3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Part I

The essentials

'La théorie des probabilités n'est que le bon sens reduit au calcul.'

— Pierre Simon, Marquis de Laplace

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

This page intentionally left blank

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

1 The basics

'There are three kinds of lies: lies, damned lies and statistics.'

Mark Twain (1924) probably had politicians in mind when he reiterated Disraeli's famous remark. Scientists, we hope, would never use data in such a selective manner to suit their own ends. But, alas, the analysis of data is often the source of some exasperation even in an academic context. On hearing comments like 'the result of this experiment was inconclusive, so we had to use statistics', we are frequently left wondering as to what strange tricks have been played on the data.

The sense of unease which many of us have towards the subject of statistics is largely a reflection of the inadequacies of the 'cook-book' approach to data analysis that we are taught as undergraduates. Rather than being offered a few clear principles, we are usually presented with a maze of tests and procedures; while most seem to be intuitively reasonable individually, their interrelations are not obvious. This apparent lack of a coherent rationale leads to considerable apprehension because we have little feeling for which test to use or, more importantly, why.

Fortunately, data analysis does not have to be like this! A more unified and logical approach to the whole subject is provided by the probability formulations of Bayes and Laplace. Bayes' ideas (published in 1763) were used very successfully by Laplace (1812), but were then allegedly discredited and largely forgotten until they were rediscovered by Jeffreys (1939). In more recent times they have been expounded by Jaynes (1983, 2003) and others. This book is intended to be an introductory tutorial to the Bayesian approach, including modern developments such as maximum entropy.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

1.1 Introduction: deductive logic versus plausible reasoning

Let us begin by trying to get a general feel for the nature of the problem. A schematic representation of deductive logic is shown in Fig. 1.1(a): given a cause, we can work out its consequences. The sort of reasoning used in pure mathematics is of this type: that is to say, we can derive many complicated and useful results as the logical consequence of a few well-defined axioms. Everyday games of chance also fall into this category. For example, if we are told that a fair coin is to be flipped ten times, we can calculate the chances that all ten tosses will produce heads, or that there will be nine heads and one tail, and so on.

Most scientists, however, face the reverse of the above situation: Given that certain effects have been observed, what is (are) the underlying cause(s)? To take a simple example, suppose that ten flips of a coin yielded seven heads: Is it a fair coin or a biased one? This type of question has to do with inductive logic, or plausible reasoning, and is illustrated schematically in Fig. 1.1(b); the greater complexity of this diagram is

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

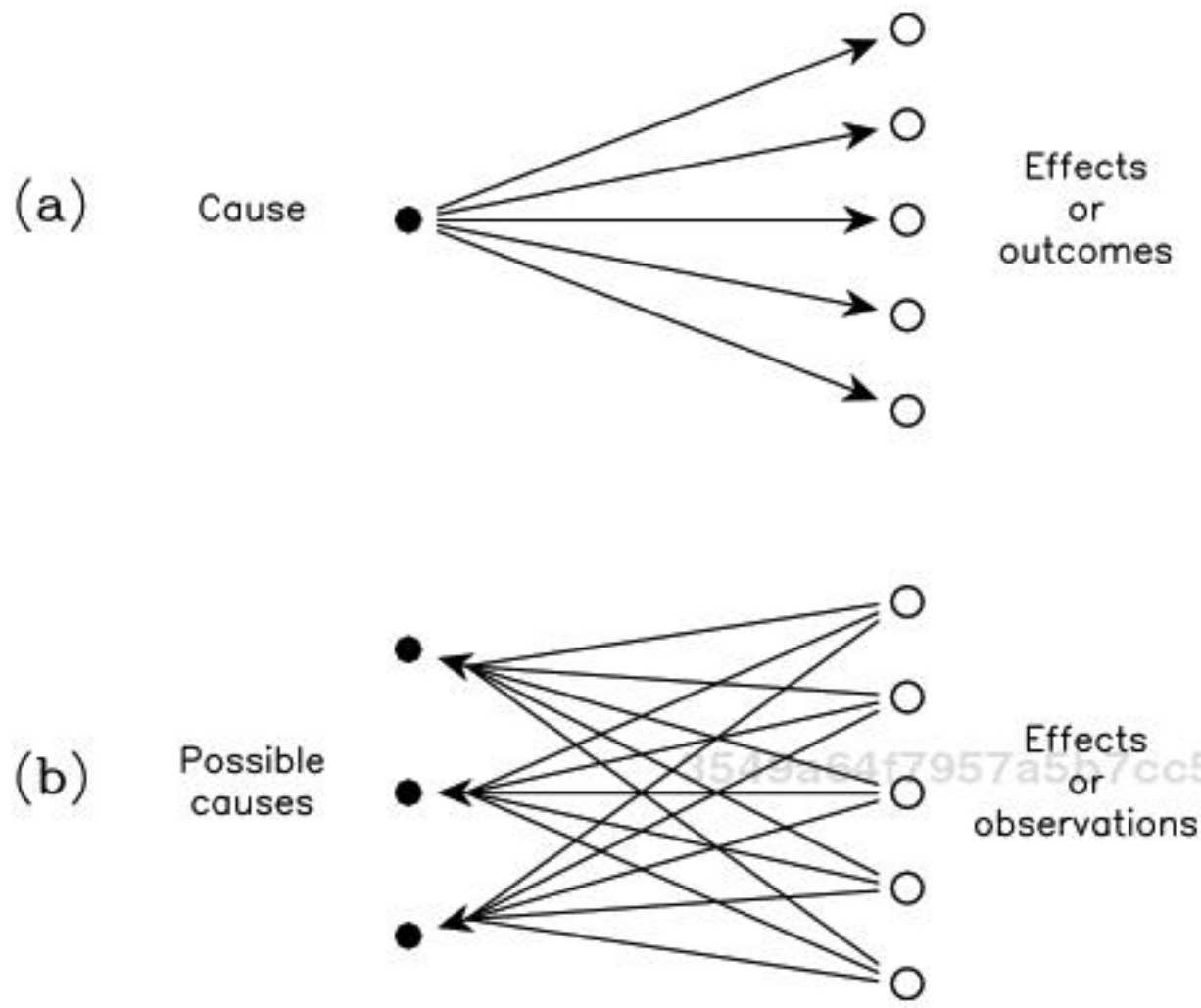


Fig. 1.1 A schematic representation of (a) deductive logic, or pure mathematics, and (b) plausible reasoning, or inductive logic.

designed to indicate that it is a much harder problem. The most we can hope to do is to make the best inference based on the experimental data and any prior knowledge that we have available, reserving the right to revise our position if new information comes to light. Around 500 BC, Herodotus said much the same thing: 'A decision was wise, even though it led to disastrous consequences, if the evidence at hand indicated it was the best one to make; and a decision was foolish, even though it led to the happiest possible consequences, if it was unreasonable to expect those consequences.'

Even though plausible reasoning is rather open-ended, are there any general quantitative rules which apply for such inductive logic? After all, this issue is central to data analysis.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

1.2 Probability: Cox and the rules for consistent reasoning

In 1946, Richard Cox pondered the quantitative rules necessary for logical and consistent reasoning. He started by considering how we might express our relative beliefs in the truth of various propositions. For example: (a) it will rain tomorrow; (b) King Harold died by being hit in the eye with an arrow at the battle of Hastings in 1066 AD; (c) this is a fair coin; (d) this coin is twice as likely to come up heads as tails; and so on. The minimum requirement for expressing our relative beliefs in the truth of these propositions in a consistent fashion is that we rank them in a *transitive* manner. In other words, if we believe (a) more than (b), and (b) more than (c), then we must necessarily believe (a) more than (c); if this were not so, we would continue to argue in circles. Such a transitive ranking can easily be obtained by assigning a *real* number to each of the propositions in a manner so that the larger the numerical value associated with a proposition, the more we believe it.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Cox actually took this much for granted — as being obvious — and wondered what rules these numbers had to obey in order to satisfy some simple requirements of logical consistency. He began by making two assertions. The first is very straightforward: if we specify how much we believe that something is true, then we must have implicitly specified how much we believe it's false. He didn't assume any particular form for this relationship, but took it as being reasonable that one existed. The second assertion is slightly more complicated: if we first specify how much we believe that (proposition) Y is true, and then state how much we believe that X is true given that Y is true, then we must implicitly have specified how much we believe that both X and Y are true. Again, he only asserted that these quantities were related but did not specify how. To work out the actual form of the relationships, Cox used the rules of *Boolean logic*, ordinary algebra, and the constraint that if there were several different ways of using the same information then we should always arrive at the same conclusions irrespective of the particular analysis-path chosen. He found that this consistency could only be ensured if the real numbers we had attached to our beliefs in the various propositions obeyed the usual rules of probability theory:

$$\text{prob}(X|I) + \text{prob}(\overline{X}|I) = 1 \tag{1.1}$$

and

$$\text{prob}(X, Y|I) = \text{prob}(X|Y, I) \times \text{prob}(Y|I)\,, \tag{1.2}$$

with $0 = \text{prob}(\text{false})$ and $1 = \text{prob}(\text{true})$ defining certainty; a proof is given in Appendix B. Here $\overline{X}$ denotes the proposition that X is false, the vertical bar '$|$' means 'given' (so that all items to the right of this conditioning symbol are taken as being true) and the comma is read as the conjunction 'and'.

We have made the probabilities conditional on I, to denote the relevant background information at hand, because there is no such thing as an absolute probability. For example, the probability we assign to the proposition 'it will rain this afternoon' will depend on whether there are dark clouds or a clear blue sky in the morning; it will also be affected by whether or not we saw the weather forecast. Although the conditioning on I is often omitted in calculations, to reduce algebraic cluttering, we must never forget its existence. A failure to state explicitly all the relevant background information, and assumptions, is frequently the real cause of heated debates about data analysis.

Equation (1.1) is called the *sum rule*, and states that the probability that X is true plus the probability that X is false is equal to one. Equation (1.2) is called the *product rule*. It states that the probability that both X and Y are true is equal to the probability that X is true given that Y is true times the probability that Y is true (irrespective of X). We can change the description of probability, as when we multiply by 100 to obtain percentages, but we are not allowed to change the content of eqns (1.1) and (1.2). Probability calculus uses the unique scale in which the rules take the form of un-adorned addition and multiplication.

1.3 Corollaries: Bayes' theorem and marginalization

The sum and product rules of eqns (1.1) and (1.2) form the basic algebra of probability theory. Many other results can be derived from them. Amongst the most useful are two

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

known as *Bayes' theorem* and *marginalization*:

$$\text{prob}(X|Y,I) = \frac{\text{prob}(Y|X,I) \times \text{prob}(X|I)}{\text{prob}(Y|I)} \tag{1.3}$$

and

$$\text{prob}(X|I) = \int_{-\infty}^{+\infty} \text{prob}(X,Y|I)\,\mathrm{d}Y\,. \tag{1.4}$$

Bayes' theorem, or eqn (1.3), follows directly from the product rule. To see this, let's rewrite eqn (1.2) with X and Y *transposed* (or interchanged):

$$\text{prob}(Y,X|I) = \text{prob}(Y|X,I) \times \text{prob}(X|I)\,.$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Since the statement that 'Y and X are both true' is the same as 'X and Y are both true', so that $\text{prob}(Y,X|I) = \text{prob}(X,Y|I)$, the right-hand side of the above can be equated to that of eqn (1.2); hence, we obtain eqn (1.3). It is invaluable because it enables us to turn things around with respect to the conditioning symbol: it relates $\text{prob}(X|Y,I)$ to $\text{prob}(Y|X,I)$. The importance of this property to data analysis becomes apparent if we replace X and Y by *hypothesis* and *data*:

$$\text{prob}(hypothesis|data,I) \propto \text{prob}(data|hypothesis,I) \times \text{prob}(hypothesis|I)\,.$$

The power of Bayes' theorem lies in the fact that it relates the quantity of interest, the probability that the hypothesis is true given the data, to the term we have a better chance of being able to assign, the probability that we would have observed the measured data if the hypothesis was true.

The various terms in Bayes' theorem have formal names. The quantity on the far right, $\text{prob}(hypothesis|I)$, is called the *prior* probability; it represents our state of knowledge (or ignorance) about the truth of the hypothesis before we have analysed the current data. This is modified by the experimental measurements through the *likelihood* function, or $\text{prob}(data|hypothesis,I)$, and yields the *posterior* probability, $\text{prob}(hypothesis|data,I)$, representing our state of knowledge about the truth of the hypothesis in the light of the data. In a sense, Bayes' theorem encapsulates the process of learning. We should note, however, that the equality of eqn (1.3) has been replaced with a proportionality, because the term $\text{prob}(data|I)$ has been omitted. This is fine for many data analysis problems, such as those involving *parameter estimation*, since the missing denominator is simply a normalization constant (not depending explicitly on the hypothesis). In some situations, like *model selection*, this term plays a crucial rôle. For that reason, it is sometimes given the special name of *evidence*. This crisp single word captures the significance of the entity, as opposed to older names, such as *prior predictive* and *marginal likelihood*, which describe how it tends to be used or calculated. Such a central quantity ought to have a simple name, and 'evidence' has been assigned no other technical meaning (apart from as a colloquial synonym of data).

The marginalization equation, (1.4), should seem a bit peculiar: up to now Y has stood for a given proposition, so how can we integrate over it? Before we answer that

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

question, let us first consider the marginalization equation for our standard X and Y propositions. It would take the form

$$\text{prob}(X|I) = \text{prob}(X, Y|I) + \text{prob}\big(X, \overline{Y}|I\big)\,. \tag{1.5}$$

This can be derived by expanding $\text{prob}(X, Y|I)$ with the product rule of eqn (1.2):

$$\text{prob}(X, Y|I) = \text{prob}(Y, X|I) = \text{prob}(Y|X, I) \times \text{prob}(X|I)\,,$$

and adding a similar expression for $\text{prob}\big(X, \overline{Y}|I\big)$ to the left- and right-hand sides, respectively, to give

$$\text{prob}(X, Y|I) + \text{prob}\big(X, \overline{Y}|I\big) = \big[\text{prob}(Y|X, I) + \text{prob}\big(\overline{Y}|X, I\big)\big] \times \text{prob}(X|I)\,.$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Since eqn (1.1) ensures that the quantity in square brackets on the right is equal to unity, we obtain eqn (1.5). Stated verbally, eqn (1.5) says that the probability that X is true, irrespective of whether or not Y is true, is equal to the sum of the probability that both X and Y are true and the probability that X is true and Y is false.

Suppose that instead of having a proposition Y, and its negative counterpart $\overline{Y}$, we have a whole set of alternative possibilities: $Y_1, Y_2, \ldots, Y_M = \{Y_k\}$. For example, let's imagine that there are M (say five) candidates in a presidential election; then Y_1 could be the proposition that the first candidate will win, Y_2 the proposition that the second candidate will win, and so on. The probability that X is true, for example that unemployment will be lower in a year's time, irrespective of whoever becomes president, is then given by

$$\text{prob}(X|I) = \sum_{k=1}^{M} \text{prob}(X, Y_k|I)\,. \tag{1.6}$$

This is just a generalization of eqn (1.5), and can be derived in an analogous manner, by putting $\text{prob}(X, Y_k|I) = \text{prob}(Y_k|X, I) \times \text{prob}(X|I)$ in the right-hand side of eqn (1.6), as long as

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\sum_{k=1}^{M} \text{prob}(Y_k|X, I) = 1\,. \tag{1.7}$$

This *normalization* requirement is satisfied if the $\{Y_k\}$ form a *mutually exclusive* and *exhaustive* set of possibilities. That is to say, if one of the Y_k's is true then all the others must be false, but one of them has to be true.

