# 7 Bayesian Inference

Concepts and tools commonly used by classical statisticians and scientists in their work, discussed in detail in the previous three chapters, may appear disconnected, and at times, somewhat arbitrary. This uneasy feeling stems largely from the very definition of the frequentist notion of probability as a limiting frequency. By contrast, within the Bayesian paradigm, data are regarded as facts, and it is the models and their parameters, globally regarded as hypotheses, that are given a probability. Indeed, the capacity to assign a degree of plausibility to hypotheses, and the integration of probability as logic enables the elaboration of a systematic and concise inference program that is easy to follow (if not to calculate) and is well adapted to the scientific method.

The basic structure and components of the Bayesian inference process are defined and summarized in §7.2. The following section, §7.3, presents a discussion of technical matters associated with choices of prior probability functions as well as probability models. Basic examples of Bayesian inference in the presence of Gaussian noise, including Bayesian fits with linear models (i.e., models that are linear in their parameters), are discussed in §7.4, while Bayesian fitting in the presence of non-Gaussian noise and nonlinear models is introduced in §7.5. A variety of numerical techniques amenable to the calculation of the mode of posterior probabilities and $\chi^2$ minimization are next discussed in §7.6. Finally, §7.7 presents an introduction to Bayesian model comparison and entity classification with two simple examples of commonly encountered model selection problems.

## 7.1 Introduction

Thanks to tremendous developments since the mid-twentieth century, frequentist statisticians, and thus by extension all scientists, are now equipped with a rather sophisticated arsenal of tools adapted to a great variety of measurements and inference problems. However, many Bayesian statisticians would argue that the frequentist methods are typically nonintuitive and form a collection of somewhat arbitrary and seemingly disconnected techniques. They would further argue that a more intuitive and better integrated approach to inference is needed and, in fact, possible within the Bayesian paradigm, and most particularly within the context of the interpretation of probability as an extension of logic. While the goals of Bayesian inference are not altogether different from those of classical inference, the Bayesian interpretation of the notion of probability as a level of plausibility (or degree of belief) does in fact enable a more intuitive, systematic, and well-integrated inference process. Indeed, the fact one can associate a prior degree of plausibility to statements or



https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

hypotheses about the real world and determine the posterior degree of plausibility of said hypotheses based on data greatly expands the realm of inference and enables a more satisfactory set of procedures, which is arguably better adapted to the scientific method.

The practical difference between frequentist and Bayesian inferences is that a treatment of statistics based on subjective probability enables statisticians to express model parameters and their uncertainties in terms of probabilities. This cannot actually be done in the frequentist approach. For instance, frequentist 95% confidence intervals imply that for repeated measurements, 95% of confidence intervals derived from data would actually cover the parameter being measured. But beyond that point, the notion of probability is not applicable in the frequentist approach. By contrast, Bayesian statistics and inference can compare hypotheses directly and choose (infer) which is best supported by the measured data. Frequentists cannot use a probability in this fashion and must instead consider long-run frequency as extreme or in excess of the observed values, which is intuitively somewhat unclear.

All said, Bayesian statistical inference can be decomposed into three types of considerations and classes of techniques. They include

1. Determination of the posterior probability of a model and its parameters based on measured data
2. Characterization of the posterior to estimate (model) parameters and their errors, as well as compare models
3. Calculation of predictive distributions

By contrast, frequentist inference is focused on the data, which it regards as instances drawn out of a sample space, and its main components include the formulation of a probability model to interpret the data, the calculation of the likelihood to distill key features of the data or model parameter values, and comparison of a null hypothesis with competing hypotheses (hypothesis testing). These three components may obviously be regarded as "special cases" of the broader and more comprehensive Bayesian inference process. Distillation of key features of the data includes, in particular, the estimation of model parameters and associated confidence intervals of interest in either approaches of probability, as is the need to test hypotheses. The Bayesian approach, however, provides a more robust and comprehensive framework for such studies. We will thus endeavor to discuss all aspects of the Bayesian inference process in detail throughout this chapter.

The centerpiece of these considerations is the determination of the posterior probability of one or several hypotheses based on measured data. That involves the identification of a suitable data probability model to describe the measurement process (e.g., fluctuations of measured observables), the choice or evaluation of a prior probability for the physical model and its parameters, the determination of a likelihood function based on the probability model, the measured data and various assumptions about the measurement process, and, finally, a straightforward use of Bayes' theorem. Once the posterior is known, one can then make specific statements about the observables (discrete or continuous hypothesis) or the parameters of the model. Such statements about the posterior, commonly known as **summaries** by Bayesian statisticians, include the determination of the most likely values of the parameters (i.e., the mode of the distribution) as well as the evaluation of credible ranges

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

for the model parameters of interest. Additionally, given the posterior, it is also possible to make predictive inferences and calculate the probability of other observables of interest or predict the outcome of further measurements (predictive inference).

While many scientists still work within the frequentist paradigm and make regular use of "frequentist techniques" to analyze their data, a growing number are embracing the Bayesian paradigm and base their statistical analyses on the subjective interpretation of probability and Bayes' theorem. It was thus tempting to focus this textbook entirely on the most recent developments in Bayesian statistics. We feel, however, that it would be unfair to the young men and women beginning their training as scientists given that a very large fraction of works published, even today, are rooted in the frequentist paradigm or, at the very least, make use of frequentist terminology. It is thus important for young scientists to be equipped with a minimal frequentist toolset so they can understand past works in their field as well as the ongoing debate between Bayesians and frequentists. This said, it is also clear that the framework provided by probability as logic and Bayesian inference enables a robust and well-structured approach to data analysis including both deductive and predictive inferences, which might eventually take precedence over classical inference. There is in fact little doubt that Bayesian inference will steadily grow in popularity within the scientific community over the next decades. It is thus then essential to also include a sound discussion of Bayesian inference.

## 7.2 Basic Structure of the Bayesian Inference Process

Bayesian inference is defined in the context of Bayes' theorem given by Eq. (2.17) and centrally relies on the Bayesian interpretation of probability. It operates in a hypothesis (or model) space where specific hypotheses, whether discrete or continuous, are associated a probability expressing their plausibility. Measured data, considered as given (facts), are used to update the degree of belief in model hypotheses, and whenever relevant, discriminate between competing hypotheses or models.

Bayesian inference involves the following components:

1. Determination of the posterior probability of a model and its parameters based on measured data:
   a. Formulation of a data probability model based on prior information
   b. Formulation of a prior probability (density) for a working hypothesis of interest
   c. Calculation of the likelihood of the measured data based on the hypothesis
   d. Calculation of the posterior probability of the working hypothesis with Bayes' theorem
2. Distillation of key features of the posterior distribution:
   a. Determination of the mode or expectation value(s) of the posterior relative to the parameter(s) of interest
   b. Calculation of credible ranges for the parameter(s) of interest
3. Comparison of the leading hypothesis with competing hypotheses (hypothesis testing)

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

4. Calculation of predictive distributions corresponding to the outcome of prospective measurements

Two important and related points are worth stressing. First, as a general scientific rule, the measurement and inference processes should be kept distinct. The gathering of "raw" data must be carried out without interference or biasing from the inference process, that is, the answer obtained by inference should not influence or bias the sampling process. The data acquisition and analysis process must thus formally be considered as distinct and carried out in such a way that one does not affect the other.[1] Second, and for essentially the same reasons, the formulation of a prior probability, the experimental process, and the calculation of a posterior should also be considered as distinct and carried out independently. More specifically, the formulation of the prior should be completed "before" the experiment, or at the very least, without knowledge of the data produced by the experiment. As we will discuss in more detail later in this section, it is the likelihood function, determined by the data, that should influence the posterior. Having the data also influence the prior probability would amount to double counting and is thus scientifically unwarranted.

We first briefly describe each of the aforementioned components of the inference process in the following paragraphs of this section. We then elaborate on selected topics in the following sections of this chapter. Evidently, the main goal of Bayesian statistics is the determination of the posterior probability density or distribution function (PDF) of a hypothesis of interest, which may be viewed as weighted average between the prior PDF and the likelihood function. Once the posterior PDF is known, several basic features of interest can be readily computed, such as the most probable value of a continuous hypothesis, a confidence interval, as well as the plausibility of the hypothesis compared to other hypotheses or models.

### 7.2.1 Probability Model of the Data

Scientific models are typically formulated to describe properties of systems or entities and their relation with other systems or entities. For instance, in classical mechanics, Newton's third law embodies the notion that the acceleration of an object is strictly proportional to the net force applied on it by external forces. Deterministic as it may be, Newton's law says little about the specificities of measurements that may be carried out to test it. Devices or sensors may be used to determine the net force and the acceleration, or these quantities may alternatively be determined from other measurements such as the compression of a spring (force) and the time evolution of the speed of the object or measurements of its position vs. time, and so on. Each of these measurements may in turn be affected by external conditions, instrumental effects, and so on. Consequently, in addition to the (physical) model of interest, one also needs a measurement model that describes how it is measured, and how it might vary, measurement by measurement. In effect, this requires the formulation

[1] An obvious exception to this rule is the need to monitor data acquisition to ensure all components of a complex apparatus are performing properly.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

of a **data probability model** describing the probability of outcomes of the measurement process.

## Formulation of a Probability Model

A data probability model embodies the stochastic nature of the measurement process and describe the probability $p(x|x_0)$ of measuring specific values $x$ given the true value $x_0$ of an observable $X$. A probability model is formulated on the basis of prior information about the system, which is either known or assumed to be true with a certain degree of belief. Imagine, as a specific example, that you run a nuclear laboratory and that you are given the task of measuring the lifetime of some radioactive compound embedded in a small material sample. The lifetime of a nuclear isotope is evidently not measurable directly. But, if it is reasonable to assume the sample is pure, that is, composed of a single radioactive element, then one expects the activity (rate of decay) of the sample to follow a decreasing exponential with a "slope" determined by the lifetime of the compound. The lifetime may then be determined based on a measurement of the activity of the sample vs. time, or explicit measurements of decay times taken over an extended period of time. A number of complications may obviously arise in the measurement: the compound may not be perfectly pure, and the measurement precision of the activity or decay times shall evidently be limited by the quality of the instrumentation as well as the experimental conditions of the measurements. A proper description of the probability model of the measurement may thus require the inclusion of a smearing function to account for the finite resolution of the measurement and some form of background distribution to reckon for sample impurities which contribute weak but finite activity, and so on.

The formulation of a probability model obviously depends on the specificities of the measurement being considered and is perhaps best described with the detailed examples we discuss in §7.4. However, at the outset, it is useful to briefly discuss the key steps involved in the formulation of such a model. The basic idea is to formulate all information, $I$, relevant to a system in terms of logical propositions, either known to be true, or whose degree of belief or plausibility can be expressed in terms of a probability distribution or probability density. Reasoning based on these propositions shall then indicate how individual measurement instances will behave in practice. One may then model the probability of an aggregate of $n$ independent measurements in terms of the individual probabilities of each of the measurements.

For illustrative purposes, let us consider a particular model, $I$, that stipulates that the outcome $y$ of a specific measurement has a PDF $p(y|\theta)$, where $\theta$ represents one or several model parameters. The probability of observing values $\vec{y} = \{y_1, y_2, \ldots, y_n\}$ in a series of $n$ *independent* measurements, written $p(\vec{y}|\theta)$, shall then be proportional to the product of the individual probabilities of each of the measurements:

$$p(\vec{y}|\theta) \propto p(y_1|\theta) \times p(y_2|\theta) \times \cdots \times p(y_n|\theta). \tag{7.1}$$

Let us consider two simple examples illustrating the calculation of the probability $p(\vec{y}|\theta)$: the first, presented in §7.2.1, is based on the Bernoulli distribution, while the second, discussed in §7.2.1, makes use of the exponential distribution.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

### Example 1: Application of the Bernoulli Distribution

Let us first revisit the example discussed in §2.2.2 involving the manufacturing and testing of auto parts by supplier 10P100Bad. Consider that a sample of $n = 500$ manufactured parts are to be examined for defects. Let the outcome $y_i$ of each observation (i.e., each part) be 1 if a defect is found, and 0 otherwise, for $i = 1, \ldots, n$. Let $\theta$ represent the probability that a randomly selected part might be defective. Each observation may then be represented with the Bernoulli distribution (see §3.1 for a formal definition of the distribution):

$$p(y_i|\theta) = \theta^{y_i} (1-\theta)^{1-y_i}, \quad \text{for } i = 1, \ldots, n. \tag{7.2}$$

The sampled data $\vec{y} = \{y_1, y_2, \ldots, y_n\}$ thus has a probability proportional to each of the $p(y_i|\theta)$ of Eq. (7.2), and one obtains the probability model:

$$p(\vec{y}|\theta) \propto \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}, \tag{7.3}$$

$$= \theta^{\sum_{i=1}^{n} y_i} (1-\theta)^{n-\sum_{i=1}^{n} y_i}, \tag{7.4}$$

$$= \theta^{n\bar{y}} (1-\theta)^{n(1-\bar{y})}, \tag{7.5}$$

where we introduced the arithmetic mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ of the measured values. The probability model amounts to the likelihood of the observed data given a hypothesis $\theta$, which when combined with a prior probability, as we discuss in more detail later in this chapter, determines the posterior PDF of the (physical) model parameter $\theta$.

### Example 2: Application of the Exponential Distribution

Let us formulate in concrete terms the example mentioned in the beginning of this section concerning a measurement of the lifetime of an unknown compound. As discussed in detail in §3.5, one can model the probability $p(t|\tau)$ of a radioactive decay at time $t$ in terms of a decreasing exponential with a parameter $\tau$, corresponding to the lifetime of the nucleus under investigation, as follows:

$$p(t|\tau) = \frac{1}{\tau} \exp\left(-\frac{t}{\tau}\right). \tag{7.6}$$

Observations of $n$ decays at times $\vec{t} = \{t_1, t_2, \ldots, t_n\}$, may then be modeled according to

$$p(\vec{t}|\tau) \propto \frac{1}{\tau^n} \prod_{i=1}^{n} \exp\left(-\frac{t_i}{\tau}\right), \tag{7.7}$$

$$= \frac{1}{\tau^n} \exp\left(-\frac{n\bar{t}}{\tau}\right), \tag{7.8}$$

where $\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$ is the mean of the measured decay times. Here again, the probability model $p(\vec{t}|\tau)$ corresponds to the likelihood of the measured decay times $\vec{t} = \{t_1, t_2, \ldots, t_n\}$ given $\tau$. It must be combined to the prior probability of $\tau$ to determine its posterior PDF.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

## Probability Models in Real Life

In general, observations may involve a number of complications associated with instrumental effects and the presence of background processes. The basic principle of the formulation of a probability model nonetheless remains the same: all relevant information about the system must be explicitly stated and encoded to determine the probability of specific datasets. For instance, in the case of the measurement of radioactive decays, one might have to account for the precision of the time measurements. Assuming, for instance, that the timer used in the measurement exhibits Gaussian fluctuations with a standard deviation $\sigma_t$, one may then write the data probability model of each time measurement $t$ as an integral of the product of two PDFs: an exponential to account for the nature of the decay process and a Gaussian to take into account the smearing imparted by the finite resolution of the timer (clock) used in the measurement.

$$p(t|\sigma_t, \tau) = \int_{-\infty}^{\infty} p_{\text{Gaus}}(t|t', \sigma_t) \times p_{\text{Exp}}(t'|\tau)\, dt', \tag{7.9}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left[-\frac{(t-t')^2}{2\sigma_t^2}\right] \frac{1}{\tau} \exp\left(-\frac{t'}{\tau}\right) dt'. \tag{7.10}$$

The method can be readily extended to joint measurements of two or more observables determined by several model parameters, as we illustrate through several examples later in this chapter.

As we saw in §2.2.4, the introduction of the notion of probability as logic naturally enlarges the scientific discourse to include the probability of models and their parameters. In this context, the choice of a data probability model implicitly involves the formulation of one or several hypotheses about the measurement process, including the choice of a particular PDF to describe fluctuations of the measurement outcome, and parameters of the PDFs. The parameters of the data model PDF may be unspecified a priori and must then be either measured explicitly, or marginalized, as appropriate in the context of a specific measurement. Detailed studies either through actual measurements or computer simulations may then be used to establish the plausibility of the measurement model. Clearly, scientists often assume, based either on their experience or understanding of the measurement process, that the model they use is appropriate, that is, has a large plausibility of properly representing the measurement process. While such an assumption may be viable in most experimental contexts, it may have to be examined in detail if the level of accuracy and precision required of a measurement is very large. As an example, consider that the use of a Gaussian distribution to describe an observable's fluctuations is perfectly adequate as long as one ignores very rare and large fluctuations. By construction, a Gaussian has finite probability density extending to infinity (on either sides of the mean). Since infinities are not possible in practice, the Gaussian model is thus obviously flawed for very large fluctuations and must thus be replaced by a more appropriate PDF whenever extreme deviations (e.g., larger than $5\sigma$) from the mean are considered of interest.

The preceding discussion illustrates that the Bayesian inference framework naturally enables discussions of hypotheses about the measurement process and its modeling. Consideration of the plausibility of a measurement model may thus be naturally and explicitly

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

included in the calculation of physical model parameters. Such inclusion is not readily possible or feasible, strictly speaking, in the context of the frequentist approach.

### 7.2.2 Formulation of Prior PDFs

Use of Bayes' theorem toward the determination of a posterior probability of a model hypothesis $H$ evidently requires assumptions about the prior probability distribution (or density) of this hypothesis. In many cases, prior distributions may be derived from the posterior distribution of previous experiments. It is indeed the very nature of the scientific process that knowledge begets knowledge and foundational experiments may be used to produce informed guesses on parameters or model hypotheses motivating additional experiments. Prior probabilities may then be assigned directly from the posterior of previous experiments, or through the use of well-established models or frameworks. However, there are also plenty of scientific cases in which no prior experimental data exist or little information is available to construct a prior distribution. One may then establish a prior based on some general guiding principles, if any, or admit total ignorance and encode one's ignorance as a prior that considers all competing hypotheses as equally probable.

Broadly speaking, there are thus two classes of approaches discussed in the literature toward the formulation of prior distributions: they yield prior distributions commonly known as **informative** and **noninformative**. Informative priors, as the name suggests, are meant to convey and encode whatever substantive knowledge may already exist about a particular system, whether derived from other data or by reasonings based on well-established scientific principles. By contrast, noninformative priors are designed to encode one's basic ignorance of a system or a specific set of hypotheses.

Critics of the Bayesian approach argue that there exists too much freedom in the choice of priors, and that as such, the notion of prior introduces a large degree of arbitrariness and subjectivity in the inference process. Although this is true to a degree, many a Bayesian statistician would counteract that a certain level of arbitrariness also exists in frequentist procedures, and more specifically, in the choice of probability models. Supporters of the Bayesian interpretation would also argue that it would be unscientific to neglect or ignore prior information pertaining to a model that can, for instance, narrow the range of plausibility of a given continuous hypothesis. Indeed, neglecting well established facts or (physics) principles would seem rather unsound a scientific approach, much like rejecting data points because they might not support one's preconceived ideas. Bayesians would additionally argue that what really matters is that equivalent states of prior knowledge properly encoded into prior probability should lead to equivalent conclusions. This means that statisticians implementing the same prior knowledge independently, and having access to the same likelihood function, should reach equivalent posteriors and obtain similar conclusions. While such a statement can be made mathematically correct if all statisticians implement prior knowledge with the same functions and parameters, the choice of priors, most particularly uninformative priors, has remained somewhat of a contentious issue for quite some time. This apparent excess of freedom has generated quite a bit of discussions and arguments among statisticians, and much research has gone into the elaboration of techniques that enable sound choices of priors.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Given a lack of prior knowledge, the goal is to represent the ignorance of a parameter before it is measured in an objective and self-consistent manner. It would seem sensible to choose a noninformative prior distribution that encodes this ignorance by assigning equal probabilities to all possible values of a discrete parameter, or a uniform (defined in §3.4) probability density to all values of a continuous parameter.

$$p(\theta|I) = \text{ constant}; \qquad \text{Uniform prior.} \tag{7.11}$$

Indeed, one could seemingly argue there is little arbitrariness involved in assigning a constant value to a noninformative prior. Unfortunately, the issue is not so simple, and there are in fact several distinct criteria that may be used to build a noninformative prior based on identical data models and phenomenological contexts. Arguably, these distinct criteria produce distinct priors that may thus lead to different conclusions, numerically, in the assignment of modes, expectation values, error intervals, and hypothesis testing. This simplistic approach may thus lead to troubling inconsistencies.

A central issue is that in many statistical analyses, one has a rather vague prior knowledge of the parameters (continuous hypotheses) of a model. But from a physical perspective, one should expect that the choice of units or scale of a parameter, in particular, should have no bearing on the inference process and the outcome of a statistical analysis. More specifically, consider for instance that if a model formulated in terms of a parameter $\theta$ is transformed to depend on a parameter $\rho \equiv \ln\theta$, one should expect the prior distributions of $\theta$ and $\rho$ to carry the same information. This is obviously quite problematic, however, because the logarithmic scaling entails that it is, by construction, impossible for both parameters to be characterized by a uniform probability density. Indeed, if $\theta$ is chosen to have a uniform prior density, the prior distribution of $\rho$ cannot be uniform, and conversely. The question arises, then, as to which of the two variables, $\theta$ or $\rho$, should be given a uniform prior. One might also wonder how to best choose the parameters that should be uniform. For instance, while using a normal distribution model (§3.9), should the standard deviation or the variance be considered more fundamental and thus assigned a uniform prior? Does the question actually make sense? Indeed, is it meaningful to express lack of knowledge by stating that the width of the normal distribution could be infinitely small ($\sigma = 0$) or infinite? To make matters worse, one should also acknowledge that arbitrarily many forms of scaling might be possible and of interest. Which expression of a physical quantity can thus be deemed most fundamental and given a uniform prior?

Although there does not exist a universal solution to this difficult question, statisticians have developed a number of "rules" to formulate prior probabilities applicable in specific contexts. We explore some of these rules in §7.3. At this point, let us just state that Bayesian statisticians often adopt a strategy that, for instance, involves the construction of nearly (but not perfectly) uniform distributions as noninformative priors. Distributions are chosen such that variations among a set of relatively flat priors should not grossly affect the outcome of the analysis. But if it does, in practice, it is likely an indication that the parameters of interest are in fact rather poorly constrained by the data (likelihood function). Indeed, it should be the data (i.e., measurements of the real world) that constrain model parameters, not a scientist's pre- or misconceptions!

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

### 7.2.3 Posterior PDF Calculation

Given a prior distribution, $p(H_i|I)$, for a hypothesis $H_i$ and the likelihood function of measured data, $p(D|H_i, I)$, the posterior probability of the hypothesis, $p(H_i|D, I)$, is readily calculated according to Bayes' theorem:

$$p(H_i|D, I) = \frac{p(H_i|I)p(D|H_i, I)}{p(D|I)}, \tag{7.12}$$

where $p(D|I)$ is the probability of the data given prior information $I$ about the system. While it can in principle be calculated on the basis of the law of total probability,

$$p(D|I) = \int p(H_i|I)p(D|H_i, I)\, dH_i, \tag{7.13}$$

the factor $p(D|I)$ may often be regarded as a normalization constant, which becomes unnecessary when, for instance, computing the ratio of the probability of two hypotheses (§7.7).

We consider practical cases of inference and calculation of posteriors in §§7.4, 7.6, and 7.7.

### 7.2.4 Posterior PDF Characterization (Summaries)

#### Bayesian Estimators

The posterior PDF $p(H_i|D, I)$ nominally embodies everything there is to know about a hypothesis $H_i$ based on measured data, $D$, and prior information, $I$, about the system. By all accounts, it is the Bayesian estimator of an observable. Although a rather wide spectrum of types of hypotheses may be formulated and put to the test, one is most often concerned with continuous hypotheses $H(\theta)$ about the value of a system or model parameter $\theta$. It is frequently the case that $\theta$ might represent a single or specific value (e.g., a constant of nature), but the measurement process and instruments invariably produce smearing effects. The outcome of a measurement of such a parameter is thus typically a posterior probability density with a finite spread across the nominal domain of $\theta$. This finite spread is merely an artifact of the measurement process and not an intrinsic characteristic of the parameter itself. For instance, repeated measurements of the Planck constant or Big G would naturally produce distinct values that gather around the actual values of these constants. The dispersion of values arise from the measurement process, not from fluctuations of these physical parameters over time.[2] One then wishes to use the calculated posterior density $p(H_i|D, I)$ to extract a best estimate of the value of the physical parameter $\theta$. Conceptually, the task of the Bayesian statistician thus becomes rather similar to that of classical statisticians: that of obtaining an optimal estimate of the parameter value and evaluating what might be the error on that value, that is, estimate how far the extracted value might lie from the actual

[2] In the context of commonly accepted and vetted physical theories, there is no reason to believe such variations or fluctuations might occur.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

value. Important differences exist, however, in the manner in which Bayesian and classical statisticians might carry out this task. Classical statisticians have no use for a prior or a posterior. Their parameter estimation, as discussed in Chapter 4, is based on a choice of statistics (an estimator actually) in which they plug in measured data values, and while they do need a probability model to calculate the likelihood of the data, the model parameter themselves are not given a probability. Bayesian statisticians, on the other hand, wish to use prior information to constrain the estimation process; they thus carry out their estimates based on posterior probabilities that combine prior information, that is, prior probability distributions of model parameters, as well the likelihood function of the data.

The distinction between frequentist and Bayesian parameter estimation extends well beyond the numerical techniques used to obtain the estimates. In the frequentist interpretation, the probability of a given value represents the frequency with which the value would be obtained if the measurement could be repeated indefinitely. For instance, a 95% frequentist confidence interval (discussed in §6.1) corresponds to a range that should cover the actual value of the parameter 95% of the time (i.e., in 95% of measurement instances), whereas in the Bayesian approach, the true value is actually believed to be within the range with a probability of 95%.

Important differences exist also in the manner in which hypotheses are to be tested. Hypothesis testing in the frequentist paradigm, discussed in §6.4, relies on the rather unintuitive notion of a test statistics exceeding a preset threshold. Indeed, recall that one should reject a hypothesis if its test statistics has a value equal or larger than a specific value determined by the (chosen) significance level of the test. In contrast, in the Bayesian approach, one computes and compares the odds of the hypotheses themselves, that is, their respective probabilities. A particular hypothesis can then be adopted (rejected) if its probability is appreciably larger (weaker) than that of competing hypotheses.

It is important to stress once again that there could be no scientific method without proper assessment of measurement errors. A sound discussion of methods to assess the magnitude of experimental errors is thus paramount. This is not an easy topic, and several aspects of the problem had to be drawn before rigorous methods of error assessment could be considered approachable by beginning students. This is why we used a staged approach and introduced the notion of error as well as error propagation, confidence intervals, and so on, in several steps beginning already in Chapter 2, §2.11, with an appeal to intuition and development of probability distributions in series. The notion of error was formalized in Chapter 4 through the introduction of the variance of estimators and in Chapter 6 with a discussion of classical confidence intervals. In this chapter, we revisit the notion of error in the context of the Bayesian interpretation of probability and introduce the notion of **credible range** based on hypothesis posteriors.

## Mode of the Posterior PDF

As already stated in the previous section, a posterior PDF $p(\theta|D, I)$ embodies everything there is to know about the model parameter $\theta$ based on the measurement $D$, and as such constitutes a complete Bayesian estimator of $\theta$. But if $\theta$ represents an observable with a single value (or at the very least is believed to have a single and unique value), one shall

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

wish to extract a specific value $\hat{\theta}$ that is most likely to be the true value of $\theta$. Given the posterior probability density $p(\theta|D, I)$ represents the degree of belief that the true value of $\theta$ be found in the interval $[\theta, \theta + d\theta]$, it then makes sense to seek and report the mode of this function, that is, the value of $\theta$ with the largest probability density, as the best estimate of the parameter or observable. Evidently, if $p(\theta|D, I)$ is a symmetric function, the mode (extremum of the function) shall also correspond to the expectation value of $\theta$. In general, however, the expectation value of $\theta$ shall not have the largest probability and thus should not be reported as the value of the observable.

Finding the extremum of a PDF $p(\theta|D, I)$ is in principle straightforward whenever it is obtained from an analytical data probability model and a conjugate prior since, by construction, it is also analytical and a member of the same family of functions as the prior. In general, however, the posterior may be obtained from a nonlinear parametric likelihood function, it may involve several "fit" parameters, or its prior PDF might have been generated by Monte Carlo simulations and thus available in the form of a histogram: finding an extremum of $p(\theta|D, I)$ may then be somewhat arduous and require use of numerical techniques. We discuss various examples of analytical cases in §§7.3.3 and 7.4, while basic principles of extremum searches based on numerical techniques are presented in §7.6.

## Credible Range

Almost invariably, a posterior $p(\theta|D, I)$ shall have a finite spread across the domain of the parameter $\theta$. While one might wish for a very narrow distribution or even an infinitely narrow distribution (i.e., a delta function $\delta(\theta - \theta_0)$), the measurement process and instrumental effects invariably lead to a distribution with a finite dispersion. The mode has the highest probability (density) but a range of other values have a finite probability (density), as well, of corresponding to the true value. It is thus meaningful to define a range of values as a **credible range** (CR) within which the value of $\theta$ is most likely to fall. Evidently, as for confidence intervals, the breadth, or width, of a credible range shall be defined by its probability content or confidence level $\alpha$ according to

$$\int_{\text{CR}} p(\theta|D, I)\, d\theta = \alpha, \tag{7.14}$$

and there exists an infinite number of ranges CR that satisfy this condition.

Given $p(\theta|D, I)$ represents the degree of belief the observable has a value in the range $[\theta, \theta + d\theta]$, it is most meaningful to include portions of the PDF that have largest values. The credible range must then include the mode and adjoining regions of the domain of $\theta$ where $p(\theta|D, I)$ is the largest. For a symmetric distribution, this naturally leads, as for confidence intervals, to the selection of a **central interval** $[\theta_{\text{min}}, \theta_{\text{max}}]$ defined according to

$$\begin{aligned} \textstyle\int_{-\infty}^{\theta_{\text{min}}} p(\theta|D, I)\, d\theta &= (1 - \alpha)/2, \\ \textstyle\int_{\theta_{\text{max}}}^{\infty} p(\theta|D, I)\, d\theta &= (1 - \alpha)/2, \end{aligned} \tag{7.15}$$

and such that $\theta_{\text{min}}$ and $\theta_{\text{max}}$ are at equal distances from the mode of the distribution. While central intervals are also commonly applied for the selection of confidence intervals in

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

the case of asymmetric PDFs, the desire to include elements of the domain of $\theta$ with the largest probabilities leads to a different range construction algorithm. Indeed, it appears more sensible to proceed similarly to the Feldman–Cousin algorithm (§6.1.8) and include elements of the domain of $\theta$ starting with the mode and with decreasing amplitude thereafter, as illustrated in Figure 7.1a. This then also leads to the shortest interval $[\theta_{\min}, \theta_{\max}]$. One obvious drawback of this approach, however, is that the interval is not invariant under a transformation $\theta \to \xi(\theta)$; that is, an interval in $\xi$ defined with the same algorithm would not map onto the interval for $\theta$.

Use of the posterior $p(\theta|D, I)$, rather than the likelihood distribution $p(D|\theta, I)$, to identify an error interval presents two very important advantages. First, no interval inversion of the type used in the frequentist approach (and discussed in §6.1.8) is required since the "inversion" is already embodied in the posterior. And second, the presence of a physical boundary is handled gracefully. For instance, in a measurement of the a priori unknown mass of a particle (e.g., the neutrino mass), one may use an uninformative and improper prior PDF of the form

$$\begin{aligned} p(m|I) &= 0 \quad \text{for } m < 0, \\ p(m|I) &= 1 \quad \text{for } 0 \le m \le M, \end{aligned} \tag{7.16}$$

where $M$ is a suitably large constant. If the detector response is Gaussian with a precision $\xi_0$, and given the aforementioned prior, one can show that the posterior (§7.4.1) has the form of a truncated Gaussian distribution

$$\begin{aligned} p(m|D, I) &= 0 && \text{for } m < 0, \\ p(m|D, I) &= N_m \sqrt{\frac{\xi_0}{2\pi}} \exp\left[-\frac{\xi_0}{2}(m - \hat{m})^2\right] && \text{for } 0 \le m \le M, \end{aligned} \tag{7.17}$$

where $\hat{m}$ is the uncorrected mass reported by the measurement, and $N_m$ is a normalization constant equal to

$$N_m = \left(\sqrt{\frac{\xi_0}{2\pi}} \int_0^M \exp\left[-\frac{\xi_0}{2}(m - \hat{m})^2\right]\right)^{-1}. \tag{7.18}$$

Because of instrumental effects, the measured mass $\hat{m}$ may be negative. However, the notion that the mass of a particle cannot be negative is ab initio built in the posterior by virtue of the prior, Eq. (7.16). One can then use the probability density sorting method mentioned previously to establish a credible range for the measured mass: if $\hat{m}$ is large and far exceeds the mass resolution $\sigma$, one will naturally obtain a symmetric central interval, as illustrated in Figure 7.1b. However, if $\hat{m}$ is very close to the physical boundary, that is, is smaller or of the order of the mass resolution, then the sorting procedure will automatically yield a credible range with a lower bound $m = 0$ and an upper bound $m_{\max}$ (effectively a one-sided interval) determined by the precision $\xi_0$ as well as the probability content $\alpha$, as schematically illustrated in Figure 7.1c. One thus readily avoids technical issues of flip-flopping, lack of coverage, and empty confidence intervals, encountered in the frequentist approach (§6.1.8).

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

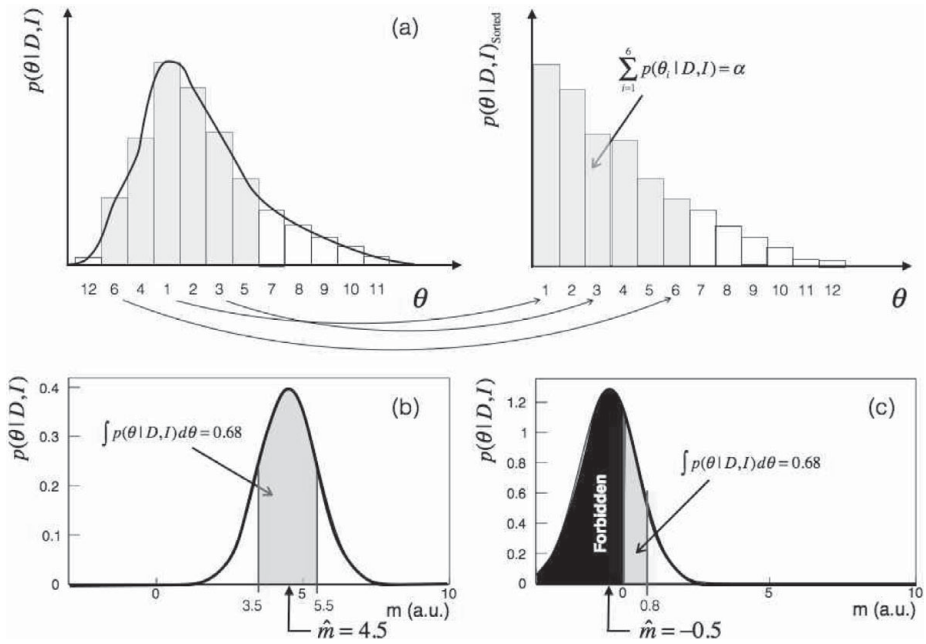


**Fig. 7.1** (a) Schematic illustration of the sorting technique used to build credible ranges with the largest probability densities. (b) Central credible region associated with a signal far from physical boundaries. (c) One-sided credible region obtained near a physical boundary (mass measurement).

### 7.2.5 Predictive Data Distributions

The Bayesian approach readily provides a framework for the computation of predictive distributions. Indeed, suppose a measurement has yielded a dataset with values $\vec{y} = (y_1, \ldots, y_n)$ resulting in the evaluation of posterior probability $p(\theta|\vec{y}, I)$ for the model parameter $\theta$. One may then use this posterior probability as the prior of a new experiment to predict the distribution of values $\vec{y}' = (y'_1, \ldots, y'_n)$ that might be obtained in that measurement. This is known as a **Posterior Predictive Distribution** (PPD) and denoted $p(\vec{y}'|\vec{y}, I)$. It can be regarded as the probability of new data $\vec{y}'$ given old data $\vec{y}$. In this context, the value of the model parameter $\theta$ might be considered of minor or no interest, and the sole purpose purpose of the procedure might then be the prediction of future data, or more aptly, the distribution of new data.

