\boldsymbol{\theta}}$. The shape of the ellipsoid is determined by the Fisher information matrix, and its size depends on the critical value $Q^{\alpha}_{\chi^2_d}$ corresponding to the desired confidence level $1-\alpha$.

The confidence region includes all parameter vectors $\boldsymbol{\theta}$ for which the Wald statistic does not exceed the critical value, meaning that those values of $\boldsymbol{\theta}$ are not significantly different (in the hypothesis testing sense) from the MLE at the chosen confidence level.

## 2.7 Fisher matrix forecasts

### Exercises

2.1 You flip a coin $n = 10$ times and obtain 8 heads.

1. What is the likelihood function for this measurement? Identify explicitly what are the data and what is the free parameter you are trying to estimate.

2. What is the Maximum Likelihood Estimate for the probability of obtaining heads in one flip, $\theta$?
3. Approximate the likelihood function as a Gaussian around its peak and derive the $1\sigma$ confidence interval for $\theta$. How would you report your result for $\theta$?
4. With how many $\sigma$ confidence can you exclude the hypothesis that the coin is fair? (*Hint: compute the distance between the MLE for $p$ and $p = 1/2$ and express the result in number of $\sigma$*).
5. You now flip the coin 100 times and obtain 80 heads. What is the MLE for $\theta$ now and what is the $1\sigma$ confidence interval for $\theta$? With how many $\sigma$ confidence can you exclude the hypothesis that the coin is fair now?

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

1 If $\lambda$ is the average number of counts per per unit time interval, write down the appropriate probability distribution function for $r$.

2 Now we seek to determine $\lambda$ 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 joint likelihood function for $\lambda$, $\mathscr{L}(\lambda) = P(\hat{r} \mid \lambda)$, given the measured sequence of counts $\hat{r}$.

3 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 $\lambda$ is just the average of the measured counts, $\hat{r}$, i.e.

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

4 By considering the Taylor expansion of $\ln \mathscr{L}(\lambda)$ to second order around $\lambda_{\mathrm{ML}}$, derive the Gaussian approximation for the likelihood $\mathscr{L}(\lambda)$ around the Maximum Likelihood point, 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.

5 Compare with the equivalent expression for $M$ Gaussian-distributed measurements to show that the variance $\sigma^2$ of the Poisson distribution is given by $\sigma^2 = \lambda$.

2.3 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$^2$):

$$34.7 \pm 5.0, \quad 28.9 \pm 2.0, \quad 27.1 \pm 3.0, \quad 30.6 \pm 4.0. \tag{2.80}$$

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

2 Write down the joint likelihood for the combination of the 4 measurements.

3 Find the MLE of the photon flux, $F_{\mathrm{MLE}}$, and show that it is given by:

$$F_{\mathrm{MLE}} = \sum_i \frac{\hat{n}_i}{\hat{\sigma}_i^2 / \bar{\sigma}^2}, \tag{2.81}$$

where

$$\frac{1}{\bar{\sigma}^2} \equiv \sum_i \frac{1}{\hat{\sigma}_i^2}. \tag{2.82}$$

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

5 Find the $1\sigma$ confidence interval for your MLE for the mean, and show that it is given by:

$$\left(\sum_i \frac{1}{\hat{\sigma}_i^2}\right)^{-1/2}. \tag{2.83}$$

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



 

# 3

# Theory of Bayesian Inference

## 3.1 Bayes Theorem as an Inference Device

As a mathematical result, Bayes theorem, Eq. (2.8), is elementary. It becomes interesting (and somewhat controversial[1]) for the purpose of inference when we replace $A \to \boldsymbol{\theta}$ (the parameters of interest one wants to infer) and $B \to \boldsymbol{d}$ (the observed data), obtaining:

$$\Pr(|\, \boldsymbol{\theta}\boldsymbol{d}) = \frac{\Pr(|\, \boldsymbol{d}\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})}{\Pr(\boldsymbol{d})} . \tag{3.1}$$

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

On the RHS, $\Pr(|\, \boldsymbol{d}\boldsymbol{\theta}) = \mathscr{L}(\boldsymbol{\theta})$ is the probability of the data given the value of the parameters. When considered as a function of the parameters for the observed data, it is the likelihood function we already encountered.

The quantity $\Pr(\boldsymbol{\theta})$ is the **prior distribution** (or "prior" for short). It represents our degree of belief in the value of $\boldsymbol{\theta}$ before we see the data (hence the name). This is an essential ingredient of Bayesian statistics.

In the denominator, $\Pr(\boldsymbol{d})$ is a normalizing constant, **the Bayesian evidence** or (in statistics) **marginal likelihood** (as all parameters have been integrated out), than ensures that the posterior is normalized to unity, and it is given by the average of the likelihood over the prior:

$$\Pr(\boldsymbol{d}) = \int \mathrm{d}\boldsymbol{\theta} \mathscr{L}(\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) . \tag{3.2}$$

The evidence is the fundamental object in Bayesian model comparison (see chapter 5), but can be ignored insofar as parameter inference is concerned, for it is a normalizing constant that does not depend on $\boldsymbol{\theta}$.

[1] Excellent popular science accounts of the history of Bayes theorem, and its modern applications, can be found in McGrayne (2012); Chivers (2024).

Bayes theorem relates the posterior probability for $\boldsymbol{\theta}$ (i.e., what we know about the parameters after seeing the data) to the likelihood and the prior (i.e., what we knew about the parameters before we saw the data). It can be thought of as a general rule to update our knowledge about a quantity from the prior to the posterior.

Remember that, in general, $\Pr(|\boldsymbol{\theta}\boldsymbol{d}) \neq \mathscr{L}(\boldsymbol{\theta})$, i.e. the posterior and the likelihood are two different quantities, with different meaning! The likelihood is the probability of making the observation if we know what the parameters are, while the posterior is the probability of the parameters given that we have made a certain observation.

**Example 3.1: Posterior vs likelihood**

We want to determine if a randomly-chosen person is male (M) or female (F)[a]. 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 $P(d = Y \mid \theta = F) = 0.03$ (i.e., there is a 3% probability that a randomly selected female is pregnant), but the posterior probability $P(\theta = F \mid d = Y) = 1.0$, 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.

[a] This example is due to Louis Lyons.

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 Cox Theorem

In this section, we address the question of how the subjective probabilities in Bayesian theory are to be assigned. What we need is a set of principles for assigning subjective probabilities by logical analysis of incomplete information

 

(Jaynes, 2003; Van Horn, 2003). Moreover, we will show that the Kolmogorov axioms of probability theory, introduced as postulates in section 2.3, have a deeper origin and are in fact *the only* consistent rules for doing inference.

In this context, one could see the Frequentist approach to probability as a particular application of these rules, in the limiting case in which:

1 there is no relevant prior information;
2 events are independent repetitions of equiprobable experiment, i.e. they are independently, identically distributed (i.i.d.); and
3 the number of repetitions tends to infinity (asymptotic limit).

Under these circumstances, the Bayesian and Frequentist approaches agree, as shown by the following

**Theorem 3.1** (Bernstein-von Mises) *Let $\boldsymbol{\theta}_0$ be the true value of the parameter of interest, and let $x_1, x_2, \ldots, x_n$ be a sequence of i.i.d. random variables with a probability density function $\Pr(X \mid \boldsymbol{\theta})$. Assuming that:*

- *The likelihood function $\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \Pr(x_i \mid \boldsymbol{\theta})$ is differentiable and satisfies certain regularity conditions.*
- *The prior distribution $\Pr(\boldsymbol{\theta})$ is positive and continuous in a neighborhood of the true parameter value $\boldsymbol{\theta}_0$.*

*Then, as the sample size $n$ increases, the posterior $\Pr(\boldsymbol{\theta} \mid x_1, \ldots, x_n)$ converges in distribution to:*

$$\Pr(\boldsymbol{\theta} \mid x_1, \ldots, x_n) \xrightarrow{\text{in d.}} \mathcal{N}\left(\hat{\boldsymbol{\theta}}_n, \frac{1}{nI(\hat{\boldsymbol{\theta}}_n)}\right),$$

*where $\hat{\boldsymbol{\theta}}_n$ is the MLE of $\boldsymbol{\theta}$ and $I(\hat{\boldsymbol{\theta}}_n)$ is the observed Fisher information evaluated at the MLE. In other words, the posterior distribution becomes approximately normal, centered around the MLE with a variance that shrinks at a rate of $1/n$ as $n \to \infty$. This shows that the Bayesian posterior converges to the Frequentist likelihood as $n$ grows, regardless of the prior $\Pr(\boldsymbol{\theta})$, provided that the prior is well-behaved. The theorem therefore guarantees that inferences based on the Bayesian posterior are asymptotically correct in the Frequentist sense, if the prior is non-zero around the true value of the parameters. For a proof, see e.g. Van der Vaart (1998).*

The Bernstein-von Mises theorem provides reassurance that inferences obtained from the likelihood will agree with the Bayesian results in the asymptotic limit. While helpful, this however does not tell us which scientific conclusions we should reach in regimes (e.g., small sample limits) where the two approaches give divergent results – we will encounter examples of such situations below.

If the Bayesian and Frequentist approaches agree asymptotically, a question arises naturally: Why should the rules of classical probability theory, which deal with relative frequencies, coincide with the rules for manipulating degrees of belief, which are in appearance a completely different object? After all, frequencies and degrees of belief (or states of information) are two very different concepts. We would like to show that a system of plausible reasoning, subject to qualitative requirements, must be isomorphic to probability theory. In fact, what we would ideally like to derive is a system of plausible reasoning that extends deductive logic (i.e., propositional calculus) to propositions that are not conclusively true or false. This is the content of **Cox (1946) Theorem**.

We start by considering a minimal set of requirements for 'plausible reasoning', i.e., a system of reasoning capable of dealing with propositions that are not conclusively true or false, but that is compatible with classical deductive logic:

R1) The "degree of plausibility" of a proposition is represented by a single[2], real number; i.e., there exists a real number $T$, such that the plausibility of proposition $A$ given $X$ satisfies: $(A \mid X) \leq T$, $\forall A, X$. In the following, we denote by $(\cdot)$ the plausibility assignment to proposition $\cdot$.

R2) Plausibility assignments are compatible with propositional calculus; denoting by "$A \wedge B$" the proposition "proposition $A$ and $B$ together" (which for calculus is equivalent to the logical "and"), we have:

1 If $A$ is equivalent[3] to $A'$, then $(A \mid X) = (A' \mid X)$.

2 If $A$ is a tautology, then $(A \mid X) = T$.

3 $(A \mid B, C, X) = (A \mid (B \wedge C), X)$ (if we know that both $B$ and $C$ are true, then $B \wedge C$ must also be true).

4 If $X$ is consistent[4] and $(\neg A \mid X) < T$, then $\{A, X\}$ is also consistent (if $\neg A$ cannot be said to be true with information $X$, then there remains a possibility that $A$ is true, hence information $X$ must not contradict $A$).

R3) Negation: there exists a non-increasing function $S_0$, such that $(\neg A \mid X) = S_0(A \mid X)$. The non-increasing requirement comes from consistency with propositional calculus, for we know that if $\neg A$ is true, then $A$ must be false. Defining $F \equiv S_0(T)$, it follows that $F \leq (A \mid X) \leq T$ for all $A$ and consistent $X$, since $(A \mid X) = S_0(\neg A \mid X) \geq S_0(T)$.

[2] That plausibility ought to be fully expressed by a single number is controversial. Some authors have suggested the use of two dimensions, to quantify, alongside the degree of belief, also the degree of doubt in a proposition. Here, however, we consider only the simplest case.

[3] $A$ is said to be 'equivalent' to $B$ if $(A \iff B)$ is a tautology, i.e., a true proposition regardless of the truth of its atomic propositions.

[4] $X$ is said to be 'consistent' if there is no proposition $B$ for which $(B \mid X) = T$ and $(\neg B \mid X) = T$. For example, a state of information $\{A, \neg A, X\}$ would be inconsistent, for it asserts the truth of two inconsistent propositions.

R4) Universality: there exists a non-empty set of real numbers $P_0$ with the following properties:

1 $P_0$ is a dense subset of $(F,T)$ , i.e., $\forall a,b \in \mathbb{R}$ so that $F \leq a < b \leq T$, there exists a $c \in P_0$ such that $a < c < b$. This ensures that there are no holes in our plausibility assignments.

2 For any $y_1, y_2, y_3 \in P_0$, there exists a consistent $X$ with at least three atomic propositions $A_1, A_2, A_3$, such that $(A_1 \mid X) = y_1, (A_2 \mid A_1, X) = y_2, (A_3 \mid A_1, A_2, X) = y_3$. This is a weaker requirement than assuming that, if the propositions $A_i$ $(i = 1,2,3)$ are unrelated, they can be assigned an arbitrary plausibility without reference to the others – but this case is included in this requirement.

R5) There exist a continuous function $\mathscr{F} : [F,T]^2 \to [F,T]$, strictly increasing in both arguments[5], giving the plausibility of the proposition $(A \wedge B \mid X)$ as a function of only two arguments, namely:

$$(A \wedge B \mid X) = \mathscr{F}\left[(A \mid B, X), (B \mid X)\right], \forall A, B \text{ and consistent } X. \tag{3.3}$$

The choice of just $(A \mid B, X)$ and $(B \mid X)$ as arguments of $\mathscr{F}[,]$, among the 15 possible combinations of the atomic propositions $A, B, X$ follows from these considerations: since $A \wedge B$ is true only if $B$ is true, the plausibility assignment $(B \mid X)$ must be relevant to determine the plausibility of $A \wedge B$; if $B$ is true, $A$ needs also to be true for $A \wedge B$ to be true; therefore, the plausibility $(A \mid B, X)$ is also relevant. All other possible arguments are either superfluous or lead to logical contradiction in the plausibility assignment.

Notice that nowhere in the above requirements for plausible reasoning have we mentioned random variables, nor relative frequency of outcomes. Instead, we have merely stated five requirements for plausible reasoning, inspired by adherence to propositional calculus and logical consistency. Very surprisingly, such requirements lead to the proof of the existence of a function, $\Pr(\cdot)$, that exhibits the exact axiomatic properties introduced by Kolmogorov That is to say, the Kolmogorov axioms can be derived from the above five requirements for plausible reasoning under uncertainty. In other words, a system of plausible reasoning that generalizes logical deduction is isomorphic to probability theory. This is the content of

**Theorem 3.2** (Cox) *Given requirements R1 to R5 above, there exists a continuous, strictly increasing function* $\Pr(\cdot)$ *such that for all propositions A, B and consistent X:*

[5] This is justified as follows: imagine that when new information $X'$ becomes available, the plausibility of $B$ increases, while that of $A$ conditional on $B$ stays the same, i.e., $(B \mid X') > (B \mid X)$, and $(A \mid B, X) = (A, B, X')$. Then the plausibility of $A \wedge B$ must also increase, i.e., $(A \wedge B \mid X') > (A \wedge B \mid X)$.

*1* $\Pr(|\, AX) = 0$ *iff* $A$ *is false given* $X$*;*

*2* $\Pr(|\, AX) = 1$ *iff* $A$ *is true given* $X$*;*

*3* $0 \leq \Pr(|\, AX) \leq 1$*;*

*4* $\Pr(|\, A \wedge BX) = \Pr(|\, AX) \Pr(|\, BA, X)$ *(i.e., the product rule is satisfied);*

*5* $\Pr(|\, \neg AX) = 1 - \Pr(|\, AX)$ *(i.e, the sum rule is satisfied).*

The function $\Pr(\cdot)$ entering the theorem is a rule for relating the plausibility of a proposition, $(|\, AX)$, with its probability $\Pr(|\, AX)$; but, since the function is invertible by definition, we can equally well work with $\Pr(|\, AX)$ or $(|\, AX)$ as far as the proof is concerned.

Here we can only sketch what the proof of the theorem involves, following Jaynes (2003, Chap. 2) and Van Horn (2003). We begin by proving the product rule (number (iv) in the list above): we want to show that there exists a continuous, monotonic and non-negative function $w$ (which at the end will be identified with probability) such that $w(|\, A \wedge BC) = w(|\, AB, C)\, w(|\, BC)$.

*Product Rule* For consistency with propositional calculus, we require the associativity property. Since $A \wedge B \wedge C = [A \wedge B] \wedge C = A \wedge [B \wedge C]$, we have

$$|\, A \wedge B \wedge CX = \mathscr{F}\left[(|\, CX), (|\, A \wedge BC, X)\right] \tag{3.4}$$

$$= \mathscr{F}\left[(|\, CX), \mathscr{F}\left[(|\, BC, X), (|\, AB, C, X)\right]\right] \text{ and also} \tag{3.5}$$

$$|\, A \wedge B \wedge CX = \mathscr{F}\left[(|\, B, CX), (|\, AB, C, X)\right] \tag{3.6}$$

$$= \mathscr{F}\left[\mathscr{F}\left[(|\, CX), (|\, BC, X)\right], (|\, AB, C, X)\right]. \tag{3.7}$$

Equating Eq. (3.5) and (3.7), we get the *associative equation*

$$\mathscr{F}\left[x, \mathscr{F}\left[y, z\right]\right] = \mathscr{F}\left[\mathscr{F}\left[x, y\right], z\right]. \tag{3.8}$$

If a function satisfies the associative equation (3.8), then Aczèl Lemma says that there exist some continuous, strictly increasing function $g$ so that

$$g(f(x, y)) = g(x) + g(y). \tag{3.9}$$

Define $w(x) \equiv \exp\big(g(x)\big)$, and by applying Aczèl Lemma with $f = \mathscr{F}$ we obtain that

$$w(\mathscr{F}\left[x, y\right]) = \exp\big(g(\mathscr{F}\left[x, y\right])\big) = \exp\big(g(x)\big) \exp\big(g(y)\big) = w(x)\, w(y). \tag{3.10}$$

We call the function $w(\cdot)$ a proto-probability – 'proto', since it has not yet been shown to be identical to probability. If we now seek to apply $w$ to the plausibility of $A \wedge B$ given $X$, we obtain that it satisfies the property:

$$w((A \wedge B \mid X)) = w(\mathscr{F}\left[(A \mid B, X), (B \mid X)\right]) = w(A \mid B, X)\, w(B \mid X), \tag{3.11}$$

where the second equality follows from Aczèl Lemma. This shows that proto-probabilities satisfy the product rule.

In order to be able to call $w$ a probability we still have to show that it is bounded by zero and one and that zero (one) corresponds to impossibility (certainty). This can easily be done in the 'Boolean limit', as we show now. Suppose that $X \implies A$; then

i $(\mid AB, X) = (\mid AX)$, i.e. $B$ does not matter; and

ii $(\mid A \wedge BX) = (\mid BC)$, i.e. $A$ doesn't matter as it is implied by $X$.

Then $w(\mid A \wedge BX) \overset{\text{(ii)}}{=} w(\mid BX) \overset{\text{p.r.}}{=} w(\mid AB, X)\, w(\mid BX) \overset{\text{(i)}}{=} w(\mid AX)\, w(\mid BX)$, where 'p.r.' stands for 'product rule', for all $w(\mid BX)$; therefore $w(\mid AX) = 1$ when $A$ is certain under $X$. Thus, certainty is represented by $w(\cdot) = 1$.

Similarly, supposing that $A$ is impossible given $X$, one can easily show that $w(\mid A \wedge BX) = w(\mid AX) \overset{\text{p.r.}}{=} w(\mid AX)\, w(\mid BX)$ for all $w(\mid BX)$, which requires $w(\mid AX) = 0$ when $A$ is impossible under $X$. That is, impossibility is represented as $w(\cdot) = 0$.

We conclude that $w$, which satisfies the product rule, is a probability; this proves this part of Cox theorem. □

In conclusion, Cox theorem is a powerful result that explains why we can manipulate degrees of plausibility $(\mid AX)$ using the same set of rules as for probability theory, where $\Pr(\cdot)$ represent relative frequencies; in the Frequentist probability theory, those rules match exactly Kolmogorov's axioms. Cox theorem connects them to plausible reasoning. The theorem also implies that any generalisation of Boolean algebra that satisfies R1–R5 must be isomorphic to probability theory.

### 3.3 De Finetti's Exchangeability Theorem

As a consequence of their different views on what can be considered a RV, Bayesians and Frequentists attribute different meaning to the question of inference: while a Frequentist aims to find an approximation for the unknown –but fixed– set of true parameters $\boldsymbol{\theta}$ (which are not random variables), a Bayesian aims to improve their knowledge, i.e. the sharpness of their degree of belief, on $\boldsymbol{\theta}$ (which are instead considered random variables). For both, data is the key. Note, however, that for a Frequentist data must come from some experiment that is repeatable, at least in principle.

 

### 3.3.1 Exchangeability

Very often, the proof of asymptotic results in Frequentist settings leans on the assumption that the RVs are assumed to be **independently identically distributed (i.i.d.)**, that is, their joint distribution is the product of their individual distributions, taken to be equal:

$$\Pr(X_1, X_2, \ldots, X_N) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{N} \Pr(X_i \mid \boldsymbol{\theta}). \tag{3.12}$$

(Notice, once again, that $\boldsymbol{\theta}$ in the above expression should be thought as a set of *fixed* parameters).

In contrast to the notion of i.i.d. random variables, Italian statistician Bruno de Finetti[6] introduced the notion of **exchangeability** of random variables.

**Definition 3.3** A finite sequence of random variables $\{X_i\}_{i=1}^{N}$ is called 'exchangeable' if the order of the random variables does not matter, i.e., when for any permutation $\pi$ of $\{1, 2, \ldots, N\}$ the distributions of $X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(N)}$ and $X_1, X_2, \ldots, X_N$ agree:

$$\Pr(X_1, X_2, \ldots, X_N) = \Pr\left(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(N)}\right). \tag{3.13}$$

An 'infinitely exchangeable' sequence $\{X_i\}_{i \in \mathbb{N}}$ is one for which the finite sequences $X_1, X_2, \ldots, X_N$ are exchangeable for any $N \in \mathbb{N}$.

Exchangeability says that the sequence in which the data are gathered does not matter: each random variable is equivalent to each other in terms of when it occurs. In this respect, it is a symmetric type of ignorance with respect to the present and past, and it is relevant in many practical settings, for example, flipping coins or measuring particle production in an accelerator. It is also a weaker assumption than i.i.d., in that exchangeable sequences need not be i.i.d. (viceversa, an i.i.d. sequence is also exchangeable).

It is clear that if a sequence is exchangeable, then the random variables must be identically distributed. Also, if the variables are i.i.d., then they are exchangeable, since

$$\begin{aligned} \Pr(X_1, X_2 \ldots, X_N) &\overset{\text{indep.}}{=} \Pr(X_1) \Pr(X_2) \ldots \Pr(X_N) \\ &= \prod_{i=1}^{N} \Pr\left(X_{\pi(i)}\right) \\ &= \Pr\left(X_{\pi(1)}, \ldots, X_{\pi(N)}\right) \end{aligned} \tag{3.14}$$

[6] Bruno de Finetti was born in Austria in 1906, and studied mathematics and the Politecnico di Milano. He became a statistician at the Instituto Nazionale di Statistica and from 1931 worked for the Assicurazioni Generali in Trieste. He won a chair in statistics at the University of Trieste in 1936, but fascist laws prevented him from being appointed until 1950. He later became professor at La Sapienza in Rome. He died in 1985.

where we used independence in the first line and fact that identically distributed variables fulfil $\Pr(X_1) = \Pr(X_2) = \cdots = \Pr(X_N)$. However, exchangeable variables need not be independent, as can be seen with a simple example: consider $N$ i.i.d. random variables, $\{X_1, \ldots, X_N\}$ and add to each the same random variable, $Y$, independent from all others. Then the resulting sequence $\{X_1 + Y, \ldots, X_N + Y\}$ is exchangeable but not independent, and therefore not i.i.d..

### *3.3.2 Representation Theorem*

De Finetti's exchangeability theorem is an important cornerstone of modern statistics, and particularly of Bayesianism. It states that any infinite sequence of exchangeable random variables can be represented as a mixture of i.i.d. random variables. This mixture can be interpreted in Bayesian terms, where the sequence of variables is governed by an unknown parameter, and the uncertainty about this parameter is described by a prior distribution. This important result was first derived by de Finetti (1937) (translation in English in de Finetti (1992)) but recognized only much later[7].

**Theorem 3.4** (de Finetti's $\{0,1\}$ representation theorem) *Consider an infinite sequence*[8] $\{X_i\}_{i=1}^{\infty}$ *of Bernoulli random variables (i.e., discrete variables taking a value of 1 with probability* $\theta$*, and 0 with probability* $1-\theta$*), with* $\{X_i\}_{i=1}^{N}$ *exchangeable for each* $N$*. Define*

$$\overline{Y}_N \equiv \sum_{i=1}^{N} X_i$$

*and denote by* $\{x_1, x_2, \ldots, x_N\}$ *the realised sequence of random variables, and by* $s_N = \sum_{i=1}^{N} x_i$ *the observed number of 1s.*

*Then there exists a cumulative distribution function* $Q(t)$ *such that*

$$\Pr(X_1 = x_1, \ldots, X_N = x_n) = \int_0^1 \left\{ \prod_{i=1}^{N} \theta^{x_i} (1-\theta)^{1-x_i} \right\} \mathrm{d}Q(\theta) \tag{3.15}$$

[7] For an overview of de Finetti's work and its impact, as well as his life, see Cifarelli and Regazzini (1996).

[8] This is an important condition for the theorem to hold, and it is easy to create finite sequences counterexamples. E.g. Diaconis (1977) gives the example for $N = 2$ and a finitely echangeable sequence of Bernoulli RV $\{X_1, X_2\}$ such that:

$$P(X_1 = 0, X_2 = 0) = P(X_1 = 1, X_2 = 1) = 0 \text{ and } P(X_1 = 1, X_2 = 0) = P(X_1 = 0, X_2 = 1) = 1/2\,.$$

If the representation theorem would hold, this means that there exists some $Q(\theta)$ such that

$$\int_0^1 \theta^2 dQ(\theta) = \int_0^1 (1-\theta)^2 dQ(\theta) = 0\,,$$

which is impossible since such a $Q(\theta)$ would have to assign probability 1 to both $\theta = 0$ and $\theta = 1$. (Bernardo and Smith, 1994, Proposition 4.19) however gives formal reassurance that in practice it is not necessary for $N$ to be infinite, merely very large.

*where*

$$Q(t) = \lim_{N\to\infty} \Pr(Y_N/N \leq t)\,.$$

*The random variable $\theta \in [0,1]$ is defined by $\theta \equiv \lim_{N\to\infty} Y_N/N$, and therefore asymptotically for $N \to \infty$, $s_N/N \to \theta$, i.e. $\theta$ is asymptotically equal to the limiting relative frequency of observed 1s.*

The theorem says that the distribution of the observable random variables, $\Pr(X_1 = x_1, \ldots, X_N = x_n)$, can be written as a mixture of conditionally independent (given $\theta$) binomial sampling distributions, $\theta^{X_i}(1-\theta)^{1-X_i}$, weighted by a marginal distribution $Q$. When the distribution $Q$ admits a density $\Pr(\theta)$, such that $\mathrm{d}Q(\theta) = \Pr(\theta)\,\mathrm{d}\theta$, the representation can be written as:

$$\Pr(X_1 = x_1, \ldots, X_N = x_n) = \int_0^1 \left\{ \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} \right\} \Pr(\theta)\,\mathrm{d}\theta \tag{3.16}$$

If instead of considering the individual random variables in the sequence we are interested only in the summary random variable $Y_N$, then, since $\Pr\left(\sum_{i=1}^{N} X_i = s_N\right) = \binom{N}{s_N} \Pr(X_1 = x_1, \ldots, X_N = x_n)$, the representation theorem takes the form:

$$\Pr\left(\sum_{i=1}^{N} X_i = s_N\right) = \int_0^1 \binom{N}{s_N} \theta^{x_i}(1-\theta)^{1-x_i} \Pr(\theta)\,\mathrm{d}\theta\,, \tag{3.17}$$

where we easily recognize the expression of the Bayesian evidence with a Binomial likelihood and prior density given by $\Pr(\theta)$.

De Finetti's theorem is a profound result that connects the subjectivist Bayesian view with the Frequentist one: it is *as if* the observed $x_1, \ldots, x_N$ are a random sample from a Bernoulli distribution with parameter $\theta$, with the parameter itself a random variable with prior distribution $\Pr(\theta)$. In other words, it is *as if* the $x_1, \ldots, x_N$ were a random sample from a joint sampling distribution

$$\Pr(X_1 = x_1, \ldots, X_N = x_n \mid \theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} = \prod_{i=1}^{N} f(x_i \mid \theta)\,,$$

which, when interpreted for fixed $x_i$'s and as a function of $\theta$ is the likelihood function. What is more, exchangeability leads to the likelihood being written as a multiplication of conditionally independent (conditioned on $\theta$) distributions, since the joint distribution factorizes into a product of individual RVs. The distribution $\Pr(\theta)$ is interpreted as the prior because it *as if* it represents beliefs about what we can anticipate observing as the limiting relative frequency of a very large number of observations. Therefore, the representation theorem justifies the usual Bayesian procedure of building a posterior out of the multiplication of prior and likelihood.

The above theorem can be generalized in de Finetti's general representation theorem:

**Theorem 3.5** (General representation) *Consider an infinitely exchangeable sequence $\{X_i\}_{i=1}^{\infty}$ of real-valued random variables. Then there exists a probability measure $Q$ in the set of all distribution functions $\mathscr{Q}$ on $\mathbb{R}$, such that the joint distribution of $\{X_i\}_{i=1}^{N}$ has the form*

$$\Pr(X_1, \ldots, X_N) = \int_{\mathscr{Q}} \prod_{i=1}^{N} F(X_i) \mathrm{d}Q(F) \tag{3.18}$$

*where $F$ is a distribution function with probability measure $Q$ defined in the set $\mathscr{Q}$ by*

$$Q(F) = \lim_{N \to \infty} \Pr(F_N) \tag{3.19}$$

*and $F_N$ is the empirical distribution function of $\{X_i\}_{i=1}^{N}$.*

For a proof of the above two theorems, see Bernardo and Smith (1994).

### *3.3.3 Learning from the past*

A key question for the problem of inference is under which circumstances past occurrences are informative with respect to future events. In the situation of sequential learning, where data are accumulated over time, we would like to be able to use past observations as a guide for what to expect in the next one. After all, if in a sequence of 50 tosses of what initially we believed was a fair coin we observed 45 heads, we would be foolish to expect with 50% probability that the 51st toss will be heads.

In sequential learning, we have gathered $n$ observations and we want to predict the probability of outcomes for the next one, observation $(n+1)$. In general, we can write the conditional probability for the $(n+1)$-th datum given the previous $n$, denoted by $\mathbf{X} = X_1, \ldots, X_n$, as:

$$\Pr(X_{n+1} \,|\, \mathbf{X}_n) = \frac{\Pr(X_{n+1}, \mathbf{X}_n)}{\Pr(\mathbf{X}_n)}. \tag{3.20}$$

If the variables are merely *independent*, then

$$\Pr(\boldsymbol{X}, X_{n+1}) = \prod_{i=1}^{n+1} \Pr(X_i) \Rightarrow \Pr(X_{n+1} \,|\, \mathbf{X}_n) \overset{(3.20)}{=} \Pr(X_{n+1}). \tag{3.21}$$

Therefore the predictive distribution for the $(n+1)$-th observation is identical to its marginal distribution, and it is not possible to use the past as a guide for

predicting the future. If the variables are *independent and identically distributed (i.i.d.)*, we additionally have:

$$\Pr(X_{n+1}) = \Pr(X_i) \quad \forall i = 1, \ldots, n \Rightarrow \mathbb{E}[X_{n+1}] = \mathbb{E}[X_i] = \mu, \tag{3.22}$$

hence an estimate of $\mu$ from past observations (say, the MLE) can be used to predict the outcome of the next, as they all have the same expectation. However, the i.i.d. condition, while sufficient, is not necessary for predictions, and we can rely on the weaker condition of exchangeability. By writing

$$\Pr(X_{n+1}, \mathbf{X}_n) = \int \mathrm{d}\theta \Pr(X_{n+1}, \mathbf{X}_n, \theta)$$

we can recast the predictive distribution for $X_{n+1}$ in Eq. (3.20) as:

$$\begin{aligned} \Pr(X_{n+1} \mid \mathbf{X}_n) &= \int \mathrm{d}\theta \frac{\Pr(X_{n+1}, \mathbf{X} \mid \theta) \Pr(\theta)}{\Pr(\mathbf{X}_n)} \\ &\overset{\text{cond. indep.}}{=} \int \mathrm{d}\theta \Pr(X_{n+1} \mid \theta) \Pr(\mathbf{X}_n \mid \theta) \frac{\Pr(\theta)}{\Pr(\mathbf{X}_n)} \\ &= \int \mathrm{d}\theta \Pr(X_{n+1} \mid \theta) \Pr(\theta \mid \mathbf{X}_n) \ , \end{aligned} \tag{3.23}$$

where we have made use of the fact that the $X_i$ are conditionally independent given $\theta$, a consequence of de Finetti's theorem under the assumption of exchangeable variables. This result says that the probability of the next observation is the average of its sampling distribution, $\Pr(X_{n+1} \mid \theta)$, over the posterior from the past $n$ observations.

This formulation includes a Frequentist viewpoint, as well: writing $\Pr(\theta \mid \mathbf{X}_n) = \delta(\theta - \theta_0)$ we can express the fact that $\theta$ has a fixed but unknown value $\theta_0$, hence $\Pr(X_{n+1} \mid \mathbf{X}_n) = \Pr(X_{n+1} \mid \theta_0)$, where in practice $\theta_0$ would be replaced by its maximum likelihood estimate from past data. Notice that this however neglects the uncertainty of the current estimate on $\theta$, which is instead captured by averaging over the posterior distribution, as in Eq. (3.23).

Hence, the assumption of exchangeability is sufficient for making predictions, but is it also necessary? Consider for example a sequence of RVs constructed as follows:

$$\begin{aligned} &\Pr(X_0) \sim \text{Bernoulli}(\theta) \text{ for } n = 0 \\ &\Pr(X_{n+1} = 1 \mid X_n = 0) = \Pr(X_{n+1} = 0 \mid X_n = 1) = 1 \text{ for } n \geq 1. \end{aligned}$$

This prescription leads to two, non-exchangeable possible sequences, namely:

$$\begin{cases} 1010101010101010\ldots & \text{with probability } \theta \\ 0101010101010101\ldots & \text{with probability } 1-\theta \end{cases} \tag{3.24}$$

For any observed subset of finite length $N \gg 1$, the observed frequency is

$$\sum_{i=0}^{N} x_i \simeq \frac{1}{2}, \tag{3.25}$$

but the probability of the next observation remains constant at 1, and does not match:

$$\Pr(X_{n+1} = 1 \mid X_n = 0) = 1 \neq \frac{1}{2}. \tag{3.26}$$

We therefore conclude that exchangeability is both necessary and sufficient for making predictions from past observations.

### 3.4 Reporting Inferences: the Bayesian Posterior Distribution

Bayesian inference on the parameters is based on the their posterior pdf, which is often approximated by numerical sampling techniques of the kind we will discuss in section **??**. The posterior $\Pr(|\,\boldsymbol{\theta}\boldsymbol{d})$ contains the full information about the parameters, and can be interpreted as a pdf in $\boldsymbol{\theta}$ (differently from the likelihood, as we emphasised above ). While the posterior distribution describes the full state of belief in the wake of the data, a point estimate is often used as a summary.

**Bayesian point estimates** for the parameters is offered by the posterior mean (which minimizes a quadratic loss function), the posterior median (which minimizes the absolute value loss) or the the value of the parameters that maximises the posterior, the so-called MAP ("maximum a posteriori") estimate:

$$\text{posterior mean: } \langle\boldsymbol{\theta}\rangle = \int \boldsymbol{\theta} \Pr(|\,\boldsymbol{\theta}\boldsymbol{d})\,\mathrm{d}\boldsymbol{\theta} \tag{3.27}$$

$$\text{posterior median: } \int_{-\infty}^{\boldsymbol{\theta}_{\text{med}}} \Pr(|\,\boldsymbol{\theta}\boldsymbol{d})\,\mathrm{d}\boldsymbol{\theta} = 0.5 \tag{3.28}$$

$$\text{MAP: } \boldsymbol{\theta}_{\text{MAP}} = \operatorname{argmax}_{\boldsymbol{\theta}} \Pr(|\,\boldsymbol{\theta}\boldsymbol{d})\,. \tag{3.29}$$

One- or two-dimensional posterior distributions for one or two parameters of interest at the time are obtained from the joint posterior by marginalizing (i.e. integrating over) the parameters that need to be eliminated, e.g., **the marginal posterior distribution** for $\theta_i$ is obtained as:

$$\Pr(|\,\theta_i\boldsymbol{d}) = \int \Pr(|\,\boldsymbol{\theta}\boldsymbol{d})\,\mathrm{d}\theta_{j\neq i}\,, \tag{3.30}$$

and similarly for two parameters at the time.

Since the marginal one-dimensional posterior is a pdf for the parameter in

question, a $100\alpha\%$ Bayesian **credible interval** $[\theta_i^a, \theta_i^b]$ is obtained simply by integrating the marginal posterior until a fraction $\alpha$ of its probability mass is contained within the interval:

$$\alpha = \int_{\theta_i^a}^{\theta_i^b} \Pr(|\, \theta_i \boldsymbol{d})\, \mathrm{d}\theta_i \,. \tag{3.31}$$

As there exist an infinite number of intervals with the above property, one needs to specify an additional rule to define the interval uniquely. One choice is to adopt a **symmetric (or equal-tailed) interval**, defined so that

$$\frac{1-\alpha}{2} = \int_{-\infty}^{\theta_i^a} \Pr(|\, \theta_i \boldsymbol{d})\, \mathrm{d}\theta_i = \int_{\theta_i^b}^{\infty} \Pr(|\, \theta_i \boldsymbol{d})\, \mathrm{d}\theta_i \,, \tag{3.32}$$

where $0 < \alpha < 1$. Another possibility is the **highest posterior density** (HPD) interval: this is the region that contains $100\alpha\%$ of probability, and chosen so that every point inside the region has higher probability density than any point outside. It is defined as the interval $R$ fulfilling:

$$\Pr(\theta \in R \mid \boldsymbol{d}) = \alpha \,, \tag{3.33}$$

with the property that, for $\theta_1 \in R$ and $\theta_2 \notin R$:

$$\Pr(\theta_1 \mid \boldsymbol{d}) \geq \Pr(\theta_2 \mid \boldsymbol{d}) \,. \tag{3.34}$$

By construction, this is also the shortest possible interval. The advantage of HPD over symmetric intervals is that the former generalize to non-unimodal posteriors, and flip automatically between 1– and 2– sided intervals.

See Jaynes and Kempthorne (1976) for an illuminating comparison between Frequentist and Bayesian intervals, as well as Loredo and Wolpert (2024) for a review of the advantages of Bayesian modelling.

### 3.5 Priors in Bayesian Inference

Priors are an essential ingredient of Bayesian inference. We can view them in many ways: de Finetti, taking a radical subjectivist approach, thought that they ought to represent each individual's degree of willingness to place a bet on the outcome of a measurement before any data was available. In this chapter, we explore the various ways of understanding priors, and practical implications for their choice.

Firstly, it's helpful to recall that, asymptotically and subject to mild regularity conditions, priors do not matter for inference (the situation is different for model comparison, as we shall explore in chapter 5), as guaranteed by the Bernstein-von Mises theorem, theorem 3.1. Since asymptotically classical theory also guarantees that $\hat{\theta}_{\mathrm{MLE}} \to \theta_0$ for $n \to \infty$, the posterior MAP converges to the true value,

 

**Example 3.2: Asymptotic disagreement between MLE and MAP**

Another case where the MLE and posterior disagree, even asymptotically, is the so-called 'Neyman-Scott problem', which arises in cases when there is a "structural" parameter, $\sigma$, common to all observations, and a number of "incidental" parameters, $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_N\}$, which grows linearly with the number of observations. For example, imagine taking $k$ measurements of the Gaussian-distributed fluxes of $N$ sources, $\mu_i$, with a measuring apparatus of unknown uncertainty, $\sigma$. In this case we have $kN$ measurements for $N+1$ parameters, and the joint likelihood is

$$\Pr\left(|\, \mathbf{x}\boldsymbol{\mu}, \sigma\right) = \prod_{i=1}^{N} (2\pi\sigma^2)^{-k/2} \exp\left[-\frac{\sum_{j=1}^{k}(x_{ij}-\mu_i)^2}{2\sigma^2}\right], \tag{3.35}$$

where $\mathbf{x} \in \mathbb{R}^{N\times k}$ is a matrix collecting the measurements $x_{ij}$. The parameter of interest is the variance $\sigma^2$, whose MLE is given by:

$$\sigma^2_{\mathrm{MLE}} = \frac{1}{kN}\sum_{i,j}(x_{ij}-\bar{x}_i)^2, \tag{3.36}$$

where $\bar{x}_i = \sum_{j=1}^{k} x_{ij}/k$. This estimator is not consistent[a], as $\lim_{N\to\infty}\sigma^2_{\mathrm{MLE}} = \sigma^2/k$. By contrast, the Bayesian MAP (maximum a posteriori) estimate is asymptotically unbiased.

[a] But note that Kiefer and Wolfowitz (1956) offer a proof of the consistency of the MLE in the particular case where the $\mu_i$ can be considered as being i.i.d. chance variables. Also, a simple solution is to re-state the problem to get rid of the incidental variables $\mu_i$ altogether.

irrespective of the prior (so long as the prior has support around the true value of the parameters). This also implies that, asymptotically, Bayesian credible intervals become identical to Frequentist confidence intervals. Eventually, the posterior will converge to a unique (objective) result even if different scientists start from different priors.

A simple illustration of this asymptotic converge is given by the case of a Gaussian prior and Gaussian likelihood (see exercise **??** and Fig. 3.1). We wish to infer the mean $\mu$ of a Gaussian, with a prior $\mu \sim \mathcal{N}\left(0, \Sigma^2\right)$, and $n$ data points with sampling distribution $x_i \sim \mathcal{N}\left(\mu_0, \sigma^2\right)$. Then the posterior for the mean is also Normal, and is given by

$$\Pr\left(\mu \mid \{x_1, \ldots, x_n\}\right) = \mathcal{N}\left(\bar{x}\frac{\Sigma^2}{\Sigma^2+\sigma^2/n}, \left[\frac{1}{\Sigma^2}+\frac{n}{\sigma^2}\right]^{-1}\right), \tag{3.37}$$

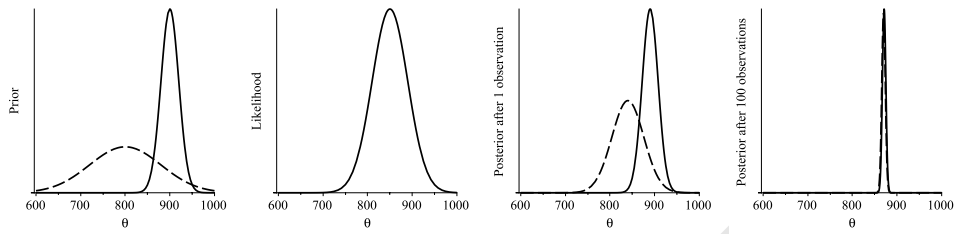


Figure 3.1 Asymptotically converging views in Bayesian inference. Two scientists having different prior beliefs $\Pr(\theta)$ about the value of a parameter $\theta$ (panel (a), the two curves representing two different priors) make one observation with likelihood $\mathscr{L}(\theta)$ (panel (b)), after which their posteriors $\Pr(\theta \mid \hat{x}_1)$ (panel (c)) represent their respective updated states of knowledge on the parameter. This posterior then becomes the prior for the next observation, and the Bayesian updating is repeated. After observing 100 data points, the two posteriors have become essentially indistinguishable (panel d).

where $\bar{x}$ is the observed sample mean, which is also the MLE for the true mean, $\mu_0$. For $n \to \infty$, the mean of the posterior converges to $\bar{x} \to \hat{x}_{\text{MLE}} \to \mu_0$, and thus the prior is forgotten.

While in this sense the prior influence disappears, it is true, however, that in most cases of interest one is not in this asymptotic regime, and indeed prior information might be highly relevant and contribute to the posterior in a meaningful way. One such example is the 'on/off problem'.

How does a Bayesian approach the on/off problem? The obvious additional information we may want to inject into the problem is to enforce the positivity of both source and background rates. Furthermore, if we are operating in a regime where the possibility that the signal might be zero cannot be ruled out, then it is important that the prior has some positive density at $s = 0$. One choice of prior that encapsulates such state of information is a uniform distribution with positive support, up to some cutoff value $c$:

$$\Pr(x) = \begin{cases} 1/c, & \text{if } 0 \le x \le c, \\ 0, & \text{otherwise;} \end{cases} \tag{3.40}$$

where $x = s, b$. The cutoff value $c$ is required to make the prior proper (i.e., normalizable), but is not very critical in this context (as long as it is sufficiently large not to cutaway significant posterior density, something that can be checked at the end of the calculation).