The actual form of the marginalization equation in (1.4) applies when we go to the *continuum limit*. For example, when we consider an arbitrarily large number of propositions about the range in which (say) the Hubble constant H_0 might lie. As long as we choose the intervals in a contiguous fashion, and cover a big enough range of values for H_0, we will have a mutually exclusive and exhaustive set of possibilities. Equation (1.4) is then just a generalization of eqn (1.6), with $M \to \infty$, where we have used the usual shorthand notation of *calculus*. In this context, Y now represents the numerical value of a parameter of interest (such as H_0) and the integrand $\text{prob}(X, Y|I)$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

is technically a *probability density* function rather than a probability. Strictly speaking, therefore, we should denote it by a different symbol, such as $\mathrm{pdf}(X, Y|I)$, where

$$\mathrm{pdf}(X, Y{=}\, y\,|I) \;=\; \lim_{\delta y \to 0} \frac{\mathrm{prob}(X,\, y \leqslant Y < y + \delta y\,|I)}{\delta y}\,, \tag{1.8}$$

and the probability that the value of Y lies in a finite range between y_1 and y_2 (and X is also true) is given by

$$\mathrm{prob}(X,\, y_1 \leqslant Y < y_2\,|I) \;=\; \int_{y_1}^{y_2} \mathrm{pdf}(X, Y|I)\,\mathrm{d}Y\,. \tag{1.9}$$

Since 'pdf' is also a common abbreviation for *probability distribution* function, which can pertain to a discrete set of possibilities, we will simply use 'prob' for anything related to probabilities; this has the advantage of preserving a uniformity of notation between the continuous and discrete cases. Thus, in the continuum limit, the normalization condition of eqn (1.7) takes the form

$$\int_{-\infty}^{+\infty} \mathrm{prob}(Y|X, I)\,\mathrm{d}Y \;=\; 1\,. \tag{1.10}$$

Marginalization is a very powerful device in data analysis because it enables us to deal with *nuisance parameters*; that is, quantities which necessarily enter the analysis but are of no intrinsic interest. The unwanted background signal present in many experimental measurements, and instrumental parameters which are difficult to calibrate, are examples of nuisance parameters. Before going on to see how the rules of probability can be used to address data analysis problems, let's take a brief look at the history of the subject.

1.4 Some history: Bayes, Laplace and orthodox statistics

About three hundred years ago, people started to give serious thought to the question of how to reason in situations where it is not possible to argue with certainty. James Bernoulli (1713) was perhaps the first to articulate the problem, perceiving the difference between the deductive logic applicable to games of chance and the inductive logic required for everyday life. The open question for him was how the mechanics of the former might help to tackle the inference problems of the latter.

Reverend Thomas Bayes is credited with providing an answer to Bernoulli's question, in a paper published posthumously by a friend (1763). The present-day form of the theorem which bears his name is actually due to Laplace (1812). Not only did Laplace rediscover Bayes' theorem for himself, in far more clarity than did Bayes, he also put it to good use in solving problems in celestial mechanics, medical statistics and even jurisprudence. Despite Laplace's numerous successes, his development of probability theory was rejected by many soon after his death.

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

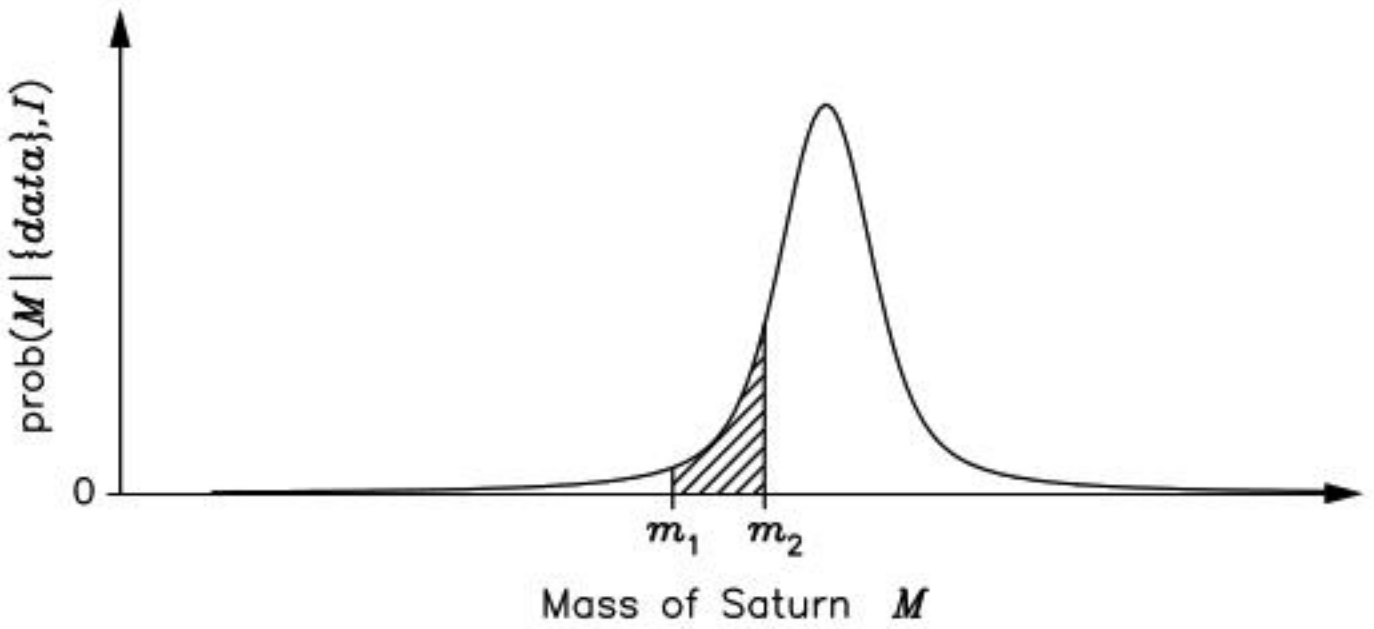


Fig. 1.2 A schematic illustration of the result of Laplace' analysis of the mass of Saturn.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

The problem was not really one of substance but concept. To the pioneers such as Bernoulli, Bayes and Laplace, a probability represented a *degree-of-belief* or plausibility: how much they thought that something was true, based on the evidence at hand. To the 19th century scholars, however, this seemed too vague and subjective an idea to be the basis of a rigorous mathematical theory. So they redefined probability as the *long-run relative frequency* with which an event occurred, given (infinitely) many repeated (experimental) trials. Since frequencies can be measured, probability was now seen as an objective tool for dealing with *random* phenomena.

Although the frequency definition appears to be more objective, its range of validity is also far more limited. For example, Laplace used (his) probability theory to estimate the mass of Saturn, given orbital data that were available to him from various astronomical observatories. In essence, he computed the posterior pdf for the mass M, given the data and all the relevant background information I (such as a knowledge of the laws of classical mechanics): $\mathrm{prob}(M|\{data\}, I)$; this is shown schematically in Fig. 1.2. To Laplace, the (shaded) area under the posterior pdf curve between m_1 and m_2 was a measure of how much he believed that the mass of Saturn lay in the range $m_1 \leqslant M < m_2$. As such, the position of the maximum of the posterior pdf represents a best estimate of the mass; its width, or spread, about this optimal value gives an indication of the uncertainty in the estimate. Laplace stated that: '. . . it is a bet of 11,000 to 1 that the error of this result is not 1/100th of its value.' He would have won the bet, as another 150 years' accumulation of data has changed the estimate by only 0.63%! According to the frequency definition, however, we are not permitted to use probability theory to tackle this problem. This is because the mass of Saturn is a constant and not a *random variable*; therefore, it has no frequency distribution and so probability theory cannot be used.

If the pdf of Fig. 1.2 had to be interpreted in terms of the frequency definition, we would have to imagine a large *ensemble* of universes in which everything remains constant apart from the mass of Saturn. As this scenario appears quite far-fetched, we might be inclined to think of Fig. 1.2 in terms of the distribution of the measurements of the mass in many repetitions of the experiment. Although we are at liberty to think about a problem in any way that facilitates its solution, or our understanding of it, having to seek a frequency interpretation for every data analysis problem seems rather perverse.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

For example, what do we mean by the 'measurement of the mass' when the data consist of orbital periods? Besides, why should we have to think about many repetitions of an experiment that never happened? What we really want to do is to make the best inference of the mass given the (few) data that we actually have; this is precisely the Bayes and Laplace view of probability.

Faced with the realization that the frequency definition of probability theory did not permit most real-life scientific problems to be addressed, a new subject was invented — *statistics*! To estimate the mass of Saturn, for example, one has to relate the mass to the data through some function called the statistic; since the data are subject to 'random' noise, the statistic becomes the random variable to which the rules of probability theory can be applied. But now the question arises: How should we choose the statistic? The frequentist approach does not yield a natural way of doing this and has, therefore, led to the development of several alternative schools of *orthodox* or *conventional* statistics. The masters, such as Fisher, Neyman and Pearson, provided a variety of different principles, which has merely resulted in a plethora of tests and procedures without any clear underlying rationale. This lack of unifying principles is, perhaps, at the heart of the shortcomings of the cook-book approach to statistics that students are often taught even today.

The frequency definition of probability merely gives the impression of a more objective theory. In reality it just makes life more complicated by hiding the difficulties under the rug, only for them to resurface in a less obvious guise. Indeed, it is not even clear that the concept of 'randomness' central to orthodox statistics is any better-defined than the idea of 'uncertainty' inherent in Bayesian probability theory. For example, we might think that the numbers generated by a call to a function like `rand` on a computer constitutes a random process: the frequency of the numbers will be distributed uniformly between 0 and 1, and their sequential order will appear haphazard. The illusory nature of this randomness would become obvious, however, if we knew the *algorithm* and the *seed* for the function `rand` (for then we could predict the sequence of numbers output by the computer). At this juncture, some might argue that, in contrast to our simple illustration above, *chaotic* and *quantum* systems provide examples of physical situations which are intrinsically random. In fact, chaos theory merely underlines the point that we are trying to make: the apparent randomness in the long-term behaviour of a classical system arises because we do not, or cannot, know its initial conditions well enough; the actual temporal evolution is entirely deterministic, and obeys Newton's Second Law of Motion. The quantum case is more difficult to address, since its interpretation (as opposed to its technical success) is still an open question for many people, but we refer the interested reader to Caves *et al.* (2002) for a very insightful viewpoint. Either way, 'randomness' is what we call our inability to predict things which, in turn, reflects our lack of knowledge about the system of interest. This is again consistent with the Bayes and Laplace view of probability, rather than the asserted physical objectivity of the frequentist approach.

To emphasize this last point, that a probability represents a state of knowledge rather than a physically real entity, consider the following example of Jaynes (1989). We are told that a dark bag contains five red balls and seven green ones. If this bag is shaken

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

well, and a ball selected at 'random', then most of us would agree that the probability of drawing a red ball is 5/12 and the probability of drawing a green one is 7/12. If the ball is not returned to the bag, then it seems reasonable that the probability of obtaining a red or green ball on the second draw will depend on the outcome of the first (because there will be one less red or green ball left in the bag). Now suppose that we are not told the outcome of the first draw, but are given the result of the second one. Does the probability of the first draw being red or green change with the knowledge of the second? Initially, many of us would be inclined to say 'no': at the time of the first draw, there were still five red balls and seven green ones in the bag; so, the probabilities for red and green should still be 5/12 and 7/12 irrespective of the outcome of the second draw. The error in this argument becomes obvious if we consider the extreme case of a bag containing only one red and one green ball. Although the second draw cannot affect the first in a physical sense, a knowledge of the second result does influence what we can infer about the outcome of the first one: if the second was green, then the first one must have been red; and vice versa. Thus (conditional) probabilities represent *logical* connections rather than *causal* ones.

The concerns about the subjectivity of the Bayesian view of probability are understandable, and the aim of creating an objective theory is quite laudable. Unfortunately, the frequentist approach does not achieve this goal: neither does its concept of randomness appear very rigorous, or fundamental, under scrutiny and nor does the arbitrariness of the choice of the statistic make it seem objective. In fact, the presumed shortcomings of the Bayesian approach merely reflect a confusion between subjectivity and the difficult technical question of how probabilities should be assigned. The popular argument goes that if a probability represents a degree-of-belief, then it must be subjective because my belief could be different from yours. The Bayesian view is that a probability does indeed represent how much we believe that something is true, but that this belief should be based on all the relevant information available. While this makes the assignment of probabilities an open-ended question, because the information at my disposal may not be the same as that accessible to you, it is not the same as subjectivity. It simply means that probabilities are always conditional, and this conditioning must be stated explicitly. As Jaynes has pointed out, objectivity demands only that two people having the same information should assign the same probability; this principle has played a key rôle in the modern development of the (objective) Bayesian approach.

In 1946, Richard Cox tried to get away from the controversy of the Bayesian versus frequentist view of probability. He decided to look at the question of plausible reasoning afresh, from the perspective of logical consistency. He found that the only rules which met his requirements were those of probability theory. Although the sum and product rules of probability are easy to prove for frequencies (with the aid of a *Venn diagram*), Cox's work shows that their range of validity goes much further. Rather than being restricted to just frequencies, probability theory constitutes the basic calculus for logical and consistent plausible reasoning; for us, that means scientific inference (which is the purpose of data analysis). So, Laplace was right: 'It is remarkable that a science, which commenced with a consideration of games of chance, should be elevated to the rank of the most important subjects of human knowledge.'

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

1.5 Outline of book

The aim of this book is to show how probability theory can be used directly to address data analysis problems in a straightforward manner. We will start, in Chapter 2, with the simplest type of examples: namely, those involving the estimation of the value of a single parameter. They serve as a good first encounter with Bayes' theorem in action, and allow for a useful discussion about *error-bars* and *confidence intervals*. The examples are extended to two, and then several, parameters in Chapter 3, enabling us to introduce the additional concepts of *correlation* and marginalization. In Chapter 4, we will see how the same principles used for parameter estimation can be applied to the problem of model selection.

Although Cox's work shows that plausibilities, represented by real numbers, should be manipulated according to the rules of probability theory, it does not tell us how to assign them in the first place. We turn to this basic question of assigning probabilities in Chapter 5, where we will meet the important principle of *maximum entropy* (MaxEnt). It may seem peculiar that we leave so fundamental a question to such a late stage, but it is intentional. People often have the impression that Bayesian analysis relies heavily on the use of clever probability assignments and is, therefore, not generally applicable if these are not available. Our aim is to show that even when armed only with a knowledge of pdfs familiar from high school (*binomial*, *Poisson* and *Gaussian*), and naïveté (a *flat*, or *uniform*, pdf), probability theory still provides a powerful tool for obtaining useful results for many data analysis problems. Indeed, we will find that most conventional statistical procedures implicitly assume such elementary assignments. Of course we might do better by thinking deeply about a more appropriate pdf for any given problem, but the point is that it is not usually crucial in practice.

In Chapter 6, we consider *non-parametric* estimation; that is to say, problems where we know so little about the object of interest that we are unable to describe it adequately in terms of a few parameters. Here we will encounter MaxEnt once again, but in the slightly different guise of *image processing*.

In Chapter 7, we focus our attention on the subject of *experimental design*. This concerns the question of 'what are the best data to collect', in contrast to most of this book, which deals with 'what is the optimal way of analysing these data'. This reciprocal question can also be addressed by probability theory, and is of great importance because the benefits of good experimental (or instrumental) design can far outweigh the rewards of the sophisticated analysis of poorer data.

Chapters 8 shows how one of the most widely used data analysis procedures in the physical sciences, *least-squares*, can be extended to deal with more troublesome measurements, such as those with 'outliers'.

Chapters 9 and 10 are concerned with modern numerical techniques for carrying out out Bayesian calculations, when analytical approximations are inadequate and a brute-force approach is impractical. In particular, they provide the first introductory account of the novel idea of *nested sampling*.

Most of the examples used in this book involve continuous pdfs, since this is usually the nature of parameters in real life. As mentioned in Section 1.3, this can simply be considered as the limiting case of an arbitrarily large number of discrete propositions.