The PPD is readily calculated on the basis of the posterior probability of the model parameter(s), $\vec{\theta}$, and the likelihood function according to

$$p(\vec{y}'|\vec{y}, I) = \int p(\vec{y}'|\theta)p(\theta|\vec{y}, I)\, d\theta, \tag{7.19}$$

where the integration is taken over the entire domain of the model parameter(s) $\vec{\theta}$.

Although the expression $p(\vec{y}'|\vec{y}, I)$ might suggest some form of causal connection between the old and the new data, it should be clear that no mechanism for such connection

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

is actually implied. The value of $\vec{y}'$ is not caused by $\vec{y}$, but the knowledge of what it could become is. The PPD of $\vec{y}'$ is **conditioned** by knowledge of the parameter $\vec{\theta}$ deduced from the old data $\vec{y}$. In other words, it becomes possible to effect a prediction of the PDF of new data because the data (old and new) can be described by a model. The old data condition the model and the model so conditioned makes a prediction of the outcome of new measurements.

Causal connections or new data conditioned by old data may occur if the existence of a particular outcome $y$ has an explicit influence (causal) influence on future outcomes $y'$. Accounting for such connections (conditioning) is possible but requires proper knowledge of the conditioning mechanism. The formalism of Kalman filters, discussed in §5.6, provides a convenient framework for calculations of the evolution of $p(\vec{y}'|\vec{y}, I)$ for cases in which causal connections exist between the old and the new data and can be described based on a model of the evolution of the parameters $\vec{\theta}$ with time or some other independent (ordering) parameter.

## 7.3 Choosing a Prior

Calculation of the posterior probability $p(H|D, I)$ of a hypothesis $H$ with Bayes' theorem requires a prior probability $p(H|I)$ be formulated before the data are processed (or become available) and the likelihood $p(D|H, I)$ calculated. A difficulty often arises that only a rather limited amount of prior information might be available about a particular hypothesis and the system it describes. In some cases, the prior information available may be sufficient to select an **informative prior**, that is, a definite probability density with specific parameters (e.g., a beta distribution with well informed parameters $\alpha$ and $\beta$), but in most cases the amount of information is too limited and, at the outset, the selection of a prior might appear as an impossible challenge. Indeed, what functional dependence should one use? Should one limit or bias the range of a model parameter? And so on. One is then compelled to choose an **uninformative prior**, that is, a prior distribution that actually carries little information about the parameter or observable of interest.

Implementing an arbitrary functional form with ad hoc model parameter boundaries constitutes a rather unsatisfactory course of action toward the selection of an uninformative prior. Fortunately, two foundational methods, guided by simple yet profound principles regarding the information carried by probability distributions, enable rational as well as practical choices of distributions. The first method was devised by Edwin Jaynes (1922–1998)[3] and relies on Shannon's entropy measure of uncertainty, while the second method, credited to Sir Harold Jeffreys (1891–1989)[4], relies on the notion of Fisher information already discussed in §4.7. In actual fact, a number of other methods and principles are discussed and used in the literature for the definition and implementation of prior probabilities.

[3] American physicist famous for his work on statistical mechanics and the foundation of probability and statistical inference.

[4] English mathematician, statistician, geophysicist, and astronomer who played an important role in the revival of the Bayesian view of probability.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Indeed, the topic of prior is still somewhat contentious and much in flux. Regardless of such concerns, we will restrict our discussion to Jaynes' method, in §7.3.1, and Jeffreys' definition of the prior based on the Fisher information of a probability distribution in §7.3.2. Readers should consult more advanced texts for a fully comprehensive discussion of this difficult topic. In particular, see the book by Harney [103] for an extensive discussion of form invariant probability distributions. This said, statistical inference operates in the real word and thus requires much practicality. The choice of priors is indeed often guided by practical needs and considerations. Priors should be relatively easy to define and use; they and the posteriors one derives from them should be integrable, and so on. Families of functions known as **conjugate priors** offer much of this needed practicality. We introduce their definition and provide selected examples of conjugate families in §7.3.3.

### 7.3.1 Choice of Prior Functional Form and Jaynes' Maximum Entropy Principle

Prior information $I$ about a system might be terse, but in some cases, it might be possible to make a testable statement concerning a specific facet of the system, a particular hypothesis, or model parameter value. For instance, a statement of the type

$$I\text{: mean value of } \theta \text{ is } \theta_0$$

is readily testable. Indeed, testing $I$ involves repeated measurements of $\theta$ to get a better estimate of its value and we will see that under such conditions, Jaynes' method yields an exponential as a prior PDF for $\theta$. However, other types of information, such as $\theta$ is smaller than $\theta_0$, might be too vague to identify a specific PDF as a prior.

Jaynes' method is based on the maximization of Shannon's entropy under specific constraints defined by testable statements about the outcomes of a particular measurement. Understanding the method requires a minimal level of familiarity with Shannon's entropy. We thus first present a brief description of the definition of entropy in the context of probability distributions in §§7.3.1–7.3.1 and proceed to describe the basic principles of the method in §7.3.1. Specific applications of the methods are presented in §7.3.1.

#### Information Entropy

The notion of information entropy stems from the basic observation that stochastic systems are, in general, not equally uncertain. Bizarre as it may seem, this statement can be readily understood by considering the random draw of red and blue balls, otherwise identical in all respect, from a large vat. Imagine the vat contains 1,000 balls, only one of which is blue. It then seems rather likely that a random draw would yield a red ball. Although one is dealing with a stochastic system, the outcome of the draw indeed seems rather certain: a red ball will most likely be picked out.

What if the vat contained 500 red balls and as many blues? The odds would then be 50 : 50, and one would be just as likely to draw a red or a blue ball. The outcome would then be far less predictable. In fact, it would be maximally uncertain! But certainty and uncertainty are relative concepts. Add yellow, white, green, and black balls in equal proportions to the vat. With six colors equally probable, the outcome of a draw is even less certain. The more

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

diverse and numerous the options are, the less certain is the outcome of a draw. This seems rather obvious. But how, then, can one quantify the degree of certainty of a particular draw or, more generally, the outcome of stochastic processes?

In a seminal paper on information theory published in 1948 [172], Claude Shanon demonstrated that the uncertainty of a discrete probability distribution $\{p_1, p_2, \ldots, p_n\}$ is embodied in its **information entropy** $S(p_1, p_2, \ldots, p_n)$, defined as

$$\text{Entropy} \equiv S(p_1, p_2, \ldots, p_n) = -\sum_{k=1}^{n} p_k \ln p_k. \tag{7.20}$$

A convenient way to visualize the meaning of Shannon's entropy is to consider the outcome of a multinomial draw, that is, a stochastic process describable by a multinomial probability distribution (§3.2). As an example of such a process, let us consider $N$ successive rolls of an $m$-side die. We denote by $p_i$ and by $n_i$ the probability of a given face $i$ and the number of times it is observed in a sequence of $n$ rolls, respectively. All rolls being independent, the probability of a specific sequence of $n$ independent rolls yielding $n_1$ times face 1, $n_2$ times face 2, and so on, is proportional to the product $P \equiv p_1^{n_1} p_2^{n_2} \ldots p_m^{n_m}$. Since there are $M \equiv n!/n_1! \ldots n_m!$ ways of generating such a sequence, its probability is thus indeed a multinomial distribution:

$$p(n_1, \ldots, n_m | N; p_1, \ldots, p_m) = MP \tag{7.21}$$

$$= \frac{n!}{n_1! \cdots n_m!} p_1^{n_1} p_2^{n_2} \cdots p_m^{n_m}, \tag{7.22}$$

with

$$\sum_{k=1}^{m} n_k = n, \tag{7.23}$$

$$\sum_{k=1}^{m} p_k = 1. \tag{7.24}$$

The number of sequence permutations, hereafter called **multiplicity**, tells us much about the (un)certainty of a sequence. For instance, let us consider the multiplicity of selected sequences of $n = 8$ rolls of an $m = 4$ face die. The number of sequences yielding eight times face 1 is

$$M_{8000}^{(8)} = \frac{8!}{8!0!0!0!} = 1, \tag{7.25}$$

whereas the number of sequences yielding each face twice is given by

$$M_{2222}^{(8)} = \frac{8!}{2!2!2!2!} = 2520, \tag{7.26}$$

and the number of sequences with $n_1 = 3$, $n_2 = 3$, $n_3 = 1$, and $n_4 = 1$ amounts to

$$M_{3311}^{(8)} = \frac{8!}{3!3!1!1!} = 1120. \tag{7.27}$$

Without prior knowledge of the probabilities $p_i$, it seems rather certain that sequences producing a single face (e.g., 8000, 0800, etc.) are far less likely than sequences involving

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

an unequal mix of all faces, and less likely still than having all four faces in equal numbers. Actual sequences involving $n$ rolls of course have probabilities also determined by the probabilities $p_i$ of each of the faces. A given selection of $n$ rolls thus provides a means to gather information about these probabilities. Quite obviously, the information shall be of limited value if only a few rolls are realized. Indeed, for a number of rolls smaller or of the order of $m$, fluctuations should dominate and observed values $n_i$ cannot provide a robust estimate of the probabilities $p_i$. However, for increasingly large values of the number of rolls $n$, the numbers $n_i$ shall narrowly cluster about their expectation values $\mathrm{E}[n_i] = \langle n_i \rangle = np_i$ thereby enabling a reasonably robust estimation of the $p_i$. Parenthetically, this also implies that (prior) statements about the means and variances (or standard deviations) of the $n_i$ are testable and can form the basis for the selection of prior distributions of the parameters $p_i$.

Let us then consider the multiplicity $M$ in the large sequence size limit. For large values $n$, calculations of factorials $n!$ yield extremely large numbers that may be impractical to handle even with modern computers. It is then convenient to consider the natural logarithm of these factorials and use the Stirling approximation

$$\ln n! \approx n \ln n - n. \tag{7.28}$$

The log of the multiplicity $M$ may then be written

$$\ln M = \ln n - \sum_{k=1}^{m} \ln n_k \tag{7.29}$$

$$= n \ln n - n - \sum_{k=1}^{m} (n_k \ln n_k - n_k) \tag{7.30}$$

$$= n \ln n - \sum_{k=1}^{m} n_k \ln n_k, \tag{7.31}$$

where in the second line, we use the Stirling approximation, and in the third, the sum $\sum n_i = n$. In the large $n$ limit here considered, it is legitimate to replace the values $n_k$ by their expectation values, and we get

$$\ln M = n \ln n - n \sum_{k=1}^{m} p_k \ln(np_k), \tag{7.32}$$

$$= n \ln n - n \ln n \sum_{k=1}^{m} p_k - n \sum_{k=1}^{m} p_k \ln p_k, \tag{7.33}$$

$$= -n \sum_{k=1}^{m} p_k \ln p_k, \tag{7.34}$$

where in the third line, we use the normalization $\sum p_i = 1$. We can then finally write

$$\ln M = nS, \tag{7.35}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

where $S$ is Shannon's entropy of the distribution

$$S = \sum_{k=1}^{m} p_k \ln p_k. \tag{7.36}$$

The multiplicity $M$ of a sequence may then be written

$$M = \exp(nS). \tag{7.37}$$

Given the multiplicity $M$ of a particular type of sequence tells us something about its (un)certainty, we conclude that the entropy $S$ corresponds to a degree of certainty per draw (so to speak). Indeed $S$ provides a measure of the level of certainty embodied in a specific probability distribution.

The multiplicity has an extremum which we label $M_{\max}$ corresponding to a maximum entropy $S_{\max}$. Defining $\Delta S \equiv S_{\max} - S$, we can then also write

$$M = M_{\max} \exp\left(-n\Delta S\right). \tag{7.38}$$

In practical situations, where $n$ is very large, one finds that the multiplicity can be extremely large for $\Delta S = 0$, but it shall quickly vanish for $\Delta S > 0$. That implies that for a given probability distribution, certain types of outcomes are far more numerous than others, and thus more "uncertain."

## Generalized Entropy

Let us once again consider the roll of an $m$-face die and assume one has estimates $\{q_i\}$ for the probabilities $\{p_i\}$ of each of the faces. Using these prior estimates, one can obtain approximate values of the probability of arbitrary sequences $\{n_i\}$ of $n$ rolls of the die based on the expression of the multinomial distribution where the $p_i$ are replaced by $q_i$:

$$p(n_1, \ldots, n_m | n; q_1, \ldots, q_m) = \frac{n!}{n_1! \cdots n_m!} q_1^{n_1} q_2^{n_2} \cdots q_m^{n_m}, \tag{7.39}$$

Let us take the logarithm of this expression. Assuming the number of rolls is very large and using the Stirling approximation, one gets

$$\ln p = \ln n! - \sum_{k=1}^{m} \ln n_k! + \sum_{k=1}^{m} n_k \ln q_k \tag{7.40}$$

$$= n \ln n - \sum_{k=1}^{m} n_k \ln n_k + \sum_{k=1}^{m} n_k \ln q_k \tag{7.41}$$

Replacing $n_k$ by their expectation values, one obtains after simplification

$$\frac{1}{n} \ln p = -\sum_{k=1}^{m} p_k \ln p_k + \sum_{k=1}^{m} p_k \ln q_k, \tag{7.42}$$

$$= -\sum_{k=1}^{m} p_k \ln\left(p_k / q_k\right), \tag{7.43}$$

$$= S_{SJ}, \tag{7.44}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

which defines a generalized entropy $S_{SJ}$ commonly known as Shannon–Jaynes entropy and Kullback entropy[5]. $S_{SJ}$ is related to the Kullback–Leibler[6] divergence, also known as information divergence or information gain, which is a measure of the difference between two probability distributions.

## Entropy of Continuous Distributions

Equation (7.43) has the right functional form for an extension of the notion of entropy to continuous distributions

$$S_c = -\int p(x) \ln \left[ \frac{p(x)}{m(x)} \right] dx, \tag{7.45}$$

where $p(x)$ is a continuous probability density and $m(x)$ is known as Lebesgue measure. The inclusion of the measure $m(x)$ insures that $S_c$ is invariant under a change of variable $z \equiv z(x)$ because both the probability $p(x)dx$ and the ratio $p(x)/m(x)$ are invariant under such transformation. One may choose the measure $m(x)$ to be a constant $k_m$ across the domain of $x$. The entropy $S_c$ then becomes

$$S_c = -\int p(x) \ln p(x)\, dx + \ln k_m \int p(x)\, dx, \tag{7.46}$$

$$= -\int p(x) \ln p(x)\, dx + \text{constant}. \tag{7.47}$$

This expression tells us that the entropy of a probability density is defined up to an inconsequential constant value. This constant should indeed have no effect on Jaynes' entropy maximization principle we discuss in the next section.

## Maximization of the Entropy

Consider a measurement yielding $m$ distinct outcomes $\{x_i\}$. Let us assume very little information is available about the phenomenon or system considered. Yet, we would like to determine the prior probability distribution $p(x_i|I)$ of observing such outcomes. If very little is known about the system of interest, there is a priori no guidance or reason to choose any particular functional form for $p(x_i|I)$. One must then seek a functional form that yields maximum entropy, that is, a probability distribution with maximum uncertainty about the outcomes $\{x_i\}$. However, if some testable (prior) information is available about the system, one may also use this information as a constraint in the maximization of the entropy. The basic idea of Jaynes' principle of entropy maximization is to carry a variational calculation to find an expression that maximizes the entropy in the presence of finitely many constraints. One thus seeks a maximum of the entropy $S$ by variation of the (unknown) probabilities $p_i$ of the observed values $x_i$. A variation of these probabilities should yield a

[5] Named after American cryptanalyst and mathematician Solomon Kullback (1907–1994).
[6] Richard A. Leibler (1914–2003), American mathematician and cryptanalyst.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

stationary solution for $S$ at the extremum:

$$dS = \sum_{k=1}^{m} \frac{\partial S}{\partial p_k} dp_k \equiv 0. \tag{7.48}$$

In the absence of further information, one might assume that the probabilities $p_k$ are mutually independent, and one would then conclude that all coefficients $\partial S/\partial p_k$ are null. $S$ would then be a constant independent of the probabilities $p_k$ and not much could be said about the functional form of the probability $p(x)$. Suppose, however, that some constraints are imposed on the $p_i$ based on prior information about the system, one should then be able to use the calculus of variation with Lagrange underdetermined multipliers to obtain a functional form for the $p_i$.

Let us illustrate the idea using a basic constraint. Since the measurement of $n$ values is known to happen, its probability is unity. The sum of the probabilities of the $m$ values $x_i$ must then be unity:

$$\sum_{k=1}^{m} p_k = 1. \tag{7.49}$$

We write the constrained (generalized) entropy as

$$S' = S - \lambda \left( \sum_{k=1}^{m} p_k - 1 \right) \tag{7.50}$$

$$= -\sum_{k=1} p_k \ln p_k - \lambda \left( \sum_{k=1}^{m} p_k - 1 \right), \tag{7.51}$$

where $\lambda$ is a Lagrange multiplier and we seek a variation of the $p_i$ that yields an extremum of $S'$:

$$dS' = \sum_{k=1}^{m} \frac{\partial S'}{\partial p_k} dp_k \equiv 0. \tag{7.52}$$

Assuming all $p_i$ are a priori independent, the coefficients $\partial S'/\partial p_k$ must all be null and we find

$$0 = \frac{\partial S'}{\partial p_i}, \tag{7.53}$$

$$= -\frac{\partial}{\partial p_i} \left[ \sum_{k=1} p_k \ln p_k + \lambda \left( \sum_{k=1}^{m} p_k - 1 \right) \right] = 0, \tag{7.54}$$

$$= \sum_{k=1} \delta_{ik} \ln p_k + \sum_{k=1} \delta_{ik} + \lambda \sum_{k=1} \delta_{ik}, \tag{7.55}$$

$$= \ln p_i + (1 + \lambda). \tag{7.56}$$

We find that the probabilities $p_i$ should be of the form

$$p_i = \exp\left[-(1 + \lambda)\right], \tag{7.57}$$

$$= \exp(-\lambda_0), \tag{7.58}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

where, for notational convenience, we introduced the constant $\lambda_0 = 1 + \lambda$. In order to determine this modified multiplier, let us insert Eq. (7.58) in the equation of the constraint

$$\sum_{k=1}^{m} \exp(-\lambda_0) = 1. \tag{7.59}$$

This yields

$$\exp(-\lambda_0) = \frac{1}{m}, \tag{7.60}$$

and we conclude that in the absence of prior information, the principle of maximum entropy tells us that all probabilities $p_i$ should be equal

$$p_i = \frac{1}{m}. \tag{7.61}$$

This result reinforces the intuitive notion that in the absence of prior information, all values of a parameter or hypothesis should be considered equally probable. Indeed, with no prior information whatsoever, the most sensible choice for a prior probability of hypothesis, that which is most uncertain, should be taken as a uniform distribution.

Let us next consider what one can learn about the $p_k$ if additional testable information is available about the observable $x$. Let us write this additional information in the form of $s$ independent constraints

$$\sum_{i=1}^{m} g_j(x_i) p_i = \langle g_j \rangle, \tag{7.62}$$

where the $g_j(x)$ represent $s$ independent functions of $x$. Adding these constraints with multipliers $\lambda_j$ to the entropy $S'$, we get

$$S' = -\sum_{i=1} p_i \ln p_i - \lambda \left( \sum_{i=1}^{m} p_i - 1 \right) - \sum_{j=1}^{s} \lambda_j \left( \sum_{i=1}^{m} g_j(x_i) p_i - \langle g_j \rangle \right) \tag{7.63}$$

Seeking once again a stationary solution for $dS' = 0$, we obtain

$$0 = \ln p_i + (1 + \lambda) + \sum_{j=1}^{s} \lambda_j g_j(x_i) \tag{7.64}$$

which yields a generic solution of the form

$$p_i = \exp(-\lambda_0) \exp\left[ -\sum_{j=1}^{s} \lambda_j g_j(x_i) \right]. \tag{7.65}$$

The first constraint (i.e., $\sum p_i = 1$) now yields

$$\exp(\lambda_0) = \sum_{i=1}^{m} \exp\left[ -\sum_{j=1}^{s} \lambda_j g_j(x_i) \right]. \tag{7.66}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

The multipliers $\lambda_j$ can similarly be determined by insertion of the $p_i$ given by Eq. (7.65) into the constraint equations (7.62). Solution of these $s$ equations for the $\lambda_j$ may in general require numerical algorithms. Analytical solutions are possible in some cases, however, as we discuss next.

## Maximization of the Entropy with Simple Constraints

Equation (7.65) provides a generic solution for the probability coefficients $p_i$ in the presence of $s$ constraints of the form given by Eq. (7.62). Let us consider two applications of this equation, each involving two simple constraints.

### An Estimate of the Mean as a Constraint

Let us first derive a specific functional form for $p_i$ in the context of a single nontrivial constraint (i.e., one constraint aside from $\sum p_i = 1$) consisting of an estimate of the mean of $x$. That is, let

$$g_1(x) = x. \tag{7.67}$$

and

$$\sum_{i=1}^{m} g_1(x_i) p_i = \sum_{i=1}^{m} x_i p_i = \langle x \rangle. \tag{7.68}$$

Equation (7.65) then reduces to an exponential distribution:

$$p_i = \exp(-\lambda_0) \exp\left[-\lambda_1 x_i\right]. \tag{7.69}$$

The normalization constraint $\sum p_i = 1$ yields

$$\exp(\lambda_0) = \sum_{i=1}^{m} \exp\left(-\lambda_1 x_i\right). \tag{7.70}$$

and the constraint imposed by the mean gives us

$$\langle x \rangle = \frac{\sum_{i=1}^{m} x_i \exp\left(-\lambda_1 x_i\right)}{\sum_{i=1}^{m} \exp\left(-\lambda_1 x_i\right)}, \tag{7.71}$$

which can be solved numerically for $\lambda_1$ and finitely many values $x_i$. In the continuum limit, sums are replaced with integrals. Restricting the domain of $x$ to $[0, \infty]$, one gets

$$\langle x \rangle = \frac{\int_0^\infty x \exp\left(-\lambda_1 x\right) dx}{\int_0^\infty \exp\left(-\lambda_1 x\right) dx} = \frac{1}{\lambda_1}. \tag{7.72}$$

The maximum entropy principle tells us that in cases for which an estimate $\langle x \rangle$ of the mean of an otherwise unconstrained parameter is known, one should use an exponential prior probability distribution

$$p(x) = \frac{1}{\langle x \rangle} \exp\left(-\frac{x}{\langle x \rangle}\right) \quad \text{for } x \geq 0. \tag{7.73}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

### An Estimate of the Variance as a Constraint

In many studies, the dispersion of a parameter may be more telling than its average. Let us thus consider a constraint based on an estimate of the variance. Let

$$g_1(x) = (x - \langle x \rangle)^2. \tag{7.74}$$

and

$$\sum_{i=1}^{m} g_1(x_i) p_i = \sum_{i=1}^{m} (x_i - \langle x \rangle)^2 p_i = \langle \Delta x^2 \rangle. \tag{7.75}$$

Equation (7.65 ) then reduces to

$$p_i = \exp(-\lambda_0) \exp\left[-\lambda_1 (x_i - \langle x \rangle)^2\right]. \tag{7.76}$$

which in the continuous limit gives us

$$p(x) = \exp(-\lambda_0) \exp\left[-\lambda_1 \left((x - \langle x \rangle)^2\right)\right]. \tag{7.77}$$

The next step is to insert Eq. (7.77) in the two constraint equations in order to obtain values for the parameters $\lambda_0$ and $\lambda_1$. In the continuum limit, one has

$$1 = \exp(-\lambda_0) \int_{-\infty}^{\infty} \exp\left[-\lambda_1 (x_i - \langle x \rangle)^2\right] dx, \tag{7.78}$$

$$\langle \Delta x^2 \rangle = \exp(-\lambda_0) \int_{-\infty}^{\infty} ((x - \langle x \rangle)^2 \exp\left[-\lambda_1 ((x - \langle x \rangle)^2\right] dx, \tag{7.79}$$

from which one gets

$$\exp(\lambda_0) = \sqrt{2\pi} \langle \Delta x^2 \rangle^{1/2} \tag{7.80}$$

$$\lambda_1 = \frac{1}{2\langle \Delta x^2 \rangle}. \tag{7.81}$$

Given a constraint determined by an estimate of the variance $\langle \Delta x^2 \rangle$, the principle of maximum entropy yields a Gaussian distribution with a mean $\langle x \rangle$ and standard deviation equal to the square root of this estimate. This means a Gaussian carries the largest uncertain in this context. This is a rather important result. It indicates, that unless additional information is available, the Gaussian distribution is the least certain and thus carries the fewest assumption about the system considered and should thus lead to the most conservative results. If the value of $\langle \Delta x^2 \rangle$ is unknown (i.e., with no available estimates), the Gaussian distribution still provides a safe prior insofar as it can be established that the dispersion of the data is finite and that its variance can be treated as nuisance parameter, that is, it can eventually be marginalized.

In some cases, the range of a model parameter can be limited a priori to a finite domain $x_L \leqslant x < x_H$. The maximum entropy principle can then be applied to the generalized entropy $S_{SJ}$, Eq. (7.42), with

$$q_i = \begin{cases} \frac{1}{x_H - x_L} & \text{for } x_L \leqslant x < x_H \\ 0 & \text{elsewhere.} \end{cases} \tag{7.82}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Because $q_i$ is a constant, the entropy maximization procedure remains essentially the same provided one replaces $p_i$ by $p_i/q_i$. The integrals of Eq. (7.78) must include a factor $1/(x_H - x_L)$ and the bounds of integration set to $x_L$ and $x_H$. The integration may then be carried out in terms of error functions, and the resulting PDFs is a truncated Gaussian (see Problem 7.1).

### 7.3.2 Scalable Priors and Jeffreys' Priors

#### Issues with Uniform Priors

Let us once again consider the problem of the determination of the detection efficiency of elementary particles in a complex detector. Recall from §4.7.4 that the problem is equivalent to that of an **unfair** coin toss determined by a binomial probability model with a probability $\varepsilon$ of yielding "head" (i.e., a success). Let us assume, perhaps dramatically, that no information whatsoever is available about the detection efficiency and that it could have a value anywhere in the range [0, 1]. Given the binomial probability model determining the outcome of $N$ repeated measurements (or coin tosses), it would make sense to use a conjugate prior, that is, a prior probability in the form of a beta distribution[7]. Unfortunately, the assumed lack of knowledge makes such a choice totally arbitrary. How indeed can one justify the use of any particular values for the shape parameters $\alpha$ and $\beta$?

Bayes and Laplace proposed that prior (total) ignorance of the value of $\varepsilon$ should be represented by a uniform PDF within the range of applicability of the parameter. In the context of our detection efficiency problem, this Bayes' prior takes the form

$$p(\varepsilon) = \begin{cases} 1 & \text{for } 0 \leq \varepsilon \leq 1 \\ 0 & \text{elsewhere.} \end{cases} \tag{7.83}$$

The posterior probability $p(\varepsilon|D)$ would then be strictly equal to the likelihood $p(D|\varepsilon)$ yielding, in the case at hand, a posterior in the form of a binomial distribution with a mode equal to the maximum likelihood estimate $\varepsilon = n/N$, where $n$ and $N$ are the number of observed and produced particles respectively. Note, parenthetically, that a uniform distribution in a finite range corresponds in fact to a beta distribution with parameters $\alpha = \beta = 1$. If the range of the parameter is unbound, the prior is said to be improper because its integral over the full range of the parameter diverges. It is relatively easy to verify, however, that the posterior remains proper, i.e., with a finite and well-defined normalization.

The notion of using a uniform distribution for a completely unknown parameter is quite appealing and sounds rather straightforward. It is also corresponds, as we saw in §7.3.1, to a functional shape with maximum uncertainty. The obvious problem arises, however, that a different choice of parameterization, $\theta \equiv \theta(\varepsilon)$, shall lead to an arbitrarily shaped prior. For instance, let the total ignorance about $\varepsilon$ be represented by Eq. (7.83) and let us choose the new parameterization $\theta = \varepsilon^2$. By definition of the notion of probability, one must have

[7] The notion of conjugate prior is formally introduced in §7.3.3.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

$p(\theta)d\theta = p(\varepsilon)d\varepsilon$ and conclude that the PDF of $\theta$ is equal to

$$p(\theta) = \frac{1}{2\theta^{1/2}}, \tag{7.84}$$

which is manifestly not a uniform distribution in the range $[0, 1]$. It then appears that the actual formulation of the parameter matters. Indeed, if one chooses $\theta$ to have a uniform prior, this prior shall be clearly incompatible with a uniform prior for $\varepsilon$. The two priors would lead to different and irreconcilable results for the posterior. This seems rather awkward at the outset although in practice, a particular formulation or choice of parameter might be better justified or natural. For instance, in the case of a binomial distribution, it seems natural and reasonable to define a uniform prior in terms of the probability of success $\varepsilon$, and any other parameterization might be considered artificial or unnatural. The choice is considerably less obvious, however, with other commonly used PDFs such as the Gaussian PDF for which one might choose to parameterize a prior in terms of a uniform standard deviation $\sigma$ or uniform variance $\sigma^2$.

Working with a binomial model, J. B. S. Haldane, a geneticist, advocated the use of an improper prior density of the form $p(\varepsilon) \propto \varepsilon^{-1}(1-\varepsilon)^{-1}$, which is a conjugate of the binomial distribution (and more specifically a particular case of the beta distribution with $\alpha = \beta = 1$). Although this distribution yield the right limiting behavior for the mean (mode), it yields an improper and thus problematic posterior if $n = 0$ or $n = N$ are encountered experimentally.

Harold Jeffrey, proposed to instead use

$$p(\varepsilon) \propto \varepsilon^{-1/2}(1-\varepsilon)^{-1/2} \tag{7.85}$$

as a conjugate prior for a binomial probability model. He based this particular form of the beta distribution on a simple invariance principle we discuss in detail in §7.3.2. But first, we consider uninformative location and scaling priors in the next two sections.

## Uninformative Location Priors

A model parameter $\theta$ may be considered a **location parameter** whenever a probability distribution may be written in the form $p(x - \theta|\theta, I)$. A sensible candidate for an uninformative prior would be a uniform prior $p(\theta|I) \propto 1$. If $\theta$ is bounded within a finite interval $\theta_{\min} \leq \theta \leq \theta_{\max}$, the prior may be properly normalized

$$p(\theta|I) = \begin{cases} (\theta_{\max} - \theta_{\min})^{-1} & \text{for } \theta_{\min} \leq \theta \leq \theta_{\max} \\ 0 & \text{elsewhere.} \end{cases} \tag{7.86}$$

However, if the domain of $\theta$ is $\mathbb{R}$, the flat prior is said to be **improper** because its integral over the domain diverges. Be it as it may, given this improper normalization appears both in the numerator and the denominator of Bayes' theorem used toward the calculation of a posterior probability for $\theta$, use of such an improper probability distribution remains technically acceptable.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

## Uninformative Scaling Priors

Consider a physical observable $X$ scalable by an arbitrary factor $\theta$. This factor might correspond to a change of units (e.g., transforming a distance from meters to kilometers) but it is more interesting to consider an actual physical scaling of the observable. Such a scaling should leave the total probability $\int p(x|I)$ of observing $X$ unchanged; one can thus write

$$p_\theta(x|I) = \frac{1}{\theta} p(x/\theta|I). \tag{7.87}$$

The actual scale factor of a phenomenon might be a priori unknown (e.g., a cross-section) and the goal of a measurement might then be to determine this factor. Since the scale factor is totally unknown, the prior on $\theta$ should then be invariant under an arbitrary rescaling of $\theta$ by a positive constant $c$, that is,

$$p(\theta) = \frac{1}{c} p(\theta/c). \tag{7.88}$$

This implies that a rescaling of $\theta$ should not change the prior, and consequently, the prior provides no information about the physical scale of the process and parameter $\theta$ of interest. Equation (7.88) is a functional equation that must admit a single solution for $p(\theta)$ up to an arbitrary scaling factor. The prior must thus be of the form

$$p(\theta) \propto \frac{1}{\theta}. \tag{7.89}$$

Applied without bounds in $\mathbb{R}$, its integral diverges, and it is thus an improper probability distribution. It may also defined with bounds

$$p(\theta) = \begin{cases} \frac{k}{\theta} & \text{for } \theta_{\min} \leq \theta \leq \theta_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{7.90}$$

The normalization constant $k$ is obtained by integration

$$1 = \int_{\theta_{\min}}^{\theta_{\max}} p(\theta|H_1, I) d\theta = k \ln \theta|_{\theta_{\min}}^{\theta_{\max}}, \tag{7.91}$$

which yields $k^{-1} = \ln \theta_{\max}/\theta_{\min}$. The bound prior may then be written

$$p(\theta) = \begin{cases} \frac{1}{\ln \theta_{\max}/\theta_{\min}} \frac{1}{\theta} & \text{for } \theta_{\min} \leq \theta \leq \theta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{7.92}$$

It is of obvious interest to seek a transformed variable $\xi(\theta)$ endowed with a flat (improper) prior. By definition of the notion of probability, one writes

$$p(\xi) = p(\theta) \left| \frac{d\theta}{d\xi} \right|, \tag{7.93}$$

where $p(\xi)$ is required to be flat and $p(\theta) \propto 1/\theta$. One thus find that

$$\frac{d\xi}{d\theta} = \frac{1}{\theta}, \tag{7.94}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

which implies that $\xi$ is of the form

$$\xi = \ln \theta. \tag{7.95}$$

We thus conclude that a flat prior formulated in terms of $\log \theta$ provides an arbitrary scalable probability distribution and the logarithm ensures that all orders of magnitude of the scaling factor are treated equally, that is, given equal probability.

As we will see later in this chapter, flat priors in $\log A$, where $A$ is the unknown amplitude of a signal, are quite useful in the study and search of weak signals of unknown amplitude.

### Improper Priors

Improper prior are based on density distributions whose integral does not exist or diverges. They are commonly used as noninformative priors. They can usually be viewed as limits of proper priors, with the corresponding posterior being the limit of the posteriors corresponding to those priors. Although they cannot formally be used as probability density, their use is generally deemed acceptable provided the posterior obtained with Bayes' theorem is a proper probability distribution, and when the data are sufficiently informative about the parameter of interest to render the prior essentially irrelevant.

### Locally Uniform Priors

Locally uniform priors are based on functions that are essentially constant over the region in which the likelihood is appreciable and do not feature large values outside that range. Because the prior is nearly constant in the range where the likelihood is large, one can write

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)\,d\theta} \approx \frac{p(x|\theta)}{\int p(x|\theta)\,d\theta} = \frac{L(x|\theta)}{\int L(x|\theta)\,d\theta}. \tag{7.96}$$

A locally uniform prior $p(\theta)$ thus effectively provides a convenient uninformative prior for the parameter of interest $\theta$. Locally uniform priors can be defined based on proper density distributions and thus eliminate the need for improper priors.

Conjugate priors (discussed in §7.3.3), such as the beta and Gaussian distributions, may be selected to yield large ranges of parameter space where they are nearly constant and thus provide, effectively, an uninformative prior. Because the posterior of a conjugate prior is in the same family of distributions as the prior, calculations of the posterior and its properties are greatly simplified. Conjugate priors thus provide double convenience: they can be suitably chosen as uninformative priors and provide for simplified calculations of posteriors and their properties.

However, the use of a locally uniform prior may be consider philosophically unsatisfying because it requires a peep at the likelihood distribution to ensure the prior's flat region in fact covers the entire domain where the likelihood is large, thereby violating the principle that a prior should be formulated based on prior knowledge only, that is, without looking at the data for which it is designed to serve as a prior.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

## Jeffreys' Invariance Principle

Jeffreys argued that equivalent propositions should have the same probability. If two or more parameterizations of a particular proposition are possible, they should be equivalent, that is, they should yield the same end result. Jeffreys additionally argued that the prior probability distribution of a parameter should be determined by the Fisher information of that parameter derived from the data probability model. This makes sense. If a parameter is unknown, its prior probability distribution should be as vague as possible but no more vague than the information that can be extracted from the data. Jeffreys therefore proposed that the prior $p(\varepsilon|I)$ be proportional to some simple power of the Fisher information. But no particular choice of parameterization should have preeminence. More specifically, given some choice of parameter $\varepsilon$, it should be possible to switch to any new model parameterization, $\theta \equiv \theta(\varepsilon)$, provided the densities satisfy

$$p(\theta) = p(\varepsilon)\frac{d\varepsilon}{d\theta}. \tag{7.97}$$

Given the transformation property of the Fisher information under a variable transform derived in §4.7.3, Eq. (4.97),

$$\mathbb{I}(\theta) = \mathbb{I}(\varepsilon)\left(\frac{\partial\varepsilon}{\partial\theta}\right)^2 \tag{7.98}$$

and the preceding requirement, one concludes that a self-consistent prior is obtained with a power one-half, that is,

$$p(\varepsilon) \propto [\mathbb{I}(\varepsilon)]^{1/2}. \tag{7.99}$$

Let us verify that this prior indeed features the desired transformation property. By virtue of the invariance principle, the prior for $\theta$ is written

$$p(\theta) \propto [\mathbb{I}(\theta)]^{1/2}. \tag{7.100}$$

Applying the transformation property of the Fisher information, Eq. (7.98), under a change of variable $\varepsilon \to \theta$, one gets

$$p(\theta) \propto [\mathbb{I}(\varepsilon)]^{1/2}\frac{\partial\varepsilon}{\partial\theta}, \tag{7.101}$$

$$= p(\varepsilon)\frac{\partial\varepsilon}{\partial\theta}, \tag{7.102}$$

which indeed satisfies the transformation, Eq. (7.97).