First, we obtain the joint posterior distribution for $s, b$ by applying the product

 

**Example 3.3: The on/off problem**

An instrument that can record single photons is aimed at a region of the sky containing a source of interest. The measured photons, $n_{\text{on}}$ photons in a time $t_{\text{on}}$, are the superposition of photons coming from the source plus "background" photons coming from other faint and unresolved sources within the field of view of the instrument (this is the "on source" measurement). The unknown background rate (also assumed Poisson) is given by $b$ (in photons per minute), which can be estimated by pointing the telescope away from the source (the "off" measurement) and measuring $n_{\text{off}}$ photons in time $t_{\text{off}}$. The MLE for the background is easily obtained by finding the maximum of the Poisson log-likelihood:

$$\frac{\partial}{\partial b}\left(n_{\text{off}}\ln(b t_{\text{off}}) - b t_{\text{off}}\right) = -t_{\text{off}} + \frac{n_{\text{off}}}{b} \stackrel{!}{=} 0 \rightarrow \hat{b}_{\text{MLE}} = n_{\text{off}}/t_{\text{off}}. \tag{3.38}$$

Since the total rate is the sum of the source plus background rate, the source rate $s$ is obtained by subtracting from the MLE for the total rate the MLE for the background rate, obtaining

$$s_{\text{MLE}} = \frac{n_{\text{on}}}{t_{\text{on}}} - \frac{n_{\text{off}}}{t_{\text{off}}}, \tag{3.39}$$

which can give a negative estimate when the Poisson counts $t_{\text{on}}$ fluctuates below $t_{\text{off}}$, which is obviously unphysical.

rule and then Bayes theorem (with $\mathbf{d} = \{n_{\text{on}}, t_{\text{on}}, n_{\text{off}}, t_{\text{off}}\}$):

$$\Pr(s, b \mid \mathbf{d}) = \Pr(s \mid b, \mathbf{d}) \Pr(b \mid t_{\text{off}}, n_{\text{off}}) \propto \Pr(n_{\text{on}} \mid s, b, t_{\text{on}}) \Pr(s) \Pr(n_{\text{off}} \mid b, t_{\text{off}}) \Pr(b) \tag{3.41}$$

$$\propto ((s+b)t_{\text{on}})^{n_{\text{on}}} (b t_{\text{off}})^{n_{\text{off}}} \exp\left(-s t_{\text{on}} - b(t_{\text{on}} + t_{\text{off}})\right) \text{ for } 0 \leq s, b \leq c. \tag{3.42}$$

Next, we integrate the joint posterior over $b$ to obtain the marginal posterior for the signal, accounting for uncertainty in the background:

$$\Pr(s \mid \mathbf{d}) = \int_0^\infty \Pr(s, b \mid \mathbf{d})\, db. \tag{3.43}$$

To carry out the integral analytically, we take $c \rightarrow \infty$ (although this amounts to an improper prior, which however in this case is harmless[9]), use the binomial

[9] We could also use a conjugate prior, which has the advantage of simplifying the calculation and has infinite support on the positive axis. This is introduced in section 3.5.5.

expansion $(s+b)^{n_{\rm on}} = \sum_{k=0}^{n_{\rm on}} \binom{n_{\rm on}}{k} s^k b^{n_{\rm on}+n_{\rm off}-k}$ and the Gamma integral $\Gamma(n) = \int_0^\infty t^{n-1}e^{-t} = (n-1)!$ for $n \in \mathbb{N}$, obtaining

$$\Pr(s \mid \mathbf{d}) \propto \sum_{k=0}^{n_{\rm on}} W_k(\boldsymbol{d}) \frac{(s t_{\rm on})^k e^{-s t_{\rm on}}}{k!}, \tag{3.44}$$

where each term in the posterior assigns $0 \le k \le n_{\rm on}$ photons to the source, each with weight

$$W_k(\boldsymbol{d}) = \frac{(n_{\rm on}+n_{\rm off}-k)!}{(n_{\rm on}-k)!}\left(1+\frac{t_{\rm off}}{t_{\rm on}}\right)^k. \tag{3.45}$$

Three examples are plotted in Fig. 3.2, which shows how the Bayesian posterior transitions from a clear detection of the signal (green) to indicating compatibility with no source ($s=0$) in the case where the number of photons per unit time observed on source is smaller than the number of photons per unit time from the background (blue). In the former case, the MAP is near the MLE value, $s_{\rm ML} = n_{\rm on}/t_{\rm on} - n_{\rm on}/t_{\rm off} = 7.5$, while in the latter the MAP is (in accord with intuition) at $s=0$, in contrast to the MLE, which gives a negative (and unphysical) estimate, $s_{\rm ML} = -2$.

When discussing priors, the issue of 'objectivity' often comes up. An objection against Bayesianism is often raised because of its reliance on priors to even start any Bayesian calculation: since there is no agreed-upon way of assigning priors, their choice is essentially subjective and this is at odds with a scientific, objective approach. This argument however can be countered from many different directions. First, we must recognize that different priors might correctly reflect different states of knowledge or belief, and there is nothing wrong with that; it is consistent reasoning to obtain different conclusions from the same data if starting from different states of beliefs. Secondly, some states of belief are in conflict with existing information or knowledge, and therefore ought to be discarded—which can be achieved by an appropriate choice of prior, which incorporates objective knowledge, as we have seen in the above example.

While there might not exist a single 'right' choice of prior, there are many wrong ones. We now introduce some of the most important considerations when selecting priors appropriately. Since there is no universally agreed-upon solution, we will put particular emphasis on why a particular rule makes sense in a given context. In any case, it is generally good practice to carry out a sensitivity analysis,i.e., varying the prior within reasonable ranges and/or choosing different possibilities and evaluating any changes in the posterior. This has the aim of checking the impact of prior choices on posterior distributions: if the posteriors remain stable, that is a good indication that the inference is dominated by the data, and one might be in an asymptotic, almost prior-free regime; if in-

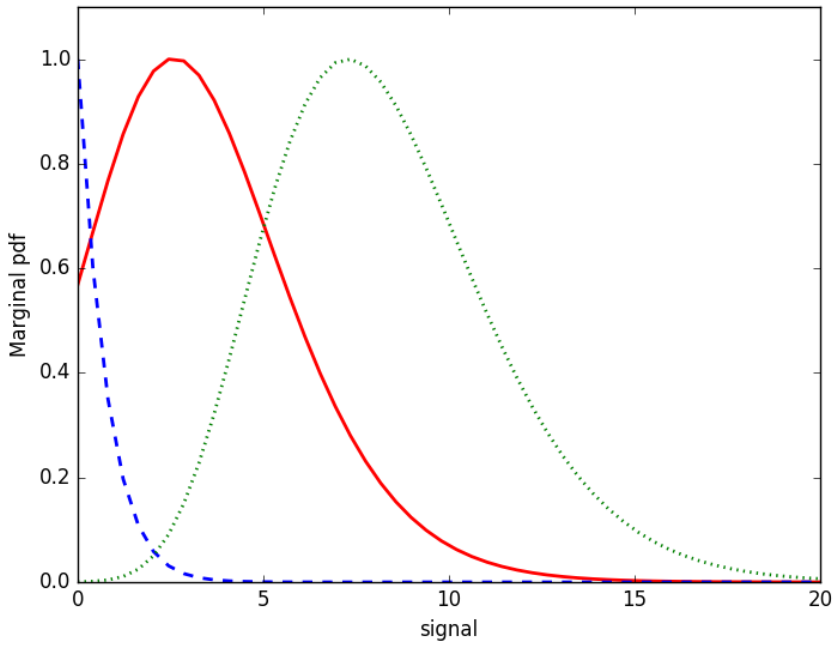


Figure 3.2 Marginal posteriors (normalized to their peak) for the signal rate $s$ for the on/off problem, with uniform priors, for three cases: $n_{\rm on} = 8, t_{\rm on} = 1, n_{\rm off} = 2, t_{\rm off} = 4$ (green); $n_{\rm on} = 5, t_{\rm on} = 1, n_{\rm off} = 2, t_{\rm off} = 1$ (red); $n_{\rm on} = 2, t_{\rm on} = 2, n_{\rm off} = 3, t_{\rm off} = 1$ (blue);

stead the posterior changes appreciably (and with it the scientific conclusion), that points to the fact that the data alone are not sufficient to override the prior, in which case extra care must be taken in ensuring that the assumptions going into choosing a particular priors are justified and a correct reflection of one's state of knowledge before seeing the data. In any case, it is good practice to justify explicitly the information that the prior is encapsulating, for this makes it transparent. The default prior choice of a 'uniform' prior in whatever parameters one is considering is often not free from problems, and needs to be assessed critically, especially in high dimensions, as we further discuss below.

It is perfectly acceptable to use the posterior from a previous observation or experiment as prior for a new measurement. Bayesian updating ensures that the final inference is independent on whether the old and new data are considered jointly (thus updating with both data sets the original prior), or sequentially. This is shown explicitly by writing the previous measurement's data as $\boldsymbol{d}_0$, and the

 

new ones as $\boldsymbol{d}_1$:

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{d}_0, \boldsymbol{d}_1) \propto \Pr(\boldsymbol{d}_0, \boldsymbol{d}_1 \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) = \Pr(\boldsymbol{d}_1 \mid \boldsymbol{\theta}) \Pr(\boldsymbol{d}_0 \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta}) = \Pr(\boldsymbol{d}_1 \mid \boldsymbol{\theta}) \Pr(\boldsymbol{\theta} \mid \boldsymbol{d}_0) \,, \tag{3.46}$$

which is valid provided that $\boldsymbol{d}_0, \boldsymbol{d}_1$ are exchangeable. The last equality shows that the posterior obtained by updating the original prior $\Pr(\boldsymbol{\theta})$ with the information in both datasets is identical to what is obtained by using only data $\boldsymbol{d}_1$ to update the posterior obtained from $\boldsymbol{d}_0$ and the original prior.

Good reviews can be found in Kass and Wasserman (1996) and Jaynes (2003) (see Chap. 11, 12 for symmetry-based arguments).

**Example 3.4: Effective likelihood for upper limits**

See hep example Also, discuss how to marginalize over errorbar rescaling parameters

### *3.5.1 Maximum Entropy Priors*

It would seem at first sight appealing to select priors based on 'total ignorance', so as to let 'the data speak for themselves'. While this is a fine principle, the fact is that 'total ignorance' about the parameters *a priori* cannot be expressed mathematically in the form of a prior, unless some further information is provided about precisely *what kind of ignorance* one wants to encode.

One powerful idea in this context is that of symmetry – often invoked in physics, and very useful in statistical analysis as well (as we saw above with the notion of exchangeability). In the case in which one has no reason to prefer any outcome over the others among a finite, discrete set of $n$ possibilities for the random variable $X$ (e.g. in a card game, or in dice tossing), then Laplace's 'Principle of insufficient reason' (1820), also called 'principle of indifference', posits that one ought to assign equal prior probabilities to each outcome:

$$\Pr(X = i) = \frac{1}{n}, \text{ for } i = 1, \ldots, n. \tag{3.47}$$

This can also be justified more formally from an information-theoretical point of view: the **maximum entropy principle** states that we should adopt the probability distribution that, subject to what is known *a priori*, maximises the Shannon information entropy, which describes the amount of randomness or uncertainty in a probability distribution.

 

**Definition 3.6** For a discrete RV $X$ which can take on $n$ possible outcomes, $x_1, \ldots, x_n$, each with probability $\Pr(x_i)$, the **Shannon entropy** is defined as:

$$H(X) \equiv -K \sum_{i=1}^{n} \Pr(x_i) \ln \Pr(x_i) \,, \tag{3.48}$$

where $K$ is a positive constant.

If there is symmetry about the possible outcomes (e.g., tossing an $n$-faced die) and nothing else is known but that the probabilities $p_i \equiv \Pr(x_i)$ must sum to unity, $\sum_{i=1}^{n} p_i = 1$, the maximum of the information entropy is obtained by introducing the Lagrange multiplier $\lambda$ and solving the variational problem:

$$\delta \left[ -K \sum_{i=1}^{n} p_i \ln p_i - \lambda \left( \sum_{i=1}^{n} p_i - 1 \right) \right] = 0 \,. \tag{3.49}$$

By setting the total derivative to 0 we obtain:

$$\sum_{i=1}^{n} \exp(\lambda/K - 1) = 1 \rightarrow \exp(\lambda/K - 1) = \frac{1}{n} \tag{3.50}$$

$$p_j = \exp(\lambda/K - 1) \rightarrow p_j = \frac{1}{n}. \tag{3.51}$$

The maximum entropy principle can also be used in cases where additional information is available: for example, if one knows *a priori* that the average number of pips that came up in a large number of tosses of a die is $m$, this can be added to the variational problem by means of an additional Lagrange multiplier term[10], $\mu\left(\sum_{i=1}^{n} i\, p_i - m\right)$.

To extend the maximum entropy principle to the case of continuous random variables, we must modify the discrete entropy expression of Eq. (3.48), since replacing $\Pr(x_i)$ (which is dimensionless) with a density $\Pr(\theta)$ (which has units) would lead to the entropy being dependent on the units in which $\theta$ is expressed, which is undesirable. Therefore, we consider instead a relative entropy measure, namely the **Kullback-Leibler (KL) divergence** between distributions $P$ and $Q$, noted $D_{\mathrm{KL}}(P||Q)$:

$$D_{\mathrm{KL}}(P||Q) = -\int \mathrm{d}\theta P(\theta) \ln\left(\frac{P(\theta)}{Q(\theta)}\right). \tag{3.52}$$

We wish to maximse the KL divergence between a prior $\Pr(\theta \mid I)$, subject to whatever information $I$ is available, and a 'reference measure' $m(\theta)$. If nothing is known about $\theta$ except that it is in the interval $[a, b]$, then maximising

[10] This problem is called 'the Brandeis dice problem', after a lecture E.T. Jaynes gave at Brandeis University in 1962; the solution is actually quite a bit more convoluted, and fuller details can be found in van Enk (2014).

$D_{\mathrm{KL}}(\Pr(\theta \mid I)||m(\theta))$ with respect to $\Pr(\theta \mid I)$ subject to the normalization constraint $\int_a^b \Pr(\theta \mid I)\,\mathrm{d}\theta = 1$, leads to the solution

$$\Pr(\theta \mid I) = m(\theta)\left[\int_a^b m(x)\mathrm{d}x\right]^{-1} \quad \text{for } a < \theta < b. \tag{3.53}$$

We have merely pushed back the problem of describing 'complete ignorance' to the problem of determining the reference measure $m(\theta)$ in the range of interest. By partitioning the interval $[a,b]$ into $n$ discrete outcomes $p_i, (i = 1,\ldots,n)$, then taking the limit $n \to \infty$, the result for the discrete case, Eq. (3.50), leads to a uniform distribution[11] for $m(\theta)$, and therefore to the **uniform prior** (often called 'flat prior'):

$$\Pr(\theta \mid I) = \begin{cases} \frac{1}{a-b} & \text{for } a < \theta < b, \\ 0 & \text{otherwise.} \end{cases} \tag{3.54}$$

What if we want to additionally specify the first and second moment of the distribution, i.e. its mean, $\mu$, and its variance, $\sigma^2$? We adopt a uniform measure $m(\theta) = 1$ (which, strictly speaking is improper as it is not normalizable; however, it only needs to apply in a finite –if large– range $[a,b]$, which effectively limits the range of $\mu$), and we seek the probability density function $\Pr(\theta)$ that maximizes $D_{\mathrm{KL}}(\Pr(\theta)||1)$ subject to these constraints:

$$\int_{-\infty}^{\infty} x \Pr(\theta)\,\mathrm{d}\theta = \mu. \tag{3.55}$$

and

$$\int_{-\infty}^{\infty} (\theta-\mu)^2 \Pr(\theta)\,\mathrm{d}\theta = \sigma^2. \tag{3.56}$$

We now have three Lagrange multipliers, $\lambda_i$ $(i = 0,1,2)$ (one for each moment constraints, plus the normalization requirement), and the Lagrangian is:

$$L = -\int_{-\infty}^{\infty} \Pr(\theta)\log\Pr(\theta)\,\mathrm{d}\theta + \lambda_0\left(\int_{-\infty}^{\infty}\Pr(\theta)\,\mathrm{d}\theta - 1\right) + \lambda_1\left(\int_{-\infty}^{\infty}\theta\Pr(\theta)\,\mathrm{d}\theta - \mu\right) + \lambda_2\left(\int_{-\infty}^{\infty}(\theta-\mu)^2\Pr(\theta)\,\mathrm{d}\theta - \sigma^2\right).$$

Taking the the functional derivative of $L$ with respect to $\Pr(\theta)$ and setting it to zero yields:

$$\Pr(\theta) = e^{\lambda_0 - 1} e^{\lambda_1 \theta} e^{\lambda_2(\theta-\mu)^2}.$$

Using the constraints on the mean, Eq. (3.55), leads to $\lambda_1 = 0$, while Eq. (3.56) gives $\lambda_2 = -1/2\sigma^2$. The normalization requirement determines $\lambda_0$ and we thus

[11] Using an invariance principle, the uniform distribution can also be derived as the correct prior for problems that are translation-invariant.

conclude that the distribution that has maximum entropy given a first and second moment is the Gaussian:

$$\Pr(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}.$$

A similar argument shows that the maximum entropy distribution subject to the constraints that $\theta \geq 0$ and with mean $\mu$ is given by the exponential distribution, $\text{Exp}(\mu)$.

### *3.5.2 Caveats about uniform priors*

From the above derivation of a uniform prior from maximum entropy it is tempting to conclude that a uniform prior is the choice that maximises ignorance, and therefore a good default in cases where little else can be established about a parameter a priori. However, it is important to remark that a uniform prior on the continuous variable $\theta$ is *not* uniform for a non-linear reparametrisation, due to the transformation law for probability densities. This lack of invariance under a general reparameterization of the problem means that the 'state of indifference' expressed by a uniform prior is only valid for the specific parameterization for which it was chosen – which often does not have a more fundamental justification than modelling convenience. We will later introduce a parameterization-invariant prior that seeks to address this problem, see section 3.5.4.

Recall the **transformation law for continuous random variables**: given a random variable $X$ with density $f_X(x)$, the probability contained within a range $[a, b]$ is

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x)\mathrm{d}x. \tag{3.57}$$

Under a one-to-one reparametrisation $Y = t(X)$, with inverse $t^{-1}$, the same probability is expressed as the integral of the density between the transformed limits:

$$\Pr(t(a) \leq Y \leq t(b)) = \int_{t(a)}^{t(b)} f_X\left(t^{-1}(y)\right) \left|\frac{\mathrm{d}t^{-1}(y)}{\mathrm{d}y}\right| \mathrm{d}y = \Pr(a \leq X \leq b).$$

The p.d.f. of $Y$ must therefore be

$$f_Y(y) = f_X\left(t^{-1}(y)\right) \left|\frac{\mathrm{d}t^{-1}(y)}{\mathrm{d}y}\right|. \tag{3.58}$$

In the general multivariate case, where $Y_i = t_i(X_1, \ldots, X_n)$, for $i = 1, \ldots, n$, similar considerations lead to the transformation law for densities:

$$f_Y(y_1, \ldots, y_n) = f_X(t_1^{-1}(\mathbf{y}), \ldots, t_n^{-1}(\mathbf{y})) \mid \det(\mathbf{J}) \mid, \tag{3.59}$$

where the matrix **J** is the Jacobian of the transformation, with entries

$$J_{ij} = \frac{\partial t_i^{-1}(\mathbf{y})}{\partial y_j}, \text{ for } \{i, j\} = 1, \ldots, n.$$

Thus, when the Jacobian is not constant, i.e., the inverse transformation $t_i^{-1}(\mathbf{y})$ is non-linear, the uniform prior on $X$ is not uniform on $Y$ (see example 3.5.2.

**Example 3.5: Success probability of coin tossing**

To see how a uniform prior can become informative under reparameterization of the problem, consider a prior $\theta \sim U(0,1)$ for the success probability of a Bernoulli variable (e.g. the toss of a coin to give heads). We may want to rephrase the question in terms of the log-odds of success, expressed by the variable $\phi \equiv t(\theta) = \log[\theta/(1-\theta)]$. This variable transformation maps the $[0,1]$ interval onto the entirety of the real axis, and it is often used in logistic regression.
The Jacobian of the inverse transformation, $t^{-1}(\phi) = \exp(\phi)/(1+\exp(\phi)) = \theta$ is given by

$$J(\phi) = \frac{\mathrm{d}t^{-1}(\phi)}{\mathrm{d}\phi} = \frac{\exp(\phi)}{(1+\exp(\phi))^2},$$

and therefore the prior $\Pr(\theta) = 1$ becomes an informative prior $\Pr(\phi) = J(\phi)$ – a logistic distribution with mean 0 and variance $\pi^2/3$.

Another important *caveat* comes with the adoption of uniform prior in high-dimensional spaces, where a phenomenon called **concentration of measure** leads to the prior density being highly localized in parameter space. Consider a parameter vector $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_n\}$ with a uniform prior over the range $[-1,1]$ for each $\theta_i$, $i = 1, \ldots, n$. What is the distribution of lengths of vectors in the $n$-dimensional space? The result is rather surprising: defining $\xi_n = \sum_{i=1}^n \theta_n^2$, we seek the expectation value of the Euclidean distance from the origin, $\mathrm{E}\left[\sqrt{\xi_n}\right]$. This is not a simple calculation, for the expectation and the square root do not commute, so an approximate result will suffice.

We consider $Y_n = \sum_{i=1}^n X_i$, where $X_i$ are i.i.d. RVs with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$ $(i = 1, \ldots, n)$. Defining $Z_n^2 = Y_n$, and rearranging the definition of variance for the variable $Z_n$, we can write

$$\mathrm{E}[Z_n] = \left(\mathrm{E}\left[Z_n^2\right] - \mathrm{Var}[Z_n]\right)^{1/2}. \tag{3.60}$$

We replace $\mathrm{E}\left[Z_n^2\right] = \mathrm{E}\left[Y_n\right] = \mu n$, and we use the Central Limit Theorem to estimate $\mathrm{Var}\left[Z_n\right] = \mathrm{Var}\left[\sqrt{Y_n}\right]$: asymptotically, we have that, for $0 \leq \alpha \leq 1$:

$$\Pr\left(\frac{Y_n - \mu n}{\sigma\sqrt{n}} \leq \alpha\right) \to \mathrm{CDF}_N(\alpha)\,,$$

where $\mathrm{CDF}_N(\alpha)$ is the CDF of the standard Normal. For $n \to \infty$ we can subtract a vanishingly small term to the argument and we will still have

$$\Pr\left(\frac{Y_n - \mu n}{\sigma\sqrt{n}} - \frac{\alpha^2\sigma}{4\mu}\frac{1}{\sqrt{n}} \leq \alpha\right) \to \mathrm{CDF}_N(\alpha)\,,$$

which can be recast as:

$$\Pr\left(Y_n \leq \left(\frac{\alpha\sigma}{2\sqrt{\mu}} + \sqrt{\mu n}\right)^2\right) \to \mathrm{CDF}_N(\alpha)\,,$$

which implies that the RV $(\sqrt{|\,Y_n\,|} - \sqrt{\mu n})/(\sigma/2\sqrt{\mu})$ converges in distribution to a standard Normal. Hence we conclude that $\mathrm{Var}\left[\sqrt{Y_n}\right] \approx \sigma^2/(4\mu)$, where we have dropped the absolute value, and we must ensure in the following that we only have positive variables. Substituting this back into Eq. (3.60), using $X_i = \theta_i^2$ and considering because of symmetry only the range $\theta_i \sim U(0,1)$, we have for each RV $X_i$ that $\mu = 1/3$, $\sigma^2 = 4/45$, and we obtain:

$$\mathrm{E}\left[\sqrt{Y_n}\right] \approx \left(\mu n - \frac{\sigma^2}{4\mu}\right)^{1/2} = \sqrt{\frac{n}{3} - \frac{1}{15}} \quad , \quad \mathrm{Var}\left[\sqrt{Y_n}\right] \approx \frac{1}{15}\,. \tag{3.61}$$

We see that the mean of the Euclidean distance scales as $\sqrt{n}$ for large $n$ with *constant variance*, which implies that samples drawn uniformly from the hypercube are increasingly concentrated on a shell, whose thickness relative to the maximum available radius, $\sqrt{n}$, vanishes as $1/\sqrt{n}$, as illustrated in Fig. 3.3. Therefore, the largest part of the available volume has essentially no samples, even though each fo the coordinates is drawn uniformly in Cartesian coordinates! For example, in $n = 20$ dimensions the origin is about 10 standard deviations away from the shell, and therefore the prior effectively disappears there (right panel in Fig. 3.3).

A related high-dimensional concept is that of **the typical set**: the idea that the locus in either parameter or data space[12] where most of the samples lie is radically different from their most probable location. A striking illustration is offered by the following example (Loredo and Wolpert, 2024). If you draw $n$ samples from a standard Normal, the most probable location is at the origin for all of them (where the Gaussian peaks). Yet, similarly to what we have seen above, their squared distance from the origin follows a $\chi^2$ distribution with $n$ degrees of

[12] From a Bayesian point of view, it does not matter which we are considering as they are both RVs.

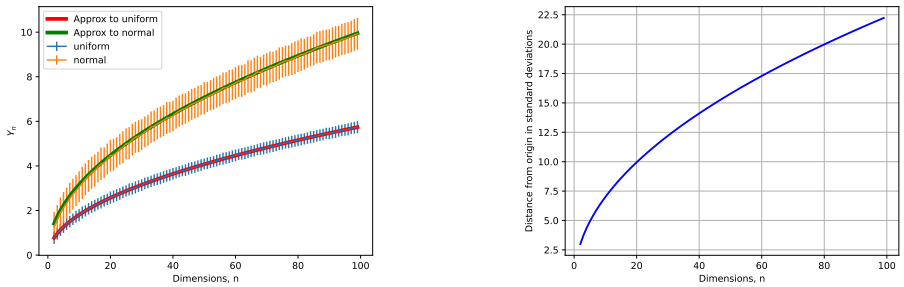


Figure 3.3 Illustration of the phenomenon of the concentration of measure in parameter spaces with a large number of dimensions, $n$. Left: distribution of the Euclidean norm from the origin for samples drawn uniformly in the range $[-1,1]$ in each Cartesian dimension, as a function of dimensionality. Right: distance of the mean Euclidean norm from the origin in units of the (constant) standard deviation.

freedom, which has expected value $n$ and variance $2n$ (the familiar 'chi-squared per degrees of freedom equals one' statement). This means that almost the totality of the samples congregate well away from the most probable location (the mode of the distribution). Similarly to the previous example, the typical set lies on a spherical shell at a distance $\sqrt{n}$ from the origin, and whose standard deviation is constant (see Fig. 3.3).

### *3.5.3 Scale-Invariant Priors*

To express in a mathematically well-defined manner a state of 'indifference' in terms of priors, we turn to a principle of invariance: we seek a prior that remains invariant under a reparameterizeation that leaves the problem invariant.

We are often in a situation where **scale invariance** is desirable in a prior, i.e., a case in which our state of ignorance (or information) on a parameter ought not to change if we perform a change of the units in which it is expressed. More precisely, consider a distribution $\Pr(\theta)$ and a change of units such that $\theta \to \psi = \theta/a$, for some constant $a > 0$. Then the distribution over $\psi$, in virtue of Eq. (3.58) is given by

$$\Pr(\psi) = \Pr(\theta) \left| \frac{\partial \theta}{\partial \psi} \right| = \Pr(\theta)\, a\,. \tag{3.62}$$

Requiring that $\Pr(\theta) = \Pr(\psi)$ leads to the condition:

$$\Pr(\theta) = \frac{1}{a} \Pr(\psi) = \frac{1}{a} \Pr\left(\frac{\theta}{a}\right), \tag{3.63}$$

which admits the solution

$$p(\theta) \propto \frac{1}{\theta}. \tag{3.64}$$

This is an improper prior, as it diverges for $\theta \to 0$, and so it requires a cutoff at some minimum value. The interpretation is illuminated by carrying out a change of variable, $\psi = \ln\theta$, whose density is given by, applying Eq. (3.58) with $t^{-1}(\psi) = \exp(\psi)$:

$$p_Y(\psi) = p_X(t^{-1}(\psi))\left|\frac{\mathrm{d}t^{-1}(\psi)}{\mathrm{d}\psi}\right| \propto \frac{1}{\theta}e^{\psi} = 1, \tag{3.65}$$

that is, the prior on $\psi$ is uniform. A scale-invariant prior thus assigns the same probability to all orders of magnitude for the parameter, as it is uniform in the logarithm of the parameter. This accords with our intuition: if we are ignorant on the scale of the problem, then all scales ought to be equally probably *a priori*. Expressed in terms of the original variable $\theta$, this prior is often called **log-uniform**.

The choice of a log-uniform prior is useful when we don't want to single out any specific order of magnitude for a quantity about which little is known. An example application in cosmology is that of the amplitude of isocurvature perturbations in the CMB. The drawback however is that, in absence of a measurement from the data, the resulting posterior gives an upper bound to the value of the parameter that depends on the lower cutoff chosen for the prior, since typically the likelihood becomes flat below a certain value. Since the lower cutoff is often arbitrary, this introduces a prior-dependence in the upper limit. This can be circumvented, for example, by choosing a lower cutoff that is motivated by the expected sensitivity of the experiment: values lower than the sensitivity cannot be detected, and therefore can be excluded from the analysis (Trotta, 2007a).

### *3.5.4 Jeffreys' Prior*

Now that we have gained familiarity with simple examples, we introduce the prior compatible with the most general formulation of the principle of invariance: Jeffreys' prior[13].

The aim is to define a prior that will leave the *posterior p.d.f.* invariant under a general, non-linear monotone reparametrisation, thus countering the criticism levelled against uniform priors – that their 'indifference' to the value of a param-

[13] After Sir Harold Jeffreys (1891–1989), British mathematician, geologist and astronomer. A fellow of St John's in Cambridge, he is credited with the revival of Bayesian analysis and major contributions to geophysics, including establishing that the Earth's core is liquid. He also held the Plumian Professorship of Astronomy and Cambridge. Notice that the final -s in 'Jeffreys' prior' is part of his name, not a possessive!

eter does not survive a non-linear transformation. Jeffreys (1946) suggested that the property encoding such an invariant prior could be obtained as follows.

Consider the sampling distribution $X \sim \Pr(x \mid \theta)$ for data $x$, conditional on a parameter $\theta$. After a monotone transformation $\psi = t(\theta)$, the sampling distribution is given by $X \sim f(x \mid t^{-1}(\psi))$. Suppose that the invariance principle, applied to both the original and the reparameterized sampling distribution, gives a prescription for the prior on their respective parameters, denoted by $\pi^J_\theta(\theta)$ and $\pi^J_\psi(\psi)$, respectively. If one started from such an invariant prior for the original sampling distribution and then carried out the variable transformation, one would obtain the transformed prior, according to Eq. (3.58):

$$\pi_\psi(\psi) = \pi^J_\theta(t^{-1}(\psi)) \left| \frac{\mathrm{d}t^{-1}(\psi)}{\mathrm{d}\psi} \right|. \tag{3.66}$$

Jeffreys then demanded that this transformed prior should be the same as the prior that one would have obtained by applying the invariance principle to the transformed sampling distribution, i.e,. that $\pi_\psi(\psi) = \pi^J_\psi(\psi)$. We shall see below that this requirement leads to posterior distributions that are invariant with respect to the chosen parameterization of the problem, i.e., one obtains the same posterior regardless of how the problem is represented.

To construct such a prior, consider (in one dimension) the **Fisher information** of a random variable X about the parameter $\theta$, defined as:

$$I_X(\theta) \equiv \int \left( \frac{\partial}{\partial \theta} \ln f(x \mid \theta) \right)^2 f(x \mid \theta) \mathrm{d}x, \tag{3.67}$$

where $f(x \mid \theta)$ is the sampling distribution for the RV $X$ given the parameter $\theta$. The derivative $\frac{\partial}{\partial \theta} \ln f(x \mid \theta)$ is called **the score function**, and it describes how the model for the RV $X$, given by the functional relation $f$, reacts to small changes in the parameter $\theta$ at any value of $\theta$. The integral over $x$ makes this into an expectation value over all possible outcomes for the RV (for more details on the Fisher information, see e.g. Ly et al. (2017)). It is interpreted as the information content of the random variable $X$ about the parameter $\theta$ in virtue of the **Cramér-Rao Lower Bound**, which states that the variance of any unbiased estimator $\hat{\theta}$ for the parameter $\theta$ of a model $f(X \mid \theta)$ is asymptotically bounded from below by:

$$\mathrm{Var}\left[\hat{\theta}\right] \geq \frac{1}{n I_X(\theta)}, \tag{3.68}$$

where $n$ is the number of i.i.d. random samples. This means that the larger the Fisher information, the smaller the variance of the estimator, and therefore the more informative $X$ is about the parameters. As we saw earlier, the MLE saturates this bound. For i.i.d. RVs $X_i \sim f(X \mid \theta)$, the Fisher information is additive, meaning that $I_{\mathbf{X^n}}(\theta) = n I_X(\theta)$, where $\mathbf{X^n} = \{X_1, \ldots, X_n\}$.

Jeffreys (1961) proposed as an invariant prior the square-root of the Fisher information, which is today called the **Jeffreys' prior**:

$$\pi^J(\theta) \propto \sqrt{I_X(\theta)}. \tag{3.69}$$

We now show that the Jeffreys' prior indeed fulfils Jeffreys' demand of invariance, by considering a general one-to-one parameter transformation $\psi = t(\theta)$, under which the sampling distribution $f(X \mid \theta) \rightarrow g(X \mid t(\theta))$, where $g(X \mid \psi) = f(X \mid t^{-1}(\psi))$. Under such a reparameterization, the Fisher information transforms as:

$$\begin{aligned} I_X(\theta) &= \int \left( \frac{\partial}{\partial \theta} \ln f(x \mid \theta) \right)^2 f(x \mid \theta) \mathrm{d}x && (3.70) \\ &= \int \left( \frac{\partial}{\partial \theta} \ln g(x \mid t(\theta)) \right)^2 g(x \mid t(\theta)) \mathrm{d}x && (3.71) \\ &= \int \left( \frac{\partial}{\partial \psi} \ln g(x \mid \psi) \frac{\partial \psi}{\partial \theta} \right)^2 g(x \mid \psi) \mathrm{d}x && (3.72) \\ &= I_X(\psi) \left( \frac{\mathrm{d}\psi}{\mathrm{d}\theta} \right)^2, && (3.73) \end{aligned}$$

where in we have used the chain rule for differentiation. With the definition of Eq.(3.69), we have that

$$\pi^J(\theta) = \pi^J(\psi) \left| \frac{\mathrm{d}\psi}{\mathrm{d}\theta} \right|, \tag{3.74}$$

from which follows, noting that, for any function $y = t(x)$ with inverse $x = t^{-1}(y)$ it holds that $\mathrm{d}t/\mathrm{d}y = 1/(\mathrm{d}t^{-1}/\mathrm{d}x)$:

$$\pi^J(\psi) = \frac{\pi^J(\theta)}{\left| \frac{\mathrm{d}t(\theta)}{\mathrm{d}\theta} \right|} = \pi^J(t^{-1}(\psi)) \left| \frac{\mathrm{d}t^{-1}(\psi)}{\mathrm{d}\psi} \right| \tag{3.75}$$

which fulfils Jeffreys' requirement of invariance, Eq. (3.66).

We can now demonstrate that the posterior obtained from a Jeffreys' prior is invariant under reparameterization of the problem: this does not mean that the posterior has the same functional form for any parameterization, but rather that the same conclusions about $\theta$ are reached (as encoded in the posterior) whether we (1) apply Jeffreys rule to construct a prior on $\theta$ and update it with the observed data, or (2) apply Jeffreys rule to construct a prior on $\psi$, update to a posterior distribution on $\psi$, and then transform this to a posterior on $\theta$. Considering once again the reparametrisation $\psi = t(\theta)$, and let us carry out (2) above. The posterior for $\theta$, obtained from transforming the posterior on $\psi$ is

given by:

$$\begin{aligned}\Pr(\theta \mid x) &\propto \Pr\left(\psi \mid x\right)\left|\frac{\mathrm{d}t(\theta)}{\mathrm{d}\theta}\right| \\ &= f(x \mid \psi)\pi^J_\psi(\psi)\left|\frac{\mathrm{d}\psi}{\mathrm{d}\theta}\right| \\ &= f(x \mid t(\theta))\pi^J_\theta(\theta)\,, \end{aligned} \tag{3.76}$$

where we have used Eq. (3.74) in the last passage. Since the likelihood is invariant under reparameterization, $f(x \mid t(\theta))$ can be interpreted as conditional on $\theta$, and therefore the RHS of Eq. (3.76) equals the posterior that would have been obtained by (1), i.e., starting from $\theta$ and its corresponding Jeffreys' prior.

Geometric interpretation of the Jeffreys' prior: use of the Fisher-Rao metric induced by the prior

**Example 3.6: Examples of Jrffreys' prior**

We now compute Jeffreys' prior for the parameters of a univariate normal likelihood and $n$ i.i.d. observations, obtaining:

$$\pi^J(\mu) = \frac{\sqrt{n}}{\sigma} = \text{const} \quad \text{and} \quad \pi^J(\sigma) \propto \frac{\sqrt{n}}{\sigma}\,, \tag{3.77}$$

which is consistent with non-informative priors for the mean (a uniform prior) and the standard deviation (a scale-invariant prior).
For $n$ i.i.d. observations from a binomial distribution, we have

$$\pi^J(\theta) \propto \frac{\sqrt{n}}{\theta(1-\theta)}\,. \tag{3.78}$$

For multidimensional parameter vectors $\boldsymbol{\theta}$, Jeffreys' prior generalizes to:

$$p^J(\boldsymbol{\theta}) \propto \left(\det \boldsymbol{I}_X(\boldsymbol{\theta})\right)^{1/2} \tag{3.79}$$

where the **Fisher information matrix** $\boldsymbol{I}_X(\boldsymbol{\theta})$ is given in Eq. (2.78).

However, when both scale and location parameters are present simultaneously, the above prior must be applied with caution. Consider as an illustration the case of a Normal likelihood, and consider both parameters at the same time: we have $\boldsymbol{\theta} = \{\mu, \sigma\}$ and the determinant of the Fisher matrix is given by:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & 2\sigma^{-2} \end{pmatrix} \Rightarrow p^J(\mu, \sigma) \propto \frac{1}{\sigma^2}, \tag{3.80}$$

 

which differs from what we obtained by considering the variance as a standalone parameter, namely $1/\sigma$. Jeffreys' viewed this feature as a shortcoming of his rule and suggested the following solution: whenever location parameters $\boldsymbol{\mu}$ are present, separate them from the other parameters, $\boldsymbol{\psi}$, and apply a uniform prior on the location parameters; then apply the 'general rule' to all remaining parameters. This leads to the following prescription:

$$p_{\mathrm{J}}\left(\boldsymbol{\mu}, \boldsymbol{\psi}\right) \propto \left(\det \boldsymbol{I}(\boldsymbol{\psi})\right)^{1/2}, \tag{3.81}$$

where $\boldsymbol{I}(\boldsymbol{\psi})$ is computed holding $\boldsymbol{\mu}$ fixed.

### *3.5.5 Conjugate Priors*

Conjugate priors are based on a principle of mathematical and analytical convenience: given a likelihood, one chooses a prior so that the posterior belongs to the same parametric family as the prior. A parametric family is a set of distribution with the same functional form, but differing in the value of the free parameters defining them. Under the assumption of a conjugate prior, the posterior distribution is readily obtained by simply updating the distributions' parameters with summary statistics from the data. For example, a binomial likelihood has as conjugate prior the Beta distribution. The prior is thus

$$\Pr(\theta) \sim \mathrm{Beta}(\alpha, \beta), \tag{3.82}$$

where $\alpha, \beta$ are called 'hyperparameters', as they parameterize the shape of the prior distribution. After $n$ i.i.d. samples with $r$ successes (associated with $\theta = 1$), the posterior is simply given by

$$\Pr(\theta \mid n, r) = \mathrm{Beta}(\alpha + r, \beta + n - r). \tag{3.83}$$

Another case of conjugate prior we already encountered is the Normal prior for a Normal likelihood, leading to a Normal posterior. A few example of conjugate priors are given in Table 3.1.

For conjugate priors to exponential families (such as the exponential distribution and the Normal), the posterior mean is a convex combination of the MLE and the prior mean, a so-called 'shrinkage' estimator, in that the posterior is 'shrunk' towards the prior. Why this can be beneficial will be explained when we encounter Bayesian hierarchical models.

### *3.5.6 Recommendations on Prior Choice*

To conclude this section on priors, here is a list of summary recommendations that arise from experience.

Table 3.1 *A few cases of conjugate priors. Here,* $\bar{x} = \sum_{i=1}^{n} x_i / n$ *is the sample mean* $r$ *is the number of successes, and* $(i = 1, \dots, n)$ *denotes the i.i.d. samples.*

| Likelihood | Conjugate prior | Posterior |
|---|---|---|
| $x_i \sim \text{Poisson}(\lambda)$ | $\lambda \sim \text{Gamma}(\alpha, \beta)$ | $\lambda \mid \boldsymbol{d} \sim \text{Gamma}(\alpha + n\bar{x}, \beta + n)$ |
| $x_i \sim \text{Bernoulli}(\theta)$ or $\text{Binomial}(n, \theta)$ | $\theta \sim \text{Beta}(\alpha, \beta)$ | $\theta \mid \boldsymbol{d} \sim \text{Beta}(\alpha + r, \beta + n - r)$ |
| $x_i \sim \text{Normal}(\mu, \sigma^2)$, $\sigma^2$ known | $\mu \sim \text{Normal}(\mu_0, \tau^2)$ | $\mu \mid \boldsymbol{d} \sim \text{Normal}\left(\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{\mu_0}{\tau^2} + \frac{n\bar{x}}{\sigma^2}\right), \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right)$ |
| $x_i \sim \text{Normal}(\mu, \sigma^2)$, $\mu$ known | $\sigma^2 \sim \text{InvGamma}(\alpha, \beta)$ | $\sigma^2 \mid \boldsymbol{d} \sim \text{InvGamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$ |
| $\boldsymbol{x}_i \sim \text{MV Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ known | $\boldsymbol{\mu} \sim \text{Multivariate Normal}(\boldsymbol{\mu}_0, \Lambda)$ | $\boldsymbol{\mu} \mid \boldsymbol{d} \sim \text{MV Normal}\left(\left(\Lambda^{-1} + n\Sigma^{-1}\right)^{-1}\left(\Lambda^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\bar{\boldsymbol{x}}\right), \left(\Lambda^{-1} + n\Sigma^{-1}\right)^{-1}\right)$ |
| $\boldsymbol{x}_i \sim \text{MV Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}$ known | $\Sigma \sim \text{InvWishart}(\boldsymbol{\Psi}, \nu)$ | $\Sigma \mid \boldsymbol{d} \sim \text{InvWishart}\left(\boldsymbol{\Psi} + \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top, \nu + n\right)$ |

1 Include all relevant information in the priors, and be explicit (for example, to your readers) about why and how you do it.
2 Be wary of so-called 'non-informative' (i.e. flat) priors, especially in high dimensions!
3 Many cases of interest have well-tested priors: search the literature and use previous result.
4 For observables that are non-linear in the parameters, the data might be Gaussian but the posterior generally won't. Take as an example

$$\Pr(x \mid \theta) \propto \exp\left(-\frac{1}{2}\frac{\left(x - f(\theta)\right)^2}{\sigma^2}\right) \tag{3.84}$$

with $f(\theta)$ giving the predicted value of an observable $x$, which is a non-linear function of the parameter of interest $\theta$: the data might tell you $x = \overline{x} \pm \sigma$ but if $f(\theta)$ is strongly non-linear and you don't realise it, you might end up with a prediction that is concentrated in the wrong region of the $x$-space. In these cases it might be useful to investigate the distribution of $f(\theta)$, i.e. of the predicted data, under the prior.
5 Differently from parameter inference (which has concerned us thus far), priors remain important even in the $n \to \infty$ limit for Bayesian model comparison; see chapter 5.
6 Priors are the 'price to pay' in Bayesian analysis; on the other hand, they force you to clarify—first of all, to yourself—and state your assumptions.
7 Priors are a *feature* of Bayesian analysis (in the sense of de Finetti).
8 There is plenty of wrong prior choices, but no single 'right' one: the prior depends on the situation, since it should express a subjective state of knowledge.
9 Remember the central *tenet* of inference:

THERE IS NO INFERENCE WITHOUT ASSUMPTIONS

Hence, if someone tells you they are using 'no prior', chances are they don't quite know what assumptions they are making.

**Exercises**

3.1 A violin soloist is nervously awaiting the start of her concert behind the curtain, wondering whether the 1000-seater auditorium is full. She cannot see the public, but she can hear them coughing intermittently. She counts 10 coughs during 2 minutes. Assuming that the average rate of coughing for each person is $\lambda_i = 0.05$ coughs/minutes, that everybody in the auditorium has the same average coughing rate, and that people cough independently from each other, estimate the number of people in the auditorium.

3.2 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.

1 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 $BB = 0$).
2 Compute the conditional probability that your friend is sick, knowing that she has tested positive, i.e., find $P(BB = 1 \mid +)$.
3 Imagine screening the general population for a very rare disease, 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., $P(+ \mid BB = 1)$) 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 $P(+ \mid BB = 0)$ (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?

4 Now we write the false positive rate as $P(+ \mid BB = 0) = 1 - P(- \mid BB = 0)$. It is reasonable to assume (although this is not true in general) that $P(- \mid BB = 0) = P(+ \mid BB = 1)$, 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., $P(+ \mid BB = 1)$) 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.

3.3 You are having lunch with a colleague in the cafeteria and notice that she is eating a salad. There are 5 different dishes on offer, and only one of them is vegetarian. Furthermore, the probability for a random person to be vegetarian is 3%. You may assume that a non-vegetarian would choose at random among the dishes on offer.

1 What is the probability that your colleague is vegetarian?
2 The day after you have lunch together again, and once again she chooses a salad for lunch. The same happens for $N$ consecutive days (including the first). Determine the smallest value of $N$ so that the probability of your colleague being vegetarian exceeds 95%.

3.4 At a party, you make a bet with one of the other 50 guests that she cannot toss a fair coin 10 times in a row and obtain 10 heads. She flips the fair coin you provided 10 times, and obtains 10 heads; you lose the bet. Later you learn that one of the guests was a famous magician, known for being able to manipulate coin flips at will.

What is the probability that the person you made the bet with was the magician?

3.5 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.

1 Given the above rules (and your full knowledge of them), should you stick with your choice or is it better to switch?
2 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.

3.6 We have $N$ i.i.d. distributed samples drawn from $\mathcal{N}(\mu, \sigma)$. Find the joint

posterior distribution $p(\mu, \sigma \mid \boldsymbol{d})$, with appropriate non-informative priors on the parameters, and derive the marginal distributions for $\mu$, $\sigma$.



 

# 4

# Bayesian Parameter Inference

This chapter deals with the practical application of Bayes Theorem, and introduces tools that enable the cosmologist to turn a neat conceptual idea into a powerful inference technique, able to deal with realistic models and complex data.

## 4.1 Bayesian Model Building

The cosmology and astrophysics communities have been embracing Bayesian methods 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. 4.1. In more recent years, machine learning has been playing an increasingly important role.