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Jaynes (2003) correctly warns us, however, that difficulties can sometimes arise if we are not careful in carrying out the limiting procedure explicitly; this is often the underlying cause of so-called paradoxes of probability theory. Such problems also occur in other mathematical calculations, which are quite unrelated to probability. All that one really needs to avoid them is basic professional care allied to common sense. Indeed, we hope that the reader will come to share Laplace's more general view that '*Probability theory is nothing but common sense reduced to calculation*'.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary
3549a64f7957a5b7cc59fd64aa8d4446
ebrary
3549a64f7957a5b7cc59fd64aa8d4446
ebrary
Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

2 Parameter estimation I

Parameter estimation is a common data analysis problem. Like Laplace, for example, we may be interested in knowing the mass of Saturn; or, like Millikan, the charge of the electron. In the simplest case, we are only concerned with the value of a single parameter; such elementary examples are the focus of this chapter. They serve as a good introduction to the use of Bayes' theorem, and allow for a discussion of error-bars and confidence intervals.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

2.1 Example 1: is this a fair coin?

Let us begin with the analysis of data from a simple coin-tossing experiment. Suppose I told you that I had been to Las Vegas for my holidays, and had come across a very strange coin in one of the casinos; given that I had observed 4 heads in 11 flips, do you think it was a *fair* coin? By fair, we mean that we would be prepared to lay an even $50:50$ bet on the outcome of a flip being a head or a tail. In ascribing the property of fairness (solely) to the coin we are, of course, assuming that the coin-tosser was not skilled enough to be able to control the initial conditions of the flip (such as the angular and linear velocities). If we decide that the coin was fair, the question which follows naturally is how sure are we that this was so; if it was not fair, how unfair do we think it was?

A sensible way of formulating this problem is to consider a large number of contiguous propositions, or hypotheses, about the range in which the *bias-weighting* of the coin might lie. If we denote the bias-weighting by H, then $H = 0$ and $H = 1$ can represent a coin which produces a tail or a head on every flip, respectively. There is a continuum of possibilities for the value of H between these limits, with $H = 1/2$ indicating a fair coin. The propositions could then be, for example: (a) $0.00 \leqslant H < 0.01$; (b) $0.01 \leqslant H < 0.02$; (c) $0.02 \leqslant H < 0.03$; and so on. Our state of knowledge about the fairness, or the degree of unfairness, of the coin is then completely summarized by specifying how much we believe these various propositions to be true. If we assign a high probability to one (or a closely-grouped few) of these propositions, compared to the others, then this would indicate that we were confident in our estimate of the bias-weighting. If there was no such strong distinction, then it would reflect a high level of ignorance about the nature of the coin.

In the light of the data, and the above discussion, our inference about the fairness of this coin is summarized by the conditional pdf: $\mathrm{prob}(H\,|\,\{data\}, I)$. This is, of course, shorthand for the limiting case of a continuum of propositions for the value of H; that is to say, the probability that H lies in an infinitesimally narrow range is given by $\mathrm{prob}(H\,|\,\{data\}, I)\,\mathrm{d}H$. To estimate this posterior pdf, we need to use Bayes' theorem (eqn 1.3); it relates the pdf of interest to two others, which are easier to assign:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(H|\{data\}, I) \propto \text{prob}(\{data\}|H, I) \times \text{prob}(H|I)\,. \tag{2.1}$$

Note that we have omitted the denominator $\text{prob}(\{data\}|I)$, as it does not involve bias-weighting explicitly, and replaced the equality by a proportionality. If required, we can evaluate the missing constant subsequently from the normalization condition of eqn (1.10):

$$\int_0^1 \text{prob}(H|\{data\}, I)\,\mathrm{d}H = 1\,. \tag{2.2}$$

The prior pdf, $\text{prob}(H|I)$, on the far right-hand side of eqn (2.1), represents what we know about the coin given only the information I that we are dealing with a 'strange coin from Las Vegas'. Since casinos can be rather dubious places, we should keep a very open mind about the nature of the coin; a simple probability assignment which reflects this is a uniform, or flat, pdf:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(H|I) = \begin{cases} 1 & 0 \leqslant H \leqslant 1\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.3}$$

This prior state of knowledge, or ignorance, is modified by the data through the likelihood function: $\text{prob}(\{data\}|H, I)$. It is a measure of the chance that we would have obtained the data that we actually observed, if the value of the bias-weighting was given (as known). If, in the conditioning information I, we assume that the flips of the coin were independent events, so that the outcome of one did not influence that of another, then the probability of obtaining the data 'R heads in N tosses' is given by the binomial distribution:

$$\text{prob}(\{data\}|H, I) \propto H^R\,(1-H)^{N-R}\,. \tag{2.4}$$

We leave a formal derivation of this pdf to Chapter 5, but point out that eqn (2.4) seems reasonable because H is the chance of obtaining a head on any flip, and there were R of them, and $1-H$ is the corresponding probability for a tail, of which there were $N-R$. For simplicity, an equality has again been replaced by a proportionality; this is permissible since the omitted terms contain factors of only R and N, rather than the quantity of interest H.

According to eqn (2.1), the product of eqns (2.3) and (2.4) yields the posterior pdf we require; it represents our state of knowledge about the nature of the coin in the light of the data. To get a feel for this result, it is instructive to see how this pdf evolves as we obtain more and more data pertaining to the coin. This is done with the aid of data generated in a computer simulation, and the results of their analyses is shown in Fig. 2.1. The panel in the top left-hand corner shows the posterior pdf for H given no data; it is, of course, the same as the prior pdf of eqn (2.3). It indicates that we have no more reason to believe that the coin is fair than we have to think that it is double-headed, double-tailed, or of any other intermediate bias-weighting.

Suppose that the coin is flipped once and it comes up heads; what can we now say about the value of H? The resulting posterior pdf, shown in the second panel of Fig. 2.1,

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

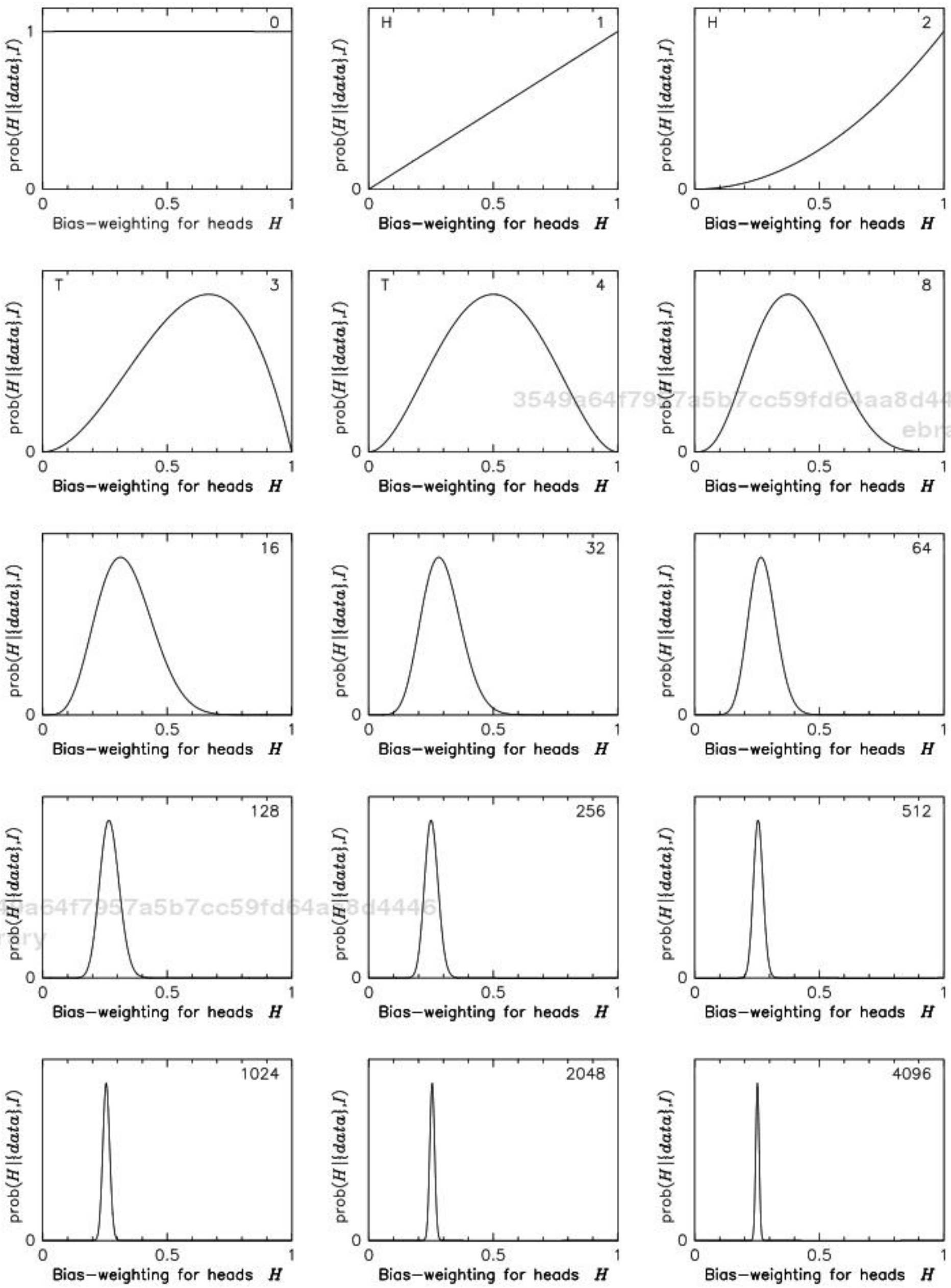


Fig. 2.1 The evolution of the posterior pdf for the bias-weighting of a coin, $\mathrm{prob}(H\,|\,\{data\}, I)$, as the number of data available increases. The figure on the top right-hand corner of each panel shows the number of data analysed; in the early panels, the H or T in the top left-hand corner shows whether the result of that (last) flip was a head or a tail.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

goes to zero for $H = 0$ and rises linearly to having the greatest value at $H = 1$. Based purely on this single datum, it is most probable that the (strange) coin has two heads; after all, we don't have any empirical evidence that it even has a tail yet. Although $H = 1$ is in some ways our 'best' estimate so far, the posterior pdf indicates that this value is not much more probable than many others. The only thing we can really be sure about is that the coin is not double-tailed; hence, the posterior pdf is zero at $H = 0$.

If the coin is flipped for a second time and again comes up heads, the posterior pdf becomes slightly more peaked towards $H = 1$ (proportional to H^2); this is plotted in the third panel of Fig. 2.1. Since we still haven't seen a tail, our inclinations following the first datum are just reinforced. As soon as a tail is obtained, however, the posterior pdf for $H = 1$ also drops to zero; we are then sure that the coin is not double-headed either. The next panel shows the resultant pdf, given that the third flip produced a tail (proportional to $(1-H)\,H^2$). If the fourth flip also comes up tails, then the maximum of the posterior pdf is at $H = 0.5$. It then becomes most probable that the coin is fair, but there is still a large degree of uncertainty in this estimate; this is indicated graphically in panel 5 of Fig. 2.1.

The remainder of Fig. 2.1 shows how the posterior pdf for H evolves as the number of data analysed becomes larger and larger. We see that the position of the maximum wobbles around, but that the amount by which it does so decreases with the increasing number of observations. The width of the posterior pdf also becomes narrower with more data, indicating that we are becoming increasingly confident in our estimate of the bias-weighting. For the coin in this example, the best estimate of H eventually converges to 0.25. This was, of course, the value chosen to simulate the flips; they can be thought of as coming from a tetrahedral coin with a head on one face and tails on the other three!

2.1.1 Different priors

Most people tend to be happy with the binomial distribution for the likelihood function, but worry about the prior pdf. The uniform assignment of eqn (2.3) was chosen mostly for its simplicity; it is just a naïve way of encoding a lot of initial ignorance about the coin. How would our inference about the fairness of the coin have changed if we had chosen a different prior?

To address this question, let's repeat the analysis of the data above with two alternative prior pdfs; the results are shown in Fig. 2.2. The solid line is for the uniform pdf used in Fig. 2.1, and is included for ease of comparison. One of the alternative priors is peaked around $H = 0.5$ and reflects our background information that most coins are fair (even in Las Vegas); it is plotted with a dashed line. It has a width which is broad enough to comfortably accommodate the possibility that H might be as low as 0.35, or as high as 0.65, but expresses considerable doubt about greater levels of biased behaviour. The second alternative, shown with a dotted line, is very sharply peaked at $H = 0$ and $H = 1$; it indicates that we would expect a 'strange coin from a casino' to be heavily biased, one way or another.

Figure 2.2 shows how the posterior pdfs for the different priors evolve, as more and more data become available; they have all been scaled vertically to have the same

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

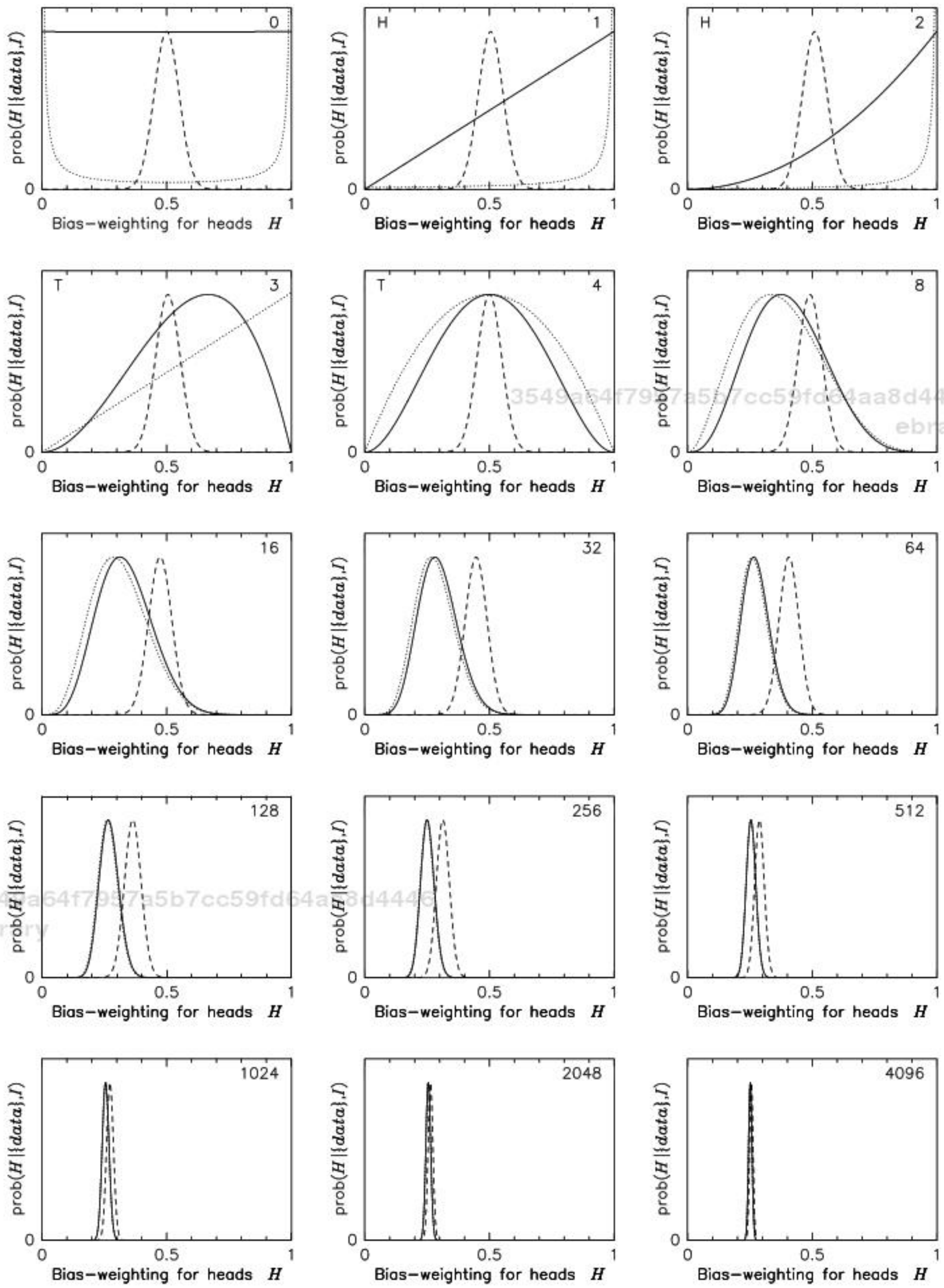


Fig. 2.2 The effect of different priors, $\mathrm{prob}(H\,|\,I)$, on the posterior pdf for the bias-weighting of a coin. The solid line is the same as in Fig. 2.1, and is included for ease of comparison. The case for two alternative priors, reflecting slightly different assumptions in the conditioning information I, are shown with dashed and dotted lines.