## Examples of Jeffreys' Prior

Let us apply Jeffreys' invariance principle with selected probability models to obtain prior probability distributions applicable in various contexts.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

### Uninformative Scaling Prior Revisited

Let us first verify that the uninformative scaling prior discussed in §7.3.2 is in fact a Jeffreys prior for the amplitude $A$ of (improper) distributions of the form

$$p(x|A) = Af(x), \tag{7.103}$$

where $f(x)$ is a nonnegative, finite, and integrable function of $x$, independent of the amplitude $A$. Using the definition, Eq. (4.89), to calculate the Fisher information, we find

$$\begin{aligned} \mathbb{I}(A) &= -\mathrm{E}\left[\frac{\partial^2}{\partial A^2}\ln p\right] \\ &= -\mathrm{E}\left[\frac{\partial}{\partial A}\frac{1}{A}\right] \\ &= \frac{1}{A^2}. \end{aligned} \tag{7.104}$$

Taking the square root of the information,

$$p(A) = \mathbb{I}(A)^{1/2} = \left(\frac{1}{A^2}\right)^{1/2} = \frac{1}{A}. \tag{7.105}$$

we find that Jeffreys prior for $A$ is indeed of the form, Eq. (7.89), obtained for scalable parameters.

### Binomial Distribution

Next, consider Jeffreys prior for the success rate, $\varepsilon$, of a binomial distribution. As we saw in §4.109, the Fisher information $\mathbb{I}(\varepsilon)$ of a binomial PDF is given by

$$\mathbb{I}(\varepsilon) = \frac{N}{\varepsilon\,(1-\varepsilon)}. \tag{7.106}$$

We thus verify that Eq. (7.85), reproduced below, indeed corresponds to Jeffreys prior for a binomial PDF, that is, it satisfies Jeffreys invariance principle

$$p(\varepsilon) \propto \varepsilon^{-1/2}(1-\varepsilon)^{-1/2}. \tag{7.107}$$

### Gaussian Distribution

Next consider the prior for the parameters of a Gaussian PDF. Given that the off-diagonal elements of the Fisher information matrix of a Gaussian are null, as shown in Eq. (4.116), one can sensibly formulate independent priors for the mean $\mu$ and variance $\sigma^2$ of the distribution. For the mean, we get

$$p(\mu) \propto \sqrt{\mathbb{I}_{\mu,\mu}(\mu,\sigma^2)}, \tag{7.108}$$

$$= \frac{k}{\sigma}, \tag{7.109}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

where in the second line, we introduced a normalization constant determined, for proper priors, by the range of applicability $\mu_{\min} \leq \mu \leq \mu_{\max}$ of the mean, that is, $k = \sigma(\mu_{\max} - \mu_{\min})$. In effect, since $\sigma$ is a constant, albeit of unknown value, we find that the prior for $\mu$ is independent of $\mu$, and thus amounts to a flat prior.

For the variance, one gets

$$p(\sigma^2) \propto \sqrt{\mathbb{I}_{\sigma^2,\sigma^2}(\mu, \sigma^2)}, \tag{7.110}$$

$$= \frac{k'}{\sigma^2}. \tag{7.111}$$

One thus concludes that the prior for $\sigma^2$ is not flat. However, it is easy to verify (Problem 7.2) that this implies that the prior for $\log \sigma^2$ is a uniform distribution. This is both convenient and meaningful. If a particular observable is really Gaussian distributed, its standard deviation must be finite and bound under some scale. Since the scale may not be known a priori, a uniform prior for $\log \sigma^2$ ensures that smaller values of $\sigma^2$ are far more probable that very large values. In essence, it preserves the Gaussian nature of the distribution, which for very large values of $\sigma$ might otherwise appear flat in a small range of the observable.

### Poisson Distribution

Jeffreys prior for the rate parameter $\lambda$ of a Poisson distribution is similarly calculated.

$$p(\lambda) \propto \sqrt{\mathbb{I}(\lambda)}, \tag{7.112}$$

$$= \sqrt{\mathrm{E}\left[\left(\frac{\partial}{\partial \lambda} \ln p(n|\lambda)\right)^2\right]}, \tag{7.113}$$

$$= \sqrt{\frac{1}{\lambda^2}\mathrm{E}\left[(n-\lambda)^2\right]}. \tag{7.114}$$

Noting that $\mathrm{E}\left[(n-\lambda)^2\right]$ corresponds to the variance of the distribution, $\mathrm{Var}[n] = \lambda$, one concludes that the prior for the rate parameter may be written

$$p(\lambda) \propto \frac{1}{\sqrt{\lambda}}. \tag{7.115}$$

### Multinomial Distribution

Finally, let us consider the prior distributions for the rate parameters, $\vec{p} = (p_1, p_2, \ldots, p_m)$, of a multinomial distribution, with the constraint $\sum_{i=1}^{m} p_i = 1$. One can show that Jeffreys prior for the coefficients $\vec{p}$ is the Dirichlet distribution with all of its parameters set to half. One can additionally show that with transformations $p_i = \phi_i^2$, the parameters $\vec{\phi}$ are uniformly distributed on a unit sphere of $m - 1$ dimensions.

## 7.3.3 Conjugate Priors

The choice of a particular functional form for a prior probability density is often guided by practical considerations. One finds, indeed, that certain functional forms are

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

particularly well suited, or convenient, for use with specific probability models. For instance, if a data sample may be described with a Bernoulli probability model, the choice of a beta distribution as prior is quite convenient because, as we demonstrate in the text that follows, the posterior obtained from the product of a beta distribution by the likelihood of a sample of values determined by a Bernoulli distribution is also a beta distribution. The beta distribution (family) is then said to be **conjugate prior** to the Bernoulli distribution (family).

Having a posterior in the same distribution family as the prior is quite convenient because it enables a dynamic and iterative improvement of the knowledge of the parameters of the data model. Indeed, once data are acquired, the product of the likelihood by the prior enables the determination of a posterior of the same family that can then be used as a prior for another sequence of measurements. The process can be iterated indefinitely because the posterior remains in the same family as the prior regardless of the observed values or the number of samples. In that sense, the family of beta distributions is then also said to be be closed under sampling from a Bernoulli distribution.

The notion of conjugate distribution can be extended to virtually any data probability models. Indeed, essentially all data probability models commonly in use in statistical analyses are associated with a family of distributions useable as conjugate priors. In this section, we introduce a basic selection of such conjugate families for illustrative purposes and as a foundation for examples of Bayesian inference discussed in later sections of this chapter. Several additional conjugate prior families, as well as Jeffreys priors, are documented in the very comprehensive compendium produced by D. Fink [82].

## Bernoulli Processes

Consider the sum, $y = \sum_{i=1}^{n} x_i$, of $n$ instances $x_i$ of random variables $X_i$[8] drawn from a common Bernoulli distribution with an unknown success parameter $\varepsilon$ defined in the range $0 < \varepsilon < 1$. The likelihood distribution of the sum $y$ is the joint PDF $p(\vec{x}|\varepsilon)$:

$$p(\vec{x}|\varepsilon) = \prod_{i=1}^{n} p(x_i|\varepsilon), \tag{7.116}$$

$$= \prod_{i=1}^{n} \varepsilon^{x_i} (1-\varepsilon)^{1-x_i}, \tag{7.117}$$

$$= \varepsilon^{y} (1-\varepsilon)^{n-y}. \tag{7.118}$$

The functional dependence of the likelihood on $\varepsilon$ and $1-\varepsilon$ suggests that the beta distribution (§3.7) might be an appropriate prior for the Bernoulli distribution. The beta distribution may be written

$$p(\varepsilon|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \varepsilon^{\alpha-1} (1-\varepsilon)^{\beta-1}, \tag{7.119}$$

[8] In a Bernoulli process the variable $X$ equals 1 for successes and 0 for failures.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

which indeed features factors in $\varepsilon$ and $1-\varepsilon$ similar to those of the likelihood. The posterior is, by construction, proportional to the product of the prior and the likelihood. Keeping exclusively the factors with an explicit dependence on $\varepsilon$, we get

$$p(\varepsilon|x) \propto \varepsilon^{y}\,(1-\varepsilon)^{n-y}\,\varepsilon^{\alpha-1}\,(1-\varepsilon)^{\beta-1}\,, \tag{7.120}$$

$$\propto \varepsilon^{y+\alpha-1}\,(1-\varepsilon)^{\beta+n-y-1}\,, \tag{7.121}$$

which up to a normalization constant may be recognized as a beta distribution of the form Eq. (7.119) with parameters

$$\alpha' = \alpha + y, \tag{7.122}$$

$$\beta' = \beta + n - y, \tag{7.123}$$

and a normalization constant equal to

$$\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+y)\Gamma(\beta+n-y)}. \tag{7.124}$$

The family of beta distributions thus indeed constitutes a conjugate to the Bernoulli distribution.

The use of a conjugate distribution has an interesting practical advantage. Suppose that the value of $\varepsilon$ is unknown but that a beta prior with specific values for $\alpha$ and $\beta$ is available. A single measurement (trial) would yield either $x=0$ or $x=1$. If the value is 1, the posterior of this measurement is a new beta distribution with $\alpha' = \alpha + 1$ and $\beta' = \beta$ whereas if the value is 0, the posterior has $\alpha' = \alpha$ and $\beta' = \beta + 1$. Effectively, $\alpha$ counts the successes while $\beta$ accounts for the number of failures. As the number of trials $N$ increases, the parameters $\alpha$ and $\beta$ will tend toward their respective expectation values: $\langle\alpha\rangle = \varepsilon N$ and $\langle\beta\rangle = (1-\varepsilon)N$ but recall from §3.7 that the expectation value of a beta distribution is $\alpha/\beta$. That means that for increasing $N$, the ratio $\alpha/\beta$ will tend toward $\varepsilon/(1-\varepsilon)$ and, as such, provides an estimator for $\varepsilon$.

$$\varepsilon = \frac{\alpha/\beta}{1+\alpha/\beta}. \tag{7.125}$$

Also recall from §3.7 that the variance of the beta distribution is $\sigma^2 = \alpha/\beta^2$. Substituting the expectation values for $\alpha$ and $\beta$, we find

$$\sigma^2 = \frac{1}{N}\frac{\varepsilon}{(1-\varepsilon)^2}, \tag{7.126}$$

which means that the width of the posterior distribution should decrease in proportion to $\sqrt{N}$. As illustrated in Figure 7.2, the posteriors of successive Bernoulli experiments (with a beta prior) provide estimates for $\varepsilon$ with an accuracy that improves as $\sqrt{N}$. However, note that the choice of a beta prior is disadvantageous for values of $\varepsilon$ very close to zero or unity given the variance of the distribution diverges in these two limits.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

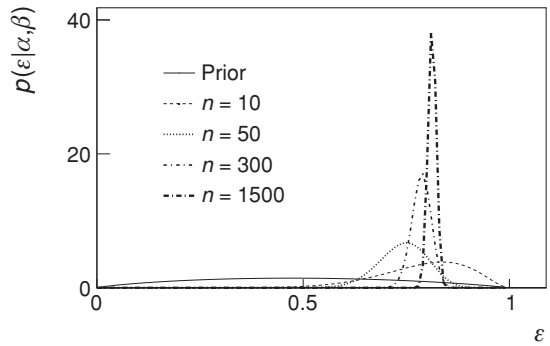


**Fig. 7.2** Illustration of the evolution of the beta posterior observed in a succession of Bernoulli measurements. Parameters of the prior (solid line) were arbitrarily set to $\alpha = 1.8$ and $\beta = 1.9$ to obtain a relatively uninformative prior. Beta posteriors are shown for $n = 10, 50, 300$, and 1,500 trials sampling of a Bernoulli distribution with $\varepsilon = 0.8$.

## Poisson Processes

Let $x_1, x_2, \ldots, x_n$ represent a random sample from a Poisson distribution with a positive definite rate parameter $\lambda$. The likelihood function of the sample is

$$p_n(\vec{x}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \quad \text{for } \lambda > 0 \tag{7.127}$$

$$= \frac{\lambda^y e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}, \tag{7.128}$$

where, once again, we defined $y = \sum_{k=0}^{n}$. Choosing a prior PDF for $\lambda$ in the form of a gamma distribution (§3.6), we write

$$p(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}. \tag{7.129}$$

The posterior $p(\lambda|\vec{x})$ is thus of the form

$$p(\lambda|\vec{x}) \propto \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \frac{\lambda^y e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}, \quad \text{for } \lambda > 0$$

$$\propto \lambda^{y+\alpha-1} e^{-(n+\beta)\lambda}, \tag{7.130}$$

which is itself a gamma distribution with parameters

$$\begin{aligned} \alpha' &= \alpha + y, \\ \beta' &= \beta + n. \end{aligned} \tag{7.131}$$

We thus conclude that the family of gamma distributions constitute a conjugate to the family of Poisson distributions.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

## Normal Processes (Known Variance)

For convenience, we define a **precision** parameter $\xi$ as the multiplicative inverse of the variance of a Gaussian distribution

$$\xi \equiv \frac{1}{\sigma^2}. \tag{7.132}$$

The Gaussian distribution may then be written

$$p(x|\mu, \xi) = \sqrt{\frac{\xi}{2\pi}} \exp\left[-\frac{1}{2}\xi(x-\mu)^2\right], \tag{7.133}$$

and the likelihood of a set of data $\vec{x} = (x_1, x_2, \ldots, x_n)$ is

$$p_n(\vec{x}|\mu, \xi) = \left(\frac{\xi_0}{2\pi}\right)^{n/2} \exp\left[-\frac{1}{2}\xi_0 \sum_{i=1}^{n}(x_i-\mu)^2\right], \tag{7.134}$$

where $\xi_0$ is the known value of the precision.

Given the Gaussian dependence of the likelihood on $\mu$, let us formulate a conjugate prior for this variable also in terms of a Gaussian distribution

$$p(\mu|\mu_p, \xi_p) = \sqrt{\frac{\xi_p}{2\pi}} \exp\left[-\frac{1}{2}\xi_p\left(\mu-\mu_p\right)^2\right], \tag{7.135}$$

where $\mu_p$ is a prior estimate of the mean and $\xi_p$ is the assumed precision of that estimate. The posterior is then of the form

$$p(\mu|\mu_p, \xi_p) \propto \exp\left\{-\frac{1}{2}\left[\xi_0 \sum_{i=1}^{n}(x_i-\mu)^2 + \xi_p\left(\mu-\mu_p\right)^2\right]\right\}. \tag{7.136}$$

We next show that the argument of Eq. (7.136) may be written in the form $-\frac{1}{2}\xi_p'(\mu-\mu_p')^2$. We first decompose the square $(x_i-\mu)^2$ as follows:

$$(x_i-\mu)^2 = (x_i-\bar{x})^2 + (\mu-\bar{x})^2 + 2\,(x_i-\bar{x})\,(\bar{x}-\mu)\,. \tag{7.137}$$

where $\bar{x}$ represents the arithmetic mean of the sample. A sum of $n$ such terms may be written

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n\,(\mu-\bar{x})^2 + 2\,(\bar{x}-\ mu)\sum_{i=1}^{n}(x_i-\bar{x})\,, \tag{7.138}$$

where the last term identically vanishes because the sum of $x_i$ equals $n\bar{x}$. Omitting factors with no dependence on $\mu$, the posterior probability becomes

$$p(\mu|\vec{x}) \propto \exp\left\{-\frac{1}{2}\left[n\xi_0\,(\mu-\bar{x})^2 + \xi_p\left(\mu-\mu_p\right)^2\right]\right\}.$$

In order to further simplify this expression, we introduce

$$Q(\mu) = n\xi_0\,(\mu-\bar{x})^2 + \xi_p\left(\mu-\mu_p\right)^2. \tag{7.139}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

It is straightforward, although somewhat tedious, to show that $Q(\mu)$ may be transformed in the form

$$Q(\mu) = \left(n\xi_0 + \xi_p\right)\left[\mu - \frac{n\xi_0\bar{x} + \mu_p\xi_p}{n\xi_0 + \xi_p}\right]^2 + K, \tag{7.140}$$

where $K$ represents a constant expression independent of $\mu$ that can be relegated to the normalization constant of the posterior (Problem 7.4). We thus finally get

$$p(\mu|\vec{x}) \propto \exp\left(-Q(\mu)/2\right), \tag{7.141}$$

$$\propto \exp\left[-\frac{1}{2}\xi_p'\left(\mu - \mu_p'\right)^2\right] \tag{7.142}$$

where we introduced the updated mean and precision

$$\mu_p' = \frac{n\xi_0}{n\xi_0 + \xi_p}\bar{x} + \frac{\xi_p}{n\xi_0 + \xi_p}\mu_p, \tag{7.143}$$

$$\xi_p' = n\xi_0 + \xi_p. \tag{7.144}$$

Note that $\mu_p'$ actually corresponds to the weighted mean of the sample mean, $\bar{x}$, and the prior estimate of the mean $\mu_p$ with weights dependent on the actual precision of the measurement $\xi_0$, the size $n$ of the sample, and the prior precision of the estimate. The precision of the estimate $\mu_p'$ is itself greatly improved as it becomes equal to the sum of the precision of $n$ measurements and the prior estimate of the precision. The variance of the posterior mean may then be written

$$\sigma_p'^2 = \frac{\sigma_0^2\sigma_p^2}{n\sigma_n^2 + \sigma_0^2}. \tag{7.145}$$

## Normal Processes (Fixed Mean)

Let us next assume that the mean of a process is known (and equal to $\mu_0$) but that its variance (or precision) is not. The likelihood, Eq. (7.134), features an exponent of $\xi$ and an exponential function of $\xi$. It is thus natural to invoke a gamma distribution as a prior. This yields a posterior of the form

$$p(\xi|\vec{x}) \propto \frac{\beta^\alpha}{\Gamma(\alpha)}\xi^{\alpha-1}e^{-\beta\xi}\left(\frac{\xi}{2\pi}\right)^{n/2}\prod_{i=1}^{n}\exp\left[-\frac{\xi}{2}(x_i - \mu_0)^2\right], \tag{7.146}$$

$$\propto \xi^{\alpha+\frac{n}{2}-1}\exp\left\{-\xi\left[\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2 + \beta\right]\right\}, \tag{7.147}$$

which is itself a gamma distribution with parameters

$$\alpha' = \alpha + n/2, \tag{7.148}$$

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2. \tag{7.149}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

We thus conclude that the gamma distribution constitutes a conjugate of the Gaussian distribution for the precision parameter $\xi$ when the mean of the normal process is fixed.

## Normal Processes (General Case)

The general case requires special care. Given both the mean $\mu$ and the precision $\xi$ are conditioned by the data $x$, the order in which they are inferred matters. The joint posterior of $\mu$ and $\xi$ may be written

$$p(\xi, \mu|\vec{x}) \propto p(\xi) \times p(\mu|\xi, I) \times p(\vec{x}|\xi, \mu, I). \tag{7.150}$$

The prior $p(\xi)$ is a beta distribution (as in the previous section). For $p(\mu|\xi, I)$, we use a Gaussian distribution, as in §7.3.3, with a precision equal to $\xi_p = n_p\xi$ where $\xi$ is the prior precision of the measurement and $n_p$ a multiplicative factor that determines the prior precision on the mean relative to the measurement precision. The joint posterior may then be written

$$\begin{aligned} p(\xi, \mu|\vec{x}) \propto\ & \xi^{\alpha-1} \exp\left[-\beta\xi\right] \\ & \times \xi^{1/2} \exp\left[-\frac{n_p\xi}{2}(\mu - \mu_p)^2\right] \\ & \times \xi^{n/2} \exp\left[-\frac{\xi}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]. \end{aligned} \tag{7.151}$$

Inserting $\bar{x} - \bar{x}$ in the $x_i - \mu$ terms, we find after some simple algebra

$$\begin{aligned} p(\xi, \mu|\vec{x}) \propto\ & \xi^{\alpha+\frac{n}{2}-1} \exp\left\{-\xi\left[\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\right\} \\ & \times \xi^{1/2} \exp\left\{-\frac{\xi}{2}\left[n_p(\mu - \mu_p)^2 + n(\bar{x} - \mu)^2\right]\right\}. \end{aligned} \tag{7.152}$$

One may then obtain a posterior for $\xi$ alone by marginalization of $\mu$. Integration of Eq. (7.152) relative to $\mu$ yields the posterior for $\xi$ (Problem 7.5):

$$\begin{aligned} p(\xi|\vec{x}) \propto\ & \xi^{\alpha+n/2-1} \exp\left\{-\xi\left[\beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right.\right. \\ & \left.\left. + \frac{nn_p}{2(n+n_p)}(\bar{x} - \mu_p)^2\right]\right\}, \end{aligned} \tag{7.153}$$

which is a beta distribution, of the form Eq. (7.119), with parameters

$$\alpha' = \alpha + n/2, \tag{7.154}$$

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nn_p}{2(n+n_p)}(\bar{x} - \mu_p)^2. \tag{7.155}$$

The posterior for $\mu$ is a Gaussian with a mean $\mu' = \frac{n}{n+n_p}\bar{x} + \frac{n_p}{n+n_p}\mu_p$ and a precision $\xi'_p = (n + n_p)\xi$. Summarizing, given a set of data $\vec{x}$, one can use Eqs. (154) and (155) to get an estimate of $\xi$ and subsequently Eq. (7.139) to get an estimate of the mean.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

### Multivariate Normal Processes

The choice of a generic conjugate distribution for a multivariate normal distribution is somewhat more involved than the examples presented in previous sections. Several parameterizations are possible and discussed in the literature (e.g., see [82] and references therein).

### Brief Epilog on Conjugates

The use of conjugate priors confers several interesting properties to a Bayesian inference analysis. The first obvious advantage derives from the definition of conjugate priors: the posterior remains in the same family as the prior, thus greatly reducing the mathematical complexity of an analysis. Furthermore, since priors of a specific family of functions have a common dependency on finitely many parameters, it is typically possible to obtain closed expressions for the evolution of these parameters with added data, as we have shown in §§7.3.3–7.3.3. This means that one can regard the inference process as an iterative process.

Given a prior and its parameters, the acquisition of new data enables an update of the parameters according to a closed form expression. These updated parameters can then be used for the definition of a prior (of the same family) for another measurement. The procedure may then be iterated arbitrarily as many times as there are new available datasets. The end result is a set of posterior parameters that seamlessly takes into account all measurements pertaining to a specific observable. Additionally note that the explicit dependence of the updated parameters on the measurement precision (e.g., normal process with known precision discussed in §7.3.3) means this iterative analysis process takes into account the actual precision of all measurements in the series. Effectively, the procedure enables full and optimal accounting of all relevant data in a sequence of multiple experiments. This iterative procedure is in fact rather similar to the Kalman filtering technique discussed in §5.6 in which a sequence of measurements are used, one after the other, to improve the knowledge of the state of a system. As for Kalman filtering, one can also treat systems with an evolving state driven by known external parameters. In effect, the notion of Kalman filtering can then be articulated and developed within the Bayesian inference paradigm (see [63] for a recent review), but such a discussion lies far beyond the scope of this textbook.

Conjugate priors may be used even when there is a very little prior information about a system or observable. Indeed, it is typically possible (as illustrated in Figure 7.2) to choose the parameters of a prior to produce a very broad and essentially uninformative prior probability density. In effect, if the parameters are chosen to yield a probability density much broader than the typical precision of measurements, the posterior parameters shall be mainly determined by the measurement and with only a very weak dependency on the prior parameter values.

## 7.4 Bayesian Inference with Gaussian Noise

Process and measurement noise can often be considered Gaussian or approximately Gaussian; it is thus of great interest to consider applications of Bayesian inference involving

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Gaussian data probability model. We begin, in §§7.4.1 and 7.4.2, with examples of Bayesian inference involving the estimation of the mean $\mu$ of a signal in the presence of Gaussian noise and the evaluation of the variance such a signal. These two examples set the stage for more elaborate problems involving the determination of model parameters by means of Bayesian fits. Linear model fits are considered in §7.4.3 whereas nonlinear model fits are discussed in §7.6.

## 7.4.1 Sample Mean Estimation

We first discuss the estimation of the mean $\mu$ of a constant signal in the presence of Gaussian noise with known and constant variance in §7.4.1. Given the noise level encountered in measurements may change, we enlarge the discussion to include varying but known noise variance in §7.4.1 and a priori unknown noise variance in §7.4.1.

### Sample Mean Estimation with Fixed Variance Noise

Consider $n$ independent measurements $\vec{x} = (x_1, x_2, \ldots, x_n)$ of a constant observable $X$. As prior information, let us assume measurements of $X$ are known to fluctuate with zero bias but Gaussian noise of fixed and known variance $\sigma_0^2$.

Our goal is to determine the posterior probability $p(\mu|X, \sigma_0, I)$. From Bayes' theorem, we get

$$p(\mu|\vec{x}, \sigma_0, I) = \frac{p(\mu|I)p(\vec{x}|\mu, \sigma_0, I)}{p(\vec{x}|\sigma_0, I)}, \tag{7.156}$$

where $p(\vec{x}|\mu, \sigma_0, I)$ is the likelihood of the data for parameters $\mu$ and $\sigma_0$, $p(\mu|I)$ is the prior of $\mu$, and $p(\vec{x}|\sigma_0, I)$ is the global likelihood of the data.

Let us first determine the likelihood function $p(X|\mu, \sigma_0, I)$. From the prior information, we know that each value $x_i$, $i = 1, \ldots, n$, represents a constant observable $X$ with Gaussian noise and null bias. We can then model each measurement according to

$$x_i = \mu + e_i, \tag{7.157}$$

where the term $e_i$ represents Gaussian noise with null mean and standard deviation $\sigma_0$:

$$p(e_i|\sigma_0, I) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{e_i^2}{2\sigma_0^2}\right] \tag{7.158}$$

Substituting $e_i = x_i - \mu$ in Eq. (7.58), one gets a data probability model for the fluctuations of the measurements $x_i$:

$$p(x_i|\mu, \sigma_0, I) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma_0^2}\right] \tag{7.159}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

For notational convenience, let us define precision factors $\xi_0 = 1/\sigma_0^2$. The likelihood of the data $\vec{x}$ may then be written

$$p(\vec{x}|\mu, \xi_0, I) = \prod_{i=1}^{n} p(x_i|\mu, \xi_0, I) = \prod_{i=1}^{n} \sqrt{\frac{\xi_0}{2\pi}} \exp\left[-\frac{\xi_0}{2}(x_i - \mu)^2\right] \tag{7.160}$$

$$= \left(\frac{\xi_0}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{\xi_0}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \tag{7.161}$$

$$= \left(\frac{\xi_0}{2\pi}\right)^{\frac{n}{2}} \exp\left[-\frac{\xi_0}{2}S\right], \tag{7.162}$$

where we defined the sum $S = \sum_{i=1}^{n}(x_i - \mu)^2$. Expanding this sum and introducing the sample mean $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$, and variance $s^2 = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = n^{-1}\sum x_i^2 - \bar{x}^2$, one gets

$$S = n\,(\mu - \bar{x})^2 + ns^2. \tag{7.163}$$

Defining $\xi_n = n\xi_0$, the likelihood can then be written

$$p(\vec{x}|\mu, \xi_0, I) = \left(\frac{1}{n}\right)^{\frac{n}{2}} \left(\frac{\xi_n}{2\pi}\right)^{\frac{n}{2}-1} \sqrt{\frac{\xi_n}{2\pi}} \exp\left[-\frac{\xi_n}{2}s^2\right] \times \sqrt{\frac{\xi_n}{2\pi}} \exp\left[-\frac{\xi_n}{2}(\mu - \bar{x})^2\right]. \tag{7.164}$$

Calculation of the posterior for $\mu$ requires a prior $p(\mu|I)$. For illustrative purposes, we will consider two cases of prior information on $\mu$. In the first case, we will assume the mean $\mu$ is bound in some range $\mu_L \leq \mu < \mu_H$ based on some theoretical considerations but without any particular preference within that range. This will require an uninformative prior. In the second case, we will assume previous experiments have reported an estimate $\mu_p$ with a standard deviation $\sigma_p$. We will then use a Gaussian conjugate prior.

## Uninformative Prior (Case 1)

Given the bounds $\mu_L \leq \mu < \mu_H$ and lack of preference in that range, we choose a uniform prior distribution

$$p(\mu|I) = \begin{cases} R_\mu^{-1} & \mu_L \leq \mu \leq \mu_H \\ 0 & \text{elsewhere,} \end{cases} \tag{7.165}$$

where the constant $R_\mu = \mu_H - \mu_L$ is determined by the normalization condition $\int p(\mu|I)\,d\mu = 1$.

Equipped with the prior, Eq. (7.165), and the likelihood, Eq. (7.164), we calculate the global likelihood of the data according to

$$p(X|I) = \int_{\mu_L}^{\mu_H} p(\mu|I)p(X|\mu, \xi_0, I)\,d\mu, \tag{7.166}$$

$$= \frac{I_{LH}}{R_\mu}\left(\frac{1}{n}\right)^{\frac{n}{2}} \left(\frac{\xi_n}{2\pi}\right)^{\frac{n}{2}-1} \sqrt{\frac{\xi_n}{2\pi}} \exp\left[-\frac{\xi_n}{2}s^2\right], \tag{7.167}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

with

$$I_{LH} = \sqrt{\frac{\xi_n}{2\pi}} \int_{\mu_L}^{\mu_H} \exp\left(-\frac{\xi_n}{2}(\mu - \bar{x})^2\right) d\mu, \tag{7.168}$$

$$= \Phi(z_H) - \Phi(z_L), \tag{7.169}$$

in which the standard normal cumulative distribution $\Phi(z)$ is evaluated at $z_H = \sqrt{\xi_n}(\mu_H - \bar{x})$ and $z_L = \sqrt{\xi_n}(\mu_L - \bar{x})$.

Computation of the posterior of $\mu$ with Eq. (7.156) finally yields

$$p(\mu|\vec{x}, \sigma_0, I) = \begin{cases} \frac{1}{I_{LH}} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{(\mu-\bar{x})^2}{2\sigma_n^2}\right] & \text{for } \mu_L \leq \mu < \mu_H, \\ 0 & \text{elsewhere,} \end{cases} \tag{7.170}$$

which corresponds to a truncated Gaussian distribution defined in the range $\mu_L \leq \mu \leq \mu_H$, with mode $\bar{x}$, standard deviation parameter $\sigma_n = \sigma_0/\sqrt{n}$, and normalization $I_{LH}$. We thus conclude that the observed sample mean $\bar{x}$ has the highest probability of being the true value of $\mu$ and the uncertainty of this estimate is inversely proportional to the square root of the sample size $n$.

### Gaussian Conjugate Prior (Case 2)

Let us next consider a Gaussian prior with mean $\mu_p$ and precision $\xi_p$.

$$p(\mu|I) = \sqrt{\frac{\xi_p}{2\pi}} \exp\left[-\frac{\xi_p}{2}\left(\mu - \mu_p\right)^2\right]. \tag{7.171}$$

From our discussion of Gaussian conjugate priors in §7.3.3, we conclude that the posterior of $\mu$ is a Gaussian

$$p(\mu|\vec{x}, \xi_0, I) = \sqrt{\frac{\xi_p'}{2\pi}} \exp\left[-\frac{\xi_p'}{2}\left(\mu - \mu_p'\right)^2\right], \tag{7.172}$$

with mean $\mu_p'$ and precision $\xi_p'$ given by

$$\mu_p' = \frac{n\xi_0}{n\xi_0 + \xi_p}\bar{x} + \frac{\xi_p}{n\xi_0 + \xi_p}\mu_p, \tag{7.173}$$

$$\xi_p' = n\xi_0 + \xi_p, \tag{7.174}$$

and the updated standard deviation may then be written

$$\sigma_p' = \frac{\sigma_p \sigma_0/\sqrt{n}}{\sqrt{\sigma_0^2/n + \sigma_p^2}}. \tag{7.175}$$

The results obtained in this section are obviously reminiscent of the classical inference properties of estimators for the mean of a sample discussed in §§5.1 and 5.5. Important differences exist, however, between the Bayesian result derived in this section and earlier results. In the frequentist approach, it is not possible to make a statement about the

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

probability of a model parameter, and one obtains a confidence interval based on the measurement. Strictly speaking, the probability content of the interval is not the probability of finding the value of $\mu$ within the interval, but the probability of the interval to contain the observable value if and when the measurement is repeated several times. Consider, for instance, a 68% confidence interval. If the experiment is repeated $m = 100$ times, $m$ distinct 68% confidence intervals will be obtained and only 68% of these intervals, on average, are expected to contain the actual value of the observable. By contrast, the Bayesian distribution obtained in this section does state the probability of finding $\mu$ in a certain interval, whether the experiment is repeated or not. This important difference stems from the definition of probability used in the Bayesian paradigm. Indeed, the Bayesian interpretation is far simpler and direct. The posterior $p(\mu|X, I)$ gives the probability density of $\mu$ and it can then be used to determine the probability of the observable value being any range of interest. Note in particular that if the bounds $\mu_L$ and $\mu_H$ are very broad and far outside the interval $\bar{x} \pm \sigma/\sqrt{n}$, then the value of $I_{LH}$ quickly converges to unity. An interval $\pm\sigma/\sqrt{n}$ then has a probability of 68% as in the frequentist paradigm. However, if the bounds $\mu_L$ and $\mu_H$ are near the interval $\bar{x} \pm \sigma/\sqrt{n}$, the integral $I_{LH}$ may deviate significantly from unity, and the probability of $\mu$ being found within $[\bar{x} - \sigma/\sqrt{n}, \bar{x} + \sigma/\sqrt{n}]$ may then far exceed 68% and even reach 100%. Note that in this case, the data do not add much to the prior information, and the posterior remains broad within the bounding range $\mu_L \leq \mu \leq \mu_H$.

## Sample Mean Estimation with Variable Noise

Let us extend the discussion of the previous section to measurements in which the variance of the noise may vary from measurement to measurement. Let us continue to assume, however, that the variances are known, that is, we assign variances $\sigma_i^2$ to each of the $n$ measurements and assume these values to be known based either on previous measurements or some other considerations.

The prior probability $p(\mu|I)$ is indifferent to the new noise conditions but the likelihood function must be slightly modified. For convenience, let us define precision factors $\xi_i = 1/\sigma_i^2$. The likelihood can then be written

$$p(\vec{x}|\mu, \sigma, I) = \prod_{i=1}^{n} \sqrt{\frac{\xi_i}{2\pi}} \exp\left[-\frac{\xi_i}{2}(x_i - \mu)^2\right] \tag{7.176}$$

$$= \left(\prod_{i=1}^{n} \sqrt{\frac{\xi_i}{2\pi}}\right) \exp\left[-\frac{1}{2}S_w\right], \tag{7.177}$$

where $n$ is the number of measurements and the sum $S_w$ is defined as

$$S_w = \sum_{i=1}^{n} \xi_i \left(x_i - \mu\right)^2. \tag{7.178}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Introducing the weighted mean $\bar{x}_w$ and the weighted variance $\overline{s_w^2}$ defined respectively as

$$\bar{x}_w = \frac{\sum_{i=1}^{n} \xi_i x_i}{\sum_{i=1}^{n} \xi_i}, \tag{7.179}$$

$$\overline{s_w^2} = \frac{\sum_{i=1}^{n} \xi_i (x_i - \bar{x}_w)^2}{\sum_{i=1}^{n} \xi_i}, \tag{7.180}$$

as well as the sum of the precision factors

$$\xi_w = \sum_{i=1}^{n} \xi_i, \tag{7.181}$$

it is relatively simple to verify that the sum $S_w$ may be written

$$S_w = \xi_w (\mu - \bar{x}_w)^2 + \xi_w \overline{s_w^2}, \tag{7.182}$$

from which we obtain the likelihood

$$p(\vec{x}|\mu, \{\xi_i\}, I) = \left( \prod_{i=1}^{n} \sqrt{\frac{\xi_i}{2\pi}} \right) \exp\left( -\frac{\xi_w}{2} \overline{s_w^2} \right) \exp\left[ -\frac{\xi_w}{2} (\mu - \bar{x}_w)^2 \right]. \tag{7.183}$$

Defining the average measurement precision $\bar{\xi}$ as

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i, \tag{7.184}$$

the precision parameter $\xi_w$ is written

$$\xi_w = n\bar{\xi}, \tag{7.185}$$

from which we conclude that the precision $\xi_w$ improves in proportion to the number $n$ of measurements. Consequently, the standard deviation of the likelihood,

$$\sigma_w = 1/\sqrt{\xi_w}, \tag{7.186}$$

decreases as the square root of the number of measurements, that is, $\sigma_w \propto 1/\sqrt{n}$.

