Bayesian Inference in Processing Experimental Data Principles and Basic Applications

G. D'Agostini
Università "La Sapienza" and INFN, Roma, Italia

Abstract

This report introduces general ideas and some basic methods of the Bayesian probability theory applied to physics measurements. Our aim is to make the reader familiar, through examples rather than rigorous formalism, with concepts such as: model comparison (including the automatic *Ockham's Razor* filter provided by the Bayesian approach); parametric inference; quantification of the uncertainty about the value of physical quantities, also taking into account systematic effects; role of marginalization; posterior characterization; predictive distributions; hierarchical modelling and hyperparameters; Gaussian approximation of the posterior and recovery of conventional methods, especially maximum likelihood and chi-square fits under well defined conditions; conjugate priors, transformation invariance and maximum entropy motivated priors; Monte Carlo estimates of expectation, including a short introduction to Markov Chain Monte Carlo methods. [0]

1 Introduction

The last decades of 20^{th} century have seen an intense expansion in the use of Bayesian methods in all fields of human activity that generally deal with uncertainty, including engineering, computer science, economics, medicine and even forensics (Kadane and Schum 1996). Bayesian networks (Pearl 1988, Cowell *et al* 1999) are used to diagrammatically represent uncertainty in expert systems or to construct artificial intelligence systems. Even venerable metrological associations, such as the International Organization for Standardization (ISO 1993), the Deutsches Institut für Normung (DIN 1996, 1999), and the USA National Institute of Standards and Technology (Taylor and Kuyatt 1994), have come to realize that Bayesian ideas are essential to provide general methods for quantifying uncertainty in measurements. A short account of the Bayesian upsurge can be found in a Science article (Malakoff 1999). A search on the web for the keywords 'Bayesian,' 'Bayesian network,' or 'belief network' gives one a dramatic impression of this 'revolution,' not only in terms of improved methods, but more importantly in terms of reasoning. An overview of recent developments in Bayesian statistics, may be found in the proceedings of the Valencia Conference series. The last published volume was (Bernardo *et al* 1999), and the most recent conference was held in June 2002. Another series of workshops, under the title of *Maximum Entropy and Bayesian Methods*, has focused more on applications in the physical sciences.

It is surprising that many physicists have been slow to adopt these 'new' ideas. There have been notable exceptions, of course, many of whom have contributed to the abovementioned Maximum Entropy workshops. One reason to be surprised is because numerous great physicists and mathematicians have played important roles in developing probability theory. These 'new' ideas actually originated long ago with Bernoulli, Laplace, and Gauss, just to mention a few

[0]Invited article to be published in *Reports on Progress in Physics*.
Email: dagostini@roma1.infn.it, URL: http://www.roma1.infn.it/~dagos.

who contributed significantly to the development of physics, as well as to Bayesian thinking. So, while modern statisticians and mathematicians are developing powerful methods to apply to Bayesian analysis, most physicists, in their use and reasoning in statistics still rely on 20$^{\text{th}}$ century 'frequentist prescriptions' (D'Agostini 1999a, 2000).

We hope that this report will help fill this gap by reviewing the advantages of using the Bayesian approach to address physics problems. We will emphasize more the intuitive and practical aspects than the theoretical ones. We will not try to cover all possible applications of Bayesian analysis in physics, but mainly concentrate on some basic applications that illustrate clearly the power of the method and how naturally it meshes with physicists' approach to their science.

The vocabulary, expressions, and examples have been chosen with the intent to correspond, as closely as possible, to the education that physicists receive in statistics, instead of a more rigorous approach that formal Bayesian statisticians might prefer. For example, we avoid many important theoretical concepts, like exchangeability, and do not attempt to prove the basic rules of probability. When we talk about 'random variables,' we will in fact mean 'uncertain variables,' and instead of referring to the frequentist concept of 'randomness' à la von Mises (1957). This distinction will be clarified later.

In the past, presentations on Bayesian probability theory often start with criticisms of 'conventional,' that is, frequentist ideas, methods, and results. We shall keep criticisms and detailed comparisons of the results of different methods to a minimum. Readers interested in a critical review of conventional frequentist statistics will find a large literature, because most introductory books or reports on Bayesian analysis contain enough material on this matter. See (Gelman *et al* 1995, Sivia 1997, D'Agostini 1999c, Jaynes 1998, Loredo 1990) and the references therein. Eloquent 'defenses of the Bayesian choice' can be found in (Howson and Urbach 1993, Robert 2001).

Some readers may wish to have references to unbiased comparisons of frequentist to Bayesian ideas and methods. To our knowledge, no such reports exist. Those who claim to be impartial are often frequentists who take some Bayesian results as if they were frequentist 'prescriptions,' not caring whether all underlying hypotheses apply. For two prominent papers of this kind, see the articles by Efron (1986a) [with follow up discussions by Lindley (1989), Zellner (1986), and Efron (1986b)] and Cousins (1995). A recent, pragmatic comparisons of frequentist and Bayesian confidence limits can be found in (Zech 2002).

Despite its lack of wide-spread use in physics, and its complete absence in physics courses (D'Agostini 1999a), Bayesian data analysis is increasingly being employed in many areas of physics, for example, in astronomy (Gregory and Loredo 1992, 1996, Gregory 1999, Babu and Feigelson 1992, 1997, Bontekoe *et al* 1994), in geophysics (Glimm and Sharp 1999), in high-energy physics (D'Agostini and Degrassi 1999, Ciuchini *et al* 2001), in image reconstruction (Hanson 1993), in microscopy (Higdon and Yamamoto 2001), in quantum Monte Carlo (Gubernatis *et al* 1991), and in spectroscopy (Skilling 1992, Fischer *et al* 1997, 1998, 2000), just to mention a few articles written in the last decade. Other examples will be cited throughout the paper.

2 Uncertainty and probability

In the practice of science, we constantly find ourselves in a state of uncertainty. Uncertainty about the data that an experiment *shall* yield. Uncertainty about the *true value* of a physical quantity, even after an experiment has been done. Uncertainty about model parameters, calibration constants, and other quantities that might influence the outcome of the experiment, and hence influence our conclusions about the quantities of interest, or the models that might have produced the observed results.

In general, we know through experience that not all the *events* that could happen, or all conceivable *hypotheses*, are equally likely. Let us consider the outcome of you measuring the temperature at the location where you are presently reading this paper, assuming you use a digital thermometer with one degree resolution (or you round the reading at the degree if you have a more precise instrument). There are some values of the thermometer display you are more confident to read, others you expect less, and extremes you do not believe at all (some of them are simply excluded by the thermometer you are going to use). Given two events E_1 and E_2, for example E_1 : "$T = 22°$C" and E_2 : "$T = 33°$C", you might consider E_2 much more *probable* than E_1, just meaning that you *believe* E_2 to happen more than E_1. We could use different expressions to mean exactly the same thing: you consider E_2 more likely; you are more confident in E_2; having to choose between E_1 and E_2 to win a price, you would promptly choose E_2; having to classify with a number, that we shall denote with P, your degree of confidence on the two outcomes, you would write $P(E_2) > P(E_1)$; and many others.

On the other hand, we would rather state the opposite, i.e. $P(E_1) > P(E_2)$, with the same meaning of symbols and referring exactly to the same events: what you are going to read at your place with your thermometer. The reason is simply because we do not share the same status of information. We do not know who you are and where you are in this very moment. You and we are uncertain about the same event, but in a different way. Values that might appear very probable to you now, appear quite improbable, though not impossible, to us.

In this example we have introduced two crucial aspects of the Bayesian approach:

1. As it is used in everyday language, the term probability has the intuitive meaning of *"the degree of belief that an event will occur."*

2. Probability depends on our state of knowledge, which is usually different for different people. In other words, probability is unavoidably *subjective.*

At this point, you might find all of this quite natural, and wonder why these intuitive concepts go by the esoteric name 'Bayesian.' We agree! The fact is that the main thrust of statistics theory and practice during the 20$^{\text{th}}$ century has been based on a different concept of probability, in which it is defined as the limit of the long-term relative frequency of the outcome of these events. It revolves around the theoretical notion of infinite ensembles of 'identical experiments.' Without entering an unavoidably long critical discussion of the frequentist approach, we simply want to point out that in such a framework, there is no way to introduce the probability of hypotheses. All practical methods to overcome this deficiency yield misleading, and even absurd, conclusions. See (D'Agostini 1999c) for several examples and also for a justification of why frequentistic test 'often work'.

Instead, if we recover the intuitive concept of probability, we are able to talk in a natural way about the probability of any kind of event, or, extending the concept, of any *proposition.* In particular, the probability evaluation based on the relative frequency of similar events occurred in the past is easily recovered in the Bayesian theory, under precise condition of validity (see Sect. 5.3). Moreover, a simple theorem from probability theory, Bayes' theorem, which we shall see in the next section, allows us to update probabilities on the basis of new information. This inferential use of Bayes' theorem is only possible if probability is understood in terms of degree of belief. Therefore, the terms 'Bayesian' and 'based on subjective probability' are practically synonyms,and usually mean 'in contrast to the frequentist, or conventional, statistics.' The terms 'Bayesian' and 'subjective' should be considered transitional. In fact, there is already the tendency among many Bayesians to simply refer to 'probabilistic methods,' and so on (Jeffreys 1961, de Finetti 1974, Jaynes 1998 and Cowell *et al* 1999).

As mentioned above, Bayes' theorem plays a fundamental role in the probability theory. This means that subjective probabilities of logically connected events are related to each other by mathematical rules. This important result can be summed up by saying, in practical terms, that *'degrees of belief follow the same grammar as abstract axiomatic probabilities.'* Hence, all formal properties and theorems from probability theory follow.

Within the Bayesian school, there is no single way to derive the basic rules of probability (note that they are not simply taken as axioms in this approach). de Finetti's principle of *coherence* (de Finetti 1974) is considered the best guidance by many leading Bayesians (Bernardo and Smith 1994, O'Hagan 1994, Lad 1996 and Coletti and Scozzafava 2002). See (D'Agostini 1999c) for an informal introduction to the concept of coherence, which in simple words can be outlined as follows. A person who evaluates probability values should be ready to accepts bets in either direction, with odd ratios calculated from those values of probability. For example, an analyst that declares to be confident 50% on E should be aware that somebody could ask him to make a 1:1 bet on E or on $\overline{E}$. If he/she feels uneasy, it means that he/she does not consider the two events equally likely and the 50% was 'incoherent.'

Others, in particular practitioners close to the Jaynes' Maximum Entropy school (Jaynes 1957a, 1957b) feel more at ease with Cox's logical consistency reasoning, requiring some consistency properties ('desiderata') between values of probability related to logically connected propositions. (Cox 1946). See also (Jaynes 1998, Sivia 1997, and Fröhner 2000, and especially Tribus 1969), for accurate derivations and a clear account of the meaning and role of information entropy in data analysis. An approach similar to Cox's is followed by Jeffreys (1961), another leading figure who has contributed a new vitality to the methods based on this 'new' point of view on probability. Note that Cox and Jeffreys were physicists. Remarkably, Schrödinger (1947a, 1947b) also arrived at similar conclusions, though his definition of event is closer to the de Finetti's one. [Some short quotations from (Schrödinger 1947a) are in order. Definition of probability: *"... a quantitative measure of the strength of our conjecture or anticipation, founded on the said knowledge, that the event comes true"*. Subjective nature of probability: *"Since the knowledge may be different with different persons or with the same person at different times, they may anticipate the same event with more or less confidence, and thus different numerical probabilities may be attached to the same event."* Conditional probability: *"Thus whenever we speak loosely of 'the probability of an event,' it is always to be understood: probability with regard to a certain given state of knowledge."*]

3 Rules of probability

We begin by stating some basic rules and general properties that form the 'grammar' of the probabilistic language, which is used in Bayesian analysis. In this section, we review the rules of probability, starting with the rules for simple propositions. We will not provide rigorous derivations and will not address the foundational or philosophical aspects of probability theory. Moreover, following an 'eclectic' approach which is common among Bayesian practitioners, we talk indifferently about probability of events, probability of hypotheses or probability of propositions. Indeed, the last expression will be often favoured, understanding that it does include the others.

3.1 Probability of simple propositions

Let us start by recalling the basic rules of probability for propositions or hypotheses. Let A and B be propositions, which can take on only two values, for example, true or false. The notation $P(A)$ stands for the probability that A is true. The elementary rules of probability for simple propositions are

$$0 \leq P(A) \leq 1; \tag{1}$$

$$
\begin{aligned}
&P(\Omega) = 1; &(2)\\
&P(A \cup B) = P(A) + P(B) - P(A \cap B)\,. &(3)\\
&P(A \cap B) = P(A \,|\, B)\, P(B) = P(B \,|\, A)\, P(A)\,, &(4)
\end{aligned}
$$

where Ω means *tautology* (a proposition that is certainly true). The construct $A \cap B$ is true only when both A and B are true (logical AND), while $A \cup B$ is true when at least one of the two propositions is true (logical OR). $A \cap B$ is also written simply as 'A, B' or AB, and is also called a *logical product*, while $A \cup B$ is also called a *logical sum*. $P(A, B)$ is called the joint probability of A and B. $P(A \,|\, B)$ is the probability of A under that condition that B is true. We often read it simply as "the probability of A, given B." .

Equation (4) shows that the joint probability of two events can be decomposed into conditional probabilities in different two ways. Either of these ways is called the *product rule*. If the status of B does not change the probability of A, and the other way around, then A and B are said to be *independent*, *probabilistically* independent to be precise. In that case, $P(A \,|\, B) = P(A)$, and $P(B \,|\, A) = P(B)$, which, when inserted in Eq. (4), yields

$$
P(A \cap B) \;=\; P(A)\, P(B) \qquad \Longleftrightarrow \qquad \textit{probabilistic independence}. \tag{5}
$$

Equations (1)–(4) logically lead to other rules which form the body of probability theory. For example, indicating the *negation* (or *opposite*) of A with $\overline{A}$, clearly $A \cup \overline{A}$ is a tautology ($A \cup \overline{A} = \Omega$), and $A \cap \overline{A}$ is a contradiction ($A \cap \overline{A} = \emptyset$). The symbol $\emptyset$ stands for contradiction (a proposition that is certainly false). Hence, we obtain from Eqs. (2) and (3)

$$
P(A) + P(\overline{A}) \;=\; 1\,, \tag{6}
$$

which says that proposition A is either true or not true.

3.2 Probability of complete classes

These formulae become more interesting when we consider a set of propositions H_j that all together form a tautology (i.e., they are *exhaustive*) and are mutually *exclusive*. Formally

$$
\begin{aligned}
\cup_i \, H_j &\;=\; \Omega &(7)\\
H_j \cap H_k &\;=\; \emptyset \quad \text{if} \quad j \neq k\,. &(8)
\end{aligned}
$$

When these conditions apply, the set $\{H_j\}$ is said to form a *complete class*. The symbol H has been chosen because we shall soon interpret $\{H_j\}$ as a set of *hypotheses*.

The first (trivial) property of a complete class is *normalization*, that is

$$
\sum_j P(H_j) \;=\; 1\,, \tag{9}
$$

which is just an extension of Eq. (6) to a complete class containing more than just a single proposition and its negation.

For the complete class H, the generalizations of Eqs. (6) and the use of Eq. (4) yield:

$$
\begin{aligned}
P(A) &\;=\; \sum_j P(A, H_j) &(10)\\
P(A) &\;=\; \sum_j P(A \,|\, H_j)\, P(H_j)\,. &(11)
\end{aligned}
$$

Equation (10) is the basis of what is called *marginalization*, which will become particularly important when dealing with uncertain variables: the probability of A is obtained by the summation over all possible *constituents* contained in A. Hereafter, we avoid explicitly writing the limits of the summations, meaning that they extend over all elements of the class. The constituents are 'A, H_j,' which, based on the complete class of hypotheses $\{H\}$, themselves form a complete class, which can be easily proved. Equation (11) shows that the probability of any proposition is given by a weighted average of all conditional probabilities, subject to hypotheses H_j forming a complete class, with the weight being the probability of the hypothesis.