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

maximum value to aid comparison between them. We find that when there are few data, the resulting posteriors are different in detail; as the number of data increases, they all become more sharply peaked and converge to the same answer. This seems quite reasonable. The outcome of only a few flips tells us little about the fairness of the coin. Our state of knowledge after the analysis of these data is, therefore, strongly dependent on what we knew or assumed before the results; hence, the posteriors are somewhat different. As the empirical evidence grows, we are eventually led to the same conclusions irrespective of our initial beliefs; the posterior pdf is then dominated by the likelihood function, and the choice of the prior becomes largely irrelevant.

Two further curious points may be noted: (i) it takes quite a lot of flips to be able to estimate the bias-weighting with some degree of confidence (about a thousand to pin it down between 0.2 and 0.3); (ii) the posterior pdfs for the solid and dotted lines converge together quite quickly, but the dashed case takes much longer. The answer to the first observation is that it just does, but the number of flips required will depend on the actual bias-weighting of the coin. If the coin was tossed ten times and came up tails on every occasion, we would rapidly conclude that it was biased; on the other hand, the result of 45 heads and 55 tails in a hundred flips would still leave us somewhat uncertain as to whether the coin was fair. With regard to the second point, both the flat (solid) and the spiky (dotted) priors encode a large degree of ignorance about the nature of the coin. Despite the peculiar shape of the latter, it is fairly flat for most values of H; the strange-looking spikes at $H=0$ and $H=1$ disappear as soon as a head and a tail have been observed. The 'fair-minded' prior (dashed line), however, claims to be moderately well-informed about the character of the coin regardless of the data. It, therefore, takes much more to be convinced that the coin is not fair. Even though the prior probability for $H=0.5$ was about a million times greater than that of $H=0.25$, a thousand flips were enough to drag it (kicking and screaming, perhaps) to the conclusion that the value of H was less than 0.3 (but more than 0.2).

2.1.2 Sequential or one-step data analysis?

If we have a set of data $\{D_k\}$ comprising the outcome of N flips of a coin ($k = 1, 2, \ldots, N$), then Bayes' theorem tells us that the posterior pdf for H is given by

$$\text{prob}(H|\{D_k\}, I) \propto \text{prob}(\{D_k\}|H, I) \times \text{prob}(H|I)\,. \tag{2.5}$$

This is a one-step process in that we consider the data collectively, as a whole. As an alternative, we could also think of analysing the data sequentially (as they arrive). That is to say, we start by computing the posterior pdf based on the first datum D_1, $\text{prob}(H|D_1, I)$, and use it as the prior for the analysis of the second datum D_2; the new posterior could then be used as the prior for the third datum, and so on. If we continue this procedure, will we obtain the same result as the one-step approach?

To simplify matters, let us consider the case of just two data. The posterior pdf for H, based on both, is merely a special case of eqn (2.5) for $N=2$:

$$\text{prob}(H|D_2, D_1, I) \propto \text{prob}(D_2, D_1|H, I) \times \text{prob}(H|I)\,. \tag{2.6}$$

Equally well, we could use Bayes' theorem to express the posterior in terms of pdfs conditional on D_1 throughout (like the background information I):

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(H|D_2, D_1, I) \propto \text{prob}(D_2|H, D_1, I) \times \text{prob}(H|D_1, I)\,. \tag{2.7}$$

This shows that the prior in eqn (2.6) can certainly be replaced by the posterior pdf based on the first datum, but the other term doesn't quite look like the likelihood function for the second datum. Implicit in the assignment of the binomial pdf of eqn (2.4), however, was the assumption (subsumed in I) that the data were *independent*. This means that, given the value of H, the result of one flip does not influence what we can infer about the outcome of another; mathematically, we write this as

$$\text{prob}(D_2|H, D_1, I) = \text{prob}(D_2|H, I)\,.$$

Substituting this in eqn (2.7) yields the desired relationship, showing that the two data can either be analysed together or one after the other. This argument can be extended to the third datum D_3, giving

$$\text{prob}(H|D_3, D_2, D_1, I) \propto \text{prob}(D_3|H, I) \times \text{prob}(H|D_2, D_1, I)\,,$$

and repeated until all the data have been included. We should not really be surprised that both the one-step and sequential methods of analysis give the same answer; after all, the requirements of such consistency was what led Cox to the rules of probability theory in the first place.

When one first learns about Bayes' theorem, and sees how the data modify the prior through the likelihood function, there is occasionally a temptation to use the resulting posterior pdf as the prior for a re-analysis of the same data. It would be erroneous to do this, and the results quite misleading. In order to justify any data analysis procedure, we must be able to relate the pdf of interest to others used in its calculation through the sum and product rules of probability (or their corollaries). If we cannot do this then the analysis will be suspect at best, and open to logical inconsistencies. For the case of this proposed 'boot-strapping' we would, in fact, be trying to relate the posterior pdf to itself; this can only be done by an equality, and not through the likelihood function. If we persist in our folly, and keep repeating it, the resulting posterior pdfs will become sharper and sharper; we will just fool ourselves into thinking that the quantity of interest can be estimated far more accurately than is warranted by the data.

2.2 Reliabilities: best estimates, error-bars and confidence intervals

We have seen how the posterior pdf encodes our inference about the value of a parameter, given the data and the relevant background information. Often, however, we wish to summarize this with just two numbers: the best estimate and a measure of its reliability. Since the probability (density) associated with any particular value of the parameter is a measure of how much we believe that it lies in the neighbourhood of that point, our best estimate is given by the maximum of the posterior pdf. If we denote the quantity of interest by X, with a posterior pdf $P = \text{prob}(X|\{data\}, I)$, then the best estimate of its value X_{O} is given by the condition

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\left.\frac{\mathrm{d}P}{\mathrm{d}X}\right|_{X_\mathrm{o}} = 0\,. \tag{2.8}$$

Strictly speaking, we should also check the sign of the *second derivative* to ensure that X_O represents a maximum rather than a minimum (or a point of inflexion):

$$\left.\frac{\mathrm{d}^2P}{\mathrm{d}X^2}\right|_{X_\mathrm{o}} < 0\,. \tag{2.9}$$

In writing the derivatives of P with respect to X we are, of course, assuming that X is a continuous parameter. If it could only take discrete values, our best estimate would still be that which gave the greatest posterior probability; it's just that we couldn't then use the calculus notation of eqns (2.8) and (2.9), because the *gradients* are not defined (since the increment in X cannot be infinitesimally small).

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

To obtain a measure of the reliability of this best estimate, we need to look at the width or spread of the posterior pdf about X_O. When considering the behaviour of any function in the neighbourhood of a particular point, it is often helpful to carry out a *Taylor series* expansion; this is simply a standard tool for (locally) approximating a complicated function by a low-order *polynomial*. Rather than dealing directly with the posterior pdf P, which is a 'peaky' and positive function, it is better to work with its *logarithm* L,

$$L = \log_\mathrm{e}\big[\,\mathrm{prob}(X\,|\,\{data\},I)\,\big]\,, \tag{2.10}$$

since this varies much more slowly with X. Expanding L about the point $X = X_\mathrm{O}$, we have

$$L = L(X_\mathrm{O}) + \frac{1}{2}\left.\frac{\mathrm{d}^2L}{\mathrm{d}X^2}\right|_{X_\mathrm{o}}(X-X_\mathrm{O})^2 + \cdots\,, \tag{2.11}$$

where the best estimate of X is given by the condition

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\left.\frac{\mathrm{d}L}{\mathrm{d}X}\right|_{X_\mathrm{o}} = 0\,, \tag{2.12}$$

which is equivalent to eqn (2.8) because L is a *monotonic* function of P.

The first term in the Taylor series, $L(X_\mathrm{O})$, is a constant and tells us nothing about the shape of the posterior pdf. The linear term, which would be proportional to $X - X_\mathrm{O}$, is missing because we are expanding about the maximum (as indicated by eqn 2.12). The *quadratic* term is, therefore, the dominant factor determining the width of the posterior pdf and plays a central rôle in the reliability analysis. Ignoring all the higher-order contributions, the exponential of eqn (2.11) yields

$$\mathrm{prob}(X\,|\,\{data\},I) \approx A\,\exp\left[\frac{1}{2}\left.\frac{\mathrm{d}^2L}{\mathrm{d}X^2}\right|_{X_\mathrm{o}}(X-X_\mathrm{O})^2\right]\,, \tag{2.13}$$

where A is a normalization constant. Although this expression looks a little weird, what we have really done is to approximate the posterior pdf by the ubiquitous Gaussian distribution. Also known as the *normal* distribution, it is usually written as

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

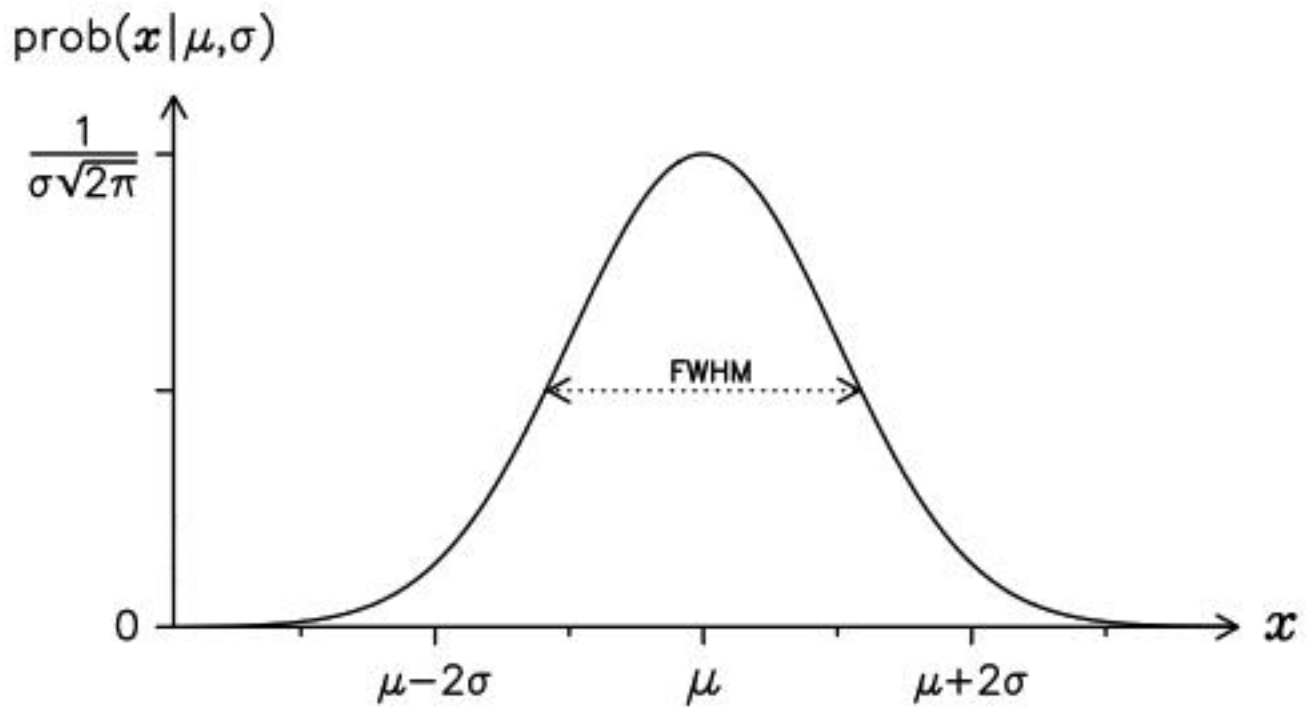


Fig. 2.3 The Gaussian, or normal, distribution. It is symmetric with respect to the maximum, at $x=\mu$, and has a full width at half maximum (FWHM) of about $2.35\,\sigma$.

$$\text{prob}(x\,|\,\mu,\sigma) \;=\; \frac{1}{\sigma\sqrt{2\pi}}\,\exp\left[-\,\frac{(x-\mu)^2}{2\,\sigma^2}\right], \tag{2.14}$$

and is plotted in Fig. 2.3; this function is symmetric about the maximum, at $x=\mu$, and has a width which is proportional to σ. Comparing the exponents of eqns (2.13) and (2.14), we are reassured to find that the posterior pdf for X has a maximum at $X=X_{\rm O}$; its width, characterized by the parameter σ, is inversely related to the square root of the second derivative of L at $X=X_{\rm O}$:

$$\sigma \;=\; \left(-\,\frac{\mathrm{d}^2 L}{\mathrm{d}X^2}\bigg|_{X_{\rm O}}\right)^{-1/2}, \tag{2.15}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary
where $\mathrm{d}^2L/\mathrm{d}X^2$ is necessarily negative following eqn (2.9). Our inference about the quantity of interest is conveyed very concisely, therefore, by the statement

$$X \;=\; X_{\rm O} \pm \sigma\,, \tag{2.16}$$

where $X_{\rm O}$ is the best estimate of the value of X, and σ is a measure of its reliability; the parameter σ is usually referred to as the *error-bar*.