The posterior PDF $p(\mu|\vec{x}, \{\xi_i\}, I)$ is calculated similarly as in the previous section.

## Bounded Uniform Prior (Case 1)

In the case of a bounded uniform prior, one gets

$$p(\mu|X, I) = \begin{cases} \frac{1}{I_w} \sqrt{\frac{\xi_w}{2\pi}} \exp\left[ -\frac{\xi_w}{2} (\mu - \bar{x}_w)^2 \right], & \text{for } \mu_L \leq \mu < \mu_H, \\ 0 & \text{elsewhere.} \end{cases} \tag{7.187}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

The posterior PDF of $\mu$ is once again a truncated Gaussian distribution with a normalization factor $I_w$ defined by

$$I_w = \sqrt{\frac{\xi_w}{2\pi}} \int_{\mu_L}^{\mu_H} \exp\left[-\frac{\xi_w}{2}(\mu - \bar{x}_w)^2\right] d\mu. \tag{7.188}$$

The expectation value and most probable value of $\mu$ is $\bar{x}_w$ and the distribution has a standard deviation parameter $\sigma_w$, which, as per Eq. (7.186), varies inversely with the number of points $x_i$ collected in the data sample. Once again, one can directly calculate the probability of the value $\mu$ be found in any particular interval with Eq. (7.187).

### Gaussian Conjugate Prior (Case 2)

Given the likelihood, Eq. (7.183), and the Gaussian conjugate prior, Eq. (7.171), the posterior PDF for $\mu$ is a regular Gaussian with posterior mean $\mu'_p$ and precision $\xi'_p$ given by

$$\mu'_p = \frac{\xi_w}{\xi_w + \xi_p}\bar{x}_w + \frac{\xi_p}{\xi_w + \xi_p}\mu_p, \tag{7.189}$$

$$\xi'_p = \xi_w + \xi_p = n\bar{\xi} + \xi_p \tag{7.190}$$

and the updated standard deviation may then be written

$$\sigma'_p = \frac{\sigma_p \bar{\sigma}/\sqrt{n}}{\sqrt{\bar{\sigma}^2/n + \sigma_p^2}}, \tag{7.191}$$

where

$$\bar{\sigma} = 1/\sqrt{\bar{\xi}}. \tag{7.192}$$

For sufficiently large $n$ (and finite $\bar{\xi}$ and $\xi_p$), the prior parameter $\xi_p$ becomes negligible, and the standard deviation of the posterior scales as

$$\sigma'_p \propto \frac{1}{\sqrt{n}}, \tag{7.193}$$

as expected. For small $n$, the mode of the distribution is very much influenced by the prior parameter $\mu_p$, but for increasing larger values of $n$, the second term of Eq. (7.189) becomes progressively smaller and eventually negligible relative to the first term. The mode of the posterior is thus indeed eventually (i.e., for suitably large $n$) entirely determined by $\bar{x}_w$.

## Sample Mean Estimation with Unknown Noise

In the two previous sections, we determined the posterior probability of the mean of a sample assuming a priori knowledge of the variance of the noise. What if the variance of the noise is in fact a priori unknown but can be assumed constant? Or what if the measured $x_i$ also contain a weak but unknown periodic signal? With such limited knowledge, one must conduct a generic analysis that disregards specific unknown details of the observables $x_i$. The CLT tells us that a sum of several processes should have a Gaussian distribution.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Alternatively, the Max Entropy principle informs us that the most conservative choice for the situation at hand is also a Gaussian distribution. Either way, as long as the noise is finite, we should once again be able use the data model

$$x_i = \mu + e_i \tag{7.194}$$

and assume the $e_i$ are Gaussian distributed with some unknown standard deviation $\sigma_0$. Since we now have two unknowns, we must first conduct our analysis to determine a posterior probability of both $\mu$ and $\sigma$. From Bayes' theorem, we get

$$p(\mu, \sigma|X, I) = \frac{p(\mu, \sigma|I)p(X|\mu, \sigma, I)}{p(X|I)}. \tag{7.195}$$

An estimate of $\mu$ can then be obtained by marginalization of this posterior probability against $\sigma$.

$$p(\mu|X, I) = \int p(\mu, \sigma|X, I)\, d\sigma. \tag{7.196}$$

Calculation of the posterior proceeds as in previous sections. Given we assumed the standard deviation of the measurement $\sigma_0$ is constant, albeit unknown, the likelihood of the data is given by Eq. (7.160). We also need a prior $p(\mu, \sigma|I)$ for $\mu$ and $\sigma$. Since little is known about the noise, it is safe to factorize the prior into a product

$$p(\mu, \sigma|I) = p(\mu|I)p(\sigma|I). \tag{7.197}$$

Let use the uninformative uniform prior PDF, Eq. (7.165), for $\mu$. Determination of $p(\sigma|I)$ requires additional considerations. The standard deviation $\sigma_0$ is positive definite and behaves, for all practical intents, as a scale parameter. It is also very unlikely to be infinite, as this would imply a signal that carries very large energy. It is thus sensible to assume a Jeffreys prior with minimum and maximum bounds $\sigma_L$ and $\sigma_H$, respectively.

$$p(\sigma|I) = \begin{cases} \frac{1}{\sigma_0 \ln(\sigma_H/\sigma_L)} & \text{for } \sigma_L \leq \sigma < \sigma_H \\ 0 & \text{elsewhere.} \end{cases} \tag{7.198}$$

Inserting the expressions for the priors and the likelihood in Eq. (7.196), the posterior $p(\mu|X, I)$ may now be calculated. One gets after simplification

$$p(\mu|X, I) = \frac{\int_{\sigma_L}^{\sigma_H} d\sigma_0 \sigma_0^{-(n+1)} \exp\left[-\frac{S}{2\sigma_0^2}\right]}{\int_{\mu_L}^{\mu_H} d\mu \int_{\sigma_L}^{\sigma_H} d\sigma_0 \sigma_0^{-(n+1)} \exp\left[-\frac{S}{2\sigma_0^2}\right]}. \tag{7.199}$$

In order to evaluate the preceding integrals, it is convenient to define $\zeta = S/2\sigma_0^2$. Insertion in Eq. (7.199) yields after a short calculation,

$$p(\mu|X, I) = \frac{\int_{\zeta_L}^{\zeta_H} S^{-n/2} \zeta^{n/2-1} e^{-\zeta} d\zeta}{\int_{\mu_L}^{\mu_H} \int_{\zeta_L}^{\zeta_H} S^{-n/2} \xi^{n/2-1} e^{-\zeta} d\zeta\, d\mu}, \tag{7.200}$$

which involves integrals that may be evaluated in terms of the incomplete gamma function. Although the integral in $\zeta$ depends on $\mu$ (through $S$ dependency on $\mu$), one can verify

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

that provided $\sigma_L \ll s$ and $\sigma_H \gg s$, it is essentially constant. Equation (7.200) thus approximately simplifies to

$$p(\mu|X, I) \approx \frac{S^{-n/2}}{\int_{\mu_L}^{\mu_H} S^{-n/2} d\mu}, \tag{7.201}$$

Substituting the value of $S$ given by Eq. (7.163), we get after some additional algebra

$$p(\mu|X, I) \approx \frac{\left[1 + \frac{(\mu-\bar{x})^2}{s^2}\right]^{-\frac{n}{2}}}{\int_{\mu_L}^{\mu_H} \left[1 + \frac{(\mu-\bar{x})^2}{s^2}\right]^{-\frac{n}{2}} d\mu}, \tag{7.202}$$

which, as it happens, features a structure rather similar to Student's $t$-distribution. To see this, first recall from §3.12.1 that Student's $t$-distribution is defined according to

$$p_t(t|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \tag{7.203}$$

where $\nu = n - 1$ is the number of degrees of freedom being sampled. Setting

$$\frac{t^2}{\nu} = \frac{(\mu - \bar{x})^2}{s^2}, \tag{7.204}$$

we find that the numerator of Eq. (7.202) is indeed structurally similar to Student's $t$-distribution. We can then write

$$p(\mu|X, I) \approx \frac{1}{I_s} \frac{1}{\sqrt{\pi(n-1)}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \left[1 + \frac{(\mu - \bar{x})^2}{s^2}\right]^{-n/2} \tag{7.205}$$

where the extra normalization factor $I_s$ is defined as an integral of Student's $t$-distribution,

$$I_s = \frac{1}{\sqrt{\pi(n-1)}} \frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})} \int_{\mu_L}^{\mu_H} \left[1 + \frac{(\mu - \bar{x})^2}{s^2}\right]^{-n/2} d\mu. \tag{7.206}$$

As in the case of the coefficients $I_{LH}$ and $I_w$ defined in previous sections, one finds that $I_s$ is essentially equal to unity for bounds of $\mu$ extending to $\pm\infty$, and $p(\mu|X, I)$ can then be strictly identified to a Student $t$-distribution. For a narrow range $\mu_L \leq \mu \leq \mu_H$, the functional shape remains the same, but the normalization is changed by the presence of the factor $I_s$. That means that the probability of $\mu$ being, say, in the interval $\mu \pm s$ in general exceeds that of the Student $t$-distribution and can even reach 100% if $\mu_L$ and $\mu_H$ are both very close to the actual value of $\mu$[9].

### 7.4.2 Sample Standard Deviation Estimation

Let us now turn to the determination of the variance (or standard deviation) of a sample $\{x_i\}$ of size $n$. As in the previous section, we can use the data to infer a joint posterior probability $p(\mu, \sigma|\vec{x}, I)$ for $\mu$ and $\sigma$, but we shall now marginalize this probability against

[9] Similarly to the case of the Gaussian distribution encountered in §7.4.1.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

$\mu$ to focus our attention on $\sigma$. We must then compute

$$p(\sigma|\vec{x}, I) = \frac{p(\sigma|I)\int p(\mu|I)p(\vec{x}|\mu, \sigma, I)\,d\mu}{p(\vec{x}|I)}, \tag{7.207}$$

where the priors and likelihood are the same as in the previous section. Substituting expressions for the priors and likelihoods, and simplifying factors common to the numerator and denominator, one gets

$$p(\sigma|\vec{x}, I) = \frac{\sigma^{-(n+1)}\exp\left(-ns^2/2\sigma^2\right)\int_{\mu_L}^{\mu_H} d\mu \exp\left(-\frac{n(\mu-\bar{x})^2}{2\sigma^2}\right)}{\int_{\sigma_L}^{\sigma_H} d\sigma\,\sigma^{-(n+1)}\exp\left(-ns^2/2\sigma^2\right)\int_{\mu_L}^{\mu_H} d\mu \exp\left(-\frac{n(\mu-\bar{x})^2}{2\sigma^2}\right)} \tag{7.208}$$

The integrals in $\mu$ would yield $\sqrt{2\pi/n}\sigma$ if the bounds $\mu_L$ and $\mu_H$ extended to $\pm\infty$. For finite values, we may then replace the integrals by $f\sqrt{2\pi/n}\sigma$, where $f$ is a constant in the range $0 \leq f \leq 1$ determined by $\mu_L$, $\mu_H$, $\sigma$, and $n$. The posterior thus simplifies to

$$p(\sigma|\vec{x}, I) = \begin{cases} K\sigma^{-n}\exp\left(-\frac{ns^2}{2\sigma^2}\right) & \text{for } \sigma_L \leq \sigma < \sigma_H \\ 0 & \text{elsewhere,} \end{cases} \tag{7.209}$$

with

$$K^{-1} = \int_{\sigma_L}^{\sigma_H} d\sigma\,\sigma^{-n}\exp\left(-\frac{ns^2}{2\sigma^2}\right) \tag{7.210}$$

It is illustrated in Figure 7.3 for $s^2 = 1$ and selected values of $n$. One observes that the function $p(\sigma|\vec{x}, I)$ is peaked at $\sigma = s$ and that it becomes progressively narrower with increasing values of $n$. Calculating and solving $\partial p/\partial\sigma = 0$ for $\sigma$, one verifies the mode of the distribution is indeed

$$\hat{\sigma} = s. \tag{7.211}$$

Moreover, computing the first and second moments of the distribution in terms of the inverse gamma integral, one finds

$$\langle\sigma\rangle = \frac{\sqrt{n}s\Gamma\left[(n-2)/2\right]}{\sqrt{2}\Gamma\left[(n-1)/2\right]} \tag{7.212}$$

$$\langle\sigma^2\rangle = \frac{ns^2}{n-1}. \tag{7.213}$$

One thus concludes that values of the mode $\hat{\sigma}$, the first moment $\langle\sigma\rangle$, and the root mean square $\sqrt{\langle\sigma^2\rangle}$ are distinct. However, one can readily verify that all three values asymptotically converge to $s$ in the large $n$ limit.

### 7.4.3 Bayesian Fitting with a Linear Model

#### Problem Definition

The basic Bayesian inference procedure we have used for the estimation of the mean and standard deviation of a dataset can be readily extended for the determination of the

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

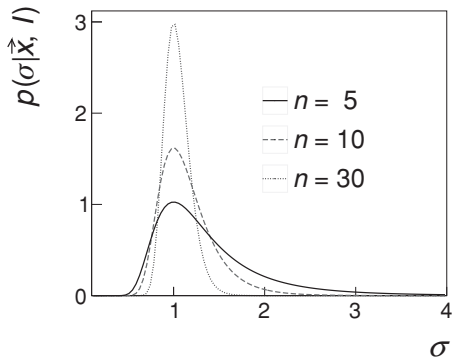


**Fig. 7.3** Evolution of the posterior $p(\sigma|\vec{x}, I)$, given by Eq. (7.209), with the sample size $n$ for $s^2 = 1$.

parameters of a model meant to describe the relation between a dependent observable $Y$ and an independent variable $X$.

Let us consider a set of $n$ measurements $\{x_i, y_i\}$, $i = 1, \ldots, n$. Let us posit that there exists a relation between the measured values $x_i$ and $y_i$ we can model according to

$$y_i = f(x_i|\vec{a}) + e_i, \tag{7.214}$$

where, as in previous sections, the terms $e_i$ represent Gaussian noise with null expectation value and known standard deviation $\sigma_i$, while $f(x_i|\vec{a})$ is a (physical) model function expressing the relation between $X$ and $Y$ but with $m$ unknown model parameters $\vec{a} = (a_1, a_2, \ldots, a_m)$.

## Calculation of the Likelihood

Bayesian fits with arbitrary models (i.e., models featuring nonlinear dependencies on parameters $\vec{a}$) will be considered in §7.6. In this section, we first consider generalized linear model functions of the form given by Eq. (5.90) encountered in §5.2.5:

$$f(x_i|\vec{a}) = \sum_{k=1}^{m} a_k f_k(x_i), \tag{7.215}$$

where the coefficients $f_k(x_i)$ are linearly independent functions of $x$ that do not depend on the model parameters $\vec{a}$. The noise terms may then be written

$$e_i = y_i - f(x_i|\vec{a}) \tag{7.216}$$

$$= y_i - \sum_{k=1}^{m} a_k f_k(x_i). \tag{7.217}$$

Here again, it is convenient to define the precision $\xi_i$ of each of the $n$ measurements according to

$$\xi_i = \frac{1}{\sigma_i^2}. \tag{7.218}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

If the noise terms $e_i$ are all mutually independent, the likelihood of the data is the product of each of their probabilities

$$p(\vec{y}|\vec{a}, I) = \prod_{i=1}^{n} \sqrt{\frac{\xi_i}{2\pi}} \exp\left[-\frac{\xi_i}{2}(y_i - f(x_i))^2\right] \tag{7.219}$$

$$= (2\pi)^{-n/2} \left(\prod_{i=1}^{n} \sqrt{\xi_i}\right) \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \xi_i (y_i - f(x_i))^2\right] \tag{7.220}$$

$$= (2\pi)^{-n/2} \left(\prod_{i=1}^{n} \sqrt{\xi_i}\right) \exp\left[-\frac{1}{2} S_w\right] \tag{7.221}$$

where, in the third line, we introduced the sum $S_w$ defined as

$$S_w = \sum_{i=1}^{n} \xi_i (y_i - f(x_i))^2 . \tag{7.222}$$

If the noise terms $e_i$ are correlated with a covariance matrix $\mathbf{V}$, the sum may instead be written

$$S_w = \sum_{i=1}^{n}\sum_{j=1}^{n} \left(y_i - \sum_{k=1}^{m} a_k f_k(x_i)\right) \left(V^{-1}\right)_{ij} \left(y_j - \sum_{k'=1}^{m} a_{k'} f_{k'}(x_i)\right), \tag{7.223}$$

which is identical to the $\chi^2$ function given by Eq. (5.92), for a linear model, discussed in §5.2.5 in the context of classical inference. For notational convenience, we once again introduce coefficients $F_{ik} \equiv f_k(x_i)$ as well as vector notations $\vec{y} = (y_1, \ldots, y_N)$, $\vec{a} = (a_1, \ldots, a_N)$, and matrix notations $\mathbf{F}$ and $\mathbf{V}$. The likelihood $p(\vec{y}|\vec{a}, I)$ then becomes

$$p(\vec{y}|\vec{a}, I) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\mathbf{V}|^{1/2}} \exp\left[-\frac{1}{2} S_w\right], \tag{7.224}$$

where $|\mathbf{V}|$ represents the determinant of the covariance matrix $\mathbf{V}$, and the quadratic form $S_w$ can now be written

$$S_w = \vec{y}^T \mathbf{V}^{-1} \vec{y} - \vec{a}^T \mathbf{F}^T \mathbf{V}^{-1} \vec{y} - \vec{y}^T \mathbf{V}^{-1} \mathbf{F} \vec{a} + \vec{a}^T \mathbf{F}^T \mathbf{V}^{-1} \mathbf{F} \vec{a}. \tag{7.225}$$

## Calculation of the Posterior

With the likelihood $p(\vec{y}|\vec{a}, I)$ and the expression, Eq. (7.225 ), for $S_w$ in hand, we now need prior probabilities for the parameters $\vec{a}$. Given the Gaussian form of the likelihood, two choices are readily possible and convenient: bounded uniform priors for all $a_i$ and conjugate Gaussian priors.

### Bounded Uniform Priors (Case 1)

Let us first consider bounded flat priors for each of the $m$ parameters:

$$p(\vec{a}|I) = \begin{cases} \prod_{i=1}^{m} R_i^{-1} & a_{i,\min} \leqslant a_i \leqslant a_{i,\max} \\ 0 & \text{elsewhere} \end{cases} \tag{7.226}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

where $R_i = a_{i,\max} - a_{i,\min}$. The posterior probability of parameters $\vec{a}$ may then be written

$$p(\vec{a}|\vec{y}, I) = K \exp(-S_w/2) \qquad \text{(flat priors)}$$

where $K$ is a constant determined by

$$K = \frac{p(\vec{a}|I)}{p(\vec{y}|I)} \tag{7.227}$$

with

$$p(\vec{y}|I) = \int_{R_i} \prod da_i p(\vec{a}|I) p(\vec{y}|\vec{a}, I) \tag{7.228}$$

$$= \prod_{i=1}^{m} R_i^{-1} (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \int_{R_i} \exp(-S_w/2). \tag{7.229}$$

The posterior $p(\vec{a}|\vec{y}, I)$, given by Eq. (7.227), constitutes the formal answer to the Bayesian inference optimization problem we set out to solve. It gives the probability density of all values of parameters $\vec{a}$ in parameter space. As such, it contains everything there is to know about the parameters $\vec{a}$ given a particular set of $n$ measured values $y_i$ and their respective measurement noise $\sigma_i$.

A full determination of $p(\vec{a}|\vec{y}, I)$ requires the cumbersome calculation of the integral in Eq. (7.229). However, much can be said about the parameters $\vec{a}$ without actually performing this integral. The PDF $p(\vec{a}|\vec{y}, I)$ typically spans a wide range of values in parameter space, but as per the physical model, Eq. (7.214), each of the $a_i$ should have a specific value, that is, $\vec{a}$ should be a specific vector in parameter space. Which vector should one choose? It stands to reason that one should choose the value $\vec{a}$ that maximizes the posterior $p(\vec{a}|\vec{y}, I)$. One is then interested in finding the (global) maximum or mode of this function in the model parameter space defined by the bounds $a_{i,\min} \leqslant a_i \leqslant a_{i,\max}$. This is achieved, as usual, by jointly setting derivatives of $p(\vec{a}|\vec{y}, I)$ to zero and solving for $\vec{a}$.

$$\frac{\partial p(\vec{a}|\vec{y}, I)}{\partial a_i} = 0 \quad \text{for } i = 1, \ldots, m. \tag{7.230}$$

Given our choice of flat prior, $\partial p(\vec{a}|\vec{y}, I)/\partial a_i$ is proportional to an exponential function of $-S_w/2$. There is thus no need to calculate the constant $K$, and optimization of the exponential amounts to minimization of $S_w$.

$$\frac{\partial S_w}{\partial a_i} = 0 \quad \text{for } i = 1, \ldots, m. \tag{7.231}$$

The optimization of the posterior $p(\vec{a}|\vec{y}, I)$ by Eq. (7.227) with uniform priors is thus strictly equivalent to the least-squares minimization problem we encountered in §5.2.5 (classical inference). Optimum estimators of the parameters $\vec{a}$ are thus given by

$$\hat{a} = \alpha^{-1} \vec{b}. \tag{7.232}$$

where, as we showed already in §5.2.5, the matrix $\alpha$ and the vector $\vec{b}$ are given by

$$\alpha = \mathbf{F}^T V^{-1} \mathbf{F}, \tag{7.233}$$

$$\vec{b} = \mathbf{F}^T V^{-1} \vec{y}. \tag{7.234}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Based on the similitude between the foregoing solution and the frequentist solution discussed in §5.2.5, it may appear as if Bayesian inference brings nothing new to the optimization problem. But that is not quite correct. Bayesian inference in fact provides at least two advantages. The first has to do with the Bayesian interpretation of probability: the posterior $p(\vec{a}|\vec{y}, I)$ expresses the degree of belief of observing the parameters $a_i$ in ranges $[a_i, a_i + da_i]$. One can then directly calculate credible ranges for the parameters $a_i$ by simple integration and marginalization of the posterior $p(\vec{a}|\vec{y}, I)$. These integrals then represent the probability of observing the parameters $a_i$ in such ranges, as we discuss in more detail in §7.4.3. Second, Bayesian inference enables the determination of posteriors when prior information about a system a priori restricts the ranges of the parameters. Bayesian inference then effectively combines old and new information toward the determination of optimum parameters. Such a combination is not formally defined in the context of the frequentist paradigm.

### Linear Model Fit with Gaussian Priors (Case 2)

Let us now consider that an experiment might be the successor of one or several prior experiments. Instead of using flat priors for the parameters $a_i$, one might then use the (combined) outcome of these previous experiments, which we may assume have yielded a Gaussian posterior with optimum values $\vec{a}_0$ and a covariance matrix $Q$. Let us then define a generic Gaussian prior as

$$p(\vec{a}|I) = \frac{1}{(2\pi)^m |\mathbf{Q}|^{1/2}} \exp\left[-\frac{1}{2}(\vec{a} - \vec{a}_0)^T Q^{-1} (\vec{a} - \vec{a}_0)\right], \tag{7.235}$$

where $m$ is the number of parameters $\vec{a}$ (i.e., the dimension of $\vec{a}$), and $|\mathbf{Q}|$ is the determinant of the covariance matrix $\mathbf{Q}$. The likelihood, Eq. (7.224), remains unchanged but the posterior must be modified to account for the Gaussian priors. One gets

$$p(\vec{a}\,|\vec{y}, I) = \frac{p(\vec{a}\,|I)p(\vec{y}\,|\vec{a}, I)}{p(\vec{y}|I)} \tag{7.236}$$

$$= K \exp\left[-\frac{1}{2}(\vec{a} - \vec{a}_0)^T Q^{-1} (\vec{a} - \vec{a}_0)\right] \exp\left(-S_w/2\right), \tag{7.237}$$

where $K$ is a constant determined by the global likelihood $p(\vec{y}|I)$. The preceding product of exponentials may be reduced to a single exponential by addition of their arguments, which we write

$$p(\vec{a}|Y, I) = K \exp\left(-S'_w/2\right) \tag{7.238}$$

with

$$S'_w = (\vec{a} - \vec{a}_0)^T Q^{-1} (\vec{a} - \vec{a}_0) + S_w. \tag{7.239}$$

Expanding the terms, and rearranging, $S'_w$ may be written

$$S'_w = \vec{y}^T \mathbf{V}^{-1} \vec{y} + \vec{a}_0^T \mathbf{Q}^{-1} \vec{a}_0 - \vec{a}^T \left(\mathbf{Q}^{-1} \vec{a}_0 + \mathbf{F}^T \mathbf{V}^{-1} \vec{y}\right) \tag{7.240}$$

$$- \left(\vec{a}_0^T \mathbf{Q}^{-1} + \vec{y}^T \mathbf{V}^{-1} \mathbf{F}\right) \vec{a} + \vec{a}^T \left(\mathbf{Q}^{-1} + \mathbf{F}^T \mathbf{V}^{-1} \mathbf{F}\right) \vec{a}. \tag{7.241}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

It is then possible (Problem 7.6) to show that the optimal solution for $\vec{a}$ is of the form of Eq. (7.232) but with modified values of the matrix $\alpha$ and vector $\vec{b}$ as follows:

$$\alpha' = \mathbf{F}^T V^{-1}\mathbf{F} + \mathbf{Q}^{-1} \tag{7.242}$$

$$\vec{b}' = \mathbf{Q}^{-1}\vec{a}_0 + \mathbf{F}^T\mathbf{V}^{-1}\vec{y}. \tag{7.243}$$

## Errors and Credible Ranges

Let $\delta a_k = a_k - \hat{a}_k$ represent the difference between an arbitrary value $a_k$ and the optimal estimate obtained by optimization of the posterior $p(\vec{a}|Y, I)$ and let $S_{\min}$ be the value of $S_w$ corresponding to that optimum. In order to make a statement about parameter errors and credible ranges, it is convenient to express the value of $S_w$ for arbitrary parameters $\vec{a}$ as a Taylor expansion:

$$S_w(\delta\vec{a}) = S_{\min} + \sum_{k=1}^{m} \left.\frac{\partial S_w}{\partial a_k}\right|_{\min} \delta a_k + \frac{1}{2}\sum_{k,k'=1}^{m} \left.\frac{\partial^2 S_w}{\partial a_k \partial a_{k'}}\right|_{\min} \delta a_k \delta a_{k'}. \tag{7.244}$$

First note that the second term of this expression is null by construction because the first derivative is evaluated at the minimum. Additionally note that the series does not involve terms of third or higher order because Eq. (7.225) does not contain such terms. Let us define $\Delta S_w = S_w - S_{\min}$. From Eq. (7.244), we get

$$\Delta S_w = \frac{1}{2}\sum_{k,k'=1}^{m} \left.\frac{\partial^2 S_w}{\partial a_k \partial a_{k'}}\right|_{\min} \delta a_k \delta a_{k'}. \tag{7.245}$$

### Flat Parameter Priors (Case 1)

For flat parameter model priors, calculation of second derivatives of the solution Eq. (7.225) yields

$$\Delta S_w = \delta\vec{a}^T \mathbf{F}^T \mathbf{V}^{-1}\mathbf{F}\delta\vec{a}. \tag{7.246}$$

We can then write the posterior $p(\vec{a}|Y, I)$ as

$$p(\vec{a}|Y, I) = K' \exp\left(-\frac{1}{2}\Delta S_w\right) \tag{7.247}$$

where $K'$, given by

$$K' = K \exp\left(-\frac{1}{2}S_{\min}\right), \tag{7.248}$$

is just a constant. It then becomes relatively straightforward to determine the covariance matrix $\mathbf{U}$ of the parameters $\vec{a}$ as it may be calculated in terms of the variance $\mathbf{V}$ of the data points $\vec{y}$ as follows:

$$\mathbf{U} = \left[\frac{d\vec{a}}{d\vec{y}}\right] \mathbf{V} \left[\frac{d\vec{a}}{d\vec{y}}\right]^T. \tag{7.249}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

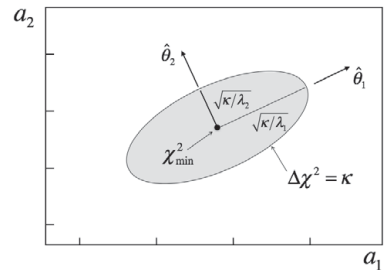


**Fig. 7.4** $\Delta\chi^2 = \kappa$ ellipse in $a_1$ vs. $a_2$ parameter space.

Using Eqs. (7.233) and (7.234), calculation of the derivative $d\vec{a}/d\vec{y}$ yields

$$\frac{d\vec{a}}{d\vec{y}} = \left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{V}^{-1}. \tag{7.250}$$

Substitution of this expression in Eq. (7.249) yields

$$\mathbf{U} = \left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{V}^{-1}\mathbf{V}\left[\left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{V}^{-1}\right]^T. \tag{7.251}$$

The matrix $\mathbf{V}$ and its inverse are symmetric by construction. Remembering that the transpose of a product of matrices is equal to the product of the transposed in reversed order (e.g., $[\mathbf{ABC}]^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T$), Eq. (7.251) simplifies and we thus conclude that

$$\mathbf{U} = \left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}\right)^{-1}. \tag{7.252}$$

We find that the expression of $\mathbf{U}^{-1} = \mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}$ is identical to $\alpha$, given by Eq. (7.233). We can then write $p(\vec{a}|\vec{y}, I)$ as a multidimensional Gaussian of the form

$$p(\vec{a}|Y, I) = K' \exp\left(-\frac{1}{2}\delta\vec{a}^T\mathbf{U}^{-1}\delta\vec{a}\right), \tag{7.253}$$

where $\mathbf{U}$ represents the covariance matrix of the model parameters $\vec{a}$ we calculated earlier.

Since $\mathbf{U}$ is a symmetric matrix, the principal axis theorem tells us there exists a transformation $\delta\vec{a} = \mathbf{O}\delta\vec{\theta}$ that transforms $\delta\vec{a}^T\mathbf{U}^{-1}\delta\vec{a}$ into $\delta\vec{\theta}^T\mathbf{\Lambda}\vec{\theta}$ and such that $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{U}$. It is thus always possible to transform the parameters $\vec{a}$ to uncorrelated parameters $\vec{\theta}$. For $m = 2$, this takes the form

$$\Delta S_w = \begin{pmatrix}\delta\theta_1 & \delta\theta_2\end{pmatrix}\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}\begin{pmatrix}\delta\theta_1\\ \delta\theta_2\end{pmatrix} \tag{7.254}$$

$$= \lambda_1\delta\theta_1^2 + \lambda_2\delta\theta_2^2, \tag{7.255}$$

where $\lambda_1$ and $\lambda_2$ are the eigenvalues of $\mathbf{U}$. Clearly, $\Delta S_w = \kappa$ defines an ellipse

$$1 = \frac{\theta_1^2}{\kappa/\lambda_1} + \frac{\theta_2^2}{\kappa/\lambda_2}, \tag{7.256}$$

as illustrated in Figure 7.4.

In general, fixed values of $\Delta S_w$ define credible regions consisting of a simple range in one dimension (i.e., a single parameter), ellipses in two dimensions, ellipsoids in three,

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

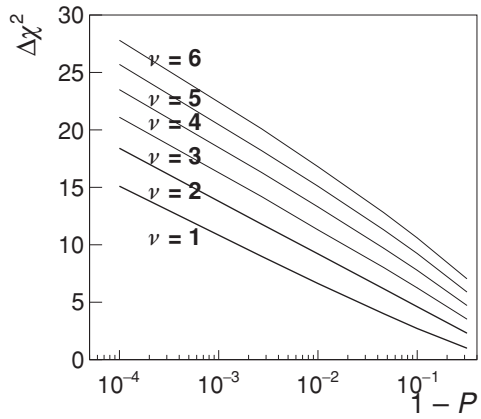


**Fig. 7.5** Increment $\Delta\chi^2$ vs. $1 - P$ for selected values of the number of degrees of freedom $\nu$ (fit parameters).

and hyper-ellipsoids in more than three dimensions. The probability $P$ associated with such credible regions is obtained by integration of the posterior within the boundary of the ellipse (or ellipsoid in $m = 3$ or hyper-ellipsoid $m \geq 4$ dimensions) defined by $\Delta S_w = \kappa_{\rm crit}$

$$P = \int_{\Delta S_w < \kappa_{\rm crit}} p(\vec{a}|Y, I) \prod_k da_i. \tag{7.257}$$

Since $S_w$ is by construction a $\chi^2$ variable, one can show that the probability $P$ is given by

$$P = 1 - \frac{\gamma(m/2, \kappa/2)}{\Gamma(m/2)}, \tag{7.258}$$

where $\gamma(\alpha, \beta)$ is the incomplete gamma function. For instance, a region containing a probability of $P = 68.3\%$ corresponds to a $\Delta\chi^2 = 1$ for $m = 1$ and $\Delta\chi^2 = 2.3$ for $m = 2$, while for $P = 90.0\%$ one has $\Delta\chi^2 = 2.71$ and 4.61 for $m = 1$ and $m = 2$, respectively. Other values of $\Delta\chi^2$ are shown as a function of $1 - P$ for several values of the number of degrees of freedom $\nu$ in Figure 7.5.

Recall from our discussion in §6.1.2 that the notion of error is not absolute but predicated by a choice of probability content. In the frequentist approach, with one parameter and Gaussian errors, a probability content of 68.3% determines a $\pm 1\sigma$ interval which may then be interpreted as the error on the parameter. Choosing the probability content to be 95.45% instead, one gets a $\pm 2\sigma$ error interval. In the Bayesian approach discussed in this section, the notion of error is also readily derived from the probability content within selected intervals. However, rather than being called confidence intervals, the selected intervals are typically named **credible intervals** and they are defined by integration of the posterior which yields a given probability content. The interpretation of the credible interval is also far more direct. Because the posterior is the probability density of the parameters $a_i$, integrals over selected intervals of the parameters (e.g., credible range) of the posterior correspond to the probability of finding the true value of the parameter within the selected interval. It then becomes possible to define credible ranges associated with probability of 68%, 96%, or any other probability. Equation (7.258) then provides a one-to-one relation

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

between the probability content and the increment $\kappa = \Delta\chi^2$ defining its associated credible range. It should be stressed that for $m \geq 2$ parameters, the credible range is not a simple interval but typically a hyper-ellipsoidal locus in parameter space. It is thus not formally possible (or correct) to identify the error on a given parameter by stating a single interval such as $\mu \pm \delta\mu$. One must produce the whole ellipsoidal region, which for $m = 2$ reduces to an ellipse, as illustrated in Figure 7.4. However, it is also possible to obtain the error on a single parameter by marginalizing all the others, as we discuss in §7.4.3.