In general, there are many ways to choose complete classes (like 'bases' in geometrical spaces). Let us denote the elements of a second complete class by E_i. The constituents are then formed by the elements (E_i, H_j) of the Cartesian product $\{E\} \times \{H\}$. Equations (10) and (11) then become the more general statements

$$P(E_i) = \sum_j P(E_i, H_j) \tag{12}$$

$$P(E_i) = \sum_j P(E_i \mid H_j)\, P(H_j) \tag{13}$$

and, symmetrically,

$$P(H_j) = \sum_i P(E_i, H_j) \tag{14}$$

$$P(H_j) = \sum_i P(H_j \mid E_i)\, P(E_i)\,. \tag{15}$$

The reason we write these formulae both ways is to stress the symmetry of Bayesian reasoning with respect to classes $\{E\}$ and $\{H\}$, though we shall soon associate them with *observations* (or *events*) and *hypotheses*, respectively.

3.3 Probability rules for uncertain variables

In analyzing the data from physics experiments, we need to deal with measurement that are discrete or continuous in nature. Our aim is to make inferences about the models that we believe appropriately describe the physical situation, and/or, within a given model, to determine the values of relevant physics quantities. Thus, we need the probability rules that apply to uncertain variables, whether they are discrete or continuous. The rules for complete classes described in the preceding section clearly apply directly to discrete variables. With only slight changes, the same rules also apply to continuous variables because they may be thought of as a limit of discrete variables, as interval between possible discrete values goes to zero.

For a discrete variable x, the expression $p(x)$, which is called a *probability function*, has the interpretation in terms of the probability of the proposition $P(A)$, where A is true when the value of the variable is equal to x. In the case of continuous variables, we use the same notation, but with the meaning of a *probability density function* (pdf). So $p(x)\,dx$, in terms of a proposition, is the probability $P(A)$, where A is true when the value of the variable lies in the range of x to $x + dx$. In general, the meaning is clear from the context; otherwise it should be stated. Probabilities involving more than one variable, like $p(x, y)$, have the meaning of the probability of a logical product; they are usually called *joint* probabilities.

Table 1 summarizes useful formulae for discrete and continuous variables. The interpretation and use of these relations in Bayesian inference will be illustrated in the following sections.

4 Bayesian inference for simple problems

We introduce the basic concepts of Bayesian inference by considering some simple problems. The aim is to illustrate some of the notions that form the foundation of Bayesian reasoning.

Table 1: *Some definitions and properties of probability functions for values of a discrete variable* x_i *and probability density functions for continuous variables* x. *All summations and integrals are understood to extend over the full range of possibilities of the variable. Note that the expectation of the variable is also called* expected value *(sometimes* expectation value*),* average *and* mean. *The square root of the variance is the* standard deviation σ.

	discrete variables	continuous variables
probability	$P(X = x_i) = p(x_i)$	$\mathrm{d}P_{[x \le X \le x+\mathrm{d}x]} = p(x)\,\mathrm{d}x$
normalization†	$\sum_i p(x_i) = 1$	$\int p(x)\,\mathrm{d}x = 1$
expectation of $f(X)$	$\mathrm{E}[f(X)] = \sum_i f(x_i)\,p(x_i)$	$\mathrm{E}[f(X)] = \int f(x)\,p(x)\,\mathrm{d}x$
expected value	$\mathrm{E}(X) = \sum_i x_i\,p(x_i)$	$\mathrm{E}(X) = \int x\,p(x)\,\mathrm{d}x$
moment of order r	$\mathrm{M}_r(X) = \sum_i x_i^r\,p(x_i)$	$\mathrm{M}_r(X) = \int x^r\,p(x)\,\mathrm{d}x$
variance	$\sigma^2 = \sum_i [x_i - \mathrm{E}(X)]^2\,p(x_i)$	$\sigma^2 = \int [x - \mathrm{E}(X)]^2\,p(x)\,\mathrm{d}x$
product rule	$p(x_i, y_j) = p(x_i \mid y_j)\,p(y_j)$	$p(x, y) = p(x \mid y)\,p(y)$
independence	$p(x_i, y_j) = p(x_i)\,p(y_j)$	$p(x, y) = p(x)\,p(y)$
marginalization	$\sum_j p(x_i, y_j) = p(x_i)$	$\int p(x, y)\,\mathrm{d}y = p(x)$
decomposition	$p(x_i) = \sum_j p(x_i \mid y_j)\,p(y_j)$	$p(x) = \int p(x \mid y)\,p(y)\,\mathrm{d}y$
Bayes' theorem	$p(x_j \mid y_i) = \frac{p(y_i \mid x_j)\,p(x_j)}{\sum_j p(y_i \mid x_j)\,p(x_j)}$	$p(x \mid y) = \frac{p(y \mid x)\,p(x)}{\int p(y \mid x)\,p(x)\,\mathrm{d}x}$
likelihood	$\mathcal{L}(x_j\,;\,y_i) = p(y_i \mid x_j)$	$\mathcal{L}(x\,;\,y) = p(y \mid x)$

†A function $p(x)$ such that $\sum_i p(x_i) = \infty$, or $\int p(x)\,\mathrm{d}x = \infty$, is called *improper*. Improper functions are often used to describe *relative beliefs* about the possible values of a variable.

4.1 Background information

As we think about drawing conclusions about the physical world, we come to realize that everything we do is based on what we know about the world. Conclusions about hypotheses will be based on our general background knowledge. To emphasize the dependence of probability on the state of background information, which we designate as I, we will make it explicit by writing $P(E \mid I)$, rather than simply $P(E)$. (Note that, in general, $P(A \mid I_1) \neq P(A \mid I_2)$, if I_1 and I_2 are different states of information.) For example, Eq. (4) should be more precisely written as

$$P(A \cap B \mid I) = P(A \mid B \cap I)\, P(B \mid I) = P(B \mid A \cap I)\, P(A \mid I)\,, \tag{16}$$

or alternatively as

$$P(A, B \mid I) = P(A \mid B, I)\, P(B \mid I) = P(B \mid A, I)\, P(A \mid I)\,. \tag{17}$$

We have explicitly included I as part of the conditional to remember that any probability relation is valid only under the same state of background information.

4.2 Bayes' theorem

Formally, Bayes' theorem follows from the symmetry of $P(A, B)$ expressed by Eq. (17). In terms of E_i and H_j belonging to two different complete classes, Eq. (17) yields

$$\frac{P(H_j \mid E_i, I)}{P(H_j \mid I)} = \frac{P(E_i \mid H_j, I)}{P(E_i \mid I)} \tag{18}$$

This equation says that the *new* condition E_i alters our belief in H_j by the same updating factor by which the condition H_j alters our belief about E_i. Rearrangement yields *Bayes' theorem*

$$P(H_j \mid E_i, I) = \frac{P(E_i \mid H_j, I)\, P(H_j \mid I)}{P(E_i \mid I)}\,. \tag{19}$$

We have obtained a logical rule to update our beliefs on the basis of new conditions. Note that, though Bayes' theorem is a direct consequence of the basic rules of axiomatic probability theory, its updating power can only be fully exploited if we can treat on the same basis expressions concerning hypotheses and observations, causes and effects, models and data.

In most practical cases, the evaluation of $P(E_i \mid I)$ can be quite difficult, while determining the conditional probability $P(E_i \mid H_j, I)$ might be easier. For example, think of E_i as the probability of observing a particular event topology in a particle physics experiment, compared with the probability of the same thing *given* a value of the hypothesized particle mass (H_j), a given detector, background conditions, etc. Therefore, it is convenient to rewrite $P(E_i \mid I)$ in Eq. (19) in terms of the quantities in the numerator, using Eq. (13), to obtain

$$P(H_j \mid E_i, I) \quad = \quad \frac{P(E_i \mid H_j, I)\, P(H_j \mid I)}{\sum_j P(E_i \mid H_j, I)\, P(H_j \mid I)}\,, \tag{20}$$

which is the better-known form of Bayes' theorem. Written this way, it becomes evident that the denominator of the r.h.s. of Eq. (20) is just a normalization factor and we can focus on just the numerator:

$$P(H_j \mid E_i, I) \quad \propto \quad P(E_i \mid H_j, I)\, P(H_j \mid I)\,. \tag{21}$$

In words

$$\text{posterior} \quad \propto \quad \text{likelihood} \times \text{prior}\,, \tag{22}$$

where the *posterior* (or *final* state) stands for the probability of H_j, based on the new observation E_i, relative to the *prior* (or *initial*) probability. (Prior probabilities are often indicated with P_0.) The conditional probability $P(E_i \,|\, H_j)$ is called the *likelihood*. It is literally the probability of the observation E_i given the specific hypothesis H_j. The term likelihood can lead to some confusion, because it is often misunderstood to mean "the likelihood that E_i comes from H_j." However, this name implies to consider $P(E_i \,|\, H_j)$ a mathematical function of H_j for a fixed E_i and in that framework it is usually written as $\mathcal{L}(H_j; E_i)$ to emphasize the functionality. We caution the reader that one sometimes even finds the notation $\mathcal{L}(E_i \,|\, H_j)$ to indicate exactly $P(E_i \,|\, H_j)$.

4.3 Inference for simple hypotheses

Making use of formulae (20) or (21), we can easily solve many classical problems involving inference when many hypotheses can produce the same single effect. Consider the case of interpreting the results of a test for the HIV virus applied to a *randomly chosen* European. Clinical tests are very seldom perfect. Suppose that the test accurately detects infection, but has a false-positive rate of 0.2%:

$$P(\text{Positive} \,|\, \text{Infected}) = 1\,, \quad \text{and} \quad P(\text{Positive} \,|\, \overline{\text{Infected}}) = 0.2\%\,.$$

If the test is positive, can we conclude that the particular person is infected with a probability of 99.8% because the test has only a 0.2% chance of mistake? Certainly not! This kind of mistake is often made by those who are not used to Bayesian reasoning, including scientists who make inferences in their own field of expertise. The correct answer depends on what we else know about the person tested, that is, the background information. Thus, we have to consider the incidence of the HIV virus in Europe, and possibly, information about the lifestyle of the individual. For details, see (D'Agostini 1999c).

To better understand the updating mechanism, let us take the ratio of Eq. (20) for two hypotheses H_j and H_k

$$\frac{P(H_j \,|\, E_i, I)}{P(H_k \,|\, E_i, I)} \;=\; \frac{P(E_i \,|\, H_j, I)}{P(E_i \,|\, H_k, I)}\,\frac{P(H_j \,|\, I)}{P(H_k \,|\, I)}\,, \tag{23}$$

where the sums in the denominators of Eq. (20) cancel. It is convenient to interpret the ratio of probabilities, given the same condition, as *betting odds*. This is best done formally in the de Finetti approach, but the basic idea is what everyone is used to: the amount of money that one is willing to bet on an event is proportional to the degree to which one expects that event will happen. Equation (23) tells us that, when new information is available, the initial odds are updated by the ratio of the likelihoods $P(E_i \,|\, H_j, I)/P(E_i \,|\, H_k, I)$, which is known as the *Bayes factor*.

In the case of the HIV test, the initial odds for an arbitrarily chosen European to be infected $P(H_j \,|\, I)/P(H_k \,|\, I)$ are so small that we need a very high Bayes' factor to be reasonably certain that, when the test is positive, the person is really infected. With the numbers used in this example, the Bayes factor is $500 = 1/0.002$. For example, if we take for the prior $P_0(\text{Infected})/P_0(\overline{\text{Infected}}) = 1/1000$, the Bayes' factor changes these odds to $500/1000 = 1/2$, or equivalently, the probability that the person is infected would be 1/3, quite different from the 99.8% answer usually prompted by those who have a standard statistical education. This example can be translated straightforwardly to physical problems, like particle identification in the analysis of a Cherenkov detector data, as done, e.g. in (D'Agostini 1999c).

5 Inferring numerical values of physics quantities — General ideas and basic examples

In physics we are concerned about models ('theories') and the numerical values of physical quantities related to them. Models and the value of quantities are, generally speaking, the hypothesis we want to infer, given the observations. In the previous section we have learned how to deal with simple hypotheses, 'simple' in the sense that they do not depend on internal parameters.

On the other hand, in many applications we have strong beliefs about what model to use to interpret the measurements. Thus, we focus our attention on the model parameters, which we consider as uncertain variables that we want to infer. The method which deals with these applications is usually referred as *parametric inference*, and it will be shown with examples in this section. In our models, the value of the relevant physical quantities are usually described in terms of a continuous uncertain variable. Bayes' theorem, properly extended to uncertain quantities (see Tab.1), plays a central role in this inference process.

A more complicate case is when we are also uncertain about the model (and each possible model has its own set of parameter, usually associated with different physics quantities). We shall analyse this problem in Sect. 7.

5.1 Bayesian inference on uncertain variables and posterior characterization

We start here with a few one-dimensional problems involving simple models that often occur in data analysis. These examples will be used to illustrate some of the most important Bayesian concepts. Let us first introduce briefly the structure of the Bayes' theorem in the form convenient to our purpose, as a straightforward extension of what was seen in Sect. 4.2.

$$p(\theta \,|\, d, I) \;=\; \frac{p(d \,|\, \theta, I)\, p(\theta \,|\, I)}{\int p(d \,|\, \theta, I)\, p(\theta \,|\, I)\, \mathrm{d}\theta}\,. \tag{24}$$

θ is the generic name of the parameter (used hereafter, unless the models have traditional symbols for their parameters) and d is the data point. $p(\theta \,|\, I)$ is the prior, $p(\theta \,|\, d, I)$ the posterior and $p(d \,|\, \theta, I)$ the likelihood. Also in this case the likelihood is is often written as $\mathcal{L}(\theta; d) = p(d \,|\, \theta, I)$, and the same words of caution expressed in Sect. 4.2 apply here too. Note, moreover, that, while $p(d \,|\, \theta, I)$ is a properly normalized pdf, $\mathcal{L}(\theta; d)$ has not a pdf meaning in the variable θ. Hence, the integral of $\mathcal{L}(\theta; d)$ over θ is only accidentally equal to unity. The denominator in the r.h.s. of Eq. (24) is called the *evidence* and, while in the parametric inference discussed here is just a trivial normalization factor, its value becomes important for model comparison (see Sect. 7).

Posterior probability distributions provide the full description of our state of knowledge about the value of the quantity. In fact, they allow to calculate all *probability intervals* of interest. Such intervals are also called *credible intervals* (at a specified level of probability, for example 95%) or *confidence intervals* (at a specified level of 'confidence', i.e. of probability). However, the latter expression could be confused with frequentistic 'confidence intervals', that are not probabilistic statements about uncertain variables (D'Agostini 1999c).

It is often desirable to characterize the distribution in terms of a few numbers. For example, mean value (arithmetic *average*) of the posterior, or its most probable value (the *mode*) of the posterior, also known as the *maximum a posteriori (MAP) estimate*. The spread of the distribution is often described in terms of its *standard deviation* (square root of the *variance*). It is useful to associate the terms mean value and standard deviation with the more inferential terms *expected value*, or simply *expectation* (value), indicated by E( ), and *standard uncertainty* (ISO 1993), indicated by $\sigma(\,)$, where the argument is the uncertain variable of interest. This will be our standard way of reporting the result of inference in a quantitative way, though, we emphasize that the full answer is given by the posterior distribution, and reporting only these summaries in case of the complex distributions (e.g. multimodal and/or asymmetrical pdf's) can be misleading, because people tend to think of a Gaussian model if no further information is provided.

5.2 Gaussian model

Let us start with a classical example in which the response signal d from a detector is described by a Gaussian error function around the *true value* μ with a standard deviation σ, which is assumed to be exactly known. This model is the best-known among physicists and, indeed, the Gaussian pdf is also known as *normal* because it is often assumed that errors are 'normally' distributed according to this function. Applying Bayes' theorem for continuous variables (see Tab. 1), from the likelihood

$$p(d\,|\,\mu, I) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(d-\mu)^2}{2\,\sigma^2}\right] \tag{25}$$

we get for μ

$$p(\mu\,|\,d, I) \;=\; \frac{\dfrac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\dfrac{(d-\mu)^2}{2\,\sigma^2}\right]\,p(\mu\,|\,I)}{\displaystyle\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(d-\mu)^2}{2\,\sigma^2}\right]\,p(\mu\,|\,I)\,\mathrm{d}\mu}\,. \tag{26}$$

Considering all values of μ equally likely over a very large interval, we can model the prior $p(\mu\,|\,I)$ with a constant, which simplifies in Eq. (26), yielding

$$p(\mu\,|\,d, I) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left[-\frac{(\mu-d)^2}{2\,\sigma^2}\right]. \tag{27}$$

Expectation and standard deviation of the posterior distribution are $\mathrm{E}(\mu) = d$ and $\sigma(\mu) = \sigma$, respectively. This particular result corresponds to what is often done intuitively in practice. But one has to pay attention to the assumed conditions under which the result is logically valid: Gaussian likelihood and uniform prior. Moreover, we can speak about the probability of true values only in the subjective sense. It is recognized that physicists, and scientists in general, are highly confused about this point (D'Agostini 1999a).