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

1 Choose a model containing a set of hypotheses in the form of a vector of parameters, $\boldsymbol{\theta}$ (e.g., the mass of an extra–solar planet or the abundance of dark matter in the Universe).
2 Specify the priors for the parameters. Priors should summarize your state of knowledge about the parameters before you consider the new data, including any relevant external source of information. State explicitly the assumptions being made about prior choices, particularly whenever those are attempting to encapsulate a state of 'ignorance' or indifference about a parameter.
3 Construct the likelihood function for the measurement, which, in the simplest of cases, directly reflects the random process by which the data are ob-

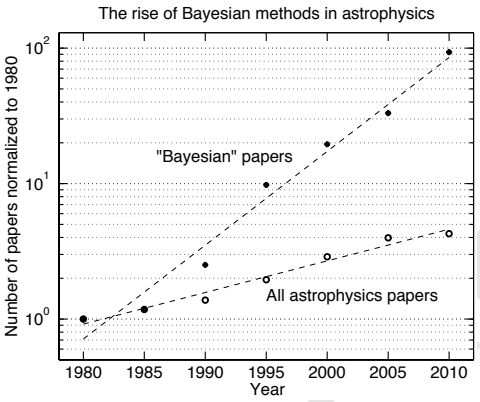


Figure 4.1 Number of articles in astronomy and cosmology with "Bayesian" in the title, as a function of publication year (upper data points) and total number of articles (lower data points) as a function of publication year. Numbers are normalized to 1980 levels for each data series (source: NASA/ADS).

tained. For example, a measurement with Gaussian noise will be represented by a Normal distribution, while $\gamma$–ray counts on a detector will have a Poisson distribution for a likelihood. Usually, the likelihood contains nuisance parameters describing all sorts of complications at play – from backgrounds to foregrounds, from non-random selection effects systematics. Such nuisance parameters need to be included in the likelihood (with appropriate priors), and we give more detailed examples of increasing complexity below: multilevel hierarchical models can be used when the objects being observed come from a population (see section 4.3), and to include important observational effects such as selection effects. If the likelihood is not known or intractable Simulation-Based Inference can still give a mean to map the posterior (see 4.6).

4 If external measurements are available for the nuisance parameters, they can be incorporated either as an informative prior on them, or as additional likelihood terms.

5 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, or SBI).

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 in-

 

formation 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[1].

## 4.2 The Gaussian Linear Model

A simple, yet widely applicable model is the Gaussian linear model, which is amenable to fully analytical solutions and applies, at least in an approximate manner, to many relevant cases of interest – for example, CMB analysis. Here we solve analytically the general problem in $d$ dimensional parameter space.

The set-up is one in which the dependent variable, $y$, is related to the independent variable $x$ via a relationship of the form:

$$f(x) = \sum_{j=1}^{d} \theta_j B_j(x)\,, \tag{4.1}$$

where the model's parameters are $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_d\} \in \mathbb{R}^d$ and the basis functions $B_j$ $(j = 1, \ldots, d)$ can be any function of $x$. Notice that the model is **linear** in the parameters $\theta_j$, not necessarily in $x$, i.e. $B_j$ can very well be a non-linear function of $x$. For example, a polynomial relationship between $y$ and $x$ could be expressed by choosing $B_j(x) = x^{j-1}$ $(j = 1, \ldots, d)$.

The dependent variable (also called 'response variable'), $y_i$, is measured with independent Gaussian noise, i.e., for each datum

$$y_i \mid \boldsymbol{\theta} \sim \mathcal{N}(f(x_i), \tau_i^2)\,.$$

where the independent variable $x_i$ is assumed known (we shall relax this in the errors-in-variables model below). We can write this model in more compact form by using matrix notation:

$$\mathbf{y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\epsilon}\,, \tag{4.2}$$

where we have collected the observed values in the data vector $\mathbf{y} \in \mathbb{R}^n$, and the 'design matrix' $\mathbf{F} \in \mathbb{R}^{n \times d}$ is a matrix of known constants which encapsulate the linear relation of Eq. (4.1), given by:

$$F_{ij} = B_j(x_i)\,,\ (i = 1, \ldots, n;\ j = 1, \ldots, d)\,.$$

Furthermore, $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a vector of RV with zero mean (the noise) which follows a multivariate Normal distribution with uncorrelated covariance matrix

[1] 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.

$C \equiv \text{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_n^2)$. The Gaussian likelihood can then be written as:

$$\Pr\left(\mathbf{y} \mid \boldsymbol{\theta}\right) = \frac{1}{(2\pi)^{n/2} \prod_j \tau_j} \exp\left[-\frac{1}{2}(b - A\theta)^\top (b - A\theta)\right], \tag{4.3}$$

where we have defined $A_{ij} = F_{ij}/\tau_i \in \mathbb{R}^{n\times d}$ and $b_i = y_i/\tau_i \in \mathbb{R}^n$. This can be recast with some simple algebra as

$$\Pr\left(\mathbf{y} \mid \boldsymbol{\theta}\right) = \mathscr{L}_0 \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{ML}})^\top \mathbf{L} (\boldsymbol{\theta} - \boldsymbol{\theta}_{\text{ML}})\right], \tag{4.4}$$

with the likelihood precision matrix (i.e., the inverse covariance matrix) $\mathbf{L} \in \mathbb{R}^{d\times d}$ given by

$$\mathbf{L} \equiv \mathbf{A}^\top \mathbf{A} \tag{4.5}$$

and a normalization constant

$$\mathscr{L}_0 \equiv \frac{1}{(2\pi)^{n/2} \prod_{i=1}^n \tau_i} \exp\left[-\frac{1}{2}(\mathbf{b} - \mathbf{A}\boldsymbol{\theta}_{\text{ML}})^\top (\mathbf{b} - \mathbf{A}\boldsymbol{\theta}_{\text{ML}})\right], \tag{4.6}$$

where $\boldsymbol{\theta}_{\text{ML}}$ is the MLE for $\boldsymbol{\theta}$, given by:

$$\boldsymbol{\theta}_{\text{ML}} = \mathbf{L}^{-1} \mathbf{A}^\top \mathbf{b}. \tag{4.7}$$

Our parameters of interest, $\boldsymbol{\theta}$, are location parameters in the Gaussian linear model, and hence we can use a conjugate prior for the mean of a multivariate Gaussian, which is given by a Normal distribution. We choose for prior a multivariate Normal with zero mean and precision matrix $\mathbf{P} \in \mathbb{R}^{d\times d}$, given by

$$\Pr\left(\boldsymbol{\theta}\right) = \frac{|\,\mathbf{P}\,|^{1/2}}{(2\pi)^{d/2}} \exp\left[-\frac{1}{2}\boldsymbol{\theta}^\top \mathbf{P} \boldsymbol{\theta}\right]. \tag{4.8}$$

Comparison with Table 3.1, with $\mathbf{P} = \Lambda^{-1}$ and $= \mathbf{L} = n\Sigma^{-1}$, gives that the posterior distribution is a Gaussian with precision matrix $\mathscr{F}$

$$\mathscr{F} = \mathbf{L} + \mathbf{P} \tag{4.9}$$

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

$$\bar{\boldsymbol{\theta}} = \mathscr{F}^{-1} \mathbf{L} \boldsymbol{\theta}_{\text{ML}}. \tag{4.10}$$

Finally, Bayesian evidence (whose properties we will analyze in detail in the next chapter), i.e., the normalizing constant in Bayes theorem, is given by

$$\begin{aligned} \Pr\left(y\right) &= \mathscr{L}_0 \frac{|\,\mathscr{F}\,|^{-1/2}}{|\,\mathbf{P}\,|^{-1/2}} \exp\left[-\frac{1}{2}\boldsymbol{\theta}_{\text{ML}}^\top (\mathbf{L} - \mathbf{L}\mathscr{F}^{-1}\mathbf{L})\boldsymbol{\theta}_{\text{ML}}\right] \\ &= \mathscr{L}_0 \frac{|\,\mathscr{F}\,|^{-1/2}}{|\,\mathbf{P}\,|^{-1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_{\text{ML}}^\top \mathbf{L}\boldsymbol{\theta}_{\text{ML}} - \bar{\boldsymbol{\theta}}^\top \mathscr{F} \bar{\boldsymbol{\theta}})\right]. \end{aligned} \tag{4.11}$$

It is worth noting that in this case, too, asymptotically for $n \to \infty$, we have that $\mathscr{F} \to \mathbf{L}$ and $\bar{\boldsymbol{\theta}} \to \boldsymbol{\theta}_{\mathrm{ML}}$, i.e., the posterior converges to the maximum likelihood result.

From the joint Normal posterior for all $d$ parameters, $\boldsymbol{\theta}$, we are often interested in deriving marginal 1– or 2–dimensional posteriors for a subset of $\boldsymbol{\theta}$. Splitting the parameter vector as $\boldsymbol{\theta} = \{\boldsymbol{\theta}_I, \boldsymbol{\theta}_U\}$, where $\boldsymbol{\theta}_I \in \mathbb{R}^k$ is the subset of parameters of interest (with $k = 1, 2$, usually) and $\boldsymbol{\theta}_U \in \mathbb{R}^{d-k}$ is the group of 'uninteresting' parameters we want to marginalize over, the marginal posterior for the parameters of interest is:

$$\Pr\left(\boldsymbol{\theta}_I \mid \mathbf{y}\right) = \int \Pr\left(\boldsymbol{\theta}_I, \boldsymbol{\theta}_U \mid \mathbf{y}\right) \mathrm{d}\boldsymbol{\theta}_U \,, \tag{4.12}$$

which is still a multivariate Normal, with the same mean as the joint posterior but with the last $d-k$ entries removed, and the same covariance matrix as the joint posterior, only with the last $d-k$ rows and columns deleted:

$$V_{ij} = \left[\mathscr{F}^{-1}\right]_{ij} \quad 0 \leq i, j \leq k \,. \tag{4.13}$$

This result can be obtained by performing explicitly the integration in Eq. (4.12) or more elegantly by using the properties of the characteristic function (Kendall and Stuart, 1963, Chap.4, Vol.1). If we write the posterior precision matrix as a combination of sub-matrices, split according to the parameters of interest as:

$$\mathscr{F} = \begin{pmatrix} \mathbf{K} & \mathbf{G} \\ \mathbf{G}^\top & \mathbf{B} \end{pmatrix} \tag{4.14}$$

where $\mathbf{K} \in \mathbb{R}^{k \times k}$, $\mathbf{B} \in \mathbb{R}^{(d-k)\times(d-k)}$ and $\mathbf{G} \in \mathbb{R}^{(d-k)\times k}$, then the covariance matrix for the marginalized posterior is:

$$\mathbf{V} = \left[\mathbf{K} - \mathbf{G}\mathbf{B}^{-1}\mathbf{G}^\top\right]^{-1} . \tag{4.15}$$

For a single parameter of interest, $k = 1$, and the posterior standard deviation is given by $\sqrt{\left(\mathscr{F}^{-1}\right)_{11}}$.

It is interesting to compare the marginal result with the profile likelihood. We write the joint likelihood precision matrix as

$$\mathbf{L} = \begin{pmatrix} \mathbf{K}_L & \mathbf{G}_L \\ \mathbf{G}_L^\top & \mathbf{B}_L \end{pmatrix} \tag{4.16}$$

where $\mathbf{K}_L \in \mathbb{R}^{k \times k}$, $\mathbf{B}_L \in \mathbb{R}^{(d-k)\times(d-k)}$ and $\mathbf{G}_L \in \mathbb{R}^{(d-k)\times k}$. Maximising over the uninteresting parameters $\boldsymbol{\theta}_U$ as in Eq. (2.69), means minimizing the quadratic form in the exponent of Eq. (4.4), $(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{ML}})^\top \mathbf{L} (\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathrm{ML}})$, w.r.t. $\boldsymbol{\theta}_U$. Differentiating with respect to $\boldsymbol{\theta}_U$, the minimum of the quadratic form lies at

$$\boldsymbol{\theta}_U = -\mathbf{B}_L^{-1}\mathbf{G}_L^\top (\boldsymbol{\theta}_I - \boldsymbol{\theta}_{\mathrm{ML},I}) \,, \tag{4.17}$$

where $\boldsymbol{\theta}_{\mathrm{ML},I}$ is the MLE with the last $k-d$ component removed. Therefore the profile likelihood for the parameters of interest is:

$$\mathscr{L}_p(\boldsymbol{\theta}_I) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}_I - \boldsymbol{\theta}_{\mathrm{ML},I})^\top \left[\mathbf{K}_L - \mathbf{G}_L \mathbf{B}_L^{-1} \mathbf{G}_L^\top\right](\boldsymbol{\theta}_I - \boldsymbol{\theta}_{\mathrm{ML},I})\right\}, \tag{4.18}$$

which is the same result as in the marginal posterior, Eq. (4.15), except that the posterior precision matrix is replaced by the likelihood precision matrix, and the posterior mean by the MLE. If instead of a Gaussian conjugate prior we choose a uniform prior on $\boldsymbol{\theta}$, then the two results would be identical.

## 4.3 Bayesian Hierarchical Models

Bayesian hierarchical models (BHMs) offer a robust statistical framework for modeling data with complex, nested structures. By layering probability distributions over parameters and incorporating prior knowledge at different levels, BHMs enable precise quantification of uncertainty, inclusion of different types of stochastic and deterministic dependencies and the sharing of information across hierarchical levels. BHMs are particularly useful when dealing with data that arise from multiple sources or groups, each with distinct but related characteristics, described by the top-level population distributions. The explicit incorporation of hierarchical structures provides a way to account for variation within groups while simultaneously estimating parameters that describe population-level effects, creating a statistical linking among them that goes under the name of 'borrowing of strength' (this will be quantified more precisely below).

At the foundation of BHMs is the concept of **partial pooling**, which represents a balance between modeling data at the population level, without distinguishing between groups (complete pooling) and at the individual level (no pooling). Through a hierarchical structure, group-level parameters are assumed to be drawn from a common population distribution, allowing the model to 'borrow strength' across groups. This hierarchical organization often takes the form of multi-level prior distributions, where lower-level parameters depend on higher-level parameters, thus introducing dependencies across levels.

### *4.3.1 Representation via direct acyclic graphs*

A BHM is usually represented via a **direct acyclic graph (DAG)**, an example of which is shown in Fig. 4.2: this is a visual, compact representation of the relationships between the variables in a BHM. In a DAG, each variable in the model is represented by a node (or vertex), and each conditional dependency is represented by a directed edge (or arrow) connecting two nodes. Each edge has a direction, indicated by an arrow pointing from one node to another, representing

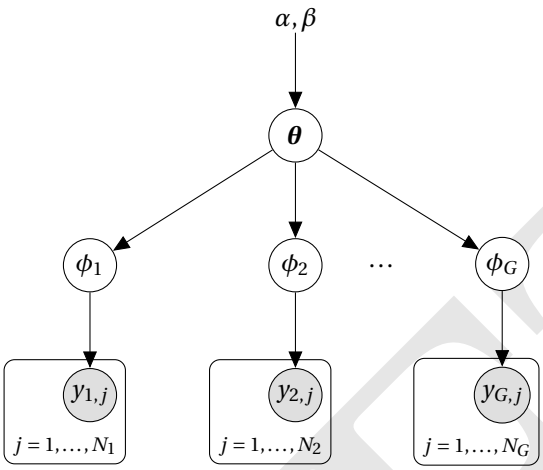


Figure 4.2 DAG representation of a Bayesian hierarchical model with observed data $y_{i,j}$ for groups 1, 2, and $G$, group-level parameters $\phi_i$, and population-level parameters $\boldsymbol{\theta}$. The hyperparameters $\alpha, \beta$ control the prior on $\boldsymbol{\theta}$.

a conditional relationship. For instance, an arrow from node $A$ to node $B$ means that $A$ is a parent of $B$, i.e., $B$ depends on $A$ in the model. A DAG has no cycles (i.e., no loops), meaning that following the direction of the edges, you cannot start at any node and return to it by following a series of edges. This ensures that there are no circular dependencies among variables.

Nodes in a DAG are styled differently depending on what kind of variable they represent: observed data nodes are shown as shaded nodes, and are variables for which we have observed data; parameter nodes are shown as unshaded nodes, denoting unknown quantities that we aim to estimate; hyperparameter nodes are unshaded and/or with additional styling, representing higher-level parameters that usually define prior distributions over the lower-level parameters. Finally, rectangular-shaped plates surround nodes that are duplicated a certain number of times, indicated by the indexing inside each plate.

The DAG of a typical BHM flows from top to bottom, with the highest-level, population parameters at the top, and data at the bottom. Following the example of Fig. 4.2, the conditional structure of the BHM shown there is, from bottom up:

$$\begin{aligned} y_{i,j} &\sim \Pr\left(y \mid \phi_i\right) \\ \phi_i &\sim \Pr\left(\phi \mid \boldsymbol{\theta}\right) \\ \boldsymbol{\theta} &\sim \Pr\left(\boldsymbol{\theta} \mid \alpha, \beta\right). \end{aligned} \tag{4.19}$$

where $y_{i,j}$ represents the observed data for the $j$-th observation ($j = 1, \ldots, N_i$) within the $i$-th group, $\phi_i$ are the group-level parameters for group $i$ ($i = 1, \ldots, G$),

 

$\boldsymbol{\theta}$ are the parameters governing the population-level distribution and $\alpha, \beta$ are hyperparameters controlling the top-level prior[2]. This hierarchical specification enables the model to capture both group-specific effects and population-wide regularities.

Once the conditional structure has been established, it is simple to obtain the posterior for the parameters of interest, usually the population-level variables: one writes the joint posterior over both population-level variables and latent ones, and marginalizes out the latent (unobserved) parameters. The joint is then re-written in terms of known conditionals:

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{d}) = \int \Pr\left(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{d}\right) \mathrm{d}\phi = \int \Pr\left(\boldsymbol{d} \mid \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\phi} \mid \boldsymbol{\theta}\right) \Pr\left(\boldsymbol{\theta} \mid \alpha, \beta\right) \mathrm{d}\boldsymbol{\phi}\,, \tag{4.20}$$

where $\boldsymbol{d} = \{y_{i,j} \mid i = 1, \ldots, G, j = 1, \ldots N_i\}$. Since the latent variables are often high-dimensional (i.e., $G \gg 1$), it is computationally advantageous to carry out the marginalization integral above analytically whenever possible (we shall encounter examples below). To this end, often conditional conjugate priors are adopted, so that all of the conditionals are analytically known and can be sampled from exactly. This greatly reduces the dimensionality of the parameter space that needs to be sampled numerically and thus increases computational efficiency (see e.g. Norton et al. (2016) for some examples). Another helpful method is Gibbs sampling (see section 4.4.3), which is particularly well suited to the conditional structure of BHM.

If the latent, group-level parameters $\boldsymbol{\phi}$ have been marginalized out analytically, it is often desirable to recover their posterior distribution for inference on the latent structure. This can be done via **conditional sampling**, which draws samples for the latent variables conditioned on both observed data and sampled values of the population-level parameters from the marginal posterior. One thus samples from the posterior distribution of the population-level distribution, obtaining posterior samples $\boldsymbol{\theta}_s$ ($s = 1, \ldots, T$). Then, for each sampled $\boldsymbol{\theta}_s$, one draws samples for the latent parameters $\boldsymbol{\phi}$ from the conditional distribution

$$\Pr\left(\boldsymbol{\phi} \mid \boldsymbol{d}, \boldsymbol{\theta}_s\right) \propto \Pr\left(\boldsymbol{d} \mid \boldsymbol{\phi}, \boldsymbol{\theta}_s\right) \Pr\left(\boldsymbol{\phi} \mid \boldsymbol{\theta}_s\right),$$

where, for the hierarchical structure of Eq. (4.19), we can drop the conditioning on $\boldsymbol{\theta}_s$ in $\Pr\left(\boldsymbol{d} \mid \boldsymbol{\phi}, \boldsymbol{\theta}_s\right)$ since the data only depend on the latent variables, $\boldsymbol{\phi}$.

BHMs are well-suited to scenarios in which data naturally fall into groups or nested structures, such as individuals within schools, measurements within regions, or time points within patients. For instance, in epidemiology, modeling disease incidence across regions can benefit from BHMs, where each region

[2] In some of the literature, the term 'hyperparameters' refers to what we have called here 'population-level parameters'. We reserve the designation of 'hyperparameters' to parameters describing the top-level prior distribution for clairiy.

might have its own incidence rate but also share certain characteristics across regions. The group-level parameters can be understood as the 'true' value of the observed data, in the absence of noise – and for this reason are called **latent variables** in the model. When possible (e.g., in Gaussian linear model), they are often marginalized over analytically, as this reduces the dimensionality of the parameter space that needs to be sampled in cases where the scientific focus is on the population-level parameters. However, inspecting their marginal posterior distribution can be instructive in many cases – this can be accessed a posteriori through the conditional structure of the model. We shall clarify these concept further with see some concrete examples below.

The advantages of BHMs are that they provide a structured approach to manage variability within groups and across groups. The partial pooling of information in BHMs balances capturing individual group characteristics with maintaining an overall model consistency. Bayesian hierarchical models also facilitate inclusion of prior information at multiple levels, enabling more informed and structured inferences; they are also capable of capturing dependence structures that are challenging to model directly, for example through higher-level parameters that implicitly induce correlation among lower-level units.

On the other hand, BHMs require attention to computational feasibility, as these models can become computationally intensive with increased model complexity and the need for sampling over multiple layers, each of which may have a large number of variables. Choosing appropriate priors and assessing model convergence is crucial. Given their hierarchical nature, the sensitivity of the posterior to hyperparameters and model assumptions should be carefully evaluated, often through sensitivity analysis or cross-validation.

### *4.3.2 Errors-in-variables Normal model*

As a concrete illustration of Bayesian hierarchical modeling, we consider here a linear errors-in-variable model: this designates the case, common in astronomy, where both the dependent and independent variable are measured with uncertainty. Examples in astronomy and cosmology include the scaling relation between richness and mass of galaxy clusters (Andreon and Hurn, 2010), empirical standardization of supernovae type Ia's magnitude (March et al., 2011) and others... . Here, we follow the treatment in Gull (1989).

The model is shown by the DAG of Fig. 4.3: a linear relationship is assumed between the latent variables, $x_i$ and $y_i$ ($i = 1, \ldots, N$), described by a slope $a$ and intercept $b$ (which are collectively denoted as a parameter vector $\boldsymbol{\theta}$):

$$y_i = a x_i + b\,. \tag{4.21}$$

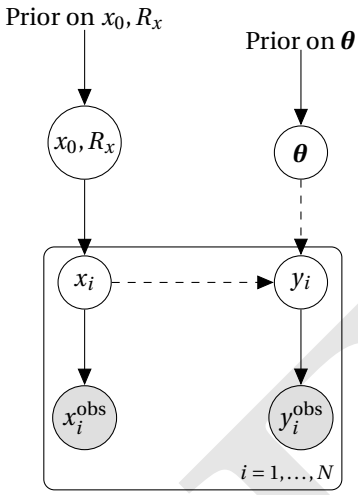


Figure 4.3 DAG for Bayesian error-in-variables model with $y = ax + b$, with $\boldsymbol{\theta} = \{a, b\}$. Solid lines indicate probabilistic connections, dashed lines represent deterministic connections.

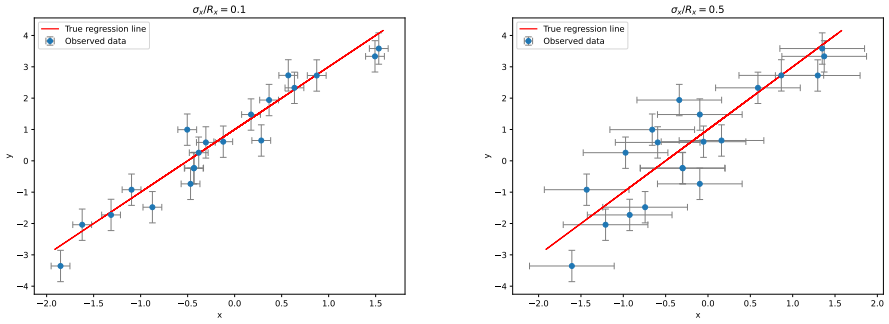


Figure 4.4 Example data sets for the linear regression with errors-in-variables model.

This being a deterministic relation, the corresponding arrow in the DAG is dashed (as opposed to solid). The observed values for the dependent and independent variables are denoted by $\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{y}_i^{\text{obs}}$, $i = 1, \ldots, N$), and are i.i.d. from the Gaussian distribution (with known variances $\sigma_x^2, \sigma_y^2$) conditional on the latent values (assuming uncorrelated noise between $x$ and $y$ for simplicity, though this would be simple to generalize with a covariance matrix instead):

$$x_i^{\text{obs}} \mid x_i \sim \mathcal{N}(x_i, \sigma_x^2) \quad \text{and} \quad y_i^{\text{obs}} \mid y_i \sim \mathcal{N}(y_i, \sigma_y^2)\,. \tag{4.22}$$

This probabilistic relationship is depicted in Fig. 4.3 by the solid arrows connect-

ing the latent variables to the observed quantities. The independent variable, $x$, is drawn from a Gaussian distribution with mean $x_0$ and variance $R_x^2$:

$$x_i \sim \mathcal{N}(x_0, R_x^2)\,, \tag{4.23}$$

which are unknown, population-level parameters that need to be estimated from the data. Their priors are discussed below, and two example data sets are shown in Fig. 4.4. The joint likelihood is given by

$$\Pr\left(\boldsymbol{x}^{\rm obs}, \boldsymbol{y}^{\rm obs} \mid \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}\right) \propto \exp\left(-\frac{1}{2}\left[\frac{\sum_i (\boldsymbol{x}_i^{\rm obs} - x_i)^2}{\sigma_x^2} + \frac{\sum_i (\boldsymbol{y}_i^{\rm obs} - y_i)^2}{\sigma_y^2}\right]\right). \tag{4.24}$$

The problem can be made more symmetric by defining rescaled versions of the data:

$$\hat{X} = \frac{\boldsymbol{x}^{\rm obs} - x_0}{R_x} \text{ and } \hat{Y} = \frac{\boldsymbol{y}^{\rm obs} - y_0}{R_y}, \tag{4.25}$$

where the variables $x_0, y_0$ are the mean latent value of $x$ and $y$, respectively, and $R_x, R_y$ the latent variances. While $x_0, R_x$ are genuine population variables, $y_0, R_y$ are related to the slope and intercept by:

$$b = y_0 - a x_0 \text{ and } a = R_y/R_x \quad \rightarrow \quad b = y_0 - \frac{R_y}{R_x} x_0\,. \tag{4.26}$$

The joint posterior for $\mathbf{x}, \mathbf{y}$ and $\boldsymbol{\phi} = \{x_0, y_0, R_x, R_y\}$ can be written as:

$$\Pr\left(\mathbf{x}, \mathbf{y}, \boldsymbol{\phi} \mid \boldsymbol{x}^{\rm obs}, \boldsymbol{y}^{\rm obs}\right) \propto \Pr\left(\boldsymbol{x}^{\rm obs}, \boldsymbol{y}^{\rm obs} \mid \mathbf{x}, \mathbf{y}, \boldsymbol{\phi}\right) \Pr\left(\mathbf{x} \mid \boldsymbol{\phi}\right) \Pr\left(\boldsymbol{\phi}\right), \tag{4.27}$$

where the first term on the r.h.s. is the likelihood of Eq. (4.24) and the second is given by Eq. (4.23), multiplied for $N$ i.i.d. variables. Crucially, both $x_0$ and $R_x$ are unknown, and are explicitly determined from the data in the joint posterior, before being marginalized out at the end. Finally, for the prior $\Pr\left(\boldsymbol{\phi}\right)$ we adopt a uniform prior on $x_0, y_0$ (as those are location variables) and a prior uniform in $\log R_x, \log R_y$ (those being scale variables). Notice that this prior transforms in a non-trivial way to a prior on $b$ via the transformation defined by Eq. (4.26).

From Eq. (4.27), $y_i$ can be eliminated by using Eq. (4.21), and $b$ replaced by $R_y$ via Eq. (4.26), leading to:

$$\begin{aligned}\Pr\left(\mathbf{x}, \boldsymbol{\phi}, a \mid \boldsymbol{x}^{\rm obs}, \boldsymbol{y}^{\rm obs}\right) \propto\ & (R_x^2)^{-N/2} \exp\left(-\frac{1}{2}\left[\frac{\sum_i (\boldsymbol{x}_i^{\rm obs} - x_i)^2}{\sigma_x^2} + \right.\right. \\ & \left.\left. + \frac{\sum_i (\boldsymbol{y}_i^{\rm obs} - a x_i - y_0 + a x_0)^2}{\sigma_y^2} + \frac{\sum_i (x_i - x_0)^2}{R_x^2}\right]\right).\end{aligned} \tag{4.28}$$

where the pre-factor $(R_x^2)^{-N/2}$ comes from the normalization of the distribution for the latent independent variables, Eq. (4.23) (and we have dropped constant

terms that do not depend on parameters). From this expression, the latent variables $\mathbf{x}$ can be marginalized out analytically, as it enters linearly in the Gaussian expression, and so can the nuisance parameters $x_0, y_0$, by using appropriate completions of the square in the Gaussian. After some algebra, we obtain:

$$\Pr\left(\log a, \log R \mid \boldsymbol{x}^{\rm obs}, \boldsymbol{y}^{\rm obs}\right) \propto (a^2\sigma_x^2 R_x^2 + \sigma_x^2\sigma_y^2 + \sigma_y^2 R_x^2)^{-\frac{N-1}{2}} \\ \exp\left(-\frac{1}{2}\frac{V_{xx}(a^2R_x^2+\sigma_y^2) - 2V_{xy}aR_x^2 + V_{yy}(R_x^2+\sigma_x^2)}{a^2\sigma_x^2R_x^2 + \sigma_x^2\sigma_y^2 + \sigma_y^2R_x^2}\right) \tag{4.29}$$

where:

$$V_{xx}^2 = \sum_i (\boldsymbol{x}_i^{\rm obs} - \bar{x})^2, \quad V_{xy}^2 = \sum_i (\boldsymbol{x}_i^{\rm obs} - \bar{x})(\boldsymbol{y}_i^{\rm obs} - \bar{y}), \quad V_{yy}^2 = \sum_i (\boldsymbol{y}_i^{\rm obs} - \bar{y})^2, \quad R = (R_x R_y)^{1/2}. \tag{4.30}$$

In the above, $\bar{x} = \sum_i \boldsymbol{x}_i^{\rm obs}/N$, and similarly for $\bar{y}$, and notice that a log-uniform prior on $R_x, R_y$ translates into a log-uniform prior for both $R$ and $a$. In Eq. (4.29), we take $\log R$ and $\log a$ as our variables, and therefore can drop the uniform prior. The marginal posterior for the logarithm of the logarithm of the slope, $\log a$, is obtained by numerical marginalization of $\log R$ from the above expression.

If instead of the Bayesian solution found above, one writes down a simple Gaussian expression for the likelihood, including error propagation from the linear relationship (4.21) for the error term, one would obtain (D'Agostini, 2003):

$$\mathscr{L}(a,b) \propto \exp\left(-\frac{1}{2}\sum_i \frac{(\boldsymbol{y}_i^{\rm obs} - a\boldsymbol{x}_i^{\rm obs} - b)^2}{\sigma_y^2 + a^2\sigma_x^2}\right), \tag{4.31}$$

from where the intercept $b$ is eliminated by profiling. It is clear that this likelihood suffers from the problem that, while Gaussian in the data, it is no longer Gaussian in the parameters. In particular, the slope $a$ appears in the denominator as a contributor to the overall variance, and in absence of a normalization term in front, a larger value of $a$ leads to larger variance (as well as to shifting the mean), and hence to a better fit. Maximising the likelihood therefore leads to overestimating the value of the slope.

We compare the resulting inference on the slope parameter $a$ from Eq. (4.31) (with $b$ eliminated via profiling) with the result obtained using the Bayesian expression, (4.29), marginalized numerically over $\log R$ (with a uniform prior on $\log R$) in Figs. 4.5 and 4.6. In each figure, the upper left panel shows simulated data ($N = 300$), with the error bars giving the size of the standard deviation in the $x$ and $y$ directions for each datum (i.e., the value of $\sigma_x, \sigma_y$). The contour plots depict joint posterior regions for $(\log a, \log R)$ from the Bayesian expression of Eq. (4.29), and confidence regions from the likelihood (4.31) in the $a, b$

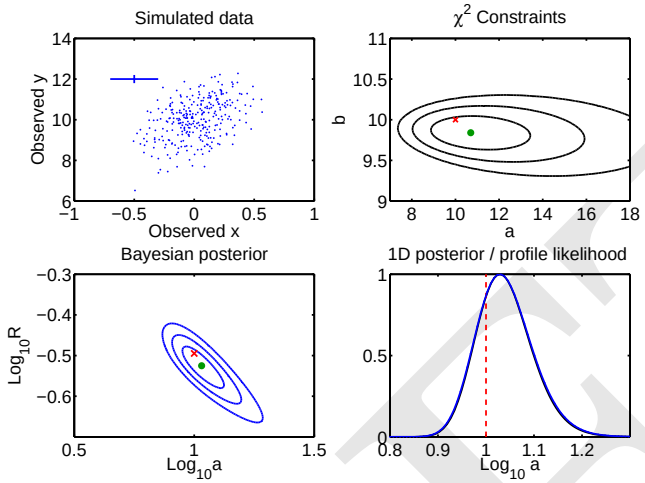


Figure 4.5 Comparison between profile likelihood and Bayesian linear fitting for the errors-in-variables model. Upper left panel: data set of $N = 300$ observations with errors both in the $x$ and $y$ directions, given by the error bar. Upper right panel: reconstruction of the slope $a$ and intercept $b$ using a linearized likelihood (red cross is the true value, green circle the maximum likelihood value). Lower left panel: Bayesian posterior in the $\log a$, $\log R$ plane, with green circle showing posterior mean. In both panels, contours enclose $1\sigma$, $2\sigma$ and $3\sigma$ regions. Lower right panel: marginalized Bayesian posterior (blue) and profiled likelihood (black, lying exactly on top of the blue curve), with the dashed line showing the true value. The two methods give essentially identical results in this case.

plane. The lower right panel shows the 1d marginalized posterior distribution for $\log a$ (blue) and the profile likelihood (black).

The two results are essentially identical when variance due to noise is much smaller than the variance of the latent values for the independent variable, i.e., when $\sigma_x/R_x \ll 1$, see Fig. 4.6. However, when the noise in the $x$ direction is of the same order or larger than the latent variables spread, as in Fig. 4.6, the profile likelihood gives a biased result for the slope parameter $a$, while the Bayesian posterior is closer to the true value, although it (correctly) shows a larger uncertainty. This illustration from a single data realization is borne out by comparing the long-term performance (i.e. coverage) of the two methods over repeated draws of the data.

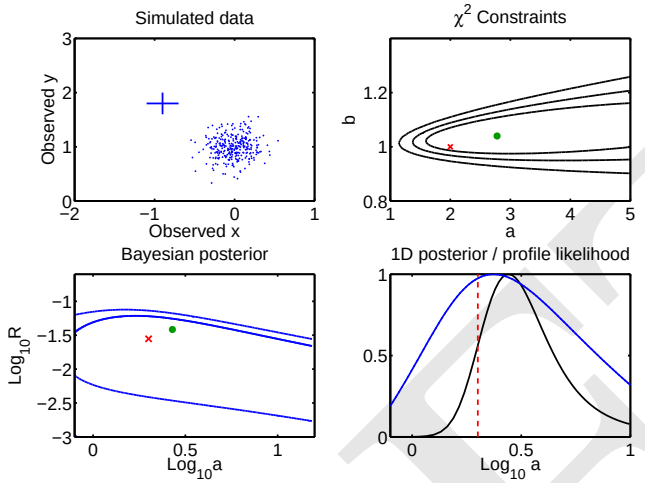


Figure 4.6 As in Fig. 4.6, but with a larger statistical uncertainty in the data compared to their latent variance. The Bayesian marginal posterior for the slope (blue, lower right panel) is closer to the true value than the profile likelihood expression (black). In the lower left panel, the $3\sigma$ contour from the Bayesian method lies outside the range of the plot.

### *4.3.3 Including intrinsic variability*

In an astrophysical context, the relationship between dependent and independent variable is never exact; rather, the independent variable exhibits **intrinsic variability** due to e.g. effects not captured by the model, variety of properties of the sources, contamination, and so on. An important example of such intrinsic scatter, as it is also called, is supernova type Ia. Scaling relations between cluster properties also exhibit such behaviour.

Incorporating such an effect in the Bayesian model is straightforward: it is sufficient to modify the hierarchical model of Eqs. (4.21) to add a stochastic dependence of the response variable $y$ on $x$, incorporating intrinsic variability w.r.t. the dependent variable, $y$:

$$y_i \mid x_i, \boldsymbol{\theta}, \sigma_{\mathrm{int}} \sim \mathcal{N}(ax_i + b, \sigma^2_{\mathrm{int}}) \tag{4.32}$$

$$x_i^{\mathrm{obs}} \mid x_i \sim \mathcal{N}(x_i, \sigma_x^2) \quad \text{and} \quad y_i^{\mathrm{obs}} \mid y_i \sim \mathcal{N}(y_i, \sigma_y^2)\,. \tag{4.33}$$

The intrinsic variability is parameterized via a Gaussian with variance $\sigma^2_{\mathrm{int}}$, itself a free parameter in the model. The corresponding DAG is depicted in Fig. 4.7.

As usual in a Bayesian setting, we obtain the desired distribution by 'expanding the discourse' to the joint distribution over unobserved (latent) variables and marginalizing those out. In this case, the **observed data likelihood** for each ob-

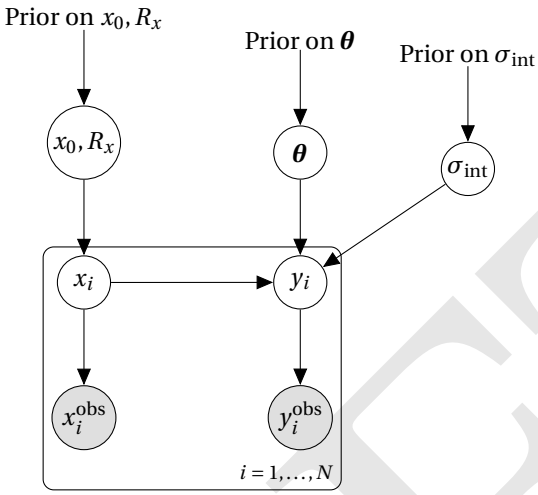


Figure 4.7 DAG for Bayesian error-in-variables model with $y = ax + b$, with intrinsic variability $\sigma_{\rm int}$. Notice that the dashed arrows into the dependent variable $y$ have now been replaced by solid lines (denoting a stochastic dependence).

servation $\Pr\left(x_i^{\rm obs}, y_i^{\rm obs} \mid \boldsymbol{\Theta}\right)$ is obtained by marginalizing over the so-called **complete likelihood** that includes the latent variables, $x_i, y_i$ $(i = 1, \ldots, n)$:

$$\Pr\left(x_i^{\rm obs}, y_i^{\rm obs} \mid \boldsymbol{\Theta}\right) = \int\int \Pr\left(x_i^{\rm obs}, y_i^{\rm obs}, x_i, y_i \mid \boldsymbol{\Theta}\right) \mathrm{d}x_i \mathrm{d}y_i \tag{4.34}$$

$$= \int\int \Pr\left(x_i^{\rm obs}, y_i^{\rm obs} \mid x_i, y_i\right) \Pr\left(y_i \mid x_i, \boldsymbol{\theta}, \sigma_{\rm int}\right) \Pr\left(x_i \mid x_0, R_x\right) \mathrm{d}x_i \mathrm{d}y_i\,, \tag{4.35}$$

where we have defined $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \sigma_{\rm int}, x_0, R_x\}$ and the first term on the RHS of the above equation is given by the measurement noise in Eq. (4.33), the second by the intrinsic variability, Eq. (4.32), and the last is the population-level distribution given by Eq. (4.23). Since observations are independent, the joint likelihood is the product over the $i$ observations, and carrying out the Gaussian integrals over the $2n$ latent variables $x_i, y_i$ leads to (Kelly, 2007):

$$\Pr\left(\boldsymbol{x}^{\rm obs}, \boldsymbol{y}^{\rm obs} \mid \boldsymbol{\Theta}\right) = \prod_{i=1}^{n} \frac{1}{2\pi \mid \boldsymbol{V} \mid^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{z}_i^{\rm obs} - \boldsymbol{\mu})^{\top} \boldsymbol{V}^{-1} (\boldsymbol{z}_i^{\rm obs} - \boldsymbol{\mu})\right), \tag{4.36}$$

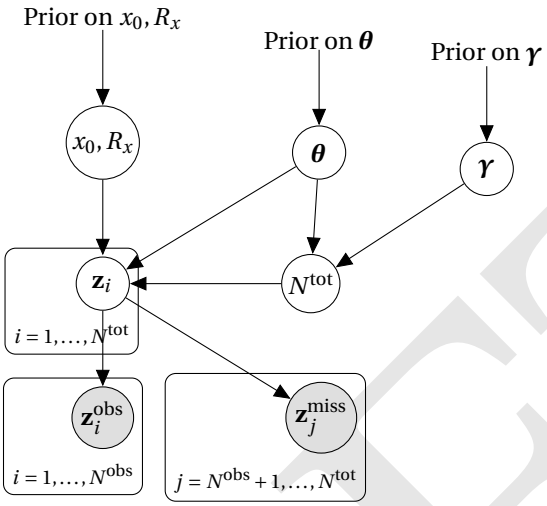


Figure 4.8 DAG for Bayesian error-in-variables model with selection effects. We denote by $\boldsymbol{z}_i = \{\mathbf{x}_i, \boldsymbol{y}_i\}$ the collection of latent variables, whose cardinality is given by $N^{\rm tot}$, itself an unknown parameter in the model that may depend both on $\boldsymbol{\theta}$ and its own parameters, $\boldsymbol{\gamma}$.

where we have defined:

$$\boldsymbol{z}_i^{\rm obs} = \begin{pmatrix} y_i^{\rm obs} \\ x_i^{\rm obs} \end{pmatrix}, \tag{4.37}$$

$$\boldsymbol{\mu}_i = \begin{pmatrix} a + bx_0 \\ x_0 \end{pmatrix}, \tag{4.38}$$

$$\boldsymbol{V} = \begin{pmatrix} b^2 R_x^2 + \sigma_{\rm int}^2 + \sigma_y^2 & bR_x^2 \\ bR_x^2 & R_x^2 + \sigma_x^2 \end{pmatrix}. \tag{4.39}$$

Armed with this likelihood, the marginal posterior for $\boldsymbol{\Theta}$ can now be computed as in Eq. (4.20), and obtained with little effort with numerical sampling methods of the kind discussed in section 4.4.

### *4.3.4 Selection Effects and Missing Data*

**Add here discussion of Eddington vs Malmquist bias and plot showing selection function of y**

We start by considering a complete population before any selection, consisting of an unknown number $N^{\rm tot}$ of objects, of which a known $N^{\rm obs}$ have been measured to yield the dataset $\boldsymbol{d}^{\rm obs} = \{d_1, \ldots, N^{\rm obs}\}$. Each of the $N^{\rm tot}$ objects is associated with a random variable $I_i$, indicating whether it was observed ($= I_i^{\rm obs}$) or missed ($= I_j^{\rm miss}$). In this context, "missed" refers to objects that could have

been detected but were not, due to specific realizations of their latent features and observational noise. The "selection procedure" encompasses all steps that determine which objects appear in the final analysis. Missed objects are labeled with indices $j \in \{N^{\text{obs}}+1,\dots,N^{\text{tot}}\}$, and we use $I_i^{\text{obs}}$ to signify "object $i$ was observed" ($I_i = I_i^{\text{obs}}$), and similarly $I_j^{\text{miss}}$ for missed or undetected objects. We denote by $\boldsymbol{I} = \{I^{\text{obs}}_{i=1,\dots,N^{\text{obs}}}, I^{\text{miss}}_{j=N^{\text{obs}}+1,\dots,N^{\text{tot}}}\}$ the complete indicator variable. Below, we follow the treatment in Karchev and Trotta (2024).