### Gaussian Conjugate Priors (Case 2)

Calculation of the (modified) covariance matrix $\mathbf{U}'$ of the model parameters $\vec{a}'$ obtained with Gaussian priors proceeds similarly to the calculation of $\mathbf{U}$. Using the modified solutions, Eq. (7.242), one finds

$$\mathbf{U}' = \left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F} + \mathbf{Q}^{-1}\right)^{-1}\mathbf{F}^T\mathbf{V}^{-1}\mathbf{V}\left[\left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F} + \mathbf{Q}^{-1}\right)^{-1}\mathbf{F}^T\mathbf{V}^{-1}\right]^T, \tag{7.259}$$

$$= \left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F} + \mathbf{Q}^{-1}\right)^{-1}\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}\left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F} + \mathbf{Q}^{-1}\right)^{-1}. \tag{7.260}$$

Substituting the inverse of the parameter covariance, $\mathbf{U}^{-1} = \mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}$, obtained for flat priors, Eq. (7.252), we get

$$\mathbf{U}' = \left(\mathbf{U}^{-1} + \mathbf{Q}^{-1}\right)^{-1}\mathbf{U}^{-1}\left(\mathbf{U}^{-1} + \mathbf{Q}^{-1}\right)^{-1}. \tag{7.261}$$

Inspection of this expression shows that the fit modifies (shifts) the prior $a_0$ and reduces the prior variances $Q_{ii}$ of the parameters. These effects are easiest to identify when reducing the scope of the problem to a single parameter. Let $Q \equiv \sigma_0^2$ represent the variance of the parameter observed by the previous experiment (prior knowledge) and let $\mathbf{U} = \left(\mathbf{F}^T\mathbf{V}^{-1}\mathbf{F}\right)^{-1} \equiv \sigma^2$ be the variance of the parameter determined with uniform priors (or from the likelihood function alone). The original solution $\vec{a} = \alpha^{-1}b$ may be written $a = \sigma^2 b$, or $b = a/\sigma^2$. The modified solution is then $a' = (\sigma_0^{-2} + \sigma^{-2})^{-1}(a_0\sigma_0^{-2} + a\sigma^{-2})$ which simplifies to

$$a' = a_0 + \zeta\,(a - a_0)\,, \tag{7.262}$$

where

$$\zeta = \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}. \tag{7.263}$$

We find, similarly to Kalman filters, that the (new) value $a'$ obtained from the fit is equal to the prior value, $a$, plus a term proportional to the difference $a - a_0$ times a "gain" factor $\zeta$. Effectively, the parameter value obtained with the inclusion of the prior information is shifted by an amount proportional to the difference between the prior value of the parameter and that obtained on the basis of the likelihood alone. The information carried by the prior carries little weight, however, if the variance of the measurement, $\sigma^2$, is much smaller than the variance $\sigma_0^2$ of the prior. Indeed, for $\sigma^2 \ll \sigma_0^2$, one gets $\zeta \to 1$ and $a' \approx a$. Similarly,

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

for a single parameter, the variance $\mathbf{U}'$, Eq. (7.261), may be written

$$\sigma'^2 = \left(\sigma^{-2} + \sigma_0^{-2}\right)^{-1} \sigma^{-2} \left(\sigma^{-2} + \sigma_0^{-2}\right)^{-1}, \tag{7.264}$$

which simplifies to

$$\sigma' = \sigma \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}. \tag{7.265}$$

The posterior variance is determined by a combination of the prior variance $\sigma_0^2$ and the variance $\sigma^2$ of the measurement. The posterior variance is reduced relative to the prior variance thanks to new knowledge acquired with the measurement. If the procedure is repeated several times, each time with a measurement variance $\sigma^2$, the posterior of one measurement can be used as the prior of the next. Initially, the prior variance $\sigma_0^2$ is likely to be much larger than the variance of the measurement $\sigma^2$ and a single measurement shall then produce a posterior variance $\sigma^2$ significantly smaller than the prior's. However, with successive iterations, the posterior variance shall eventually become smaller than measurement variance and successive iterations shall then produce only a small gain in precision. In effect, the precision shall then increase in proportion to the number $n$ of measurements, which means $\sigma' \propto \sigma/\sqrt{n}$.

For fits involving $m \geq 2$ parameters, shifts in parameter values and width reductions are similarly obtained, albeit somewhat less transparently.

## Parameter Marginalization and Errors

Marginalization is required whenever one wishes to determine the probability distribution of one parameter (or several) irrespective of the other parameters. It enables the determination of the credible range, and thus the error, of the parameter(s) of interest. For a model involving $m$ parameters, marginalization may be achieved iteratively by integration of Eq. 7.253 one parameter at a time.

Let us exemplify the procedure for the marginalization of a single parameter, which, for the sake of notational simplicity, we choose to be the first one. The posterior of $m$ parameters marginalized against the first parameter is given by

$$p(a_2, \ldots, a_m|\vec{y}, I) = \int_{\Delta a_1} p(a_1, \ldots, a_m|\vec{y}, I)\, da_1, \tag{7.266}$$

where $\vec{y}$ represents the measured data and the integration is carried over the domain $\Delta a_1$ of the parameter $a_1$. We now substitute the posterior, Eq. (7.247), obtained in the previous section and write

$$p(a_2, \ldots, a_m|\vec{y}, I) = K' \int_{\Delta a_1} \exp\left(-\frac{1}{2}\Delta\chi^2\right) da_1, \tag{7.267}$$

$$= K' \int_{\Delta a_1} \exp\left(-\frac{1}{2}\delta\vec{a}^T \mathbf{U}^{-1} \delta\vec{a}\right) d\delta a_1, \tag{7.268}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

In order to carry out the integral in $\delta a_1$, we must factorize the terms in $a_1$ from other terms. To accomplish this, let us first focus on the argument $\Delta\chi^2$ of the exponential and write

$$\Delta\chi^2 = \delta\vec{a}^T \mathbf{U}^{-1} \delta\vec{a} \tag{7.269}$$

$$= \sum_{k,k'=1}^{m} \delta a_k (U^{-1})_{kk'} a_{k'} \tag{7.270}$$

$$= (U^{-1})_{11}\delta a_1^2 + 2\delta a_1 \sum_{k=2}^{m} (U^{-1})_{1k}\delta a_k + \sum_{k,k'=2}^{m} \delta a_k (U^{-1})_{1k}\delta a_{k'}, \tag{7.271}$$

where in the third line, we have explicitly separated terms in $k, k' = 1$ from others. Completing the square formed by the first two terms of this expression, we get

$$\begin{aligned}\Delta\chi^2 &= (U^{-1})_{11}\delta a_1^2 + 2\delta a_1 \sum_{k=2}^{m} (U^{-1})_{1k}\delta a_k + \frac{1}{(U^{-1})_{11}} \left( \sum_{k=2}^{m} (U^{-1})_{1k}\delta a_k \right)^2 \\ &\quad - \frac{1}{(U^{-1})_{11}} \left( \sum_{k=2}^{m} (U^{-1})_{1k}\delta a_k \right)^2 + \sum_{k,k'=2}^{m} \delta a_k (U^{-1})_{1k}\delta a_{k'}, \\ &= (U^{-1})_{11} \left[ \delta a_1 + \frac{1}{(U^{-1})_{11}} \sum_{k=2}^{m} (U^{-1})_{1k}\delta a_k \right]^2 + \Delta\chi_r^2, \end{aligned} \tag{7.272}$$

where we have introduced a "reduced" $\Delta\chi^2$ defined as

$$\Delta\chi_r^2 = -\frac{1}{(U^{-1})_{11}} \left( \sum_{k=2}^{m} (U^{-1})_{1k}\delta a_k \right)^2 + \sum_{k,k'=2}^{m} \delta a_k (U^{-1})_{kk'} a_{k'}. \tag{7.273}$$

The marginalized posterior we seek is thus

$$p(a_2, \ldots, a_m | Y, I) = K' \exp\left( -\frac{1}{2}\Delta\chi_r^2 \right) \tag{7.274}$$

$$\times \int_{\Delta a_1} \exp\left( -\frac{1}{2}(U^{-1})_{11} \left[ \delta a_1 + \frac{1}{(U^{-1})_{11}} \sum_{k=2}^{m} (U^{-1})_{1k}\delta a_k \right)^2 \right] d\delta a_1.$$

The integrand is a Gaussian in $\delta a_1$ with a variance $\sigma^2 = 1/(U^{-1})_{11}$. For an infinite range of integration, the integral yields $\sqrt{2\pi}\sigma = \sqrt{2\pi/(U^{-1})_{11}}$. If the range is finite but far exceeds the region where the integrand is finite, $\sqrt{2\pi/(U^{-1})_{11}}$ constitutes a very good approximation of the true value of the integral. We can then write

$$p(a_2, \ldots, a_m | Y, I) = K'' \exp\left( \frac{1}{2}\Delta\chi_r^2 \right), \tag{7.275}$$

with $K'' = K'\sqrt{2\pi/(U^{-1})_{11}}$.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Let us consider a specific case involving $m = 2$ parameters only. The reduced $\chi^2$ is then

$$\Delta\chi_r^2 = -\frac{1}{(U^{-1})_{11}}\left[(U^{-1})_{12}\delta a_2\right]^2 + (U^{-1})_{22}\delta a_2^2 \tag{7.276}$$

$$= \delta a_2^2 \frac{\left[(U^{-1})_{11}(U^{-1})_{22}\right] - (U^{-1})_{12}^2}{(U^{-1})_{11}}. \tag{7.277}$$

Since the inverse of an inverse matrix is the matrix itself, the ratio of the right-hand side may be recognized as the multiplicative inverse of the matrix element $U_{22}$ which one may write $\sigma_2^2$. We thus conclude that the marginal probability $p(a_2|Y, I)$ is a Gaussian with a variance $U_{22} = \sigma_2^2$. The matrix $U^{-1}$ is indeed the inverse of the covariance matrix of the parameters $a_1$ and $a_2$. After marginalization of $a_1$, the estimate of $a_2$ and its associated error may then be written

$$a_2 \pm \sigma_2. \tag{7.278}$$

Repeating the reasoning for marginalization of parameter $a_2$, we conclude similarly that the estimate of $a_1$ and associated error may then be written

$$a_1 \pm \sigma_1. \tag{7.279}$$

The error $\sigma_k$ reflect the credible range obtained for a specific probability content $P$ after marginalization of the other parameter. Indeed, it is important to reiterate that the errors $\pm\sigma_1$ and $\pm\sigma_2$ are applicable only after marginalization, that is, they apply only when a single parameter is considered after the other has been marginalized. If no marginalization is performed, the joint error on the parameters is described by the full ellipse, Eq. (7.256), and the individual ranges $\pm\sigma_1$ and $\pm\sigma_2$ are not strictly applicable as (joint) errors on $a_1$ and $a_2$.

## 7.5 Bayesian Inference with Nonlinear Models and Non-Gaussian Processes

As we saw in the previous section, Bayesian fits of linear models of the form given by Eq. (7.215), with Gaussian measurement noise, reduce to problems of $\chi^2$ minimization that can be expressed in terms of linear matrix equations readily solvable by standard linear algebra techniques. Alas, linear models are not the norm in scientific studies. One indeed often encounters models that cannot be linearized, that is, models whose free parameters $\vec{\theta} = (\theta_1, \ldots, \theta_m)$ cannot be expressed in terms of linear equations of the form Eq. (7.215). Commonly encountered examples of such nonlinear models include Gaussian and Breit–Wigner distributions with polynomial backgrounds. Additionally, one finds that measurements fluctuations, or noise, may also be strictly non-Gaussian. This is the case, for instance, for measurements of rare particle production cross sections in nuclear scattering experiments, which can be described in terms of Poisson and Bernoulli statistics. Poisson statistics is used to describe the stochastic nature of rare particle production

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

processes whereas Bernoulli statistics (binomial distribution) is used to account for instrumental effects such as particle losses (detection efficiency). We must thus seek and develop techniques of Bayesian inference amenable to both nonlinear physical models and non-Gaussian data probability models. The basic goal of Bayesian inference nonetheless remains the same and involves calculation of the posterior probability of the model parameters, $p(\vec{\theta}|D, M, I)$. Unfortunately, the determination of the mode (maximum probability density) of the posterior and credible ranges of the model parameters typically becomes nontrivial problem that must be solved by numerical techniques.

We consider Bayesian inference with nonlinear physical models in §7.5.1, whereas non-Gaussian data probability models are introduced in §7.5.2. General optimization techniques to determine the mode of posterior PDFs are discussed in §7.6.

### 7.5.1 Bayesian Inference with Nonlinear Models

Consider an experiment reporting $n$ data points $(x_i, y_i)$, $i = 1, \ldots, n$. We shall assume the data may be described with a data model $M$ of the form

$$y_i = f(x_i|\vec{\theta}) + e_i, \tag{7.280}$$

where $f(x_i|\vec{\theta})$ is a physical model expressing a relation between an independent observable $X$ and a dependent observable $Y$ with $m$ free or unspecified model parameters $\vec{\theta} = (\theta_1, \ldots, \theta_m)$. The model $f(x_i|\vec{\theta})$ is considered nonlinear in its parameter $\vec{\theta}$ if it cannot be linearized and expressed in the form of Eq. (7.215). Let us further assume that the noise terms $e_i$ have null expectation values and are Gaussian distributed with a known covariance $\mathbf{V}$. The posterior may then be written

$$p(\vec{\theta}|\vec{y}, M, I) = \frac{p(\vec{\theta}|M, I)}{p(\vec{y}|M, I)} \frac{1}{(2\pi)^{N/2}|\mathbf{V}|^{1/2}} \times \exp\left[\frac{1}{2}\sum_{i,j}(y_i - f(x_i))\, V_{ij}^{-1}\left(y_j - f(x_j)\right)\right], \tag{7.281}$$

where $p(\vec{y}|M, I)$ is the global likelihood of the data. This probability density should yield a maximum for the optimal value of the parameters $\vec{\theta}$, denoted $\hat{\theta}$, corresponding to the mode of the distribution. Finding this maximum is an optimization problem, which may be addressed with the numerical techniques presented in §7.6. Given such an optimum $\hat{\theta}$, and based on our treatment of linear models, it is possible to write

$$p(\vec{\theta}|\vec{y}, M, I) = K \exp\left(-\frac{1}{2}\delta\vec{\theta}^{\,T}\mathbf{I}\delta\vec{\theta}\right) + O(3), \tag{7.282}$$

where $\delta\hat{\theta} = \vec{\theta} - \hat{\theta}$, $O(3)$ represents contributions of higher orders admissible for a nonlinear model, while the constant $K$ is determined by the prior $p(\hat{\theta}|M, I)$ and likelihood $\mathbb{L}(\hat{\theta}) \equiv p(\vec{y}|\hat{\theta}, M, I)$ at the mode, as well as the global likelihood $p(\vec{y}|M, I)$:

$$K = \frac{p(\vec{\theta}|M, I)}{p(\vec{y}|M, I)}\mathbb{L}(\hat{\theta}), \tag{7.283}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

and $\mathbf{I}$ is known as Fisher information matrix. $\mathbf{I}$ may be numerically evaluated using Eqs. (7.281, 7.281). One first uses Eq. (7.281), based on the measured data $(x_i, y_i)$, $i = 1, \ldots, n$, to compute $p(\vec{\theta}|\vec{y}, M, I)$. Taking the logarithm of Eq. (7.282), one next gets

$$\ln\left[p(\vec{\theta}|\vec{y}, M, I)\right] = \ln(K) - \frac{1}{2}\left(\delta\vec{\theta}^T \mathbf{I} \delta\vec{\theta}\right). \tag{7.284}$$

Second derivatives of $\ln[p(\vec{\theta}|\vec{y}, M, I)]$, evaluated at $\hat{\theta}$, then yield

$$\mathbf{I}_{ij} = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln\left[p(\vec{\theta}|\vec{y}, M, I)\right]\Big|_{\hat{\theta}}. \tag{7.285}$$

In the context of a linear model, as we saw in previous sections, the inverse of $\mathbf{I}$ would correspond to the covariance matrix of the model parameters. For nonlinear models, Eq. (7.282) is only approximate, and the neglect of higher-order terms implies $\mathbf{I}^{-1}$ is not equal, strictly speaking, to the covariance matrix of the parameters. In practice, however, one finds that for sufficiently large datasets, the Gaussian approximation embodied in Eq. (7.282) may be sufficient and yield an inverse matrix, $\mathbf{I}^{-1}$, which provides a useful approximation of the covariance matrix of the model parameters $\vec{\theta}$.

Equation (7.282) may also be used to evaluate the global hypothesis posterior $p(M|\vec{y}, I)$ by marginalization of all model parameters. This is useful, for instance, in hypotheses testing based on the odds ratio of the global posteriors of competing hypotheses. However, given the Fisher information matrix $\mathbf{I}$ is in general nondiagonal, integration over all parameters $\vec{\theta}$ may become nontrivial and rather tedious, particularly in the presence of finite bounds on each of the parameters. Integration may be achieved, nonetheless, based on an eigenvalue decomposition of the matrix. Since $\mathbf{I}$ is by definition real and symmetric, there indeed exists a transformation $\delta\vec{\theta} = \mathbf{O}\delta\vec{\xi}$ such that

$$\delta\vec{\theta}^T \mathbf{I} \delta\vec{\theta} = \delta\vec{\xi}^T \mathbf{\Lambda} \delta\vec{\xi}, \tag{7.286}$$

where $\Lambda$ is a diagonal matrix consisting of eigenvalues $\lambda_k$ of $\mathbf{I}$. The integral we seek is thus

$$\begin{aligned}
\mathbb{I} &= \int \prod_k d\theta_k \exp\left(-\frac{1}{2}\delta\vec{\theta}^T \mathbf{I} \delta\vec{\theta}\right) \\
&= J \int \prod_k d\xi_k \exp\left(-\frac{1}{2}\delta\vec{\xi}^T \mathbf{\Lambda} \delta\vec{\xi}\right) \\
&= \prod_k \int_{\xi_{k,\min}}^{\xi_{k,\max}} d\xi_k \exp\left(-\frac{1}{2}\delta\xi_k \lambda_k \delta\xi_k\right),
\end{aligned} \tag{7.287}$$

where, in the third line, we used $J = \left|\frac{\partial\theta}{\partial\xi}\right| = \det\mathbf{O} = 1$, and the fact $\mathbf{\Lambda}$ is a diagonal matrix consisting of values $\lambda_k$. The boundaries of integration $\xi_{k,\min}$ and $\xi_{k,\max}$ are obtained from those on $\vec{\theta}$ based on the inverse transformation $\delta\vec{\xi}_{\min,\max} = \mathbf{O}^{-1}\delta\vec{\theta}_{\min,\max}$.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

### 7.5.2 Bayesian Inference with Non-Gaussian Processes

There exists a plurality of phenomena (or systems) for which a Gaussian data probability model is not suitable or applicable. Of particular interest in astronomy and high-energy physics are processes involving Poisson, binomial, or multinomial sampling. Unfortunately, the difficulty arises that, for such data probability models, it is typically not possible to reduce the calculation of the posterior to the evaluation of a single quantity such as a $\chi^2$ function, and one must then handle the full posterior probability distribution. Considerable simplifications are possible, however, if the parameters of the probability data model are constant and the prior probability of these parameters may be expressed as conjugate priors. A more general example involving model parameters dependencies on a control variable is considered in §7.5.3.

We limit our discussion to Poisson processes but the techniques introduced in the following can be straightforwardly extended to other types of statistical processes.

#### Finding the Rate Parameter of a Poisson Process

Poisson sampling describes phenomena with a specific rate, that is, an average number of instances, or events, per unit of time (or dependency on a specific control variable). Let $r$ be the process rate. If observations are carried out over a time period of $T$, one then expects, on average, to observe $\langle n \rangle \equiv \mu = rT$ instances of the phenomenon. The actual number of instances observed in any given interval, $n$, should fluctuate (i.e., vary interval by interval) according to a Poisson distribution

$$p(n|\mu, I) = \frac{\mu^n e^{-\mu}}{n!}. \tag{7.288}$$

If the rate $r$ is a priori unknown but can be sensibly expected to be constant over extended time periods, one can use measurements of the number of occurrences, $n_k$, in $m$ different time intervals of duration $T$ to estimate $r$. For notational convenience, let us denote these $m$ measurements in terms of a vector $\vec{n} = (n_1, \ldots, n_m)$.

The rate $r$ may be trivially computed on the basis of the mean number of instances, $\hat{\mu}$, and the duration of the time interval $T$ during which events are counted:

$$r = \hat{\mu}/T. \tag{7.289}$$

We thus focus on the determination of $\hat{\mu}$. A frequentist evaluation of $\hat{\mu}$ could of course be readily obtained from the mean of the measured values $n_k$, but we here consider a Bayesian estimation instead. That means we need to calculate the posterior probability of $\mu$:

$$p(\mu|\vec{n}, I) = \frac{p(\mu|I)p(\vec{n}|\mu, I)}{p(\vec{n}|I)}, \tag{7.290}$$

where $p(\mu|I)$ is the prior probability of $\mu$, $p(\vec{n}|\mu, I)$ is the likelihood of the data given $\mu$, and $p(\vec{n}|I)$ is the global likelihood of the data.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

The likelihood $p(\vec{n}|\mu, I)$ is readily evaluated by multiplying the probabilities of each of the observed values $n_k$, $k = 1, \ldots, m$:

$$\begin{aligned} p(\vec{n}|\mu, I) &= \prod_{k=1}^{m} p(n_k|\mu, I), \\ &= \prod_{k=1}^{m} \frac{\mu^{n_k} e^{-\mu}}{n_k!}, \\ &= \frac{\mu^z e^{-m\mu}}{\prod_{k=1}^{m} n_k!}, \end{aligned} \tag{7.291}$$

where we have introduced the variable $z$ defined as a sum of the measured $n_k$

$$z = \sum_{k=1}^{m} n_k. \tag{7.292}$$

Recall from §7.3.3 that for a Poisson data probability model, it is convenient to use a conjugate prior for the rate parameter in the form of a gamma distribution (§3.6). We thus write

$$p(\mu|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\beta\mu}, \tag{7.293}$$

where the distribution parameters $\alpha$ and $\beta$ may be chosen to reflect an uninformative scaling prior (e.g., $\alpha = 1.0$ and $\beta = 1.0$) or may be constrained by prior measurements or knowledge of the phenomenon of interest. And as we showed in §7.3.3, this readily implies that the posterior $p(\mu|\vec{x})$ is then also a gamma distribution. Including proper normalization, one gets

$$\begin{aligned} p(\mu|\vec{n}, I) &= p_\gamma(\mu|z + \alpha, m + \beta) \\ &= \frac{(m + \beta)^{z+\alpha}\, \mu^{z+\alpha-1} e^{-(m+\beta)\mu}}{\Gamma(z + \alpha)}. \end{aligned} \tag{7.294}$$

with a mode given by

$$\hat{\mu} = \frac{z + \alpha - 1}{\beta + m}, \tag{7.295}$$

and a variance equal to

$$\sigma_\mu^2 = \frac{z + \alpha}{(\beta + m)^2}. \tag{7.296}$$

Credible ranges $[\mu_{\min}, \mu_{\max}]$, with probability content $C$, are determined by simultaneously solving

$$P_\gamma(\mu_{\max}|z + \alpha, m + \beta) - P_\gamma(\mu_{\min}|z + \alpha, m + \beta) = C \tag{7.297}$$

and

$$p_\gamma(\mu_{\min}|z + \alpha, m + \beta) = p_\gamma(\mu_{\max}|z + \alpha, m + \beta). \tag{7.298}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

$P_\gamma(\mu|\alpha, \beta)$ represents the cumulative density function (CDF) of the gamma distribution with parameters $\alpha$ and $\beta$, given by Eq. (3.113), which may be computed according to

$$P_\gamma(\mu|\alpha, \beta) = \frac{\gamma(\alpha, \mu)}{\Gamma(\alpha)}, \tag{7.299}$$

in which $\Gamma(x)$ and $\gamma(\alpha, x)$ are the gamma function and the (lower) incomplete gamma function. Evaluation of these functions is possible within commonly available math packages such as Mathematica®, MATLAB®, or ROOT.

## Rate Parameter of a Poisson Process with Fixed Background

In practical situations, the observed process rate $r$ may result from a combination of the signal of interest, with rate $r_s$, and some background process, with rate $r_b$, such that $r = r_s + r_b$. In a time interval $T$, one then expects average numbers of signal and background instances equal to $\mu_s = r_s T$ and $\mu_b = r_b T$, respectively.

If the background rate $r_b$ or number of background instances $\mu_b$ are precisely known (and thus fixed), the posterior probability $p(\mu|\vec{n}, I)$ determines the probability distribution of $\mu_s$ uniquely:

$$p(\mu_s|\vec{n}, \mu_b, I) = p(\mu|\vec{n}, I). \tag{7.300}$$

One can then compute $p(\mu_s|\vec{n}, \mu_b, I)$ based on Eq. (7.294) provided one replaces $\mu$ by $\mu_b + \mu_s$:

$$\begin{aligned} p(\mu_s|\vec{n}, \mu_b, I) &= p_\gamma(\mu_s + \mu_b|z + \alpha, m + \beta) \\ &= \frac{(m+\beta)^{z+\alpha}\,(\mu_s + \mu_b)^{z+\alpha-1}\,e^{-(m+\beta)(\mu_s+\mu_b)}}{\Gamma(z+\alpha)} \end{aligned} \tag{7.301}$$

The most probable value of $\mu_s$ is then

$$\hat{\mu}_s = \frac{z + \alpha - 1}{\beta + m} - \mu_b, \tag{7.302}$$

and the credible range for $\hat{\mu}_s$ is obtained by shifting the range of $\hat{\mu}$ calculated earlier by $-\mu_b$.

## Rate Parameter of a Poisson Process with Unknown Background

If the background rate is a priori unknown, it may be possible to turn off the signal in order to estimate the background rate. One should then proceed to carry out $m'$ measurements, with signal turned off, of the number of instances $n_k^{(B)}$ observed during intervals of equal duration $T$. Evidently, the number of background instances shall fluctuate and the background rate thus cannot be determined with absolute precision. Repeating the aforementioned reasoning, one obtains a posterior probability for the number of background instance $\mu_b$ per time interval $T$:

$$\begin{aligned} p\big(\mu_b|\vec{n}^{(B)}, I\big) &= p_\gamma(\mu_b|z_b + \alpha, m' + \beta) \\ &= \frac{(m'+\beta)^{z_b+\alpha}\,\mu_b^{z_b+\alpha-1}e^{-(m'+\beta)\mu_b}}{\Gamma(z_b+\alpha)}, \end{aligned} \tag{7.303}$$

where $z_b = \sum_{k=1}^{m'} n_k^{(B)}$.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Given the preceding posterior for $\mu_b$, one can then return to the evaluation of the posterior probability of $\mu_s$. For a known background rate, we found that the posterior of $\mu_s$ is given by Eq. (7.300). The background rate is not known precisely, however, and one must then account for all possible values of $\mu_b$. To accomplish this, one can think of the samples $n_k$ and $n_k^{(B)}$ as a joint measurement of the background and signal rates, yielding a joint posterior we denote $p(\mu_s, \mu_b|\vec{n}, \vec{n}^{(B)}, I)$. We are not actually interested in the background rate; we thus marginalize this posterior for $\mu_b$ and obtain a posterior for $\mu_s$ by integration over all values of $\mu_b$. But since the measurements of the samples $n_k$ and $n_k^{(B)}$ are in fact disjoint (or independent), the joint posterior $p(\mu_s, \mu_b|\vec{n}, \vec{n}^{(B)}, I)$ is simply the product of the posteriors for $\mu_s$ (at fixed $\mu_b$) and $\mu_b$ (with signal turned off). It then suffices to integrate this product over all physically possible values of $\mu_b$. One gets

$$p(\mu_s|\vec{n}, \vec{n}^{(B)}) = \int_0^\infty p(\mu_s|\vec{n}, \mu_b, I)p(\mu_b|\vec{n}^{(B)}, I)\, d\mu_b, \tag{7.304}$$

which can be readily computed with standard numerical techniques.

### 7.5.3 Bayesian Fitting with Non-Gaussian Processes

Bayesian fitting in the context of non-Gaussian data probability models involving one or several unknown parameters cannot be reduced, in general, to the evaluation and optimization of a $\chi^2$ function but relies on the evaluation of the joint posterior probability of the model parameters. Generally, the evaluation of the posterior proceeds similarly as for cases involving Gaussian noise, and once the mode of the posterior is found, with any of the techniques presented in §7.6, nuisance parameters can be eliminated by marginalization. Overall, the treatment of problems involving non-Gaussian data probability models is thus not very different from the methods already discussed in this chapter. One must acknowledge, however, that the calculation and optimization of non-Gaussian posteriors may be computationally rather intensive and has thus been readily feasible only since the advent of high-performance computers. A number of model-specific methods have been developed to streamline and simplify calculations. Examples of such methods are presented in [97].

## 7.6 Optimization Techniques for Nonlinear Models

Inference methods seeking the optimal values of a model, that is, the values that best describe measured data, require the optimization of an objective function, often called goal or merit function. Depending on the circumstances, the objective function may be the posterior probability of model parameters, a likelihood function, or a $\chi^2$ function. Either way, one seeks an extremum (also called optimum) of these functions in the $n$-dimension space defined by the parameters of the model. Evidently, in the case of probabilities, the extremum must be a maximum whereas for $\chi^2$ functions, it must be a minimum.

We saw in §7.4.3 that optimization problems involving linear models in the presence of Gaussian noise can be reduced to linear equations readily solved, at least in principle, by matrix inversion. Data modeling, however, as we discuss in §7.5, often involves nonlinear

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

equations or non-Gaussian probability models not amenable to linear solutions. Finding the extremum of such objective functions must thus, in general, be achieved with numerical techniques.

There is no general solution to the global optimization encountered in classical and Bayesian inference. However, a plurality of methods have been developed over the years to tackle the problem. The techniques most often used include

1. Hill climbing and gradient methods
   a. Extended Newton methods
   b. Levenberg–Marquardt algorithm
   c. Powell's algorithm
   d. Simplex algorithm
2. Stochastic methods
   a. Simulated annealing
   b. Genetic algorithms
3. Hybrid methods consisting of a combination of hill climbing, gradient based, and stochastic methods

We discuss the principle and applicability of these and related techniques throughout this section beginning in §7.6.1 with hill climbing and gradient methods. Stochastic methods are introduced in §7.6.5. Several other methods and variants of the aforementioned methods exist and are documented in the computing literature but a comprehensive discussion of all these methods is beyond the scope of this book. In-depth discussions and implementation of several optimization algorithms are discussed by Press et al. [156], Besset [36], and Gregory [97]. It should also be clear at the outset that many of the methods discussed in this section find applications in a variety of optimization problems and are not restricted to "fitting" problems, but our discussion will focus on optimization of continuous functions in the context of statistical inference and model fitting.

### 7.6.1 Hill Climbing and Gradient Methods

Consider a function $f(\vec{x})$ defined over a space of $n$-dimensions, $\mathbb{R}^n$. Our goal is to find a global extremum of the function, that is, the position $\vec{x}$ that globally minimizes or maximizes the function as appropriate. The function $f(\vec{x})$ may be rather generic and is not restricted to objective functions measuring the level of fitness of a model to measured data. Many of the techniques discussed in the text that follows indeed apply to a wide range of functions. In the context of inference problems, however, the vector $\vec{x}$ shall here stand for a collection of model parameters that need to be optimized or fitted to measured data (and not the data itself).

Let us assume the function $f(\vec{x})$ is continuous everywhere in the search space and derivable relative to all components of $\vec{x}$. Extrema of the function may then be sought for by searching values of $\vec{x}$ where the gradient of the function vanishes.

$$\nabla f(\vec{x}) = \frac{df(\vec{x})}{d\vec{x}} = 0 \tag{7.305}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

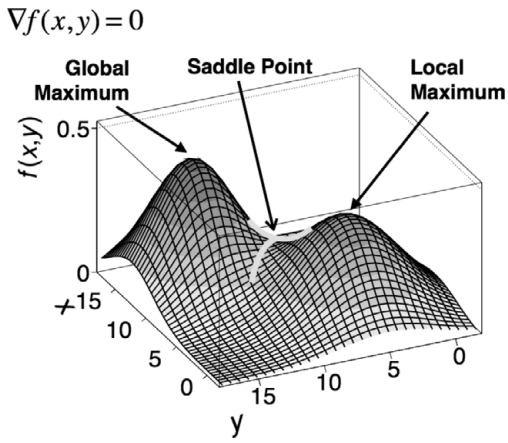


**Fig. 7.6** A vanishing gradient, $\nabla f(\vec{x}) = 0$, is a necessary but insufficient condition in the search for a function's global extremum.

Unfortunately, this condition does not guarantee the existence of an extremum. In $n > 1$ dimensions, points or regions where $\nabla f$ vanish may indeed include saddle points as well as minima and maxima, as illustrated in Figure 7.6. The condition Eq. (7.305) thus provides a necessary but not sufficient condition for the existence of extrema. Solutions $\vec{x}$ of Eq. (7.305) are, however, easily tested to verify whether they correspond to a minimum, a maximum, or a saddle point region by considering second derivatives of the functions along all of its components. Positive and negative second derivatives identify minima and maxima, respectively, while a mix of negative and positive second derivatives (or all null second derivatives) reveals saddle point regions.[10] A more difficult issue arises from the fact that, in general, the function $f(\vec{x})$ may feature several minima or maxima as well as saddle points. Extrema search algorithms may then end up finding local minima or maxima, not the lowest (global minimum) or highest value (global maximum) of the function. In fitting and inference problems, this entails that the solutions obtained may not be truly optimal, that is, they may not represent the model parameters that best fit the data. We will see that optimization algorithms vary greatly in their capacity to identify a global extremum. Hill climbing and gradient based algorithms are typically very fast and efficient but tend, by their very nature, to settle onto the first extremum they find whether global or not. By contrast, Monte Carlo algorithms are typically designed to provide a better and fuller exploration of the search space, albeit at the cost of requiring many function evaluations and thus greater CPU time.

## Newton Optimization Methods

Newton optimization methods implement a generalized version of Newton's zero-finding algorithm to find an extremum of a function $f(\vec{x})$. Starting from an initial position $\vec{x}^{(0)}$,

[10] Note that in one dimension a "saddle point" is observed whenever both the first and second derivative of a function vanish at the same point.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

they use derivatives of the function $f(x)$ to quickly home in on an extremum but provide no guarantee whatsoever that the extremum found is a global minimum or maximum.

Newton's method and related techniques are based on a Taylor expansion of the function $f(\vec{x})$ in the vicinity of a point $\vec{x}^{(k=0)}$ believed to be near the sought for optimum of the function:

$$f(\vec{x}) = f(\vec{x}^{(k)}) + \sum_{i=1}^{n} \left.\frac{\partial f(\vec{x})}{\partial x_i}\right|_{\vec{x}^{(k)}} \left(x_i - x_i^{(k)}\right) + O(2), \tag{7.306}$$

where higher-order terms $O(2)$ are neglected. Inserting this expression in Eq. (7.305), we get

$$\sum_{i=1}^{n} \left.\frac{\partial^2 f(\vec{x})}{\partial x_i \partial x_j}\right|_{\vec{x}^{(k)}} \left(x_i - x_i^{(k)}\right) + \left.\frac{\partial f(\vec{x})}{\partial x_j}\right|_{\vec{x}^{(k)}} = 0, \tag{7.307}$$

which can readily be written as an approximate linear equation of the form

$$\alpha^{(k)} \delta\vec{x}^{(k)} = \vec{b}^{(k)}, \tag{7.308}$$

where

$$\delta\vec{x}^{(k)} = \vec{x} - \vec{x}^{(k)}, \tag{7.309}$$

$$\alpha_{ij}^{(k)} = \left.\frac{\partial^2 f(\vec{x})}{\partial x_i \partial x_j}\right|_{\vec{x}^{(k)}}, \tag{7.310}$$

$$b_i^{(k)} = -\left.\frac{\partial f(\vec{x})}{\partial x_j}\right|_{\vec{x}^{(k)}}. \tag{7.311}$$

Multiplying Eq. (7.308) on both sides by the inverse $(\alpha_{ij}^{(k)})^{-1}$, one gets

$$\delta\vec{x}^{(k+1)} = \left(\alpha_{ij}^{(k)}\right)^{-1} \vec{b}^{(k)}. \tag{7.312}$$

Given an initial estimate of the extremum, $\vec{x}^{(k)}$, also called prior estimate, one obtains an updated and more accurate estimate of the extremum's position, $\vec{x}^{(k+1)}$, by addition of $\delta\vec{x}^{(k+1)}$:

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} + \delta\vec{x}^{(k+1)}. \tag{7.313}$$

We labeled the solution with an index $(k+1)$ to indicate that the procedure can be iterated. Indeed, given an initial value $\vec{x}^{(0)}$, the solution provided by Eq. (7.312) is strictly exact only if higher-order terms in Eq. (7.306) are truly negligible. However, one can use the preceding equations iteratively to obtain an approximation of the sought for extremum. Iterations should be terminated when $|\delta\vec{x}^{(k+1)}| < \varepsilon$ or $|f(\vec{x}^{(k+1)}) - f(\vec{x}^{(k)})| < \xi$, where $\varepsilon$ and $\xi$ define the required precision of the search.