A noteworthy case of a prior for which the naive inversion gives paradoxical results is when the value of a quantity is constrained to be in the 'physical region,' for example $\mu \geq 0$, while d falls outside it (or it is at its edge). The simplest prior that cures the problem is a step function $\theta(\mu)$, and the result is equivalent to simply renormalizing the pdf in the physical region (this result corresponds to a 'prescription' sometimes used by practitioners with a frequentist background when they encounter this kind of problem).

Another interesting case is when the prior knowledge can be modeled with a Gaussian function, for example, describing our knowledge from a previous inference

$$p(\mu\,|\,\mu_0, \sigma_0, I) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_0}\exp\left[-\frac{(\mu-\mu_0)^2}{2\,\sigma_0^2}\right]. \tag{28}$$

Inserting Eq. (28) into Eq. (26), we get

$$p(\mu\,|\,d, \mu_0, \sigma_0, I) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma_1}\exp\left[-\frac{(\mu-\mu_1)^2}{2\,\sigma_1^2}\right], \tag{29}$$

where

$$\mu_1 = \mathrm{E}(\mu) = \frac{d/\sigma^2 + \mu_0/\sigma_0^2}{1/\sigma^2 + 1/\sigma_0^2} \tag{30}$$

$$= \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}\, d + \frac{\sigma^2}{\sigma^2+\sigma_0^2}\,\mu_0 = \frac{\sigma_1^2}{\sigma^2}\, d + \frac{\sigma_1^2}{\sigma_0^2}\,\mu_0 \tag{31}$$

$$\sigma_1^2 = \mathrm{Var}(\mu) = \left(\sigma_0^{-2} + \sigma^{-2}\right)^{-1} \tag{32}$$

We can then see that the case $p(\mu \mid I) =$ constant corresponds to the limit of a Gaussian prior with very large σ_0 and finite μ_0. The formula for the expected value combining previous knowledge and present experimental information has been written in several ways in Eq. (31).

Another enlighting way of writing Eq. (30) is considering μ_0 and μ_1 the estimates of μ at times t_0 and t_1, respectively before and after the observation d happened at time t_1. Indicating the *estimates* at different times by $\hat{\mu}(t)$, we can rewrite Eq. (30) as

$$\hat{\mu}(t_1) = \frac{\sigma_\mu^2(t_0)}{\sigma_d^2(t_1)+\sigma_\mu^2(t_0)}\, d(t_1) + \frac{\sigma_d^2(t_1)}{\sigma_d^2(t_1)+\sigma_\mu^2(t_0)}\,\hat{\mu}(t_0)$$

$$= \hat{\mu}(t_0) + \frac{\sigma_\mu^2(t_0)}{\sigma_d^2(t_1)+\sigma_\mu^2(t_0)}\,[d(t_1) - \hat{\mu}(t_0)] \tag{33}$$

$$= \hat{\mu}(t_0) + K(t_1)\,[d(t_1) - \hat{\mu}(t_0)] \tag{34}$$

$$\sigma_\mu^2(t_1) = \sigma_\mu^2(t_0) - K(t_1)\,\sigma_\mu^2(t_0)\,, \tag{35}$$

where

$$K(t_1) = \frac{\sigma_\mu^2(t_0)}{\sigma_d^2(t_1)+\sigma_\mu^2(t_0)}\,. \tag{36}$$

Indeed, we have given Eq. (30) the structure of a *Kalman filter* (Kalman 1960). The new observation 'corrects' the estimate by a quantity given by the *innovation* (or *residual*) $[d(t_1) - \hat{\mu}(t_0)]$ times the *blending factor* (or *gain*) $K(t_1)$. For an introduction about Kalman filter and its probabilistic origin, see (Maybeck 1979 and Welch and Bishop 2002).

As Eqs. (31)–(35) show, a new experimental information reduces the uncertainty. But this is true as long the previous information and the observation are somewhat consistent. If we are, for several reasons, sceptical about the model which yields the combination rule (31)–(32), we need to remodel the problem and introduce possible systematic errors or underestimations of the quoted standard deviations, as done e.g. in (Press 1997, Dose and von der Linden 1999, D'Agostini 1999b, Fröhner 2000).

5.3 Binomial model

In a large class of experiments, the observations consist of counts, that is, a number of things (events, occurrences, etc.). In many processes of physics interests the resulting number of counts is described probabilistically by a binomial or a Poisson model. For example, we want to draw an inference about the efficiency of a detector, a branching ratio in a particle decay or a rate from a measured number of counts in a given interval of time.

The binomial distribution describes the probability of randomly obtaining n events ('successes') in N independent trials, in each of which we assume the same probability θ that the event will happen. The probability function is

$$p(n \mid \theta, N) = \binom{N}{n}\,\theta^n (1-\theta)^{N-n}\,, \tag{37}$$

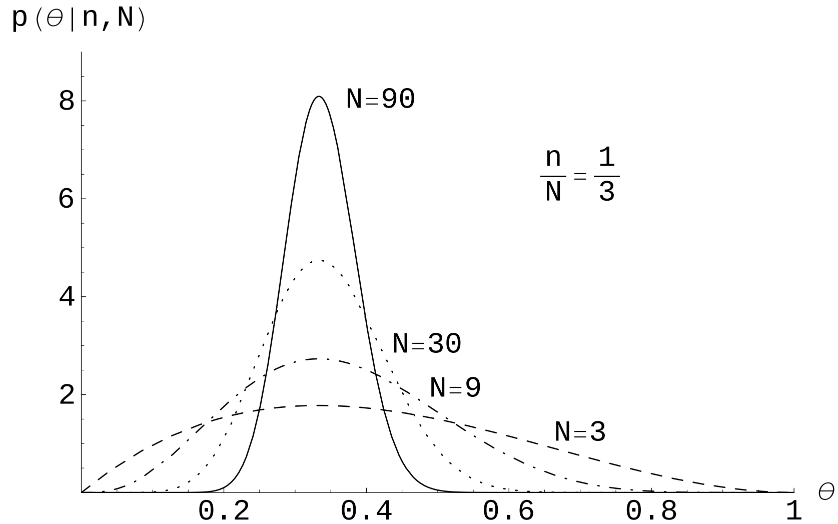

Figure 1: Posterior probability density function of the binomial parameter θ, having observed n successes in N trials.

where the leading factor is the well-known binomial coefficient, namely $N!/n!(N-n)!$. We wish to infer θ from an observed number of counts n in N trials. Incidentally, that was the "problem in the doctrine of chances" originally treated by Bayes (1763), reproduced e.g. in (Press 1992). Assuming a uniform prior for θ, by Bayes' theorem the posterior distribution for θ is proportional to the likelihood, given by Eq. (37):

$$
\begin{aligned}
p(\theta\,|\,n,N,I) &= \frac{\theta^n\,(1-\theta)^{N-n}}{\int_0^1 \theta^n\,(1-\theta)^{N-n}\,\mathrm{d}\theta} && (38)\\
&= \frac{(N+1)!}{n!\,(N-n)!}\,\theta^n\,(1-\theta)^{N-n}\,. && (39)
\end{aligned}
$$

Some examples of this distribution for various values of n and N are shown in Fig. 1. Expectation, variance, and mode of this distribution are:

$$
\begin{aligned}
\mathrm{E}(\theta) &= \frac{n+1}{N+2} && (40)\\
\sigma^2(\theta) &= \frac{(n+1)(N-n+1)}{(N+3)(N+2)^2} = \frac{\mathrm{E}(\theta)\ (1-\mathrm{E}(\theta))}{N+3} && (41)\\
\theta_{\mathrm{m}} &= \frac{n}{N}\,, && (42)
\end{aligned}
$$

where the mode has been indicated with θ_{m}. Equation (40) is known as the Laplace formula. For large values of N and $0 \ll n \ll N$ the expectation of θ tends to θ_{m}, and $p(\theta)$ becomes approximately Gaussian. This result is nothing but a reflection of the well-known asymptotic Gaussian behavior of $p(n\,|\,\theta, N)$. For large N the uncertainty about θ goes like $1/\sqrt{N}$. Asymptotically, we are practically certain that θ is equal to the relative frequency of that class of events observed in the past. This is how the frequency based evaluation of probability is promptly recovered in the Bayesian approach, under well defined assumptions.

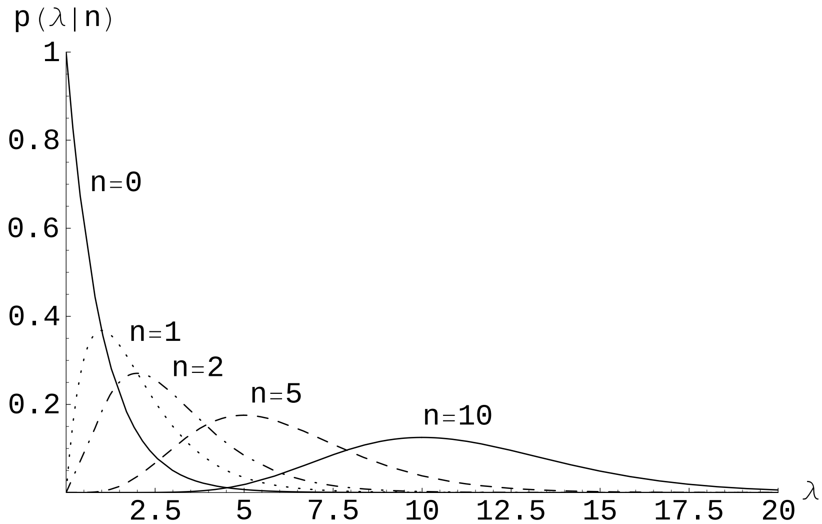

Figure 2: The posterior distribution for the Poisson parameter λ, when n counts are observed in an experiment.

5.4 Poisson model

The Poisson distribution gives the probability of observing n counts in a fixed time interval, when the expectation of the number of counts to be observed is λ:

$$p(n\,|\,\lambda) \;=\; \frac{\lambda^n e^{-\lambda}}{n!}\,. \tag{43}$$

The inverse problem is to infer λ from n counts observed. (Note that what physically matters is the rate $r = \lambda/\Delta T$, where ΔT is the observation time.) Applying Bayes' theorem and using a uniform prior $p(\lambda\,|\,I)$ for λ, we get

$$p(\lambda\,|\,n, I) = \frac{\dfrac{\lambda^n\,e^{-\lambda}}{n!}}{\displaystyle\int_0^\infty \frac{\lambda^n\,e^{-\lambda}}{n!}\,\mathrm{d}\lambda} = \frac{\lambda^n\,e^{-\lambda}}{n!}\,. \tag{44}$$

As for the Gaussian model, the same mathematical expression holds for the likelihood, but with interchanged role of variable and parameter. Expectation and variance of λ are both equal to $n+1$, while the most probable value is $\lambda_m = n$. For large n, the extra '+1' (due to the asymmetry of the prior with respect to $\lambda = 0$) can be ignored and we have $\mathrm{E}(\lambda) = \sigma^2(\lambda) \approx n$ and, once again, the uncertainty about λ follows a Gaussian model. The relative uncertainty on λ decreases as $1/\sqrt{n}$.

When the observed value of n is zero, Eq. (44) yields $p(\lambda\,|\,n=0) = e^{-\lambda}$, giving a maximum of belief at zero, but an exponential tail toward large values of λ. Expected value and standard deviation of λ are both equal to 1. The 95% probabilistic upper bound of λ is at $\lambda_{95\%UB} = 3$, as it can be easily calculated solving the equation $\int_0^{\lambda_{95\%UB}} p(\lambda\,|\,n=0)\,\mathrm{d}\lambda = 0.95$. Note that also this result depends on the choice of prior, though Astone and D'Agostini (1999) have shown that the upper bound is insensitive to the exact form of the prior, if the prior models somehow what they call "positive attitude of rational scientists" (the prior has not to be in contradiction with what one could actually observe, given the detector sensitivity). In particular, they show that a uniform prior is a good practical choice to model this attitude. On the other hand, talking about 'objective' probabilistic upper/lower limits makes no sense, as discussed in detail and with examples in the cited paper: one can at most speak about conventionally defined non-probabilistic *sensitivity bounds*, which separate the measurement region from that in which experimental sensitivity is lost (Astone and D'Agostini 1999, D'Agostini 2000, Astone *et al* 2002).

5.5 Inference from a data set and sequential use of Bayes formula

In the elementary examples shown above, the inference has been done from a single *data point* d. If we have a set of observations (data), indicated by $\boldsymbol{d}$, we just need to insert in the Bayes formula the likelihood $p(\boldsymbol{d}\,|\,\theta, I)$, where this expression indicates a multi-dimensional joint pdf.

Note that we could think of inferring θ on the basis of each newly observed datum d_i. After the one observation:

$$p(\theta\,|\,d_1, I) \propto p(d_1\,|\,\theta, I)\, p(\theta\,|\,I) \tag{45}$$

and after the second:

$$\begin{aligned} p(\theta\,|\,d_1, d_2, I) &\propto p(d_2\,|\,\theta, d_1, I)\, p(\theta\,|\,d_1, I) && (46)\\ &\propto p(d_2\,|\,\theta, d_1, I)\, p(d_1\,|\,\theta, I)\, p(\theta\,|\,I)\,. && (47) \end{aligned}$$

We have written Eq. (47) in a way that the dependence between observables can be accommodated. From the product rule in Tab. 1, we can rewrite Eq. (47) as

$$p(\theta\,|\,d_1, d_2, I) \;\propto\; p(d_1, d_2\,|\,\theta, I)\, p(\theta\,|\,I)\,. \tag{48}$$

Comparing this equation with (47) we see that the sequential inference gives exactly the same result of a single inference that properly takes into account all available information. This is an important result of the Bayesian approach.

The extension to many variables is straightforward, obtaining

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{d}, I) \;\propto\; p(\boldsymbol{d}\,|\,\boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}\,|\,I)\,. \tag{49}$$

Furthermore, when the d_i are independent, we get for the likelihood

$$\begin{aligned} p(\boldsymbol{d}\,|\,\boldsymbol{\theta}, I) &= \prod_i p(d_i\,|\,\boldsymbol{\theta}, I) && (50)\\ \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{d}) &= \prod_i \mathcal{L}(\boldsymbol{\theta}; d_i)\,, && (51) \end{aligned}$$

that is, the combined likelihood is given by the product of the individual likelihoods.

5.6 Multidimensional case — Inferring μ and σ of a Gaussian

So far we have only inferred one parameter of a model. The extension to many parameters is straightforward. Calling $\boldsymbol{\theta}$ the set of parameters and $\boldsymbol{d}$ the data, Bayes' theorem becomes

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{d}, I) \;=\; \frac{p(\boldsymbol{d}\,|\,\boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}\,|\,I)}{\int p(\boldsymbol{d}\,|\,\boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}\,|\,I)\, \mathrm{d}\boldsymbol{\theta}}\,. \tag{52}$$

Equation (52) gives the posterior for the full parameter vector $\boldsymbol{\theta}$. *Marginalization* (see Tab. 1) allows one to calculate the probability distribution for a single parameter, for example, $p(\theta_i\,|\,\boldsymbol{d}, I)$, by integrating over the remaining parameters. The marginal distribution $p(\theta_i\,|\,\boldsymbol{d}, I)$ is then the complete result of the Bayesian inference on the parameter θ_i. Though the characterization of the marginal is done in the usual way described in Sect. 5.1, there is often the interest to summarize some characters of the multi-dimensional posterior that are unavoidably lost in the marginalization (imagine marginalization as a kind of geometrical projection). Useful quantities are the covariances between parameters θ_i and θ_j, defined as

$$\mathrm{Cov}(\theta_i, \theta_j) \;=\; \mathrm{E}[(\theta_i - \mathrm{E}[\theta_i])\,(\theta_j - \mathrm{E}[\theta_j])]\,. \tag{53}$$

As is well know, quantities which give a more intuitive idea of what is going on are the correlation coefficients, defined as $\rho(\theta_i, \theta_j) = \mathrm{Cov}(\theta_i, \theta_j)/\sigma(\theta_i)\sigma(\theta_j)$. Variances and covariances form the covariance matrix $\mathbf{V}(\boldsymbol{\theta})$, with $V_{ii} = \mathrm{Var}(\theta_i)$ and $V_{ij} = \mathrm{Cov}(\theta_i, \theta_j)$. We recall also that convenient formulae to calculate variances and covariances are obtained from the expectation of the products $\theta_i\theta_j$, together with the expectations of the parameters:

$$V_{ij} \;=\; \mathrm{E}(\theta_i\theta_j) - \mathrm{E}(\theta_i)\,\mathrm{E}(\theta_j) \tag{54}$$

As a first example of a multidimensional distribution from a data set, we can think, again, at the inference of the parameter μ of a Gaussian distribution, but in the case that also σ is unknown and needs to be determined by the data. From Eqs. (52), (50) and (25), with $\theta_1 = \mu$ and $\theta_2 = \sigma$ and neglecting overall normalization, we obtain

$$p(\mu, \sigma \,|\, \boldsymbol{d}, I) \;\propto\; \sigma^{-n}\, \exp\left[-\frac{\sum_{i=1}^{n}(d_i - \mu)^2}{2\,\sigma^2}\right] p(\mu, \sigma \,|\, I) \tag{55}$$

$$p(\mu \,|\, \boldsymbol{d}, I) \;=\; \int p(\mu, \sigma \,|\, \boldsymbol{d}, I)\, \mathrm{d}\sigma \tag{56}$$

$$p(\sigma \,|\, \boldsymbol{d}, I) \;=\; \int p(\mu, \sigma \,|\, \boldsymbol{d}, I)\, \mathrm{d}\mu\,. \tag{57}$$

The closed form of Eqs. (56) and 57) depends on the prior and, perhaps, for the most realistic choice of $p(\mu, \sigma \,|\, I)$, such a compact solution does not exists. But this is not an essential issue, given the present computational power. (For example, the shape of $p(\mu, \sigma \,|\, I)$ can be easily inspected by a modern graphical tool.) We want to stress here the conceptual simplicity of the Bayesian solution to the problem. [In the case the data set contains some more than a dozen of observations, a flat $p(\mu, \sigma \,|\, I)$, with the constraint $\sigma > 0$, can be considered a good practical choice.]