With a hierarchical model and parameters collectively denoted by $\boldsymbol{\Xi}$, and assuming independence between objects, the **complete data likelihood** consists of the probability of selecting objects $i \in \{1,\dots,N^{\text{obs}}\}$ and the probability density of their observed data, multiplied by the probability of missing objects $j \in \{N^{\text{obs}}+1,\dots,N^{\text{tot}}\}$, and the combinatorial factor $\binom{N^{\text{tot}}}{N^{\text{obs}}}$, accounting for the permutation invariance of labels, giving:

$$\Pr(\boldsymbol{d},\mathbf{I}\mid\boldsymbol{\Xi}) = \binom{N^{\text{tot}}}{N^{\text{obs}}}\prod_{i=1}^{N^{\text{obs}}}\Pr\left(I_i^{\text{obs}}\mid \boldsymbol{z}_i^{\text{obs}},\boldsymbol{\Xi}\right)\Pr\left(\boldsymbol{z}_i^{\text{obs}}\mid\boldsymbol{\Xi}\right)\prod_{j=1}^{N^{\text{tot}}-N^{\text{obs}}}\Pr\left(I_j^{\text{miss}}\mid\boldsymbol{\Xi}\right).$$

If we multiply and divide by $\prod_{i=1}^{N^{\text{obs}}} p(I_i^{\text{obs}}\mid\boldsymbol{\Xi})$, and replacing

$$\frac{\Pr\left(I_i^{\text{obs}}\mid\boldsymbol{z}_i^{\text{obs}},\boldsymbol{\Xi}\right)\Pr\left(\boldsymbol{z}_i^{\text{obs}}\mid\boldsymbol{\Xi}\right)}{\Pr\left(I_i^{\text{obs}}\mid\boldsymbol{\Xi}\right)} = \Pr\left(\boldsymbol{z}_i^{\text{obs}}\mid I_i^{\text{obs}},\boldsymbol{\Xi}\right), \tag{4.40}$$

we obtain:

$$\Pr(\boldsymbol{d},\mathbf{I}\mid\boldsymbol{\Xi}) = \prod_{i=1}^{N^{\text{obs}}}\Pr\left(I_i^{\text{obs}}\mid\boldsymbol{z}_i^{\text{obs}},\boldsymbol{\Xi}\right)\Pr\left(N^{\text{obs}}\mid N^{\text{tot}},\boldsymbol{\Xi}\right), \tag{4.41}$$

where we have introduced the probability of collecting a dataset with exactly $N^{\text{obs}}$ objects from a population of $N^{\text{tot}}$:

$$\Pr\left(N^{\text{obs}}\mid N^{\text{tot}},\boldsymbol{\Xi}\right) = \binom{N^{\text{tot}}}{N^{\text{obs}}}\times\prod_{i=1}^{N^{\text{obs}}}\Pr\left(I_i^{\text{obs}}\mid\boldsymbol{\Xi}\right)\times\prod_{j=1}^{N^{\text{tot}}-N^{\text{obs}}}\Pr\left(I_j^{\text{miss}}\mid\boldsymbol{\Xi}\right). \tag{4.42}$$

This binomial distribution represents the probability of $N^{\text{obs}}$ selections from $N^{\text{tot}}$ trials, based on individual detection probabilities that may vary with observing conditions and noise. Since $N^{\text{tot}}$ is unknown, it needs to be treated as a parameter, as apparent from the DAG in Fig. 4.8, which shows that in general $N^{\text{tot}}$ may also depend on its own parameters, $\boldsymbol{\gamma}$. Assigning it a prior $\Pr\left(N^{\text{tot}}\mid\boldsymbol{\Xi},\boldsymbol{\gamma}\right)$ and marginalizing over $N^{\text{tot}}$ in the joint distribution results in:

$$\Pr\left(N^{\text{obs}}\mid\boldsymbol{\Xi},\boldsymbol{\gamma}\right) = \int\Pr\left(N^{\text{obs}}\mid N^{\text{tot}},\boldsymbol{\Xi},\boldsymbol{\gamma}\right)\Pr\left(N^{\text{tot}}\mid\boldsymbol{\Xi},\gamma\right)\mathrm{d}N^{\text{tot}}. \tag{4.43}$$

We further adopt two simplifying assumptions that often apply: first, the prior

for the cardinality of the total population follows a Poisson distribution, so that $\Pr\left(N^{\text{tot}} \mid \Xi, \boldsymbol{\gamma}\right) = \text{Poiss}\left[\lambda(\Xi, \gamma)\right]$; second, objects are a priori indistinguishable, so that in Eq. (4.42) $\Pr\left(I_i^{\text{obs}} \mid \Xi\right) = \Pr\left(I^{\text{obs}} \mid \Xi\right) \equiv v$ for all $i = 1, \ldots, N^{\text{obs}}$, and therefore $\Pr\left(I_j^{\text{miss}} \mid \Xi\right) = 1 - v$ for all $j = N^{\text{obs}} + 1, \ldots, N^{\text{tot}}$. Under these conditions, the marginalization becomes:

$$\Pr\left(N^{\text{obs}} \mid \Xi, \boldsymbol{\gamma}\right) = \int \binom{N^{\text{tot}}}{N^{\text{obs}}} v^{N^{\text{obs}}} (1-v)^{1-N^{\text{obs}}} \frac{\lambda^{N^{\text{tot}}}}{N^{\text{tot}}!} \exp(-\lambda) \mathrm{d}N^{\text{tot}}, \tag{4.44}$$

leading to the result:

$$\Pr\left(N^{\text{obs}} \mid \Xi, \boldsymbol{\gamma}\right) = \text{Poiss}\left[\lambda(\Xi, \gamma) v\right]. \tag{4.45}$$

Therefore, the **observed data likelihood** can be expressed as follows:

$$\Pr\left(\boldsymbol{d}^{\text{obs}}, N^{\text{obs}} \mid \Xi, \boldsymbol{\gamma}\right) = \Pr\left(N^{\text{obs}} \mid \Xi, \boldsymbol{\gamma}\right) \prod_{i=1}^{N^{\text{obs}}} \frac{\Pr\left(I_i^{\text{obs}} \mid \boldsymbol{z}_i^{\text{obs}}, \Xi\right) \Pr\left(\boldsymbol{z}_i^{\text{obs}} \mid \Xi\right)}{\Pr\left(I^{\text{obs}} \mid \Xi\right)}. \tag{4.46}$$

which factorizes into the probability of observing $N^{\text{obs}}$ sources, given by Eq. (4.45), times the probability for the observed objects, corrected for selection effects. To further understand this result, we need to distinguish among different cases for the selection probability, $\Pr\left(I_i^{\text{obs}} \mid \boldsymbol{z}_i^{\text{obs}}, \Xi\right)$:

1 When $\Pr\left(I_i^{\text{obs}} \mid \boldsymbol{z}_i^{\text{obs}}, \Xi\right)$ is independent of $\boldsymbol{z}_i^{\text{obs}}$, then it must also be independent of $\Xi$, since a dependency on $\Xi$ necessarily entails also a dependency on $\boldsymbol{z}_i^{\text{obs}}$. The selection-effects correction terms in Eq. (4.46) are therefore constant, since

$$\Pr\left(I^{\text{obs}} \mid \Xi\right) = \int \Pr\left(I^{\text{obs}} \mid \boldsymbol{z}, \Xi\right) \Pr\left(\boldsymbol{z} \mid \Xi\right) \mathrm{d}\mathbf{z} \propto \int \Pr\left(\boldsymbol{z} \mid \Xi\right) \mathrm{d}\mathbf{z} = \text{const}.$$

Therefore **selection effects can be ignored**. This case is called 'missing at random'.

2 When $\Pr\left(I_i^{\text{obs}} \mid \boldsymbol{z}_i^{\text{obs}}, \Xi\right) = \Pr\left(I_i^{\text{obs}} \mid x_i^{\text{obs}}\right)$ depends on the observed independent variable, $x_i^{\text{obs}}$, then selection effects can be ignored insofar as the regression parameters $\boldsymbol{\theta}$ are concerned, with the caveat that the posterior distribution for the population-level parameters $x_0, R_x$ now describes **the distribution of the observed (as opposed to latent) objects**.

3 When $\Pr\left(I_i^{\text{obs}} \mid \boldsymbol{z}_i^{\text{obs}}, \Xi\right)$ additionally depends on the observed response variable $y_i^{\text{obs}}$, then an explicit dependence on the regression parameters $\boldsymbol{\theta}$ is introduced. In this case, from Eq. (4.46) we see that the observed data likelihood is equal to the standard likelihood for the observed data times a **'selection correction' term** given by $\Pr\left(I_i^{\text{obs}} \mid \boldsymbol{z}_i^{\text{obs}}, \Xi\right) \Pr\left(I^{\text{obs}} \mid \Xi\right)^{-1}$. This case is called 'truncation'.

For a more complete discussion, see Kelly (2007), where the general case of the Gaussian linear model is treated, including the case of censoring (where some data points only have upper limits). We illustrate below the issues with the correction terms with a simple one-dimensional case, following Karchev and Trotta (2024).

**Example 4.1: Toy selection effects model**

We simplify the hierarchical model to the extreme and adopt the following setup: $x_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$, $x_i^{\rm obs}|x_i \sim \mathcal{N}\left(x_i, \sigma_x^2\right)$, subject to the selection criterion $x_i^{\rm obs} > 0$. We aim to infer $\boldsymbol{\theta} \equiv \{\mu, \sigma\}$, with priors $\Pr\left(\mu\right) = \text{Uniform}(-1,1)$ and $\Pr\left(\log\sigma\right) = \text{Uniform}(0,1)$ and fixed measurement noise $\sigma_x = 0.2$. This is an example of case 3) above, for $x$ here plays the role of response variable since it is directly estimating the parameter of interest, $\mu$.
The standard likelihood for the observed data is:

$$\Pr\left(x_i^{\rm obs} \mid \boldsymbol{\theta}\right) = \frac{\exp\left(-(x_i^{\rm obs} - \mu)^2/2(\sigma^2 + \sigma_x^2)\right)}{\sqrt{2\pi(\sigma^2 + \sigma_x^2)}}. \tag{4.47}$$

The per-object selection probability $\Pr\left(I_i^{\rm obs} \mid x_i^{\rm obs}, \boldsymbol{\theta}\right) = \Pr\left(I_i^{\rm obs} \mid x_i^{\rm obs}\right)$ is independent of parameters, and is unity for selected objects. Integrating the data likelihood within $x_i^{\rm obs} > 0$ gives for the selection probability:

$$\Pr\left(I_i^{\rm obs} \mid \boldsymbol{\theta}\right) = \frac{1}{2}\left[1 + \text{erf}(\mu/\sqrt{2(\sigma^2 + \sigma_x^2)})\right]. \tag{4.48}$$

Using a data set with fiducial values $\mu = 0$, $\sigma = 0.5$, and $N^{\rm obs} = 100$, we evaluate all terms in (4.40), and depict them separately and combined in Fig. 4.9. The true values lie in low-likelihood regions of each of the terms. However, combined, they cancel nearly perfectly, giving the green contours for the observed data likelihood. The cancellation is very significant in this example: across the 2-$\sigma$ region, each term varies by a factor $\sim \text{e}^{100}$.

The example above highlights a challenge of the numerical modeling of selection effects: as data grows and observed objects differ more from the overall population, parameter constraints using only observed data become increasingly biased. At the same time, the correction (compounded $N^{\rm obs}$ times) grows to adjust the maximum likelihood region toward the true parameters. Each term?s log-values grow linearly with the number of observations, whereas the high-

 

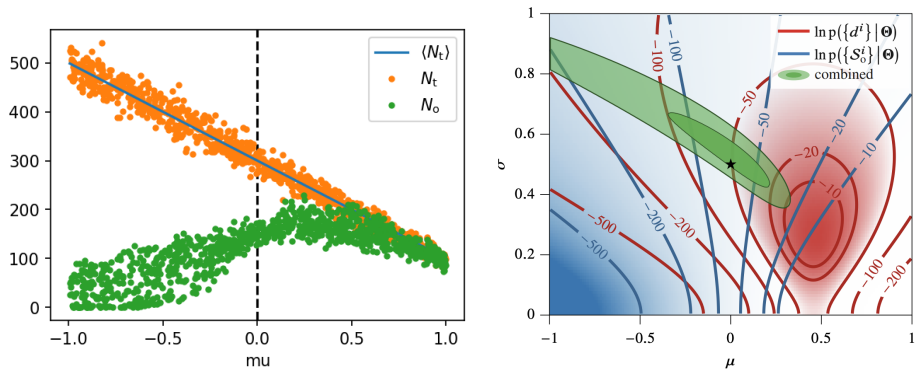


Figure 4.9 Left: cardinality of datasets for example 4.3.4 as a function of true $\mu$ value, showing total number of objects ($N^{\text{tot}}$, orange) and of selected objects (green), under a selection probability of unity for $x_i^{\text{obs}} > 0$ (and 0 otherwise). Right: Cancellation of likelihood terms for 100 selected objects. Red and blue shading with iso-log-likelihood contours represent the standard likelihood for the observed data, Eq. (4.47), and the selection probability correction term, Eq. (4.48). In this arrangement, they nearly cancel leading to a combined likelihood $\Pr\left(x_i^{\text{obs}} \mid \boldsymbol{\theta}\right) / \Pr\left(I_i^{\text{obs}} \mid \boldsymbol{\theta}\right)$ (green 1- and 2-$\sigma$ contours) peaking in the tails of both. Figure from Karchev and Trotta (2024).

combined-likelihood region contracts with the square root of data size, worsening the problem for large datasets. Thus, if we approximate the two terms separately, the estimates must be highly precise in both terms' tails. From a computational perspective, the selection probability term requires integration over the latent variables, which is typically expensive, and breaks the conditional dependency structure that enables efficient Gibbs sampling in a hierarchical model. Methods that have been developed to address these issues include using normalizing flows for the selection probability, and deep sets in simulation-based inference.

## 4.4 Posterior Sampling

Once a model has been built, the question becomes how to approximate numerically the posterior, $\Pr(\boldsymbol{\theta} \mid \boldsymbol{d})$, in cases where no closed-form expression exists. One simple solution would be to generate $n$ samples from the desired distribution (here, the posterior), and replace e.g. the expectation value with the sample mean, and the variance with the sample variance, i.e. for any function $h(\boldsymbol{\theta})$ and for samples $\boldsymbol{\theta}_i \sim \Pr(\boldsymbol{\theta} \mid \boldsymbol{d})$, for $i = 1, \ldots, n$ we have that

$$\mathrm{E}\left[h(\boldsymbol{\theta})\right] = \int \Pr(\boldsymbol{\theta} \mid \boldsymbol{d})\, h(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \leftarrow \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{\theta}_i) \text{ for } n \to \infty.$$

The difficulty with this so-called 'Monte Carlo method' (a name inspired by the famous gambling city in Southern France[3]) is that the posterior target needs to be fully specified and samples must be drawn from it. This is often unfeasible in complex models of the kind we saw above, where the conditionals are known but the full posterior is not analytically tractable.

An option for numerical sampling from a 1D posterior, $\pi(\theta)$ that is not analytically tractable is **rejection sampling**. Rejection sampling is a technique to generate samples from a target distribution $\pi(\theta)$ in one dimension by using a proposal distribution $q(\theta)$ which is easier to sample from. It works as follows:

1 Select a proposal distribution $q(\theta)$ such that

$$q(\theta) \geq \frac{\pi(\theta)}{M} \quad \text{for all} \quad \theta,$$

where $M$ is a constant scaling factor.
2 For each sample, first draw $\theta \sim q(\theta)$, then generate a uniform random variable $u \sim U(0,1)$.
3 Accept $\theta$ as a sample from $\pi(\theta)$ if

$$u < \frac{\pi(\theta)}{M \cdot q(\theta)};$$

otherwise, reject and repeat.

The accepted samples approximate the target distribution $\pi(\theta)$. The efficiency depends on how well $M \cdot q(\theta)$ fits $\pi(\theta)$, as larger values of $M$ increase the rejection rate.

Another technique is **slice sampling** (Neal, 2003), illustrated in Fig. 4.10. Denoting the target 1D posterior to be sampled from by $\pi(\theta)$, slice sampling proceed as follows:

1 Start from a point $\theta_0$. Draw $y \sim \text{Uniform}(0, \pi(\theta_0))$ and define a horizontal slice $S : \{x \,|\, y < \pi(\theta_0)\}$ (grey horizontal line in Fig. 4.10), to which $\theta_0$ belongs by construction;
2 starting from $\theta_0$, place an interval of size $w$ at random around $\theta_0$; expand it in steps of $w$ until all of the slice is contained in the interval—this defines an interval $[L, R]$, dot-dashed vertical lines in Fig. 4.10;
3 draw $\theta_1 \sim \text{Uniform}(L, R)$, repeating until you obtain a point on the slice.

Of course rejection and slice sampling only work in one dimension, which is

[3] The name appears to have been coined by Nick Metropolis (Metropolis, 1987). In one of the early expositions of the then-new Monte Carlo method, Metropolis and Ulam (1949) quaintly (to our contemporary ears) write: 'We want now to point out that modern computing machines are extremely well suited to perform the procedure described. In practice, the set of values of parameters characterizing a particle is represented, for example, *by a set of numbers punched on a card*' (my emphasis).

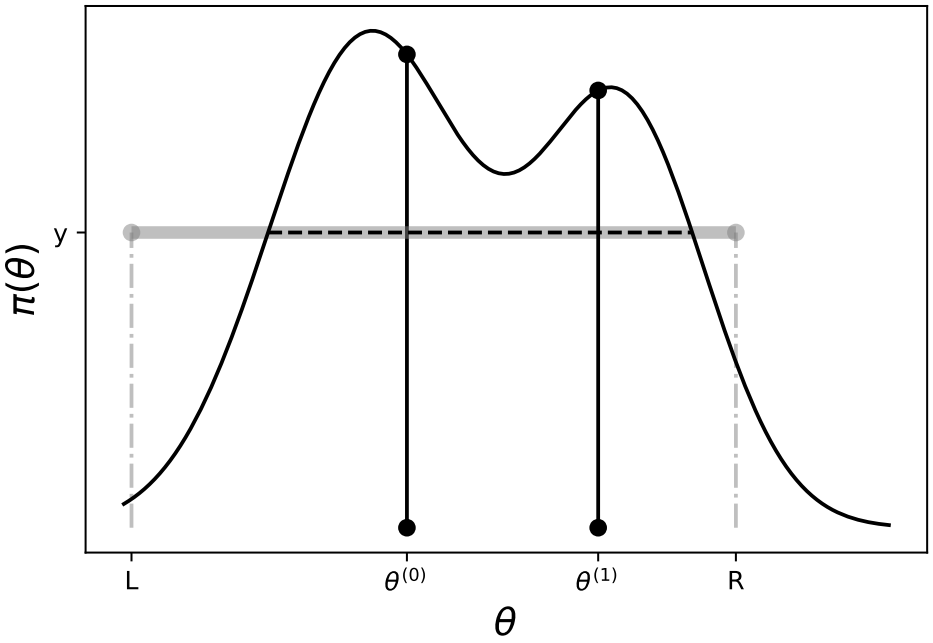


Figure 4.10 Illustration of slice sampling in 1D.

hardly helpful in the majority of inference problems (though this method can be integrated in a Gibbs sampler, as we shall see below). A better, more flexible approach to the problem of high-dimensional posterior sampling is represented by Markov Chain Monte Carlo (MCMC) methods — a suite of algorithms for sampling from complex probability distributions of the kind we have encountered above. Here, too, the eventual goal is to obtain samples from a target distribution[4], in this case the posterior, but circumventing the need to perform direct sampling. The fundamental idea is to construct a simple Markov chain (a sequence of random variables where each state depends only on the previous one[5] such that the posterior distribution is the limiting distribution of the chain. We turn to the necessary theory before introducing some of the algorithms suited to this task.

[4] In this context, statisticians sometimes talk about 'simulating from the posterior', by which they mean 'sampling from the posterior'; this meaning of 'simulation' is quite different from the one physicists usually employ.

[5] Nafter the Russian mathematician Andrey Markov, who studied chains in the early 20th century.

### 4.4.1 Markov Chain Monte Carlo Methods

We begin by introducing the concept of **Markov Chain (MC)**: a sequence of random variables $\{X_0, X_2, \dots\}$ such that the probability of the element $t+1$ only depends on the value of the $t$-th element. A MC over a discrete, countable space state $S = \{s_1, s_2, \dots, s_N\}$ is defined by its **transition probability matrix** $\boldsymbol{T} \in \mathbb{R}^{N\times N}$ for jumping from state $i$ to state $j$:

$$t_{ij} = \Pr\left(X_{t+1} = s_j \mid X_t = s_i\right).$$

A MC is called **irreducible**, if any point in the state space can be reached starting from any other point, i.e. $\Pr\left(X_n = s_j \mid X_0 = s_i\right) > 0 \,\forall\, s_i, s_j$.

The discrete probability distribution $\pi = \{\pi, \dots, \pi_N\}$ is **stationary** (or **invariant**) for a transition probability $t_{ij}$ (and its associated MC, $\{X_n\}_{n\geq 0}$) if the probability distribution for $X_0$ is the same as for all other $i > 0$, i.e. (for the discrete case):

$$\Pr\left(X_1 = s_j\right) = \sum_i \Pr\left(X_1 = s_j \mid X_0 = s_i\right) \Pr\left(X_0 = s_i\right) = \sum_i t_{ij}\pi_i \overset{!}{=} \Pr\left(X_0 = s_j\right) \quad (4.49)$$

Tthis means that the stationary distribution fulfils:

$$\pi = \pi \boldsymbol{T}, \quad (4.50)$$

where $\boldsymbol{T}$ is the matrix with entries $t_{ij}$.

For a continuous state space, a probability distribution $\pi$ with density $p_\pi(x)$ is stationary for the transition kernel $\Pr(x \to \cdot)$ if

$$\pi(A) = \int_A p_\pi(x)\mathrm{d}x = \int_S \Pr(x \to A)\, p_\pi(x)\mathrm{d}x \quad \forall A \subset S. \quad (4.51)$$

Finally, a MC is called **aperiodic**, if no partition of the state space exists such that the MC exhibits periodicity over that partition.

**Theorem 4.1** (Law of Large Numbers for Markov Chains (discrete state space)) *Let $\{X_n\}_{n\geq 0}$ be an irreducible MC with transition matrix $t_{ij}$ and stationary distribution $\pi = \{\pi_i : s_i \in S\}$. Then for any bounded function $h : S \to \mathbb{R}$ and for any initial distribution of $X_0$*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} h(X_i) \xrightarrow{in\,p.} \sum_j h(s_j)\pi_j. \quad (4.52)$$

This means that, for large $n$, the estimate $\frac{1}{n}\sum_{i=0}^{n-1} h(X_i)$ will converge in probability to the expectation value of $h(x)$ under the Markov chain's stationary distribution $\pi$. For the proof, see e.g. Rosenthal (2001).

This result opens the door to a new way of approaching the problem of sampling from the posterior: if we can construct a suitable MC with the posterior

as its target, we need not sample directly from the posterior, but merely from the transition matrix, which is hopefully a much simpler task. Our task is thus to construct a MC with a given target distribution $\pi$ as its stationary state, which later will be identified with the posterior.

The next theorem further guarantees that the distribution of $X_n$ for $n \to \infty$ converges to the stationary distribution, irrespective of the chain's starting point.

**Theorem 4.2** (Asymptotic convergence of Markov Chains) *Let $\boldsymbol{T}$ be the transition matrix for an irreducible, aperiodic MC with stationary distribution $\pi$. Then for any initial distribution of $X_0$,*

$$\lim_{n\to\infty} |\Pr\left(X_n = s_j\right) - \pi_j| = 0\,. \tag{4.53}$$

For a proof, see e.g. Rosenthal (2001).

We now proceed to construct a MC with a given target distribution $\pi$ as its stationary state, which we will later identify with the posterior. We start from the discrete case for simplicity, and later generalize to a continuous state space. Let $Q = \{q_{ij}\}$ be a transition probability matrix, called **proposal transition probability**, such that it is simple to generate a sample from the distribution defined by $\{q_{ij} \mid j \in S\}$ over the state space $S$. Start from $X_t = s_i$, and build the next step in the MC as follows: sample $Y_t$ from $\{q_{ij} \mid j \in S\}$, then choose $X_{t+1}$ according to:

$$\Pr\left(X_{t+1} = Y_t \mid X_t, Y_t\right) = \rho(X_t, Y_t) \text{ (move with probability } \rho)\,, \tag{4.54}$$

$$\Pr\left(X_{t+1} = X_t \mid X_t Y_t\right) = 1 - \rho(X_t, Y_t) \text{ (stay with probability } 1-\rho)\,, \tag{4.55}$$

where the **acceptance probability** $\rho(X_t, Y_t)$ is given by:

$$\rho(X_t = s_i, Y_t = s_j) = \min\left\{\frac{\pi_j}{\pi_i}\frac{q_{ji}}{q_{ij}}, 1\right\}. \tag{4.56}$$

This generates a MC $\{X_n\}_{n\geq 0}$ with transition probability matrix $t_{ij}$ given by

$$t_{ij} = \begin{cases} q_{ij}\rho(s_i, s_j)\,, & j \neq i \\ 1 - \sum_{k\neq i} t_{ik}\,, & j = i\,. \end{cases} \tag{4.57}$$

Because of the symmetry in the indexes in Eq. (4.56), this transition mechanism satisfies by construction

$$\pi_i t_{ij} = \pi_j t_{ji}\,. \tag{4.58}$$

This crucial property means that, summing both sides over $i$:

$$\sum_i \pi_i t_{ij} = \pi_j \sum_i t_{ji} = \pi_j\,, \tag{4.59}$$

which hold for all $j$, and we have used that $\sum_i t_{ji} = 1$. Therefore $\pi = \{\pi_1, \ldots, \pi_N\}$ satisfies Eq. (4.50) and it is the stationary distribution of the MC, as desired.

If we now consider a continuous space state instead, we replace $\{s_1, \ldots, s_N\}$ with a continuous (possibly multivariate) RV $\boldsymbol{\theta}$, and the transition matrix defining the MC is replaced by a **transition kernel** $T(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t)$, which gives the probability for the chain to move from $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1}$, and is a density in the first argument. In analogy with Eq. (4.58), a sufficient condition for the chain to have target density $\pi$ is that its transition probability satisfies the **detailed balance** condition:

$$\pi(\boldsymbol{\theta}_t)\, T(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t) = \pi(\boldsymbol{\theta}_{t+1})\, T(\boldsymbol{\theta}_t; \boldsymbol{\theta}_{t+1}) \tag{4.60}$$

If detailed balance is satisfied, the chain is called **reversible**, since the product of the ratio of target densities between two neighbouring states is the same as the ratio of their respective transition probabilities.

Similarly to how we constructed the MC above for the discrete case, starting from $\boldsymbol{\theta}_t$ we draw a candidate point $\boldsymbol{\theta}_c \sim Q(\cdot \mid \boldsymbol{\theta}_t)$ from a **proposal density** (also called "candidate kernel" or "jumping kernel") $Q$, which may only depend on the current location, $\boldsymbol{\theta}_t$ to fulfil the Markov condition. The acceptance probability for the move, Eq. (4.56), becomes:

$$\rho(\boldsymbol{\theta}_c, \boldsymbol{\theta}_t) = \min\left[1, \frac{\pi(\boldsymbol{\theta}_c)\, Q(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_c)}{\pi(\boldsymbol{\theta}_t)\, Q(\boldsymbol{\theta}_c \mid \boldsymbol{\theta}_t)}\right]. \tag{4.61}$$

This means that we move $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_c$ with probability $\rho$, in which case $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_c$; and we stay in $\boldsymbol{\theta}_t$ with probability $1-\rho$, thus assigning $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t$. It is important to notice that only the ratio of the target probabilities appear in Eq. (4.56); therefore, when replacing $\pi$ with the target posterior, it is sufficient to use the unnormalized posterior, i.e., the. product of the prior with the likelihood, since the normalizing constants cancel. This is a major computational advantage. We highlight once again that in this scheme there is no sampling from the posterior, which only needs to be evaluated (up to a normalization constant) at the current location of the Markov chain and at the candidate point in parameter space.

The transition probability for a move $\boldsymbol{\theta}_t \to \boldsymbol{\theta}_{t+1}$ can be written as, in analogy with Eq. (4.57):

$$\begin{aligned} T(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t) &= Q(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t)\, \rho(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) \\ &= \min\left[\frac{\pi(\boldsymbol{\theta}_t)\, Q(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t)}{\pi(\boldsymbol{\theta}_t)}, \frac{\pi(\boldsymbol{\theta}_{t+1})\, Q(\boldsymbol{\theta}_t; \boldsymbol{\theta}_{t+1})}{\pi(\boldsymbol{\theta}_t)}\right]. \end{aligned} \tag{4.62}$$

From this, it follows that

$$T(\boldsymbol{\theta}_{t+1}; \boldsymbol{\theta}_t)\, \pi(\boldsymbol{\theta}_t) = \min\left[\pi(\boldsymbol{\theta}_t)\, Q(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t), \pi(\boldsymbol{\theta}_{t+1})\, Q(\boldsymbol{\theta}_t; \boldsymbol{\theta}_{t+1})\right]$$

is explicitly symmetric under the exchange $\boldsymbol{\theta}_t \leftrightarrow \boldsymbol{\theta}_{t+1}$, hence detailed balance, Eq. (4.60), is satisfied.

### 4.4.2 Metropolis-Hastings

We now have all the ingredients necessary to build a MCMC algorithm to obtain samples from the posterior as the chain's stationary distribution. The simplest example of a reversible transition probability is Metropolis–Hastings[6]: given an arbitrary proposal density $Q$, the **Metropolis-Hasting algorithm**(see e.g. Robert (2015) and references therein) proceeds as follows:

1 Initialize $t = 0$, $\boldsymbol{\theta}_0 \sim \Pr(\boldsymbol{\theta})$.
2 Draw the next candidate point: $\boldsymbol{\theta}_c \sim Q(\cdot;\boldsymbol{\theta}_t)$.
3 Accept $\boldsymbol{\theta}_c$ with probability given by Eq. (4.61); if the candidate is accepted, set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_c$; else set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$.
4 Increase counter $t \to t+1$; loop back to step 2.

If the candidate point is not accepted in the accept/reject step, the current sample value, $\boldsymbol{\theta}_t$ is retained in the chain; in practice, this is more efficiently accomplished by increasing the so-called **multiplicity** $m_t$ of the sample by one unit; the multiplicity of each sample is thus equal to:

$$m_t = 1 + r_{t+1}\,, \tag{4.63}$$

where $r_{t+1)}$ is the number of rejections when proposing a new candidate sample for step $t+1$.

The above holds true for any choice of proposal density $Q$; in the special case in which the proposal density is symmetric in its arguments, i.e., $Q(\boldsymbol{\theta}_t;\boldsymbol{\theta}_{t+1}) = Q(\boldsymbol{\theta}_{t+1};\boldsymbol{\theta}_t)$, the algorithm is called **Metropolis**. It is also worth noticing that a convex sum $\sum_i w_i Q_i$ of $N$ Metropolis-Hastings proposal densities $Q_1, Q_2, \ldots, Q_N$ with positive weights $\sum_i w_i = 1$ is also a valid proposal kernel.

For any choice of $Q$, the convergence to the true posterior is only guaranteed asymptotically, i.e. for $t \to \infty$ by the convergence and law of large numbers for Markov chains. We thus need to address the questions of how to choose $Q$ optimally, i.e., in order to achieve convergence after a finite (and as as small as possible) number of steps.

It is useful to review some of the potential issues that Markov Chains can run into before addressing the question of optimality. Let us consider the simplest case of a Metropolis kernel in the form of a Normal proposal density with fixed covariance matrix $\boldsymbol{C}$, i.e.

$$Q(\cdot;\boldsymbol{\theta}_t) = \mathcal{N}(\boldsymbol{\theta}_t, \boldsymbol{C})$$

As the chain progresses, it is useful to monitor its **acceptance rate** $\mathscr{R}$, defined

[6] Originally introduced in Metropolis et al. (1953) using a symmetric proposal distribution and later generalized in the form presented here allowing for non-symmetric proposal distributions by Hastings (1970).

as the number of accepted candidate steps divided by the number of proposed steps. As illustrated by the sketches in Fig. 4.11, the following situations might arise:

- Top left: when the proposal density $Q$ is too concentrated when compared with the target density (the posterior), this leads to diffusion and hence slow exploration. In this case, $\mathscr{R} \to 1$.
- Top right: if the principal directions of $Q$ are not aligned with the target density, many proposals will be rejected and the chain will be stuck. In this case, $\mathscr{R} \to 0$.
- Middle left: when the posterior exhibits curved degeneracies, a fixed proposal density will not be able to efficiently explore the target in all regions, as its shape might be well suited in a certain region of parameter space but quite inefficient in another.
- Middle right: when the target is multi-modal, it is often difficult to explore all the modes jointly. A proposal that is well adapted to one of the modes might be very inefficient for the other, and jumping across modes may be difficult. Some modes might be missed altogether.
- Bottom left: in the presence of chimney-like targets (where the density drops to 0 in the middle), there are two very different scales to the target, which makes exploration difficult with a single, fixed proposal density.
- Bottom right: hard prior boundaries (e.g., positivity constraints) might lead to under-exploration of the target near the boundary, as candidates are rejected when they encroach onto the region where the prior has no support. This typically leads to approximate posteriors with smaller density near the boundary than the true posterior.

*Adaptive Metropolis*

A way to address some of the issues outlined above consists in dynamically adapting $Q$ to the target, based on information learnt from the chain up to that point. This results in the **adaptive Metropolis algorithm**, which is discussed in further detail in Haario et al. (2001).

The algorithm introduces a proposal distribution in Metropolis–Hastings that explicitly depends on the previous $t_0$ steps:

$$Q^{(t+1)} = \mathcal{N}\left(\boldsymbol{\theta}_t, \boldsymbol{C}_t\right) \tag{4.64}$$

with covariance matrix given by

$$\boldsymbol{C}_t = \begin{cases} \mathbf{1}_d, & \text{for } t \leq t_0 \\ s_d \left[\mathbf{Cov}_t\left(\boldsymbol{\theta}_{t-t_0}, \ldots, \boldsymbol{\theta}_t\right) + \epsilon \mathbf{1}_d\right], & \text{for } t > t_0 \,. \end{cases} \tag{4.65}$$

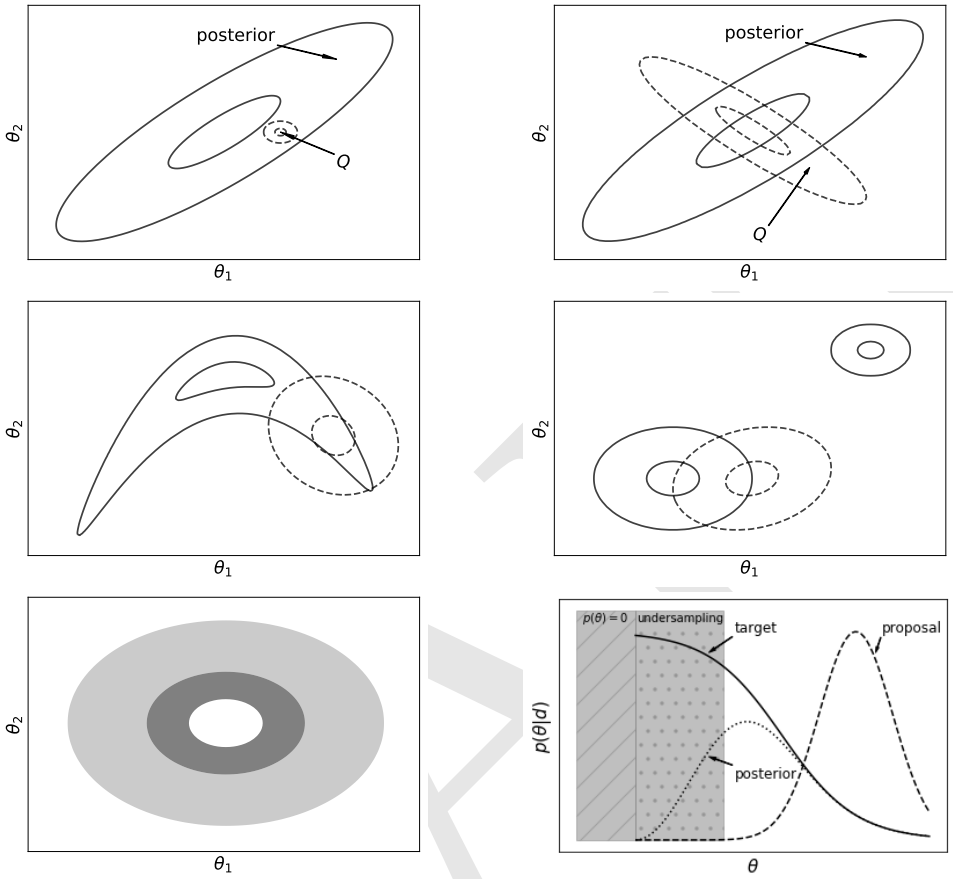


Figure 4.11 Potential issues with Metropolis MCMC proposal densities. Top left: too small proposal density, $Q$ (dashed contours), w.r.t. the size of the target posterior (solid contours). Top right: $Q$ not well aligned with the target density. Middle left: degenerate, e.g. banana-shaped target, which is difficult to explore for a fixed proposal density. Middle right: multiple modes in the target, difficult to explore jointly. Bottom left: chimney-like targets, which are difficult to explore reliably. Bottom right: hard prior boundaries might lead to under-exploration of the posterior near the boundary.

Here $\epsilon \ll 1$ and the term proportional to the identity matrix are related to the need of ensuring ergodicity of the chain and protects from singularities; $s_d$ is a numerical coefficient that depends on the dimensionality of the parameter space $d = \dim(\boldsymbol{\theta})$, see Eq. (4.69) for a justification:

$$s_d = \frac{(2.4)^2}{d} \tag{4.66}$$

The covariance matrix is estimated from the samples gathered up to step $t$, and

 

updated at each step, or every $k$ steps. The behaviour of the algorithm is illustrated in Fig. 4.12.

In principle, such a dynamic updating appears potentially problematic, as it breaks the Markov condition and hence asymptotic convergence to the target is no longer guaranteed by the law of large numbers for Markov chains. Nevertheless, the following theorem (Haario et al., 2001) assuages such doubts:

**Theorem 4.3** (Adaptive Metropolis) *Let $\pi$ be a target distribution, with compact support $S$ and bounded from above. Then the adaptive Metropolis algorithm simulates properly from the target, i.e.,*

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} h(\boldsymbol{\theta}_i) = \int_S h(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \tag{4.67}$$

*almost surely for all bounded and measurable $h : S \to \mathbb{R}$.*

While adaptive Metropolis is useful in adjusting the proposal to the target distribution's shape, it does not help to mitigate the difficulty of sampling from a multimodal distribution.

We now return to the numerical rescaling factor $s_d$ introduced in Eq. (4.66) above. Its role is to rescale the principal components of the proposal distribution with respect to the target density's covariance by a dimension-dependent factor, $\sqrt{s_d}$, which varies as the inverse of the square root of the dimensionality of the parameter space. Its justification lies in the following result by Gelman et al. (1996).

**Theorem 4.4** (Optimal jumping kernel) *Consider a multivariate Metropolis proposal with factorizable target*[7]

$$\pi(\boldsymbol{\theta}) = \prod_{i=1}^{d} f_i(\boldsymbol{\theta}_i) \tag{4.68}$$

*for some densities $f_i$, $(i = 1,\ldots,d)$; then the speed of diffusion of a MH MCMC is maximised by a Normal proposal distribution, centered on the current value, with covariance matrix given by:*

$$\frac{(2.38)^2}{d}\mathbf{1}_d \tag{4.69}$$

*and asymptotic jumping probability (i.e. acceptance rate) of 0.234.*

Modern MCMC algorithms use different approaches to adapt to the shape of

[7] Writing the target density as a product of independent densities is equivalent to rotating the coordinate system to align with the principal directions of the target's covariance matrix, for the case of a Gaussian linear model.

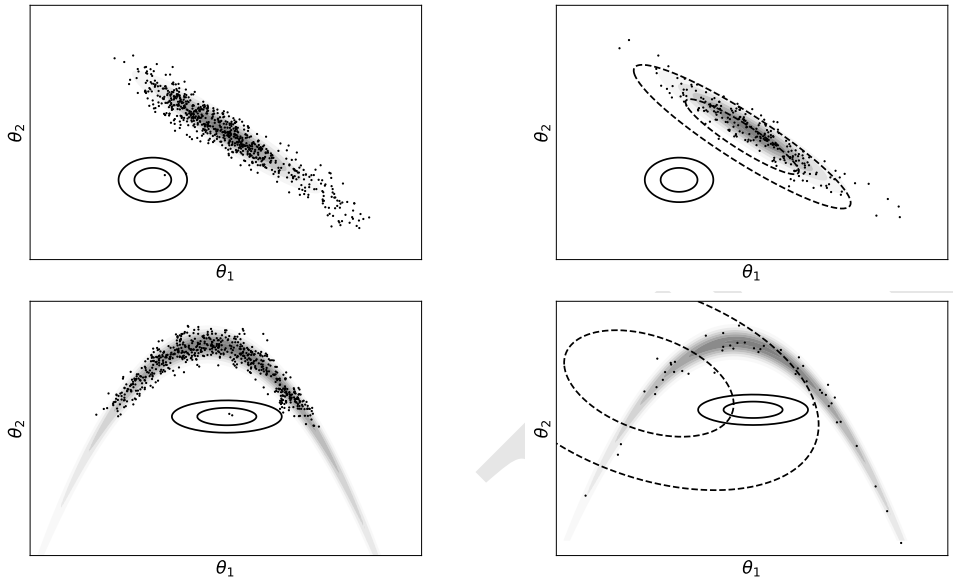


Figure 4.12 Illustration of the Adaptive Metropolis algorithm (right), compared with non-adaptive MH (left) for the same target densities (gray shaded regions) and the same number of accepted steps (dots, representing single-weight samples). The solid ellipses give the initial Gaussian proposal density (or the fixed proposal for MH), the dashed ellipses the final proposal after adaptation. A Gaussian target (top) and a curved degeneracy (bottom panels) are explored with non-adaptive MH, leading to under-exploration. The same targets are much better explored using Adaptive Metropolis.

the target dynamically. The widely-used `emcee` package[8], for example, implements an affine-invariant ensemble sampling MCMC introduced by Goodman and Weare (2010), which is particularly effective for highly anisotropic posterior distributions. Unlike traditional MCMC, `emcee` uses multiple "walkers" (i.e, independent chains) that jointly sample the parameter space by learning the relative positions of walkers in an affine-invariant way. Affine invariance refers to an algorithm?s ability to adapt to transformations of the parameter space without requiring parameter rescaling or axis alignment. In particular, `emcee` excels at sampling from distributions stretched or elongated in one direction. It achieves this by allowing each walker to move by "stretching" or shifting relative to other walkers. This independence from scaling and rotation transformations means the sampler can converge efficiently even in anisotropic distributions, main-

[8] Available as a Python implementation at `https://emcee.readthedocs.io/en/stable/` and described in detail in Foreman-Mackey et al. (2013).

taining invariant sampling behaviour across different shapes and orientations of the target posterior.

### *4.4.3 Gibbs sampling*

In cases such as hierarchical models, when the model's structure is reflected in the joint distribution breaking down naturally into a product of conditionals, it is helpful and effective to use Gibbs sampling – see Smith and Roberts (1993) for a review[9]. The idea underpinning the method is to divide the sampling steps by cycling through a series of so-called **full conditional distributions** (i.e., where each variable is conditioned upon all others), until all variables have been sampled from.

Let us denote by $\pi\left(\theta^{(1)},\dots,\theta^{(d)}\right)$ the target posterior density, where the parameters are $\boldsymbol{\theta} = \left(\theta^{(1)},\dots,\theta^{(d)}\right)$; and $\pi\left(\theta^{(j)}|\boldsymbol{\theta}^{(-j)}\right)$ the conditional posterior distribution of $\theta^{(j)}$ given all the other coordinates in the parameter vector except for the $j$-th, i.e., $\boldsymbol{\theta}^{(-j)} = \{\boldsymbol{\theta}^{(i)}|1 \leq i \neq j \leq d\}$. Here, we depart from the notation used elsewhere and we use a superscript in parentheses to denote the components of a parameter vector, while a subscript will denote the element in the MC, as before.

Then the **Gibbs sampler** proceeds as follows:

0 Initialize: $\boldsymbol{\theta}_0 \sim \Pr(\boldsymbol{\theta})$, sampling from the prior density.
1 Sample $\theta_1^{(1)} \sim \pi\left(\theta^{(1)} \mid \boldsymbol{\theta}_0^{(-1)}\right)$;
2 sample $\theta_1^{(2)} \sim \pi\left(\theta^{(2)} \mid \theta_1^{(1)}, \theta_0^{(3)}, \dots, \theta_0^{(d)}\right)$;
3 sample $\theta_1^{(3)} \sim \pi\left(\theta^{(3)} \mid \theta_1^{(1)}, \theta_1^{(2)}, \theta_1^{(4)}, \dots, \theta_0^{(d)}\right)$;
$\vdots$
d sample $\theta_1^{(d)} \sim \pi\left(\theta_d \mid \boldsymbol{\theta}_1^{(-d)}\right)$;
d+1 transition from $\boldsymbol{\theta}_0 \to \boldsymbol{\theta}_1$. Loop back to step 1 and increase counters by 1 unit.

Note that the order in which the full conditional densities are sampled is unimportant for the result, as long as one cycles through all of the coordinate updates, although the choice of ordering might affect efficiency or simplicity of the algorithm. It is important to remember that the full conditional distribution for each variable only depends on its **Markov blanket**, i.e., the nodes that are directly connected to it via edges (that is, its parents and children).

It is possible and indeed at times desirable to bundle two or more updates

[9] The name comes from analogy with the Gibbs distribution in lattice-like physical systems, which was considered in the original paper by Geman and Geman (1984). It is however only with the advent of relatively fast and more easily programmable computers in the 1990s that the method really took off. In 1993, statistician Peter Clifford commented on the recent adoption of 'Metropolis methods' by scientists: '[...] from now on we can compare our data with the model we actually want to use rather than with a model which has some mathematical convenient form. This is surely a revolution' (Clifford, 1993).

 

together in a single, joint update, for efficiency or convenience (called **blocking**). For example, blocking directions 1 and 2 means jointly updating

$$\left\{\theta_{t+1}^{(1)}, \theta_{t+1}^{(2)}\right\} \sim \pi\left(\theta^{(1)}, \theta^{(2)} \mid \theta_t^{(3)}, \ldots, \theta_t^{(d)}\right); \tag{4.70}$$

this is a useful trick when the variables being blocked are strongly correlated.

The Gibbs sampler can be seen as a special case of Metropolis-Hastings, with a transition probability given by:

$$T\left(\boldsymbol{\theta}_t; \boldsymbol{\theta}_{t+1}\right) = \prod_{k=1}^{d} \pi\left(\theta_{t+1}^{(k)} \mid \boldsymbol{\theta}_t^{(j \neq k)}\right). \tag{4.71}$$

That cycling through the full conditionals for each variables gives at the end a sample from the joint posterior is not trivial (see Casella and George (1992) for a simple argument based on a 2D example). Under the condition that the joint posterior is strictly positive, the Hammersley-Clifford theorem[10] (see e.g. Winkler (2002)) guarantees that any distribution compatible with a graphical model structure can be simulated via Gibbs sampling.

If the conditional are analytically tractable, the acceptance rate is equal to 1, which leads to perfect efficiency (no moves are ever rejected). When sampling from one or more conditionals is not analytically tractable, then we can resort to using e.g. Metropolis-Hastings to obtain a sample from the intractable direction(s), or use rejection sampling or slice sampling for 1D conditionals.

As an explicit example, we return to the errors-in-variables model introduced in section 4.3.2 (Fig 4.3), which has the following conditional structure, with known $\sigma_x, \sigma_y$ (and no intrinsic scatter) for $i = 1, \ldots, N$:

$$y_i^{\text{obs}} \mid \boldsymbol{\theta}, x_i \sim \mathcal{N}\left(\theta_0 + \theta_1 x_i, \sigma_y^2\right), \tag{4.72}$$

$$x_i^{\text{obs}} \mid x_i \sim \mathcal{N}\left(0, \sigma_x^2\right), \tag{4.73}$$

$$x_i \sim \mathcal{N}(x_0, R_x^2), \tag{4.74}$$

$$x_0 \sim \mathcal{N}(\mu_{x_0}, \sigma_{x_0}^2), \tag{4.75}$$

$$R_x^2 \sim \text{Inv-Gamma}(\alpha_R, \beta_R), \tag{4.76}$$

where we have chosen conjugate priors for the population-level parameters $x_0, R_x^2$ (with fixed hyper-parameters $\mu_{x_0}, \sigma_{x_0}^2, \alpha_R, \beta_R$) to enable analytical computation of the relevant conditional distributions. The Gibbs sampler proceeds as follows (remembering that after each conditional sampling step, the value of the variable just sampled is updated in the conditional for the next step):

[10] While J.M. Hammersley and P.E. Clifford formulated the theorem in an unpublished paper in 1971, the first published version is due to Julian Besag, in 1974.