The expression of eqn (2.16) is, of course, just shorthand for the Gaussian approximation of eqn (2.13). The integral properties of the normal distribution tell us that the probability that the true value of X lies within $\pm 1\sigma$ of $X=X_{\rm O}$ is 67%:

$$\text{prob}(X_{\rm O}-\sigma \leqslant X < X_{\rm O}+\sigma\,|\,\{data\},I) = \int_{X_{\rm O}-\sigma}^{X_{\rm O}+\sigma} \text{prob}(X,\{data\}\,|\,I)\,\mathrm{d}X \;\approx\; 0.67\,.$$

Similarly, the probability that X lies within $\pm 2\sigma$ of $X_{\rm O}$ is 95%; we would be quite surprised, however, if our best estimate of X was wrong by more than about 3σ.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

2.2.1 The coin example

As a concrete example of the above analysis, let's consider the case of the coin-tossing experiment of the previous section. Multiplying the uniform prior of eqn (2.3) with the binomial likelihood function of eqn (2.4), we obtain the posterior pdf for the bias-weighting:

$$\mathrm{prob}(H|\{data\},I) \propto H^R\,(1-H)^{N-R}\,,$$

where $0 \leqslant H \leqslant 1$. Taking its natural logarithm, according to eqn (2.10), gives

$$L = \mathrm{constant} + R\log_{\mathrm{e}}(H) + (N-R)\log_{\mathrm{e}}(1-H)\,. \tag{2.17}$$

For the best estimate of H and its error-bar, we need both the first and the second derivative of L; differentiating twice, with respect to H, we have

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\frac{\mathrm{d}L}{\mathrm{d}H} = \frac{R}{H} - \frac{(N-R)}{(1-H)} \qquad \text{and} \qquad \frac{\mathrm{d}^2L}{\mathrm{d}H^2} = -\frac{R}{H^2} - \frac{(N-R)}{(1-H)^2}\,. \tag{2.18}$$

However, eqn (2.12) states that, at the optimal value for the bias-weighting H_{O}, the first derivative is equal to zero:

$$\left.\frac{\mathrm{d}L}{\mathrm{d}H}\right|_{H_{\mathrm{O}}} = \frac{R}{H_{\mathrm{O}}} - \frac{(N-R)}{(1-H_{\mathrm{O}})} = 0\,.$$

Hence, algebraic rearrangement of this equation tells us that the best estimate of the bias-weighting is given by the relative frequency of outcomes of heads:

$$H_{\mathrm{O}} = \frac{R}{N}\,. \tag{2.19}$$

According to eqn (2.15), the associated error-bar is related to the second derivative of L evaluated at $H = H_{\mathrm{O}}$. Substituting for R from eqn (2.19), into the right-hand formula in eqn (2.18), and after some rearrangement and cancellation, we obtain

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\left.\frac{\mathrm{d}^2L}{\mathrm{d}H^2}\right|_{H_{\mathrm{O}}} = -\frac{N}{H_{\mathrm{O}}(1-H_{\mathrm{O}})}\,.$$

Equation (2.15) then tells us that the error-bar for the estimate of the bias-weighting is given by

$$\sigma = \sqrt{\frac{H_{\mathrm{O}}(1-H_{\mathrm{O}})}{N}}\,. \tag{2.20}$$

Since H_{O} does not vary a lot after a moderate amount of data have been analysed, the numerator tends to a constant value; thus the width of the posterior becomes inversely proportional to the square root of the number of data, as can be seen in Fig. 2.1. This formula for the error-bar also confirms our earlier assertion that it is easier to identify a highly biased coin than it is to be confident that it's fair (because the numerator in eqn (2.20) has its greatest value when $H_{\mathrm{O}} = 0.5$). The posterior pdf for H, given the data 9 heads in 32 flips, and the Gaussian approximation to it resulting from the above analysis ($H = 0.28 \pm 0.08$), are shown in Fig. 2.4.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

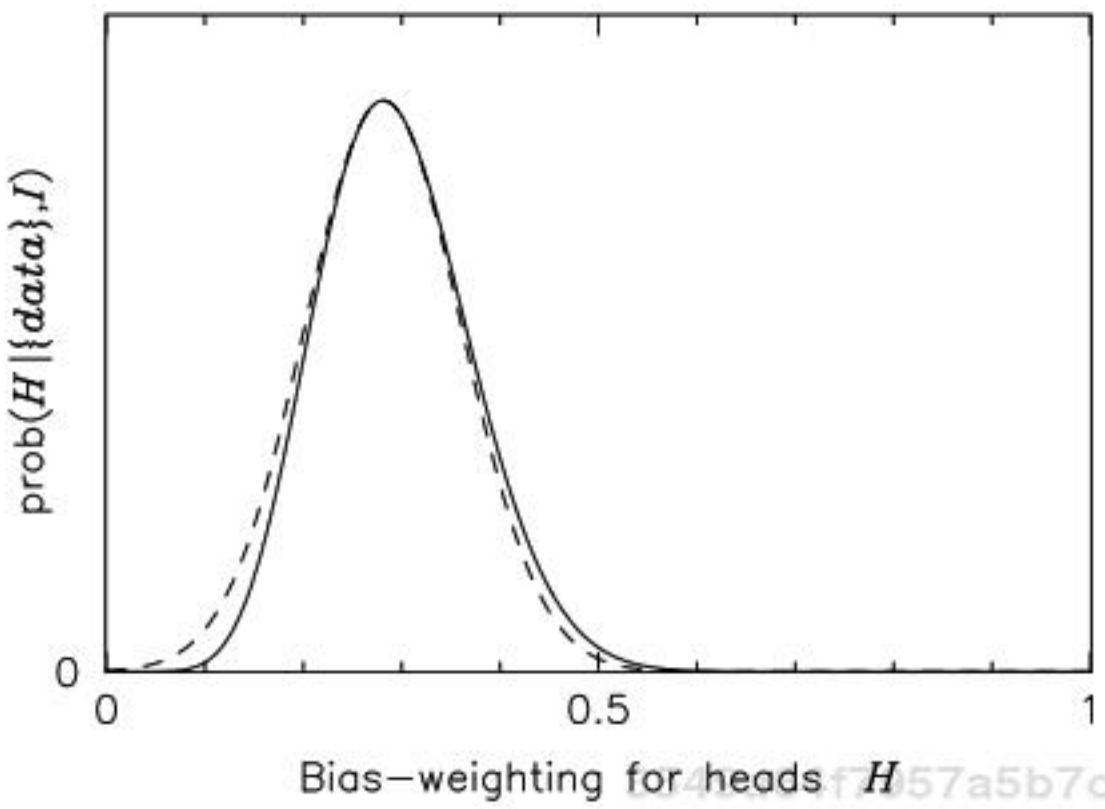


Fig. 2.4 The Gaussian, or quadratic, approximation (dashed line) to the posterior pdf for the bias-weighting of the coin, given 9 heads in 32 flips.

2.2.2 Asymmetric posterior pdfs

The preceding analysis, leading to the error-bar formula of eqn (2.15), relies on the validity of the quadratic expansion of eqn (2.11). Figure 2.4 illustrates that this is usually a reasonable approximation; in fact, the agreement with the Gaussian pdf tends to improve steadily as the number of data increases. There are times, however, when the posterior pdf is markedly asymmetric; this can be seen in the early panels of Figs. 2.1 and 2.2. While the maximum of the posterior (X_{O}) can still be regarded as giving the best estimate, the true value is now more likely to be on one side of this rather than the other. The concept of an error-bar does not seem appropriate in this case, as it implicitly entails the idea of symmetry.

A good way of expressing the reliability with which a parameter can be inferred, for an asymmetric posterior pdf, is through a *confidence interval*. Since the area under the posterior pdf between X_1 and X_2 is proportional to how much we believe that X lies in that range, the *shortest interval* that encloses 95% of the area represents a sensible measure of the uncertainty of the estimate. Assuming that the posterior pdf has been normalized, to have unit area, we need to find X_1 and X_2 such that

$$\mathrm{prob}(X_1 \leqslant X < X_2 \,|\{data\}, I) = \int_{X_1}^{X_2} \mathrm{prob}(X|\{data\}, I)\,\mathrm{d}X \approx 0.95\,, \qquad (2.21)$$

where the difference $X_2 - X_1$ is as small as possible. The region $X_1 \leqslant X < X_2$ is called the shortest 95% confidence interval, and is illustrated in Fig. 2.5.

A natural question which arises is 'Why have we chosen the 95% confidence level?' The answer is that this is traditionally seen as a reasonable value, as it provides a respectably conservative estimate of the reliability. There is nothing to stop us from giving the shortest 50%, 70%, 99%, or any other, confidence interval. Indeed, there is something to be said for listing a whole set of nested intervals since this provides a more

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

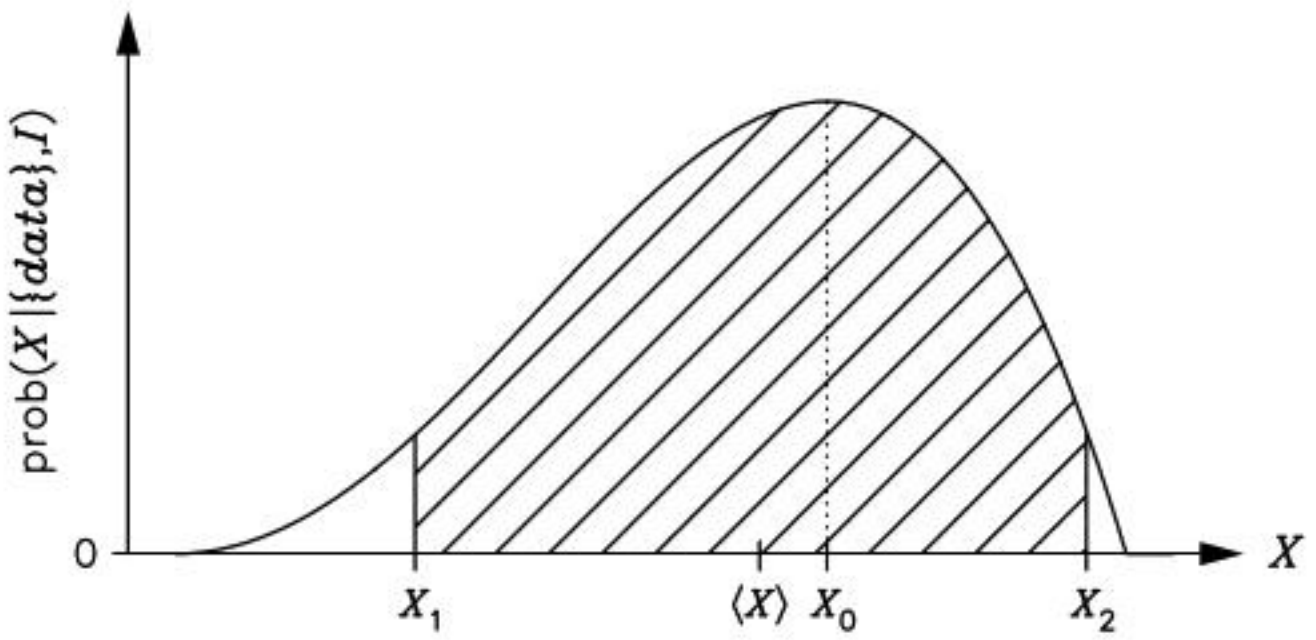


Fig. 2.5 The shortest 95% confidence interval, shown by the shaded region.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

complete picture of the reliability analysis; by doing so, however, we are merely reconstructing the posterior pdf!

When dealing with a highly asymmetric posterior pdf, the question of what we mean by the 'best' estimate is rather open-ended. The maximum still indicates the single most probable value for the parameter of interest, but the *mean*, or *expectation*, can be thought of as being more representative as it takes into account the skewness of the pdf. For a normalized posterior, this weighted average $\langle X \rangle$ is given by

$$\langle X \rangle = \int X \, \mathrm{prob}(X|\{data\}, I) \, \mathrm{d}X \,, \tag{2.22}$$

and is sometimes denoted by $E(X)$ or $\overline{X}$ (the latter not to be confused with 'not X'); as usual, the integral is replaced by a sum when X can only take discrete values. If the posterior pdf has not been normalized then, of course, the right-hand side must be divided by $\int \mathrm{prob}(X|\{data\}, I) \, \mathrm{d}X$. For the case of the pdf shown in Fig. 2.5, the mean lies slightly to the left of X_O. If the posterior pdf is symmetric about the maximum, as in the Gaussian distribution, then the mean and the maximum become coincident ($\langle X \rangle = X_\mathrm{O}$).

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

2.2.3 Multimodal posterior pdfs

So far, we have only considered posterior pdfs which have a single maximum. Depending on the nature of the experimental data to be analysed, we can sometimes obtain posteriors which are *multimodal*; this is depicted schematically in Fig. 2.6. There is no difficulty when one of the maxima is very much larger than the others: we can simply ignore the subsidiary solutions, to a good approximation, and concentrate on the global maximum. The problem arises when there are several maxima of comparable magnitude. What do we now mean by a best estimate, and how should we quantify its reliability?

Well, to a large extent, the difficulty is of our own making. The posterior pdf gives a complete description of what we can infer about the value of the parameter in the light of the data, and our relevant prior knowledge. The idea of a best estimate and an error-bar, or even a confidence interval, is merely an attempt to summarize the posterior with just two or three numbers; sometimes this just can't be done, and so these concepts

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

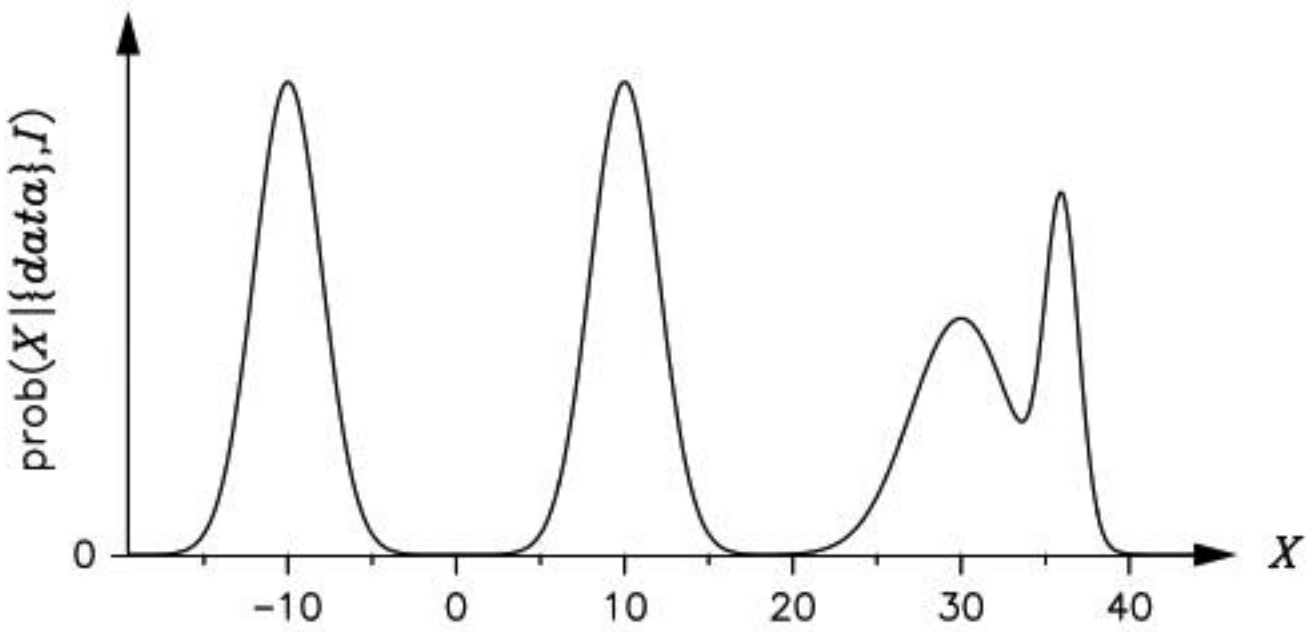


Fig. 2.6 A multimodal posterior pdf. Since its shape cannot be summarized by just a couple of numbers, the concept of a best estimate and an error-bar is inappropriate.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

are not valid. The posterior pdf still exists, and we are free to draw from it whatever conclusions are appropriate.

Consider the special case of a multimodal distribution having two (roughly) equal-sized probability bumps; such a *bimodal* pdf is seen in Fig. 2.6, if the structure to the right of $X = 20$ is ignored. This posterior conveys the message that the value of X is either -10, give or take a little bit, or about $+10$; we could write it as: '$X = -10 \pm 2$ or $X = +10 \pm 2$'. Since the mean of the posterior pdf is still unique, it is sometimes suggested that this should be used as a (single) best estimate. The problem with that argument is that the expectation $\langle X \rangle \approx 0$, a value which the posterior pdf indicates is very improbable! Even if we turned a blind eye to this difficulty, and used the mean as the best estimate, we would have to assign a sizeable error-bar to encompass the most probable values of X; this would again be somewhat misleading, as it is not a good reflection of the information inherent in the posterior pdf. For the bimodal case we can characterize the posterior in terms of a few numbers: two best estimates and their associated error-bars, or disjoint confidence intervals. For a general multimodal pdf, the most honest thing we can do is just display the posterior itself.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

2.3 Example 2: Gaussian noise and averages

For our second example, let us consider the problem of estimating the mean of a Gaussian process. The normal distribution, which was defined in eqn (2.14), and plotted in Fig. 2.3, is often used as a theoretical model for describing the *noise* (or imperfections) associated with experimental data. We leave its formal derivation with maximum entropy to Chapter 5, but note that its use is traditionally justified by appealing to the *central limit* theorem. For our present purposes, we need merely say that the probability of the kth datum having a value x_k is given by

$$\mathrm{prob}(x_k|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_k-\mu)^2}{2\,\sigma^2}\right], \tag{2.23}$$

where μ is the true value of the parameter of interest, and σ is a measure of the error in its measurement. Given a set of data $\{x_k\}$, what is the best estimate of μ and how

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

confident are we of this prediction?