5.7 Predictive distributions

A related problem is to 'infer' what an experiment will observe given our best knowledge of the underlying theory and its parameters. Infer is within quote marks because the term is usually used for model and parameters, rather than for observations. In this case people prefer to speak about *prediction* (or *prevision*). But we recall that in the Bayesian reasoning there is conceptual symmetry between the uncertain quantities which enter the problem. Probability density functions describing not yet observed event are referred to as *predictive distributions*. There is a conceptual difference with the likelihood, which also gives a probability of observation, but under different hypotheses, as the following example clarifies.

Given μ and σ, and assuming a Gaussian model, our uncertainty about a 'future' d_f is described by the Gaussian pdf Eq. (25) with $d = d_f$. But this holds only under that particular hypothesis for μ and σ, while, in general, we are also uncertain about these values too. Applying the decomposition formula (Tab. 1) we get:

$$p(d_f \,|\, I) \;=\; \int p(d_f \,|\, \mu, \sigma, I)\, p(\mu, \sigma \,|\, I)\, \mathrm{d}\mu\, \mathrm{d}\sigma \tag{58}$$

Again, the integral might be technically difficult, but the solution is conceptually simple. Note that, though the decomposition formula is a general result of probability theory, it can be applied to this problem only in the subjective approach.

An analytically easy, insightful case is that of experiments with well-known σ's. Given a past observation d_p and a vague prior, $p(\mu \,|\, d_p, I)$ is Gaussian around d_p with variance σ_p^2 [note that, with respect to $p(\mu, \sigma \,|\, I)$ of Eq.(58), it has been made explicit that $p(\mu)$ depend on d_p]. $p(d_f \,|\, \mu)$ is Gaussian around μ with variance σ_f^2. We get finally

$$p(d_f \,|\, d_p, I) = \int p(d_f \,|\, \mu, I)\, p(\mu \,|\, d_p, I)\, \mathrm{d}\mu \tag{59}$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_p^2 + \sigma_f^2}} \exp\left[-\frac{(d_f - d_p)^2}{2\,(\sigma_p^2 + \sigma_f^2)}\right]. \tag{60}$$

5.8 Hierarchical modelling and hyperparameters

As we have seen in the previous section, it is often desirable to include in a probabilistic model one's uncertainty in various aspects of a pdf. This is a natural feature of the Bayesian methods, due to the uniform approach to deal with uncertainty and from which powerful analysis tools are derived. This kind of this modelling is called *hierarchical* because the characteristics of one pdf are controlled by another pdf. All uncertain parameters from which the pdf depends are called *hyperparameter*. An example of use of hyperparameter is described in Sect. 8.3 in which the prior to infer θ in a binomial model are shown to be controlled by the parameters of a Beta distribution.

As an example of practical importance, think of the combination of experimental results in the presence of *outliers*, i.e. of data points which are somehow in mutual disagreement. In this case the combination rule given by Eqs. (30)–(32), extended to many data points, produces unacceptable conclusions. A way of solving the problem (Dose and von der Linden 1999, D'Agostini 1999b) is to model a scepticism about the quoted standard deviations of the experiments, introducing a pdf $f(r)$, where r is a rescaling factor of the standard deviation. In this way the σ's that enter the r.h.s. of Eqs. (30)–(32) are hyperparameters of the problem. An alternative approach, also based on hierarchical modelling, is shown in (Fröhner 2000). For a more complete introduction to the subject see e.g. (Gelman *et al* 1995).

5.9 From Bayesian inference to maximum-likelihood and minimum chi-square model fitting

Let us continue with the case in which we know so little about appropriate values of the parameters that a uniform distribution is a practical choice for the prior. Equation (52) becomes

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{d}, I) \;\propto\; p(\boldsymbol{d} \,|\, \boldsymbol{\theta}, I)\, p_0(\boldsymbol{\theta}, I) \propto p(\boldsymbol{d} \,|\, \boldsymbol{\theta}, I) = \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{d}), \tag{61}$$

where, we recall, the likelihood $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{d})$ is seen as a mathematical function of $\boldsymbol{\theta}$, with parameters $\boldsymbol{d}$.

The set of $\boldsymbol{\theta}$ that is most likely is that which maximizes $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{d})$, a result known as the *maximum likelihood principle*. Here it has been obtained again as a special case of a more general framework, under clearly stated hypotheses, without need to introduce new ad hoc rules. Note also that the inference does not depend on multiplicative factors in the likelihood. This is one of the ways to state the *likelihood principle*, ideally desired by frequentists, but often violated. This 'principle' always and naturally holds in Bayesian statistics. It is important to remark that the use of unnecessary principles is dangerous, because there is a tendency to use them uncritically. For example, formulae resulting from maximum likelihood are often used also when non-uniform reasonable priors should be taken into account, or when the shape of $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{d})$ is far from being multi-variate Gaussian. (This is a kind of ancillary default hypothesis that comes together with this principle, and is the source of the often misused '$\Delta(-\ln\mathcal{L}) = 1/2$' rule to determine probability intervals.)

The usual least squares formulae are easily derived if we take the well-known case of pairs $\{x_i, y_i\}$ (the generic $\boldsymbol{d}$ stands for all data points) whose true values are related by a deterministic function $\mu_{y_i} = y(\mu_{x_i}, \boldsymbol{\theta})$ and with Gaussian errors only in the ordinates, i.e. we consider $x_i \approx \mu_{x_i}$. In the case of independence of the measurements, the likelihood-dominated result becomes,

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \quad\propto\quad \prod_i \exp\left[-\frac{(y_i - y(x_i,\boldsymbol{\theta}))^2}{2\,\sigma_i^2}\right] \tag{62}$$

or

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \quad\propto\quad \exp\left[-\frac{1}{2}\chi^2\right]\,, \tag{63}$$

where

$$\chi^2(\boldsymbol{\theta}) \quad=\quad \sum_i \frac{(y_i - y(x_i,\boldsymbol{\theta}))^2}{\sigma_i^2} \tag{64}$$

is called 'chi-square,' well known among physicists. Maximizing the likelihood is equivalent to minimizing χ^2, and the most probable value of $\boldsymbol{\theta}$ is easily obtained (i.e. the *mode* indicated with $\boldsymbol{\theta}_m$), analytically in easy cases, or numerically for more complex ones.

As far as the uncertainty in $\boldsymbol{\theta}$ is concerned, the widely-used evaluation of the covariance matrix $\mathbf{V}(\boldsymbol{\theta})$ (see Sect. 5.6) from the Hessian,

$$(V^{-1})_{ij}(\boldsymbol{\theta}) \quad=\quad \frac{1}{2}\frac{\partial^2\chi^2}{\partial\theta_i\partial\theta_j}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_m}\,, \tag{65}$$

is merely consequence of an *assumed* multi-variate Gaussian distribution of $\boldsymbol{\theta}$, that is a parabolic shape of χ^2 (note that the '$\Delta(-\ln\mathcal{L}) = 1/2$' rule, and the from this rule resulting '$\Delta\chi^2 = 1$ rule,' has the same motivation). In fact, expanding $\chi^2(\boldsymbol{\theta})$ in series around its minimum, we have

$$\chi^2(\boldsymbol{\theta}) \quad\approx\quad \chi^2(\boldsymbol{\theta}_m) + \frac{1}{2}\,\boldsymbol{\Delta}\theta^T\,\mathbf{H}\,\boldsymbol{\Delta}\theta \tag{66}$$

where $\boldsymbol{\Delta}\theta$ stands for the the set of differences $\theta_i - \theta_{m_i}$ and $\mathbf{H}$ is the Hessian matrix, whose elements are given by twice the r.s.h. of Eq. (65). Equation (63) becomes then

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \quad\approx\quad \propto\exp\left[-\frac{1}{2}\boldsymbol{\Delta}\theta^T\,\mathbf{H}\,\boldsymbol{\Delta}\theta\right]\,, \tag{67}$$

which we recognize to be a multi-variate Gaussian distribution if we identify $\mathbf{H} = \mathbf{V}^{-1}$. After normalization, we get finally

$$p(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \quad\approx\quad (2\pi)^{-n/2}\,(\det\mathbf{V})^{-1/2}\,\exp\left[-\frac{1}{2}\boldsymbol{\Delta}\theta^T\,\mathbf{V}^{-1}\boldsymbol{\Delta}\theta\right]\,, \tag{68}$$

with n equal to the dimension of $\boldsymbol{\theta}$ and $\det\mathbf{V}$ indicating the determinant of $\mathbf{V}$. Holding this approximation, $\mathrm{E}(\boldsymbol{\theta})$ is approximately equal to $\boldsymbol{\theta}_m$. Note that the result (68) is exact when $y(\mu_{x_i}, \boldsymbol{\theta})$ depends linearly on the various θ_i.

In routine applications, the hypotheses that lead to the maximum likelihood and least squares formulae often hold. But when these hypotheses are not justified, we need to characterize the result by the multi-dimensional posterior distribution $p(\boldsymbol{\theta})$, going back to the more general expression Eq. (52).

The important conclusion from this section, as was the case for the definitions of probability in Sect. 3, is that Bayesian methods often lead to well-known conventional results, but without introducing them as new ad hoc rules as the need arises. The analyst acquires then a heightened sense of awareness about the range of validity of the methods. One might as well use these 'recovered' methods within the Bayesian framework, with its more natural interpretation of the results. Then one can speak about the uncertainty in the model parameters and quantify it with probability values, which is the usual way in which physicists think.

5.10 Gaussian approximation of the posterior distribution

The substance of the results seen in the previous section holds also in the case in which the prior is not flat and, hence, cannot be absorbed in the normalization constant of the posterior. In fact, in many practical cases the posterior exhibits an approximately (multi-variate) Gaussian shape, even if the prior was not trivial. Having at hand an *un-normalized* posterior $\tilde{p}()$, i.e.

$$\tilde{p}(\boldsymbol{\theta}\,|\,\boldsymbol{d},I) \;\;=\;\; p(\boldsymbol{d}\,|\,\boldsymbol{\theta},I)\,p_0(\boldsymbol{\theta},I)\,, \tag{69}$$

we can take its *minus-log* function $\varphi(\boldsymbol{\theta}) = -\ln\tilde{p}(\boldsymbol{\theta})$. If $\tilde{p}(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I)$ has approximately a Gaussian shape, it follows that

$$\varphi(\boldsymbol{\theta}) \;\;\approx\;\; \frac{1}{2}\,\boldsymbol{\Delta}\theta^T\,\mathbf{V}^{-1}\boldsymbol{\Delta}\theta + constant. \tag{70}$$

$\mathbf{V}$ can be evaluated as

$$(V^{-1})_{ij}(\boldsymbol{\theta}) \;\;\approx\;\; \left.\frac{\partial^2\varphi}{\partial\theta_i\partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_m}, \tag{71}$$

where $\boldsymbol{\theta}_m$ was obtained from the minimum of $\varphi(\boldsymbol{\theta})$.

6 Uncertainties from systematic effects

The uncertainty described in the previous section are related to the so-called *random*, or *statistical* errors. Other important sources are, generally speaking (see ISO 1993 for details), related to uncertain values of *influence variables* on which the observed values, or the data-analysis process, might depend. In physics, we usually refer to these as systematic effects or errors. They can be related to the parameters of the experiment, like a particle beam energy or the exposure time, or to environmental variables, like temperature and pressure, calibration constants of the detector, and all other parameters, 'constants' (in the physical sense), and hypotheses that enter the data analysis. The important thing is that we are unsure about their precise value. Let us indicate all the influence variables with the vector $\boldsymbol{h} = \{h_1, h_2, \ldots, h_n\}$, and their joint pdf as $p(\boldsymbol{h}\,|\,I)$.

The treatment of uncertainties due to systematic errors has traditionally been lacking a consistent theory, essentially due the unsuitability to standard statistical methods of dealing with uncertainty in the most wide sense. Bayesian reasoning becomes crucial to handle these sources of uncertainty too, and even metrological organizations (ISO 1993) had to recognized it. For example, the ISO *type B* uncertainty is recommended to be *"evaluated by scientific judgment based on all the available information on the possible variability"* (ISO 1993) of the influence quantities (see also D'Agostini 1999c).

6.1 Reweighting of conditional inferences

The values of the influence variables and their uncertainties contribute to our background knowledge I about the experimental measurements. Using I_0 to represent our very general background knowledge, the posterior pdf will then be $p(\mu \,|\, \boldsymbol{d}, \boldsymbol{h}, I_0)$, where the dependence on all possible values of $\boldsymbol{h}$ has been made explicit. The inference that takes into account the uncertain vector $\boldsymbol{h}$ is obtained using the rules of probability (see Tab. 1) by integrating the joint probability over the uninteresting influence variables:

$$\begin{aligned} p(\mu \,|\, \boldsymbol{d}, I_0) &= \int p(\mu, \boldsymbol{h} \,|\, \boldsymbol{d}, I_0)\, \mathrm{d}\boldsymbol{h} && (72) \\ &= \int p(\mu \,|\, \boldsymbol{d}, \boldsymbol{h}, I_0)\, p(\boldsymbol{h} \,|\, I_0)\, \mathrm{d}\boldsymbol{h}\,. && (73) \end{aligned}$$

As a simple, but important case, let us consider a single influence variable given by an additive instrumental offset z, which is expected to be zero because the instrument has been calibrated as well as feasible and the remaining uncertainty is σ_z. Modelling our uncertainty in z as a Gaussian distribution with a standard deviation σ_z, the posterior for μ is

$$\begin{aligned} p(\mu \,|\, d, I_0) &= \int_{-\infty}^{+\infty} p(\mu \,|\, d, z, \sigma, I_0)\, p(z \,|\sigma_z, I_0)\, \mathrm{d}z && (74) \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{(\mu-(d-z))^2}{2\,\sigma^2}\right] \times \\ &\qquad \frac{1}{\sqrt{2\pi}\,\sigma_z} \exp\left[-\frac{z^2}{2\,\sigma_z^2}\right] \mathrm{d}z && (75) \\ &= \frac{1}{\sqrt{2\pi}\,\sqrt{\sigma^2+\sigma_z^2}} \exp\left[-\frac{(\mu-d)^2}{2\,(\sigma^2+\sigma_z^2)}\right]. && (76) \end{aligned}$$

The result is that the net variance is the sum of the variance in the measurement and the variance in the influence variable.