1 Sample the latent variables, $x_i$ $(i = 1, \dots, N)$: the conditional distribution for $x_i$ is obtained as, with $\boldsymbol{d}_i = \{x_i^{\text{obs}}, y_i^{\text{obs}}\}$, $\Xi = \{\boldsymbol{\theta}, x_0, R_x\}$:

$$\Pr(x_i \mid \boldsymbol{d}, \Xi) \propto \Pr(x_i, \boldsymbol{d}, \Xi) \propto \Pr(\boldsymbol{d} | x_i, \Xi) \Pr(x_i, \Xi) = \Pr(\boldsymbol{d} | x_i, \boldsymbol{\theta}) \Pr(x_i | x_0, R_x) ,$$

where the first term can be understood as a 'likelihood' (in that it pulls the conditional value of $x_i$ towards the observations), and the second as a 'prior'. The first term comes from the edges connecting the $x_i$ node to its children, and is given by

$$\Pr(\boldsymbol{d} | x_i, \boldsymbol{\theta}) = \mathcal{N}_{y_i^{\text{obs}}}\left(\theta_0 + \theta_1 x_i, \sigma_y^2\right) \mathcal{N}_{x_i^{\text{obs}}}\left(x_i, \sigma_x^2\right)$$

where in the above the value of $x_i$ being conditioned upon is that from the previous iteration of the Gibbs' sampler. Altogether, the full conditional for the latent $x_i$ takes the form:

$$\Pr(x_i \mid \text{rest}) \propto \mathcal{N}(y_i^{\text{obs}} \mid \theta_0 + \theta_1 x_i, \sigma_y^2) \cdot \mathcal{N}(x_i^{\text{obs}} \mid x_i, \sigma_x^2) \cdot \mathcal{N}(x_i \mid x_0, R_x^2) .$$

This can be recast as:

$$x_i \sim \mathcal{N}\left(\mu_x, \sigma_x^2\right) ,$$

where:

$$\sigma_x^2 = \left(\frac{\theta_1^2}{\sigma_y^2} + \frac{1}{\sigma_x^2} + \frac{1}{R_x^2}\right)^{-1} ,$$

$$\mu_x = \sigma_x^2 \left(\frac{\theta_1 (y_i^{\text{obs}} - \theta_0)}{\sigma_y^2} + \frac{x_i^{\text{obs}}}{\sigma_x^2} + \frac{x_0}{R_x^2}\right).$$

2 By a similar argument, the regression coefficients $\theta_0$ and $\theta_1$ only depend on their parent node (the top-level prior) and their children node, $y_i^{\text{obs}}$ ($y_i$ can be eliminated as it is deterministically related to $x_i$ and $\boldsymbol{\theta}$).Therefore, in the second step of the sampler the regression coefficient are sampled jointly (in blocked step) from:

$$\Pr(\boldsymbol{\theta} \mid \text{rest}) \propto \prod_i \mathcal{N}(y_i^{\text{obs}} \mid \theta_0 + \theta_1 x_i, \sigma_y^2) \cdot \mathcal{N}(\theta_0 \mid \mu_\theta, \sigma_\theta^2) \cdot \mathcal{N}(\theta_1 \mid \mu_\theta, \sigma_\theta^2)$$

where we have chosen a Normal prior for both $\theta_0$ and $\theta_1$, out of analytical convenience. Using the results of section 4.2, this distribution can be written as bivariate normal:

$$\Pr(\boldsymbol{\theta} \mid \text{rest}) \propto \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right) ,$$

where:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \left(\frac{1}{\sigma_y^2} \boldsymbol{A}^\top \boldsymbol{y}^{\text{obs}} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right),$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \left( \frac{1}{\sigma_y^2} \boldsymbol{A}^\top \boldsymbol{A} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1},$$

where the normalized design matrix has entries $A_{ij} = (\delta_{i1} + \delta_{i2} x_i)/\sigma_y^2$ for $j = 1, \ldots, N$, $\boldsymbol{\mu}_0 = \{\mu_\theta, \mu_\theta\}$ and $\boldsymbol{\Sigma}_0 = \text{diag}(\sigma_\theta^2, \sigma_\theta^2)$.

3 Sample the population-level mean of $x_i$, $x_0$:

$$\Pr(x_0 \mid \text{rest}) \propto \prod_i \mathcal{N}(x_i \mid x_0, R_x^2) \cdot \mathcal{N}(x_0 \mid \mu_{x_0}, \sigma_{x_0}^2),$$

which can be written as:

$$x_0 \sim \mathcal{N}(\mu'_{x_0}, \sigma'^2_{x_0}),$$

where:

$$\sigma'^2_{x_0} = \left( \frac{N}{R_x^2} + \frac{1}{\sigma_{x_0}^2} \right)^{-1},$$

$$\mu'_{x_0} = \sigma'^2_{x_0} \left( \frac{\sum_i x_i}{R_x^2} + \frac{\mu_{x_0}}{\sigma_{x_0}^2} \right).$$

4 Sample the variance of $x_i$, $R_x^2$:

$$\Pr\left(R_x^2 \mid \text{rest}\right) \propto \prod_i \mathcal{N}(x_i \mid x_0, R_x^2) \cdot \text{Inv-Gamma}(\alpha_R, \beta_R),$$

which is:

$$R_x^2 \sim \text{Inv-Gamma}\left( \alpha_R + \frac{N}{2}, \beta_R + \frac{1}{2} \sum_i (x_i - x_0)^2 \right).$$

### *4.4.4 Hamiltonian Monte Carlo*

Originally developed within the context of lattice quantum chromodynamics and then called **hybrid MC** (Duane et al., 1987), Hamiltonian Monte Carlo (HMC) exploits gradient information and Hamiltonian dynamics to efficiently sample from the posterior and obtain low auto-correlation between the samples and in parameter space of high dimensionality. The method was made more accessible to a wider audience by an influential paper by Neal (2011). The fundamental idea behind HMC is to use the target distribution as the kinetic energy term in a Hamiltonian, while introducing a suitable potential energy that is used as a device to facilitate the exploration of the parameter space of interest by following trajectories of constant energy along the Hamiltonian dynamic.

To introduce HMC, in this section we adopt a notation that brings forth the

 

analogy with Hamiltonian dynamics. We will use $\boldsymbol{q} \in \mathbb{R}^d$ to denoted the vector of the variables of interest (previously denoted $\boldsymbol{\theta}$), which in a Hamiltonian context correspond to generalised positions for the system; $\boldsymbol{p} \in \mathbb{R}^d$ is a set of auxiliary variables, corresponding to generalised momenta, which will be discarded at the end and that are used for the purpose of constructing a potential energy term. Doubling the number of variables might look at first as an unnecessary complication but it actually improves the scaling of the MCMC with the dimensionality $d$, from $\mathcal{O}\left(d^2\right)$ for random walk MCMC to $\mathcal{O}\left(d^{5/4}\right)$ for HMC (Hoffman and Gelman, 2014).

Let us briefly recall some notions from statistical mechanics. For a physical system in state $\boldsymbol{X} = \{\boldsymbol{q}, \boldsymbol{p}\}$ with energy $E(\boldsymbol{X})$, the **canonical distribution** over states has probability density

$$\Pr(\boldsymbol{X}) = \frac{1}{Z} \exp\left\{-\frac{E(\boldsymbol{X})}{T}\right\}, \tag{4.77}$$

where $Z$ is a normalisation constant and $T$ is the temperature. If we are interested in $\Pr(\boldsymbol{X})$, we can obtain it from the canonical distribution by setting $T \equiv 1$.

The **Hamiltonian** is the energy function in phase space:

$$\Pr\left(\boldsymbol{q}, \boldsymbol{p}\right) = \frac{1}{Z} \exp\left\{-\frac{H\left(\boldsymbol{q}, \boldsymbol{p}\right)}{T}\right\}; \tag{4.78}$$

the energy, as expressed by the Hamiltonian, is given by the sum of kinetic and potential energies:

$$\begin{aligned} H &= U\left(\boldsymbol{q}\right) + K\left(\boldsymbol{p}\right) \\ &= \text{potential energy} + \text{kinetic energy} \end{aligned} \tag{4.79}$$

As the parameters of interest, $\boldsymbol{\theta}$, are denoted by $\boldsymbol{q}$, we define the "potential energy" as

$$U\left(\boldsymbol{q}\right) = -\ln \tilde{\pi}(\boldsymbol{q}), \tag{4.80}$$

where $\tilde{\pi}(\boldsymbol{q})$ is the unnormalized posterior. We can ignore the normalizing constant of the posterior, as this has no influence on the shape of its density as a function of the parameters of interest (it merely changes the potential energy by a constant factor), and, for the same reason, set $T = 1$.

Finally, we need to choose a kinetic energy. Since the $\boldsymbol{p}$ are auxiliary variables, this function can be chosen arbitrarily—that is for convenience and to maximise the efficiency of the exploration. A simple choice is that of a **Euclidean–Gaussian mass matrix**, $\boldsymbol{M}$:

$$\frac{1}{2} \boldsymbol{p}^\top \boldsymbol{M}^{-1} \boldsymbol{p}. \tag{4.81}$$

When the mass matrix is diagonal, this simplifies to the kinetic energy of $N$ 'particles', each of mass $m_i$:

$$K(\boldsymbol{p}) = \frac{1}{2}\sum_{i=1}^{d}\frac{p_i^2}{m_i}, \quad m_i > 0 \quad \text{for } i = 1,\ldots,d. \tag{4.82}$$

So far, we have doubled the dimensionality of parameter space by introducing the auxiliary variable $\boldsymbol{p}$ and expressed the target as a canonical distribution characterized by the energy $H(\boldsymbol{q},\boldsymbol{p})$. If we can sample from the joint distribution

$$\begin{aligned}\Pr(\boldsymbol{q},\boldsymbol{p}) &\propto e^{-\frac{1}{2}U(\boldsymbol{q})}e^{-\frac{1}{2}K(\boldsymbol{p})} \\ &\propto \tilde{\pi}(\boldsymbol{q})e^{-\frac{1}{2}K(\boldsymbol{p})},\end{aligned} \tag{4.83}$$

we can marginalise over $\boldsymbol{p}$ and thus obtain the target density, namely the posterior:

$$\int \Pr(\boldsymbol{q},\boldsymbol{p})\,\mathrm{d}\boldsymbol{p} \propto \tilde{\pi}(\boldsymbol{q}). \tag{4.84}$$

The reason why this procedure performs particularly well is that we have set it up as a Hamiltonian system, whose dynamics conserves the energy. If we follow the solution to Hamilton's equations of motion

$$\frac{\mathrm{d}q_i}{\mathrm{d}t} = \frac{\partial H}{\partial p_i} \quad \text{and} \quad \frac{\mathrm{d}p_i}{\mathrm{d}t} = -\frac{\partial H}{\partial q_i} \quad (i = 1,\ldots,d), \tag{4.85}$$

we have

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \sum_{i=1}^{d}\left[\frac{\mathrm{d}q_i}{\mathrm{d}t}\frac{\partial H}{\partial q_i} + \frac{\mathrm{d}p_i}{\mathrm{d}t}\frac{\partial H}{\partial p_i}\right] = 0. \tag{4.86}$$

Thus, trajectories of $H(\boldsymbol{q},\boldsymbol{p}) = \text{const.}$ follow constant energy contours and therefore constant probability density contour of the canonical distribution. Hence, if we start from $(\boldsymbol{q},\boldsymbol{p})$ and move to $(\boldsymbol{q}',\boldsymbol{p}')$ using Hamilton's equation exactly (i.e. without numerical error) and use $(\boldsymbol{q}',\boldsymbol{p}')$ as a deterministic proposal in a MH step, this will be accepted with unit probability. In reality, the equations of motion are not solved exactly and one still needs an accept/reject step at the end to account for any numerical error, as we shall see. In this way, the proposed steps are going to travel a long way in phase space and hence reach distant values of $\boldsymbol{q}'$ from $\boldsymbol{q}$. A further important property of the Hamiltonian evolution is that it is **volume-preserving** (i.e., symplectic), which, for our purposes, means that the volume of phase space is the same at the start and at the end of the trajectory. Therefore, no Jacobian is required to adjust for any changes in the volume element between the start and the end.



 

**Example 4.2: HMC for a harmonic oscillator**

We consider a simple example in one dimension, $d = 1$, where the posterior is a standard normal, $\mathcal{N}(0,1)$. We have for the kinetic and potential energies, choosing for the mass value $m = 1$:

$$U(q) = \frac{q^2}{2}$$
$$K(p) = \frac{p^2}{2}. \tag{4.87}$$

Hamilton's equations correspond to those for a harmonic oscillator:

$$\frac{\mathrm{d}q}{\mathrm{d}t} = p, \quad \frac{\mathrm{d}p}{\mathrm{d}t} = -q, \tag{4.88}$$

with analytical solution:

$$q(t) = r\cos(a+t)$$
$$p(t) = -r\sin(a+t) \tag{4.89}$$

with $a$, $r$ constants that are set by the initial conditions. In phase space, trajectories look like circles: by following such constant energy lines, we can move very long distances in $q$ parameter space.

*HMC algorithm*

The algorithmic implementation of the ideas we laid out above consists of the following three steps, after initialization of a starting value of $\boldsymbol{q}_0$, drawn as usual from the prior:

1 Draw a new value for the momentum, $\boldsymbol{p}_0$, from the (un-normalized) distribution given by:

$$\boldsymbol{p}_0 \sim \exp\left[-K(\boldsymbol{p})\right], \tag{4.90}$$

independently of $\boldsymbol{q}$. This distribution is proportional to a Gaussian in the Euclidean mass matrix case;

2 use the Hamiltonian dynamic of Eq. (4.85) to deterministically propose a new candidate point, evolved from $\{\boldsymbol{q}_0, \boldsymbol{p}_0\}$ for a time $t$ according to:

$$\{\boldsymbol{q}_0, \boldsymbol{p}_0\} \to \{\boldsymbol{q}_t, \boldsymbol{p}_t\} = H_t\left(\boldsymbol{q}_0, \boldsymbol{p}_0\right), \tag{4.91}$$

where $H_t\left(\boldsymbol{q}_0, \boldsymbol{p}_0\right)$ denotes the (potentially approximate) solution of Hamilton's equations evolved from $\{\boldsymbol{q}_0, \boldsymbol{p}_0\}$ after a time $t$. Formally, at this point the

value of $\boldsymbol{p}_t$ is reversed (i.e., $\boldsymbol{p}_t \to -\boldsymbol{p}_t$) to guarantee detailed balance (see below); in practice, this can be omitted as the potential energy is a quadratic from in $\boldsymbol{p}_t$ and therefore the sign change makes no difference.

3 accept this proposal with acceptance probability:

$$\rho(\{\boldsymbol{q}_t, \boldsymbol{p}_t\}, \{\boldsymbol{q}_0, \boldsymbol{p}_0\}) = \min\left(1, \exp\left[-H(\boldsymbol{q}_t, -\boldsymbol{p}_t) + H(\boldsymbol{q}_0, \boldsymbol{p}_0)\right]\right), \tag{4.92}$$

where the argument of the exponential is the difference in the energy between the final and initial state. For a perfect (or analytical) integration, the argument of the exponential equals 0 since energy would be exactly conserved. This accept/reject step is necessary to ensure detailed balance. Loop back to step 1.

Importantly, when looping back to step 1, a new momentum is drawn irrespective of whether the proposed step has been accepted or not. This scrambling of momentum ensures that new regions of $\boldsymbol{q}$ space become accessible, and that the chain performs a random walk across different energy levels. The process is illustrated in a cartoon in Fig. 4.13.

It is clear that the chain is Markovian, since the next proposed point only depends on the current one. In order to show that it correctly samples from the target, Eq. (4.83), we need to demonstrate detailed balance, Eq. (4.60). To see that, consider HMC as a MH method combining a deterministic evolution (the Hamiltonian phase) with a MH accept/reject step at the end, given by Eq. (4.92). Since the Hamiltonian evolution is deterministic, the transition probability coincides with the acceptance probability, Eq. (4.92). The detailed balance equation, (4.60), is thus:

$$\begin{aligned}
&\pi(\boldsymbol{q}_0)\exp\big(-K(\boldsymbol{p})\big)\min\left(1, \exp\left[-H(\boldsymbol{q}_t, -\boldsymbol{p}_t) + H(\boldsymbol{q}_0, \boldsymbol{p}_0)\right]\right) && (4.93)\\
=&\min\left(\exp\big(-H(\boldsymbol{q}_0, \boldsymbol{p}_0)\big), \exp\left[-H(\boldsymbol{q}_t, -\boldsymbol{p}_t)\right]\right) && (4.94)\\
=&\min\left(1, \exp\left[-H(\boldsymbol{q}_0, \boldsymbol{p}_0) + H(\boldsymbol{q}_t, -\boldsymbol{p}_t)\right]\right)\exp\big(-H(\boldsymbol{q}_t, -\boldsymbol{p}_t)\big) && (4.95)\\
=&\min\left(1, \exp\left[-H(\boldsymbol{q}_0, \boldsymbol{p}_0) + H(\boldsymbol{q}_t, -\boldsymbol{p}_t)\right]\right)\pi(\boldsymbol{q}_t)\exp\big(-K(\boldsymbol{p}_t)\big). && (4.96)
\end{aligned}$$

In the last line, the first term is the acceptance probability for the time evolution returning back to the starting point, followed by momentum sign flip: $H(\boldsymbol{q}_t, -\boldsymbol{p}_t) \to H(\boldsymbol{q}_0, -\boldsymbol{p}_0) \to H(\boldsymbol{q}_0, \boldsymbol{p}_0)$, hence detailed balance holds. For proofs of ergodicity and invariance of the resulting chain, see Neal (2011).

The numerical integration of the Hamiltonian dynamics introduces numerical errors, which are minimized by using **symplectic integrators**, such as the 'leapfrog' method, which prevent the solution from drifting off from constant-energy trajectories (see fig. 4.14). The symplectic integrator requires choosing a time step, $\epsilon$, and a number of steps, $L$; because of the finite integration accuracy, after $L$ steps the state of the system will have accumulated an energy error

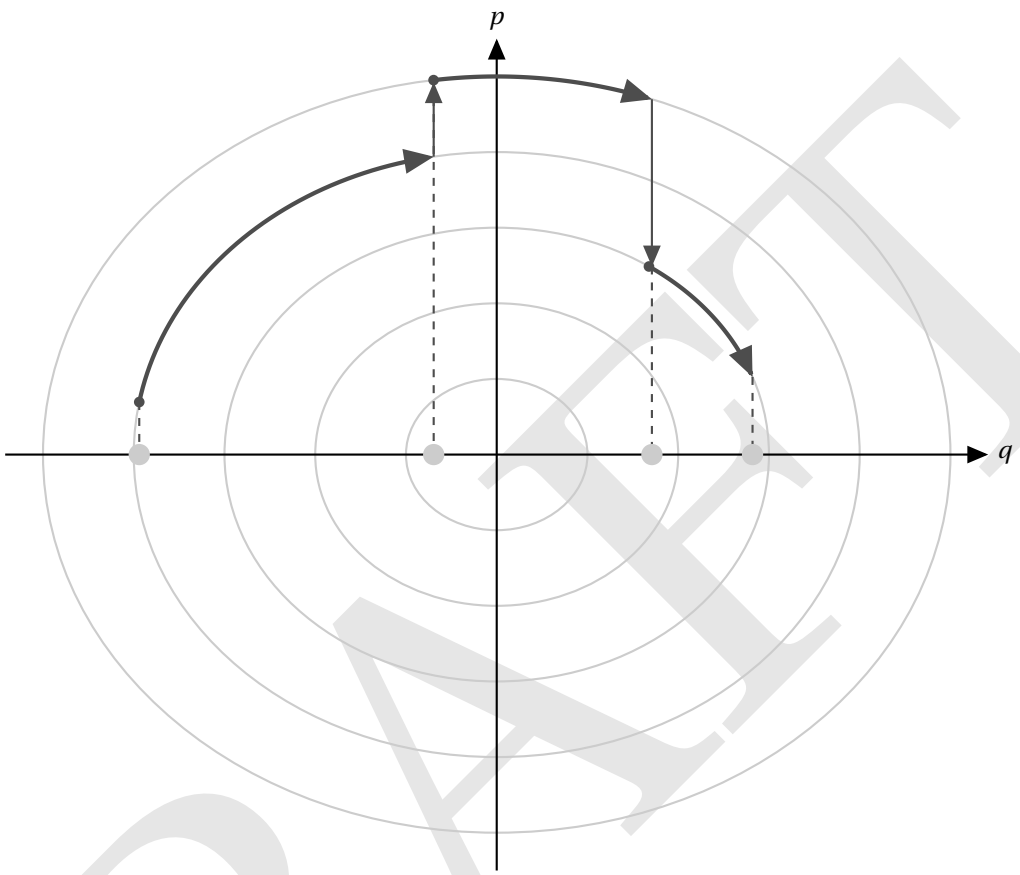


Figure 4.13 Illustration of HMC in phase space. The Hamiltonian evolution is represented by the thick, dark lines along constant energy surfaces (grey ellipses). At the end of each deterministic trajectory, a new momentum is randomly sampled and the chain jumps to a different energy level (vertical arrows). The momentum coordinates are discarded to give samples from the target for the parameters of interest (light grey dots).

$\Delta E$. The leapfrog method proceeds in these three steps for each $i = 1, \ldots, d$ (and assuming a diagonal mass matrix):

$$\begin{aligned}
&\text{(a)} \quad p_i\left(t+\frac{\epsilon}{2}\right) = p_i(t) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_i}\left(\boldsymbol{q}(t)\right), \\
&\text{(b)} \quad q_i(t+\epsilon) = q_i(t) + \epsilon\frac{p_i}{m_i}\left(t+\frac{\epsilon}{2}\right), \\
&\text{(c)} \quad p_i(t+\epsilon) = p_i\left(t+\frac{\epsilon}{2}\right) - \frac{\epsilon}{2}\frac{\partial U}{\partial q_i}\left(\boldsymbol{q}(t+\epsilon)\right).
\end{aligned} \tag{4.97}$$

This makes it clear that gradient of the potential energy, and therefore of the log-

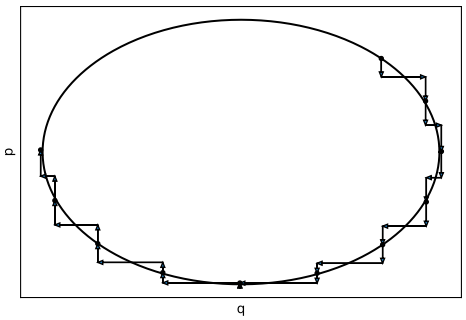


Figure 4.14 Numerical integration in phase space using the leapfrog method. Each arrow represents one of the three-steps updates to the position and momentum variables; the end of a leapfrog integration is shown by the dots. The smooth trajectory is the exact solution.

likelihood, is used to guide the HMC exploration, in the first and third step: the direction of the momentum is adjusted to point towards a region of large target density: when the gradient of the log-likelihood changes sign, the negative sign in front of the momentum update ensures that the next step is directed back towards the peak.

The efficiency of the HMC sampler depends on the choice of a few hyperparameters, which need to be tuned to the specific problem:

1 The kinetic energy mass matrix, $\boldsymbol{M}^{-1}$. The Hamiltonian structure of phase space means that a change of coordinate in momentum space which diagonalizes the mass matrix, $\boldsymbol{p} \rightarrow \boldsymbol{A}\boldsymbol{p}$, where $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{M}^{-1}$, induces the opposite transformation in coordinate space, i.e., $\boldsymbol{q} \rightarrow \boldsymbol{B}\boldsymbol{q}$, where $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{M}$. As a consequence, the optimal choice for the mass matrix is one that aligns well with the covariance of the target parameters, $\boldsymbol{q}$, as this creates energy levels that are more uniformly spaced and hence easier to explore. This can be estimated from the covariance of the target parameters obtained from a preliminary run. Betancourt (2013) also discusses a generalisation of this idea to to Riemannian metrics, where the coordinate transformation and hence the mass matrix is $\boldsymbol{q}$-dependent.
2 The the leapfrog step size, $\epsilon$. If $\epsilon$ is chosen too small, then the accuracy of the integration is good but a large amount of time is lost in integrating the equations of motion; if instead $\epsilon$ is too large, large inaccuracies in the Hamiltonian evolution creep in and this results in a small acceptance rate in the acceptance step at the end.
3 The leapfrog number of steps, $L$: the goal is to go 'as far as possible' in phase space; this can be achieved by adapting $L$ as the exploration progresses. In

 

order to decide when to stop integrating, a useful termination criterion is implemented in the so-called **NUTS** algorithm (No U-Turns Sampler). The idea is that the integrator has taken a sufficient number of steps when the trajectory in phase space starts "doubling back onto itself", and to head back towards its origin — at which point it is optimal to stop integrating (Hoffman and Gelman, 2014).

A very well established and maintained HMC package is `Stan`[11], which implements HMC with NUTS, as well as variational inference, differentiable probability functions, differentiable algebra and much more; it interfaces with R, Python (PyStan) and MATLAB. A recent development in Microcanonical Hamiltonian Monte Carlo (MCHMC) (Robnik et al., 2022), designed so that the target posterior is obtained as the marginal of the uniform distribution on the constant-energy-surface over the momentum variables. This new method follows a constant-energy trajectory in phase space, i.e., $\Pr(\boldsymbol{q},\boldsymbol{p}) \propto \delta(H(\boldsymbol{q},\boldsymbol{p}) - E)$, with additional 'momentum bounces' (i.e. momentum-conserving directionality changes) added to ensure ergodicity of the chain. This approach has been shown to outperform HMC in several benchmark problems.

### *4.4.5 Importance Sampling*

This technique is often useful for post-processing chains to adjust the posterior e.g. when adding new data to the likelihood with which the chains had been originally obtained, as well as in variational inference.

In the situation where one wishes to sample from a target density $\pi(\boldsymbol{\theta})$, but has samples from some other density, $q(\boldsymbol{\theta})$—for instance because $q$ is analytically tractable, or a fast approximation, or simpler to sample from than $\pi$. Then the expectation under $\pi$ for a function $h$ is

$$\begin{aligned}\mathbb{E}\left[h(\boldsymbol{\theta})\right]_\pi &= \int h(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= \int h(\boldsymbol{\theta})\,\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\,q(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \\ &= \mathbb{E}\left[\frac{\pi(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\,h(\boldsymbol{\theta})\right]_q . \end{aligned} \tag{4.98}$$

We thus obtain the expectation value under $\pi$ from the expectation value under $q$ by re-weighting the function $h$ by the ratio of densities $\pi(\boldsymbol{\theta})/q(\boldsymbol{\theta})$.

A sampling estimate of the expectation value of the function $h$ from an MCMC

[11] Available from: `mc-stan.org`.

 

chain $(\boldsymbol{\theta}_t)_{t=1,\ldots,n}$ is therefore given by

$$\mathbb{E}[h(\boldsymbol{\theta})] \approx \frac{1}{n} \frac{\sum_{t=1}^{n} w_t h(\boldsymbol{\theta}_t)}{\sum_{t=1}^{n} w_t}, \tag{4.99}$$

with **importance sampling** weights $w_t$ given by:

$$w_t = \frac{\pi(\boldsymbol{\theta}_t)}{q(\boldsymbol{\theta}_t)}, \quad (t = 1, \ldots, n). \tag{4.100}$$

The difficulty with this approach is that if $\pi$ and $q$ have different support this density ratio might be small and/or very noisy, leading to large noise in the importance weights and hence potential inaccuracies in the reweighted chain.

## 4.5 Running MCMC

We now turn to practical considerations concerning the implementation of MCMCs to real-world problems. An important question is when to stop the chain, i.e., how to check whether the target posterior has been thoroughly explored.

*Burn-in* A more benign issue concerns the initial phase of the chain, the so-called **burn-in** (or warm-up) of the chain[12]. When the chain is started at $t = 0$ from a sample drawn from the prior, it is typically in a region of parameter space with a very small posterior density, and hence it rapidly diffuses towards regions of appreciable density, when it then starts oscillating around the mode of the posterior. In this phase, the chain is sampling a region that has a very small probability of being reached – it is only being seen because the chain happened to start from there. If the chain was to be run for a very long time, eventually the initial samples of the chain would be representative of the target, but this is never happens because this would require an extremely large number of steps. Therefore, samples seen in the initial, so-called burn-in phase, are not representative of the target density, i.e., their relative multiplicity is much larger than it ought to be.

The burn-in phase can be assessed by looking at the values of the negative log-posterior[13] as a function of chain step, $t$. During the burn-in phase, the negative log-posterior drops steeply as the chain moves towards the mode of the target distribution, where it then stabilizes (see Fig. 4.15). If one runs multiple chains and starts them from different points in parameter space (a practice that

[12] The term 'burn-in' might have originated from the electronic industry's practice to subject its components to testing before shipping them to customers in order to discard those that fail early in their lifetime; or it could have to do with the testing of aeroplanes' breaks.

[13] More precisely, one considers the un-normalized negative log-posterior values, which are identical to the negative log-likelihood (up to a constant) for the case of uniform priors.

 

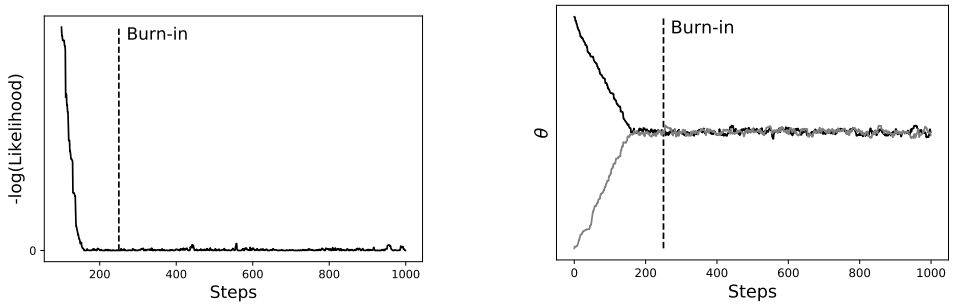


Figure 4.15 Left: The burn-in phase can be assessed by looking at the evolution of the negative unnormalized log-posterior as a function of the number of steps in the chain. Right: trace plot for one dimension of the parameter vector; two chains (dark and light grey), starting from different regions in parameter space, converge to the same location. In both plots, burn-in samples to the left of the vertical, dashed line need to be discarded.

is is always recommended, among other reasons to check for convergence to the same mode), burn-in can be identified from each chain's **trace plot** (i.e., chain location as a function of step number) as the period during which the various chains are not in the same location in parameter space.

The standard approach is to remove samples during the burn-in phase, i.e. the first $M$ samples in a chain of length $N$. Statistical texts often suggest using $M = n/2$, i.e., discarding the first half of the samples; out of experience, though, $M = 0.1n$ is often sufficient.

*Convergence* Let us now come to some strategies to assess **convergence** in practice. One of the most-often used criteria is the **Gelman–Rubin number** (Gelman and Rubin, 1992), which compares the variability within each chain to the variability between chains. The Gelman-Rubin diagnostic is based on the idea that if all chains have converged to the same distribution, the variability within each chain (intra-chain variability) should be comparable to the variability between the chains (inter-chain variability).

Let $\pi(\boldsymbol{\theta})$ be the target distribution, which is assumed to be univariate. Consider one parameter coordinate at the time, $\theta^{(k)}$ $(k = 1, \ldots, d)$, and let the target mean in that coordinate be $\mu^{(k)}$, and the variance $\sigma^{(k)^2}$. Run $m$ chains for $n$ steps each (after burn-in), each with sample mean $\overline{\theta}_i^{(k)}$, $(i = 1, \ldots, m)$. Let $\overline{\theta}^{(k)}$ be the sample mean for coordinate $k$ across all $m$ chains and $W^{(k)}$ the average sample

 

variance (the 'intra-chain variance', where $W$ stands for 'within chain'):

$$\overline{\theta}^{(k)} = \frac{1}{m}\sum_{i=1}^{m}\overline{\theta}_i^{(k)}, , \tag{4.101}$$

$$W^{(k)} = \frac{1}{m}\sum_{i=1}^{m} s_i^{(k)^2}, \tag{4.102}$$

$$\text{where} \quad s_i^{(k)^2} = \frac{1}{n-1}\sum_{t=1}^{n}\left(\theta_{i,t}^{(k)} - \overline{\theta}_i^{(k)}\right)^2 \tag{4.103}$$

is the the sample variance for chain $i$, and $\theta_{i,t}^{(k)}$ denotes step $t$ for chain $i$ and parameter coordinate $k$. Then define

$$B^{(k)} = n\sum_{i=1}^{m}\frac{\left(\overline{\theta}_i^{(k)} - \overline{\theta}^{(k)}\right)^2}{m-1}, \tag{4.104}$$

which is the variance between the chains' sample means (the 'inter-chain variance'). Notice that the factor of $n$ on the RHS, which corrects for the reduction in variance ($\propto 1/n$) induced by the fact that each chain's mean is computed by averaging over $n$ samples. This ensures that $B^{(k)}$ reflects the actual variance of the parameter distribution, rather than just the variance of the chain means.

The total variance of the target, $V^{(k)}$, is estimated as the sum of the inter-chain and the intra-chain variances, each with a sample-size dependent weight:

$$\hat{V}^{(k)} = \frac{n-1}{n}W^{(k)} + \frac{B^{(k)}}{n}, \tag{4.105}$$

where the $(n-1)/n$ weight adjusts for the $1/(1-n)$ term of each chain's variance estimator, and the $1/n$ weight adjusts for the factor of $n$ in the definition of $B$. One therefore monitors the so-called **scale reduction factor**

$$\hat{R}^{(k)} = \sqrt{\frac{\hat{V}^{(k)}}{W^{(k)}}} \cong 1 + \frac{1}{2}\left(\frac{B^{(k)}}{nW^{(k)}}\right), \tag{4.106}$$

where the approximate equality holds for large $n$. Because for finite $n$, $\hat{V}^{(k)}$ overestimates the total variance while $W^{(k)^2}$ underestimates it, it follows that $\hat{R}^{(k)} \to 1$ from above. Therefore, monitoring $\hat{R}^{(k)}$ as it approaches 1 from above leads to an assessment of the chains' convergence to the target as a function of $n$.

The heuristic convergence diagnostics proposed by (Gelman and Rubin, 1992) is as follows:

1 run $m$ chains for $2n$ steps each;
2 discard the first $n$ steps in each chain (burn-in phase);
3 compute $\hat{R} = \max_{k=1,\ldots,d}\hat{R}^{(k)}$;



 

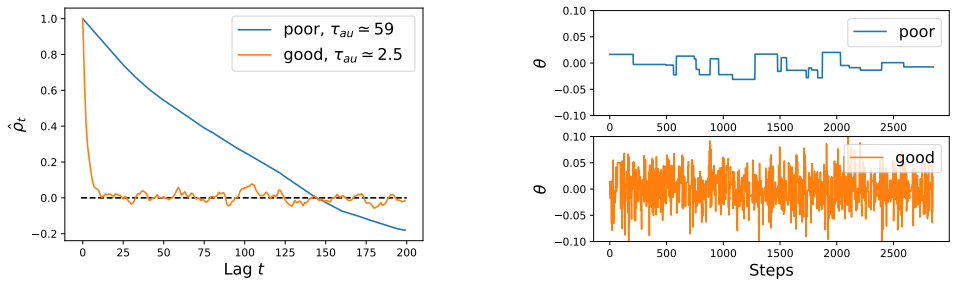


Figure 4.16 Left: Comparison of the estimated autocorrelation, $\hat{\rho}_t$, as a function of lag $t$ for a chain exhibiting short correlation length (orange) and one with large correlation length (blue). Right: Trace plot for each chain. Large autocorrelations correspond to poor mixing and a characteristic step-like appearance in the trace plot.

4 stop the chains when $\hat{R} < \delta$, with the threshold value (chosen heuristically) $\delta = 1.1$.

*Autocorrelation and effective sample size* Samples obtained via MCMC are typically correlated for small lag (i.e., distance in the chain). This means that the effective sample size (i.e., the number of equivalent uncorrelated samples from the target) is smaller than the number of accepted samples. We can estimate the autocorrelation at lag $t$, $\hat{\rho}_t^{(k)}$, as:

$$\hat{\rho}_t^{(k)} = \frac{\sum_{i=1}^{n-t}\left(\theta_i^{(k)} - \overline{\theta}^{(k)}\right)\left(\theta_{i+t}^{(k)} - \overline{\theta}^{(k)}\right)}{\sum_{i=1}^{n}\left(\theta_i^{(k)} - \overline{\theta}^{(k)}\right)^2}. \tag{4.107}$$

From the theory of Markov Chains, one expects that asymptotically the autocorrelation takes an exponential form:

$$\rho_t^{(k)} \propto \exp\left\{-\frac{t}{\tau_{\text{ac}}^{(k)}}\right\}, \tag{4.108}$$

where $\tau_{\text{ac}}^{(k)}$ is the characteristic timescale after which autocorrelations decay. The value of the autocorrelation scale depends on how the chain is constructed: in general, one should aim to obtain a small value of $\tau_{\text{ac}}^{(k)}$, as it leads to faster independence between samples (see fig. 4.16 for an example).

*Thinning and mixing* If one wishes to obtain approximately independent samples from the target, it is helpful to keep only one out of every $\max_k \tau_{\text{ac}}^{(k)}$ samples

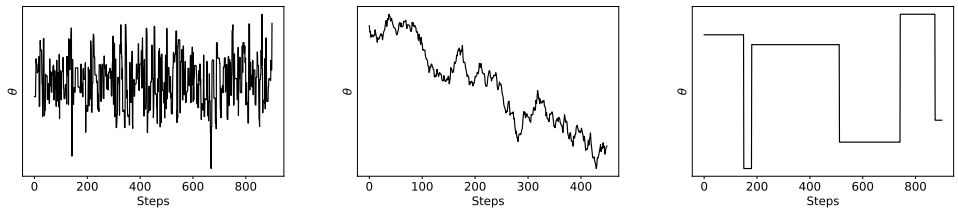


Figure 4.17 Illustration of good and poor mixing in MCMC. Left panel: a well-mixed chain looks like a 'hairy caterpillar'; central panel: too small a step size leads to diffusion, and slow convergence; right panel: a proposal density poorly aligned with the target leads to the chain being stuck (plateau regions), and hence slow convergence.

in the chain —a technique called **thinning of the chain**. Thinning reduces autocorrelation among sampled points, and indeed one expects that above the correlation length, samples become nearly uncorrelated. However, MacEachern and Berliner (1994) advise against thinning, for the variance of the thinned chain, with its smaller number of samples, is generally larger than that of the original one; hence, MC estimates of the thinned chain tend to be more noisy and less precise.

A chain with small autocorrelation scale and that explores the target efficiently (i.e., without diffusing nor getting stuck) is called 'well mixed'. The trace plot of such an ideal chain looks like a 'hairy caterpillar', an example of which is shown in Fig. 4.17. Monitoring the appearance of the trace plots helps to diagnose a poor choice of proposal distribution (central and right panels in the figure).

*Equal weight samples* If **equal weights samples** are required (i.e., samples with the same multiplicity), e.g. for visualization purposes in scatter plots (where the multiplicity of each sample does not show), these can be obtained from the chain by renormalizing the samples' multiplicities into weights $w_t = m_t / \max_t m_t \in (0,1]$, $t = 1,\dots,n$. Then, equal weight samples are obtained by drawing $\gamma \sim \text{Uniform}[0,1]$: if $w_t > \gamma$, sample $\theta_t$ is retained, otherwise it is discarded and one moves $t \to t+1$. This produces a shorter chain, in which all retained samples are equally weighted.

It is always recommendable to inspect the raw samples (before and after burn-in, thinning and single-sample reweighting) before producing more elaborate plots of the posterior (e.g. by using modern packages that employ kernel density smoothing to prettify the resulting distributions).

*Recommendations when running MCMC* We conclude this section on MCMC with some heuristic recommendations when employing MCMC methods. To re-

assure yourself that the posterior samples obtained via MCMC are a fair representation of the target posterior, make sure to:

- Check that subsegments of the chain(s) give similar results as the chain(s) as a whole; different chains behaving differently, and converging to different regions of parameter space is usually a sign of under-explored multi-modality;
- monitor the acceptance rate, aiming for the optimal $\mathscr{R} \sim 0.24$; monitor autocorrelation, and tweak the sampler to achieve small autocorrelation length.
- check that the chain is well-mixed' — see Fig. 4.17 for an illustration;
- try running multiple chains from different starting points, and check that they all converge to the same mode; if multiple modes are present, make sure that each chain mixes well across modes.
- use different samplers and check if the results are compatible;
- pay attention to the accuracy of sampling into the tails of the target. This typically requires special attention if you want to achieve high accuracy above 2 or 3 $\sigma$ into the tails. Investigate unexpected posterior shape or glitches, trying to determine whether those are real features of the posterior, numerical artefacts or boundary effects of the parameter space being cut off by the prior. It is good practice to build an intuition for the likelihood before starting any complex MCMC sampling in high dimensions. This can be achieved by plotting the shape of the likelihood function one dimension at a time (with all other parameters fixed).

## 4.6 Simulation-based Inference

In many problems, the likelihood function is either unavailable in closed form or computationally expensive to evaluate. This is particularly common in complex models, such as those involving $N$-body simulations, nonlinear systems or observational effects. In such cases, traditional Bayesian inference methods, which rely on explicit evaluation of the likelihood, become infeasible, or the approximations needed to write down an explicit likelihood may introduce systematic bias in the inference (e.g., Karchev et al. (2022)).

### *4.6.1 Approximate Bayesian Computation*

Originally introduced in the context of population genetics in the late 1990s, Approximate Bayesian Computation (ABC) is a general-purpose method that aims at circumventing the need to explicitly evaluate the likelihood in situations where it may be unavailable, but we have access to simulators that can generate synthetic data given parameter inputs. Instead of computing the likelihood,

 

ABC works by comparing simulated data with observed data using a measure of similarity (e.g., summary statistics or distance metrics). This allows Bayesian inference without explicit likelihood calculation.

ABC provides a way to approximate the posterior distribution of model parameters given observed data, through the following steps:

1 Sampling of parameters: draw $\boldsymbol{\theta} \sim \Pr(\boldsymbol{\theta})$ from their prior.
2 Simulate data: generate synthetic data, $\boldsymbol{d}$, from the model using the sampled parameters; if available, compute the chosen (lower dimensional) summary $\boldsymbol{t}(\boldsymbol{d})$.
3 Compare with the observed data, $\boldsymbol{d}_{\text{obs}}$, using a similarity measure (e.g., Euclidean distance between summary statistics): if $\|\boldsymbol{t}(\boldsymbol{d}) - \boldsymbol{t}(\boldsymbol{d}_{\text{obs}})\| < \epsilon$, keep the sample, otherwise discard it.
4 Repeat from step 1, to construct the approximate posterior:

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{d}_{\text{obs}})_{\text{ABC}} \propto \Pr(\boldsymbol{\theta}) \mathbb{I}[|\boldsymbol{t}(\boldsymbol{d}) - \boldsymbol{t}(\boldsymbol{d}_{\text{obs}})\| < \epsilon],$$

where $\mathbb{I}$ is the indicator function. Through this procedure, ABC seeks to produce posteriors that, while not conditional on $\boldsymbol{d}_{\text{obs}}$, are averaged over data that are 'close' to $\boldsymbol{d}_{\text{obs}}$. ABC delivers, in principle, an exact posterior for $\epsilon \to 0$, because in that case $\boldsymbol{d} \to \boldsymbol{d}_{\text{obs}}$, but the procedure is wasteful for large-dimensional datasets and parameter spaces, since a great deal of simulation is expended on parameters that lead to data very dissimilar to the observations. The efficiency of ABC decreases as the dimension of the data and parameter space increases, and choosing a low-dimensional summary statistics is therefore critical, as well as the choice of the tolerance parameter $\epsilon$. Usually, iterative refinement of the proposal (i.e., restricting the prior in order to guide simulation towards $\boldsymbol{d}_{\text{obs}}$) is necessary to obtain sufficiently accurate posteriors. Finally, another drawback fo traditional ABC is that the whole procedure is targeted to the specific data realization $\boldsymbol{d}_{\text{obs}}$, and needs to be repeated for new data. More modern versions of ABC, which collectively go under the moniker of 'Simulation-Based Inference' (SBI) seek to improve on these limitations.

### *4.6.2 Neural SBI*

With the recent rise of neural network-based methods, it becomes possible to train a machine learning model as a surrogate for the **joint distribution** between data and parameters, $\Pr(\boldsymbol{\theta}, \boldsymbol{d})$. This is appealing in many ways: first, it treats data and parameters symmetrically; from the joint distribution, one can derive the marginals $\Pr(\boldsymbol{d})$ (the evidence) and $\Pr(\boldsymbol{\theta})$ (the parameters' prior), as well as the conditionals $\Pr(\boldsymbol{d} \mid \boldsymbol{\theta}_0)$ (sampling distribution for the data at a given parameters

| Method | What is learnt? | Characteristics |
|---|---|---|
| **NPE** | $\Pr(\boldsymbol{\theta} \mid \boldsymbol{d})$ | Direct posterior access, flexible |
| **NLE** | $\Pr(\boldsymbol{d} \mid \boldsymbol{\theta})$ | Likelihood explicit, reusable |
| **NRE** | $\Pr(\boldsymbol{d} \mid \boldsymbol{\theta}) / \Pr(\boldsymbol{d})$ | Simulation-efficient, scalable |

Table 4.1 *Comparison of Neural Posterior Estimation (NPE), Neural Likelihood Estimation (NLE), and Neural Ratio Estimation (NRE).*

value $\boldsymbol{\theta}_0$) and $\Pr(\boldsymbol{\theta} \mid \boldsymbol{d}_0)$ (posterior for the parameters for given data $\boldsymbol{d}_0$). Secondly, once the joint is learnt over a range of possible data $\boldsymbol{d}$, inference can be obtained by evaluating it at any data $\boldsymbol{d}_0$ one wishes to specify, without needing to retrain the model. This is called **amortization** – i.e., the computational effort spent training the model is paid off by the fact that the same model can then be deployed on any observed data set (within the domain of amortization). Apart from computational efficiency at the evaluation stage (for evaluation of a trained neural network is typically very fast), amortization also opens the door to sophisticated **calibration** of the resulting posteriors, for example by ensuring that they have exact Frequentist coverage (see below), and perform Bayesian checks.