In this example, we will restrict ourselves to the relatively easy case when the value of σ is known; the more realistic situation, when this condition is relaxed, is left to Section 3.3. Our inference about the value of μ is expressed, therefore, by the posterior pdf $\mathrm{prob}(\mu|\{x_k\},\sigma,I)$; to help us calculate it, we use Bayes' theorem:

$$\mathrm{prob}(\mu|\{x_k\},\sigma,I) \propto \mathrm{prob}(\{x_k\}|\mu,\sigma,I) \times \mathrm{prob}(\mu|\sigma,I)\,. \tag{2.24}$$

Although a knowledge of σ could be subsumed into the background information I, we will continue to state this conditioning explicitly (for future reference). If we assume that the data are independent, so that the measurement of one datum does not influence what we can infer about the outcome of another (when given the values of μ and σ), then the likelihood function is given by the product of the probabilities of obtaining the N individual data:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\mathrm{prob}(\{x_k\}|\mu,\sigma,I) = \prod_{k=1}^{N} \mathrm{prob}(x_k|\mu,\sigma,I)\,. \tag{2.25}$$

This follows from the repeated use of eqn (1.2), which reduces to

$$\mathrm{prob}(X,Y|I) = \mathrm{prob}(X|I) \times \mathrm{prob}(Y|I)\,,$$

since the notion of independence implies that

$$\mathrm{prob}(X|Y,I) = \mathrm{prob}(X|I)\,.$$

As a knowledge of the width of a Gaussian tells us nothing about the position of its centre, let us assign a simple uniform pdf for the prior:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\mathrm{prob}(\mu|\sigma,I) = \mathrm{prob}(\mu|I) = \begin{cases} A & \mu_{\min} \leqslant \mu \leqslant \mu_{\max}\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{2.26}$$

where the normalization constant A is given by the reciprocal of the range $\mu_{\max} - \mu_{\min}$, determined by the relevant background information I. According to eqn (2.24), if we multiply this prior by the likelihood function resulting from eqns (2.23) and (2.25), we obtain the following for the logarithm of the posterior pdf L:

$$L = \log_{\mathrm{e}}\big[\mathrm{prob}(\mu|\{x_k\},\sigma,I)\big] = \text{constant} - \sum_{k=1}^{N} \frac{(x_k-\mu)^2}{2\,\sigma^2}\,, \tag{2.27}$$

where the constant includes all terms not involving μ; the posterior pdf is, of course, zero outside the range $\mu_{\min}$ to $\mu_{\max}$.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

To find the best estimate μ_o, eqn (2.12) tells us that we need to differentiate L with respect to μ and set the derivative to zero:

$$\left.\frac{\mathrm{d}L}{\mathrm{d}\mu}\right|_{\mu_o} = \sum_{k=1}^{N} \frac{x_k - \mu_o}{\sigma^2} = 0 .$$

Since σ is a constant which is independent of k, we can take it outside the summation sign and rearrange the equation as follows:

$$\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} \mu_o = N\mu_o .$$

Thus the best estimate of the value of μ is given by the simple arithmetic average of the sample of measurements $\{x_k\}$:

$$\mu_o = \frac{1}{N} \sum_{k=1}^{N} x_k , \tag{2.28}$$

irrespective of the magnitude of the error in the measurement process. Our confidence in the reliability of this best estimate, however, does depend on the size of σ, as indicated by the second derivative of L:

$$\left.\frac{\mathrm{d}^2L}{\mathrm{d}\mu^2}\right|_{\mu_o} = -\sum_{k=1}^{N} \frac{1}{\sigma^2} = -\frac{N}{\sigma^2} .$$

According to eqn (2.15), the error-bar we should associate with μ_o is given by the square root of the inverse of (minus) this quantity; thus we can summarize our inference about the value of μ by the statement

$$\mu = \mu_o \pm \frac{\sigma}{\sqrt{N}} . \tag{2.29}$$

Just as in the case of the coin-tossing experiment, we encounter the familiar result that the reliability of the estimate is proportional to the square root of the number of data.

In Section 2.2, we noted that the error-bar analysis relies on the validity of the quadratic expansion of eqn (2.11). For the case of Gaussian noise considered above, this is not only a reasonable approximation but an exact identity (because all higher derivatives of L are zero). Thus, the posterior pdf is completely defined by the parameters of eqn (2.29). The only proviso concerns the values of $\mu_{\min}$ and $\mu_{\max}$. We could, in principle, express an inordinate amount of prior ignorance by letting them tend to $\pm\infty$ (although not at the outset!) but, as long as the range they span is large enough, their actual value has no effect on the posterior pdf. If the best estimate and error-bar allow for values of μ outside the range $\mu_{\min}$ to $\mu_{\max}$, then it is most honest simply to display the posterior itself with its cut-offs; in that case, probability theory would just be warning us that our prior knowledge was at least as important as the data.

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

2.3.1 Data with different-sized error-bars

In the above analysis, we assumed that the magnitude of the error-bar for each datum was the same; this may well be reasonable if all the measurements were made with a given experimental set-up. Suppose, however, that the data were obtained from several laboratories using equipment of varying sophistication. How should we then combine the evidence from observations of differing quality?

Let us assume that the measurement error can still be modelled through a Gaussian pdf, so that the probability of the kth datum having a value x_k is

$$\mathrm{prob}(x_k|\mu,\sigma_k) = \frac{1}{\sigma_k\sqrt{2\pi}}\exp\left[-\frac{(x_k-\mu)^2}{2\,\sigma_k^2}\right],$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

where the error-bar for each datum σ_k need not be of the same size. Following the arguments between eqns (2.23) and (2.27), we find that the logarithm of the posterior pdf for μ is now given by

$$L = \log_e\left[\,\mathrm{prob}(\mu|\{x_k\},\{\sigma_k\},I)\,\right] = \text{constant} - \sum_{k=1}^{N}\frac{(x_k-\mu)^2}{2\,\sigma_k^2}\,.$$

As usual, we obtain the expression for the best estimate by setting the first derivative of L equal to zero; since the resulting equation does not simplify as much as before, the formula for μ_o is somewhat more complicated:

$$\mu_o = \sum_{k=1}^{N} w_k x_k \Big/ \sum_{k=1}^{N} w_k\,, \qquad \text{where } w_k = \frac{1}{\sigma_k^2}\,. \tag{2.30}$$

Rather than the best estimate being given by the arithmetic mean of the data, we must now calculate their weighted average. Less reliable data will have larger error-bars σ_k and correspondingly smaller weights w_k. The second derivative of L yields the error-bar for the best estimate and allows us to summarize our inference about μ as

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\mu = \mu_o \pm \left(\sum_{k=1}^{N} w_k\right)^{-1/2}\,. \tag{2.31}$$

Note that if all the data were of comparable quality, so that $\sigma_k = \sigma$, then eqns (2.30) and (2.31) would reduce to the simpler forms of eqns (2.28) and (2.29).

2.4 Example 3: the lighthouse problem

For the third example, we follow Gull (1988) in considering a very instructive problem found on a problems sheet for first-year undergraduates at Cambridge: 'A lighthouse is somewhere off a piece of straight coastline at a position α along the shore and a distance β out at sea. It emits a series of short highly collimated flashes at random intervals and hence at random azimuths. These pulses are intercepted on the coast by photo-detectors

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

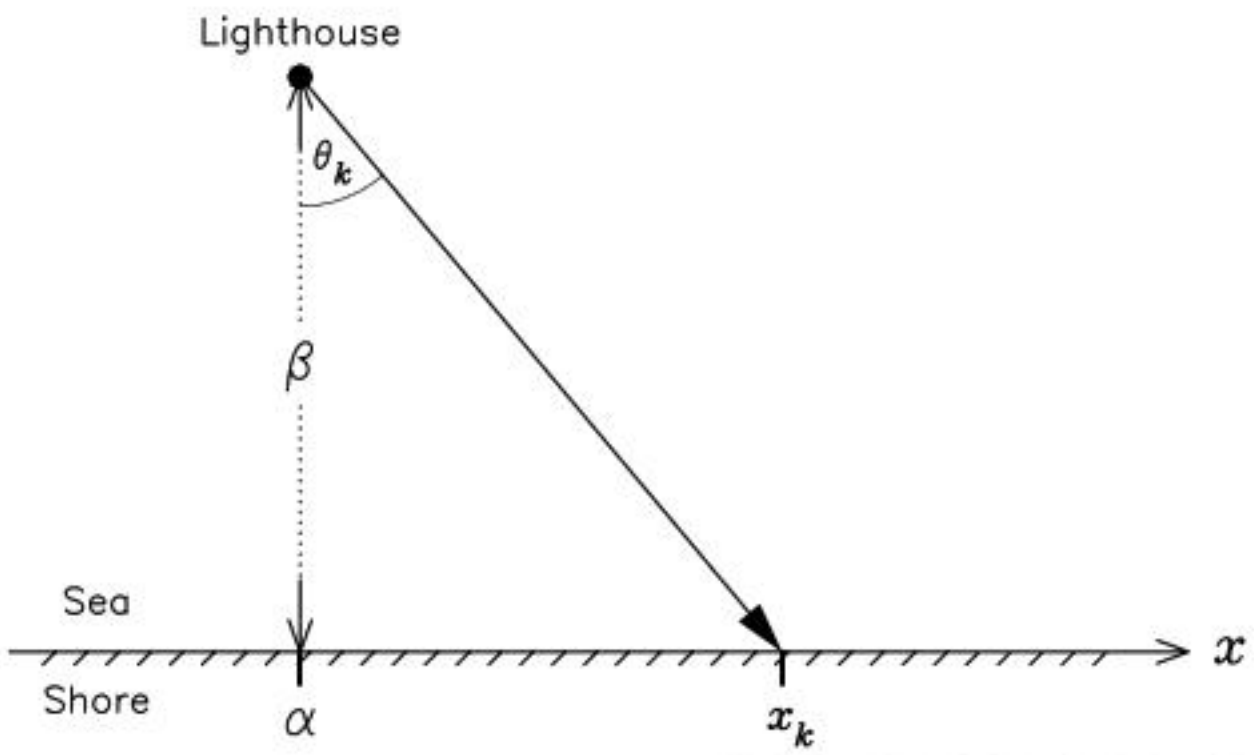


Fig. 2.7 A schematic illustration of the geometry of the lighthouse problem.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

that record only the fact that a flash has occurred, but not the angle from which it came. N flashes have so far been recorded at positions $\{x_k\}$. Where is the lighthouse?'

The geometry of the problem above is illustrated in Fig. 2.7. Given the nature of the lighthouse emissions, it seems reasonable to assign a uniform pdf for the azimuth of the kth datum θ_k:

$$\text{prob}(\theta_k|\alpha, \beta, I) = \frac{1}{\pi}, \tag{2.32}$$

where the angle must lie between $\pm\pi/2$ radians (or $\pm 90°$) to have been detected. Since the photo-detectors are only sensitive to position along the coast rather than direction, we must relate θ_k to x_k. An inspection of Fig. 2.7, and elementary *trigonometry*, allows us to write:

$$\beta \tan\theta_k = x_k - \alpha . \tag{2.33}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

As we shall see in Section 3.6, when dealing with the subject of *changing variables*, the relationship of eqn (2.33) can be used to rewrite the pdf of eqn (2.32) as

$$\text{prob}(x_k|\alpha, \beta, I) = \frac{\beta}{\pi\left[\beta^2 + (x_k - \alpha)^2\right]} . \tag{2.34}$$

This tells us that the probability that the kth flash will be recorded at position x_k, knowing the coordinates of the lighthouse (α, β), is given by a *Cauchy* distribution. The functional form of this pdf is met frequently in physics and is often called a *Lorentzian*. It is symmetric about the maximum, at $x_k = \alpha$, and has a FWHM of 2β; the distribution is plotted in Fig. 2.8.

Inferring the position of the lighthouse from the data involves the estimation of both α and β; as a two-parameter problem, this is outside the scope of the present chapter. We will therefore assume that the distance out to sea is known and reduce it to a single parameter example; the reader should have no difficulty, however, in tackling the full problem after the material of the next chapter. Our inference about the position

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

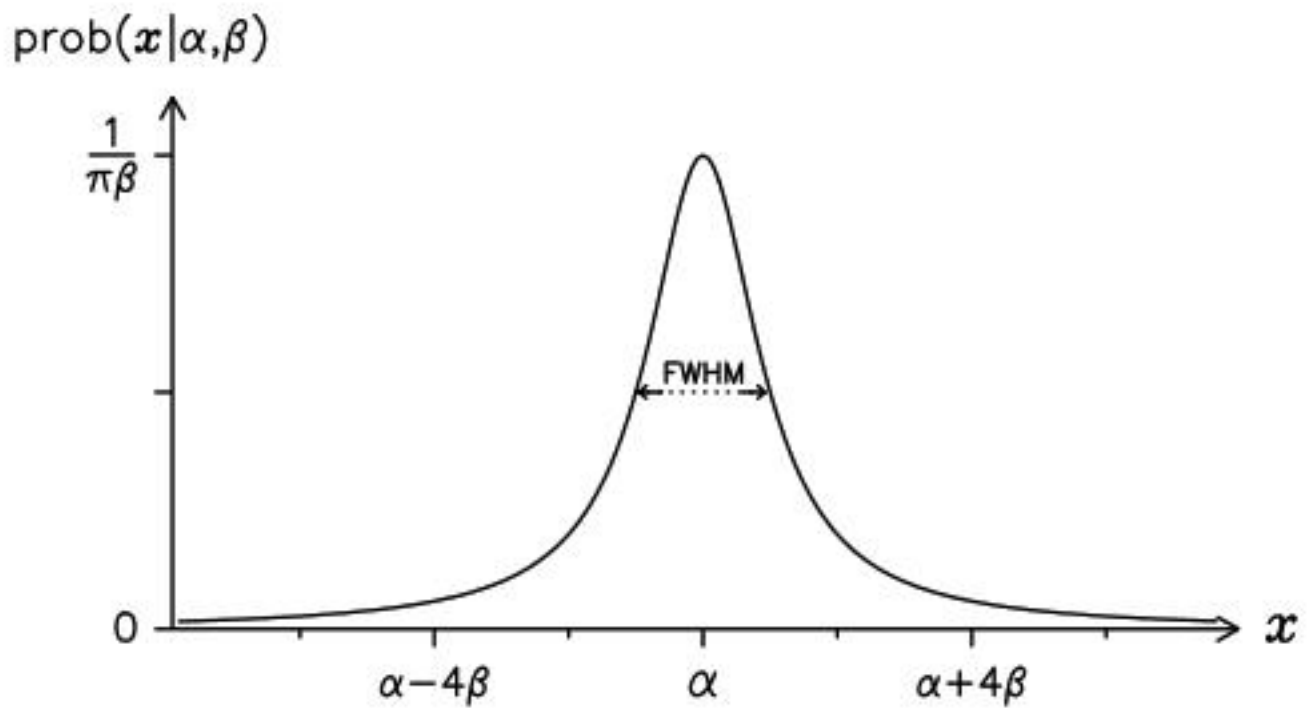


Fig. 2.8 The Cauchy, or Lorentzian, distribution. It is symmetric with respect to the maximum, at $x=\alpha$, and has a FWHM of twice β.

of the lighthouse is then expressed by the posterior pdf $\text{prob}(\alpha|\{x_k\},\beta,I)$; to help us calculate it, we begin by writing down Bayes' theorem:

$$\text{prob}(\alpha|\{x_k\},\beta,I) \propto \text{prob}(\{x_k\}|\alpha,\beta,I) \times \text{prob}(\alpha|\beta,I)\,. \tag{2.35}$$