Newton's algorithm is commonly used in optimization engines and packages such as Minuit [115] for extremum searches, but it is important to be aware of its limitations. The algorithm is efficient and precise but will typically home in on the closest extremum because it does not have built-in capabilities to differentiate between local extrema and a global extremum. Indeed, the algorithm can be rather quickly locked in a local minimum. Other techniques must then be used to first identify a region of global extremum and

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

select a seed value $\vec{x}^{(0)}$ with which to initiate the algorithm. It should also be noted that the matrix $\alpha$ may be ill conditioned and not invertible for seed values $\vec{x}^{(0)}$ "far" from an extremum. Yet another limitation of the technique is that it requires the evaluation of first and second derivatives. If such analytical derivatives are not available or possible, numerical techniques must be used, which may reduce the accuracy and the efficiency of the technique. Numerical calculations of second derivatives, in particular, require several calls to evaluate the function and may then considerably increase the CPU time needed to achieve solutions of required precision. It may then be preferable to make use of techniques that do not require calculations of derivatives such as the hill climbing techniques introduced in §7.6.3.

### 7.6.2 Levenberg–Marquardt Algorithm

The Levenberg–Marquardt algorithm is a popular and efficient variant of Newton's algorithm used in $\mathbb{LS}$ minimization problems. The objective function is the $\chi^2$ function defined as

$$\chi^2(\vec{\theta}) = \sum_{i=1}^{N} \left[ \frac{y_i - f(x_i|\vec{\theta})}{\sigma_i} \right]^2 , \tag{7.314}$$

in which $N$ is the number of data points $(x_i, y_i)$, $f(x|\vec{\theta})$ represents the model, and $\vec{\theta}$ are the unknown (free) parameters to be determined by $\chi^2$ minimization. As for Newton's algorithm, one assumes the $\chi^2$-function can be represented by a quadratic form in the vicinity of the optimal values of the parameters $\vec{\theta}_{\min}$:

$$\chi^2(\vec{\theta}) \approx \chi^2_{\min} + \frac{1}{2} \left( \vec{\theta} - \vec{\theta}_{\min} \right)^T \mathbf{D} \left( \vec{\theta} - \vec{\theta}_{\min} \right) , \tag{7.315}$$

where $\chi^2_{\min}$ is the minimum of the function achieved for the optimal value of the parameters $\vec{\theta}_{\min}$; $\mathbf{D}$ is an $m \times m$ symmetric matrix that depends on the data points and the model to be fitted; and $\vec{\theta}$ is an $1 \times m$ column vector one can vary to explore the shape of the $\chi^2$-function.

Let us assume a prior estimate $\vec{\theta}^{(0)}$ of the model parameters is known. If the preceding quadratic form is a reasonable approximation of the actual $\chi^2$-function, then one can use the gradient of the $\chi^2$-function at $\vec{\theta}^{(0)}$ to seek the minimum

$$\left. \frac{\partial \chi^2}{\partial \theta_k} \right|_{\vec{\theta}^{(0)}} = \left. \frac{1}{2} \sum_{ij} \frac{\partial}{\partial \theta_k} \left[ (\theta_i - \theta_{\min,i}) D_{ij} \left( \theta_j - \theta_{\min,j} \right) \right] \right|_{\vec{\theta}^{(0)}} , \tag{7.316}$$

$$= \sum_j D_{kj} \left( \theta_j^{(0)} - \theta_{\min,j} \right) , \tag{7.317}$$

where we have used $\partial \theta_i / \partial \theta_k = \delta_{ik}$, the symmetric nature of $\mathbf{D}$, and carried sums to eliminate the delta functions. It is convenient to write this expression in matrix form:

$$\nabla \chi^2 \big|_{\vec{\theta}^{(0)}} = \mathbf{D} \cdot \left( \vec{\theta}^{(0)} - \vec{\theta}_{\min} \right) . \tag{7.318}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

One solves for $\vec{\theta}_{\min}$ by multiplying both sides of this expression by the inverse $\mathbf{D}^{-1}$:

$$\vec{\theta}_{\min} = \vec{\theta}^{(0)} - \mathbf{D}^{-1} \left. \nabla \chi^2 \right|_{\vec{\theta}^{(0)}} . \tag{7.319}$$

In principle, the values $\vec{\theta}_{\min}$ should correspond to the minimum of the $\chi^2$-function and provide optimal values of the parameters $\vec{\theta}$. In practice, the quadratic form, Eq. (7.315), may be a poor approximation of the actual shape of the $\chi^2$-function. It also relies on numerical calculations of the gradient $\left. \nabla \chi^2 \right|_{\vec{\theta}^{(0)}}$ and the matrix $\mathbf{D}$, which may not be perfectly accurate. One may, however, replace the matrix $\mathbf{D}$ by a constant and achieve what is commonly known as the **steepest descent method**. Given a prior estimate $\theta^{(i-1)}$, a posterior (i.e., an improved estimate) can be evaluated as

$$\theta^{(i)} = \theta^{(i-1)} - k \times \left. \nabla \chi^2 \right|_{\vec{\theta}^{(i-1)}} , \tag{7.320}$$

in which $k$ is a suitably chosen constant.

Computation of Eq. (7.319) requires knowledge of the gradient $\nabla \chi^2$ and the matrix $\mathbf{D}$. The gradient $\nabla \chi^2$ may be calculated according to

$$\frac{\partial \chi^2}{\partial \theta_k} = -2 \sum_{i=1}^{N} \left( \frac{y_i - f(x_i|\vec{\theta})}{\sigma_i^2} \right) \frac{\partial f(x_i|\vec{\theta})}{\partial \theta_k} . \tag{7.321}$$

Calculation of the matrix $\mathbf{D}$, called the **Hessian matrix**[11], involves second-order derivatives of the $\chi^2$-function with respect to $\theta_k$ at any value of $\theta_k$:

$$\frac{\partial \chi^2}{\partial \theta_k \partial \theta_l} = 2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[ \frac{\partial f(x_i|\vec{\theta})}{\partial \theta_l} \frac{\partial f(x_i|\vec{\theta})}{\partial \theta_k} - \left( y_i - f(x_i|\vec{\theta}) \right) \frac{\partial^2 f(x_i|\vec{\theta})}{\partial \theta_l \partial \theta_k} \right] . \tag{7.322}$$

It is convenient to define coefficients $\beta_k$ and $\alpha_{kl}$ as follows:

$$\beta_k \equiv -\frac{1}{2} \frac{\partial \chi^2}{\partial \theta_k} , \tag{7.323}$$

$$\alpha_{kl} \equiv \frac{1}{2} \frac{\partial^2 \chi^2}{\partial \theta_l \partial \theta_k} . \tag{7.324}$$

The coefficients $\alpha_{kl}$ form a matrix $\alpha = \mathbf{D}/2$. Equation (7.319) can thus be recast as

$$\sum_{l=1}^{m} \alpha_{kl} \delta \theta_l^{(i)} = \beta_k . \tag{7.325}$$

Starting with a prior estimate $\theta^{(i-1)}$, solution of the linear equation (7.325) yields an increment $\delta \theta_l^{(i)}$, which one can use to obtain an improved estimate $\theta^{(i)}$:

$$\vec{\theta}^{(i)} = \vec{\theta}^{(i-1)} + \delta \vec{\theta}^{(i)} . \tag{7.326}$$

[11] The Hessian matrix is named after the nineteenth-century German mathematician Ludwig Otto Hesse (1811–1874), who used the determinant of the matrix as a measure of the local curvature of a function of many variables.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

$\alpha$ is commonly called **curvature matrix**, as well as Hessian matrix. Its inverse and Eq. (7.325) provide for the **inverse-Hessian** formula

$$\delta\vec{\theta} = \alpha^{-1}\vec{\beta}. \tag{7.327}$$

As per Eq. (7.322), calculation of $\alpha$ in principle involves both first- and second-order derivatives of the functions $f(x|\vec{\theta})$ with respect to parameters $\vec{\theta}$. Second-order derivatives arise because the gradient, Eq. (7.321), involves first-order derivatives $\partial f/\partial\theta_k$. However, these are often ignored in practice because the term containing second-order derivatives in Eq. (7.322) also includes the coefficients $(y_i - f(x_i|\vec{\theta}))$. These coefficients are expected to be small and oscillate around zero when the fit converges to a minimum $\chi^2$. The sum involving the second-order derivatives thus tends to vanish or at the very least be negligible compared to the first term, which involves a product of first-order derivatives. The calculation of the curvature matrix $\alpha$ is thus reduced to

$$\alpha_{kl} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[\frac{\partial f(x_i|\vec{\theta})}{\partial\theta_l}\frac{\partial f(x_i|\vec{\theta})}{\partial\theta_k}\right]. \tag{7.328}$$

The inverse-Hessian formula usually converges toward the true minimum of the $\chi^2$-function, but the convergence may be slow because of the relatively small steps produced by the method. The calculation of the Hessian and its inverse can also be computationally intensive. Early computers were not as fast and powerful as those in use today. Finding ways to improve the rate of convergence and reduce the computation involved in numerical fits was thus highly desirable. Clearly, the steepest descent method, Eq. (7.320), has the advantages of simplicity and speed. Calculation of the matrix $\alpha$ is not required, and given a suitably chosen constant provides for large increments $\delta\vec{\theta}^{(i)}$ and a fast descent toward the minimum of the $\chi^2$-function. The use of large increments, however, implies the method can easily overshoot the position of the minimum and end up oscillating around it without ever reaching it. It is thus useful to "turn on" the slower but more precise inverse-Hessian method when nearing the minimum. This idea was first suggested in 1944 by K. Levenberg [136] and further developed by D. W. Marquardt in 1963 [140]. It is thus now commonly known as the **Levenberg–Marquardt algorithm** (LMA) or alternatively as the **Damped Least-Squares** (DLS) method. It forms the basis of many fitting packages commonly available today.

In order to use Eq. (7.320), one must identify a proper scale for the constant $k$ that controls the rate of descent. We first note that the $\chi^2$ is by definition nondimensional; in other words, it is a pure number. The fit parameters $\theta_k$, on the other hand, generally have dimensions and units such as cm, kg, and so on. Since the coefficients $\beta_k$ are defined as first-order derivatives of $\chi^2$ with respect to to $\theta_k$, they must carry dimensions of $1/\theta_k$. Equation (7.320) thus implies the constants $k$ have units of $1/\theta_k^2$. The diagonal elements of the Hessian matrix $\alpha$ have the same dimension and as such constitute a natural choice to determine the scale of the constant $k$. The diagonal elements might, however, be too large. Marquadt thus had the insight to include a "fudge" factor $\lambda$:

$$\delta\theta_k = \frac{1}{\lambda\alpha_{kk}}\beta_k. \tag{7.329}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Since the coefficients $\alpha_{kk}$ are by construction positive definite, the increment $\delta\theta_k$ remains proportional to the gradient $\beta_k$ and provides the intended behavior. Marquardt also realized he could combine the preceding steepest descent with the inverse-Hessian formula by introducing modified matrix elements $\alpha'_{ij}$ as follows:

$$\alpha'_{ii} \equiv \alpha_{ii}\,(1+\lambda)\,, \tag{7.330}$$

$$\alpha'_{ij} \equiv \alpha_{ij}, \tag{7.331}$$

and replace $\alpha_{ij}$ by $\alpha'_{ij}$ in Eq. (7.327), thereby yielding

$$\delta\vec{\theta} = \alpha'^{-1}\vec{\beta}. \tag{7.332}$$

When the scale parameter $\lambda$ is very large, the matrix $\alpha'$ is effectively diagonal and so is its inverse. Equation (7.332) thus yields the steepest descent, Eq. (7.320). On the other hand, if $\lambda$ vanishes, one recovers the inverse-Hessian formula, Eq. (7.327). The LMA method is iterative. It should converge relatively rapidly to a region of the parameter space $\vec{\theta}$ that yields near-minimum $\chi^2$ values. However, the search should be stopped if (1) it fails to converge or (2) incremental reduction (decreases) of the $\chi^2$ become negligible (e.g., smaller than 0.01 in absolute value), or some small fractional change, like $10^{-3}$. One should not stop the procedure when the $\chi^2$ increases, unless, of course, the method fails to converge in a reasonable number of iterations.

The LMA can be implemented as follows:

1. Obtain an initial rough guess $\vec{\theta}^{(i=0)}$ of the fit parameters.
2. Compute $\chi^2(\vec{\theta}^{(i=0)})$.
3. Select a small value for $\lambda$, for example, $\lambda = 0.001$.
4. Solve the linear equation (7.332) for $\delta\theta^{(i)}$.
5. Evaluate $\chi^2(\vec{\theta}^{(i)} + \delta\theta^{(i)})$ and calculate the change $\Delta\chi^2 = \chi^2(\vec{\theta}^{(i)} + \delta\theta^{(i)}) - \chi^2(\vec{\theta}^{(i)})$.
6. If $\Delta\chi^2 \geq 0$, increase $\lambda$ by a factor of 10, and repeat step 3.
7. If $|\Delta\chi^2| < \epsilon$, with $\epsilon$ set to be a reasonably small number (e.g., 0.01), stop the search.
8. Otherwise decrease $\lambda$ by a factor of 10, update the parameters: $\vec{\theta}^{(i+1)} = \vec{\theta}^{(i)} + \delta\theta$, and repeat step 3.

Once an acceptable minimum has been reached, one should set $\lambda = 0$ and compute $C = \alpha^{-1}$ in order to obtain an estimate of the covariance matrix of the standard errors of the final parameters $\vec{\theta}^{(i)}$.

While in general the LMA works rather well, it may be slow to converge when the number of fitted parameters is very large. A variety of modern methods have thus been developed to improve on its basic principles, some of which are presented in other sections of this text. Techniques have also been developed to carry out fits with user-defined constraints on the parameter space. This is particularly useful to avoid nonphysical parameter regions or pathological behaviors. The minimization and fitting package **Minuit** [115], popular in the high-energy physics community, uses a combination of Monte Carlo search methods; the simplex method of Nelder and Mead [147], discussed in §7.6.4; and the variable metric

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

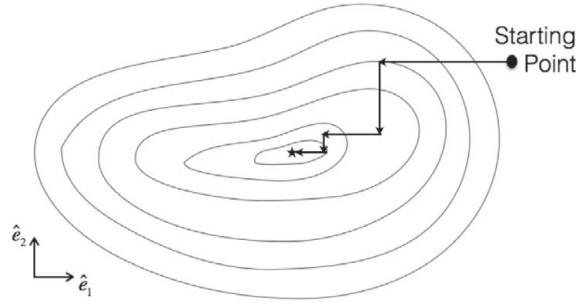


Fig. 7.7 Basic principle of hill climbing methods: Thin solid lines represent iso-contours of the objective function; thick arrows represent paths followed by the hill climbing algorithm along successive directions until a local extremum of the function is reached.

method of Fletcher [84]. It also gives the ability to impose constraints on the parameter space accessible to a search and to control how errors are handled and calculated.

### 7.6.3 Hill Climbing Methods

Hill climbing methods provide convenient objective function optimization techniques that do not require evaluations of function derivatives. They seek a function's extremum in $n$-dimensions by sequentially finding extrema along straight lines, as schematically illustrated in Figure 7.7. Indeed, rather than seeking an extremum simultaneously in several dimensions, the hill climbing carries out the search along straight lines of arbitrary direction in $n$-dimension space. Once a particular extremum is found, the direction is changed, and an improved extremum sought for. The process is repeated iteratively until a true extremum is found within the desired precision. Several distinct techniques may be used for the selection of the direction $\vec{d}^{(k)}$ and the 1D optimization. We will restrict our discussion to a few illustrative techniques in this and the following sections.

The basic idea is to reduce the objective function $f(\vec{x})$ to a single variable function $g(z)$ defined as follows:

$$g(z) = f(\vec{x}^{(k)} + z\vec{d}^{(k)}). \tag{7.333}$$

where the vectors $\vec{x}^{(k)}$ and $\vec{d}^{(k)}$ represent the starting point and direction, respectively, of the $k$th iteration of the search algorithm. A 1D search algorithm is invoked to find the extremum of $g(z)$ at each iteration. The extremum $z$ is then used to calculate the starting point of the next iteration. The algorithm may thus be summarized as follows:

1. Set $k = 0$, and choose a starting point $\vec{x}^{(k)}$ by any appropriate method.
2. Select a search direction $\vec{d}^{(k)}$ in $n$-dimensions.
3. Find an extremum $z^{(k)}$ of $g(z)$ using a suitable 1D search technique.
4. Calculate the corresponding point in full space according to $\vec{x}^{(k+1)} = \vec{x}^{(k)} + z^{(k)}\vec{d}^{(k)}$.
5. If convergence is achieved, terminate the search and produce $\vec{x}^{(k+1)}$ as the sought for extremum.
6. Otherwise, set $\vec{x}^{(k+1)}$ as a new starting point, set $k = k + 1$, and proceed to step 2.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

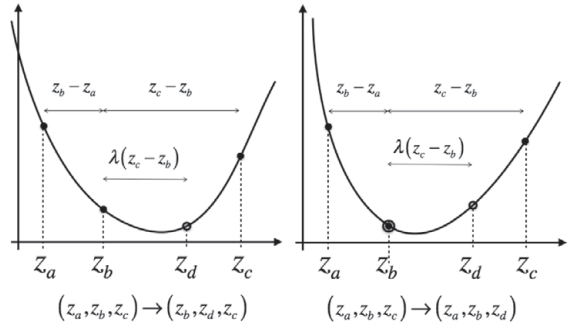


**Fig. 7.8** Illustration of the bisection algorithm. The initial triplet $(z_a, z_b, z_c)$ is replaced by $(z_b, z_d, z_c)$ or $(z_a, z_b, z_d)$ depending on whether $g(z_d)$ is better or worse than $g(z_b)$.

## Bisection and Bracketing Algorithms

The principle of the bisection algorithm is illustrated in Figure 7.8. One seeks an extremum of a continuous function $g(z)$ assuming its derivatives are unknown or not readily calculable. As a starting point, the algorithm requires a triplet $(z_a, z_b, z_c)$ such that $z_a < z_b < z_c$. Let us assume $g(z_b)$ is "better" than $g(z_a)$ and $g(z_c)$. In this context, better means a higher value if searching for a maximum, or a lower value if searching for a minimum. Given the function is continuous, the fact that $g(z_b)$ is better than $g(z_a)$ and $g(z_c)$ implies the sought for extremum is necessarily within the range $z_a \leq z \leq z_c$. Let us thus seek a new triplet that narrows this interval. On general grounds, one can assume there is a larger probability to find the extremum in the larger of the two intervals $[z_a, z_b]$ and $[z_b, z_c]$. For illustrative purposes, let us assume $z_b - z_a < z_c - z_b$ and select a point $z_d = z_b + \lambda(z_c - z_b)$ where $\lambda$ is typically chosen to be the golden ratio $(3 - \sqrt{5})/2$. If $g(z_d)$ is better than $g(z_b)$, the new triplet is identified as $(z_b, z_d, z_c)$; otherwise it is set to $(z_a, z_b, z_d)$. The procedure is then iterated with the new triplet. The search proceeds iteratively and identifies successively narrower intervals. It is terminated when $\Delta z = z_c - z_a \leq \varepsilon$, that is, when the width of the triplet is narrower than the required precision $\varepsilon$. By construction, the sought for extremum is somewhere in the interval $[z_a, z_c]$, the solution $z_b$ achieved thus has a precision of order $\pm(z_c - z_a)/2$.

In the preceding discussion, we assumed the starting triplet $(z_a, z_b, z_c)$ featured a point $z_b$ better than both $z_a$ and $z_c$. If the three points of the triplet are chosen randomly, there is a good chance that $z_a$ or $z_c$ might be better than $z_b$. There is thus no guarantee that the interval $[z_a, z_c]$ brackets the extremum; that is, the extremum is not necessarily within the interval. One must then seek a new a new triplet $(z_a, z_b, z_c)$, with $g(z_b) < g(z_a), g(z_c)$, that brackets the extremum.

Bracketing of the extremum may be accomplished with a rather simple algorithm as follows. Consider two given points $z_b$ and $z_c$. Let us assume that $g(z_c)$ is better than $g(z_b)$ as illustrated in Figure 7.9. Let $\Delta z = z_c - z_b$. We then choose a point $z_e$ such that $z_e = z_c + 2\Delta z = 3z_c - 2z_b$ in the direction of the optimum. If $g(z_c)$ is better than $g(z_e)$, then the extremum is bracketed in the interval $[z_b, z_e]$, and one can proceed with the bisection algorithm discussed earlier. Otherwise, iterate the extension, that is, let $z_b^{(k+1)} = z_c^{(k)}$,

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

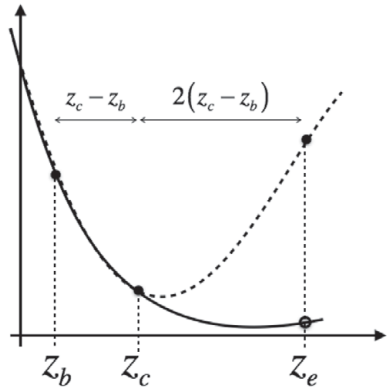


Fig. 7.9 Extension algorithm used to bracket an extremum.

$z_c^{(k+1)} = z_e^{(k)}$, and $z_e^{(k+1)} = 3z_c^{(k+1)} - 2z_b^{(k+1)}$. The width of the interval increases by a factor of 2 at each iteration and is thus guaranteed to eventually contain the extremum, unless an extremum does not exist for finite $z$ values. Implementations of this algorithm should thus check for floating point exceptions.

Insofar as the function $g(z)$ is continuous, the bisection algorithm cannot fail and will always return a valid approximation of the extremum. However, its convergence rate is rather slow, and obtaining a solution of high precision may require a large number of iterations. The required precision $\varepsilon$ should thus be set to a practical value reflecting the needs of the application. If extremely high precision is required, one can complete the search with the faster and more precise Newton (gradient) method, which is essentially guaranteed to work well in the immediate vicinity of the extremum.

## Powell's Algorithm

Powell's conjugate direction method, commonly known as Powell's algorithm [154], provides a straightforward implementation of the hill climbing method toward the search for a function's extremum. The algorithm relies on the notion that once an optimum has been obtained along a particular direction $\vec{d}_k$, the likelihood for greatest improvement lies in a new direction $\vec{d}_{(k+1)}$ perpendicular to the original direction $\vec{d}_k$, that is, such that $\vec{d}_{(k+1)} \cdot \vec{d}_k = 0$. However, since this condition does not uniquely define $\vec{d}_{(k+1)}$, the algorithm also includes a specific recipe to decide on the next best direction at each step. The algorithm may be summarized as follows:

1. Select the required precision of the search, $\varepsilon$.
2. Let $k = 0$.
3. Given a best point $\vec{x}^{(k)}$ at iteration $k$, initialize $n$ unit vectors $\hat{e}_i$ to form a complete basis of the search space. The initial set of vectors may be defined as $\hat{e}_1 = (1, 0, \ldots, 0)$, $\hat{e}_2 = (0, 1, \ldots, 0)$, …, $\hat{e}_n = (0, \ldots, 0, n)$.
4. Initiate the search with $k = 1$.
5. Determine the optimum $\vec{x}^{(k)}$ of the function $f(\vec{x})$ along the direction $\hat{e}_k$ starting from $\vec{x}^{(k-1)}$.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

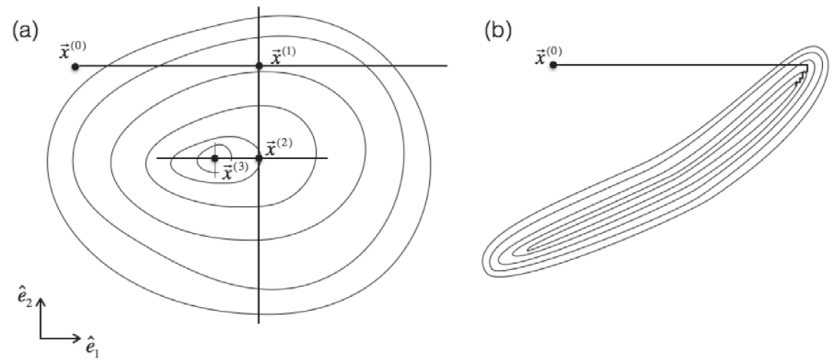


**Fig. 7.10** Illustration of Powell's algorithm for (a) a function nearly quadratic in the vicinity of its extremum and (b) a function exhibiting a narrow and elongated extremum structure.

6. Increment $k$ by one unit and if $k \leq n$, proceed to step 5.
7. Otherwise, set $\hat{e}_k = \hat{e}_{k-1}$.
8. Set $\hat{e}_n = (\vec{x}^{(n)} - \vec{x}^{(0)})/|\vec{x}^{(n)} - \vec{x}^{(0)}|$.
9. Find a new extremum $\vec{x}^{(n+1)}$ of $f(\vec{x})$ along $\hat{e}_n$.
10. If $|\vec{x}^{(n)} - \vec{x}^{(0)}| < \varepsilon$, terminate the algorithm.
11. Otherwise, set $\vec{x}^{(0)} = \vec{x}^{(n+1)}$, and proceed to step 3.

Powell's algorithm is robust for functions exhibiting a (nearly) quadratic behavior near their extremum, as illustrated in Figure 7.10a, but may be extremely slow to converge (or in some cases not converge at all) if the extremum is located within a very narrow and elongated hill (or valley), as schematically shown in Figure 7.10b.

### 7.6.4 Simplex Algorithm

The simplex algorithm, invented by Nelder and Mead [147], uses the notion of **simplex** to subdivide a large parameter space and carry out a search for the extremum of a function. Defined in $n$-dimension space, an $n$-simplex is a polytope figure formed by joining $n + 1$ vertices (or summits) by straight lines. The $n + 1$ vertices $\vec{x}_k$ of an $n$-simplex must be affinely independent, that is, the differences $\vec{x}_1 - \vec{x}_0, \vec{x}_2 - \vec{x}_0, \ldots, \vec{x}_n - \vec{x}_0$ must be linearly independent. Formally, a point may be considered a 1-simplex and a line segment a 2-simplex. But the notion of simplex applied to searches for function extremum is of interest mostly for $n \geq 2$. A 3-simplex is a regular triangle in two dimensions, a 4-simplex is a tetrahedron in three dimensions, and so on, as illustrated in Figure 7.11.

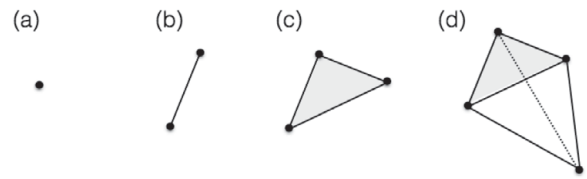


**Fig. 7.11** Illustration of basic simplex of order $n = 1$–4: (a) single vertex, (b) line segment, (c) triangle, (d) tetrahedron.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

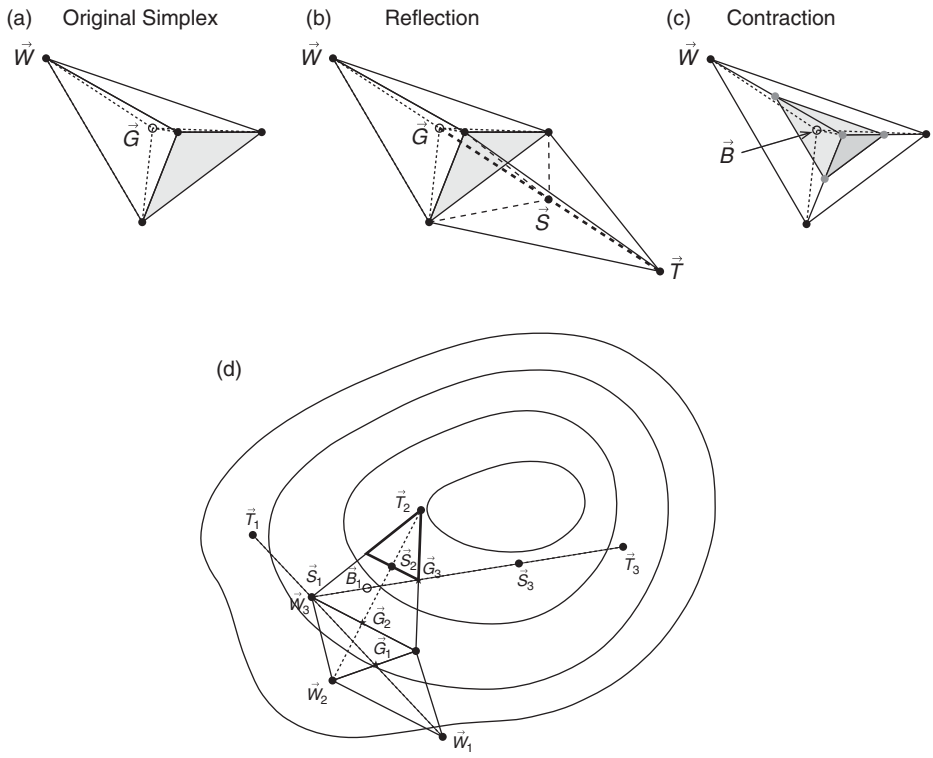


**Fig. 7.12** Illustration of the simplex algorithm: (a) original simplex, (b) simplex obtained by reflection of the weakest vertex; (c) simplex contraction relative to the best vertex $\vec{B}$; (d) example of the evolution of a 3-simplex involving two reflections and one contraction.

The bisection technique presented in §7.6.3 has guaranteed success in one dimension but is not applicable in $n \geq 2$ dimensions. It is indeed not possible to robustly bracket the extremum of a function in two or more dimensions. Segmentation of the search space nonetheless remains possible with the notion of simplex. Given the edges and faces of an $n$-simplex cut across all dimensions of the search space, one may naturally consider how a function of interest behaves across these faces and seek regions where the function grows (search for a maximum) or decreases (search for a minimum).

The algorithm is relatively simple and involves no derivatives. The search for an extremum is based solely on evaluations of the function at the vertices of a simplex while its shape, orientation, and size are varied. The value of the function at a vertex is said to be better if larger (smaller) than those at other vertices in searches for a maximum (minimum). Although bracketing of an extremum is not possible in $n \geq 2$, the shape of the simplex can be varied to stretch or shrink in directions where the function is better. It is thus possible to progressively converge toward an extremum of the function. The simplex algorithm, schematically illustrated in Figure 7.12, may be decomposed as follows:

1. Select a simplex consisting of $n + 1$ vertices $\vec{x}_i$.
   a. If no prior guess of the function's extremum exists, the vertices of the simplex may be selected randomly in the search space of interest. For instance, if the search

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

consists of an $n$-dimensions box $B$ (i.e., a hyperrectangle), the vertices may be generated according to

$$\vec{x}_i = r \times (x_{\max,i} - x_{\min,i})\,\hat{e}_i, \tag{7.334}$$

where $r$ is a random number in the range $0 \leq r \leq 1$; $x_{\max,i}$ and $x_{\min,i}$ represent the upper and lower boundaries of the box, respectively, along dimension $i$; and $\hat{e}_i$ are unit vectors along each of the dimensions $i$.

b. Alternatively, given an initial guess $\vec{x}^{(0)}$, one can use $\vec{x}_0 \equiv \vec{x}^{(0)}$ as the first vertex of the simplex and generate $n$ additional vertices according to

$$\vec{x}_i = \vec{x}_0 + k\hat{e}_i, \tag{7.335}$$

where $k$ is constant representative of the scale of the search, and $\hat{e}_i$ are once again unit vectors along each of the dimensions $i$. If the search space is of considerably different sizes along each of the dimensions, the constant $k$ may be replaced by scale factors for each of the coordinates.

2. Evaluate the objective function at each of the vertices.
3. Tag the worst vertex as $\vec{W}$.
4. If the size of the simplex is smaller than the desired precision $\varepsilon$, terminate the algorithm.
5. Determine the center of gravity, $\vec{G}$, of the simplex according to

$$\vec{G} = \frac{1}{n}\sum_{i\neq W}^{n+1} \vec{x}_i, \tag{7.336}$$

where the worst vertex is excluded.

6. Determine a point $\vec{S}$ located symmetrically to $\vec{W}$ relative to $\vec{G}$,

$$\vec{S} = \vec{G} - (\vec{W} - \vec{G}) = 2\vec{G} - \vec{A}. \tag{7.337}$$

7. Evaluate the function at $\vec{S}$, and skip to step 11 if $f(\vec{S})$ is not better than all other points of the simplex.
8. Otherwise, calculate a new point $\vec{T}$ that is twice as far from $\vec{G}$ than $\vec{S}$

$$\vec{T} = \vec{G} - 2(\vec{S} - \vec{G}) = 3\vec{G} - 2\vec{A}. \tag{7.338}$$

9. Evaluate the function at $\vec{T}$, and if $f(\vec{T})$ is better than $f(\vec{S})$, update the current simplex by replacing $\vec{W}$ by $\vec{T}$ and proceed to step 4.
10. Otherwise, replace $\vec{W}$ by $\vec{S}$ and proceed to step 4.
11. Calculate the point $\vec{B} = (\vec{G} + \vec{W})/2$.
12. If $f(\vec{B})$ is better than all vertices, update the simplex by replacing $\vec{W}$ by $\vec{T}$, and proceed to step 4.
13. Otherwise, contract the simplex by a factor of 2 relative to the best vertex, that is, reduce in half all edges leading to the best vertex, and proceed to step 4.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Step 8 expands the simplex in the direction of the best point and guarantees that the next step will be in a different direction, thereby enabling a complete exploration of the surrounding space. Step 13 shrinks the simplex and thus guarantees an eventual convergence onto an optimum.

The simplex algorithm is reasonably efficient and is guaranteed to converge to an extremum. It is particularly well suited for searches in multidimensional space of large dimensionality, that is, large values of $n$. Unfortunately, the algorithm may easily get trapped in a local extremum and thus fail to find the sought for global extremum. The algorithm is also found to progress somewhat slowly in the immediate vicinity of an extremum. The simplex algorithm should thus be used with a relatively low precision $\varepsilon$, and once an approximate extremum has been identified, one should switch to more robust and efficient algorithms that work best in the immediate vicinity of an extremum (e.g., Newton algorithm, Levenberg–Marquardt, etc.). The algorithm may also be used in conjunction with the simulated annealing method to reduce the risk of terminating the algorithm at a local extremum.

### 7.6.5 Random Search Methods

Objective functions involved in modeling of data with nonlinear models often feature a large number of extrema. An extremum search conducted with a gradient or hill climbing type algorithm and initiated totally at random thus runs the risk of homing-in on the "closest" local extremum, thereby missing the true global optimum and the best model parameter set. A technique capable of scanning the entire parameter space that does not get stuck on a local extremum and is capable of finding the best extremum (i.e., the global extremum) of the function is thus needed. Such a global extremum finding method should be particularly useful for modeling problems where the number of parameters is very large or whenever a sensible initialization of the model parameters, either algorithmically or by a user, is not feasible. We discuss two illustrative examples of random model space scanning: simulated annealing and genetic algorithms.

### 7.6.6 Simulated Annealing Methods

The numerical method of simulated annealing is inspired by the annealing technique used in metallurgy to control the size of crystals and minimize the number and size of defects in materials produced. Studies in metallurgy have shown that the size of crystals and defects are largely determined by the thermodynamic free energy and the rate of cooling used in the production of materials: slow cooling enables uniform temperature decrease and the atoms of the material are then less likely to settle in local configurations that produce "local" minima of the free energy. The numerical method of simulated annealing parallels the slow cooling process by introducing a modified objective function that depends on a temperature parameter $T$. The technique involves a random exploration of the parameter space in which a migration to better parameters (in the sense of producing a higher likelihood or minimum $\chi^2$) is always accepted, but such that occasional jumps to worse local solutions are also possible. It is those random jumps away from a local minimum that

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

allow, eventually, finding stronger (better) regions of the objective function and, ultimately, the identification of a true global extremum.

Initial developments of the simulated annealing technique are commonly attributed to S. Kirkpatrick et al. [126] and V. Cerny [62], who independently adapted the Metropolis–Hastings algorithm [143] used in Monte Carlo simulations of thermodynamic systems. Several variants of the technique have been developed that find a wide range of applications in physical sciences, life sciences, and manufacturing. A detailed discussion of these variants, their range of applicability, strengths, and weaknesses is beyond the scope of this text and can be found in the computing literature. Our discussion is thus limited to a brief conceptual introduction of the principle and merits of the method.

The basic principle of the simulated annealing technique is to replace the posterior $p(\vec{\theta}|D, I)$ by a modified posterior $p_T(\vec{\theta}|D, I)$ defined as follows:

$$p_T(\vec{\theta}|D, I) = \exp\left(\frac{\ln p(\vec{\theta}|D, I)}{T}\right), \tag{7.339}$$

where $T$ plays the role of a temperature parameter. For $T = 1$, the function $p_T(\vec{\theta}|D, I)$ is obviously strictly equivalent to the posterior $p(\vec{\theta}|D, I)$. However, for increasingly larger values of the temperature, the ratio of the logarithm of $p(\vec{\theta}|D, I)$ to $T$ produces progressively flatter and flatter profiles of probability. Random motions in parameter space thus do not result in large changes in probability. It is then possible to randomly explore the entire parameter space at low cost and run a lower risk of being stuck in a local minimum.