Three main approaches exist to neural SBI, summarized in Table 4.1:

1 **neural posterior estimation (NPE)** aims at approximating the posterior via a neural conditional estimator such as normalizing flows, trained on simulated data model that performs inference on empirical data;
2 **neural likelihood estimation (NPE)** is similar, but targets the likelihood instead;
3 **neural ratio estimation (NRE)** approximates the likelihood-to-evidence ratio by training a classifier-based neural network using binary cross-entropy, thus turning the inference problem into a classification problem.

### *4.6.3 Truncated Marginal Neural Ratio Estimation (TMNRE)*

Here we focus particularly on a ratio estimation method called Truncated Marginal Ratio Estimation (TMNRE), introduced by Miller et al. (2021). The fundamental idea (often known as 'the likelihood ratio trick') is to learn the likelihood-to evidence ratio (Cranmer et al., 2016, 2020):

$$r(\boldsymbol{d};\boldsymbol{\theta}) = \frac{\Pr(\boldsymbol{d} \mid \boldsymbol{\theta})}{\Pr(\boldsymbol{d})} = \frac{\Pr(\boldsymbol{\theta} \mid \boldsymbol{d})}{\Pr(\boldsymbol{\theta})},$$

 

where the equality follows from Bayes' Theorem. In virtue of the second equality, once we have learnt $r$ we can immediately obtain the posterior by evaluating $r(\boldsymbol{d} = \boldsymbol{d}_{\text{obs}}; \boldsymbol{\theta})$ and multiplying by the prior density:

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{d}) = \Pr(\boldsymbol{\theta})\, r(\boldsymbol{d} = \boldsymbol{d}_{\text{obs}}; \boldsymbol{\theta}) .$$

The fundamental idea is to turn the problem of learning the ratio $r$ into a **supervised classification problem** – a task at which neural networks excel. The ratio can be written as:

$$r(\boldsymbol{d}; \boldsymbol{\theta}) = \frac{\Pr(\boldsymbol{d} \mid |\boldsymbol{\theta})}{\Pr(\boldsymbol{d})} \frac{\Pr(\boldsymbol{d})}{\Pr(\boldsymbol{d})} = \frac{\Pr(\boldsymbol{d}, \boldsymbol{\theta})}{\Pr(\boldsymbol{d}) \Pr(\boldsymbol{\theta})} .$$

If we can generate pairs of parameters-data, $\{\boldsymbol{\theta}_i, \boldsymbol{d}_i\}$, we can train a binary classifier to distinguish between pairs drawn from the **joint** distribution in the numerator, $\Pr(\boldsymbol{d}, \boldsymbol{\theta})$, vs scrambled pairs, $\{\boldsymbol{\theta}_i, \boldsymbol{d}_j\}$ drawn from the product of the marginals in the denominator.

This is a special case of a more general idea, namely, that of representing the target distribution as a conditional density in a model with an additional categorical (discrete) class label, $c = 1,2$. The target, $p_1(\boldsymbol{x})$ is then expressed as $\Pr(\boldsymbol{x} \mid c = 1)$, while a second, tractable distribution $p_2(\boldsymbol{x}) = \Pr(\boldsymbol{x} \mid c = 2)$ is also introduced. Since the posterior probability for class 1 is (assuming equal class priors):

$$\Pr(c = 1 \mid \boldsymbol{x}) = \frac{\Pr(\boldsymbol{x} \mid c = 1)}{\Pr(\boldsymbol{x} \mid c = 1) + \Pr(\boldsymbol{x} \mid c = 2)} = \frac{p_1(\boldsymbol{x})}{p_1(\boldsymbol{x}) + p_2(\boldsymbol{x})} , \tag{4.109}$$

it follows that the target can be obtained from the class 1 posterior and the ratio $r(\boldsymbol{x}) = p_1(\boldsymbol{x}) / p_2(\boldsymbol{x})$ in terms of the tractable $p_2(\boldsymbol{x})$ as:

$$p_1(\boldsymbol{x}) = \Pr(c = 1 \mid \boldsymbol{x})\, p_2(\boldsymbol{x})(1 + r(\boldsymbol{x})) . \tag{4.110}$$

We wish to train a neural network, described by weights $\boldsymbol{\Phi}$, so that its output, $f_{\boldsymbol{\Phi}}$, is an approximation to the quantity $\ln r$. This can be obtained by minimizing the **binary cross-entropy loss function**, which measures the difference between the class probability as predicted by the neural network, $q_{\boldsymbol{\Phi}}$, and the true class (i.e, the ground truth labels from the training set). Minimizing this loss ensures that the neural network's predictions are as close as possible to the true labels:

$$\text{BCE} = -\left( \mathrm{E}\left[ \ln q_{\boldsymbol{\Phi}}(c = 1 \mid \boldsymbol{x}) \right]_{p_1(\boldsymbol{x})} + \mathrm{E}\left[ \ln q_{\boldsymbol{\Phi}}(c = 2 \mid \boldsymbol{x}) \right]_{p_2(\boldsymbol{x})} \right)$$

From Eq. (4.109), we can express the neural network's predicted class 1 probability as:

$$q_{\boldsymbol{\Phi}}(c = 1 \mid \boldsymbol{x}) = \frac{1}{1 + \hat{r}_{\boldsymbol{\Phi}}(\boldsymbol{x})} = \sigma(\ln \hat{r}_{\boldsymbol{\Phi}}) = \sigma(f_{\boldsymbol{\Phi}}) ,$$

where $\sigma$ is the logistic function, given by:

$$\sigma(x) = \frac{1}{1+\exp(-x)},$$

and similarly $q_{\boldsymbol{\Phi}}(c=2 \mid \boldsymbol{x}) = 1/(1+\hat{r}_{\boldsymbol{\Phi}}) = \sigma(-\ln(\hat{r}_{\boldsymbol{\Phi}})) = \sigma(-f_{\boldsymbol{\Phi}})$. Here, $\hat{r}_{\boldsymbol{\Phi}}$ is the neural network approximation for the true ratio, $r$. We are therefore led to defining the loss function:

$$L_{\text{BCE}}(\boldsymbol{\Phi}) = -\sum_{i\in\text{class 1}} \ln\big(\sigma(f_{\boldsymbol{\Phi}}(\boldsymbol{x}_i))\big) - \sum_{j\in\text{class 2}} \ln\big(\sigma(-f_{\boldsymbol{\Phi}}(\boldsymbol{x}_j))\big), \tag{4.111}$$

where the first term maximises the probability of predicting that training examples drawn from class 1 are indeed from that class, and the second term maximises the probability of correctly predicting examples from class 2.

For the specific case of NRE, we choose for $p_1$ examples from the joint distribution, $p_1 \to \Pr(\boldsymbol{\theta}, \boldsymbol{d})$, while $p_2$ represents examples drawn from each marginal. During training, we draw two pairs, $\{\boldsymbol{\theta}_i, \boldsymbol{d}_i\}$ $(i = 1,2)$ from the joint distribution (i.e., we simulate $\boldsymbol{d}_i$ from parameters $\boldsymbol{\theta}_i$), and we build two matching pairs $\{\boldsymbol{\theta}_1, \boldsymbol{d}_1\}$ and $\{\boldsymbol{\theta}_2, \boldsymbol{d}_2\}$ from class 1, and two mismatched pairs, $\{\boldsymbol{\theta}_1, \boldsymbol{d}_2\}$ and $\{\boldsymbol{\theta}_2, \boldsymbol{d}_1\}$, as examples from class 2. The loss function is then:

$$\begin{aligned} 2L_{\text{NRE}}(\boldsymbol{\Phi}) &= -\ln\big(\sigma(f_{\boldsymbol{\Phi}}(\boldsymbol{\theta}_1, \boldsymbol{d}_1))\big) - \ln\big(\sigma(f_{\boldsymbol{\Phi}}(\boldsymbol{\theta}_2, \boldsymbol{d}_2))\big) &(4.112)\\ &= -\ln\big(\sigma(-f_{\boldsymbol{\Phi}}(\boldsymbol{\theta}_2, \boldsymbol{d}_1))\big) - \ln\big(\sigma(-f_{\boldsymbol{\Phi}}(\boldsymbol{\theta}_1, \boldsymbol{d}_2))\big), &(4.113)\end{aligned}$$

which can then be optimized via stochastic gradient descent and backpropagation.

We now show that, asymptotically in the infinite data and network expressivity limit, optimizing this loss function leads to $f_{\boldsymbol{\Phi}} \to \ln r$. In this continuum limit, we can write the loss function as:

$$L(f_{\boldsymbol{\Phi}})_{\text{NRE}} \to -\int \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{d}\left[\Pr(\boldsymbol{\theta}, \boldsymbol{d}) \ln\sigma(f_{\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{d})) + \Pr(\boldsymbol{\theta})\Pr(\boldsymbol{d}) \ln\big(\sigma(-f_{\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{d}))\big)\right].$$

Setting the functional derivative of the loss function to zero to find its minimum gives:

$$\frac{\delta L(f_{\boldsymbol{\Phi}})}{\delta f_{\boldsymbol{\Phi}}} = -\Pr(\boldsymbol{\theta}, \boldsymbol{d})\,\sigma(-f_{\boldsymbol{\Phi}}) + \Pr(\boldsymbol{\theta})\Pr(\boldsymbol{d})\,\sigma(f_{\boldsymbol{\Phi}}) = 0.$$

Where we used:

$$\frac{d\sigma(x)}{dx} = \sigma(x)\big(1-\sigma(x)\big) = \sigma(x)\sigma(-x).$$

and

$$\frac{d\ln\sigma(f)}{df} = \frac{1}{\sigma(f)}\frac{d\sigma(f)}{df} = \frac{\sigma(f)\sigma(-f)}{\sigma(f)} = \sigma(-f).$$

We therefore obtain that the value of $\boldsymbol{\Phi}$ that minimizes the loss leads to:

$$\frac{\Pr(\boldsymbol{\theta}, \boldsymbol{d})}{\Pr(\boldsymbol{\theta})\Pr(\boldsymbol{d})} = \frac{\sigma(f_{\boldsymbol{\Phi}})}{\sigma(-f_{\boldsymbol{\Phi}})} = \exp\big(f_{\boldsymbol{\Phi}}\big) = r(\boldsymbol{\theta}, \boldsymbol{d}),$$

which implies:

$$f_{\boldsymbol{\Phi}}(\boldsymbol{\theta}, \boldsymbol{d}) \to \ln r(\boldsymbol{\theta}, \boldsymbol{d}).$$

A major improvement in NRE is to consider only some **parameters groups** at the time: i.e., we don't need to perform joint inference on the entirety of $\boldsymbol{\theta}$, but we can instead focus the ratio estimator only on a low-dimensional subgroup of interest (usually, 1- or 2-dimensional). Learning the parameter groups marginally involves using separate ratio estimators tailored to each group. This approach motivates an inference network structure that consists of a data pre-processor (head) and multiple tails, where each tail includes a parameter pre-processor and a ratio estimator. During each training step, the data is pre-processed once and provided as input to all tails. The loss is then evaluated independently for each parameter group, and the resulting losses are combined before performing the gradient descent step. Consequently, the head network is updated based on its performance across all parameter groups, requiring it to extract relevant information for all of them. This leads to **marginal inference** for each of the groups of interest. Consequently, at no point in the inference chain the full, high-dimensional posterior is learnt or evaluated.

Finally, a suitable scheme needs to be identified to guide the parameter sampling towards a region of parameter space that leads to 'similar' data as observed, in order to increase the efficiency of training. One possibility (inspired by Papamakarios and Murray (2018) and adopted in Miller et al. (2020)) is to carry out a **parameter truncation scheme** – a particular example of sequential training (where the output of previous training iterations is used to guide the next). A possibility is to use a truncation scheme in the prior range, that does not alter the shape of the prior distribution but simply limits its support, introducing only a uniform re-normalization. At each stage of training, the truncated (excluded) region –determined independently for each parameter group– is where the posterior density is considered negligible: the prior is restricted to a rectangular box that encloses the HPD region containing 99.99% of the probability mass from the current approximate posterior, evaluated for the target data. After each truncation, a new network is randomly initialized and trained on samples generated using the newly constrained priors[14]. The process is repeated until the next round of truncation shrinks the prior range by less than a pre-set factor (typically, a factor of 2). This scheme does require the choice of a reference

[14] A trivial adjustment in the ensuing estimated posterior density is required, see Karchev et al. (2022) for details.

data set, $\boldsymbol{d}_0$ (often taken to be the observed data, $\boldsymbol{d}_{\text{obs}}$), that is used during training as target of the truncation scheme. However, the resulting trained inference network remains **locally amortized** within a restricted prior box around $\boldsymbol{d}_0$.

### *4.6.4 Validation and Calibration of Amortized Posteriors*

As mentioned above, one major advantage of SBI is that, in many implementations, the posterior is (at least locally) amortized. This enables inference on a large number of simulated data sets, which in turns allows to validate and calibrate the long-run performance of the parameter estimation procedure.

One such verification tool is that of the **Bayesian P-P (Probability-Probability) plot**. Once the inference network has been trained, one generates test parameters from the prior, and simulates a realization of the data from the model: $\boldsymbol{\theta}_t \sim \Pr(\boldsymbol{\theta})$, then $\boldsymbol{\theta}_t \xrightarrow{\text{simulator}} \boldsymbol{d}_t$. From the test data, the HPD credible region of nominal (approximated) probability content $\alpha_*$, called $\Gamma_{\alpha_*}$ is obtained as $\Gamma_{\alpha_*} = \{\boldsymbol{\theta} : \alpha_* > \gamma(\boldsymbol{\theta}_t, \boldsymbol{d}_t)\}$, where $\gamma(\boldsymbol{\theta}_0, \boldsymbol{d}_0)$ is a function that associates with each parameter value $\boldsymbol{\theta}_0$ the HPD region of credibility $\gamma$ that has $\boldsymbol{\theta}_0$ on its boundary:

$$\gamma(\boldsymbol{\theta}_0, \boldsymbol{d}_0) = \int_{R(\boldsymbol{\theta}_0, \boldsymbol{d}_0)} q(\boldsymbol{\theta} \mid \boldsymbol{d}_0) \mathrm{d}\boldsymbol{\theta}\,,$$

where $R(\boldsymbol{\theta}_0, \boldsymbol{d}_0)$ is the parameter space region where the approximate posterior density (obtained from the trained neural network after sequential truncation), $q(\boldsymbol{\theta}|\boldsymbol{d}_0)$, is larger than $\boldsymbol{\theta}_0$:

$$R(\boldsymbol{\theta}_0, \boldsymbol{d}_0) = \{\boldsymbol{\theta} : q(\boldsymbol{\theta} \mid \boldsymbol{d}_0) > q(\boldsymbol{\theta}_0 \mid \boldsymbol{d}_0)\}\,.$$

Therefore $\gamma(\boldsymbol{\theta}_t, \boldsymbol{d}_t)$ is the HPD region that just encloses the true value of the parameters from which the data were generated at a credible level $\gamma$.

We want to check how often the true value of the parameters is inside $\gamma(\boldsymbol{\theta}_t, \boldsymbol{d}_t)$, averaged over many data realizations, in order to evaluate the coverage properties of the approximate Bayesian marginal posteriors obtained via NRE. Taking $\boldsymbol{\theta}_t$ to be the true parameters used to generate $\boldsymbol{d}_t$, we plot the frequency $F(\gamma)$ with which HPD regions, of different credibility $\gamma$ include (i.e., cover) $\boldsymbol{\theta}_t$. This is the **empirical coverage**. If the empirical coverage is larger than the credibility, i.e., if $F(\gamma) > \gamma$, our posterior approximation covers more frequently than its nominal credibility, and the posterior is thus conservative (wider than necessary). If instead the empirical coverage is smaller than the credibility (i.e., for $F(\gamma) < \gamma$), the approximate posterior undercovers. This can be assess by plotting $F(\gamma)$ vs $\gamma$, in what is called a 'P–P plot'. If we repeat this procedure for many values of $\boldsymbol{\theta}_t$, drawn from the prior, we obtain a Bayesian P–P plot, where the empirical coverage is averaged over representative parameter values from the prior. A diago-

nal Bayesian P–P plot is a necessary but not sufficient condition for establishing convergence of the approximate posterior to the true posterior because of the data-averaging: one can imagine the approximation conspiring to be conservative for some data and undercovering for others, in such a way that on average $F(\gamma) = \gamma$.

In a Frequentist context, we cannot perform the average over the parameters' prior, as this is not defined. Instead, we can define the **empirical conditional coverage**, i.e., conditioned on the value used to generate the data, as:

$$F_F(\gamma_* \mid \boldsymbol{\theta}_t) = \int \mathrm{d}\boldsymbol{d} \Pr(\boldsymbol{d} \mid \boldsymbol{\theta}_t) \left[ \int_0^{\gamma_*} \delta(y - \gamma(\boldsymbol{\theta}_t, \boldsymbol{d})) \mathrm{d}y \right] \tag{4.114}$$

We can then use the approximate Bayesian credibilities to construct confidence regions with exact coverage, as follows: we define the **required credibility** $\gamma_R(\boldsymbol{\theta}_t, \gamma_*)$ as the credibility of approximate posterior HPD regions that exhibit Frequentist coverage equal to $\gamma_*$. This is obtained by solving:

$$F_F(\gamma_R \mid \boldsymbol{\theta}_t) = \int_0^{\gamma_R} \Pr\left(\gamma' \mid \boldsymbol{\theta}_t\right) \mathrm{d}\gamma' = \gamma_* \tag{4.115}$$

i.e. the required credibility is the value of the credibility that ensures that the true value $\boldsymbol{\theta}_t$ is included in the posterior credible interval with empirical frequency $\gamma_*$.

From $\gamma_R$ we can construct Frequentist confidence intervals with exact coverage by proceeding as follows: For observed data $\boldsymbol{d}_{\mathrm{obs}}$, we select (on a low-dimensional grid; posteriors are only 1- or 2-d marginals anyhow) values of $\boldsymbol{\theta}_t$ and we build $\gamma(\boldsymbol{\theta}_t, \boldsymbol{d}_{\mathrm{obs}})$. If the observed credibility $\gamma(\boldsymbol{\theta}_t, \boldsymbol{d}_{\mathrm{obs}})$ for the parameter value $\boldsymbol{\theta}_t$ is lower than the required credibility (for a fixed desired coverage $\gamma_*$), the parameter value $\boldsymbol{\theta}_t$ is included in the confidence region:

$$C_{\boldsymbol{\theta}}(\boldsymbol{d}_{\mathrm{obs}}, \gamma_*) = \{\boldsymbol{\theta}_t : \gamma(\boldsymbol{\theta}_t, \boldsymbol{d}_{\mathrm{obs}}) \leq \gamma_R(\boldsymbol{\theta}_t, \gamma_*)\}. \tag{4.116}$$

This means that if the approximate posterior already has exact frequentist coverage, i.e., if $\gamma(\boldsymbol{\theta}_t, \boldsymbol{d}_{\mathrm{obs}}) = \gamma_R$, no modification to the posterior region will occur. Regions of parameters that are undercovered, i.e., regions where the required credibility is larger than the desired coverage, $\gamma_R(\boldsymbol{\theta}_t, \gamma_*) > \gamma_*$, will be added to the confidence region, see Fig. 4.18 for an illustration.

### Exercises

4.1 We investigate here in more detail the coin tossing problem. A coin is tossed $N$ times and heads come up $H$ times. Let $\theta$ denote the probability of heads in one flip.

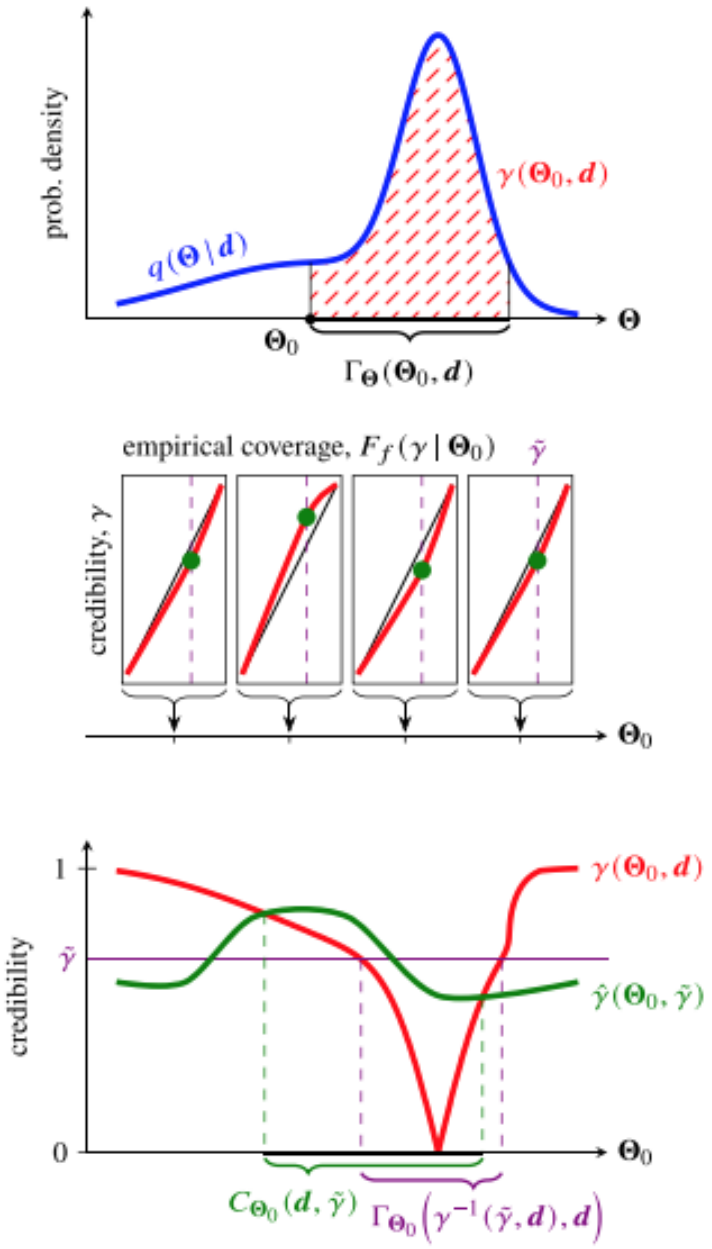


Figure 4.18 Calibration and validation of SBI posteriors. Top panel: definition of $\gamma(\boldsymbol{\theta}_0, \boldsymbol{d})$ as the HPD region that just encloses the true value of the parameters from which the data were generated at a credible level $\gamma$ (dashed area). Middle panel: By repeated draws of $\boldsymbol{d}$ at fixed $\boldsymbol{\theta}_0$, we obtain samples for $\gamma(\boldsymbol{\theta}_0, \boldsymbol{d})$, from its empirical conditional cumulative distribution, $F_F(\gamma \mid \boldsymbol{\theta}_0)$ is built (red lines). The credibilities $\gamma_R(\boldsymbol{\theta}_0, \gamma_*)$ required for regions to cover $\boldsymbol{\theta}_0$ with a given frequency $\gamma_*$ are indicated with green dots, obtained as the $\gamma_*$-th quantiles of $F_F$. Bottom panel: $C_{\boldsymbol{\theta}}(\boldsymbol{d}, \gamma_*)$, the region with confidence level $\gamma_*$, is that in which $\gamma(\boldsymbol{\theta}_0, \boldsymbol{d})$ (red line) is lower than $\gamma_R(\boldsymbol{\theta}_0, \gamma_*)$ (green line). Note: notation in the figure (from Karchev et al. (2022) is at the moment inconsistent).

1 Compute analytically the posterior probability for $\theta$, for the case of a Jeffreys' prior and a uniform prior on $\theta$. This integral will prove useful:

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

2 Determine the posterior mean and posterior standard deviation of $\theta$ for both choices of priors.

3 Plot the posterior (for both choices of prior) as a function of $\theta$, assuming that the true value of the parameter is $\theta_t = \{0.1, 0.5, 0.8\}$, and for $N = \{10, 100, 1000\}$. Compute numerically the 68.3% (i.e., $1\sigma$) highest posterior density (HPD) interval in each case, and compare it with the posterior standard deviation found above.

4 By simulating a large number of pseudo-data, compare the coverage properties of the 68.3% ($1\sigma$) and the 95.4% ($2\sigma$) HPD intervals for both choices of prior, as well as with the coverage properties of the likelihood-based MLE with Wald confidence intervals. Do this for the same choices of $\theta_t$ and $N$ as above.

4.2 We wish to obtain a numerical posterior over all the variables of the errors-in-variables model. The model has the following conditional structure, with known $\sigma_x, \sigma_y$ (and no intrinsic scatter) for $i = 1, \ldots, N$:

$$y_i^{\rm obs} \mid \boldsymbol{\theta}, x_i \sim \mathcal{N}\left(\theta_0 + \theta_1 x_i, \sigma_y^2\right), \tag{4.117}$$

$$x_i^{\rm obs} \mid x_i \sim \mathcal{N}\left(0, \sigma_x^2\right), \tag{4.118}$$

$$x_i \sim \mathcal{N}(x_0, R_x^2), \tag{4.119}$$

$$x_0 \sim \mathcal{N}(\mu_{x_0}, \sigma_{x_0}^2), \tag{4.120}$$

$$R_x^2 \sim \text{Inv-Gamma}(\alpha_R, \beta_R), \tag{4.121}$$

where we have chosen conjugate priors for the population-level parameters $x_0, R_x^2$ (with fixed hyper-parameters $\mu_{x_0}, \sigma_{x_0}^2, \alpha_R, \beta_R$) to enable analytical computation of the relevant conditional distributions.

Choose fiducial values of the variables $\boldsymbol{\theta}, x_0, R_x$ (and sensible values of $\sigma_x^2, \sigma_y^2$, ensuring that $\sigma_x \approx R_x$) and generate synthetic data from the model for $N = 100$. Then perform numerical inference on *all* variables, using one of the two following approaches:

1 Sample from the 4-dimensional marginal posterior $\Pr(\boldsymbol{\theta}, x_0, R_x | \boldsymbol{d})$ after analytical marginalization over the latent $x_i$; then obtain marginal posterior distributions for $x_i$ via conditional sampling. Simple Metropolis-Hastings should suffice in this case or you could use e.g. the `emcee` package.

2 Sample from the full $4 + N$ dimensional posterior using Gibbs sampling.

Plot posterior marginals for all parameters, and compare the posterior for the latent variables with the likelihood.



 

# 5

# Bayesian Model Comparison

This section is concerned with how to decide whether a model is a good explanation for the data. We will compare and contrast the Frequentist and Bayesian approach to this important question at the heart of the scientific method.

## 5.1 The Bayesian Evidence

Bayesian model comparison (BMC) is fundamentally different from Frequentist hypothesis testing. We begin by recalling salient aspects of the latter in order to highlight the shift in perspective that a Bayesian approach entails. Frequentist hypothesis testing focuses on the concept of **falsification**: the goal is to *reject* a null hypothesis (e.g., 'there is no signal in these data') because, roughly speaking, the data are improbable under the null (this is not quite accurate; we return to this in more detail below). In BMC, by contrast, the aim is to identify the 'best' model among a set of alternatives. In a Bayesian sense, the 'best' model is the one that combines at the same time a high quality of fit and a notion of simplicity or parsimony (the so-called **'Occam's razor'**[1]).

We can contextualize BMC in a stack of three levels of inference:

*Level 1: Parameter inference* We assume a model $M$, with parameter $\boldsymbol{\theta}_M$ and a prior $\Pr(\boldsymbol{\theta}_M \mid M)$. Then, given data $\boldsymbol{d}$, we determine the posterior for the parameters from Bayes' theorem:

$$\Pr(\boldsymbol{\theta}_M \mid \boldsymbol{d}, M) = \frac{\Pr(\boldsymbol{d} \mid M, \boldsymbol{\theta}_M) \Pr(\boldsymbol{\theta}_M \mid M)}{\Pr(\boldsymbol{d} \mid M)}. \tag{5.1}$$

[1] Named after William of Ockham, an English Franciscan friar who died in 1347 and was famous for expounding the principle, already variously expressed by others starting from Aristotle, that the simplest explanation for a given observed phenomenon was the best, or "entities should not be multiplied beyond necessity". For a history of Occam's razor in science, see McFadden (2023).

*Level 2: Model comparison* Given $N$ competing models $M_1, \ldots, M_N$, each with parameters $\boldsymbol{\theta}_{M_1}, \ldots, \boldsymbol{\theta}_{M_N}$ and priors $\Pr\left(\boldsymbol{\theta}_{M_1} \,\middle|\, M_1\right), \ldots, \Pr\left(\boldsymbol{\theta}_{M_N} \,\middle|\, M_N\right)$, and a prior distribution over models $\Pr(M_i)$ satisfying

$$\sum_i \Pr(M_i) = 1 \quad (i = 1, \ldots, N), \tag{5.2}$$

i.e. under the assumption that the list of models is complete the 'best' model is identified using the posterior odds between pairs of models:

$$\frac{\Pr(M_i \mid \boldsymbol{d})}{\Pr\left(M_j \,\middle|\, \boldsymbol{d}\right)}, \tag{5.3}$$

where the posterior probability for model $j$ is

$$\Pr\left(M_j \,\middle|\, \boldsymbol{d}\right) = \frac{\Pr\left(\boldsymbol{d} \,\middle|\, M_j\right) \Pr\left(M_j\right)}{\sum_{i=1}^{N} \Pr(\boldsymbol{d} \mid M_i) \Pr(M_i)}. \tag{5.4}$$

Given all the priors, the key quantity of the previous expression is the so-called **model likelihood** or **evidence** $\Pr\left(\boldsymbol{d} \,\middle|\, M_j\right)$ i.e., the probability of the data under each model. The evidence is also the predictive probability for the data under the model. Of course, in general we can expect that the list of models being considered is incomplete, or that the true model (if such a thing exists) is missing from the list, which is often the case in frontiers science. Different disciplines deal with this unquantifiable uncertainty in different ways: in the social sciences, the understanding is that 'all models are wrong, but some are useful' (a famous aphorism attributed to statistician George Box[2] ), while in the physical sciences, and particularly in fundamental physics, the working epistemological assumption is that, when correct, our model should somehow be a true mathematical reflection of how nature 'is'. Whatever one's epistemological stance, Bayesian model comparison gives a quantitative way of assessing each model performance in explaining the data, in a way that will become clear below.

*Level 3: Model averaging* If none of the models under consideration clearly stands out as the best (on a scale to be introduced later), we can still take this model uncertainty into account in parameter inference by performing model averaging. Model averaging is meaningful in the case where all the models share a common set of parameters of interest (plus additional, different parameters for each). Denoting by all $\boldsymbol{\theta}_C$ the common set of parameters across models, then the model-

[2] Box (1976) actually merely stated that 'all models are wrong' in his paper on the scientific method and the work of Fisher. Interestingly for the topic of this chapter, he also went on to say: "[...] following William of Occam [the scientist] should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity." (ibid).

 

averaged posterior is given by:

$$\Pr(\boldsymbol{\theta}_C \mid \boldsymbol{d}) = \sum_{i=1}^{N} \Pr(M_i \mid \boldsymbol{d}) \Pr(\boldsymbol{\theta}_C \mid \boldsymbol{d}, M_i) \,, \tag{5.5}$$

where $\Pr(\boldsymbol{\theta}_C \mid d, M_i)$ is the marginal posterior on the common parameters under model $M_i$. The weighing factor is given by each model's posterior.

### *5.1.1 The Bayes Factor*

The Bayesian approach to model selection is conceptually appealing, as it unifies inference and hypothesis testing in a single framework—namely, evaluation of posterior probabilities obtained via Bayes theorem. It is also free from the misinterpretations that plague Frequentist hypothesis testing. However, it is not entirely free from issues, either, the most thorny of which is arguably its intrinsic (and unavoidable) prior dependency, as we shall see.

Bayesian model selection is based on the **model likelihood (or Bayesian evidence)** :

$$\Pr(\boldsymbol{d} \mid M) = \int_{S_M} \Pr(\boldsymbol{d} \mid \boldsymbol{\theta}_M, M) \Pr(\boldsymbol{\theta}_M) \,; \tag{5.6}$$

where the domain of integration is over the model's parameter space, $S_M$. From here, we can use Bayes Theorem to obtain the posterior probability for the model:

$$\Pr(M \mid \boldsymbol{d}) = \frac{\Pr(\boldsymbol{d} \mid M) \Pr(M)}{\Pr(\boldsymbol{d})}, \tag{5.7}$$

where $\Pr(M)$ is the model's prior probability.

As the name implied, the Bayesian approach to model selection is one of comparison, i.e., we must always specify at least an alternative model to which to compare a baseline model of interest. This is achieved by considering the **posterior odds** between models $M_i$ and $M_j$ $(i, j = 1, \ldots, N)$:

$$\frac{\Pr(M_i \mid \boldsymbol{d})}{\Pr\left(M_j \,\middle|\, \boldsymbol{d}\right)} = \frac{\Pr(\boldsymbol{d} \mid M_i)}{\Pr\left(\boldsymbol{d} \,\middle|\, M_j\right)} \frac{\Pr(M_i)}{\Pr\left(M_j\right)}, \tag{5.8}$$

where the ratio

$$B_{ij} = \frac{\Pr(\boldsymbol{d} \mid M_i)}{\Pr\left(\boldsymbol{d} \,\middle|\, M_j\right)} \tag{5.9}$$

is the **Bayes factor**, the factor by which our relative prior degree of belief (or prior odds) in the models under consideration has been changed by the data. Therefore, the Bayes factor updates the prior odds (between the two models) to their posterior odds – a similar role played by the likelihood in inference, except here the update is in the odds (ratio of probabilities). If one reports the

Bayes factor, the posterior odds can be computed for any prior odds one wishes to specify. Hence Bayesian model comparison is fully characterised by the pairwise Bayes factors between one reference model (say, model $M_k$) and all the others, since the Bayes factor between models $i$ and $j$ can always be computed as $B_{ij} = B_{ik}B_{kj} = B_{ik}B_{jk}^{-1}$; hence, knowledge of $B_{ik}$ for all $i \neq k$ and fixed $k$ enables the computation of all the other Bayes factors.

The posterior probability for model $k$ can then be obtained from Eq. (5.4) in terms of Bayes factors between $M_k$ and the other models:

$$\Pr(M_k \mid \boldsymbol{d}) = \frac{1}{1 + \sum_{i \neq k}^{N} B_{ik} \frac{\Pr(M_i)}{\Pr(M_k)}}. \tag{5.10}$$

In the absence of strong reasons for preferring one model over another a priori, the models' priors can be chosen equal, so that the prior ratio $\Pr(M_i) / \Pr(M_k)$ is unity.

The idea underpinning BMC is radically different from hypothesis testing: the fundamental outlook is that there is no point in rejecting a model unless a better alternative is available—hence the emphasis on *comparison* as opposed to rejection. Through Eq. (5.7), we obtain the quantity that we care about, namely, the posterior probability for each model under scrutiny. Also, evidence can and does accumulate in favour of a model if its predictions are borne out by observation. Therefore, the Popperian notion of falsification is seen to be limited. In reality, when a model makes a prediction that is confirmed by the data, our degree of belief in that model increases, which is correctly reflected in the outcome of Bayesian model comparison. Consider, for example, Einstein's prediction of the deflection angle of the light around the Sun. When his prediction was confirmed by the May 29th 1919 solar eclipse observation, general relativity was considered verified, and became the *de facto* new standard model of gravity. Nobody said that general relativity merely failed to be falsified by Eddington's data!

We can also see the evidence as the predictive probability for the data under model $M$ and prior $\Pr(\boldsymbol{\theta}_M \mid M)$. This is illustrated in Fig. 5.1, where the evidence for two models, $M_0$ and $M_1$, is represented with two different densities in data space (horizontal axis, where the data are shown in one dimension for illustration. In reality the data space is of course typically very high dimensional). Model $M_0$ concentrates its prediction for the observation in the vicinity of $d_1$ (where the density is largest); model $M_1$ spreads out its predictions more widely in data space, typically as a consequence of having a larger number of tunable parameters. When the data arrive (vertical dashed lines, indicating two possible data sets), the relative densities of their predictions are compared: if the data are $d_1$, then the Bayes factor will favour model $M_0$, as its quite sharp prediction has been verified by the data; on the contrary, if the data are given by $d_2$, this is evi-

 

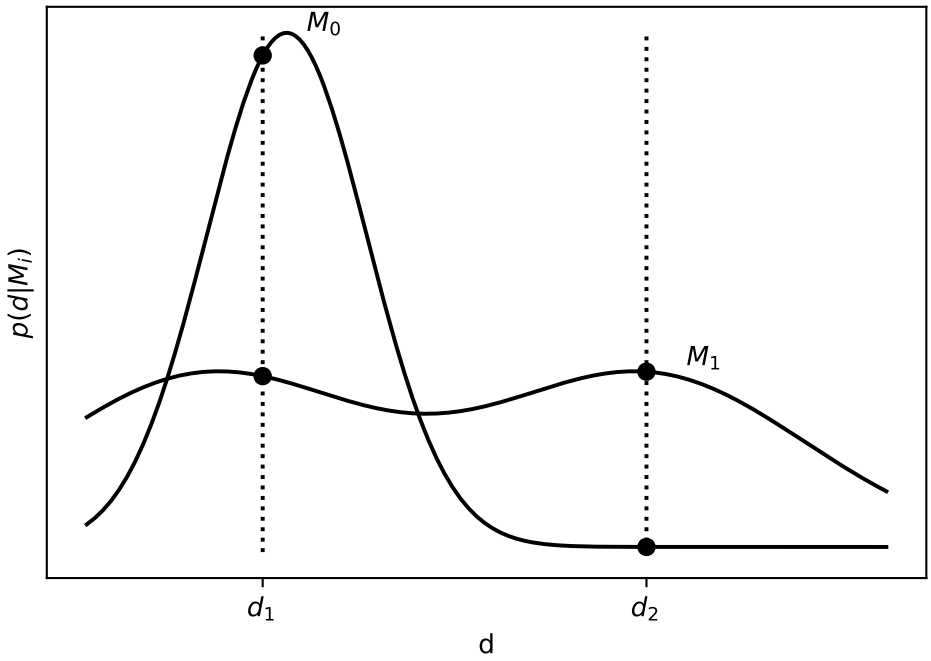


Figure 5.1 Illustration of Bayesian model comparison. If the observations fall on $d_1$, the Bayesian evidence (ratio between the densities) favours model $M_0$, as the more predictive of the two; if the data are given by $d_2$, the more flexible model $M_1$ is favoured.

dence that the additional flexibility of $M_1$ is required, and correspondingly it will be favoured by the Bayes factor (as the ratio of the evidences at $d = d_2$ favours model $M_1$).

### *5.1.2 The Occam's Factor*

The 'best' model in a Bayesian sense is one that balances quality of fit (as measured by the likelihood) and simplicity, or parsimony of parameters, a quantitative implementation of the concept of Occam's razor through the so-called **Occam's razor**. It is important to highlight that in model comparison the role of the parameters' prior never disappears, not even asymptotically, as it controls the strength of the Occam's penalty. This is fundamentally different from inference, where as we saw in the previous chapter, the prior influence vanishes for large sample sizes.

For simplicity, we consider first a one-parameter case, but the generalization

 

**Example 5.1: Fair coin and Bayesian model comparison**

A coin is tossed $N = 12$ times, $r = 3$ of which yield heads. Consider the two models, each with prior $\Pr(M_i) = 1/2$:

- $M_0$: $\theta = 1/2$ (the coin is fair); i.e., $\Pr(\theta \mid M_0) = \delta(\theta - 1/2)$
- $M_1$: $\theta \in \text{Uniform}(0,1)$, i.e. $\Pr(\theta \mid M_1) = 1$ (for $0 \leq \theta \leq 1$). This represents a model where we have no idea about the fairness of the coin; we could also choose a more complex prior, such as for example a symmetric Beta density, $\text{Beta}(\alpha,\alpha), \alpha > 0$; an asymmetric prior, $\text{Beta}(\alpha,\beta), \alpha,\beta > 0$ would instead represent a state of belief that the coin is biased one way or the other.

We wish to compute the Bayes factor between the two models. The evidence for each is:

$$\begin{aligned}\Pr(\boldsymbol{d} \mid M_1) &= \int \mathrm{d}\theta \Pr(r, N \mid \theta) \Pr(\theta \mid M_1) \\ &= \int \mathrm{d}\theta \binom{N}{r} \theta^r (1-\theta)^{N-r} \cdot 1 \\ &= \binom{N}{r} \frac{r!\,(N-r)!}{(N+1)!}, \end{aligned} \tag{5.11}$$

$$\Pr(\boldsymbol{d} \mid M_0) = \binom{N}{r} \frac{1}{2^N}; \tag{5.12}$$

hence the Bayes factor is

$$B_{01} = \frac{\Pr(\boldsymbol{d} \mid M_0)}{\Pr(\boldsymbol{d} \mid M_1)} = \frac{(N+1)!}{r!\,(N-r)!} \approx 0.7\,. \tag{5.13}$$

which translates in a posterior for $M_1$ of about 58% (up from a 50% prior). There is no evidence for bias – despite the fact that a Frequentist hypothesis test would reject the null (that $\theta = 0.5$) at the 5% significance level.

to any number of dimensions is straightforward and will be presented below. Consider a model with one parameter $\theta$, endowed with a prior with variance $\Sigma^2$, and a likelihood with variance $\sigma^2$. Here, prior and likelihood need not be Gaussian, as in this example we are just considering order-of-magnitude estimates; we are implicitly assuming that they are both unimodal, as in the illustration in Fig. 5.2). In a typical situation, $\Sigma > \sigma$, as the prior ought to be more diffuse than the likelihood when the data are informative. We can estimate the evidence for

 

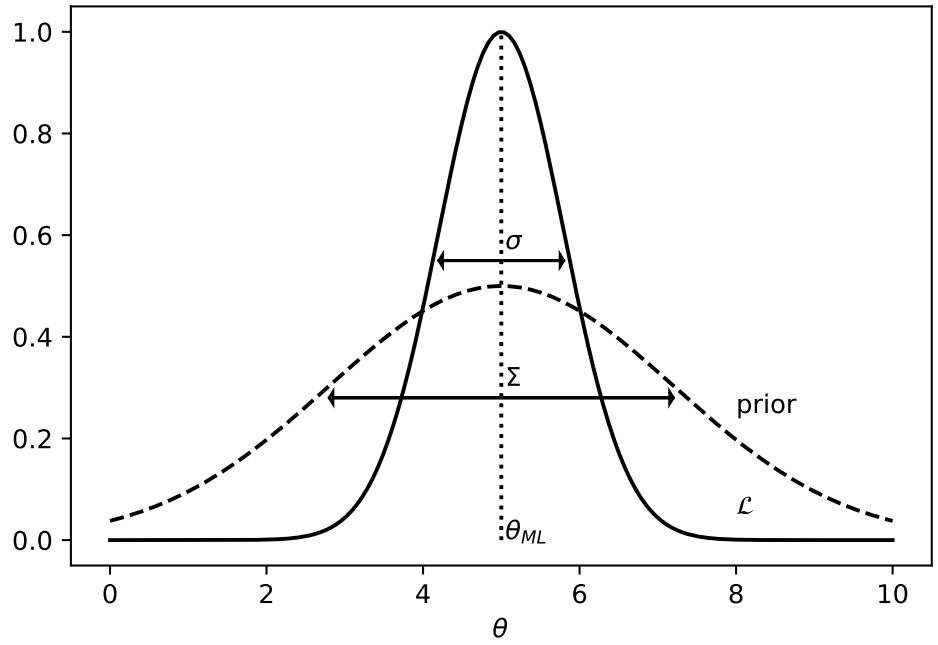


Figure 5.2 Illustration of the Occam's factor in Bayesian model comparison.

this model as:

$$\begin{aligned}\Pr(\boldsymbol{d}\,|\,M) &= \int \mathrm{d}\theta\, \mathscr{L}(\theta)\Pr(\theta) \\ &\approx \sigma\, \mathscr{L}(\boldsymbol{\theta}_{\mathrm{MLE}})\Pr(\boldsymbol{\theta}_{\mathrm{MLE}}) \\ &\approx \mathscr{L}(\boldsymbol{\theta}_{\mathrm{MLE}})\,\frac{\sigma}{\Sigma}\,; \end{aligned} \tag{5.14}$$

here we have approximated the likelihood as its value at the MLE point times its characteristic width (the square root of its variance), and approximated the prior as a constant given by its density at the MLE, which in turn is proportional to its standard deviation because of the normalization constraint. We see from Eq. (5.14) that the evidence is large when the maximum likelihood values is large, that is, when the fit to the data under the model is good, but it is penalised by the so-called **Occam's factor**, given by the ratio of the likelihood to prior volumes, here simply $\sigma/\Sigma < 1$:

$$\text{Occam's factor} = \frac{\text{likelihood volume}}{\text{prior volume}} \leq 1\,. \tag{5.15}$$

The interpretation of the Occam's factor is that a model is penalized for having

 

access (under a diffuse prior) to a large volume of parameter space, which is then ruled out (one could say, sliced away) by the likelihood. In this sense, 'wasteful models' (i.e., ones that have a large number of tunable parameters, each with a diffuse prior indicating lack of predictivity) are penalized as overly complex.

If a model has a parameter that is not constrained by the data (in which case, $\sigma > \Sigma$, and the posterior is prior-dominated) then the integral of Eq. (5.14) can be approximated as $\mathscr{L}(\boldsymbol{\theta}_{\mathrm{MLE}}) \frac{\Sigma}{\Sigma} = \mathscr{L}(\boldsymbol{\theta}_{\mathrm{MLE}})$, because the integration domain is limited by the support of the prior, rather than that of the likelihood. If therefore see that such a parameter does not attract any Occam's razor penalty. If a parameter is irrelevant (or unconstrained), Bayesian model comparison does not penalize a model for it.

### *5.1.3 Strength of Evidence*

Many scientists are familiar with the 'number of sigma' significance for rejection in Frequentist hypothesis testing (more on which below), but not so acquainted perhaps with a scale on which to assess the strength of evidence in favour or against a model as quantified by the Bayes factor. Following Goodman (1999), we can develop an intuitive understanding of the scale of evidence by studying the change it induces on the corresponding posterior probabilities for models.