As a knowledge of β, without the data, tells us nothing new about the position of the lighthouse along the shore, let us assign a simple uniform pdf for the prior:

$$\text{prob}(\alpha|\beta,I) = \text{prob}(\alpha|I) = \begin{cases} A & \alpha_{\min} \leqslant \alpha \leqslant \alpha_{\max}\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{2.36}$$

where $\alpha_{\min}$ and $\alpha_{\max}$ could represent the limits of the coastline, if these were known, or be made arbitrarily large (up to the size of the Earth!) if we were very ignorant; the normalization constant A is, of course, just the reciprocal of this prior range. Since the recording of a signal at one photo-detector does not influence what we can infer about the position of another measurement (when given the location of the lighthouse), the likelihood function for these independent data is just the product of the probabilities for obtaining the N individual detections:

$$\text{prob}(\{x_k\}|\alpha,\beta,I) = \prod_{k=1}^{N} \text{prob}(x_k|\alpha,\beta,I)\,. \tag{2.37}$$

Substituting the prior of eqn (2.36) and the likelihood function resulting from eqns (2.34) and (2.37) into Bayes' theorem of eqn (2.34), allows us to write the logarithm of the posterior pdf as

$$L = \log_{\mathrm{e}}\big[\,\text{prob}(\alpha|\{x_k\},\beta,I)\,\big] = \text{constant} - \sum_{k=1}^{N} \log_{\mathrm{e}}\Big[\beta^2 + (x_k - \alpha)^2\Big]\,, \tag{2.38}$$

where the constant includes all terms not involving α. For the purposes of this example, we will assume that the prior range is so large that we need not worry about any cut-offs imposed on the posterior.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

The best estimate of the position α_o is given by the maximum of the posterior pdf; differentiating eqn (2.38) once, with respect to α, we obtain the condition

$$\left.\frac{\mathrm{d}L}{\mathrm{d}\alpha}\right|_{\alpha_o} = 2\sum_{k=1}^{N} \frac{x_k - \alpha_o}{\beta^2 + (x_k - \alpha_o)^2} = 0\,. \tag{2.39}$$

Unfortunately, it is difficult to rearrange this equation so that α_o is expressed in terms of $\{x_k\}$ and β. Although an analytical solution may confound us, there is nothing to stop us tackling the problem numerically. The most straightforward method is to use brute force and ignorance: simply evaluate L, from eqn (2.38), for a whole series of different possible values of α; the number giving rise to the largest L will be the best estimate. This is a tedious procedure if carried out manually, but can be performed painlessly if automated through a short computer program. If we plot the exponential of L, $\exp(L)$, on the vertical axis, as a function of α, on the horizontal axis, then we obtain the posterior pdf for the position of the lighthouse. Not only does this give us a complete visual representation of our inference, it has the advantage that we don't need to worry whether the posterior pdf is asymmetric or multimodal.

Before illustrating the use of such an analysis, it is worth making a brief practical point: it is better to calculate L first, as a function of α (on a uniform grid), and then take exponentials, rather than working directly with the posterior pdf itself. The reason is purely numerical and stems from the limit on the size of the number which can be stored in a computer; typically, anything smaller than 10^{-300} tends to be indistinguishable from zero and numbers larger than 10^{+300} are effectively infinite. The reader may be forgiven for failing to see why this should be a problem, since 600 orders of magnitude is an enormous range; after all, probabilities which are more than a factor of 1000 down (10^{-3}) on the maximum are almost imperceptible on a graph and of little real interest anyway. This is certainly true, but it assumes that the maximum value of the posterior probability has already been scaled to a sensible number. In practice, this is often not the case as we usually ignore the normalization constant until the end of the calculation. In terms of eqn (2.38), we tend to omit the constant term from the sum because it affects neither the best estimate nor its error-bar; its value does, however, translate into a multiplicative scale factor for the posterior pdf. A good procedure, therefore, is to evaluate L for a series of different α's, find the maximum $L_{\max}$ and subtract it from all the L's; taking the exponential of these numbers will give a sensible maximum value of one. As well as being useful for plotting purposes, it then enables us to calculate the normalization constant safely, if required, without fear of underflows or overflows in the execution of the program.

Figure 2.9 shows how the posterior pdf for the position of the lighthouse evolves as the number of flashes detected on the coast increases. A random number generator was used to create a uniform sample of azimuths $\{\theta_k\}$, according to eqn (2.32), which were converted into data positions $\{x_k\}$ with eqn (2.33); the lighthouse was taken to be 1 km out at sea ($\beta = 1$), and assumed to be known. In the early panels, the positions of the flashes are marked by small open circles at the top of the graphs; the number of data analysed is shown in the top right-hand corner. Not only is the posterior pdf very broad

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

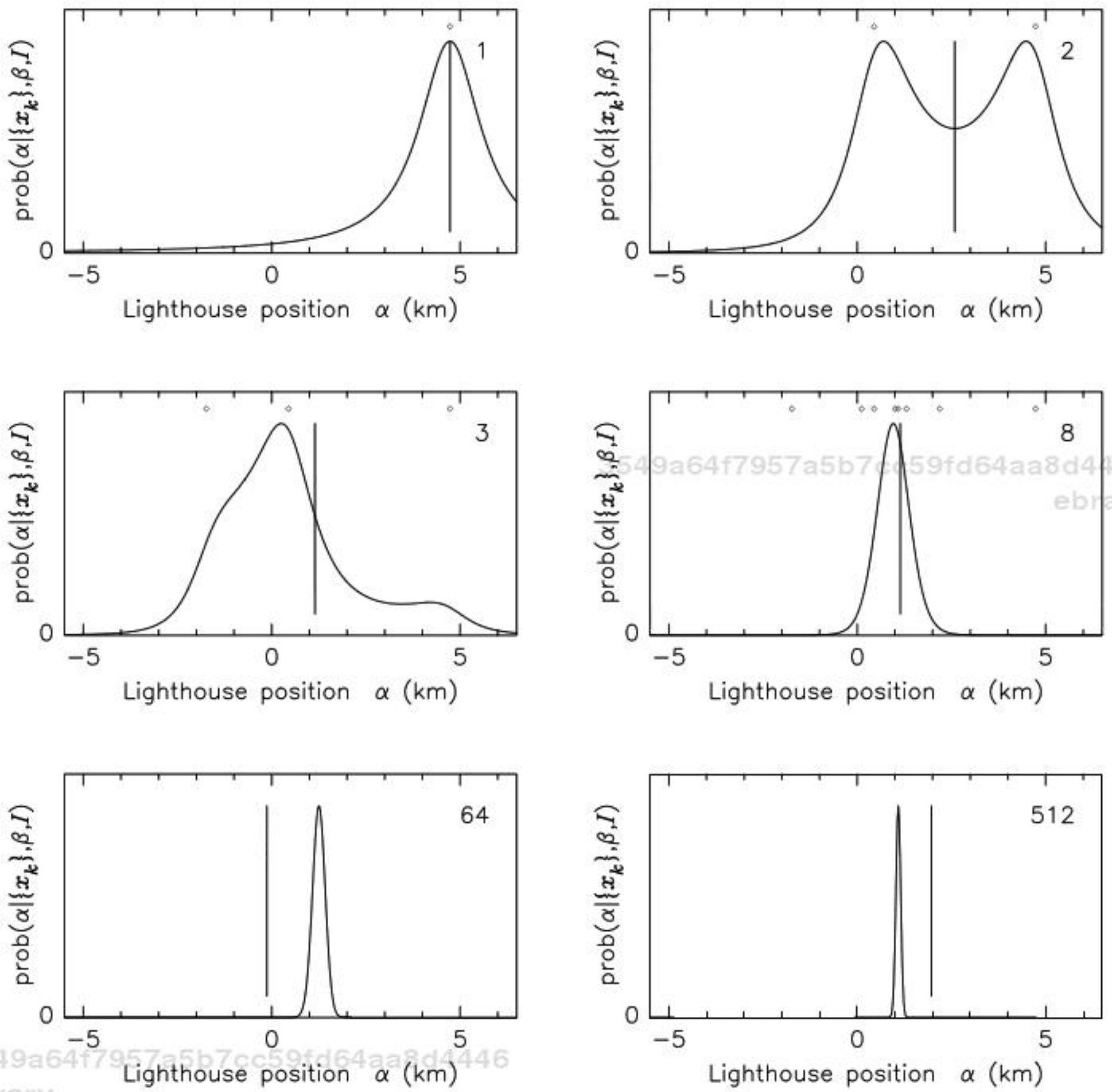


Fig. 2.9 The posterior pdf for the position of the lighthouse along the coast, given that it is 1 km out to sea. The number of data available is shown in the top right-hand corner of the graphs, and their positions are marked by small open circles at the top of the first four panels. The average value of the data is indicated by a long vertical line.

when there are only a few data, it can easily be multimodal if their locations are sufficiently discordant. By about a dozen measurements, however, the posterior becomes a well behaved Gaussian-like pdf. As the experimental evidence mounts, the posterior pdf becomes narrower, and narrower (inversely proportional to $\sqrt{N}$), with the peak converging towards a value of $\alpha = 1$; this is as expected, and quite reassuring, as the data were generated with the lighthouse 1 km along the shore.

2.4.1 The central limit theorem

Before leaving this problem, it is worth pointing out a curious feature about the mean value of the data. In the previous section, we found that the best estimate of the pa-

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

rameter μ of a normal distribution, as in eqn (2.23), was given by the simple arithmetic average of samples drawn from it; stated formally in eqn (2.28), this result seems intuitively reasonable as μ is the mean of the Gaussian pdf. In the present case, the data are drawn from a Cauchy pdf; since this too is symmetric about the point of interest α, we might think that the average value of the data would also provide a good estimate of the position of the lighthouse. We should be a little concerned by this proposal, however, as the sample mean is not the solution to eqn (2.39) for the best estimate α_o. Our suspicions are confirmed in Fig. 2.9, where the average value of the data is indicated by the long vertical line. We see that the sample mean is not a good *estimator* for this problem, since it often lies well outside the range allowed by the posterior pdf.

At first sight, the observation above seems very peculiar; it appears to conflict with our training if not experience. The everyday respectability afforded to averages hinges, formally, on the central limit theorem: this states that if samples are drawn randomly from (almost) any pdf with mean μ, then in the limit of many data N, their average will tend to this value; the error-bar for the difference between μ and the sample mean will also go down like $\sqrt{N}$ (as in eqn 2.29). Technically, it can be shown that the Cauchy distribution is an exception to this rule because it violates one of the underlying assumptions of the central limit theorem. This follows from the fact that means μ and variances σ^2 are additive. On dividing by N to recover averages, the error-bar on the mean diminishes like $\sqrt{N}$. But for this to hold, μ and σ have to exist. For a Cauchy distribution, its very wide wings make σ^2 infinite and μ undefined. In fact, the variability of the mean of the data does not decrease with increasing number of measurements, and is likely to be as 'wrong' after a thousand, or a million, data as it is after just one!

Although the central limit theorem is not applicable to the Cauchy likelihood, this did not prevent us from being able to estimate the position of the lighthouse. To infer the value of α, given β and the detection of flashes at $\{x_k\}$, we needed to consider the posterior pdf $\mathrm{prob}(\alpha|\{x_k\},\beta,I)$; as can be seen from Fig. 2.9, the latter continues to get sharper as the number of data increases. The point is that the central limit theorem is a statement about the pdf for the sample mean: $\mathrm{prob}(\overline{x}|\alpha,\beta,I)$, where $\overline{x}=\sum x_k/N$. If the maximum of the posterior pdf happens to be equal to the average value of the data, then this might be relevant; otherwise, it's of little concern. The moral of this tale is that, despite its everyday respectability, the sample mean is not always a useful number — let the posterior pdf decide what is best.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3 Parameter estimation II

In the previous chapter, we were concerned with the most elementary type of data analysis problem: parameter estimation involving just one unknown variable. Now let us progress to the more common case in which there are several parameters; some of these will be of interest to us, but others merely a nuisance. As well as being a more challenging optimization problem, we will be led to generalize the idea of error-bars to include correlations and to the use of marginalization to deal with the unwanted variables. We will also see how certain approximations naturally give rise to some of the most frequently used analysis procedures, and discuss the so-called propagation of errors.

3.1 Example 4: amplitude of a signal in the presence of background

In many branches of science, we are often faced with the task of having to estimate the amplitude of a signal in the presence of a background. For example, an X-ray diffraction pattern from a crystalline sample will contain both the distinctive Bragg peaks of interest and a general contribution from diffuse scattering processes; in an astronomical setting, the emission spectrum of a galaxy may be contaminated by stray light from the night sky. An idealized one-dimensional situation is illustrated in Fig. 3.1, where the horizontal x-axis represents the measurement variable (such as the scattering angle, wavelength, or whatever). In the simplest case, the background can be taken as being flat, and of unknown magnitude B, and the signal of interest assumed to be the amplitude A of a peak of known shape and position. The data for these problems are frequently integer-valued, and correspond to things like the number of photons detected from a quasar at a certain wavelength or the number of neutrons scattered by a sample in a given direction. Given such a set of counts $\{N_k\}$, measured at experimental settings $\{x_k\}$, what is our best estimate of the amplitude of the signal peak and the background?

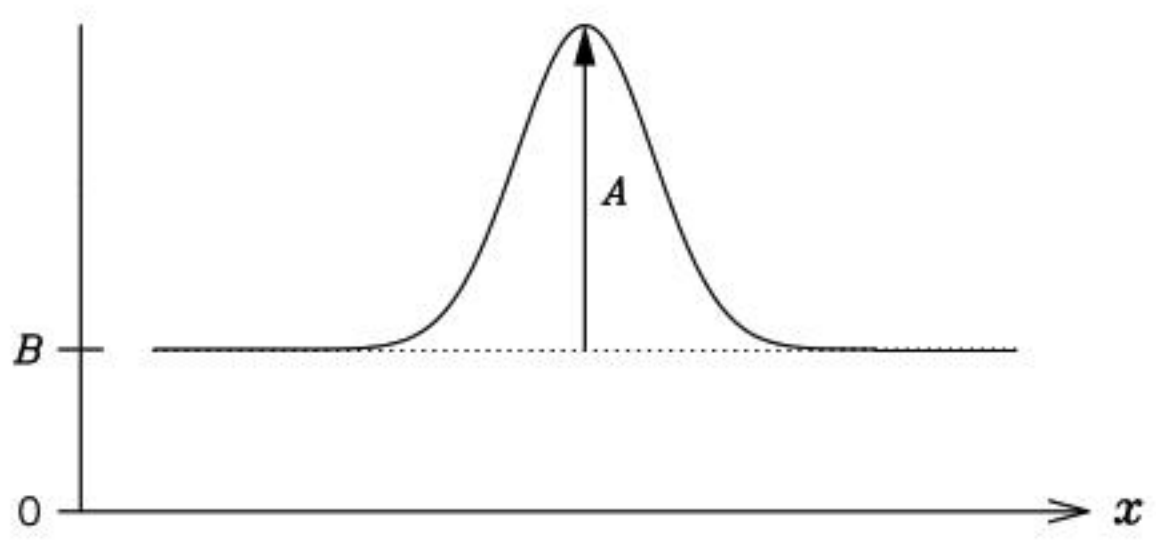


Fig. 3.1 A signal peak of amplitude A sitting on a flat background of magnitude B.