6.2 Joint inference and marginalization of nuisance parameters

A different approach, which produces identical results, is to think of a joint inference about both the quantities of interest and the influence variables:

$$p(\mu, \boldsymbol{h} \,|\, \boldsymbol{d}, I_0) \;\propto\; p(\boldsymbol{d} \,|\, \mu, \boldsymbol{h}, I_0)\, p_0(\mu, \boldsymbol{h} \,|\, I_0)\,. \qquad (77)$$

Then, marginalization is applied to the variables that we are not interested in (the so called *nuisance parameters*), obtaining

$$\begin{aligned} p(\mu|\, \boldsymbol{d}, I_0) &= \int p(\mu, \boldsymbol{h} \,|\, \boldsymbol{d}, I_0)\, \mathrm{d}\boldsymbol{h} && (78) \\ &\propto \int p(\boldsymbol{d} \,|\, \mu, \boldsymbol{h}, I_0)\, p_0(\mu, \boldsymbol{h} \,|\, I_0)\, \mathrm{d}\boldsymbol{h}\,. && (79) \end{aligned}$$

Equation (77) shows a peculiar feature of Bayesian inference, namely the possibility making an inference about a number of variables larger than the number of the observed data. Certainly, there is no magic in it, and the resulting variables will be highly correlated. Moreover, the prior cannot be improper in all variables. But, by using *informative priors* in which experts feel confident, this feature allows one to tackle complex problems with missing or corrupted parameters. In the end, making use of marginalization, one can concentrate on the quantities of real interest.

The formulation of the problem in terms of Eqs. (77) and (79) allows one to solve problems in which the influence variables might depend on the true value μ, because $p_0(\mu, \boldsymbol{h} \,|\, I_0)$ can model dependences between μ and $\boldsymbol{h}$. In most applications, $\boldsymbol{h}$ does not depend on μ, and the prior factors into the product of $p_0(\mu \,|\, I_0)$ and $p_0(\boldsymbol{h} \,|\, I_0)$. When this happens, we recover exactly the same results as obtained using the reweighting of conditional inferences approach described just above.

6.3 Correlation in results caused by systematic errors

We can easily extend Eqs. (73), (77), and (79) to a joint inference of several variables, which, as we have seen, are nothing but parameters $\boldsymbol{\theta}$ of suitable models. Using the alternative ways described in Sects. 6.1 and 6.2, we have

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{d}, \boldsymbol{h}, I_0) \;\propto\; p(\boldsymbol{d} \,|\, \boldsymbol{\theta}, \boldsymbol{h}, I_0)\, p_0(\boldsymbol{\theta} \,|\, I_0) \tag{80}$$

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{d}, I_0) \;=\; \int p(\boldsymbol{\theta} \,|\, \boldsymbol{d}, \boldsymbol{h}, I_0)\, p(\boldsymbol{h} \,|\, I_0)\, \mathrm{d}\boldsymbol{h} \tag{81}$$

and

$$p(\boldsymbol{\theta}, \boldsymbol{h} \,|\, \boldsymbol{d}, I_0) \;\propto\; p(\boldsymbol{d} \,|\, \boldsymbol{\theta}, \boldsymbol{h}, I_0)\, p_0(\boldsymbol{\theta}, \boldsymbol{h} \,|\, I_0) \tag{82}$$

$$p(\boldsymbol{\theta} | \, \boldsymbol{d}, I_0) \;=\; \int p(\boldsymbol{\theta}, \boldsymbol{h} \,|\, \boldsymbol{d}, I_0)\, \mathrm{d}\boldsymbol{h}\,, \tag{83}$$

respectively. The two ways lead to an identical result, as it can be seen comparing Eqs. (81) and (83).

Take a simple case of a common offset error of an instrument used to measure various quantities μ_i, resulting in the measurements d_i. We model each measurement as μ_i plus an error that is Gaussian distributed with a mean of zero and a standard deviation σ_i. The calculation of the posterior distribution can be performed analytically, with the following results (see D'Agostini 1999c for details):

- The uncertainty in each μ_i is described by a Gaussian centered at d_i, with standard deviation $\sigma(\mu_i) = \sqrt{\sigma_i^2 + \sigma_z^2}$, consistent with Eq. (76).
- The joint posterior distribution $p(\mu_1, \mu_2, \ldots)$ does not factorize into the product of $p(\mu_1)$, $p(\mu_2)$, etc., because correlations are automatically introduced by the formalism, consistent with the intuitive thinking of what a common systematic should do. Therefore, the joint distribution will be a multi-variate Gaussian that takes into account correlation terms.
- The *correlation coefficient* between any pair $\{\mu_i, \mu_j\}$ is given by

$$\rho(\mu_i, \mu_j) \;=\; \frac{\sigma_z^2}{\sigma(\mu_i)\,\sigma(\mu_j)} = \frac{\sigma_z^2}{\sqrt{(\sigma_i^2 + \sigma_z^2)\,(\sigma_j^2 + \sigma_z^2)}}\,. \tag{84}$$

  We see that $\rho(\mu_i, \mu_j)$ has the behavior expected from a common offset error; it is non-negative; it varies from practically zero, indicating negligible correlation, when $(\sigma_z \ll \sigma_i)$, to unity $(\sigma_z \gg \sigma_i)$, when the offset error dominates.

6.4 Approximate methods and standard propagation applied to systematic errors

When we have many uncertain influence factors and/or the model of uncertainty is non-Gaussian, the analytic solution of Eq. (73), or Eqs. (77)–(79) can be complicated, or not existing at all.

Then numeric or approximate methods are needed. The most powerful numerical methods are based on *Monte Carlo* (MC) techniques (see Sect. 9 for a short account). This issue goes beyond the aim of this report. In a recent comprehensive particle-physics paper by Ciuchini *et al* (2001), these ideas have been used to infer the fundamental parameters of the Standard Model of particle physics, using all available experimental information.

For routine use, a practical approximate method can be developed by thinking of the value inferred for the expected value of $\boldsymbol{h}$ as a *raw* value, indicated with μ_R, that is, $\mu_R = \mu\,\Big|_{\boldsymbol{h}=\mathrm{E}(\boldsymbol{h})}$ ('raw' in the sense that it needs later to be 'corrected' for all possible value of $\boldsymbol{\theta}$, as it will be clear in a while). The value of μ, which depends on the possible values of $\boldsymbol{h}$, can be seen as a function of μ_R and $\boldsymbol{h}$:

$$\mu \;=\; f(\mu_R, \boldsymbol{h})\,. \tag{85}$$

We have thus turned our inferential problem into a standard problem of evaluation of the pdf of a function of variables, of which are particularly known the formulae to obtain approximate values for expectations and standard deviations in the case of independent *input quantities* (following the nomenclature of ISO 1993):

$$\mathrm{E}(\mu) \;\approx\; f(\mathrm{E}(\mu_R), \mathrm{E}(\boldsymbol{h})) \tag{86}$$

$$\begin{aligned} \sigma^2(\mu) \;\approx\; & \left(\left.\frac{\partial f}{\partial \mu_R}\right|_{\mathrm{E}(\mu_R),\mathrm{E}(\boldsymbol{h})}\right)^2 \sigma^2(\mu_R) \\ & + \sum_i \left(\left.\frac{\partial f}{\partial h_i}\right|_{\mathrm{E}(\mu_R),\mathrm{E}(\boldsymbol{h})}\right)^2 \sigma^2(h_i)\,. \end{aligned} \tag{87}$$

Extension to multi-dimensional problems and treatment of correlations is straightforward (the well-known covariance matrix propagation) and we refer to (D'Agostini and Raso 1999) for details. In particular, this reference contains approximate formulae valid up to second order, which allow to take into account relatively easily non linearities.

7 Comparison of models of different complexity

We have seen so far two typical inferential situations:

1. Comparison of simple models (Sect. 4), where by simple we mean that the models do not depend on parameters to be tuned to the experimental data.

2. Parametric inference given a model, to which we have devoted the last sections.

A more complex situation arises when we have several models, each of which might depend on several parameters. For simplicity, let us consider model A with n_A parameters $\boldsymbol{\alpha}$ and model B with n_B parameters $\boldsymbol{\beta}$. In principle, the same Bayesian reasoning seen previously holds:

$$\frac{P(A\,|\,\mathrm{Data}, I)}{P(B\,|\,\mathrm{Data}, I)} \;=\; \frac{P(\mathrm{Data}\,|\,A, I)}{P(\mathrm{Data}\,|\,B, I)}\,\frac{P(A\,|\,I)}{P(B\,|\,I)}\,, \tag{88}$$

but we have to remember that the probability of the data, given a model, depends on the probability of the data, given a model and any particular set of parameters, weighted with the prior beliefs about parameters. We can use the same decomposition formula (see Tab. 1), already applied in treating systematic errors (Sect. 6):

$$P(\mathrm{Data}\,|\,M, I) \;=\; \int P(\mathrm{Data}\,|\,M, \boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}\,|\,I)\,\mathrm{d}\boldsymbol{\theta}\,, \tag{89}$$

with $M = A, B$ and $\boldsymbol{\theta} = \boldsymbol{\alpha}, \boldsymbol{\beta}$. In particular, the Bayes factor appearing in Eq. (88) becomes

$$\frac{P(\mathrm{Data}\,|\,A, I)}{P(\mathrm{Data}\,|\,B, I)} = \frac{\int P(\mathrm{Data}\,|\,A, \boldsymbol{\alpha}, I)\, p(\boldsymbol{\alpha}\,|\,I)\, \mathrm{d}\boldsymbol{\alpha}}{\int P(\mathrm{Data}\,|\,B, \boldsymbol{\beta}, I)\, p(\boldsymbol{\beta}\,|\,I)\, \mathrm{d}\boldsymbol{\beta}} \tag{90}$$

$$= \frac{\int \mathcal{L}_A(\boldsymbol{\alpha};\,\mathrm{Data})\, p_0(\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\alpha}}{\int \mathcal{L}_B(\boldsymbol{\beta};\,\mathrm{Data})\, p_0(\boldsymbol{\beta})\, \mathrm{d}\boldsymbol{\beta}}\,. \tag{91}$$

The inference depends on the *marginalized likelihood* (89), also known as the *evidence*. Note that $\mathcal{L}_M(\boldsymbol{\theta};\,\mathrm{Data})$ has its largest value around the maximum likelihood point $\boldsymbol{\theta}_{ML}$, but the evidence takes into account all prior possibilities of the parameters. Thus, it is not enough that the best fit of one model is superior to its alternative, in the sense that, for instance,

$$\mathcal{L}_A(\boldsymbol{\alpha}_{ML};\,\mathrm{Data}) \;>\; \mathcal{L}_B(\boldsymbol{\beta}_{ML};\,\mathrm{Data})\,, \tag{92}$$

and hence, assuming Gaussian models,

$$\chi^2_A(\boldsymbol{\alpha}_{min\,\chi^2};\,\mathrm{Data}) \;<\; \chi^2_B(\boldsymbol{\beta}_{min\,\chi^2};\,\mathrm{Data})\,, \tag{93}$$

to prefer model A. We have already seen that we need to take into account the prior beliefs in A and B. But even this is not enough: we also need to consider the space of possibilities and then the adaptation capability of each model. It is well understood that we do not choose an $(n-1)$ order polynomial as the best description – 'best' in inferential terms – of n experimental points, though such a model always offers an exact pointwise fit. Similarly, we are much more impressed by, and we tend *a posteriori* to believe more in, a theory that absolutely predicts an experimental observation, within a reasonable error, than another theory that performs similarly or even better after having adjusted many parameters.

This intuitive reasoning is expressed formally in Eqs. (90) and (91). The evidence is given integrating the product $\mathcal{L}(\boldsymbol{\theta})$ and $p_0(\boldsymbol{\theta})$ over the parameter space. So, the more $p_0(\boldsymbol{\theta})$ is concentrated around $\boldsymbol{\theta}_{ML}$, the greater is the evidence in favor of that model. Instead, a model with a volume of the parameter space much larger than the one selected by $\mathcal{L}(\boldsymbol{\theta})$ gets disfavored. The extreme limit is that of a hypothetical model with so many parameters to describe whatever we shall observe. This effect is very welcome, and follows the *Ockham's Razor* scientific rule of discarding unnecessarily complicated models (*"entities should not be multiplied unnecessarily"*). This rule comes out of the Bayesian approach automatically and it is discussed, with examples of applications in many papers. Berger and Jefferys (1992) introduce the connection between Ockham's Razor and Bayesian reasoning, and discuss the evidence provided by the motion of Mercury's perihelion in favor of Einstein's general relativity theory, compared to alternatives at that time. Examples of recent applications are Loredo and Lamb 2002 (analysis of neutrinos observed from supernova SN 1987A), John and Narlikar 2002 (comparisons of cosmological models), Hobson *et al* 2002 (combination of cosmological datasets) and Astone *et al* 2003 (analysis of coincidence data from gravitational wave detectors). These papers also give a concise account of underlying Bayesian ideas.

After having emphasized the merits of model comparison formalized in Eqs. (90) and (91), it is important to mention a related problem. In parametric inference we have seen that we can make an easy use of improper priors (see Tab. 1), seen as limits of proper priors, essentially because they simplify in the Bayes formula. For example, we considered $p_0(\mu\,|\,I)$ of Eq. (26) to be a constant, but this constant goes to zero as the range of μ diverges. Therefore, it does simplify in Eq. (26), but not, in general, in Eqs. (90) and (91), unless models A and B depend on the same number of parameters defined in the same ranges. Therefore, the general case of model comparison is limited to proper priors, and needs to be thought through better than when making parametric inference.

8 Choice of priors – a closer look

So far, we have considered mainly likelihood-dominated situations, in which the prior pdf can be included in the normalization constant. But one should be careful about the possibility of uncritically use uniform priors, as a 'prescription,' or as a rule, though the rule might be associated with the name of famous persons. For instance, having made N interviews to infer the proportion θ of a population that supports a party, it is not reasonable to assume a uniform prior of θ between 0 and 1. Similarly, having to infer the rate r of a Poisson process (such that $\lambda = r\,T$, where T is the measuring time) related, for example, to proton decay, cosmic ray events or gravitational wave signals, we do not believe, strictly, that $p(r)$ is uniform between zero and infinity. Besides natural physical cut-off's (for example, very large proton decay r would prevent Life, or even stars, to exist), $p(r) = \text{constant}$ implies to believe more high orders of magnitudes of r (see Astone and D'Agostini 1999 for details). In many cases (for example the mentioned searches for rare phenomena) our uncertainty could mean indifference over several orders of magnitude in the rate r. This indifference can be parametrized roughly with a prior uniform $\ln r$ yielding $p(r) \propto 1/r$ (the same prior is obtainable using invariant arguments, as it will be shown in a while).

As the reader might imagine, the choice of priors is a highly debated issue, also among Bayesians. We do not pretend to give definitive statements, but would just like to touch on some important issues concerning priors.

8.1 Logical and practical role of priors

Priors are pointed to by those critical of the Bayesian approach as *the* major weakness of the theory. Instead, Bayesians consider them a crucial and powerful key point of the method. Priors are logically crucial because they are necessary to make probability inversions via Bayes' theorem. This point remains valid even in the case in which they are vague and apparently disappear in the Bayes' formula. Priors are powerful because they allow to deal with realistic situations in which informative prior knowledge can be taken into account and properly balanced with the experimental information.

Indeed, we think that one of the advantages of Bayesian analysis is that it explicitly admits the existence of prior information, which naturally leads to the expectation that the prior will be specified in any honest account of a Bayesian analysis. This crucial point is often obscured in other types of analyses, in large part because the analysts maintain their method is 'objective.' Therefore, it is not easy, in those analyses, to recognize what are the specific assumptions made by the analyst — in practice the analyst's priors — and the assumptions included in the method (the latter assumptions are often unknown to the average practitioner).

8.2 Purely subjective assessment of prior probabilities

In principle, the point is simple, at least in one-dimensional problem in which there is good perception of the possible range in which the uncertain variable of interest could lie: try your best to model your prior beliefs. In practice, this advice seems difficult to follow because, even if we have a rough idea of what the value of a quantity should be, the representation of the prior in mathematical terms seems very committal, because a pdf implicitly contains an infinite number of precise probabilistic statements. (Even the uniform distribution says that we believe *exactly* in the same way to all values. Who believes exactly that?) It is then important to understand that, when expressing priors, what matters is not the precise mathematical formula, but the gross value of the probability mass indicated by the formula, how probabilities are intuitively perceived and how priors influence posteriors. When we say, intuitively, we believe something with a 95% confidence, it means "we are almost sure," but the precise value (95%, instead of 92% or 98%)

is not very relevant. Similarly, when we say that the prior knowledge is modeled by a Gaussian distribution centered around μ_0 with standard deviation σ_0 [Eq. (28)], it means means that we are quite confident that μ is within $\pm\sigma_0$, very sure that it is within $\pm 2\sigma_0$ and *almost certain* that it is within $\pm 3\,\sigma_0$. Values even farther from μ_0 are possible, though we do not consider them very likely. But all models should be taken with a grain of salt, remembering that they are often just mathematical conveniences. For example, a textbook-Gaussian prior includes infinite deviations from the expected value and even negative values for physical quantities positively defined, like a temperature or a length. All absurdities, if taken literally. On the other hand, we think that all experienced physicists have in mind priors with low probability long tails in order to accommodate strong deviation from what is expected with highest probability. (Remember that where the prior is zero, the posterior must also be zero.)