When considering pairs of models, odds and probabilities are related by the following equations, special cases of Eq. (5.10):

$$\text{odds} = \frac{\text{prob}}{1-\text{prob}}, \quad \text{and} \tag{5.16}$$

$$\text{prob} = \frac{\text{odds}}{1+\text{odds}}. \tag{5.17}$$

If we have only two models, then their priors are related by $\Pr(M_1) = 1 - \Pr(M_0) \equiv P_1$, and the posterior for $M_1$ is given by:

$$\Pr(M_1 \mid d) = \left(1 + B_{01}\frac{1-P_1}{P_1}\right)^{-1} = \frac{B_{10}P_1}{1+P_1(B_{10}-1)}. \tag{5.18}$$

This is illustrated in Fig. 5.3, which shows the resulting posterior probability, $\Pr(M_1 \mid d)$, as a function of the model's prior, $P_1$, and the Bayes factor $B_{01}$. By way of illustration, consider the hypothetical case in which we start from a prior belief of 90% in a model, corresponding to prior odds of 9 : 1; then a Bayes factor of, for example, 1 : 10 leads to a posterior probability of 47%: If instead we started from a non-committal prior probability of 50%, a Bayes factor of 1 : 10 leads to a 9% posterior probability. This kind of reasoning leads to an empirically calibrated **Jeffreys' scale** for the strength of evidence, shown in Table 5.1.

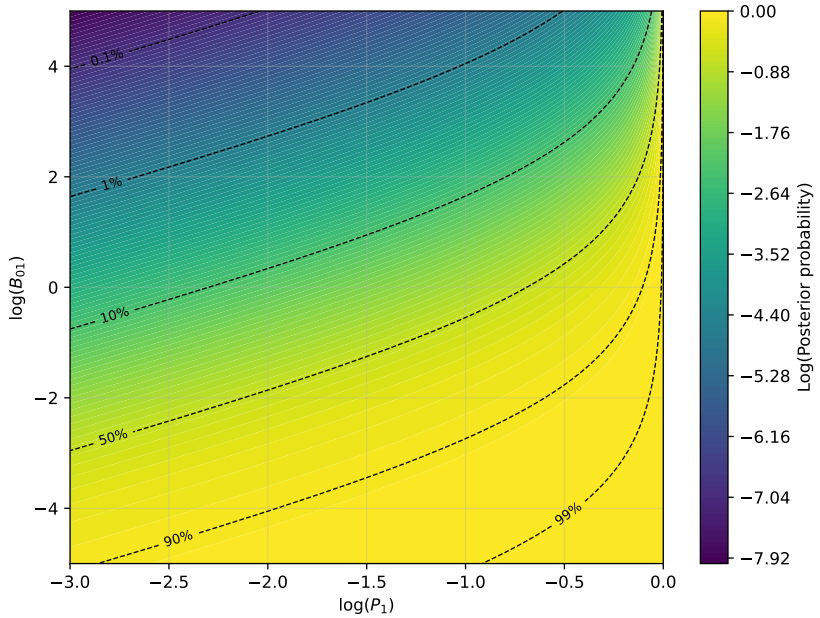


Figure 5.3 Posterior probability for a two-ways model comparison as a function of the model's prior and Bayes factor.

| $\lvert\ln B_{01}\rvert$ | BF odds | $\Pr(M_1 \mid d)$ for $\Pr(M_1) = 0.5$ | $\Pr(M_1 \mid d)$ for $\Pr(M_1) = 0.1$ | Designation |
|---|---|---|---|---|
| 1.0 | $\sim 3:1$ | 0.73 | 0.23 | weak |
| 2.5 | $\sim 12:1$ | 0.924 | 0.58 | moderate |
| 5.0 | $\sim 150:1$ | 0.993 | 0.94 | strong |

Table 5.1 *Scale for the strength of evidence. The indicative thresholds are following Jeffreys, while the description as weak/moderate/strong is my own.*

## 5.2 Nested Models

Let us consider the special (but very common) case of nested models , i.e., where we have a baseline model $M_0$ and we want to check for evidence of one (or more) extra parameter. This is often the case when we have an extended model $M_1$, which might for example feature a signal on top of the background described by model $M_0$. We consider the situation where the extended model has one single additional parameter, while the simpler model has no free parameters at all. While this might appear quite restrictive, it does not imply lack of generality for

 

in Section 5.2.2 we shall see that the common parameters between the models can be entirely ignored (under very mild assumptions) in the model comparison,

### *5.2.1 Gaussian Nested Models*

Consider a simple Normal case, where the extended model is $M_1$, with one additional parameter with prior $\Pr(\theta \mid M_1) = \mathcal{N}(0, \Sigma^2)$. The likelihood for the extra parameter[3] is also a Gaussian with variance $\sigma^2$, and mean given by the maximum likelihood value, $\mathcal{L}(\theta) \propto \mathcal{N}(\theta_{\rm MLE}, \sigma^2)$. Define

$$\lambda = \frac{|\theta_{\rm MLE}|}{\sigma}, \tag{5.19}$$

which is the distance of the MLE for the parameter from the origin, measured in units of the likelihood's standard deviation, $\sigma$. The Bayes factor can be computed easily:

$$B_{01} = \frac{\Pr(\boldsymbol{d} \mid M_0)}{\Pr(\boldsymbol{d} \mid M_1)} = \sqrt{1 + \left(\frac{\sigma}{\Sigma}\right)^{-2}} \exp\left(-\frac{1}{2}\frac{\lambda^2}{1 + \left(\frac{\sigma}{\Sigma}\right)^2}\right). \tag{5.20}$$

Let's consider the limiting cases:

1 if $\lambda \gg 1$ we have a 'many-$\sigma$ detection'; the exponential dominates, so $B_{01} \to 0$ and $M_1$ is favoured; in this regime, Bayesian model comparison and Frequentist hypothesis testing agree.
2 if $\lambda \sim 2-3$ and $\sigma/\Sigma \ll 1$ we have a sharply-peaked likelihood in the vicinity of the origin and $B_{01} \approx \Sigma/\sigma$, dominated by the Occam's factor; note that this is linear in $\Sigma$ (and inversely proportional to $\sigma$), in contrast to the previous case in which the dependence was exponential in $\sigma$; in this case, the simpler model remains favoured despite a potentially 'significant' rejection of the null hypothesis in Frequentist terms;
3 if $\sigma/\Sigma \gg 1$, the likelihood is broader than the prior, then $B_{01} \to 1$; this is the prior-dominated regime, where Bayesian model selection returns an undecided result (no change in the models' odds).

In the 'informative data limit' (i.e. when $\sigma/\Sigma \ll 1$), we can approximate

$$\ln B_{01} \approx \ln\left(\frac{\Sigma}{\sigma}\right) - \frac{\lambda^2}{2}. \tag{5.21}$$

The first term, which is positive, corresponds to the Occam's factor, and provides

[3] In the case of common parameters between the two models, this should be interpreted as the marginal likelihood for the extra parameter.

evidence in favour of $M_0$; while the second factor, quadratic in $\lambda$, is negative in virtue of the minus sign in front, thus contributing to evidence in favour of $M_1$.

We can gain further insight by considering the Kullback-Leibler (KL) divergence between prior and posterior, as defined in Eq. (3.52), which measures the information gain obtained through the data in going from the prior to the posterior. For this example, denoting the Gaussian posterior by $\mathcal{N}\left(\mu, \tau^2\right)$, we have that

$$D_{\mathrm{KL}}(\Pr(\theta|\boldsymbol{d}) \,||\, \Pr(\theta)) = -\ln\frac{\tau}{\Sigma} + \frac{1}{2}\left(\frac{\tau}{\Sigma}\right)^2\left[\left(\frac{\mu}{\tau}\right)^2 + 1\right] - \frac{1}{2}. \tag{5.22}$$

When the data are non-informative, the posterior is equal to the prior, hence $\tau = \Sigma$ and $\lambda = 0$, and therefore $D_{\mathrm{KL}} = 0$. If the data are informative (i.e., in the case $\tau/\Sigma < 1$), unless the extra parameter has been detected at high significance (when $\mu/\tau \gg 1$, i.e., the posterior peaks far away from 0), the second term is subdominant w.r.t. the first, and we can therefore approximate

$$D_{\mathrm{KL}} = \mathrm{const} + \ln\frac{\Sigma}{\tau}, \tag{5.23}$$

where the basis for the logarithm defines the units in which the information gain in going from prior to posterior is measured. In this informative data limit, the posterior standard deviation is approximately equal to the likelihood's, therefore $\tau \approx \sigma$ and we see that the KL divergence (up to an irrelevant constant, which depends on the basis used for the logarithm) corresponds to the Occam's factor, Eq. (5.21): a large information gain is a signal of a great deal of wasted parameter space, which is therefore penalized.

We can therefore approximately describe (up to a constant) the information gain by the quantity $I_{10} = \log_{10}\Sigma/\sigma$, where the subscript indicates that we are using base 10 logarithm. A summary of Bayesian model comparison for Gaussian nested model is shown and illustrated in Fig. 5.4, showing values of $\ln B_{01}$ in the plane spanned by $I_{10}$ and $\lambda$. The figure demonstrates that the outcome of Bayesian model comparison is a two-parameters game: the 'significance' of the detection (vertical axis) and the Occam's razor penalty. Large significance (large values of $\lambda$) translates into a model comparison result favouring the more complex model ($\ln B_{01} \ll 0$, blue region in the plot). However, 'detection' at 2-3$\sigma$ significance can result in a model comparison outcome that can favour either model, or be undecisive (white region), depending on the choice of prior for the extra parameter. Recall that, for fixed likelihood, a wider prior on the extra parameter corresponds to a horizontal shift towards the right of the figure. Finally, if the data are less informative than the prior, i.e., for $I_{10} < 0$, the model comparison result is undecided, according to intuition, i.e., $\ln B_{01} = 0$ (grey region on the left).

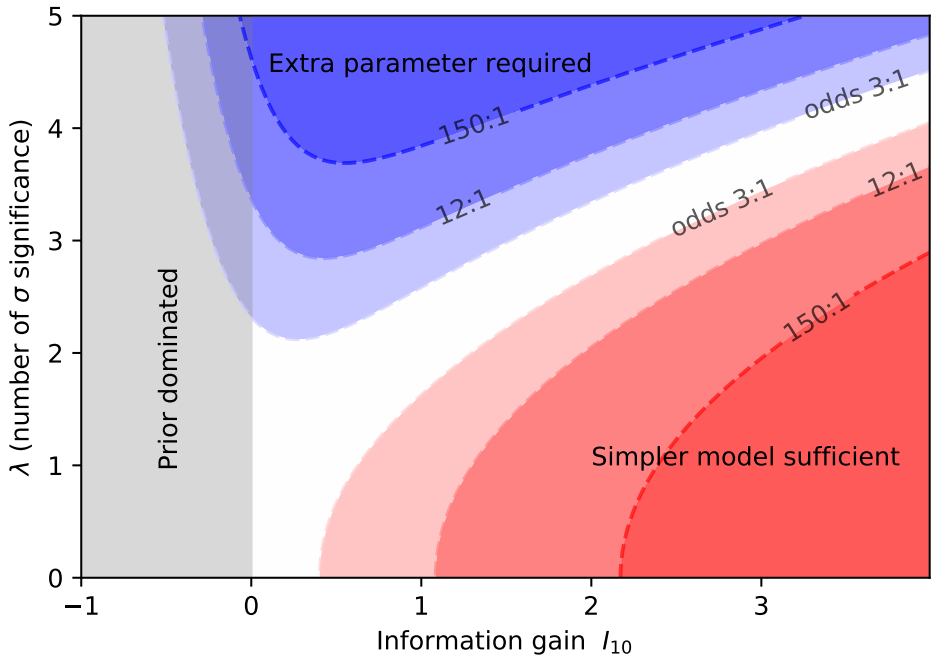


Figure 5.4 Summary of Bayesian model comparison result in the case of Gaussian nested models. The blue-shaded region is where the extra parameter is required by the data, the red-shaded region is where the simpler model is sufficient, while the white region corresponds to an undecided outcome (adapted from Trotta (2008)).

The above can be generalized to the multivariate Gaussian case, where the extended model $M_1$ has additional parameters $\boldsymbol{\theta}$ with prior $\Pr(\boldsymbol{\theta} \mid M_1) = \mathcal{N}(\mathbf{0}, \boldsymbol{P}^{-1})$ and likelihood $\mathscr{L}(\boldsymbol{\theta}) = \mathscr{L}_0 \mathcal{N}\left(\boldsymbol{\theta}_{\rm MLE}, \boldsymbol{L}^{-1}\right)$, $\boldsymbol{P}$ and $\boldsymbol{L}$ being the Fisher matrices of the prior and likelihood, respectively, and $\mathscr{L}_0$ a normalization constant, the value of the likelihood at the MLE. The evidence is

$$\Pr(\boldsymbol{d} \mid M) = \mathscr{L}_0 \frac{|\boldsymbol{F}|^{-1/2}}{|\boldsymbol{P}|^{-1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}_{\rm max}^{\top} \boldsymbol{L} \boldsymbol{\theta}_{\rm max} - \overline{\boldsymbol{\theta}}^{\top} \boldsymbol{F} \overline{\boldsymbol{\theta}}\right)\right] \tag{5.24}$$

where $\boldsymbol{F} = \boldsymbol{L} + \boldsymbol{P}$ is the posterior Fisher matrix and $\overline{\boldsymbol{\theta}} = \boldsymbol{F}^{-1} \boldsymbol{L} \boldsymbol{\theta}_{\rm max}$ the posterior mean. Here, determinants encode the volume factors, describing the volume of parameter space enclosed by the prior and the posterior:

$$|\boldsymbol{P}| \leq |\boldsymbol{F}| \Rightarrow \frac{|\boldsymbol{F}|^{-1/2}}{|\boldsymbol{P}|^{-1/2}} \leq 1\,, \tag{5.25}$$

so this multi-dimensional Occam's factor penalises models with $|\boldsymbol{F}| \ll |\boldsymbol{P}|$ i.e. 'wasted' parameter space.

### 5.2.2 General Nested Models: The Savage-Dickey Density Ratio

We now consider the more general case of two nested models but without making any assumption about the form of the prior and the likelihood. In general, we therefore wish to investigate the outcome of Bayesian model comparison in the case where we have a 'baseline model', $M_0$, with parameters $\boldsymbol{\theta}$ and with prior $\Pr(\boldsymbol{\theta} \mid M_0)$, vs an 'extended model' $M_1$, with parameters $\{\boldsymbol{\theta}, \omega\}$ and prior $\Pr(\boldsymbol{\theta}, \omega \mid M_1)$. Here, the extended model has a single additional parameter, $\omega$, but generalizing to more additional parameters is straightforward. Notice that the common parameters, $\boldsymbol{\theta}$, have no restriction on their cardinality.

In nested models, the extended model reverts to the baseline for a particular value of the extra parameter, here $\omega = \omega_*$. If, for example, $\omega$ describes the amplitude of a signal being searched for, $\omega_* = 0$. Should there be additional extended model parameters (e.g., characterizing the spectrum or location of a source whose amplitude is described by $\omega$), such parameters become unidentifiable (i.e., meaningless) for $\omega = 0$, which however is not a problem for Bayesian model comparison. The Bayes factor between the two models is:

$$B_{01} = \frac{\int \mathrm{d}\boldsymbol{\theta} \Pr(\boldsymbol{\theta} \mid M_0) \Pr(\boldsymbol{d} \mid \boldsymbol{\theta}, \omega = \omega_*)}{\Pr(\boldsymbol{d} \mid M_1)} . \tag{5.26}$$

Using the equality:

$$\Pr(\omega_*, \boldsymbol{\theta} \mid \boldsymbol{d}, M_1) = \Pr(\boldsymbol{\theta} \mid \omega_*, \boldsymbol{d}, M_1) \Pr(\omega_* \mid \boldsymbol{d}, M_1) \tag{5.27}$$

and multiplying and dividing by $\Pr(\omega_* \mid \boldsymbol{d}, M_1)$, we obtain:

$$B_{01} = \frac{\Pr(\omega_* \mid \boldsymbol{d}, M_1)}{\Pr(\boldsymbol{d} \mid M_1)} \int \mathrm{d}\boldsymbol{\theta} \frac{\Pr(\boldsymbol{\theta} \mid M_0) \Pr(\boldsymbol{d} \mid \boldsymbol{\theta}, w_*) \Pr(\boldsymbol{\theta} \mid \omega_*, \boldsymbol{d}, M_1)}{\Pr(\omega_*, \boldsymbol{\theta} \mid \boldsymbol{d}, M_1)} . \tag{5.28}$$

And, since

$$\Pr(\omega_*, \boldsymbol{\theta} \mid \boldsymbol{d}, M_1) = \frac{\Pr(\boldsymbol{d} \mid \omega_*, \boldsymbol{\theta}) \Pr(\omega_*, \boldsymbol{\theta} \mid M_1)}{\Pr(\boldsymbol{d} \mid M_1)} , \tag{5.29}$$

finally

$$B_{01} = \Pr(\omega_* \mid \boldsymbol{d}, M_1) \int \mathrm{d}\boldsymbol{\theta} \frac{\Pr(\boldsymbol{\theta} \mid M_0) \Pr(\boldsymbol{\theta} \mid \omega_*, \boldsymbol{d}, M_1)}{p_1(\omega_*, \boldsymbol{\theta})} . \tag{5.30}$$

So far we only manipulated probabilities, without making any assumption about their properties. To proceed further, wenow assume that the priors are separable[4], i.e. that the prior on the extended model's parameter is independent of the baseline model parameters' prior:

$$\Pr(\omega, \boldsymbol{\theta} \mid M_1) = \Pr(\omega \mid M_1) \Pr(\boldsymbol{\theta} \mid M_0) . \tag{5.31}$$

[4] This is actually a condition stronger than necessary, but which is usually met in practice.

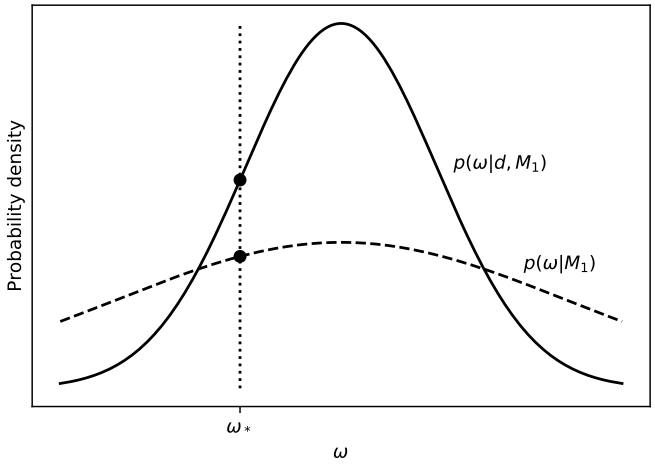


Figure 5.5 Illustration of the Savage-Dickey Density Ratio (SDDR). The SDDS is given by the ratio of the marginal (normalized!) posterior to prior densities for the additional parameter, evaluated at the value that makes the extended model revert to the simpler one (vertical, dotted line).

Then

$$B_{01} = \frac{\Pr(\omega_* \,|\, \boldsymbol{d}, M_1)}{\Pr(\omega_* \,|\, M_1)} \int \mathrm{d}\boldsymbol{\theta} \frac{Pr\boldsymbol{\theta} \,|\, M_0 \Pr(\boldsymbol{\theta} \,|\, \omega_*, \boldsymbol{d}, M_1)}{\Pr(\boldsymbol{\theta} \,|\, M_0)}$$
$$= \frac{\Pr(\omega_* \,|\, \boldsymbol{d}, M_1)}{\Pr(\omega_* \,|\, M_1)} . \tag{5.32}$$

The last expression is known as the Savage–Dickey Density Ratio (SDDR) (by Dickey, who attributed it to Savage); a generalised version was introduced in Verdinelli and Wasserman (1995).

As illustrated in Fig. 5.5, the SDDR says that the Bayes factor between nested models is simply the ratio between the posterior marginal density under the extended model on $\omega$ evaluated at $\omega_*$ and its prior density. Besides being useful as a computational tool (a question to which we return below), it is illuminating theoretically: it clarifies that parameters that are common to both models (namely, $\boldsymbol{\theta}$) do not matter in the outcome for model selection — nor does the common prior choice on them, as long as it does not alter appreciably the marginal posterior for $\omega$ (for example, doubling the range of a uniform prior on $\boldsymbol{\theta}$ has no influence on the marginal posterior for $\omega$, provided the likelihood support is within the original prior range). We highlight once again that the SDDR does not assume (nor require) Gaussianity in any of the distributions.

However, the prior on $\omega$ is important and does not disappear from the result, not even asymptotically, as it appears in the denominator of Eq. (5.32). The

 

influence of $\Pr(\omega \mid M_1)$ on the marginal posterior for $\omega$ will vanish asymptotically, but the prior density in the denominator controls the strength of the Occam's razor. Because the prior needs to be normalizable, its density is of order $\Pr(\omega_* \mid M_1) \sim 1/\Sigma$, where $\Sigma$ is the characteristic scale of the prior, and therefore

$$B_{01} \sim \Pr(\omega_* \mid M_1)^{-1} \sim \Sigma, \tag{5.33}$$

meaning that one can arbitrarily increase the strength of the Occam's factor favouring the baseline model simply by making $\Sigma$ arbitrarily large. This is at the heart of the so-called **Jeffreys–Lindley paradox**. This paradox[5] (first presented by Lindley (1957)), says that there are cases in which we cannot reconcile $p$-values and Bayes factors, even asymptotically (when the number of data points $n \to \infty$).This is an important case, for the scientific conclusion that one reaches from the same data might depend crucially on one's approach to inference (for an overview, see Cousins (2013); Lyons and Demortier (2014)).

## 5.3 Calibration of Bayes Factors

To garner intuition on the difference, it is useful to compare the outcome of Bayesian model selection with Frequentist hypothesis testing. We therefore begin by reviewing the latter. We begin by briefly reviewing the basics of Frequentist hypothesis testing, restricting the discussion to simple hypotheses (i.e., with no free parameters) for the sake of simplicity, though the considerations below can be generalized to so-called "composite" hypotheses (which depend on the value of unknown parameters).

### *5.3.1 Frequentist hypothesis testing*

In the Frequentist setting of the Neyman-Pearson school of thought[6], a null hypothesis $H_0$ is defined and contrasted with an alternative hypothesis $H_1$. frequentist hypothesis testing assesses whether observed data provides enough evidence to reject $H_0$ in favor of $H_1$, using a predefined significance threshold, usually denoted as $\alpha$. Usually, the null hypothesis is what we would like to reject (e.g., no signal in the data). One then considers the distribution of a suitably defined test statistics $TS$ in each case, assuming that either $H_0$ or $H_1$ is true. Based on $\alpha$, a **rejection (or critical) region** is defined in outcome space, a so-called decision rule: if the observed value of the test statistics falls within the

[5] Also known as "Bartlett's paradox", because of a crucial correction by Bartlett (1957).

[6] This is as opposed to the view exposed by Fisher, according to whom the alternative $H_1$ does not need to be explicitly stated, but the test is about rejecting $H_0$. Another important difference is that in hypothesis testing à la Fisher, the $p$-value is a measure of the strength of evidence against the null, rather than used against a pre-determined significance threshold to decide whether or not to reject it.

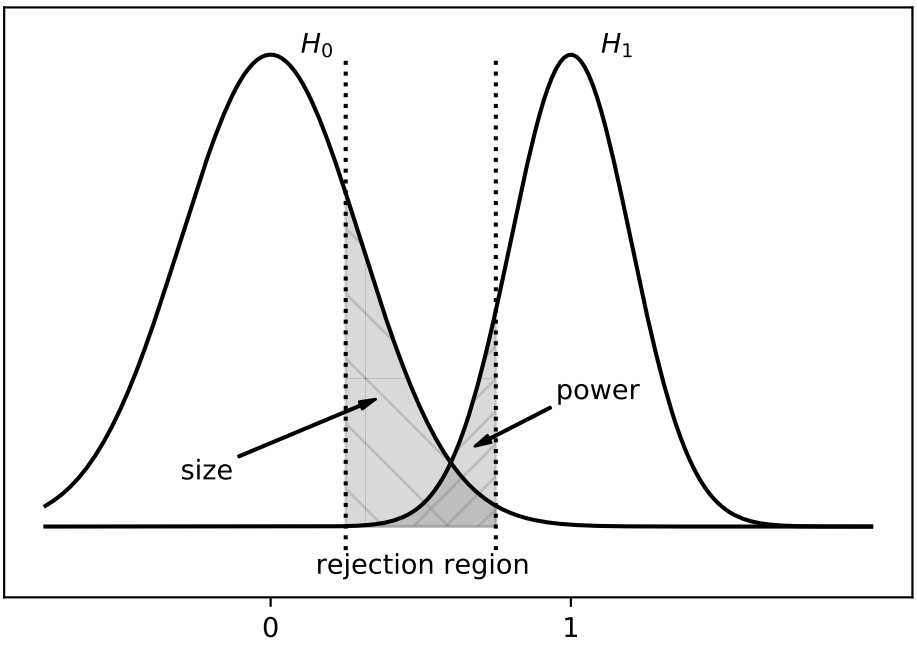


Figure 5.6 Frequentist hypothesis testing: definition of size and power of the test for a simple null hypothesis, $H_0: \mu = 0$, vs a simple alternative, $H_1: \mu = 1$, for the rejection region $R$ bounded by the two vertical, dotted lines. The horizontal axis is the value of the sample mean, a sufficient statistic in this case.

rejection region, the null hypothesis $H_0$ is rejected and the alternative $H_1$ is accepted. Note, however, that accepting $H_1$ does *not* mean that $H_1$ is true! For a review of hypothesis testing in the context of high energy physics, see Cowan et al. (2011).

The **size** of the rejection region is the probability of getting an outcome in that region if $H_0$ is true. Since we aim at rejecting $H_0$, we want a small size. The **power** of the rejection region is the probability of getting an outcome in that region if $H_1$ is true. We therefore aim for large power. The power and size of a test with rejection region $R$ for the sample mean are illustrated in Fig. 5.6. Denoting the rejection region by $R$, **type I errors** are false positives (incorrectly rejecting the null when it is, in fact, true), while **type II errors** are false negatives (incorrectly failing to reject the null):

$$\Pr\big(\text{type I error}\big) = \Pr\big(\text{accepting } H_1 \mid H_0 \text{ true}\big) = \Pr\left(TS \in R \mid H_0 \text{ true}\right). \tag{5.34}$$

The probability of type II error is:

$$\Pr\big(\text{type II error}\big) = \Pr\big(\text{accepting } H_0 \mid H_1 \text{ true}\big) = \Pr\left(TS \notin R \mid H_1 \text{ true}\right), \tag{5.35}$$

 

and the power is then given by:

$$\text{power} = 1 - \Pr\big(\text{type II error}\big)\,. \tag{5.36}$$

For a simple hypothesis, the **level of significance** $\alpha$ of the test is equal to the probability of committing a type I error[7]. The four possible outcomes are summarized in Table 5.2.

| | $H_1$ is true | $H_0$ is true |
|---|---|---|
| Reject $H_0$, accept $H_1$ | Correct decision | Type I error (false positive) |
| Accept $H_0$, reject $H_1$ | Type II error (false negative) | Correct decision |

Table 5.2 *Hypothesis testing outcomes table.*

The question is then how to choose the rejection region in order to maximise the power for a fixed probability of type I error (i.e., fixed significance level). This is the content of the Neyman–Pearson lemma, which says that the best choice of test statistics is the likelihood ratio.

**Theorem 5.1** (Neyman–Pearson Lemma) *For data $\boldsymbol{X}$ and test statistic $TS(\boldsymbol{X})$, define the likelihood ratio test as:*

$$\Lambda(\boldsymbol{X}) := \frac{\Pr\left(TS(\boldsymbol{X}) \mid H_1 \text{ true}\right)}{\Pr\left(TS(\boldsymbol{X}) \mid H_0 \text{ true}\right)}\,. \tag{5.37}$$

*The rejection region $R_\alpha$ is defined from:*

$$R_\alpha = \{TS(\boldsymbol{X}) : \Lambda(\boldsymbol{X}) > c_\alpha\} \tag{5.38}$$

*with $c_\alpha$ chosen so that the test has size $\alpha = \Pr\left(TS(\boldsymbol{X}) \in R_\alpha \mid H_0 \text{ true}\right)$. Then, among all tests of size $\tilde{\alpha} \leq \alpha$, the likelihood ratio test is the most powerful one, i.e.*

$$\Pr\big(TS(\boldsymbol{X}) \in \tilde{R}_{\tilde{\alpha}} \mid H_1 \text{ true}\big) \leq \Pr\left(TS(\boldsymbol{X}) \in R_\alpha \mid H_1 \text{ true}\right), \tag{5.39}$$

*where $\tilde{R}_{\tilde{\alpha}}$ is the rejection region corresponding to the alternative test of size $\tilde{\alpha}$.*

When the hypotheses are composite (i.e, they have free parameters), one should instead use the **generalized likelihood ratio test**, with the quantity:

$$\Lambda(\boldsymbol{X}) = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} \Pr\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)}{\sup_{\boldsymbol{\theta}\in\Theta_1} \Pr\left(\boldsymbol{x} \mid \boldsymbol{\theta}\right)}$$

replacing the likelihood ratio, where $\Theta_i$ $(i = 0, 1)$ are the parameter spaces for model $i$.

[7] For a composite hypothesis, the level of significance is the maximum probability of committing a type I error, maximised over the parameter values that are considered under the null.

**Example 5.2: Likelihood ratio test for Gaussian**

Consider the Gaussian model with i.i.d. data drawn from a Gaussian: $X_i \sim \mathcal{N}(\mu, \sigma^2)$, with fixed and known $\sigma$, $i = 1, \ldots . n$. Under the null hypothesis, $H_0 : \ \mu = 0$; under the alternative hypothesis, $H_1 : \mu = \mu_1$. Defining $t = \sum_{i=1}^{n} x_i$, we have that under the null $t \sim \mathcal{N}(0, n\sigma^2)$. We reject $H_0$ if the likelihood ratio exceeds $c_\alpha$, i.e.:

$$\Lambda(x_1, \ldots, x_n) = \exp\left(\frac{\mu_1}{\sigma^2} t - \frac{n\mu_1^2}{2\sigma^2}\right) > c_\alpha \,.$$

Taking the log, this condition translates to:

$$t > \frac{\sigma^2}{\mu_1} \ln(c_\alpha) + \frac{n\mu_1}{2} \,.$$

The critical threshold $c_\alpha$ is set such that the test has a significance level $\alpha$. Standardizing $z = t/(\sqrt{n}\sigma)$ we have that $z \sim \mathcal{N}(0, 1)$, and the type I error is

$$\Pr(Z > z_\alpha \mid H_0 \text{ true}) = \alpha,$$

The $\alpha$-quantile of the standard normal distribution determines the critical value, which gives $c_\alpha = z_\alpha \sqrt{n}\sigma$. This threshold ensures the test controls the Type I error rate at $\alpha$.

### *5.3.2 Meaning of $p$-values*

We are led to the topic of the infamous ***p*-value**: this is the probability, *assuming the null hypothesis to be true*, of obtaining a value of the test statistics $TS(\boldsymbol{X})$ as extreme or more extreme than the one that has been observed, $TS^{\text{obs}} \equiv TS(\boldsymbol{x}^{\text{obs}})$:

$$p\text{-value} = \Pr\Big(TS(\boldsymbol{X}) \geq TS^{\text{obs}}) \mid H_0 \text{ true}\Big) \,. \tag{5.40}$$

Comparing with the definition of type I error, Eq. (5.34), we see that the $p$-value is the upper bound to the significance level $\alpha$ for which the null would be rejected. This is important, since the special quality of it being the *largest* value at which the null would just be rejected for a significance level $\alpha$ makes it identifiable among all the values that would reject $H_0$ at that same pre-determined significance level. This is at the heart of the disagreement between the Neyman-Pearson approach to testing (where the threshold for rejection, $\alpha$, is fixed beforehand, and the null is rejected *at the $\alpha$ significance level* irrespective of how far below the $p$-value falls with respect to the pre-determined significance level)

 

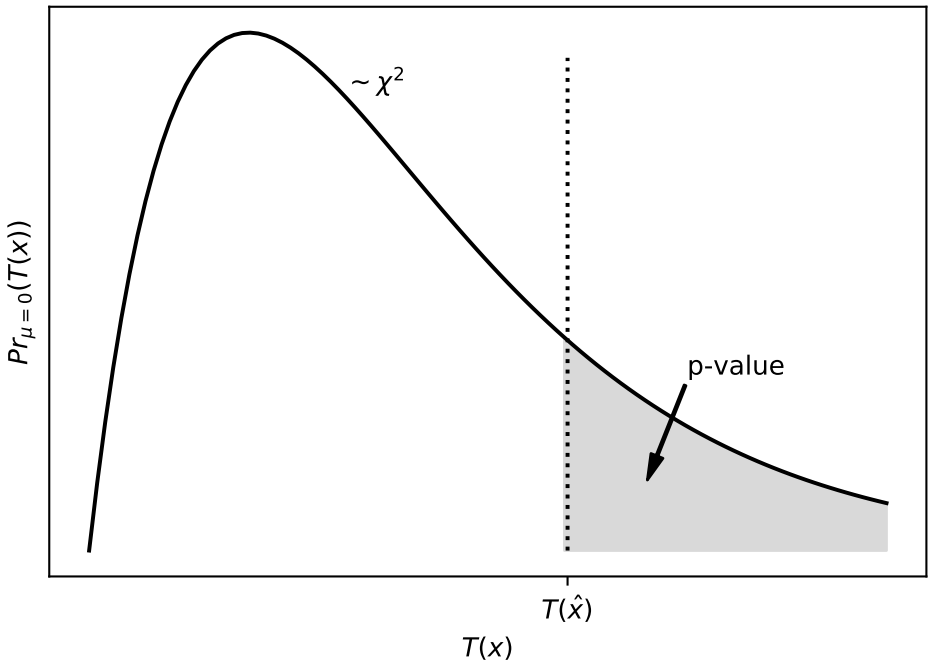


Figure 5.7 Definition of the $p$-value as the tail probability of the test statistics' distribution (shaded area), above the observed value (dashed vertical line).

and that advocated by Fisher (who argued that one should report the *observed* $p$-value instead, and then reject the null at any convenient significance level, provided that $\alpha \geq p$-value), see Berger (2003); Greenland et al. (2016) for details of this important distinction. In particular, while Neyman-Pearson testing controls the type I error, Fisher testing does not.

We introduced $p$-values by calling them infamous because, even though they are ubiquitous in the sciences, they are often used improperly and without a clear understanding of their meaning. Indeed, Jeffreys (1961) wrote that "$p$-values are absurd [since] a model is rejected [under a small $p$-value] for not predicting values of the test statistics that have not been observed". The review article by Greenland et al. (2016) lists and corrects 25 common misinterpretations concerning $p$-values (see also Wasserstein and Lazar (2016)). Among the most widespread misconceptions worth recalling (and dispelling), we mention:

1 The $p$-value is the probability that the null is true. Incorrect! Instead, it *assumes* that the null is true.

2 The $p$-value is the probability that chance *alone* has produced the data. In-

 

correct! It is the probability that the data are more extreme than observed, *assuming chance alone*.

3 A significant test (e.g., $p < 0.05$) means the null is false. Incorrect! It only says that the data are unusual *if all the assumptions* (including $H_0$ being true) *are correct*.

4 A large $p$-value is evidence in favour of $H_0$. Incorrect! It is merely a failure to reject.

5 The $p$-value is the probability of the data to occur if $H_0$ is true. Incorrect! This goes against its definition as a tail probability.

These are problems stemming from improper use and misunderstanding, and thus can be corrected with careful education. However, $p$-values are clearly counterintuitive, and more generally classical hypothesis testing entails other conceptual problems that are difficult to address. These include $p$-hacking, i.e., multiple, repeated testing until a 'highly significant' result is found (and then published) without mentioning the non-significant tests (this is sometimes called 'the look-elsewhere effect'); the stopping rule paradox, i.e., the fact that the outcome of a hypothesis test depends on what other data the researcher thought could have been produced (but weren't), as the $p$-value depends on the quantity one regards as the random variable; and ultimately the fact that hypothesis testing does not give us the answer to the scientific question we asked, that is, i.e., it cannot provide the probability of the hypothesis given the data.

### *5.3.3 Evidence upper bounds for the extended model*

As we have seen in Eq. (5.33), the Bayes factor for nested models scales irreducibly with the prior width of the extended model's additional parameter, so the Occam's razor term in favour of $M_0$ can be made arbitrarily large, in principle, by making the prior on the additional parameter arbitrarily diffuse. However, the reverse is not true: we could imagine selecting a prior on the additional parameter in model $M_1$ such as to maximise the evidence in favour of that model – this would give us a lower bound on $\ln B_{01}$, with no other choice of prior (within a certain restricted family of prior density) able to give larger support for the alternative model. This is a useful approach to find prior-independent lower evidence bounds to compare with the evidence against the null represented by $p$-values (Sellke et al., 2001).

The intuition about this procedure can be obtain by considering a horizontal slice across Fig. 5.4 (i.e., for constant $\lambda$): for a fixed likelihood variance (i.e., for fixed number of data points), this is equivalent to varying the standard deviation of the Normal prior at fixed significance level. This is shown in Fig. 5.8,

 

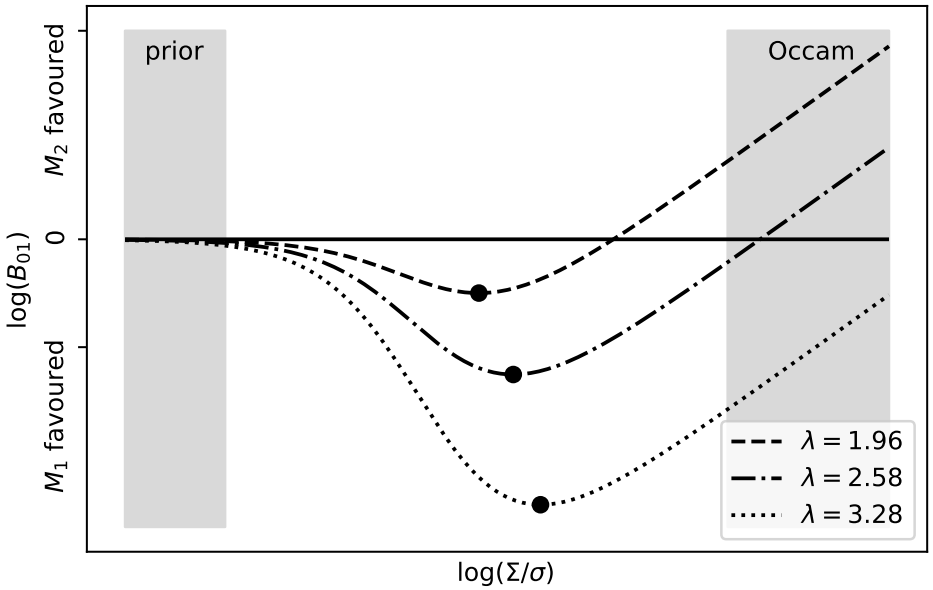


Figure 5.8 Maximum evidence against the null hypothesis in a Bayesian setting. For a fixed significance level (i.e., fixed $\lambda$, here chosen to correspond to 5%, 1% and 0.1% significance against the null from top to bottom), there is a value of the prior width (horizontal axis) that maximises the evidence agains the null (indicated in each case by the black dot). The "prior-dominated" region shaded at the left corresponds to posterior odds of 1 : 1 between the models, while the shaded region labeled "Occam" on the right shows the linear scaling of the log-evidence with the information gain in the Occam's razor regime.

where the vertical axis is the resulting log Bayes factor, and the horizontal axis the information gain. For a given choice of $\lambda$, there is a choice of prior width $\Sigma$ (for fixed likelihood width $\sigma$) that maximises the evidence in favour of model 1 (indicated by the black dot in each case). If even this prior choice, maximally favourable to the extended model, does not lead to a strong evidence against model 0, one can be confident that *no other prior choice in the same parametric family* (in this case, a Gaussian) will lead to a larger evidence against model 0.

Before determining a quantitative measure of such an evidence lower bound, let us find the simpler absolute lower bound to the Bayes factor, i.e., we would like to choose the prior for the additional parameter in such a way as to give the maximum possible evidence against the null hypothesis. This is obviously obtained by setting the 'prediction' for the alternative model to be exactly what has been observed, i.e., by apportioning the whole prior density to a delta function centred at the MLE: $\Pr(\theta|M_1) = \delta(\theta - \theta_{\mathrm{MLE}})$.

 

| $p$-value | $\lambda$ | $\underline{\underline{B}}_{01}$ | $\underline{\underline{\Pr(M_0 \mid \boldsymbol{d})}}$ for $\Pr(M_0) = 0.5$ |
|---|---|---|---|
| 0.05 | 1.96 | 0.26 | 0.21 |
| 0.01 | 2.58 | 0.036 | 0.035 |
| 0.001 | 3.28 | 0.005 | 0.005 |
| $3 \times 10^{-7}$ | 5.0 | $3.7 \times 10^{-6}$ | $3.7 \times 10^{-6}$ |

Table 5.3 *Absolute lower bound for the Bayes factor and the posterior probability for the simple model (null hypothesis) from Bayesian model comparison.*

This is obviously not a very Bayesian choice, for we are giving $M_1$ the full advantage of 'predicting' the maximum likelihood value of the parameter *after* we have observed it — a choice that maximally favours the alternative model. Using a Gaussian likelihood, we thus find the following absolute lower bound to the Bayes factor (denoted by a double underline) in this case:

$$\underline{\underline{B}}_{01} = \frac{e^{-\frac{1}{2}\frac{(0-\boldsymbol{\theta}_{\mathrm{MLE}})^2}{\sigma^2}}}{e^{-\frac{1}{2}\frac{(\boldsymbol{\theta}_{\mathrm{MLE}}-\boldsymbol{\theta}_{\mathrm{MLE}})^2}{\sigma^2}}} = e^{-\frac{1}{2}\frac{\boldsymbol{\theta}^2_{\mathrm{MLE}}}{\sigma^2}} = \exp\left(-\frac{\lambda^2}{2}\right). \tag{5.41}$$

Our choice for the alternative model's prior has set the Occam's razor factor to zero. Using this result, we obtain the conversion table 5.3, where we obtained the absolute lower bound for the posterior of the null hypothesis,

$$\underline{\underline{\Pr(M_0 \mid \boldsymbol{d})}} = \left(1 + B_{01}^{-1}\right)^{-1},$$

obtained via Eq. (5.18) and using equal models' priors.

The agreement between the numerical $p$-value and the lower bound on the actual posterior probability tends to improve as we go down the table—showing that high-significance claims do correspond to low probability for the model. However, even in this extreme case, $p$-values are lower than the corresponding lower bound for the posterior probability of the null: a $p$-value of 5%, a common standard in the social sciences, tends to strongly overestimate the evidence against the null. Table 5.3 tells us that a null hypothesis that is rejected with a significance level of 5% has *at least* 21% posterior probability of being true, even having chosen the most unfavourable prior against it.

This can be made more quantitative by computing

$$\underline{B}_{01}(\wp) = -e\wp\ln(\wp), \text{ for } \wp < 1/e \simeq 0.37, \tag{5.42}$$

where $\wp$ above is the $p$-value, and interpret this as a lower bound to $B_{01}$ (a justification is presented below). The corresponding lower bound to the posterior probability for $M_0$ (i.e., in favour of the null) is obtained by using $\underline{B}_{01}(\wp)$, prior

odds of 1 : 1 and Eq. (5.18), obtaining:

$$\underline{\Pr(M_0 \mid \boldsymbol{d})} = \frac{1}{1-(e\wp \ln \wp)^{-1}}\,, \text{ for } \wp < 1/e \simeq 0.37, \tag{5.43}$$

In this way, we obtain the calibration between $p$-values and lower bounds for the posterior in favour of the null given in Table 5.4.

Another very interesting result, proved in Berger and Sellke (1987), shows that

$$\lim_{\lambda\to\infty} \underline{\Pr(M_0 \mid \boldsymbol{d})}/(\wp\lambda) = \sqrt{\frac{\pi}{2}} = 1.25\,, \tag{5.44}$$

where $\wp$ is the $p$-value. This means that the posterior probability for the null is *always* at least $(1.25) \times (\wp\lambda)$ for any prior. So, even for a $5\sigma$ detection, for which $\wp = 3\times10^{-7}$, the null has a posterior probability of at least $1.9\times10^{-6}$ — an order of magnitude larger than a naive mis-interpretation of the $p$-value might suggest.

| $\wp$ | $\underline{B}_{01}(\wp)$ | $\underline{\Pr(M_0 \mid \boldsymbol{d})}$ |
|---|---|---|
| 0.05 | 0.407 | 0.29 |
| 0.01 | 0.125 | 0.109 |
| 0.001 | 0.0188 | 0.018 |

Table 5.4 *Calibration table for Bayes factors. The p-values is given by $\wp$, the lower value of the calibrated Bayes factor by $\underline{B}_{01}(\wp)$, and the lower bound to the posterior probability of the null hypothesis, obtained by maximising the evidence against the null over all prior in the class of unimodal and symmetric priors.*

The last column reports another lower bound, obtained by maximising the evidence against the null over all prior in the class of unimodal and symmetric priors . The rule of thumb that we gain from this comparison is that we should downgrade the number of sigma evidence by one unit to align the intuition with the Bayesian lower bound calibration. This is illustrated in Fig. 5.9.

In closing, we give a simplified derivation for the calibration given by Eq. (5.42). Under the null, the distribution of the $p$-values is Uniform$(0,1)$. We now take this to define the null

$$M_0 : \wp \sim \text{Uniform}(0,1)\,. \tag{5.45}$$

The alternative is therefore defined by the $p$-value having another distribution

$$M_1 : \wp \sim f_\psi(x)\,, \tag{5.46}$$

where $\psi$ parametrises the family of distributions $f_\psi$. Choosing a test statistics

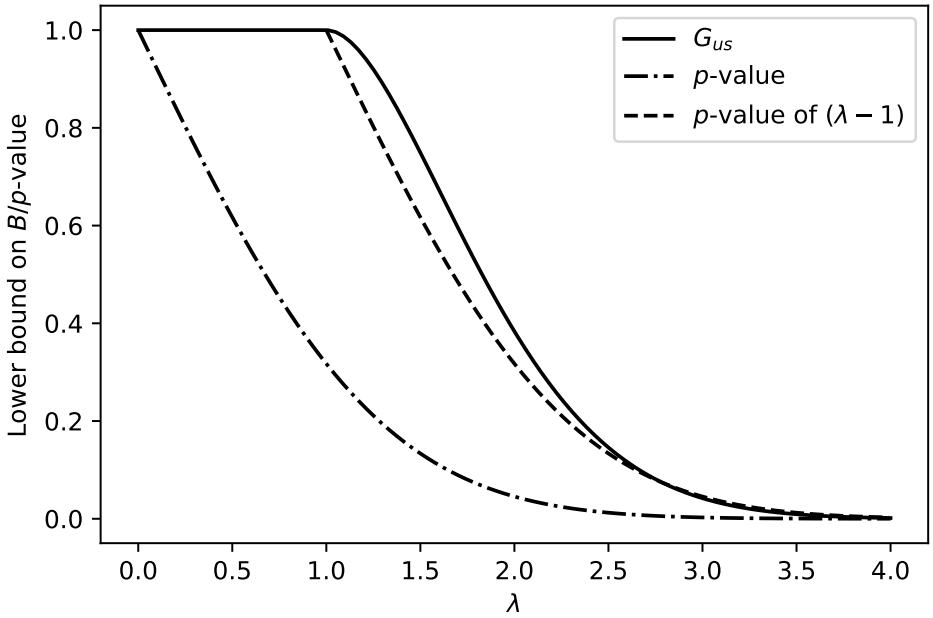


Figure 5.9 Lower bound to the Bayes factor $B_{01}$, obtained from maximising the evidence against the null over all prior in the class of unimodal and symmetric priors ($G_{us}$ line). This is compared to the $p$-value and to the $p$-value shifted by one standard deviation.