The algorithm achieves an extensive exploration of the space by using random jumps from its "current" position:

$$\vec{\theta}^{(k+1)} = \vec{\theta}^{(k)} + \Delta\vec{\theta}, \tag{7.340}$$

where the size of the jump, $\Delta\vec{\theta}$, is determined randomly for each jump. One then computes the difference of the probabilities at $\vec{\theta}^{(k+1)}$ and $\vec{\theta}^{(k)}$

$$\Delta p = p_T(\vec{\theta}^{(k+1)}|D, I) - p_T(\vec{\theta}^{(k)}|D, I). \tag{7.341}$$

A new value $\vec{\theta}^{(k+1)}$ is considered more advantageous (better), and thus readily accepted as a current estimate of the solution, if $\Delta p > 0$. If it is not better, the jump may nonetheless be randomly accepted with an acceptance rate $\alpha$ determined by the temperature $T$. The search begins at high temperature and yields a high acceptance rate for bad jumps. This enables a wide (though not systematic) exploration of the parameter space. It is then likely that the algorithm can stumble onto better regions of the parameter space. The temperature is slowly and steadily decreased toward unity as the search proceeds, and with it, the acceptance rate of bad jumps. As the rate of bad jumps decreases, the search becomes more focused on the local region surrounding the current value of the parameter. The hope is then that the random exploration of the space has enabled the identification of the region of the true extremum of $p_T(\vec{\theta}|D, I)$, so that, as the temperature finally converges to unity, the search yields a true global extremum of $p(\vec{\theta}|D, I)$.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

The simulated annealing technique involves several parameters and features of its own, including

1. A model for the rate at which $T$ is reduced relative to the number of iterations $k$ completed
2. A model for the bad jumps acceptance rate $\alpha$
3. A model for the (maximum) size of random jumps $\Delta\vec{\theta}$

Various annealing models are discussed in the computing literature, which are shown to have varying degrees of success and performance. See [156] and references therein for a more in-depth discussion of these models.

An important drawback of the basic annealing algorithm is that random jumps can be quite ineffective at finding a better region of the objective function. Indeed, a large fraction of jumps carried out may be nonadvantageous and the CPU time spent on useless parameter values (or regions) may be quite large. The problem may be particularly acute if the objective function has narrow valleys or peaks. A very large fraction of the jumps may then end up in the flat and nondiscriminating landscape of the function, thus rendering the algorithm rather ineffective. Fortunately, the concept of simulated annealing does not require the jumps to be totally random and one may in fact couple the annealing technique to more structured search algorithms. Press et al. [156] present such an algorithm in which the search parameter $\vec{\theta}^{(k)}$ is replaced by a simplex. The search for an extremum is then largely driven by the structure of the objective function. However, in order to avoid being stuck on local extrema, the annealed simplex of Press et al. accepts disadvantageous changes of the simplex with an acceptance rate that depends on the temperature. It is thus in principle possible to avoid getting stuck in a local extremum of the objective function, while, relatively speaking, limiting the number of useless function calls and steps across the search space. Other variants of the annealing algorithm have also been considered.

### 7.6.7 Genetic Algorithms

Genetic algorithms (GAs) use a search heuristic that simulates the process of natural selection toward the solution of optimization and search problems. They belong to a larger class of algorithms known as evolutionary algorithms used most particularly in artificial intelligence problems. Many GA variants have been developed and reported in the computing literature. In this section, we focus on the application of GAs in problems of model parameter optimization. As for simulated annealing algorithms, our discussion is meant to be a conceptual introduction: readers interested in implementing and using GAs in practical applications should consult specialized works on this topic [36, 107, 144, 145].

The concept of the GA was introduced in 1975 by John Holland [107] as an alternative optimization technique capable of avoiding getting stuck on local extrema. The technique is called genetic because it mimics the evolutionary process of all known living species. In the context of GAs, parameter values within a search space are considered as chromosomes of individuals. The objective function is then regarded as a measure of

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

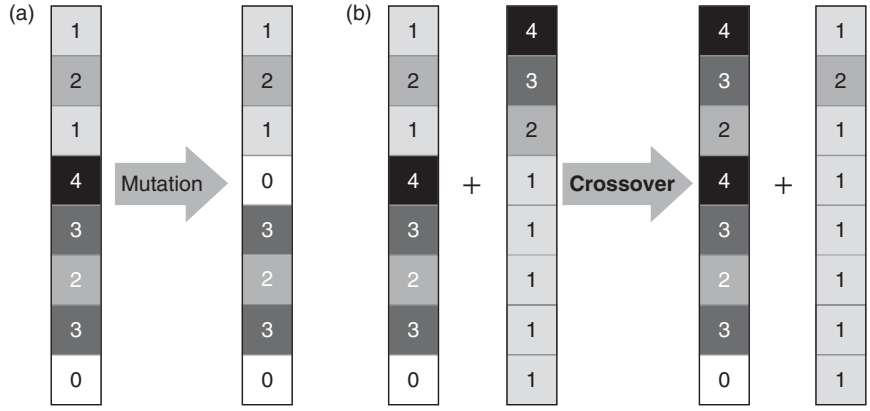


**Fig. 7.13** Schematic illustration of the (a) gene mutation and (b) crossover processes involved in genetic algorithms used toward objective function optimization.

the fitness of individuals toward adaptation to their environment. The GA proceeds by successive iterations. At each iteration, gene mutations and cross-overs are produced and examined. The fittest individuals (i.e., in this context, chromosomes with best objective function values) are selected. They "reproduce" and thus make it into the next iteration. A mutation involves a change of one chromosome at reproduction time whereas a crossover occurs when two chromosomes randomly split and recombine with components (genes) of each other, as schematically illustrated in Figure 7.13. The choice of which individuals (i.e., which chromosomes) survive and are selected for reproduction is carried out randomly. This enables a full exploration of the parameter space and local extrema are thus avoided.

In the context of optimization problems, GA is typically implemented by considering individual elements $\theta_i$ of a parameter vector $\vec{\theta} = (\theta_1, \ldots, \theta_m)$ as individual genes. The reproduction of an individual involves a perfect copy of all elements of the vector but the selection of which individuals are reproduced and/or genetically modified is a random process. The notion of survival of the fittest is implemented by assigning a selection probability larger for more fit individuals, that is, those with better values of the objective function. Mutations create individuals with genes that span the entire parameter space. Evidently, most mutants are found to have poor objective function value and discarded; that is, they do not make it to the next generation (iteration). However, occasionally, mutations reveal more fit individuals, and those are more likely to make it into the next generation. Similarly, crossovers mix good genes with the intent of producing more fit individuals. They complement mutations toward a full exploration of the parameter space. Their production and selection involves a stochastic process as well. Many crossovers lead to unfit individuals but few produce better adapted individuals that survive into the following generation. Successive iterations eventually yield parameter values that constitute a global extremum of the objective function.

An elegant implementation of the genetic algorithm for parameter estimation in SmallTalk and Java is presented by D. Besset [36].

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

# 7.7 Model Comparison and Entity Classification

## 7.7.1 Problem Definition

Model selection and entity classification problems are encountered in all scientific disciplines. And, although they have rather different practical goals, that of determining which of several models is best and identifying the type or class of a specific observed entity, they are mathematically equivalent. The two types of problems can thus be treated with the same tools and inference framework. We will here restrict our discussion to the Bayesian approach, having already discussed in §6.4 the notion of which test can be used, within the frequentist paradigm, to evaluate the goodness of models and classify entities. One may argue, however, that the Bayesian paradigm provides tools and methods far more intuitive and better adapted to either tasks than the frequentist approach.

As an example of model selection drawn from nuclear physics, consider a measurement of the momenta, $\vec{p}$, of particles produced by some nuclear interaction (e.g., proton on proton at 7 TeV). The goal shall be to determine and use the shape of the momentum spectrum in order to compare and hopefully falsify[12] competing models of the particle production dynamics. Let $D = \{p_1, p_2, \ldots, p_n\}$, where $n$ is a very large integer corresponding to the number of observed particles, represent the set of measured momenta. The inference task shall then be to determine which of $m > 1$ model hypotheses $H_k$, best represent the data $D$. For example, one could consider whether the data are best represented by an exponential distribution, a Maxwell–Boltzmann distribution, a Blast–Wave model [170], a power law, and so on. Evidently, the models $H_k$ may have distinct parameters $\vec{\theta}^{(k)}$. The model selection problem shall then consist in finding which of the $m$ models $H_k$ features the largest posterior probability $p(H_k|D, I)$ irrespective of these parameters.

Data classification problems involve a rather similar logic and process. For instance, in nuclear collision studies, one might be interested in identifying the species of measured particles such as pions, kaons, protons, and so on, based on their energy loss in a detector (e.g., a Time Projection Chamber, TPC), or one might wish to use shower shapes observed in a calorimeter system to distinguish photons and electrons from hadrons, and so on. One then formulates $m > 1$ distinct hypotheses $H_k$ corresponding to the several classes, categories, or types the measured entities (e.g., particles or calorimeter showers) might belong to. The probabilities $p(H_k|D, I)$ a given entity might belong to each of the types are then evaluated on the basis of prior probabilities and likelihood functions of each of the models. The hypothesis $H_k$ with the largest posterior probability $p(H_k|D, I)$ may then be selected as the most likely type of the measure entity based on the data $D$.

## 7.7.2 Comparing the Odds of Two Hypotheses

Let us consider a specific example of the model selection problem within the Bayesian paradigm. Let us assume that we have carried out a measurement and obtained a dataset

[12] Demonstrate as false or invalid.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

$D$. Our task shall then be to assess which of several models (or model hypotheses) $H_k$ best match or fit the data by comparing the odds, that is, the probabilities, of the models.

The comparison of the odds of the two or more models is readily formulated in terms of ratios of their posterior probabilities. For instance, in order to compare two models $H_0$ and $H_1$, with model parameters $\vec{\theta}^{(H_0)}$ and $\vec{\theta}^{(H_1)}$, respectively, we define the odds ratio of model $H_1$ in favor of model $H_0$ as

$$O_{10} = \frac{p(H_1|D,I)}{p(H_0|D,I)}, \tag{7.342}$$

where $p(H_0|D)$ and $p(H_1|D)$ represent the posterior probabilities of the models $H_0$ and $H_1$, respectively, regardless of the specific parameter values $\vec{\theta}^{(H_j)} = (\theta_1^{(H_j)}, \ldots, \theta_{m_j}^{(H_j)})$ that best fit the data. The posteriors $p(H_j|D,I)$, $j = 1, 2$, are obtained by marginalization of all parameters $\vec{\theta}^{(H_j)}$ of the models according to

$$p(H_j|D,I) = \int_{\Omega_{\vec{\theta}}} d\theta_1^{(H_j)} \cdots d\theta_{m_j}^{(H_j)} p\big(\theta_1^{(H_j)}, \ldots, \theta_{m_j}^{(H_j)}|D,I\big), \tag{7.343}$$

where the integration is carried over all model parameters $\vec{\theta}^{(H_j)}$ and across the entire domain $\Omega_{\vec{\theta}}$ of these parameters. Typically, the hypothesis $H_0$ shall be considered as the null hypothesis representing a commonly adopted model. A large value of the odds ratio, $O_{10} \gg 1$, shall indicate that the alternative hypothesis $H_1$ is a far more suitable representation (explanation) of the data than the null hypothesis $H_0$, while a value of order unity or smaller would indicate $H_1$ is not particularly favored by the data, and thus should not be adopted in favor of the null hypothesis.

The comparison of the two hypotheses is thus, in principle, rather straightforward as it suffices to take the ratio of their respective probabilities. One may in fact forgo the normalization of the posteriors by the global likelihood of the data, $p(D|I)$, and write

$$O_{10} = \frac{p(H_1|I)p(D|H_1,I)/p(D|I)}{p(H_0|I)p(D|H_0,I)/p(D|I)} = \frac{p(H_1|I)p(D|H_1,I)}{p(H_0|I)p(D|H_0,I)}, \tag{7.344}$$

which involves ratios of priors and likelihoods of the data according to the two hypotheses. Introducing the priors odds ratio

$$O_{10}^{\text{prior}} = \frac{p(H_1|I)}{p(H_0|I)}, \tag{7.345}$$

and the **Bayes factor**

$$B_{10} = \frac{p(D|H_1,I)}{p(D|H_0,I)} \tag{7.346}$$

as the ratio of the likelihoods of the data based on the two hypotheses, one can determine the posterior odds ratio in terms of a product of the priors odds ratio and the Bayes factor

$$O_{10} = O_{10}^{\text{prior}} B_{10}. \tag{7.347}$$

Examples of computation of Bayes factors and odds ratio are discussed later in this section.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

### 7.7.3 Comparing the Odds of Hypotheses with a Different Number of Parameters

A commonly encountered class of problems involves the fitting of data and comparison of models with different numbers of free parameters. For such comparisons, it is often possible (although not essential for the following discussion) to represent measured data with generic linear models of the form

$$f(x|\vec{a}) = \sum_{k=1}^{m} a_k f_k(x), \tag{7.348}$$

where the coefficients $a_k$ are free model parameters and the $f_k(x)$ represent a set of linearly independent functions. Examples of applications of such linear models include fits of polynomials, orthogonal polynomials, and Fourier decompositions. By construction, the inclusion of several functions $f_k(x)$ in a model may enable representation of a wide range of functional dependencies on $x$, and thus provide the capacity to obtain arbitrarily good fits of the data. Indeed, unless there is prior knowledge dictating how many terms $a_k f_k(x)$ should be included in the sum, one could, in principle, add arbitrarily many such terms and obtain arbitrarily good fits of measured data. However, given finite measurement errors, it may be unclear, a priori, what the true functional shape of the data should be, and what number $m$ of functions $f_k(x)$ should actually be included in the parameterization of the data. One thus wishes for a mathematical procedure capable of enabling an objective decision as to whether the addition of one or several model parameters might be justified.

The problem arises, obviously, that by arbitrarily increasing the number of parameters, one might be able to fit the data exactly. Indeed, a polynomial of order $n$ shall perfectly fit a dataset consisting of $n + 1$ points, but such a perfect fit should not be construed as meaningful or representative of the data. Evidently, the presence of (random) measurement errors can give the illusion of high-frequency components in the measured data. Such high-frequency components should not be included, however, unless they are motivated by a physical understanding of the observed phenomenon. The decision-making mathematical procedure we seek should thus have a built-in mechanism to disfavor complicated hypotheses with too many parameters. The goal of this section is to show that odds ratios, equipped with the notion of Occam factors introduced in the text that follows, do in fact provide a more or less objective procedure to compare the merits of models of the form, Eq. (7.348), that penalizes overly complicated models.

For simplicity's sake, let us begin our discussion with the comparison of a null hypothesis (a model) $H_0$ determined by a single and fixed parameter, $\theta = \theta_0$, and an alternative hypothesis $H_1$ functionally identical to $H_0$, but in which the parameter $\theta$ is allowed to vary within some specific domain $\theta_{\min} \leq \theta \leq \theta_{\max}$. We will see, later in this section, how Bayesian model comparison may be trivially extended to any finite number of parameters and any finite number of models. By virtue of the fact that $H_1$ has an adjustable parameter $\theta$, one might naively expect that it should readily provide a better fit of the data and thus have a larger probability. However, we shall next calculate the odds ratio $O_{10}$ and show that the posterior probability of $H_1$ is in fact suppressed, relative to $H_0$, by the added parameter space.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

Let us assume that the models $H_0$ and $H_1$ span the entire space of models so one can write

$$p(H_0|D, I) + p(H_1|D, I) = 1. \tag{7.349}$$

Dividing this expression by $p(H_0|D, I)$, and rearranging the terms, one gets

$$p(H_1|D, I) = 1 - p(H_0|D, I) = \frac{1}{1 + (O_{10})^{-1}}, \tag{7.350}$$

which tells us that the probability of the alternative hypothesis is determined entirely by the posterior odds ratio $O_{10}$. If $O_{10}$ is very large, $p(H_1|D, I) \to 1$, whereas if $O_{10}$ is very small, one gets $p(H_1|D, I) \approx 0$. In the specific context of the comparison of two models of similar functional dependence, but different numbers of free parameters, one may assume there are no a priori reasons to favor one model or the other, and choose $O_{10}^{\text{prior}} = 1$. Equation (7.350) then reduces to

$$p(H_1|D, I) = \frac{1}{1 + (B_{10})^{-1}}. \tag{7.351}$$

Let us thus focus our attention on the Bayes factor $B_{10}$. This requires marginalization of the posterior probability, $p(\theta|D, H_1, I)$, of the parameter $\theta$. In turn, it also necessitates the choice of a prior for $\theta$ and calculation of the likelihood of hypothesis $H_1$. Let us first consider the prior of $\theta$.

Assuming the range of $\theta$ may be sensibly bound to $\theta_{\min} \leq \theta \leq \theta_{\max}$, and given the prior lack of information about this parameter, it is reasonable to choose a flat prior for $\theta$ and write

$$p(\theta|H_1, I) = \begin{cases} \frac{1}{\Delta\theta} & \text{for } \theta_{\min} \leqslant \theta \leqslant \theta_{\max} \\ 0 & \text{elsewhere,} \end{cases} \tag{7.352}$$

where $\Delta\theta = \theta_{\max} - \theta_{\min}$. A calculation of the likelihood $p(D|\theta, H_1, I)$ evidently requires knowledge of the data $D$ and the specificities of the model $H_1$, but, in general, one may write

$$\int_{\Delta\theta} p(D|\theta, H_1, I)\, d\theta \equiv \delta\theta \times p(D|\hat{\theta}, H_1, I) = \delta\theta \times L_{\max}, \tag{7.353}$$

where $L_{\max} \equiv p(D|\hat{\theta}, H_1, I)$ corresponds to the maximum value of the likelihood function achieved at the mode $\hat{\theta}$, and $\delta\theta$ thence corresponds to a "characteristic width" of the likelihood function, as illustrated in Figure 7.14. For a Gaussian likelihood, the integral in Eq. (7.353) would yield $\delta\theta = \sqrt{2\pi}\sigma_\theta$, where $\sigma_\theta$ is the error on the parameter $\theta$. In the more general case of a non-Gaussian likelihood, one may write $\delta\theta = \alpha\sigma_\theta$, where $\alpha$ is a constant determined by the specific shape of the likelihood function. This is immaterial, however, for the calculation of the global likelihood of the data given $H_1$

$$\begin{aligned} L(H_1) \equiv p(D|H_1, I) &= \int p(\theta|H_1, I) p(D|\theta, H_1, I)\, d\theta \\ &= \frac{1}{\Delta\theta} \int_{\Delta\theta} p(D|\theta, H_1, I)\, d\theta \\ &= p(D|\hat{\theta}, H_1, I) \frac{\delta\theta}{\Delta\theta}. \end{aligned} \tag{7.354}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

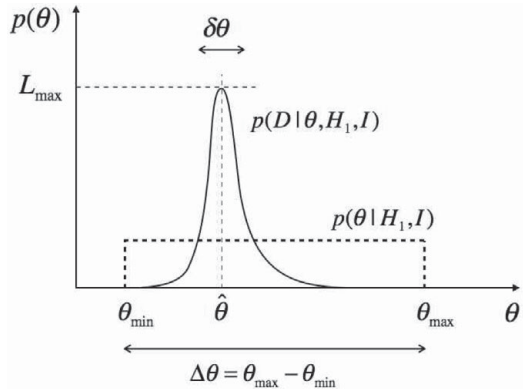


**Fig. 7.14** Schematic illustration of the notion of Occam factor, $\Omega_\theta = \delta\theta/\Delta\theta$. The dashed and solid curves represent the prior probability of the model parameter $\theta$ and the likelihood distribution of the data $D$ given hypothesis $H_1$, respectively.

The global likelihood, $L(H_1)$, of model $H_1$ is thus equal to the maximum likelihood of the model, denoted $L_{\max} \equiv L(\hat{\theta})$, achieved at the mode $\hat{\theta}$ multiplied by a purely geometrical factor $\delta\theta/\Delta\theta$, involving the width $\delta\theta$ of the likelihood distribution as well as the width $\Delta\theta$ of the nominal domain of the parameter $\theta$. Noting that the likelihood of hypothesis $H_0$ is simply

$$p(D|H_0, I) = p(D|\theta_0, H_0, I) \equiv L(\theta_0) \tag{7.355}$$

We can then complete the calculation of the Bayesian factor $B_{10}$ and write

$$\begin{aligned} B_{10} &= \frac{p(D|\hat{\theta}, H_1, I)}{p(D|\theta_0, H_0, I)} \frac{\delta\theta}{\Delta\theta}, \\ &= \frac{L(\hat{\theta})}{L(\theta_0)} \frac{\delta\theta}{\Delta\theta}. \end{aligned} \tag{7.356}$$

In general, the nominal parameter range $\Delta\theta$ is much wider than the posterior width $\delta\theta$. The ratio $\delta\theta/\Delta\theta \ll 1$ thus penalizes the model $H_1$ for the presence of its (extra) parameter $\theta$ relative to $H_0$, which has no free parameter. Indeed, the Bayes factor must satisfy

$$\frac{L(\hat{\theta})}{L(\theta_0)} \gg \frac{\Delta\theta}{\delta\theta} \tag{7.357}$$

in order to favor the hypothesis $H_1$. Since $\delta\theta/\Delta\theta \ll 1$, the maximum likelihood $L_{\max} \equiv L(\hat{\theta})$ must then be much larger than $L(\theta_0)$ to favor $H_1$. The search domain associated with the parameter $\theta$ thus indeed imposes a penalty on $H_1$ that the model must overcome by featuring a very large maximum likelihood $L(\hat{\theta})$ before it can be deemed a better description of the data. Effectively, the ratio $\delta\theta/\Delta\theta$ plays the role of Occam's razor, as it disfavors the more "complicated model," that is, the model with an extra parameter. The ratio is thus known as the **Occam factor** and commonly denoted $\Omega_\theta$. The global likelihood $p(D|H_1, I)$ may then be written

$$p(D|H_1, I) = L(\hat{\theta})\Omega_\theta. \tag{7.358}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

If the hypothesis $H_1$ has $m$ extra parameters $\theta_1, \ldots, \theta_m$, one similarly finds

$$p(D|H_1, I) = \int p(D|\theta_1, \ldots, \theta_m, H_1, I) \prod_{k=1}^{m} p(\theta_k|H_1, I)\, d\theta_k \tag{7.359}$$

$$= L(\hat{\theta}_1, \ldots, \hat{\theta}_m) \frac{\delta\theta_1}{\Delta\theta_1} \cdots \frac{\delta\theta_m}{\Delta\theta_m}$$

$$= L(\hat{\theta}_1, \ldots, \hat{\theta}_m) \Omega_{\theta_1} \cdots \Omega_{\theta_m}, \tag{7.360}$$

where $\Omega_1, \ldots, \Omega_m$ represent the Occam factors associated with the parameters $\theta_1, \ldots, \theta_m$. Under ideal circumstances, each Occam factor should be much smaller than 0.1. Consider, for instance, a case involving three (extra) fit parameters, each with an Occam factor of the order of 0.01. The ratio $L(\hat{\theta})/L(\theta_0)$ would then be required to be much in excess of $10^6$ to favor the hypothesis $H_1$. Effectively, the Occam factors collectively disfavor the alternative hypothesis, $H_1$, unless it is, in fact, correct and well supported by precise data. As we shall discuss with examples in §§7.7.4 and 7.7.4, measurement errors must in general be sufficiently small to render the test conclusive. If the data have large errors, the likelihood $L(\hat{\theta}_1, \ldots, \hat{\theta}_m)$ might be too small to overcome the Occam factors, and model $H_1$ will remain of questionable value. In the context of a linear model with Gaussian noise, the integral Eq. (7.353) implies $\delta\theta = \sqrt{2\pi}\sigma_\theta$, where $\sigma_\theta$ is the error on the parameter $\theta$. By virtue of Eq. (7.252), one then indeed expects the error $\sigma_\theta$ to scale as $\sigma_y/\sqrt{n}$, where $\sigma_y$ is the typical measurement error, and $n$ is the number of points in the dataset. The value of the Occam factor $\Omega_\theta = \delta\theta/\Delta\theta$ is thus dependent on the size of the measurement errors as well as the size of the data sample. Availability of better and larger datasets shall thus lead to smaller Occam factors, and thus more stringent demands on the maximum likelihood of the model $H_1$, and unless the ratio $L(\hat{\theta})/L(\theta_0)$ is sufficiently large, the Bayes factor will consistently favor the simpler model.

In general, there is little interest in calculating the Occam factors explicitly, but because they naturally arise as a result of the marginalization procedure, complex models (with many extra parameters) shall be transparently and automatically disfavored unless strongly supported by the data.

We illustrate the aforementioned principle of model comparison, and the effect of the Occam factors, with two examples: the first, presented in §7.7.4, involves the identification of a (new) signal amidst a noisy background, while the second, in §7.7.5, illustrates the comparison of models that differ by the addition of one or few parameters.

## 7.7.4 Example 1: Is There a Signal?

### Problem Definition

Let us consider a problem often encountered in physics: a search for a predicted signal amidst background noise. For simplicity's sake, we will here neglect much of the experimental considerations involved in such searches and reduce the complexity of the problem to one unknown: the amplitude of the signal (assuming the signal does exist). Let us

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

assume the measurement $D$ is reported in terms of a Fourier spectrum $\vec{n} = (n_1, n_2, \ldots, n_m)$ consisting of $m$ bins (channels) of equal width $\Delta f$ in the range $f_{\min} = 0$ to $f_{\max} = 100$ (in arbitrary units). Let us further assume, to keep the problem relatively simple, that the signal is expected to have a Gaussian line shape centered at a frequency $f_0 = 50.5$ and of width $\sigma_0 = 1.5$, as illustrated in Figure 7.15a. For illustrative purposes, we will carry out the analysis assuming four different measurement outcomes displayed in Figure 7.15b–e, which feature the same Gaussian noise characterized by a null expectation value, a standard deviation of $\sigma_n = 0.1$, and no bin-to-bin cross correlations. Spectra (b–e) have been generated with a signal of amplitude $A$ equal to 15.0, 5.0, 1.0, and 0.1 (arbitrary units), respectively but we will of course pretend we have no knowledge whatsoever of these amplitudes in our analysis. In fact, we will assume that a prior search of the signal has found the signal is weaker than some upper bound $A_{\max} = 100$ and that the theory stipulates the signal should have a minimum amplitude $A_{\min} = 0.01$ (again in arbitrary units).

## Comparing the Odds of Models $H_1$ and $H_0$

Our goal is to establish whether the observed spectra (b–e) provide evidence in support of a model $H_1$ stating the existence of the signal, or a null hypothesis $H_0$ asserting there is no such new signal. The presence of a signal is clearly obvious in panels b and c but rather difficult to establish visually in panels d and e. Our goal is to show how Bayesian inference can support this visual impression quantitatively when the signal seems obvious but can also provide odds of one model against the other even when the signal is not visually evident. In order to realize this goal, we will compare the posterior probabilities $p(H_1|D, I)$ and $p(H_0|D, I)$ of models $H_1$ and $H_0$, respectively, and calculate the posterior odds ratio of the two hypotheses

$$O_{10} \equiv \frac{p(H_1|D, I)}{p(H_0|D, I)}. \tag{7.361}$$

Recall, from §7.7.2, that the odds ratio $O_{10}$ may be expressed in terms of the prior odds ratio $O_{10}^{\text{prior}}$ and the Bayes factor $B_{10}$ defined as

$$O_{10} = O_{10}^{\text{prior}} B_{10} \tag{7.362}$$

with

$$O_{10}^{\text{prior}} = \frac{p(H_1|I)}{p(H_0|I)}, \tag{7.363}$$

and

$$B_{10} = \frac{p(D|H_1, I)}{p(D|H_0, I)} = \frac{\mathbb{L}(H_1)}{\mathbb{L}(H_0)}, \tag{7.364}$$

where $\mathbb{L}(H_1)$ and $\mathbb{L}(H_0)$ are the global likelihoods of the models $H_1$ and $H_0$, respectively.

Given the limited amount of prior information about the existence of the signal, it appears legitimate to set the prior odds ratio to unity, $O_{10}^{\text{prior}} = 1$. The posterior odds ratio shall

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

thus be entirely determined by the Bayes factor, that is, the ratio of the global likelihood of the two models.

It is important to note that we actually seek posterior probabilities for the models $H_1$ and $H_0$ irrespective of their "internal" parameter values. In the context of this problem, only $H_1$ has an internal parameter, and this parameter is the unknown amplitude $A$ of the purported signal. Since the actual value of this amplitude is irrelevant for the central goal of our study (i.e., establish the existence of the signal), we obtain $p(D|H_1, I)$ by marginalization according to

$$p(D|H_1, I) = \int_{\Delta A} p(A|H_1, I)p(D|H_1, A, I)\, dA \tag{7.365}$$

where $p(D|H_1, A, I)$ is the probability of observing the data $D$ given $H_1$ is valid and for a specific amplitude $A$, and $p(A|H_1, I)$ is the prior degree of belief in the strength $A$ of the signal given $H_1$ is true.

## Probability Model and Likelihoods

Let us first identify and discuss each of the components required in our Bayesian analysis of the spectra. The measured data $D$ (one the four spectra in Figure 7.15) are represented as a vector $\vec{n} = (n_1, n_2, \ldots, n_m)$ and values $n_i$ correspond to the number of counts (after background subtraction) in bins $i = 1, \ldots, m$. By construction, the bin content $n_i$ is the sum of the fraction of the signal $s_i$ expected in bin $i$ and the noise $e_i$ in that bin:

$$n_i = s_i + e_i. \tag{7.366}$$

Model $H_1$ tells us that, for a fixed amplitude $A$, the fraction of the signal $s_i$ is equal to

$$s_i = Ap_i \tag{7.367}$$

where

$$p_i = \int_{f_{i,\min}}^{f_{i,\max}} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(f - f_0)^2}{2\sigma_0^2}\right] df \tag{7.368}$$

is the relative fraction of the signal in each bin $i$, not a random number, determined by the model parameters $f_0$, $\sigma_0$, and the bin boundaries $f_{i,\min}$ and $f_{i,\max}$. The signal yield $s_i$ is thus a random number only insofar as $A$ might itself be considered a random number. The noise terms $e_i$, however, are random Gaussian deviates, which we assume all have the same standard deviation $\sigma_n$. The probability density that the noise $e_i$ be found in the range $[e_i, e_i + de_i]$ is

$$p(e_i|\sigma_n, I) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{e_i^2}{2\sigma_n^2}\right). \tag{7.369}$$

But, given the noise $e_i$ may be written as

$$e_i = n_i - s_i, \tag{7.370}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

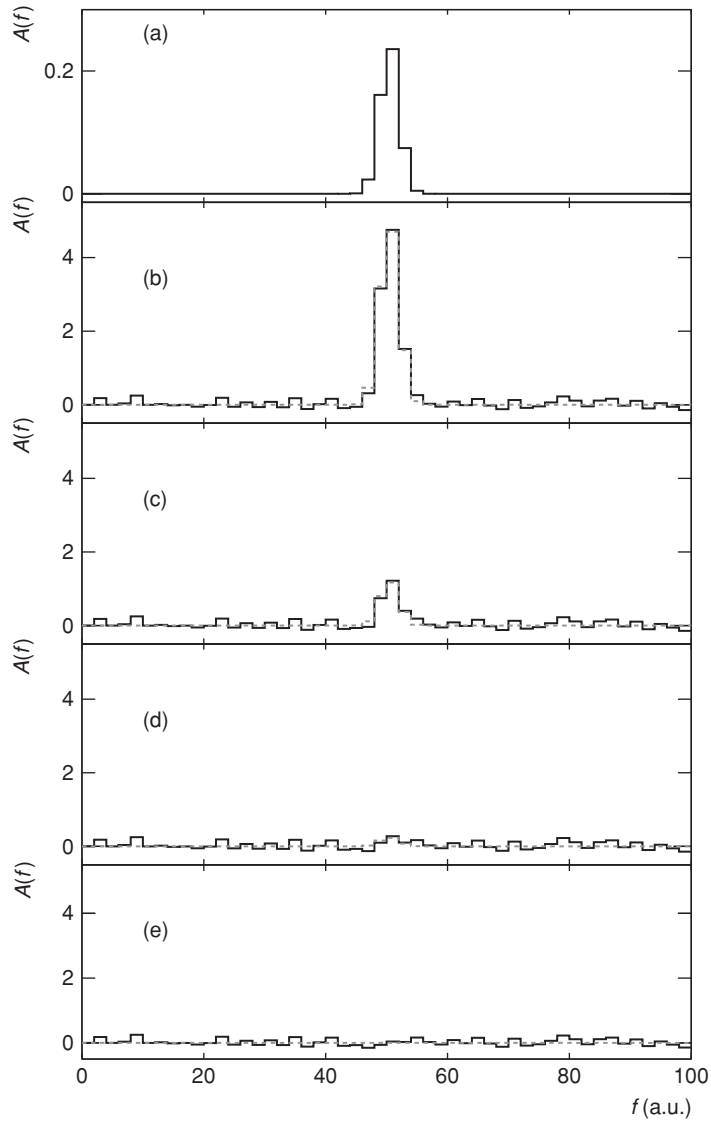


**Fig. 7.15** Simulated spectra used in the example of Bayesian inference presented in §7.7.4. (a) Signal line shape and position predicted by model $H_1$. (b–e) Simulated measurement outcomes with amplitudes $A = 20, 5, 1, 0.005$, and a Gaussian noise with standard deviation $\sigma = 0.1$. Best fits (posterior modes) are shown with dashed lines.

this implies that for a fixed signal amplitude $A$, the probability that the bin content $n_i$ will be found in the range $[n_i, n_i + dn_i]$ is equal to the probability the noise will be found in $[e_i, e_i + de_i]$, and we thus obtain the data probability model

$$p(n_i|H_1, A, \sigma_n, I) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{(n_i - Ap_i)^2}{2\sigma_n^2}\right]. \tag{7.371}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

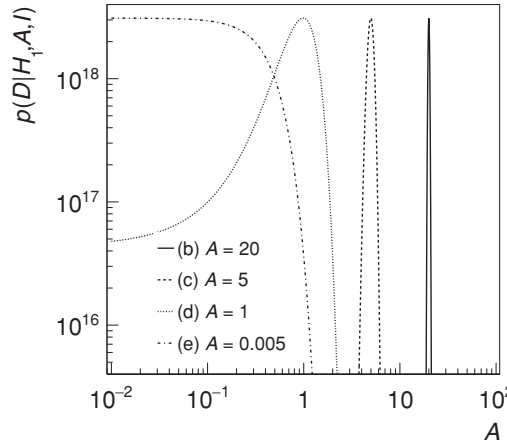


**Fig. 7.16** Likelihood distributions $p(D|H_1, A, I)$ vs. $A$ for the measurement outcomes shown in Figure 7.15b–e, with respective true amplitudes 20, 5, 1, and 0.005.

The likelihood of the data $D$ given the model $H_1$ and a fixed amplitude $A$ may then be written

$$\begin{aligned} p(D|H_1, A, I) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{(n_i - Ap_i)^2}{2\sigma_n^2}\right], \\ &= (2\pi)^{-m/2}\sigma^{-m} \exp\left[-\frac{\sum_{i=1}^{m}(n_i - Ap_i)^2}{2\sigma_n^2}\right], \\ &= (2\pi)^{-m/2}\sigma^{-m} \exp\left[-\frac{1}{2}\chi^2\right], \end{aligned} \tag{7.372}$$

where we introduced the $\chi^2$ function defined by

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - Ap_i)^2}{\sigma_n^2}. \tag{7.373}$$

The likelihood of the model $H_0$ is readily obtained by setting $A = 0$ in the preceding expressions.

$$p(D|H_0, \sigma_n, I) = (2\pi)^{-m/2}\sigma^{-m} \exp\left[-\frac{1}{2}\chi_0^2\right], \tag{7.374}$$

with

$$\chi_0^2 = \sum_{i=1}^{m} \frac{n_i^2}{\sigma_n^2}. \tag{7.375}$$

The likelihood distributions $p(D|H_1, A, \sigma_n, I)$ corresponding to spectra (b–e) are displayed in Figure 7.16 using a log-log scale to facilitate visualization of the tails of the distributions. Spectra (b) and (c) are characterized by very narrow $p(D|H_1, A, \sigma_n, I)$ distributions, with an approximate Gaussian shape, centered at the expected amplitudes and a maximum likelihood value of $3.1 \times 10^{18}$. Distributions associated with spectra (d) and

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

(e) have similar maximum likelihood values but feature non-Gaussian shapes. Somewhat wider, the likelihood distribution of spectrum (d) features a sizable low-side tail while the likelihood distribution of spectrum (e) is peaked near the origin, thereby indicating it is consistent with a signal of null amplitude (i.e., no signal).