Summing up this point, it is important to understand that a prior should tell where the *probability mass* is concentrated, without taking too seriously the details, especially the tails of the distribution (which should be, however, enough extended to accommodate 'surprises'). The nice feature of Bayes' theorem is the ability of trasform such vague, fuzzy priors into solid estimates, if a sufficient amount of good quality data are at hand. For this reason, the use of *improper priors* is not considered to be problematic. Indeed, improper priors can just be considered a convenient way of modelling relative beliefs.

In the case we have doubts about the choice of the prior, we can consider a family of functions with some hyperparameters. If we worry about the effect of the chosen prior on the posterior, we can perform a *sensitivity analysis*, i.e. to repeat the analysis for different, *reasonable* choices of the prior and check the variation of the result. The final uncertainty could, then, take into account also the uncertainty on the prior. Finally, in extreme cases in which priors play a crucial role and could dramatically change the conclusions, one should refrain to give probabilistic result, providing, instead, only Bayes factors, or even just likelihoods. An example of a recent result about gravitational wave searches presented in this way, see Astone *et al* (2002).

Having clarified meaning and role of priors, it is rather evident that the practical choice of a prior depends on what is appropriate for the application. For example, in the area of imaging, smoothness of a reconstructed image might be appropriate in some situations. Smoothness may be imposed by a variety of means, for example, by simply setting the logarithm of the prior equal to an integral of the square of the second derivative of the image (von der Linden *et al* 1996b). A more sophisticated approach goes under the name of Markov random fields (MRF), which can even preserve sharp edges in the estimated images (Bouman and Sauer 1993, Saquib *et al* 1997). A similar kind of prior is often appropriate for defomable geometric models, which can be used to represent the boundaries between various regions, for example, organs in medical images (Cunningham *et al* 1998).

A procedure that helps in choosing the prior, expecially important in the cases in which the parameters do not have a straightforwardly perceptible influence on data, is to build a *prior predictive* pdf and check if this pdf would produce data conform with our prior beliefs. The prior predictive distribution is the analogue of the (*posterior*) predictive distribution we met in Sect. 5.7, with $p(\boldsymbol{\theta}\,|\,\boldsymbol{d}, I)$ replaced by $p(\boldsymbol{\theta}\,|\,I)$ (note that the example of Sect. 5.7 was one-dimensional, with $d_1 = d_f$ and $\theta_1 = \mu$), i.e. $p(\boldsymbol{d}\,|\,I) = \int p(\boldsymbol{d}\,|\,\boldsymbol{\theta}, I)\,p(\boldsymbol{\theta}\,|\,I)\,\mathrm{d}\boldsymbol{\theta}$.

Often, expecially in complicated data analyses, we are not sufficiently knowledgable about the details of the problem. Thus, informative priors have to be modelled that capture the judgement of experts. For example, Meyer and Booker (2001) show a formal process of prior elicitation which has the aim at reducing, as much as possible, the bias in the experts' estimates of their confidence limits. This approach allows one to combine the results from several experts. In short, we can suggest the use of the 'coherent bet' (Sect. 2) to force experts to access their values of probability, asking them to provide an interval in which they feel 'practically sure', intervals on which they could wager 1:1, and so on.

8.3 Conjugate priors

Because of computational problems, modelling priors has been traditionally a compromise between a realistic assessment of beliefs and choosing a mathematical function that simplifies the analytic calculations. A well-known strategy is to choose a prior with a suitable form so the posterior belongs to the same functional family as the prior. The choice of the family depends on the likelihood. A prior and posterior chosen in this way are said to be *conjugate*. For instance, given a Gaussian likelihood and choosing a Gaussian prior, the posterior is still Gaussian, as we have seen in Eqs. (25), (28) and (29). This is because expressions of the form

$$K\,\exp\left[-\frac{(x_1-\mu)^2}{2\sigma_1^2}-\frac{(x_2-\mu)^2}{2\sigma_2^2}\right]$$

can always be written in the form

$$K'\,\exp\left[-\frac{(x'-\mu)^2}{2\sigma'^2}\right]\,,$$

with suitable values for x', σ' and K'. The Gaussian distribution is auto-conjugate. The mathematics is simplified but, unfortunately, only one shape is possible.

An interesting case, both for flexibility and practical interest is offered by the binomial likelihood (see Sect. 5.3). Apart from the binomial coefficient, $p(n\,|\,\theta,N)$ has the shape $\theta^n(1-\theta)^{N-n}$, which has the same structure as the *Beta distribution*, well known to statisticians:

$$\mathrm{Beta}(\theta\,|\,r,s)=\frac{1}{\beta(r,s)}\theta^{r-1}(1-\theta)^{s-1}\qquad\left\{\begin{array}{l}0\le\theta\le 1\\ r,\,s>0\end{array}\right.\,,\tag{94}$$

where $\beta(r,s)$ stands for the Beta function, defined as

$$\beta(r,s)=\int_0^1\theta^{r-1}(1-\theta)^{s-1}\,\mathrm{d}\theta\tag{95}$$

which can be expressed in terms of Euler's Gamma function as $\beta(r,s)=\Gamma(r)\,\Gamma(s)/\Gamma(r+s)$. Expectation and variance of the Beta distribution are:

$$\mathrm{E}(\theta)\;=\;\frac{r}{r+s}\tag{96}$$

$$\sigma^2(\theta)\;=\;\frac{r\,s}{(r+s+1)(r+s)^2}=\mathrm{E}^2(\theta)\,\frac{s}{r}\,\frac{1}{r+s+1}\,.\tag{97}$$

If $r>1$ and $s>1$, then the mode is unique, and it is at $\theta_m=(r-1)/(r+s-2)$. Depending on the value of the parameters the Beta pdf can take a large variety of shapes. For example, for large values of r and s, the function is very similar to a Gaussian distribution, while a constant function is obtained for $r=s=1$. Using the Beta pdf as prior function in inferential problems with a binomial likelihood, we have

$$p(\theta\,|\,n,N,r,s)\;\propto\;\left[\theta^n(1-\theta)^{N-n}\right]\left[\theta^{r-1}(1-\theta)^{s-1}\right]\tag{98}$$

$$\propto\;\theta^{n+r-1}(1-\theta)^{N-n+s-1}\,.\tag{99}$$

The posterior distribution is still a Beta with $r'=r+n$ and $s'=s+N-n$, and expectation and standard deviation can be calculated easily from Eqs. (96) and (97). These formulae demonstrate how the posterior estimates become progressively independent of the prior information in the limit of large numbers; this happens when both $m\gg r$ and $n-m\gg s$. In this limit, we get the same result as for a uniform prior ($r=s=1$).

Table 2 lists some of the more useful conjugate priors. For a more complete collection of conjugate priors, see e.g. (Bernardo and Smith 1994, Gelman *et al* 1995).

Table 2: *Some useful conjugate priors. x and n stand for the observed value (continuous or discrete, respectively) and θ is the generic symbol for the parameter to infer, corresponding to μ of a Gaussian, θ of a binomial and λ of a Poisson distribution.*

likelihood $p(x \mid \theta)$	conjugate prior $p_0(\theta)$	posterior $p(\theta \mid x)$
Normal(θ, σ)	Normal(μ_0, σ_0)	Normal(μ_1, σ_1) [Eqs. (30)–(32)]
Binomial(N, θ)	Beta(r, s)	Beta$(r+n, s+N-n)$
Poisson(θ)	Gamma(r, s)	Gamma$(r+n, s+1)$
Multinomial$(\theta_1, \ldots, \theta_k)$	Dirichlet$(\alpha_1, \ldots, \alpha_k)$	Dirichlet$(\alpha_1+n_1, \ldots, \alpha_k+n_k)$

8.4 General principle based priors

Many who advocate using the Bayesian approach still want to keep 'subjective' contributions to the inference to a minimum. Their aim is to derive prior functions based on 'objective' arguments or general principles. As the reader might guess, this subject is rather controversial, and the risk of trasforming arguments, which might well be reasonable and useful in many circumstances, into dogmatic rules is high (D'Agostini, 1999e).

8.4.1 Transformation invariance

An important class of priors arises from the requirement of transformation invariance. We shall consider two specific cases, translation invariance and scale invariance.

Translation invariance

Let us assume we are indifferent over a transformation of the kind $\theta' = \theta + b$, where θ is our variable of interest and b a constant. Then $p(\theta)\ \mathrm{d}\theta$ is an infinitesimal mass element of probability for θ to be in the interval $\mathrm{d}\theta$. Translation invariance requires that this mass element remains unchanged when expressed in terms of θ', i.e.

$$\begin{aligned} p(\theta)\ \mathrm{d}\theta &= p(\theta')\ \mathrm{d}\theta' &(100)\\ &= p(\theta + b)\ \mathrm{d}\theta\,, &(101)\end{aligned}$$

since $\mathrm{d}\theta = \mathrm{d}\theta'$. It is easy to see that in order for Eq. (101) to hold for any b, $p(\theta)$ must be equal to a constant for all values of θ from $-\infty$ to $+\infty$. It is therefore an improper prior. As discussed above, this is just a convenient modelling. For practical purposes this prior should always be regarded as the limit for $\Delta\theta \to \infty$ of $p(\theta) = 1/\Delta\theta$, where $\Delta\theta$ is a large finite range around the values of interest.

Scale invariance

In other cases, we could be indifferent about a scale transformation, that is $\theta' = \beta\,\theta$, where β is a constant. This invariance implies, since $\mathrm{d}\theta' = \beta\,\mathrm{d}\theta$ in this case,

$$p(\theta)\ \mathrm{d}\theta = p(\beta\,\theta)\,\beta\,\mathrm{d}\theta\,, \qquad (102)$$

i.e.

$$p(\beta\,\theta) = \frac{p(\theta)}{\beta}\,. \qquad (103)$$

The solution of this functional equation is

$$p(\theta) \propto \frac{1}{\theta}\,, \qquad (104)$$

as can be easily proved using Eq. (104) as test solution in Eq. (103). This is the famous *Jeffreys' prior*, since it was first proposed by Jeffreys. Note that this prior also can be stated as $p(\log\theta) = \text{constant}$, as can be easily verified. The requirement of scale invariance also produces an improper prior, in the range $0 < \theta < \infty$. Again, the improper prior must be understood as the limit of a proper prior extending several orders of magnitude around the values of interest. [Note that we constrain θ to be positive because, traditionally, variables which are believed to satisfy this invariance are associated with positively defined quantities. Indeed, Eq. (104) has a symmetric solution for negative quantities.]

According to the supporters of these invariance motivated priors (see e.g. Jaynes 1968, 1973, 1998, Sivia 1997, and Fröhner 2000, Dose 2002) variables associated to translation invariance are *location parameters*, as the parameter μ in a Gaussian model; variables associated to scale invariance are *scale parameters*, like σ in a Gaussian model or λ in a Poisson model. For criticism about the (mis-)use of this kind of prior see (D'Agostini 1999d).

8.4.2 Maximum-entropy priors

Another principle-based approach to assigning priors is based on in the *Maximum Entropy principle* (Jaynes 1957a, also 1983, 1998, Tribus 1969, von der Linden 1995, Sivia 1997, and Fröhner 2000). The basic idea is to choose the prior function that maximizes the Shannon-Jaynes *information entropy*,

$$S = -\sum_i^n p_i \ln p_i\,, \tag{105}$$

subject to whatever is assumed to be known about the distribution. The larger S is, the greater is our ignorance about the uncertain value of interest. The value $S = 0$ is obtained for a distribution that concentrates all the probability into a single value. In the case of no constraint other than normalization, ($\sum_i^n p_i = 1$), S is maximized by the uniform distribution, $p_i = 1/n$, which is easily proved using Lagrange multipliers. For example, if the variable is an integer between 0 and 10, a uniform distribution $p(x_i) = 1/11$ gives $S = 2.40$. Any binomial distribution with $n = 10$ gives a smaller value, with a maximum of $S = 1.88$ for $\theta = 1/2$ and a limit of $S = 0$ for $\theta \to 0$ or or $\theta \to 1$, where θ is now the parameter of the binomial that gives the probability of success at each trial.

Two famous cases of maximum-entropy priors for continuous variables are when the only information about the distribution is either the expected value or the expected value and the variance. Indeed, these are special cases of general constraints on the moments of the distribution M_r (see Tab. 1). For $r = 0$ and 1, M_r is equal to unity and to the expected value, respectively. First and second moment together provide the variance (see Tab. 1 and Sect. 5.6). Let us sum up what the assumed knowledge on the various moments provides [see e.g. (Sivia 1997, Dose 2002)].

Knowledge about M_0

Normalization alone provides a uniform distribution over the interval in which the variable is defined:

$$p(\theta \mid M_0 = 1) = \frac{1}{b-a} \qquad a \le \theta \le b\,. \tag{106}$$

This is the extension to continuous variables of the discrete case we saw above.

Knowledge about M_0 and M_1 [i.e. about $E(\theta)$]

Adding to the constraint $M_0 = 1$ the knowledge about the expectation of the variable, plus the requirement that all non-negative values are allowed, an exponential distribution is obtained:

$$p(\theta \,|\, M_0 = 1, M_1 \equiv \mathrm{E}(\theta)) \;=\; \frac{1}{\mathrm{E}(\theta)} e^{-\theta/\mathrm{E}(\theta)} \qquad 0 \le \theta < \infty\,. \tag{107}$$

Knowledge about M_0, M_1 and M_2 [i.e. about $E(\theta)$ and $\sigma(\theta)$]

Finally, the further constraint provided by the standard deviation (related to first and second moment by the equation $\sigma^2 = M_2 - M_1^2$) yields a prior with a Gaussian shape independently of the range of θ, i.e.

$$p(\theta \,|\, M_0 = 1, \mathrm{E}(\theta), \sigma(\theta)) \;=\; \frac{\exp\left[-\frac{(\theta-\mathrm{E}(\theta))^2}{2\,\sigma^2(\theta)}\right]}{\int_a^b \exp\left[-\frac{(\theta-\mathrm{E}(\theta))^2}{2\,\sigma^2(\theta)}\right]\,\mathrm{d}\theta} \qquad a \le \theta \le b\,. \tag{108}$$

The standard Gaussian is recovered when θ is allowed to be any real number.

Note, however, that the counterpart of Eq. (105) for continuous variables is not trivial, since all p_i of Eq. (105) tend to zero. Hence the analogous functional form $\int p(\theta) \ln p(\theta)\,\mathrm{d}\theta$ no longer has a sensible interpretation in terms of uncertainty, as remarked by Bernardo and Smith (1994). The Jaynes' solution is to introduce a 'reference' density $m(\theta)$ to make entropy invariant under coordinate transformation via $\int p(\theta) \ln[p(\theta)/m(\theta)]\,\mathrm{d}\theta$. (It is important to remark that the first and the third case discussed above are valid under the assumption of a unity reference density.) This solution is not universally accepted (see Bernardo and Smith 1994), even though it conforms to the requirements of dimensional analysis. Anyhow, besides formal aspects and the undeniable aid of Maximum Entropy methods in complicate problems such as image reconstruction (Buck and Macauly 1991), we find it very difficult, if not impossible at all, that a practitioner holds that status of knowledge which give rise to the two celebrated cases discussed above. We find more reasonable the approach described in Sect. 8.2, that goes the other way around: we have a rough idea of where the quantity of interest could be, then we try to model it and to summarize it in terms of expected value and standard deviation. In particular, we find untenable the position of those who state that Gaussian distribution can only be justified by Maximum Entropy principle.