$TS$ so that large values of $TS$ are in favour of $M_1$, it follows that $f_\psi$ should be decreasing in $x$. So we can parametrise

$$f_\psi(x) = \text{Beta}(\psi, 1) = \psi x^{\psi-1} \quad \text{for} \quad 0 < \psi < 1 \tag{5.47}$$

(the uniform distribution is obtained for $\psi = 1$).

Then the Bayes factor is

$$B_{01} = \frac{f_{\psi=1}(x)}{\int f_\psi(x) \Pr(\psi) \, \mathrm{d}\psi} \,. \tag{5.48}$$

The supremum is obtained by setting $\Pr(\psi) = \delta(\psi - \psi_{\text{MLE}})$. The MLE for the parameter defining the Beta distribution is $\psi_{\text{MLE}} = \ln x$ for $x < 1/e$ and 1 otherwise, so (after some algebra)

$$\underline{B_{01}} = \frac{1}{f_\psi(x)} = \frac{1}{\psi x^{\psi-1}}\bigg|_{\psi_{\text{MLE}}} = \frac{1}{x^{\ln x - 1} \ln x} = -ex \ln x \,. \tag{5.49}$$

Replacing this result into the expression for $\Pr(M_0 \,|\, \boldsymbol{d})$, and using 1:1 prior odds,

we obtain:

$$\underline{\Pr(M_0 \mid \boldsymbol{d})} = \frac{1}{1 - (ex \ln x)^{-1}}, \tag{5.50}$$

where $x$ is the observed $p$-value.

## 5.4 Other approaches to BF prior selection

There are several other approaches to selecting the prior for Bayes factors calculation — most are reviewed in Consonni et al. (2018). The bottom line is that no single approach works in all circumstances. And model checking remains important: checking residuals, predictive distributions, predicted data (from the model) are all important tools that can lead the scientist to discover problems between the model and the data in a way that a formal approach might not.

1 **Unit information prior** (we will see more when we will talk about BIC).
2 **Intrinsic BF (IBF):** Use a subset of the data to set the prior. This approach is conceptually not well motivated and it is difficult to select the subset, see Berger and Pericchi (1996).
3 **Fractional BF (FBF):** Use a fraction of the data by raising the likelihood $\mathscr{L} \rightarrow \mathscr{L}^b$.
4 **Conventional priors:** use default priors for certain likelihoods.

## 5.5 Computational methods

Computing the evidence, Eq. (5.6), is often challenging, as it involves a multidimensional integral over parameter space. Fortunately, a series of tools are available, with various levels of approximation and sophistication, depending on the problem at hand.

The **Laplace approximation** replaces the posterior with a Normal distribution centered at the MAP:

$$\Pr(\boldsymbol{\theta} \mid \boldsymbol{d}) \approx \mathcal{N}\left(\boldsymbol{\theta}_{\text{MAP}}, \hat{\boldsymbol{\Sigma}}^{-1}\right) \tag{5.51}$$

where $\hat{\boldsymbol{\Sigma}}^{-1}$ is the observed Fisher matrix of the posterior. The normalization constant – corresponding to the evidence – is thus known analytically:

$$\Pr(\boldsymbol{d}) = \int \Pr(\boldsymbol{d} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \simeq (2\pi)^{\frac{d}{2}} \left|\hat{\boldsymbol{\Sigma}}\right|^{1/2} \mathscr{L}\left(\hat{\boldsymbol{\theta}}\right) \pi\left(\hat{\boldsymbol{\theta}}\right) \tag{5.52}$$

where $d = \dim(\boldsymbol{\theta})$. The approximation is accurate to $\mathcal{O}\left(n^{-1}\right)$ for $n \rightarrow \infty$, see Kass and Wasserman (1995, Sec. 4.1) for details, as well as so-called Bartlett corrections to go beyond the Gaussian approximation. See Diciccio et al. (1997). It

works for nested as well as non-nested models, but its accuracy could be poor (and you might not be able to assess its error without comparing with some other method).

### *5.5.1 Bayesian Information Criterion (BIC)*

The Bayesian Information Criterion is a "shortcut" for the computation of the evidence, subject to specific assumptions and a sort of "default prior" choice. It is also known as the Schwarz (1978) criterion, after the name of its proposer, who derived it for i.i.d. observations and linear models, further assuming that the likelihood is from the exponential family. While the BIC can be a crude asymptotic approximation to the actual Bayesian evidence, it is asymptotically consistent, and very simple to compute. For more details about the derivation, see e.g. Neath and Cavanaugh (2011).

We approximate the log-likelihood to quadratic order in the parameters, with $\hat{\boldsymbol{\Sigma}}^{-1}$ the Hessian of the negative log-likelihood,

$$\hat{\Sigma}^{-1}_{kk'} = -\left.\frac{\partial^2 \ln \mathscr{L}}{\partial\theta_k \partial\theta_{k'}}\right|_{\boldsymbol{\theta}_{\mathrm{MLE}}} . \tag{5.53}$$

so that, to second order:

$$\ln \mathscr{L} \approx \ln \mathscr{L}(\boldsymbol{\theta}_{\mathrm{MLE}}) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MLE}})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_{\mathrm{MLE}}), \tag{5.54}$$

For $n$ i.i.d. data points, the Hessian can be rewritten as $n\cdot\overline{\boldsymbol{I}}$, where $\overline{\boldsymbol{I}}$ is the average observed Fisher matrix for one datum. Using the Laplace approximation, a second-order expansion of the log-posterior (see Eq. (5.51)), we get

$$\Pr(\boldsymbol{d}) \simeq (2\pi)^{\frac{d}{2}} \left|n\overline{\boldsymbol{I}}\right|^{-1/2} \mathscr{L}(\boldsymbol{\theta}_{\mathrm{MLE}}) \Pr(\boldsymbol{\theta}_{\mathrm{MLE}}) , \tag{5.55}$$

or, after taking the logarithm:

$$2\ln\Pr(\boldsymbol{d}) = d\ln(2\pi) - \ln\left|n\overline{\boldsymbol{I}}\right| + 2\ln\mathscr{L}_{\mathrm{MLE}} + 2\ln\Pr(\boldsymbol{\theta}_{\mathrm{MLE}}) , \tag{5.56}$$

where

$$\ln\left|n\overline{\boldsymbol{I}}\right| = d\ln(n) + \ln\left|\overline{\boldsymbol{I}}\right| ; \tag{5.57}$$

asymptotically (for $n\to\infty$) we can drop all terms that do not scale with $n$; in this limit we can also assume a Jeffreys' prior $\Pr(\boldsymbol{\theta}_{\mathrm{MLE}}) = 1$, and obtain an asymptotic approximation for the evidence as:

$$2\ln\Pr(\boldsymbol{d}) \to 2\ln\mathscr{L}_{\mathrm{MLE}} - d\ln(n) \text{ for } n\to\infty. \tag{5.58}$$

The negative of this quantity is called the BIC (Bayesian Information Criterion), minimizing which is equivalent to maximising the evidence in the above limit:

$$\mathrm{BIC} \equiv -2\ln \mathscr{L}_{\mathrm{MLE}} + d\ln(n)\,, \tag{5.59}$$

The 'best' model is therefore one that minimises the BIC, in which every additional parameter space dimension incurs a penalty term that depends only on the logarithm of the number of data points.

Denoting by $\Delta_{01} = \mathrm{BIC}(M_0) - \mathrm{BIC}(M_1)$, for $n \to \infty$ we have that this quantity converges to the Bayes factor between the two models:

$$\frac{-2\log B_{01} - \Delta_{01}}{-2\log B_{01}} \to 0 \tag{5.60}$$

but with error of $\mathscr{O}(1)$ (Kass and Wasserman, 1995). Kass and Wasserman (1995) also show convergence can be improved to to $\mathscr{O}\left(1/\sqrt{n}\right)$ if one is willing to assume as a prior the distribution:

$$p(\boldsymbol{\theta}) = \mathscr{N}(\boldsymbol{\theta}_*, \bar{\boldsymbol{I}})\,, \tag{5.61}$$

where $\boldsymbol{\theta}_*$ are the parameters' value for which $M_1$ reduces to $M_0$—i.e. one assumes the same amount of prior information as the average information in one datum, which can be seen as a kind of reference prior. See Liddle (2007) for further details and a comparison with the Akaike Information Criterion (AIC).

### *5.5.2 Nested sampling*

Nested sampling was originally proposed by Skilling (2004), and is more fully explained in Skilling (2006). This algorithm has become very popular because it allows to obtain posterior samples without much tuning by the users, also for multi-modal and curved posteriors, as well as evidence estimation at the same time. In particular *multimodal nested sampling* (Feroz and Hobson, 2008) has become the de-facto standard in astrophysics and cosmology. It works well for parameter spaces of up to ~30 dimensions, and is often a more efficient alternative to MCMC. Many other implementations exists, including `dynesty` (Speagle, 2020), using a dynamical number of live points (as introduced by Higson et al. (2019)), `nessai` (Williams et al., 2021), which uses normalizing flows, and `PolyChord` (Handley et al., 2015), which uses slice sampling and scales to larger number of dimensions (see **?** for a review).

The fundamental idea behind nested sampling is to transform the multidimensional evidence integral into a one-dimensional integral via a change of variables. Consider the prior volume element:

$$\mathrm{d}X = \Pr(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}, \text{ with } \dim(\boldsymbol{\theta}) = d\,. \tag{5.62}$$

 

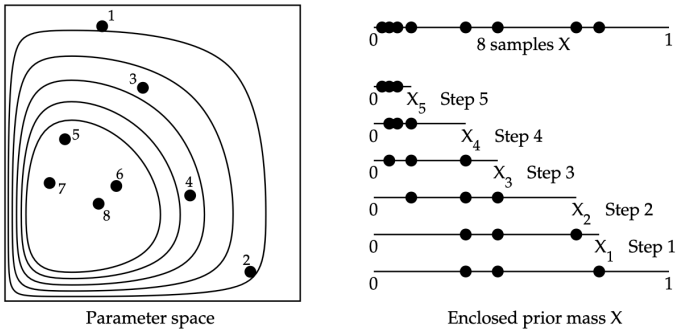


Figure 5.10 Illustration of nested sampling. Left: 2-dimensional parameter space, with lines indicating iso-likelihood contours. The box is the boundary of the uniform prior, and the filled circles give the location of the $n_{\text{live}} = 3$ live points, with numbers indicating the sequence in which they are produced. Right: the compression of the prior volume fraction $X$ as sampling progresses, from bottom to top. Figure reproduced from Skilling (2006) with kind permission from the International Society for Bayesian Analysis (ISBA).

We are going to assume that the prior $\Pr(\boldsymbol{\theta})$ can be recast, with a suitable parameter transformation, into a $d$-dimensional hypercube of unit side (see Alsing and Handley (2021) for a method to transform any prior into such a form). The quantity

$$X(\lambda) = \int_{\mathscr{L}(\boldsymbol{\theta})>\lambda} \Pr(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{5.63}$$

is the integral of the prior above an iso-likelihood contour defined by the value $\lambda$. Now, let $L(X)$ be the inverse of $X(\lambda)$, then

$$\Pr(\boldsymbol{d}) = \int \mathscr{L}(\boldsymbol{\theta}) \Pr(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} = \int_0^1 L(X)\mathrm{d}X, \tag{5.64}$$

which is a one-dimensional integral. The value $X = 1$ corresponds to the whole prior distribution, while for $X \to 0$ the prior shrinks to the point where the likelihood peaks.

 

If we can evaluate $L_j = L(X_j)$ for a sequence $0 < X_M < \cdots < X_1 < X_0 = 1$; then the evidence is given by

$$\Pr(\boldsymbol{d}) = \sum_{i=1}^{M} w_i L_i \quad \text{with} \tag{5.65}$$

$$w_i = \frac{1}{2}(X_{i-1} - X_{i+1}) \quad \text{(trapezoidal rule)}. \tag{5.66}$$

The nested sampling algorithm can be described as follows:

1 While at $X_0$ (i.e., the full prior volume, usually taken as a hypercube of side 1), lay down $n_{\text{live}}$ 'live points' by drawing $\boldsymbol{\theta}_i \sim \Pr(\boldsymbol{\theta})$, $i = 1, \ldots, n_{\text{live}}$; a typical value for $n_{\text{live}}$ is between 400 and a few thousands (the larger the value, the slower the algorithm).
2 Evaluate $\mathscr{L}_i = \mathscr{L}(\boldsymbol{\theta}_i)$ for at all points, $i = 1, \ldots, n_{\text{live}}$.
3 Remove the point with the smallest likelihood value, $\mathscr{L}_* = \min_i \mathscr{L}_i$, and replace it with a new point drawn uniformly from the prior, subject to the hard constraint that for this new point $\mathscr{L}(\boldsymbol{\theta}_{i+1}) > \mathscr{L}_*$. This is the most difficult part of the algorithm, as performing efficiently this constrained sampling is in general difficult: we don't know where the iso-likelihood contours run in parameter space. Many solutions exist, including performing MCMC, approximating the remaining prior region with overlapping ellipsoids (MultiNest's solution), performing slice sampling, and others.
4 Set $X_i = \exp(-i/n_{\text{live}})$, $w_i = (X_{i-1} - X_{i+1})/2$, and increment the estimated evidence, $Z$, by $w_i \mathscr{L}_*$.
5 Go back to 3. until the stopping criterion is met.

At each iteration, the prior volume is successively shrunk in the constraining sampling step. While we don't know the exact value of the new prior volume fraction enclosed by the replaced live point, it can be estimated probabilistically. The prior fraction $X_i$ enclosed by the new live point is given by:

$$X_i = t_i \cdot X_{i-1}, \tag{5.67}$$

where the compression factor $t_i$ is a random variable that follows a Beta($n_{\text{live}}$, 1) distribution, as it describes the probability of selecting the largest between $n_{\text{live}}$ random numbers (the live points), each uniformly distributed between $[0,1]$:

$$\Pr(t_i) = n_{\text{live}} t_i^{n_{\text{live}}-1}. \tag{5.68}$$

Given the above recursive relation, it is clear that $X_i = t_i t_{i-1} \ldots t_1 X_0$ shrinks geometrically towards the peak of the likelihood, thus rapidly traversing large regions of the prior where the likelihood is small and thus contributes little to the

evidence integral. In terms of $\log t_i$, at each step the expectation value is

$$\mathbb{E}\left[\log t_i\right] = -\frac{1}{n_{\text{live}}} \tag{5.69}$$

with standard deviation

$$\text{Var}\left[\log t_i\right] = \frac{1}{n_{\text{live}}}\,. \tag{5.70}$$

Therefore, after $i$ independent steps each of expected log-width $1/n_{\text{live}}$, the prior mass fraction will have shrunk to the expected value (see Fig. 5.10 for an illustration):

$$\mathbb{E}\left[\log X_i\right] = -\frac{i \pm \sqrt{i}}{n_{\text{live}}}\,. \tag{5.71}$$

This means that, in expectation, $X_i = \exp(-i/n_{\text{live}})$. The stopping criterion is given by selecting a maximum tolerance, $\Delta$, in the evidence increase at each iteration. The maximum remaining contribution to the evidence is approximately bound by the value $\mathscr{L}_{\text{max},i} X_i$, where $\mathscr{L}_{\text{max},i}$ is the maximum likelihood value among the remaining live points at step $i$. With this approximation, the algorithm stops when

$$\mathscr{L}_{\text{max},i} X_i < \Delta\,. \tag{5.72}$$

The same procedure also delivers posterior samples, obtained from the live points that are replaced in the process, with importance weights $W_i$ given by:

$$W_i = \frac{\mathscr{L}_i w_i}{Z}, \quad \sum_i W_i = 1\,, \tag{5.73}$$

where $Z$ is the final estimate of the evidence once the run has finished.

As far as the user is concerned, there are only two tunable parameters:

- $n_{\text{live}}$, the number of live points to work with (typically chosen in the range $\sim 400 - 1000$); the higher the number, the more accurate (as the stepping in is smaller) the posterior exploration and evidence estimation, but also the slower the algorithm; in dynamic nested sampling, the number of live points is changed dynamically as the nesting progresses (Higson et al., 2019; Speagle, 2020), in order to shrink less (and therefore be more accurate) in regions of high posterior density or where the contribution to the evidence integral is largest.
- The evidence tolerance, $\Delta$, which determines the stopping threshold.

The replacement step may fail when the likelihood exhibits flat plateaus, so that it is difficult or impossible to find a new point with a strictly larger value of

 

the likelihood than the one that is being replaced, a problem which can be addressed with a simple modification to the algorithm (Schittenhelm and Wacker, 2020; Fowlie et al., 2020b). Checks of the statistics of replacement points can also be conducted in order to verify whether the nested sampling run successfully sampled from the posterior (Fowlie et al., 2020a).

### *5.5.3 Other methods*

Several other approached are available for the numerical computation of the Bayesian evidence and/or the Bayes Factor. We will give below a selection; which tool to choose depends on the problem at hand, the computational effort and the purpose of the comparison. Note however that in general the in-built prior dependency means high-accuracy evidence computations is usually not required: there is not point in computing a Bayes factor to several digits numerical accuracy, when a mild change of prior can produce a much larger shift in value.

*Savage-Dickey Density Ratio* As we have seen in Eq. (5.32), this expression is useful for nested models and separable priors, and it can be evaluated from the marginal posterior for the additional parameters only, which should be possible to evaluate from MCMC samples. However, one must be careful to sample with sufficient accuracy in the tails of the distribution in order to obtain a reliable estimate of the density at the point $\omega_*$, when that value is located away from the mode. This is the case if

$$\left|\frac{\omega_* - \langle\omega\rangle}{\sigma_\omega}\right| \gg 1\,, \tag{5.74}$$

where $\langle\omega\rangle$ is the posterior mean of the additional parameter $\omega$, and $\sigma_\omega^2$ its posterior variance. This is illustrated in Fig. 5.11. See Trotta (2007b, Appendix C) for further discussion of this point.

*Bridge sampling* This is a general version of importance sampling , introduced by Bennett (1976); Meng and Wong (1996). Denote the posteriors for the two models being compared by $s_i = t_i/c_i$, where

$$t_i = \mathscr{L}_i \pi_i \ (i = 1,2) \quad \text{unnormalised posterior, and} \tag{5.75}$$

$$c_i = \int t_i \mathrm{d}\boldsymbol{\theta}_i \quad \text{evidence.} \tag{5.76}$$

Let $\gamma(\boldsymbol{\theta})$ be a function so that

$$0 < \left|\int \gamma(\boldsymbol{\theta}) s_1(\boldsymbol{\theta}) s_2(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}\right| < \infty\,. \tag{5.77}$$

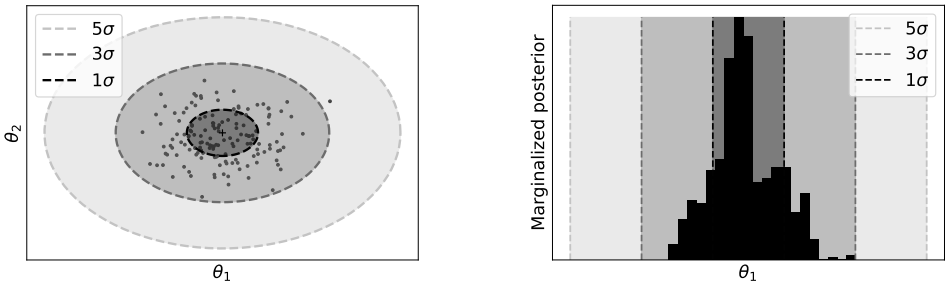


Figure 5.11 Illustration of the difficulty of accurately sampling far into the tails of a Gaussian distribution. Standard Metropolis-Hastings MCMC does not usually reach past $\sim 3\sigma$ in the tails (with performance worsening with increasing dimensionality of the parameter space) which makes an accurate estimate of the marginal posterior density entering in the SDDR challenging if $\omega_*$ is far into the tails.

Then we can write

$$\begin{aligned} B_{12} = \frac{c_1}{c_2} &= \frac{c_1}{c_1}\frac{\int s_1 \gamma s_2 \mathrm{d}\boldsymbol{\theta}}{\int s_1 \gamma s_2 \mathrm{d}\boldsymbol{\theta}} \\ &= \frac{c_1}{c_1}\frac{\int t_1/c_1 \gamma s_2 \mathrm{d}\boldsymbol{\theta}}{\int s_1 \gamma t_2/c_2 \mathrm{d}\boldsymbol{\theta}} \\ &= \frac{\int t_1 \gamma s_2 \mathrm{d}\boldsymbol{\theta}}{\int s_1 \gamma t_2 \mathrm{d}\boldsymbol{\theta}}. \end{aligned} \tag{5.78}$$

Now, write $t_1 = h = \mathscr{L}_1 \pi_1$, $c_1 = c$. Choose for $t_2$ a convenient, analytically normalizable density $q(\boldsymbol{\theta})$ that is an approximation to the density of interest, $t_1$ (for example, a Gaussian) and set $t_2 = q$, so that (as $q$ is already normalized) $c_2 = 1$; then

$$\frac{c_1}{c_2} = c = \frac{\int h(\boldsymbol{\theta})\gamma(\boldsymbol{\theta})q(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}{\int q(\boldsymbol{\theta})\gamma(\boldsymbol{\theta})\Pr(\boldsymbol{\theta}\,|\,\boldsymbol{d})\,\mathrm{d}\boldsymbol{\theta}}. \tag{5.79}$$

Since $q$ is a simple density, we can easily draw samples $\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_M$ from $q(\boldsymbol{\theta})$; also, suppose we have samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ (obtained e.g. by MCMC) from $\Pr(\boldsymbol{\theta}\,|\,\boldsymbol{d})$. Hence

$$c = \frac{\langle h\gamma \rangle_q}{\langle q\gamma \rangle_{\Pr(\boldsymbol{\theta}\,|\,\boldsymbol{d})}} \simeq \frac{\frac{1}{M}\sum_i h\left(\tilde{\boldsymbol{\theta}}_i\right)\gamma\left(\tilde{\boldsymbol{\theta}}_i\right)}{\frac{1}{N}\sum_j q\left(\boldsymbol{\theta}_j\right)\gamma\left(\boldsymbol{\theta}_j\right)}. \tag{5.80}$$

So far we haven't specified $\gamma$: different choices determine different schemes. For instance, setting $\gamma = q^{-1}$, the above expression simplifies to:

$$c = \frac{1}{M}\sum_i \frac{h\left(\tilde{\boldsymbol{\theta}}_i\right)}{q\left(\tilde{\boldsymbol{\theta}}_i\right)}, \tag{5.81}$$

 

i.e. ordinary importance sampling; in this case, the main issue is to reduce the variance in the ratio. An optimal choice (Meng and Wong, 1996) is

$$\gamma \propto \left[ \frac{Nh(\boldsymbol{\theta})}{c} + Mq(\boldsymbol{\theta}) \right]^{-1}, \tag{5.82}$$

which is to be understood iteratively since it contains the quantity $c$ we want to estimate. Other choices are

- local importance sampling,
- reciprocal importance sampling,

see Diciccio et al. (1997) for a review.

*Density estimation* In density estimation the evidence is computed from the ratio

$$\Pr(\boldsymbol{d}) = \frac{\mathscr{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\Pr(\boldsymbol{\theta} \mid \boldsymbol{d})} . \tag{5.83}$$

This requires to estimate the normalized denominator of the ratio, e.g. via kernel density estimation methods, which becomes more and more difficult as the dimensionality of the parameter space increases.

Heavens et al. (2017) proposed a method to estimate instead the constant of proportionality $a$ between the MCMC sample density, $n(\boldsymbol{\theta} \mid \boldsymbol{d})$ and the unnormalized posterior, $h(\boldsymbol{\theta}) \equiv \mathscr{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, which leads to an estimate of the evidence. Writing $h = an(\boldsymbol{\theta} \mid \boldsymbol{d})$, we have that

$$\Pr(\boldsymbol{d}) = \int h = a \int n(\boldsymbol{\theta} \mid \boldsymbol{d}) = aN , \tag{5.84}$$

where $N$ is the length of the MCMC chain. The constant can be estimated from $a = h/n$, by using a $k$ nearest-neighbour method to estimate the density of samples $n(\boldsymbol{\theta} \mid \boldsymbol{d})$. This method can be applied at the post-processing stage of any MCMC chain, and it works with an accuracy of about one percent for a parameter space of up to 10 dimensions.

### Exercises

5.1 Consider again the coin tossing problem, but this time from the point of view of Bayesian model comparison: we wish to compare a model $M_0$ of a fair coin ($\theta = 1/2$) with an alternative $M_1$, where $\theta$ is a free parameter (with a prior).

1 Using a conjugate Beta($\alpha, \beta$) prior for $\theta$, compute the Bayes factor after

$N$ flips. Starting from equal model probabilities, plot the posterior probability for model $M_1$ as a function of the number of successes $r$ for $N = \{10, 100, 1000\}$ for the following choices of hyperparameters: $\alpha = \beta = 1$ (uniform prior); $\alpha = \beta = 1/2$ (Jeffreys prior); $\alpha = 2, \beta = 1$ (skewed prior).

2 Wilks' theorem says that, asymptotically, the likelihood ratio test statistics

$$\lambda = -2\ln\left(\frac{L(\theta = 1/2)}{L(\theta_{\rm MLE})}\right),$$

is approximately distributed as a $\chi^2$ distribution with one degree of freedom. Use this to construct a hypothesis test for $H_0 : \theta = 1/2$ vs $H_1 : \theta \neq 1/2$. Compare the strength of evidence against $H_0$ as measured by the $p$-value to the posterior model probability for $M_1$ (for the case of a uniform prior) by plotting the latter vs the former for the above cases. Comment on the difference, paying particular attention to whether you believe that the asymptotic limit has been reached (how can you check for that?).

3 To compare the two approaches, proceed as follows. We assume a uniform prior for $\theta$ under $M_1$, and repeat the below procedure for $N = \{10, 100, 1000\}$:

1 generate a pair of 'true' values of $\theta$ under each model (assuming equal prior probabilities for each):

$$\theta_0 \mid H_0 = \frac{1}{2} \quad \text{and} \quad \theta_1 \mid H_1 \sim U(0,1)\,;$$

2 for each $\theta_i$ $(i = 1, 2)$, generate $K = 100$ mock data realizations, $d_{ik} \sim \text{Binomial}(N, \theta_i)$ $(k = 1, \ldots, K)$;

3 for each data realization $d_{ik}$, compute the Bayes factor, $\ln B_{ik}$ and the likelihood ratio, $\Lambda_{ik}$ and save them;

4 repeat steps $[1-3]$ $L = 1000$ times.

5 select a decision threshold, $\ln B_{\rm th}$, along a uniform grid in the range spanned by the distribution of $\ln B_{ik}$; for the likelihood ratio, select a significance value $\alpha$ along a log-uniform grid in the range $\alpha \in [10^{-5}, \ldots, 10^{-1}]$. For each $\theta_i$ pair, use the decision threshold (Bayesian) or significance value (Frequentist) to decide whether to select $M_0$ or $M_1$ (Bayesian) or whether to reject/fail to reject the null (Frequentist) for each data realization.

6 Once you have made the decision for all the $K = 100$ data realizations for a given $\theta_i$ pair, count the number of false positives under $H_0$ (i.e., how many times you wrongly rejected the null when it was true), and the number of true positives under $H_1$ (i.e., how many times you have correctly rejected the null when $H_1$ was true). Average these numbers of the $L$ choices of $\theta_i$. This will give you the false positive rate and the

*expected* true positive rate (i.e., averaged over the prior under $H_1$) for that decision threshold. Save those values.

7 Move to a different decision threshold and repeat steps [5 – 6].

8 Plot the expected true positive rate and the false positive rate for each different decision threshold, both Bayesian and Frequentist. This is the Receiver-operating characteristic (ROC) curve for each method, and the area under the curve (AUC) is a measure of the method's performance: the closer to 1, the better.

9 On the basis of the AUC for each value of $N$, comments on which method is best.

5.2 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 $1.61 \pm 0.40$ arc seconds (based on the position of 5 stars), while the team headed by Crommelin reported $1.98 \pm 0.16$ 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.



 

# Bibliography


Algeri, Sara, Aalbers, Jelle, MorÃě, Knut Dundas, and Conrad, Jan. 2020. Searching for new phenomena with profile likelihood ratio tests. *Nature Reviews Physics*, **2**(5), 245–252.

Alsing, Justin, and Handley, Will. 2021. Nested sampling with any prior you like. *Monthly Notices of the Royal Astronomical Society*, **505**, L95–L99.

Amanullah, R., Lidman, C., Rubin, D., Aldering, G., Astier, P., et al. 2010. Spectra and Light Curves of Six Type Ia Supernovae at 0.511 < z < 1.12 and the Union2 Compilation. *Astrophys.J.*, **716**, 712–738.

Anderson, Lauren, et al. 2014. 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.

Andreon, S., and Hurn, M. A. 2010. The scaling relation between richness and mass of galaxy clusters: a Bayesian approach. *Mon. Not. Roy. Astron. Soc.*, **404**, 1922.

Bartlett, M. S. 1957. A Comment on D. V. Lindley's Statistical Paradox. *Biometrika*, **44**(3/4), 533–534.

Bayes, Thomas, and Price, Richard. 1763. 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. Reproduced in: Biometrika, **45**, 293-315 (1958).

Bennett, C. H. 1976. Efficient Estimation of Free Energy Differences from Monte Carlo Data. *Journal of Computational Physics*, **22**(2), 245–268.

Berger, James O. 2003. Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statist. Sci.*, **18**(1), 1–32.

Berger, James O., and Pericchi, Luis R. 1996. The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, **91**(433), 109–122.

Berger, James O., and Sellke, Thomas. 1987. Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, **82**(397), 112–122.

Bernardo, J. M., and Smith, A. F. M. 1994. *Bayesian Theory*. Wiley.

Betancourt, M. J. 2013. Generalizing the No-U-Turn Sampler to Riemannian Manifolds. *arXiv e-prints*, Apr., arXiv:1304.1920.

Betoule, M., et al. 2014. Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples. *Astron. Astrophys.*, **568**, A22.

Box, G. E. P., and Tiao, G. C. 1992. *Bayesian Inference in Statistical Analysis*. Chicester, UK: John Wiley & Sons.

Box, George E. P. 1976. Science and statistics. *Journal of the American Statistical Association*, **71**(356), 791–799. Accessed via PDF.

Carroll, Sean M., Press, William H., and Turner, Edwin L. 1992. The Cosmological constant. *Ann.Rev.Astron.Astrophys.*, **30**, 499–542.

Casella, George, and George, Edward I. 1992. Explaining the Gibbs Sampler. *The American Statistician*, **46**(3), 167–174.

Chernoff, H. 1954. On the Distribution of the Likelihood Ratio. *The Annals of Mathematical Statistics*, **25**, 573–578.

Chivers, Tom. 2024. *Everything is Predictable: How Bayesian Statistics Explain Our World*. New York: One Signal Publishers/Atria. Illustrations; 24 cm.

Cifarelli, Donato Michele, and Regazzini, Eugenio. 1996. De Finetti's Contribution to Probability and Statistics. *Statistical Science*, **11**(4), 253–282.

Clifford, Peter. 1993. Discussion on the Meeting on the Gibbs Sampler and Other Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, **55**, 53–102.

Consonni, Guido, Fouskakis, Dimitris, Liseo, Brunero, and Ntzoufras, Ioannis. 2018. Prior Distributions for Objective Bayesian Analysis. *Bayesian Analysis*, **13**(2), 627–679, 53.

Cousins, Robert D. 2013 (October 01, 2013). *The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics*.

Cowan, Glen, Cranmer, Kyle, Gross, Eilam, and Vitells, Ofer. 2011. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, **71**, 1554.

Cox, R. T. 1946. Probability, Frequency and Reasonable Expectation. *American Journal of Physics*, **14**(1), 1–13.

Cranmer, Kyle, Pavez, Juan, and Louppe, Gilles. 2016. *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*.

Cranmer, Kyle, Brehmer, Johann, and Louppe, Gilles. 2020. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, **117**(December), 30055–30062.

D'Agostini, Giulio. 2003. *Bayesian Reasoning in Data Analysis: A Critical Introduction*. Singapore: World Scientific Publishing.

de Finetti, Bruno. 1937. La prévision : ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, **7**(1), 1–68.

de Finetti, Bruno. 1992. *Foresight: Its Logical Laws, Its Subjective Sources*. New York, NY: Springer New York. Pages 134–174.

Diaconis, Persi. 1977. Finite Forms of De Finetti?s Theorem on Exchangeability. *Synthese*, **36**(2), 271–281.

Diciccio, Thomas J., Kass, Robert E., Raftery, Adrian, and Wasserman, Larry. 1997. Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, **92**(439), 903–915.

Duane, Simon, Kennedy, A. D., Pendleton, Brian J., and Roweth, Duncan. 1987. Hybrid Monte Carlo. *Physics Letters B*, **195**(2), 216–222.

Feldman, G. J., and Cousins, R. D. 1998. A Unified Approach to the Classical Statistical Analysis of Small Signals. *Physical Review D*, **57**(7), 3873–3889.

Feroz, F., and Hobson, M. P. 2008. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, **384**, 449.

 

Foreman-Mackey, Daniel, Hogg, David W., Lang, Dustin, and Goodman, Jonathan. 2013. `emcee`: The MCMC Hammer. *Publications of the Astronomical Society of the Pacific*, **125**(925), 306?312.

Fowlie, Andrew, Handley, Will, and Su, Liangliang. 2020a. Nested sampling cross-checks using order statistics. *Monthly Notices of the Royal Astronomical Society*, **497**, 5256–5263.

Fowlie, Andrew, Handley, Will, and Su, Liangliang. 2020b (October 01, 2020). *Nested sampling with plateaus*.

Gelman, A., Roberts, G. O., and Gilks, W. R. 1996. Efficient Metropolis jumping rules. Pages 599–608 of: Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds), *Bayesian Statistics*. Oxford University Press, Oxford.

Gelman, Andrew, and Rubin, Donald B. 1992. Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.*, **7**(4), 457–472.

Geman, S., and Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**(6), 721–741.

Goodman, Jonathan, and Weare, Jonathan. 2010. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science*, **5**(1), 65–80.

Goodman, Steven N. 1999. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of internal medicine*, **130**(12), 1005–1013.

Greenland, Sander, Senn, Stephen J., Rothman, Kenneth J., Carlin, John B., Poole, Charles, Goodman, Steven N., and Altman, Douglas G. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, **31**(4), 337–350.

Gull, Stephen F. 1989. Bayesian Data Analysis: Straight-line Fitting. Pages 511–518 of: Skilling, J. (ed), *Maximum Entropy and Bayesian Methods*. Dordrecht, the Netherlands: Kluwer Academic Publishers. (871Kb).

Haario, Heikki, Saksman, Eero, and Tamminen, Johanna. 2001. An adaptive Metropolis algorithm. *Bernoulli*, **7**(2), 223–242.

Handley, W. J., Hobson, M. P., and Lasenby, A. N. 2015. polychord: next-generation nested sampling. *Monthly Notices of the Royal Astronomical Society*, **453**(4), 4385âĂŞ4399.

Hastings, W. K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, **57**(1), 97–109.

Heavens, Alan, Fantaye, Yabebal, Mootoovaloo, Arrykrishna, Eggers, Hans, Hosenie, Zafiirah, Kroon, Steve, and Sellentin, Elena. 2017 (April 01, 2017). *Marginal Likelihoods from Monte Carlo Markov Chains*.

Higson, Edward, Handley, Will, Hobson, Mike, and Lasenby, Anthony. 2019. Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation. *Statistics and Computing*, **29**, 891–913.

Hoffman, Matthew D., and Gelman, Andrew. 2014. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, **15**(1), 1593?1623.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Jaynes, E. T., and Kempthorne, Oscar. 1976. *Confidence Intervals vs Bayesian Intervals*. Dordrecht: Springer Netherlands. Pages 175–257.

Jeffreys, Harold. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, **186**(1007), 453–461.

 

Jeffreys, Harold. 1961. *Theory of Probability*. 3rd edn. Oxford University Press.
Karchev, Konstantin, and Trotta, Roberto. 2024. *STAR NRE: Solving supernova selection effects with set-based truncated auto-regressive neural ratio estimation*.
Karchev, Konstantin, Trotta, Roberto, and Weniger, Christoph. 2022. SICRET: Supernova Ia Cosmology with truncated marginal neural Ratio EsTimation. *Monthly Notices of the Royal Astronomical Society*, **520**(1), 1056–1072.
Kass, Robert E., and Wasserman, Larry. 1995. A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion. *Journal of the American Statistical Association*, **90**(431), 928–934.
Kass, Robert E., and Wasserman, Larry. 1996. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, **91**(435), 1343–1370.
Kelly, Brandon C. 2007. Some Aspects of Measurement Error in Linear Regression of Astronomical Data. *ApJ*, **665(2)**, 1489–1506.
Kendall, M. G., and Stuart, A. 1963. *The Advanced Theory of Statistics*. 2nd edn. Vol. 3. Griffin & Co.
Kessler, Richard, Becker, Andrew, Cinabro, David, Vanderplas, Jake, Frieman, Joshua A., et al. 2009. First-year Sloan Digital Sky Survey-II (SDSS-II) Supernova Results: Hubble Diagram and Cosmological Parameters. *Astrophys.J.Suppl.*, **185**, 32–84.
Kiefer, J., and Wolfowitz, J. 1956. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. *The Annals of Mathematical Statistics*, **27**(4), 887–906.
Kowalski, M., et al. 2008. Improved Cosmological Constraints from New, Old and Combined Supernova Datasets. *Astrophys.J.*, **686**, 749–778.
Laplace, Pierre Simon. 1820. *A philosophical essay on probabilities*. John Wiley and Son.
Liddle, A. R. 2007. Information criteria for astrophysical model selection. *Monthly Notices of the Royal Astronomical Society: Letters*, **377**(1), L74âĂŞL78.
Lindley, D. V. 1957. A Statistical Paradox. *Biometrika*, **44**(1/2), 187–192.
Loredo, Thomas J., and Wolpert, Robert L. 2024. *Bayesian inference: More than Bayes's theorem*.
Ly, Alexander, Marsman, Maarten, Verhagen, Josine, Grasman, Raoul, and Wagenmakers, Eric-Jan. 2017. *A Tutorial on Fisher Information*.
Lyons, Louis, and Demortier, Lucrezia. 2014. Testing Hypotheses in Particle Physics: Plots of p0 Versus p1.
MacEachern, Steven N, and Berliner, L Mark. 1994. Subsampling the Gibbs sampler. *The American Statistician*, **48**(3), 188–190.
March, M. C., Trotta, R., Berkes, P., Starkman, G. D., and Vaudrevange, P. M. 2011. Improved constraints on cosmological parameters from Type Ia supernova data: Improved constraints from SNIa. *Monthly Notices of the Royal Astronomical Society*, **418**(4), 2308?2329.
McFadden, J. 2023. Razor sharp: The role of Occam?s razor in science. *Annals of the New York Academy of Sciences*, **1530**(1), 8–17. Downloaded from Wiley Online Library on [28/12/2023] by CochraneItalia.
McGrayne, Sharon Bertsch. 2012. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Paperback edn. New Haven, CT: Yale University Press. 6.12 x 9.25 in, $16.00.
Meng, Xiao-Li, and Wong, Wing Hung. 1996. SIMULATING RATIOS OF NORMALIZING CONSTANTS VIA A SIMPLE IDENTITY: A THEORETICAL EXPLORATION. *Statistica Sinica*, **6**(4), 831–860.

 

Metropolis, N. 1987. The beginning of the Monte Carlo method. *Los Alamos Science*. Special Issue in memory of Stan Ulam.

Metropolis, N., and Ulam, S. 1949. The Monte Carlo method. *Journal of the American Statistical Association*, **44**, 335–341. PDF available here (local copy).

Metropolis, Nicholas, Rosenbluth, Arianna W., Rosenbluth, Marshall N., Teller, Augusta H., and Teller, Edward. 1953. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, **21**(6), 1087–1092.

Miller, Benjamin K., Cole, Alex, Forré, Patrick, Louppe, Gilles, and Weniger, Christoph. 2021. Truncated Marginal Neural Ratio Estimation. Pages 129–143 of: *Advances in Neural Information Processing Systems*, vol. 34. Curran Associates, Inc.

Miller, Benjamin Kurt, Cole, Alex, Louppe, Gilles, and Weniger, Christoph. 2020. Simulation-efficient marginal posterior estimation with swyft: stop wasting your precious time. *Machine Learning and the Physical Sciences at NeurIPS 2020*. arXiv:2011.13951.

Neal, Radford. 2011. *MCMC Using Hamiltonian Dynamics*. Chapman & Hall/CRC. Pages 113–162.

Neal, Radford M. 2003. Slice sampling. *Ann. Statist.*, **31**(3), 705–767.

Neal, Radford M. 2011. MCMC Using Hamiltonian Dynamics. Pages 113–162 of: Brooks, Steve, Gelman, Andrew, Jones, Galin L., and Meng, Xiao-Li (eds), *Handbook of Markov Chain Monte Carlo*, 1st edn. Chapman and Hall/CRC.

Neath, Andrew A., and Cavanaugh, Joseph E. 2011. The Bayesian information criterion: background, derivation, and applications. *WIREs Computational Statistics*, **4**(2), 199–203.

Norton, Richard A., Christen, J. Andres, and Fox, Colin. 2016. *Sampling hyperparameters in hierarchical models: improving on Gibbs for high-dimensional latent fields and large data sets*.

Papamakarios, George, and Murray, Iain. 2018. *Fast $\epsilon$-free Inference of Simulation Models with Bayesian Conditional Density Estimation*.

Rest, A., et al. 2014. 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.

Robert, Christian P. 2015. The Metropolis-Hastings algorithm. *arXiv e-prints*, Apr., arXiv:1504.01896.

Robnik, Jakob, Luca, G. Bruno De, Silverstein, E., and Seljak, Urovs. 2022. Microcanonical Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, **24**, 311:1–311:34.

Rosenthal, S. 2001. A review of asymptotic convergence for general state space Markov chains. *Far East Journal of Theoretical Statistics*, **5**.

Schittenhelm, Doris, and Wacker, Philipp. 2020 (May 01, 2020). *Nested Sampling And Likelihood Plateaus*.

Schwarz, Gideon. 1978. Estimating the Dimension of a Model. *Annals of Statistics*, **6**(2), 461–464.

Sellke, Thomas, Bayarri, M. J., and Berger, James O. 2001. Calibration of ? Values for Testing Precise Null Hypotheses. *The American Statistician*, **55**(1), 62–71.

Shariff, H., Jiao, X., Trotta, R., and van Dyk, D. A. 2016. BAHAMAS: New Analysis of Type Ia Supernovae Reveals Inconsistencies with Standard Cosmology. *Astrophys. J.*, **827**(1), 1.

Skilling, John. 2004. Nested Sampling. *AIP Conference Proceedings*, **735**(1), 395–405.

Skilling, John. 2006. Nested sampling for general Bayesian computation. *Bayesian Analysis*, **1**(4), 833–859, 27.

 

Smith, A. F. M., and Roberts, G. O. 1993. Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, **55**(1), 3–23.
Speagle, Joshua S. 2020. DYNESTY: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*, **493**, 3132–3158.
Spergel, D. N., et al. 2007. Wilkinson Microwave Anisotropy Probe (WMAP) three year results: implications for cosmology. *Astrophys. J. Suppl.*, **170**, 377.
Trotta, R. 2007a. Applications of Bayesian model selection to cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, **378**(1), 72?82.
Trotta, R. 2007b. Applications of Bayesian model selection to cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, **378**(1), 72–82. 177tn Times Cited:195 Cited References Count:60.
Trotta, R. 2008. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, **49**(2), 71–104.
Van der Vaart, Aad W. 1998. *Asymptotic Statistics*. Cambridge University Press.
van Enk, Steven J. 2014. The Brandeis Dice Problem and Statistical Mechanics. *Stud. Hist. Phil. Sci. B*, **48**, 1–6.
Van Horn, Kevin S. 2003. Constructing a logic of plausible inference: a guide to Cox's theorem. *International Journal of Approximate Reasoning*, **34**(1), 3–24.
Vardanyan, M., Trotta, R., and Silk, J. 2011. Applications of Bayesian model averaging to the curvature and size of the Universe. *Mon.Not.Roy.Astron.Soc.*, **413**, L91–L95.
Vardanyan, Mihran, Trotta, Roberto, and Silk, Joe. 2009. How flat can you get? A model comparison perspective on the curvature of the Universe. *Mon.Not.Roy.Astron.Soc.*, **397**, 431–444.
Verdinelli, Isabella, and Wasserman, Larry. 1995. Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio. *Journal of the American Statistical Association*, **90**(430), 614–618.
Wasserstein, Ronald L., and Lazar, Nicole A. 2016. The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, **70**(2), 129–133.
Wilks, S. S. 1938. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, **9**(1), 60–62.
Williams, Michael J., Veitch, John, and Messenger, Chris. 2021. Nested sampling with normalizing flows for gravitational-wave inference. *Physical Review D*, **103**, 103006.
Winkler, Gerhard. 2002. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. 2 edn. Stochastic Modelling and Applied Probability. Springer Berlin, Heidelberg. Springer-Verlag GmbH Germany 2003.
Zech, G. 2001 (June 01, 2001). *Frequentist and Bayesian Confidence Intervals*.