We now wish to determine and compare the posteriors $p(H_0|D, I)$ and $p(H_1|D, I)$. Since $H_0$ has no amplitude parameter, its posterior may be obtained directly from the likelihood $p(D|H_1, I)$ of the measured spectra. Determination of $p(H_1|D, I)$, however, requires one marginalizes the $p(H_1, A|D, I)$ for $A$. This implies we must first formulate $p(A|H_1, I)$ expressing the prior probability density of the signal amplitude $A$ given $H_1$ is assumed valid. Note that to keep everybody honest, the prior should be selected before the measurement is carried out so that one is not tempted to look at the answer (i.e., the data) in order to make that choice. Here, the problem is academic, and we will actually explore what happens under two different choices of prior.

## Comparison of Uniform and Scaling Priors for $p(A|H_1, I)$

Obviously, the problem arises that we have only a rather limited amount of information about $H_1$ and more specifically the signal amplitude $A$. We do know, based on a prior experiment, that $A$ is smaller than some upper bound $A_{\max}$ and that it should, according to the model $H_1$, be larger than some minimal value $A_{\min}$. But that's it! How then should we formulate the prior $p(A|H_1, I)$? Given the lack of additional information, it seems natural to express our ignorance about $A$ by choosing a uniform prior

$$p(A|H_1, I) = \begin{cases} \frac{1}{\Delta A} & \text{for } A_{\min} \le A \le A_{\max} \\ 0 & \text{otherwise,} \end{cases} \tag{7.376}$$

with $\Delta A = A_{\max} - A_{\min}$. It is worth noting, however, that this choice implies the top range of values is far more probable than the bottom range. For illustrative purposes, assume $A_{\min} = 0.01$ and $A_{\max} = 10.0$, and let us calculate the ratio of the probabilities of finding $A$ in the ranges $1 \le A \le 10.0$ and $0.01 \le A \le 0.1$:

$$\frac{\int_1^{10} p(A|H_1, I)\, dA}{\int_{0.01}^{0.1} p(A|H_1, I)\, dA} = \frac{\int_1^{10} dA}{\int_{0.01}^{0.1} dA} = \frac{A|_1^{10}}{A|_{0.01}^{0.1}} = 100. \tag{7.377}$$

We find, indeed, that a uniform prior implies the upper range is far more probable than the lower range. But considering the signal sought for has never been observed and thus may qualify as a rare process, it would seem sensible to expect lower ranges of values should be more probable than larger values. A prior that gives higher probability to lower amplitudes therefore sounds far more reasonable. We thus choose[13] to use an uninformative scaling prior of the form, Eq. (7.92), introduced in §7.3.2, and write

$$p(A|H_1, I) = \begin{cases} \frac{1}{\ln(A_{\max}/A_{\min})} \frac{1}{A} & \text{for } A_{\min} \le A \le A_{\max} \\ 0 & \text{otherwise,} \end{cases} \tag{7.378}$$

[13] For comparative purposes, we actually use both uniform and scaling priors in the following.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

where the logarithmic factor ensures the proper normalization by integration over the range $A_{\min} \leq A \leq A_{\max}$. Also recall from §7.3.2 that integrals of this prior in the form of logarithms imply the log of the amplitude, $\ln A$, has a uniform distribution. Effectively, ranges such as $0.01 \leq A \leq 0.1$, $0.1 \leq A \leq 1$, and $1 \leq A \leq 10$ have equal (prior) probability, and ranges of higher values are not more probable than ranges of smaller values, as required. Such a prior thus seems far more appropriate, for a search for a weak signal, than the flat prior given by Eq. (7.376).

## Posteriors of $H_0$ and $H_1$

The hypothesis $H_0$ has no free parameters. Its posterior is thus obtained simply by application of Bayes' theorem:

$$p(H_0|D, I) = \frac{1}{p(D|I)} p(D|H_0, I), \tag{7.379}$$

with

$$p(D|I) = p(D|H_0, I) + p(D|H_1, I), \tag{7.380}$$

where $p(D|H_0, I)$ and $p(D|H_1, I)$ are the global likelihoods of hypotheses $H_0$ and $H_1$, respectively. Note that $p(H_0|I)$ was omitted in the numerator of Eq. (7.379) because we have chosen equal global model priors, that is, $p(H_0|I) = p(H_1|I)$. The global model prior $p(H_1|I)$ is likewise omitted in the following. Calculation of the posterior of $H_1$ requires marginalization of its free parameter $A$. With the uniform amplitude prior, one gets

$$p(H_1|D, I) = \frac{1}{p(D|I)} \frac{1}{\Delta A} \int_{A_{\min}}^{A_{\max}} p(D|H_1, A, \sigma_n, I)\, dA \tag{7.381}$$

whereas the scaling prior yields

$$p(H_1|D, I) = \frac{1}{p(D|I)} \frac{1}{\ln A_{\max}/A_{\min}} \int_{A_{\min}}^{A_{\max}} \frac{1}{A} p(D|H_1, A, \sigma_n, I)\, dA. \tag{7.382}$$

Results from the evaluations of the posteriors $p(H_0|D, I)$ and $p(H_1|D, I)$, with Eqs. (7.379, 7.381, 7.382) are shown in Figure 7.17 for the four measurement outcomes introduced in Figure 7.15b–e, while odds ratios calculated from these probabilities are listed in Table 7.1 for both uniform and scaling priors. The posterior probability, $p(H_1|D, I)$, is unequivocally equal to unity for the clear signals shown in Figure 7.16b and c and the choice of prior makes no significant difference in these two cases. However, in the case of the much weaker signals displayed in Figure 7.16d and e, the choice of a scaling prior for the amplitude $A$ appreciably increases the posterior probability $p(H_1|D, I)$ and the odds ratios $O_{10}$. A prior favoring weaker signals should indeed yield a larger posterior for weak signals. It should be noted, however, that an odds ratio of 8.0 constitutes a tantalizing but somewhat weak indication that there might be a signal in the spectrum (d). Indeed, the use of a scaling prior significantly increases the odds ratio, from 0.6 to 8.0, but the latter value is not sufficiently large to provide incontrovertible evidence for a signal. In the case of spectrum (e), the scaling prior produces a significant increase of the odds ratio in favor

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

**Table 7.1** Odds ratios $O_{10}$ in favor of model $H_1$ relative to model $H_0$ calculated with uniform ($u$) and scaling ($s$) priors, for data samples shown in Figure 7.16

| Dataset | $O_{10}^{(u)}$ | $O_{10}^{(s)}$ |
|---|---|---|
| $A = 20$ | $\infty$ | $\infty$ |
| $A = 5$ | $1.6 \times 10^{45}$ | $3.6 \times 10^{45}$ |
| $A = 1$ | 0.60 | 8.0 |
| $A = 0.005$ | 0.004 | 0.38 |

of model $H_1$, but most scientists would likely regard a value of 0.38 as far too small to conclude the spectrum contains a signal, even though it actually does in the simulation.

Although not absolutely essential, it is interesting to evaluate the Occam factor of model $H_1$. Based on Eq. (7.353) and assuming a uniform prior for $A$, we write

$$\int_{\Delta A} p(D|\theta, H_1, I)\, dA = \delta A \times L_{\max}, \tag{7.383}$$

We carry out numerical evaluations of the integral in Eq. (7.383) for all four spectra ($A = 20, 5, 1, 0.005$) and divide by the observed maximum likelihood values $L_{\max} \approx 3.1 \times 10^{18}$ to find widths $\delta A \approx 0.84$ for spectra $A = 20, 5, 1$ and a width $\delta A \approx 0.41$ for $A = 0.005$. Dividing these by the width of the search domain, $\Delta A = 100$, we find Occam factors $\Omega = \delta A / \Delta A$ of order 0.008 and 0.004, respectively. Using a scaling prior does not appreciably changes these values. Clearly, these small Occam factor values strongly disfavor hypothesis $H_1$ and given hypothesis $H_0$ has a likelihood of order $3 \times 10^{18}$ for null signals (noise only), the posterior of $H_1$ is given a small probability for spectra (d) and (e).

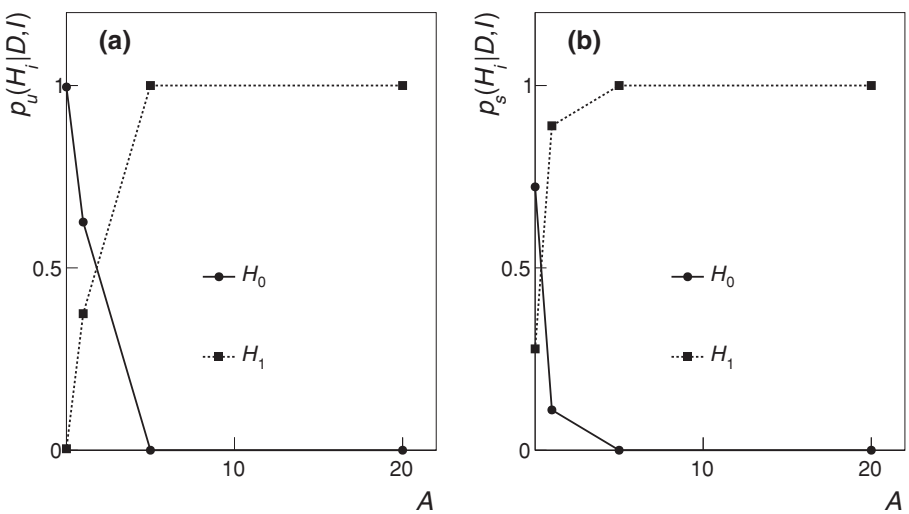


**Fig. 7.17** Posteriors $p(H_0|D, I)$ and $p(H_1|D, I)$ for the measurement outcomes shown in Figure 7.15b–e, with respective true amplitudes $A = 20, 5, 1$, and 0.005, based on (a) uniform and (b) scaling priors.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

However, the likelihood of $H_0$ falls dramatically for strong signals and $H_1$ is then ascribed a high probability in spite of the small value of the Occam factor.

### 7.7.5 Example 2: Comparison of Linear and Quadratic Models

#### Problem Definition

An issue often encountered in practical data analyses is whether one or several parameters should be added to a model to obtain a better fit of the data. This is the case, for instance, with the study of Fourier decompositions of angular correlations, or fits of data with polynomials. In either case, one must decide whether the data warrant the addition of one or more additional parameters, that is, higher-order Fourier terms or higher degree terms in a polynomial.

As a practical case, let us revisit the comparison of linear and quadratic models toward the description of a dataset $D$ introduced in §6.6.5 under the frequentist paradigm. For the purpose of this example, we simulated datasets consisting of $n = 25$ measurements $(x_i, y_i \pm \sigma_i)$ of an observable $Y$ measured as a function of a control (independent) observable $X$ generated based on the data model

$$y(x) = \mu_y(x) + R, \tag{7.384}$$

where $R$ represents random Gaussian deviates with standard deviation $\sigma_i = \gamma\sqrt{\mu_y(x)}$, and $\mu_y(x)$ expresses the physical relation between observables $Y$ and $X$, here taken to be a second-degree polynomial:

$$\mu_y(x) = b_0 + b_1 x + b_2 x^2, \tag{7.385}$$

with fixed coefficients

$$\begin{aligned} b_0 &= 500.0, \\ b_1 &= 10.0, \\ b_2 &= -1.0. \end{aligned}$$

The scaling factor $\gamma$ was set to arbitrary values $\{2.0, 1.0, 0.5, \ldots\}$ in order to generate datasets of different precision levels. Each of the produced datasets was then analyzed to establish whether they could be properly represented as straight lines or whether a quadratic term should be included. Evidently, in the analysis, we assumed the quadratic dependence of $Y$ on $X$ is not established or known. We thus formulated a null hypothesis, $H_0$, stipulating that a linear model is sufficient to "explain" the data while the alternative hypothesis, $H_1$, requires the use of quadratic model, that is, a fit with a second-degree polynomial.

$$\begin{aligned} H_0 &: f_0(x) = a_0 + a_1 x, \\ H_1 &: f_1(x) = a_0 + a_1 x + a_2 x^2. \end{aligned}$$

#### Determination of the Odds Ratio of Global Posteriors

We saw in the context of the frequentist paradigm (§6.6.5) that the $\chi^2$ of fits based on two distinct hypotheses, a linear and a quadratic model, may be used as a statistical test to

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

challenge and reject the linear hypothesis (null hypothesis) provided its $\chi^2$ exceeds some minimal value determined by the significance level of the test. The null hypothesis shall be retained, however, if the $\chi^2$ falls below the cut. The problem, of course, is that the $\chi^2$ obtained with the quadratic fit might also be relatively small. The test then features a relatively small power and thus cannot be deemed significant.

The necessity to consider both the significance level of a test and its power raises both technical and philosophical issues. Is it possible, in particular, to identify a statistic whose value might combine both the significance level and the power of the test? Can a statistic be found that is capable of rejecting the null hypothesis while showing that the alternative hypothesis has strong merits? Can such a statistic disfavor overly complicated models not properly supported by data? The short answer is that an odds ratio based on the global posteriors of alternative and null hypotheses in fact satisfy (mostly) all these criteria.

Let us indeed reexamine a Bayesian test based on the odds ratio of the global posteriors of hypotheses $H_1$ and $H_0$ in the context of the comparison of linear and quadratic models and show that the odds ratio provides an effective tool to meaningfully sustain or reject a null hypothesis challenged by a more complicated model, that is, a model with more free parameters.

The odds ratio is defined and calculated according to

$$O_{10} = \frac{p(H_1|D)}{p(H_0|D)} = O_{10}^{\text{prior}} B_{10} \tag{7.386}$$

with

$$O_{10}^{\text{prior}} = \frac{p(H_1|I)}{p(H_0|I)}, \tag{7.387}$$

and

$$B_{10} = \frac{p(D|H_1, I)}{p(D|H_0, I)} \equiv \frac{\mathbb{L}(H_1)}{\mathbb{L}(H_0)}, \tag{7.388}$$

where $\mathbb{L}(H_1)$ and $\mathbb{L}(H_0)$ are the global likelihoods of the models $H_1$ and $H_0$, respectively. These can be calculated by marginalization of the model likelihoods $p(D|H_0, a_0, a_1, I)$ and $p(D|H_1, a_0, a_1, a_2, I)$, respectively.

Given there are no tangible reasons to favor either hypothesis,[14] we set the prior odds ratio $O_{10}^{\text{prior}}$ to unity. The posterior odds ratio is thus determined solely by integrals of the likelihood functions $p(D|H_0, a_0, a_1, I)$ and $p(D|H_1, a_0, a_1, a_2, I)$ over the domains of parameters $a_0$, $a_1$, and $a_2$. In order to determine sensible ranges of integration for these parameters, we first plot the likelihood functions $p(D|H_0, a_0, a_1, I)$ and $p(D|H_1, a_0, a_1, a_2, I)$, in Figure 7.18, as a function of their parameters. The likelihood functions are calculated numerically according to

$$p(D|H_0, a_0, a_1, I) \propto \exp\left[-\frac{1}{2}\chi_0^2(a_0, a_1)\right], \tag{7.389}$$

$$p(D|H_1, a_0, a_1, a_2, I) \propto \exp\left[-\frac{1}{2}\chi_1^2(a_0, a_1, a_2)\right], \tag{7.390}$$

[14] Remember that we pretend to ignore the datasets were generated with a quadratic model.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

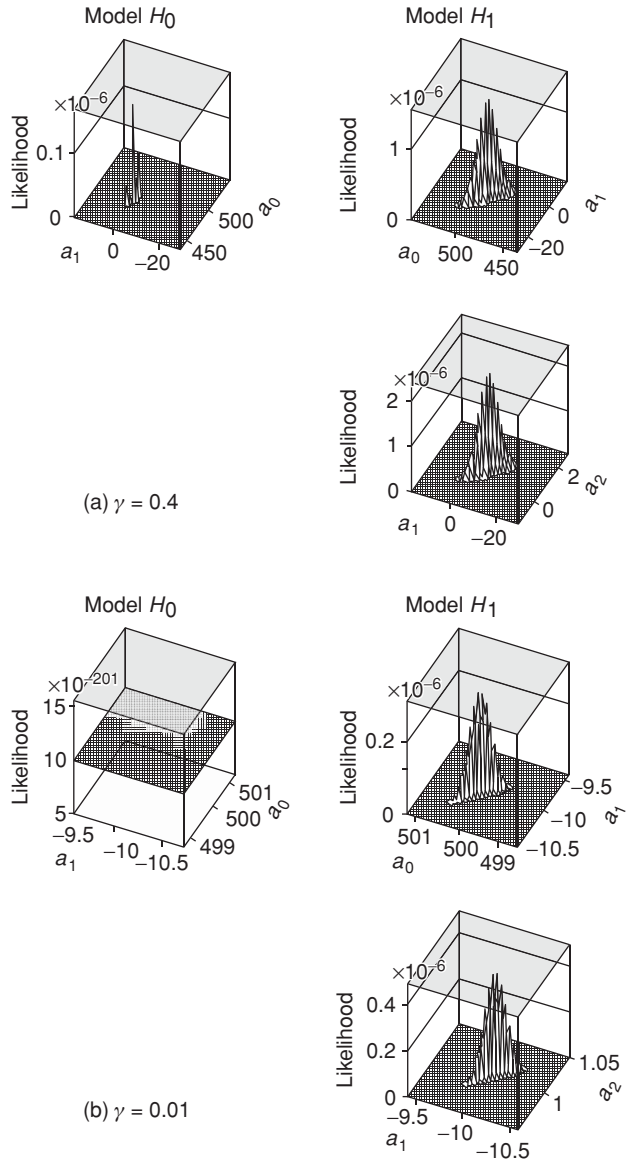


**Fig. 7.18** Examples of the likelihoods $p(D|H_0, a_0, a_1, I)$ of a linear model and $p(D|H_0, a_0, a_1, a_2, I)$ of a quadratic model obtained with (a) large noise ($\gamma = 0.4$) and (b) very small noise ($\gamma = 0.01$).

where

$$\chi_0^2(a_0, a_1) = \sum_{k=1}^{n} \frac{\left(y_i - \sum_{n=0}^{1} a_n x^n\right)^2}{\sigma_i^2},$$

$$\chi_1^2(a_0, a_1, a_2) = \sum_{k=1}^{n} \frac{\left(y_i - \sum_{n=0}^{2} a_n x^n\right)^2}{\sigma_i^2}. \tag{7.391}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

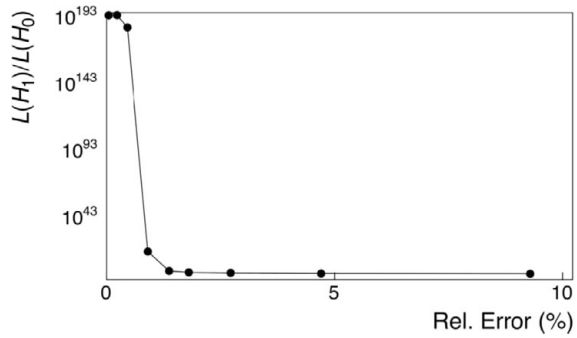


**Fig. 7.19** Bayes factor $B_{10} = \mathbb{L}(H_1)/\mathbb{L}(H_0)$ as a function of the relative error size (determined by the error scaling factor $\gamma$).

Bounds of integration are chosen to approximately correspond to $\pm 6$ times the widths of the distributions along each axis, and thus such that the functions are negligible (relative to the maximum) beyond the bounds. Marginalization of the likelihood functions $p(D|H_0, a_0, a_1, I)$ and $p(D|H_1, a_0, a_1, a_2, I)$ is obtained by numerical integration within these bounds. The Bayes factors and odds ratios are then computed according to Eqs. (7.388, 7.386). Calculated values of the Bayes factor $B_{10} = \mathbb{L}(H_1)/\mathbb{L}(H_0)$ are presented in Figure 7.19 for simulated datasets produced with $\gamma$ values ranging from 0.01 to 2. One finds that for large relative errors (i.e., obtained with large values of $\gamma$), the odds ratio is smaller than or of order unity and thus favors the null hypothesis (i.e., a linear model of the data). For small errors, obtained with $\gamma \leq 0.3$, the odds ratios increases significantly and unambiguously favors the alternative hypotheses, that is, the notion that the signal $Y$ involves a quadratic dependence on $X$.

The posteriors $p(H_1|D, I)$ and $p(H_0|D, I)$ are statistics (i.e., functions of the measured data points). As such, they would fluctuate sample by sample, much like a $\chi^2$ function, if it were in fact possible to acquire several distinct data samples. As a ratio of these two statistics, the posterior odds ratio $O_{10}$ is thus also a statistic, and it too shall fluctuate sample to sample. This is illustrated in the three sets of panels of Figure 7.20. The left set of panels displays linear and quadratic model fits of typical simulated spectra obtained with selected values of the error scaling parameter $\gamma$, while the central and right sets of panels display $\chi^2$/DoF distributions and histograms of the logarithm of the odds ratio determined for successions of 1,000 Monte Carlo generated samples of $n = 25$ points. From top to bottom, the samples are generated with error scale parameter values of $\gamma = 2.0$, 0.4, 0.2, and 0.01. One finds that for $\gamma = 2.0$ and $\gamma = 0.4$ corresponding to relative errors of the order of 10% and 2%, respectively, the $\chi^2$ distributions of the two hypotheses overlap considerably and the log of the odds ratio of the global posteriors is sharply distributed near zero, implying the odds ratio is of order 1. For smaller relative errors, the $\chi^2$ distributions no longer overlap and the odds ratio grows increasingly large, thereby indicating that the alternative hypothesis $H_1$ is strongly favored by the data.

It is interesting to note that much like the $\chi^2$ of the fits, the odds ratio fluctuates quite considerably. This stems from the fact that these quantities are in fact not independent of one another. Indeed, Eqs. (7.389,7.390) tell us that for Gaussian deviates, the likelihood of

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

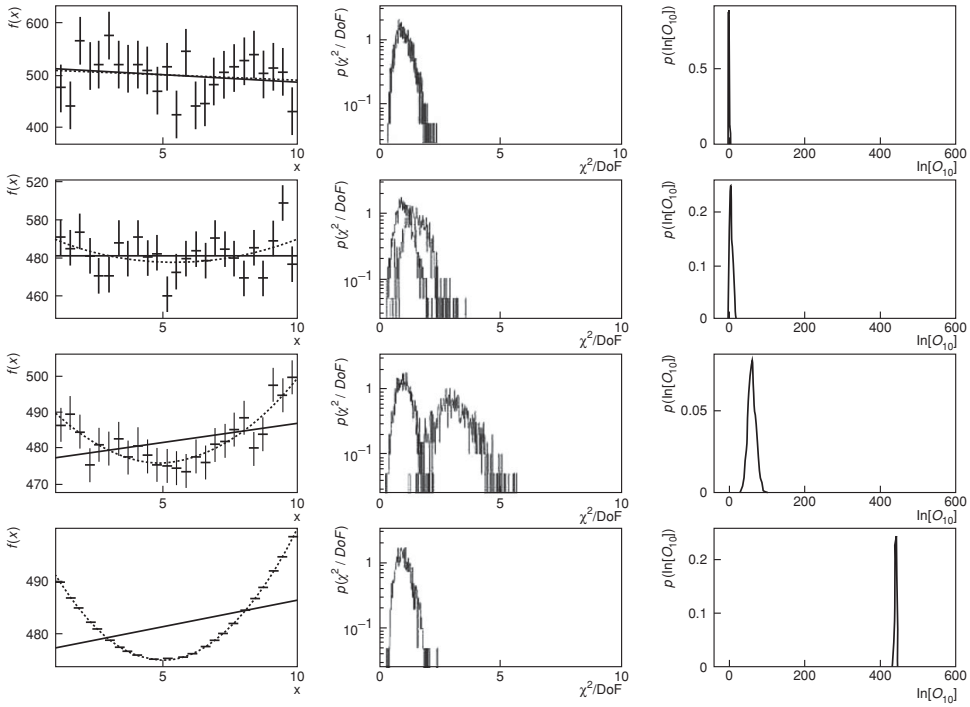


**Fig. 7.20** (Left) Representative fits of the linear and quadratic models obtained with, from top to bottom, $\gamma = 2.0, 0.4, 0.2,$ and 0.01. (Middle) Histograms of the $\chi^2/\mathrm{DoF}$ obtained with the two models for 1000 distinct samples of 25 points (as described in the text) generated with same values of $\gamma$. (Right) Histograms of the logarithm of the odds ratios obtained with the same samples and $\gamma$ values.

a dataset, given specific models parameters $a_0, a_1, \ldots,$ is strictly related to the $\chi^2$ obtained with those parameters. And, given the models $H_0$ and $H_1$ are linear in their coefficients, one can write

$$\begin{aligned}\chi_1^2(a_0, a_1, a_2) &= \sum_{i=1}^{n} \frac{\left([y_i - a_0 - a_1 x] - a_2 x_i^2\right)^2}{\sigma_i^2} \\ &= \chi_0^2(a_0, a_1) - 2a_2 \sum_{i=1}^{n} \frac{[y_i - a_0 - a_1 x]\, x_i^2}{\sigma_i^2} \\ &\quad + a_2^2 \sum_{i=1}^{n} \frac{x_i^4}{\sigma_i^2}\end{aligned}$$

The Bayes factor may thus be considered a function of the model parameters

$$\begin{aligned}B_{10} &= \exp\left[-\frac{1}{2}\left(\chi_1^2 - \chi_0^2\right)\right] \\ &= a_2 \sum_{i=1}^{n} \frac{[y_i - a_0 - a_1 x]\, x_i^2}{\sigma_i^2} - \frac{1}{2} a_2^2 \sum_{i=1}^{n} \frac{x_i^4}{\sigma_i^2},\end{aligned} \tag{7.392}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

which, evidently, peak for some specific values of the model parameters $a_0$, $a_1$, and $a_2$. Effectively, there is no new information contained in the odds ratio $O_{10}$ or the Bayes factor $B_{10}$, given the one-to-one relation between the global likelihoods of hypotheses $H_1$ and $H_0$ and their respective $\chi^2$ functions. However, the Bayes factor and odds ratio provide a practical and simple tool to decide whether the null hypothesis should sensibly be rejected.

With a single value, the odds ratio enables the rejection/acceptance of the null hypothesis and effectively provides a minimal statement about the power of the test. If the odds ratio is small (i.e., not much larger than unity), then the null hypothesis cannot be rejected. This can be the case either because the data are very precise and the null hypothesis happens to be correct, or because the data are poor (large errors) and the power of the test is too weak to enable a decision. In this latter case, a single number suffices to make the statement, and one is not required to explicitly calculated the power $\beta$ of the test, according to Eq. (6.79). If, instead, the value of the odds ratio is very large, one immediately learns that the data are incompatible with the null hypothesis and simultaneously in favor of the alternative hypothesis. Again, there is no need to calculate the power of the test. There is, of course, a possibility that the $\chi^2$ of the null hypothesis might be large due to fluctuations, but the large value of the odds ratio readily indicates that the $\chi^2$ of the alternative hypothesis is in fact very good. It then makes sense to (tentatively) reject the null hypothesis because the large value of the odds ratio guarantees the alternative hypothesis is a good fit to the data.

The calculation of an odds ratio thus has several advantages, both conceptually and in practice. It enables the rejection/acceptance of the null/alternative hypothesis on the basis of a single number without the need for possibly complicated and cumbersome calculations of the power of the test. The decision may then be taken rapidly and efficiently, either by a human or by an algorithm (machine). The value of the odds ratio also readily informs us about the quality of the decision and thus the power of the test. Clearly, an odds ratio of 1000 is far more significant than a ratio of 10 or 100. And likewise, a ratio in excess of $10^6$ or $10^9$ readily indicates that the alternative hypothesis is very strongly favored by the data.

### 7.7.6 Minimal Odds Ratio for a Discovery

As we discussed in the context of the frequentist interpretation of probability (see also §6.6.2), scientists commonly require a null hypothesis to be incompatible with the data with a significance level of $3 \times 10^{-7}$ to claim and announce a discovery. That means the $p$-value of the null hypothesis must be equal to or less than $3 \times 10^{-7}$. For instance, a peak in an invariant mass plot must be deemed incompatible with a background fluctuation because it has a probability equal to or smaller than $3 \times 10^{-7}$ before it can be considered acceptable to claim a discovery of a new particle. What is then the Bayesian equivalent of this significance level? What indeed should be the minimal odds ratio required to claim a discovery?

Although, ab initio, there is no simple or correct answer to this question, one can establish a minimal criterion, which can be deemed both necessary and sufficient by most scientists, and applicable in most situations. Clearly, the larger the odds ratio is, the more likely is the alternative hypothesis to be correct. One should thus define a minimal value

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

for the odds ratio. What should this minimum value be to provide a significance level of $3 \times 10^{-7}$?

In the frequentist approach, a test of the null hypothesis is based on a chosen statistic, called a test statistic. The data are considered incompatible with the null hypothesis if the probability to obtain a statistic in excess of the observed value is smaller than $3 \times 10^{-7}$, while the probability of the observed statistic for the alternative hypothesis is large (but often unspecified). The probability $p_t$ to obtain a statistic $t$ value in excess of the observed value $t_o$ is given by

$$p_t = \int_{t_o}^{\infty} p(t|H_i)\, dt, \tag{7.393}$$

where $p(t|H_i)$ represents the PDF of $t$ given the hypothesis $H_i$ is true. The integration to infinity sums over the probability of all suboptimal parameter values. In effect, it corresponds to the marginalization of the model parameters, and thus corresponds to the probability of the null hypothesis given any parameter value.

In the Bayesian paradigm, there is no need for an *extra* statistical test; the global posterior probability of an hypothesis *is* the test. It expresses the probability of a hypothesis to be true by marginalization (as necessary) of all model parameters, that is, given any parameter values. Effectively, there is then a one-to-one correspondence between the probability of a (frequentist) test statistic to be found in excess of a specific value and the posterior probability. It is thus reasonable to require the posterior probability of the null hypothesis to be equal or smaller than $3 \times 10^{-7}$. If the alternative and the null hypotheses can be considered an exhaustive set of options, the posterior probability of the alternative hypothesis is equal to $1 - 3 \times 10^{-7}$. The discovery criterion thus corresponds to an odds ratio with a minimum value of $3 \times 10^6$:

$$O_{10}^{\text{discovery}} = \frac{1 - 3 \times 10^{-7}}{3 \times 10^{-7}} = 3 \times 10^6. \tag{7.394}$$

Indication of a new phenomenon, which usually requires a $3\sigma$ significance level, shall have a correspondingly smaller minimum odds ratio value.

A few remarks are in order. First, one must recognize that an odds ratio is a statistic, as are the respective global posterior probabilities of the hypotheses used in the calculation of the ratio. This means that the odds ratio obtained from several samples extracted from a common parent population (using the same experimental procedure and apparatus) shall have values that appear to fluctuate from sample to sample, as illustrated in Figure 7.20 for the curve fitting example discussed in §7.7.5. It is thus conceivable, owing to the vagaries of the measurement protocol, that the odds ratio of a particular data sample might meet the discovery criterion threshold, while some others do not. Indeed, because of fluctuations (i.e., experimental errors) a given data sample might be less of a match to a specific hypothesis.

Second, it is important to realize that the notion of odds ratio does not require calculation of an additional variable to determine the power of a test. If two hypotheses are truly exhaustive (i.e., with no other options in the model space of a particular phenomenon), the probability of one is the complement of the other (i.e., $p(H_1|D, I) = 1 - p(H_0|D, I)$) and the odds ratio is perfectly well defined and determined. There is no need for a test

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

threshold and thus for the notion of power as defined by Eq. (6.79). In fact, the odds ratio itself provides an assessment of the discriminating power of the test. On the one hand, if the odds ratio is very large, $O_{10} \gg 1$, or very small, $O_{10} \ll 1$, one knows immediately that the test strongly favors a specific hypothesis. The measured data thus provide strong discriminatory power in the selection of the best hypothesis. On the other hand, if the ratio is of order unity, $O_{10} \sim 1$, it is clear that given the experimental errors, neither the null nor the alternative hypotheses are favored by the data. The "test" thus has essentially no discriminatory power toward the selection or adoption of $H_0$ and $H_1$.

Third, and last, one may contrast the notion of odds ratio with the Neyman–Pearson test (as a most powerful test for simple hypotheses). Recall from §6.7.2 that a Neyman–Pearson test is based on the ratio of the likelihoods of the alternative and null hypotheses. For instance, for two competing hypotheses about the value of a specific parameter $\theta$, one writes

$$r_n(\vec{x}|\theta_0, \theta_1) = \frac{f_n(\vec{x}|\theta_1)}{f_n(\vec{x}|\theta_0)}, \tag{7.395}$$

where $\vec{x}$ represents a set of $n$ measured points, $f_n(\vec{x}|\theta_0)$ and $f_n(\vec{x}|\theta_1)$ are the likelihoods of these data points given parameter values $\theta = \theta_0$ (hypothesis $H_0$), and $\theta = \theta_1$ (hypothesis $H_1$). The power of the test is then calculated as the expectation value of $r_n(\vec{x}|\theta_0, \theta_1)$ given the null hypothesis, $H_0$, is true, a rather unintuitive notion to say the least. However, note that the notion of using a ratio of likelihoods (for the alternative and null hypotheses) is totally in line with a Bayesian test of the null hypothesis based on the odds ratio $O_{10}$, defined as a ratio of the global posterior probabilities of the two hypotheses. The advantage of the odds ratio is that it is conceptually far easier to visualize and interpret. The only difference, of course, is that posterior probabilities may also include nontrivial (i.e., nonuniform) prior probabilities of the model parameters, which are subsequently marginalized to obtain the global posteriors. In essence, the odds ratio thus provides a Bayesian extension of the Neyman–Pearson test, and because its calculation involves all facets of the model and all of the data, it is guaranteed to be the most powerful test. Consequently, if an odds ratio of order unity, $O_{10} \sim 1$, is obtained in the comparison of two hypotheses $H_1$ and $H_0$, no manipulation of the data or use of other statistics shall improve the power of the test and discrimination of the hypotheses. Rejection of the null hypothesis $H_0$ in favor of the alternative $H_1$ thus rests entirely on the quality of the data (i.e., the size of the errors) used to carry out the test.

## Exercises

**7.1** Show that the use of the measure $q_i$ given by Eq. (7.82), with a constraint on the variance according to Eq. (7.74), leads to a truncated Gaussian prior PDF of the form

$$p(x|\mu, \sigma) = \begin{cases} \frac{1}{\Phi(z_L) - \Phi(z_H)} \frac{1}{\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] & \text{for } x_L \leq x \leq x_H \\ 0 & \text{elsewhere,} \end{cases} \tag{7.396}$$

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press

where $z_L = (x_L - \mu)/\sigma$ and $z_H = (x_H - \mu)/\sigma$, corresponding to the boundaries within which the observable is allowed (or defined).

**7.2** Verify that if the variance parameter $\sigma^2$ of a Gaussian PDF is given a prior of the form $p(\sigma^2) = k/\sigma^2$, then the log of the variance has a flat prior.

**7.3** Show that Jeffreys' prior for a multinomial distribution with rate parameters $\vec{p} = (p_1, p_2, \ldots, p_m)$, with the constraint $\sum_{i=1}^{m} p_i = 1$, is the Dirichlet distribution (§3.8) with all its parameters set to half.

**7.4** Demonstrate that the expression for $Q(\mu)$, given by Eq. (7.139), may be transformed to yield

$$Q(\mu) = \left(n\xi_0 + \xi_p\right)\left[\mu - \frac{n\xi_0\bar{x} + \mu_p\xi_p}{n\xi_0 + \xi_p}\right]^2 + K,$$

where $K$ represents a constant expression independent of $\mu$.

**7.5** Derive the expression, Eq. (7.153), for the posterior of the precision $\xi$ based on Eq. (7.152).

**7.6** Derive the expression, Eqs. (7.240), (7.242), and (7.243), for the Bayesian solution of a linear model fit with Gaussian priors for the model parameters.

https://doi.org/10.1017/9781108241922.009 Published online by Cambridge University Press