8.4.3 Reference priors

We conclude this discussion on priors by mentioning 'reference analysis,' which is an area of active research among statisticians. The intention is, similarly to that for other priors motivated by basic principles, that of "characterizing a '*non-informative*' or '*objective*' prior distribution, representing '*prior ignorance*,' '*vague prior knowledge*,' and '*letting the data speak for themselves*' " (Bernardo and Smith 1994). However, "the problem is more complex than the apparent intuitive immediacy of these words and phrases would suggest" (Bernardo and Smith 1994, p. 298):

> *"Put bluntly: data cannot ever speak entirely for themselves: every prior specification has some informative posterior or predictive implications; and 'vague' is itself much too vague an idea to be useful. There is no 'objective' prior that represents ignorance.*
>
> *On the other hand, we recognize that there is often a pragmatically important need for a form of prior to posterior analysis capturing,* in some well-defined sense, *the notion of the prior having a minimal effect, relative to the data, on the final inference. Such a* reference analysis *might be required as an approximation to actual beliefs; more typically, it might be required as a limiting 'what if?' baseline in considering a range of prior to posterior analyses, or as a* default *option when there are insufficient resources for detailed elicitation of actual prior knowledge.*

... From the approach we adopt, it will be clear that the reference prior component of the analysis is simply a mathematical tool. It has considerable pragmatic importance in implementing a reference analysis, whose role and character will be precisely defined, but it is not a privileged, 'unique non-informative' or 'objective' prior."

The curious reader may take a look at the (Bernardo and Smith 1994) and references cited therein, as well as at Bernardo (1997).

9 Computational issues

The application of Bayesian ideas leads to computational problems, mostly related to the calculation of integrals for normalizing the posterior pdf and for obtaining credibility regions, or simply the moments of the distribution (and, hence, expectations, variances and covariances). The difficulties become challenging for problems involving many parameters. This is one of the reasons why Bayesian inference was abandoned at the beginning of the 20^{th} century in favor of simplified – and simplistic – methods. Indeed, the Bayesian renaissance over the past few decades is largely due to the emergence of new numerical methods and the dramatic increases in computational power, along with clarifying work on the foundations of the theory.

9.1 Overview of approximate computational strategies

In previous sections we have already seen some 'tricks' for simplifying the calculations. The main topic of this section will be an introduction to Monte Carlo (MC). But, before doing that, we think it is important to summarize the various 'tricks' here. Much specialized literature is available on several aspects of computation in statistics. For an excellent review paper on the subject see (Smith 1991).

Conjugate priors

We discussed this topic in Sect. 8.3, giving a couple of typical simple examples and references for a more detailed list of famous conjugate distributions. We want to remark here that a conjugate prior is a special case of the class of priors that simplify the calculation of the posterior (the uniform prior is the simplest of this kind of prior).

Gaussian approximation

For reasons that are connected with the central limit theorem, when there is a large amount of consistent data the posterior tends to be Gaussian, practically independently of the exact shape of the prior. The (multi-variate) Gaussian approximation, which we encountered in Sect. 5.10, has an important role for applications, either as a reasonable approximation of the 'true' posterior, or as a starting point for searching for a more accurate description of it. We also saw that in the case of practically flat priors this method recovers the well-known minimum chi-square or maximum likelihood methods.

Numerical integration

In the case of low dimensional problems, standard numerical integration using either scientific library functions or the interactive tools of modern computer packages provide an easy solution to many problems (thanks also to the graphical capabilities of modern programs which allow the shape of the posterior to be inspected and the best calculation strategy decided upon). This is a vast and growing subject, into we cannot enter in any depth here, but we assume the reader is familiar with some of these programs or packages.

Monte Carlo methods

Monte Carlo methodology is a science in itself and it is way beyond our remit to provide an exhaustive introduction to it here. Nevertheless, we would like to introduce briefly some 'modern' (though the seminal work is already half a century old) methods which are becoming extremely popular and are often associated with Bayesian analysis, the so called *Markov Chain Monte Carlo* (*MCMC*) methods.

9.2 Monte Carlo methods

9.2.1 Estimating expectations by sampling

The easy part of the Bayesian approach is to write down the un-normalized distribution of the parameters (Sect. 5.10), given the prior and the likelihood. This is simply $\tilde{p}(\boldsymbol{\theta}\,|\,\boldsymbol{d}, I) = p(\boldsymbol{d}\,|\,\boldsymbol{\theta}, I)\, p(\boldsymbol{\theta}\,|\,I)$. The difficult task is to normalize this function and to calculate all expectations in which we are interested, such as expected values, variances, covariances and other moments. We might also want to get marginal distributions, credibility intervals (or hypervolumes) and so on. As is well-known, if we were able to *sample* the posterior (even the un-normalized one), i.e. to generate points of the parameter space according to their probability, we would have solved the problem, at least approximately. For example, the one-dimensional histogram of parameter θ_i would represent its marginal and would allow the calculation of $\mathrm{E}(\theta_i) \approx \langle\theta_i\rangle$, $\sigma(\theta_i) \approx \langle\theta_i^2\rangle - \langle\theta_i\rangle^2$ and of probability intervals ($\langle\cdot\rangle$ in the previous formulae stand for arithmetic averages of the MC sample).

Let us consider the probability function $p(\boldsymbol{x})$ of the discrete variables $\boldsymbol{x}$ and a function $f(\boldsymbol{x})$ of which we want to evaluate the expectation over the distribution $p(\boldsymbol{x})$. Extending the one-dimensional formula of Tab. 1 to n dimension we have

$$\begin{aligned} \mathrm{E}[f(\boldsymbol{x})] &= \sum_{x_1}\cdots\sum_{x_n} f(x_1,\ldots,x_n)\, p(x_1,\ldots,x_n) && (109)\\ &= \sum_i f(\boldsymbol{x}_i)\, p(\boldsymbol{x}_i)\,, && (110)\end{aligned}$$

where the summation in Eq. (109) is over the components, while the summation in Eq. (110) is over possible points in the n-dimensional space of the variables. The result is the same.

If we are able to sample a large number of points N according to the probability function $p(\boldsymbol{x})$, we expect each point to be generated m_i times. The average $\langle f(\boldsymbol{x})\rangle$, calculated from the sample as

$$\langle f(\boldsymbol{x})\rangle = \frac{1}{N}\sum_t f(\boldsymbol{x}_t)\,, \tag{111}$$

(in which the index is named t as a reminder that this is a sum over a 'time' sequence) can also be rewritten as

$$\langle f(\boldsymbol{x})\rangle = \sum_i f(\boldsymbol{x}_i)\,\frac{m_i}{N}\,, \tag{112}$$

just grouping together the outcomes giving the same $\boldsymbol{x}_i$. For a very large N, the ratios m_i/N are expected to be 'very close' to $p(\boldsymbol{x}_i)$ (Bernoulli's theorem), and thus $\langle f(\boldsymbol{x})\rangle$ becomes a good approximation of $\mathrm{E}[f(\boldsymbol{x})]$. In fact, this approximation can be good (within tolerable errors) even if not all m_i are large and, indeed, even if many of them are null. Moreover, the same procedure can be extended to the continuum, in which case 'all points' (∞^n) can never be sampled.

For simple distributions there are well-known standard techniques for generating pseudo-random numbers starting from pseudo-random numbers distributed uniformly between 0 and 1 (computer libraries are available for sampling points according to the most common distributions). We shall not enter into these basic techniques, but will concentrate instead on the calculation of expectations in more complicated cases.

9.2.2 Rejection sampling

Let us assume we are able to generate points according to some function $g(\boldsymbol{x})$, such that, given a constant c, $p(\boldsymbol{x}) \leq c g(\boldsymbol{x})$. We generate $\boldsymbol{x}^*$ according to $g(\boldsymbol{x})$ and decide to accept it with probability $p(\boldsymbol{x}^*)/cg(\boldsymbol{x}^*)$ (i.e. we extract another random number between 0 and 1 and accept the point if this number is below that ratio). It is easy to show that this procedure reshapes $g(\boldsymbol{x})$ to $p(\boldsymbol{x})$ and that it does not depend on the absolute normalization of $p(\boldsymbol{x})$ (any normalization constant can be absorbed in the multiplicative constant c). A trivial choice of $g(\boldsymbol{x})$, especially for simple one-dimensional problems, is a uniform distribution (this variation is known as the *hit or miss* method), though clearly it can be very inefficient.

9.2.3 Importance sampling

In this method, too, we start from an 'easy' function $g(\boldsymbol{x})$, which 'we hope' will approximate $p(\boldsymbol{x})$ of interest, of which in fact we know only its un-normalized expression $\tilde{p}(\boldsymbol{x})$. However, there is no requirement about how $g(\boldsymbol{x})$ approximates $p(\boldsymbol{x})$ (but the goodness of approximation will have an impact on the efficacy of the method), apart from the condition that $g(\boldsymbol{x}_i)$ must be positive wherever $p(\boldsymbol{x}_i)$ is positive .

The function $g(\boldsymbol{x})$ can be used in the calculation of $\mathrm{E}[f(\boldsymbol{x})]$, if we notice that $\mathrm{E}[f(\boldsymbol{x})]$ can be rewritten as follows:

$$\begin{aligned}
\mathrm{E}[f(\boldsymbol{x})] &= \frac{\int f(\boldsymbol{x})\, \tilde{p}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}}{\int \tilde{p}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}} && (113)\\
&= \frac{\int f(\boldsymbol{x})\, [\tilde{p}(\boldsymbol{x})/g(\boldsymbol{x})]\, g(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}}{\int [\tilde{p}(\boldsymbol{x})/g(\boldsymbol{x})]\, g(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}} && (114)\\
&= \frac{\mathrm{E}_g\left[f(\boldsymbol{x})\, \tilde{p}(\boldsymbol{x})/g(\boldsymbol{x})\right]}{\mathrm{E}_g\left[\tilde{p}(\boldsymbol{x})/g(\boldsymbol{x})\right]}\,, && (115)
\end{aligned}$$

where the the symbol E_g is a reminder that the expectation is calculated over the distribution $g(\boldsymbol{x})$. Finally, the strategy can be implemented in the Monte Carlo using Eq. (111) for the two expectations:

$$\mathrm{E}[f(\boldsymbol{x})] \approx \frac{\sum_t f(\boldsymbol{x}_t)\, \tilde{p}(\boldsymbol{x}_t)/g(\boldsymbol{x}_t)}{\sum_t \tilde{p}(\boldsymbol{x}_t)/g(\boldsymbol{x}_t)}\,. \tag{116}$$

From the same sample it is also possible to evaluate the normalization constant, given by the denominator of Eq. (113), i.e.

$$Z = \int \tilde{p}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \approx \frac{1}{N} \sum_t \frac{\tilde{p}(\boldsymbol{x}_t)}{g(\boldsymbol{x}_t)}\,. \tag{117}$$

The computation of this quantity is particularly important when we are dealing with model comparison and Z has the meaning of 'evidence' (Sect. 7).

It easily to see that the method works well if $g(\boldsymbol{x})$ overlaps well with $p(\boldsymbol{x})$. Thus, a proper choice of $g(\boldsymbol{x})$ can be made by studying where the probability mass of $p(\boldsymbol{x})$ is concentrated (for example finding the mode of the distribution in a numerical way). Often a Gaussian function is used for $g(\boldsymbol{x})$, with parameters chosen to approximate $p(\boldsymbol{x})$ in the proximity of the mode, as described in Sect. 5.10. In other cases, other functions can be used which have more pronounced tails, like t-Student or Cauchy distributions. Special techniques, into which we cannot enter here, allow n independent random numbers to be generated and, subsequently, by proper rotations, turned into other numbers which have a correlation matrix equal to that of the multi-dimensional Gaussian which approximates $p(\boldsymbol{x})$.

Note, finally, that, contrary to the rejection sampling, importance sampling is not suitable for generate samples of 'unweighted events', such as those routinely used in the planning and the analysis of many kind experiments, especially particle physics experiments.

9.2.4 Metropolis algorithm

A different class of Monte Carlo methods is based on Markov chains and is known as Markov Chain Monte Carlo. The basic difference from the methods described above is that the sequence of generated points takes a kind of random walk in parameter space, instead of each point being generated, one independently from another. Moreover, the probability of jumping from one point to an other depends only on the last point and not on the entire previous history (this is the peculiar property of a Markov chain). There are several MCMC algorithms. One of the most popular and simple algorithms, applicable to a wide class of problems, is the *Metropolis* algorithm (Metropolis *et al* 1953). One starts from an arbitrary point $\boldsymbol{x}_0$ and generates the sequence by repeating the following cycle, with $\boldsymbol{x}_t$ being the previously selected point at each iteration:

1. Select a new trial point $\boldsymbol{x}^*$, chosen according to a symmetric *proposal* pdf $q(\boldsymbol{\theta}^* \,|\, \boldsymbol{\theta}_t)$.

2. Calculate the *acceptance probability*

$$A(\boldsymbol{x}^* \,|\, \boldsymbol{x}_t) = \min\left[1, \frac{\tilde{p}(\boldsymbol{\theta}^*)}{\tilde{p}(\boldsymbol{\theta}_t)}\right] . \tag{118}$$

3. Accept $\boldsymbol{x}^*$ with probability $A(\boldsymbol{x}_t, \boldsymbol{x}^*)$, i.e.

   - if $\tilde{p}(\boldsymbol{\theta}^*) \geq \tilde{p}(\boldsymbol{\theta}_t)$, then accept $\boldsymbol{x}^*$;
   - if $\tilde{p}(\boldsymbol{\theta}^*) < \tilde{p}(\boldsymbol{\theta}_t)$, extract a uniform random number between 0 and 1 and accept $\boldsymbol{x}^*$ if the random number is less then $\tilde{p}(\boldsymbol{\theta}^*)/\tilde{p}(\boldsymbol{\theta}_t)$.

   If the point is accepted, then $\boldsymbol{x}_{t+1} = \boldsymbol{x}^*$. Otherwise $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$

This algorithm allows a jump $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t+1}$ with probability $T(\boldsymbol{x}_{t+1} \,|\, \boldsymbol{x}_t)$ (the *transition kernel*) equal to $A(\boldsymbol{x}^* \,|\, \boldsymbol{x}_t)\, q(\boldsymbol{\theta}^* \,|\, \boldsymbol{\theta}_t)$. The algorithm has a stationary asymptotic distribution equal to $p(\boldsymbol{x})$ (i.e. the chain is *ergodic*) and ensures *detailed balance*:

$$p(\boldsymbol{x}_{t+1})\, T(\boldsymbol{x}_t \,|\, \boldsymbol{x}_{t+1}) = p(\boldsymbol{x}_t)\, T(\boldsymbol{x}_{t+1} \,|\, \boldsymbol{x}_t) . \tag{119}$$

By construction, the algorithm does not depend on the normalization constant, since what matters is the ratio of the pdf's. The variation of the algorithm in which the proposal pdf $q()$ is not symmetric is due to Hasting (1970) and for this reason the algorithm is often also called Metropolis-Hasting. Moreover, what has been described here is the *global* Metropolis algorithm, in contrast to the *local* one, in which a cycle affects only one component of $\boldsymbol{x}$.

The fact that this algorithm belongs to the class of MCMC gives rise to two problems. First, each point in the chain has some correlation with the points which immediately preceded it, and usually the chain moves slowly (and irregularly) from one region in the variable space to another (note also that, if a proposed point is not accepted, the chain keep the same position in the next step, and this circumstance can happen several times consecutively). Second, the initial part of the sequence is strongly influenced by the arbitrary starting point. Therefore, it is necessary to remove the initial part of the chain.

Using an MCMC for a complex problem is not an automatic procedure and some tuning is needed. One of the important things to choose with care is the proposal function. If too small jumps are proposed, the chain moves too slowly and, can even remain trapped in a subregion and never sample the rest of the parameter space, if the probability distribution is defined over disconnected regions. If too large steps are proposed, the proposed points could often fall in very low probability regions and not be accepted, in which case the chain remains stuck in a point for many cycles.

For an interesting, insightful introduction to principles and applications of MCMC see (Kass *et al* 1998). A nice tutorial is given by (Hanson 2000). A recent application of Bayesian methods in cosmology, which uses MCMC and contains a pedagogical introduction too, can be found in (Lewis and Bridle 2002). For a good treatise, freely available on the web, (Neel 1993) is recommended. The reader will find that MCMC techniques are used to solve complex problems graphically represented in terms of *Bayesian networks* (also known as *belief networks* or simply *probabilistic network*). This subject, which has revolutionized the way of thinking artificial intelligence and the uncertainty issues related to it, does beyond the purpose of this article. The interested reader can find more information in (Pearl 1988, BUGS 1996, Cowell *et al* 1999 and Cozman 2001) and references therein.

10 Conclusions

The gain in popularity Bayesian methods have enjoyed in recent years is due to various conceptual and practical advantages they have over other approaches, among which are:

- the recovery of the intuitive idea of probability as a valid concept for treating scientific problems;
- the simplicity and naturalness of the basic tool;
- the capability of combining prior knowledge and experimental information;
- the property permitting automatic updating as soon as new information becomes available;
- the transparency of the methods, which allow the different assumptions upon which an inference may depend to be checked and changed;
- the high degree of awareness the methods give to the user.

In this article we have seen how to build a theory of uncertainty in measurement as a straightforward application of the basic Bayesian ideas, without unnecessary principles or *ad hoc* prescriptions. In particular, the uncertainty due to systematic errors can be treated in a consistent and powerful way.

Providing an exact solution for inferential problems can easily lead to computational difficulties. We have seen several ways to overcome such difficulties, either by using suitable approximations, or by using modern computational methods. In particular, it has been shown that the approximate solution often coincides with a 'conventional' method, but only under well defined conditions. Thus, for example, minimum χ^2 formulae can be used, with a Bayesian spirit and with a natural interpretation of the results, in all those routine cases in which the analyst considers as reasonable the conditions of their validity.

A variety of examples of applications have been shown, or mentioned, in this paper. Nevertheless, the aim of the author was not to provide a complete review of Bayesian methods and applications, but rather to introduce those Bayesian ideas that could be of help in understanding more specialized literature. Compendia of the Bayesian theory are given in

(Bernardo and Smith 1994, O'Hagan A 1994 and Robert 2001). Classic, influential books are (Jeffreys 1961, de Finetti 1974, Jaynes 1998). Among the many books introducing Bayesian methods, (Sivia 1996) is particularly suitable for physicists. Other recommended texts which treat general aspects of data analysis are (Box and Tiao 1973, Bretthorst 1988, Lee 1989, Gelman *et al* 1995, Cowell *et al* 1999, Denison *et al* 2002, Press 2002). More specific applications can be found in the proceedings of the conference series and several web sites. Some useful starting points for web navigation are given:

ISBA book list	http://www.bayesian.org/books/books.html
UAI proceedings	http://www2.sis.pitt.edu/ dsl/UAI/uai.html
BIPS	http://astrosun.tn.cornell.edu/staff/loredo/bayes/
BLIP	http://www.ar-tiste.com/blip.html
IPP Bayesian analysis group	http://www.ipp.mpg.de/OP/Datenanalyse/
Valencia meetings	http://www.uv.es/~bernardo/valenciam.html
Maximum Entropy resources	http://omega.albany.edu:8008/maxent.html
MCMC preprint service	http://www.statslab.cam.ac.uk/~mcmc/

I am indebted to Volker Dose and Ken Hanson for extensive discussions concerning the contents of this paper, as well as for substantial editorial help. The manuscript has also benefited from comments by Tom Loredo.

References

Astone P *et al* 2002 Search for correlation between GRB's detected by BeppoSAX and gravitational wave detectors EXPLORER and NAUTILUS *Phys. Rev.* **66** 102002.

Astone P and D'Agostini G 1999 Inferring the intensity of Poisson processes at limit of detector sensitivity (with a case study on gravitational wave burst search) *CERN-EP/99-126*

Astone P, D'Agostini G and D'Antonio 2003 Bayesian model comparison applied to the Explorer-Nautilus 2001 coincidence data, arXiv:gr-qc/0304096

Babu G J and Feigelson E D 1992 eds *Statistical Challenges in Modern Astronomy I* (New York: Springer)

Babu G J and Feigelson E D 1997 eds *Statistical Challenges in Modern Astronomy II* (New York: Springer)

Berger J O and Jefferys W H 1992 Sharpening Ockham's razor on a Bayesian strop *Am. Scientist* **89** 64–72 and *J. Ital. Stat. Soc.* **1** 17

Bernardo J M 1999 ed *Bayesian Methods in the Sciences*, special issue of *Rev. Acad. Cien. Madrid*, **93**(3)

Bernardo J M, J O Berger, A P Dawid and A F M Smith 1999 eds *Bayesian Statistics 6* (Oxford: Oxford University)

Bernardo J M and Smith F M 1994 *Bayesian Theory* (Chichester: John Wiley & Sons)

Bernardo J M 1997 Non-informative priors do not exist *J. Stat. Planning and Infer.* **65** 159

Bontekoe T R, Koper E and Kester D J M 1994 Pyramid maximum entropy images of IRAS survey data *Astron. Astrophys.* **284** 1037

Bouman C A and Sauer K 1993 A generalized Gaussian image model for edge-preserving MAP estimation *IEEE Trans. on Image Processing* **2** 296

Box G E P and Tiao G C 1973 *Bayesian inference in statistical analysis* (Chichester: J. Wiley & Sons)

Bretthorst G L 1988 *Bayesian spectrum analysis and parameter estimation* (Berlin: Springer)

BUGS 1996 http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

Buck B and Macauly V A eds 1991 *Maximum Entropy in action*, (Oxford: Oxford University Press)

Ciuchini M *et al* 2001 2000 CKM-Triangle Analysis: A critical review with updated experimental inputs and theoretical parameters *J. High Energy Phys.* **0107** 013

Coletti G and Scozzafava R 2002 *Probabilistic logic in a coherent setting*", (Dordrecht: Kluwer Academic)

Cousins R D 1995 Why isn't every physicist a Bayesian? *Am. J. Phys.* **63** 398

Cowell R G, Dawid A P, Lauritzen S L and Spiegelhalter D J 1999 *Probabilistic Networks and Expert Systems*, (New York: Springer)

Cox R T 1946 Probability, Frequency and Reasonable Expectation *Am. J. Phys.* **14** 1

Cozman F B 2001 JavaBayes version 0.346 – Bayesian networks in Java http://www-2.cs.cmu.edu/~javabayes/Home/

Cunningham G S, Hanson K M and Battle X L 1998 Three-dimensional reconstructions from low-count SPECT data using deformable models *Opt. Expr.* **2** 227

D'Agostini G 1999a Bayesian Reasoning versus Conventional Statistics in High Energy Physics *Maximum Entropy and Bayesian Methods* ed von der Linden W *et al* (Dordrecht: Kluwer Academic)

D'Agostini G 1999b Sceptical combination of experimental results: General considerations and application to epsilon-prime/epsilon *CERN-EP/99-139*

D'Agostini G 1999c *Bayesian reasoning in high-energy physics: principles and applications CERN Report 99-03* (an extended version of this report is going to be published as *Bayesian reasoning in data analysis – A critical introduction* by World Scientific Publishing)

D'Agostini G 1999d Teaching statistics in the physics curriculum: Unifying and clarifying role of subjective probability *Am. J. Phys.* **67** 1260

D'Agostini G 1999e Overcoming prior Anxiety, *Bayesian Methods in the Sciences* ed J. M. Bernardo; special issue of *Rev. Acad. Cien. Madrid* **93**(3), 311

D'Agostini G 2000 Confidence limits: what is the problem? Is there *the* solution? *CERN Report 2000-005* ed James F and Lyons L (Geneva: CERN) 3

D'Agostini G 2002 Minimum bias legacy of search results 2002 *Nucl Phys Proc Suppl* **109** 148

D'Agostini G and Degrassi G 1999 Constraints on the Higgs Boson Mass from Direct Searches and Precision Measurements *Eur. Phys. J.* **C10** 663

D'Agostini G and Raso M Uncertainties due to imperfect knowledge of systematic effects: general considerations and approximate formulae *CERN-EP/2000-026*

de Finetti B 1974 *Theory of Probability* (Chichester: J. Wiley & Sons)

Denison D G T, Holmes C C, Mallick B K and Smith A F M 2002 *Bayesian methods for nonlinear classification and regression* (Cichester: J. Wiley & Sons)

DIN (Deutsches Institut für Normung) 1996 *Grundlagen der Messtechnik - Teil 3: Auswertung von Messungen einer einzelnen Messgre, Messunsicherheit* DIN 1319-3 (Berlin: Beuth Verlag)

DIN (Deutsches Institut für Normung) 1999 Grundlagen der Messtechnik - Teil 4: Auswertung von Messungen, Messunsicherheit DIN 1319-4 (Berlin: Beuth Verlag)

Dose V 2002 Bayes in five days, CIPS, MPI für Plasmaphysik, Garching, Germany, Reprint 83, May 2002

Dose V and von der Linden W 1999 Outlier tolerant parameter estimation *Maximum Entropy and Bayesian Methods* ed von der Linden W *et al* (Dordrecht: Kluwer Academic) 47

Efron B 1986a Why isn't everyone a Bayesian? *Am. Stat.* **40** 1

Efron B 1986b reply to Zellner 1986 *Am. Stat.* **40** 331

Fischer R, Mayer M, von der Linden W and Dose V 1997 Enhancement of the energy resolution in ion-beam experiments with the maximum-entropy method *Phys. Rev. E* **55** 6667

Fischer R, Mayer M, von der Linden W and Dose V 1998 Energy resolution enhancement in ion beam experiments with Bayesian probability theory *Nucl. Instr. Meth.* **136-138** 1140

Fischer R, Hanson K M, Dose V and von der Linden W 2000 Background estimation in experimental spectra *Phys. Rev.* **E61** 1152

Fröhner F H 2000 *Evaluation and Analysis of Nuclear Resonance Data* JEFF Report 18 (Paris: OECD Publications)

Gelman A, Carlin J B, Stern H S and Rubin D B 1995 *Bayesian Data Analysis* (London: Chapman and Hall)

Glimm J and Sharp D H 1999 Prediction and the quantification of uncertainty *Physica D* **133** 152

Gregory P C and Loredo T J 1992 A new method for the detection of a periodic signal of unknown shape and period *Astr. J.* **398** 146

Gregory P C and Loredo T J 1996 Bayesian periodic signal detection II – Bayesian periodic signal detection: analysis of ROSAT observations of PSR 0540-693 *Astr. J.* **473** 1059

Gregory P C 1999 Bayesian periodic signal detection I – analysis of 20 years of radio flux measurements of the x-ray binary LS I $+61°303$ *Astr. J.* **520** 361

Gubernatis J E, Jarrell M, Silver R N and Sivia D S 1991 Quantum Monte-Carlo simulations and maximum-entropy: dynamics from imaginary-time data *Phys. Rev. B* **44** 6011

Hanson K M 1993 Introduction to Bayesian image analysis *Medical Imaging: Image Processing* Loew M H ed *Proc. SPIE* **1898** 716

Hanson K M 2000 Tutorial on Markov Chain Monte Carlo, *http://public.lanl.gov/kmh/talks/maxent00b.pdf*

Hasting W K 1970 Monte Carlo sampling methods using Markov chains and their applications *Biometrica* **57** 97

Higdon D M and Yamamoto S Y 2001 Estimation of the head sensitivity function in scanning magnetoresistance microscopy *J. Amer. Stat. Assoc.* **96** 785

Hobson M P, Bridle S L and Lahav 2002 Combining cosmological datasets: hyperparameters and Bayesian evidence arXiv:astro-ph/0203259

Howson C and Urbach P 1993 *Scientific reasoning — the Bayesian approach* (Chicago and La Salle: Open Court)

ISO (International Organization for Standardization) 1993 *Guide to the Expression of Uncertainty in Measurement* (Geneva: ISO)

Jaynes E T 1957a Information theory and statistical mechanics *Phys. Rev.* **106** 620

Jaynes E T 1957b Information theory and statistical mechanics II *Phys. Rev.* **108** 171

Jaynes E T 1968 Prior probabilities *IEEE Trans. Syst. Cybern.* **SSC-4** 227, reprinted in (Jaynes 1983)

Jaynes E T 1973 The well-posed problem *Found. Phys.* **3** 477, reprinted in (Jaynes 1983)

Jaynes E T 1983 *Papers on Probability, Statistics and Statistical Physics* ed Harper W L and Hooker C A (Dordrecht: Reidel)

Jaynes E T 1998 *http://omega.albany.edu:8008/JaynesBook.html*

Jeffreys H 1961 *Theory of Probability* (Oxford: Oxford University)

John M V and Narlikar J V 2002 Comparison of cosmological models using Bayesian theory *Phys. Rev.* **D65** 43506

Kadane J B and Schum D A 1996 *A Probabilistic Analysis of the Sacco and Vanzetti Evidence* (Chichester: Wiley and Sons)

Kalman R E 1960 A new approach to linear filtering and prediction problems *Trans. ASME Journal of Casic Engineering* **82** 35

Kass R E, Carlin B P, Gelman A and Neal R M 1998 Markov Chain Monte Carlo in practice: A roundtable discussion *Am. Stat.* **52** 93

Lad F 1996 *Operational Subjective Statistical Methods – a Mathematical, Philosophical, and Historical Introduction* (Chichester:J. Wiley & Sons)

Lee P M 1989 *Bayesian statistics - an introduction* (Chichester:J. Wiley & Sons)

Lewis A and Bridle S 2002 Cosmological parameters from CMB and other data: a Monte-Carlo approach *Phys.Rev.* **D66** 103511

von der Linden W 1995 Maximum-entropy data analysis *Appl. Phys.* **A60** 155

von der Linden W, Dose V and Fischer R 1996b Spline-based adaptive resolution image reconstruction *Proceedings of the 1996 Maximum Entropy Conference* ed Sears M *et al* (Port Elizabeth: N.M.B. Printers) 154

Lindley D V 1986 Discussion to Efron 1986a *Am. Stat.* **40** 6

Loredo T J 1990 *Maximum Entropy and Bayesian Methods* ed Fougére P F (Dordrecht: Kluwer Academic) 81

Loredo T J and Lamb D Q 2002 Bayesian analysis of neutrinos observed from supernova SN 1987A *Phys. Rev.* **D65** 063002

Malakoff D 1999 Bayes Offers a 'New' Way to Make Sense of Numbers *Science* **286** 1460

Maybaeck P S 1979 *Stochastic models, estimation and control*, Vol. 1 (New York: Academic Press).

Metropolis H, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 Equations of state calculations by fast computing machines *Journal of Chemical Physics* **21** 1087

von Mises R 1957 *Probability, Statistics, and Truth* (St Leonards: Allen and Unwin); reprinted in 1987 by Dover

Neal R M 1993 *Probabilistic inference using Markov chain Monte Carlo methods* (Toronto: Technical Report CRG-TR-93-1)

O'Hagan A 1994 *Kendall's Advanced Theory of Statistics: Vol. 2B - Bayesian Inference* (New York: Halsted)

Pearl J 1988 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (San Mateo: Morgan Kaufmann)

Press W H 1997 Understanding data better with Bayesian and global statistical methods *Unsolved problems in astrophysics*49–60 ed Bahcall J N and Ostriker J P (Princeton: Princeton University) 49

Press S J 2002 *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications* 2nd Edition (Chichester: John Wiley & Sons)

Robert C P 2001 *The Bayesian Choice* (New York: Springer)

Saquib S S, Hanson K M, and Cunningham G S 1997 Model-based image reconstruction from time-resolved diffusion data *Proc. SPIE* **3034** 369

Schrödinger E 1947a The Foundation of the Theory of Probability – I *Proc. R. Irish Acad.* **51A** 51; reprinted in *Collected papers* Vol. 1 (Vienna 1984: Austrian Academy of Science) 463

Schrödinger E 1947b The Foundation of the Theory of Probability – II *Proc. R. Irish Acad.* **51A** 141; reprinted in *Collected papers* Vol. 1 (Vienna 1984: Austrian Academy of Science) 479

Sivia D S 1997 *Data Analysis – a Bayesian Tutorial* (Oxford: Clarendon)

Skilling J 1992 Quantified maximum entropy *Int. Spectr. Lab.* **2** 4

Smith A F M 1991 Bayesian numerical analysis *Phyl. Trans. R. Soc. London* **337** 369

Taylor B N and Kuyatt C E 1994 *Guidelines for Evaluating and Expressing Uncertainty of NIST Measurement Results* (Gaithersburg: NIST Technical Note 1297); available on line at *http://physics.nist.gov/*

Tribus M 1969 *Rational Descriptions, Decisions, and Designs* (Elmsford: Pergamon)

Welch G and Bishop G 2002 An introduction to Kalman filter (*http://www.cs.unc.edu/~welch/kalman/*)

Zech G 2002 Frequentist and Bayesian confidence limits *Eur. Phys. J. direct* **C12** 1

Zellner A 1986 Bayesian solution to a problem posed by Efron *Am. Stat.* **40** 330