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Let us begin by thinking about the nature of the data a little more carefully. It seems reasonable that we should expect the number of counts in the kth data-channel to be proportional to the sum of the signal and the background at x_k; taking the peak to be Gaussian in shape, for the sake of argument, with a width of w and centred at x_o, the ideal datum D_k ought to be given by:

$$D_k = n_o\left[A\,\mathrm{e}^{-(x_k-x_o)^2/2w^2} + B\right], \tag{3.1}$$

where n_o is a constant related to the amount of time for which the measurements were made. Unlike the number of counts N_k, however, the D_k in eqn (3.1) will not generally be a whole number. Therefore, the actual datum will be an integer ($\geqslant 0$) in the vicinity of this expected value. A pdf which incorporates this property, and is usually invoked in such counting experiments, is a Poisson distribution:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\mathrm{prob}(N|D) = \frac{D^N\,\mathrm{e}^{-D}}{N!}, \tag{3.2}$$

and is illustrated in Fig. 3.2. We leave the derivation of this strange-looking formula to Chapter 5 but note the result that its formal expectation value, as defined by the discrete version of eqn (2.22), is indeed equal to D:

$$\langle N\rangle = \sum_{N=0}^{\infty} N\,\mathrm{prob}(N|D) = D. \tag{3.3}$$

Equation (3.2) is, of course, an assignment for the likelihood of the datum N_k:

$$\mathrm{prob}(N_k|A,B,I) = \frac{D_k^{N_k}\,\mathrm{e}^{-D_k}}{N_k!}, \tag{3.4}$$

where the background information I includes a knowledge of the relationship between the expected number of counts D_k and the parameters of interest A and B; for the case of the Gaussian peak-shape model of eqn (3.1), this means that x_o, w and n_o are taken as given (as well as x_k). If the data are independent, so that, when given the values of

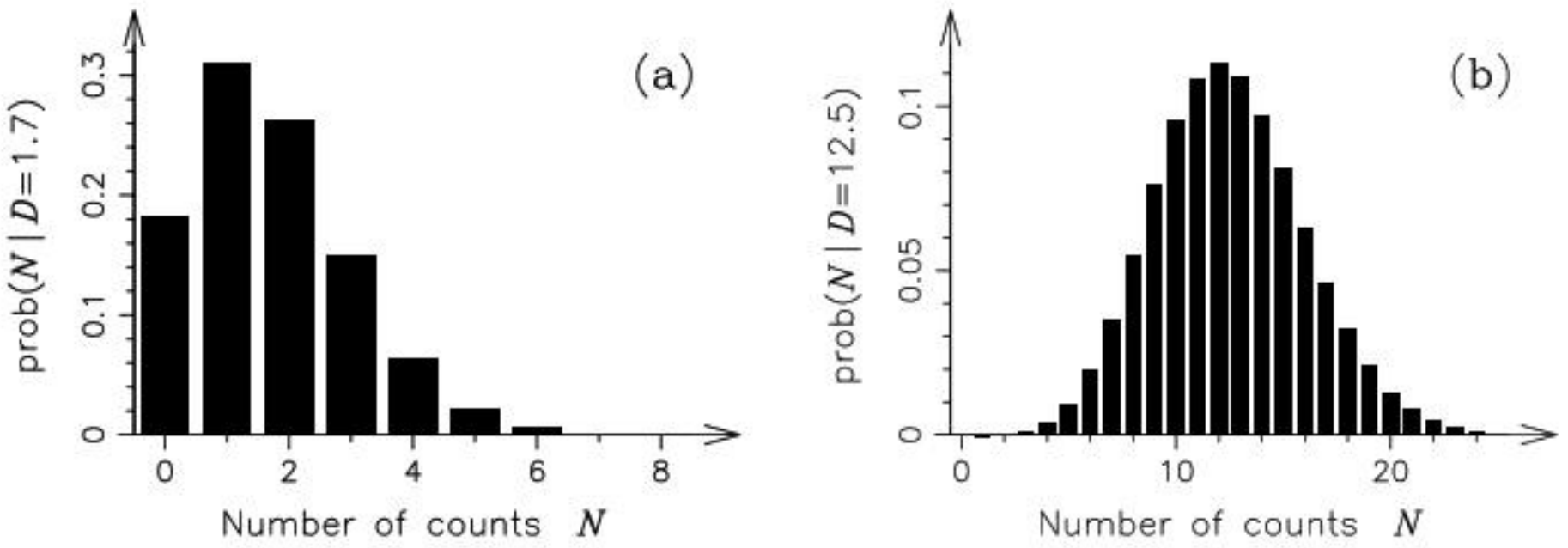


Fig. 3.2 The Poisson distribution $\mathrm{prob}(N|D)$ for: (a) $D=1.7$, and (b) $D=12.5$.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

A and B, the number of counts observed in one channel does not influence what we would expect to find in another, then the likelihood function is just the product of the probabilities for the individual measurements:

$$\text{prob}(\{N_k\}|A,B,I) = \prod_{k=1}^{M} \text{prob}(N_k|A,B,I) \ , \tag{3.5}$$

where there are M data.

Our inference about the amplitude of the signal and the background is embodied in the posterior pdf $\text{prob}(A,B|\{N_k\},I)$; as usual, we use Bayes' theorem to help us calculate it:

$$\text{prob}(A,B|\{N_k\},I) \propto \text{prob}(\{N_k\}|A,B,I) \times \text{prob}(A,B|I) \ . \tag{3.6}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Having already dealt with the likelihood function above, all that remains is the assignment of the prior pdf. Irrespective of the data, the one thing we do know is that the amplitude of neither the signal nor the background can be negative; the most naïve way of encoding this is through a uniform pdf in the positive quadrant:

$$\text{prob}(A,B|I) = \begin{cases} \text{constant} & \text{for } A \geqslant 0 \text{ and } B \geqslant 0 \ , \\ 0 & \text{otherwise} \ . \end{cases} \tag{3.7}$$

Later, in Chapter 5, we will see that this simple flat pdf is not necessarily the one that represents the most prior ignorance for this problem, but it is quite adequate for our present needs. Although we should formally have upper bounds $A_{\max}$ and $B_{\max}$, which determine the normalization constant in eqn (3.7), we will just assume they are sufficiently large as not to impose a cut-off on the posterior pdf; that is to say, in eqn (3.6), $\text{prob}(\{N_k\}|A,B,I)$ will have diminished (gradually) to a negligibly small value before these limits are reached.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Multiplying the Poisson likelihood resulting from eqns (3.4) and (3.5) by the prior of eqn (3.7), according to Bayes' theorem of eqn (3.6), yields the posterior pdf; its logarithm L is given by

$$L = \log_{\mathrm{e}}\big[\text{prob}(A,B|\{N_k\},I)\big] = \text{constant} + \sum_{k=1}^{M} \big[N_k \log_{\mathrm{e}}(D_k) - D_k\big] \ , \tag{3.8}$$

where the constant includes all terms not involving A and B, and the latter are both positive. Our best estimate of the amplitude of the signal and the background is given by the values of A and B which maximize L; its reliability is indicated by the width, or sharpness, of the posterior pdf about this optimal point. Let us illustrate the use of this analysis with some computer examples.

Four sets of data and the posterior pdfs which result from them are shown in Fig. 3.3. The data were generated with a Poisson random number generator, in accordance with eqn (3.4), from the simple 'Gaussian peak on flat background model' of eqn (3.1). They

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

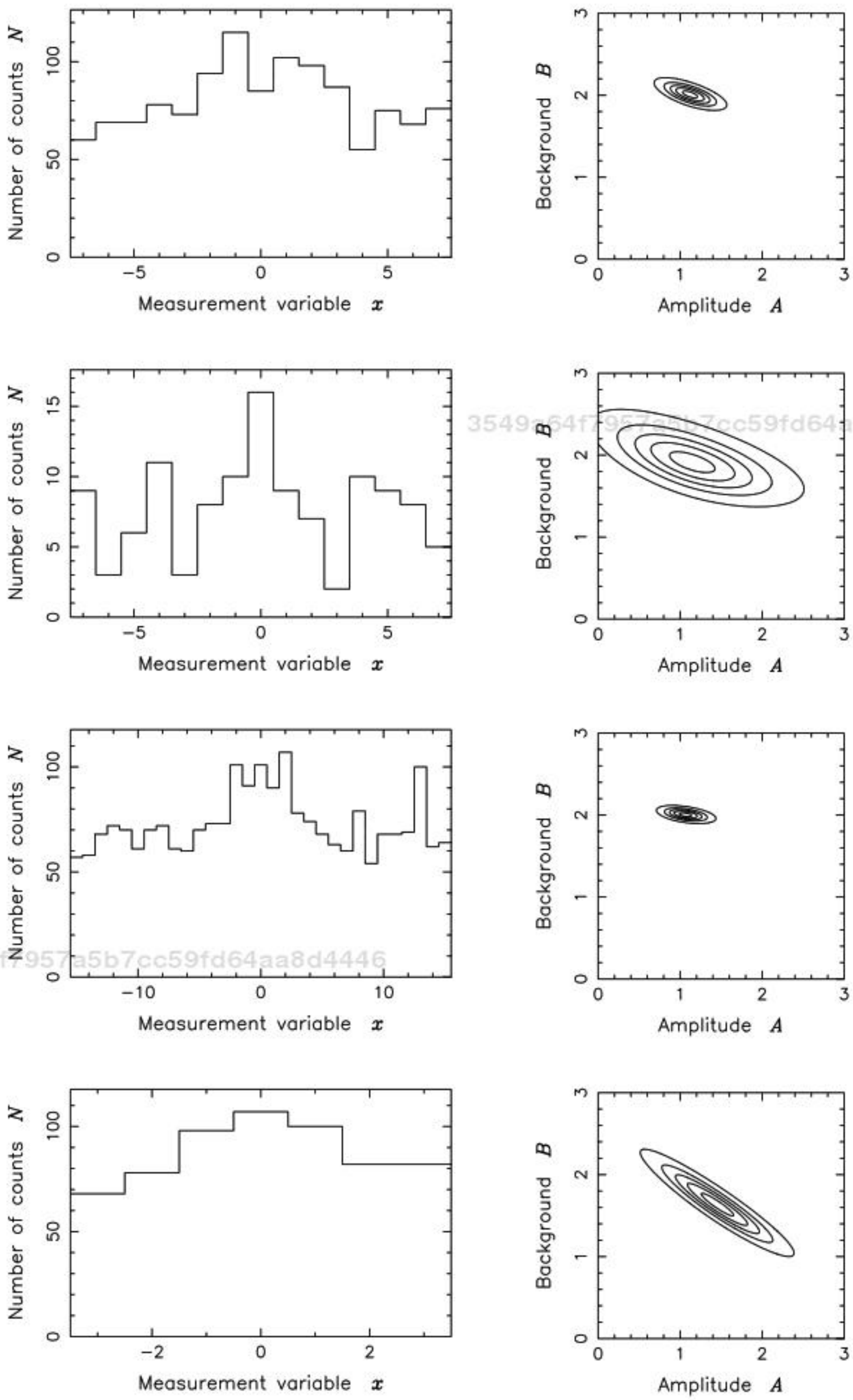


Fig. 3.3 Poisson data and the resulting posterior pdfs for the amplitude A of a Gaussian signal peak, centred at the origin with a FWHM of 5 units, and the flat background B, for four different experimental set-ups.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

are plotted as *histograms*, as experimental measurements would normally correspond to the number of counts detected in finite-sized channels; for simplicity, we have chosen the bin-width to be unity. In all cases, the underlying signal is centred at the origin, so that $x_o = 0$, and has a FWHM of 5 units; this is assumed to be known for the analysis. The posterior pdf is now two-dimensional, since it is a function of both A and B. We can display it using *contours*, which are lines joining points of equal probability density, just like hills and valleys are represented on a topographic map. In Fig. 3.3, the contours correspond to 10%, 30%, 50%, 70% and 90% of the maximum probability.

The first panel shows the number of counts detected in 15 data-bins, where the parameter n_o was chosen to give a maximum expectation of 100 counts; its value is taken as given. The corresponding posterior pdf, resulting from the exponential of eqn (3.8), is plotted in the second panel on the top; it indicates that our best estimate of the amplitude of the signal is approximately equal to one, and is about half the magnitude of the background. The second panel down shows the data for the same set-up, but where the experiment was conducted for only one-tenth of the time; the number of counts is down by a factor of ten, and the data appear more noisy. The corresponding posterior pdf is about three times as broad as before, in both directions, and is in line with the $\sqrt{10}$ that we might have guessed, on the basis of our experience from the examples of the previous chapter. The posterior is truncated (or suddenly equal to zero) for negative values of A and reflects the importance of the prior when the data are of poor quality. In the third panel down we return to the original count-rate, but have 31 data spread over twice the measurement range. With the doubling of the number of data, we might expect the reliability with which A and B can be estimated would improve by a factor of $\sqrt{2}$; although a little hard to tell from these diagrams, this only seems to be true for the background. Some further thought reveals this to be reasonable: measurements far away from the origin only tell us about the background, since they contain no information about the signal peak. The last two panels illustrate the case when there are just 7 data spread over half the original range in x; the posterior pdf is noticeably broadened, by more than a simple factor of $\sqrt{2}$, and is distinctly skewed. These features are indicative of a strong correlation between our estimates of A and B: as the range of x over which the data are collected is severely restricted, it becomes difficult to tell the signal apart from the background!

3.1.1 Marginal distributions

The two-dimensional posterior pdf above describes completely our joint inference about the values of A and B. Often, however, we are not interested in the background. It has to be included in our calculation, because it is needed for the likelihood function, but we have no intrinsic interest in knowing its value: its presence is merely a nuisance. What we would really like to estimate is just the amplitude of the signal, irrespective of the size of the background; in other words, we require the pdf $\mathrm{prob}(A|\{N_k\}, I)$. According to the marginalization rule of eqn (1.4), this can be obtained simply by integrating the joint posterior pdf with respect to B:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(A|\{N_k\}, I) = \int_0^\infty \text{prob}(A, B|\{N_k\}, I)\, \mathrm{d}B\,. \tag{3.9}$$

There may, of course, be situations where the background is of primary interest and the presence of a peak is a nuisance. This seemingly perverse case does sometimes occur, because the phenomena that give rise to broad features in the data may be of as much scientific importance as the ones resulting in a sharp signal. All we need to do then, to obtain the pdf $\text{prob}(B|\{N_k\}, I)$, is integrate the joint posterior pdf with respect to A:

$$\text{prob}(B|\{N_k\}, I) = \int_0^\infty \text{prob}(A, B|\{N_k\}, I)\, \mathrm{d}A\,. \tag{3.10}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

The four pairs of marginal distributions corresponding to the data-sets, and posterior pdfs, of Fig. 3.3 are plotted in Fig. 3.4; it is now much easier to see the effects of the different experimental set-ups on the reliability of the inferred values A and B, which were mentioned a little earlier.

To avoid any confusion, we should emphasize that the marginal distribution for A, $\text{prob}(A|\{N_k\}, I)$, is not the same as the conditional pdf $\text{prob}(A|\{N_k\}, B, I)$; although both describe our inference about the amplitude of the signal peak, they correspond to different circumstances. The former pdf takes into account our prior ignorance of the value of B, while the latter is appropriate when the magnitude of the background has already been determined reliably from a calibration experiment. This conditional pdf, given that $B = 2$, is plotted as a dotted line in Fig. 3.4; compared with its marginal counterpart, we see that there is a significant narrowing of the posterior probability for the last data-set, and least difference for the case when measurements have been made far beyond the tails of the signal peak. When we are able to distinguish between the background and the signal peak fairly well, there is little to be gained from a separate calibration experiment for B; if the data-set is severely truncated, or the background is highly structured, this additional information can be very beneficial.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

For the above analysis, we have assumed that the shape and position of the signal peak is known; in the context of the Gaussian model of eqn (3.1), we took the values of w and x_o as implicitly given in the background information I. How could this condition be relaxed if we were not quite so fortunate? Well, from the discussion about marginalization, it follows that we should integrate out the relevant variables as nuisance parameters:

$$\text{prob}(A, B|\{N_k\}, I) = \iint \text{prob}(A, B, w, x_o|\{N_k\}, I)\, \mathrm{d}w\, \mathrm{d}x_o\,, \tag{3.11}$$

where I still assumes that the model of eqn (3.1) is adequate, but does not necessarily include a knowledge of the width and position of the Gaussian peak. The four-parameter posterior pdf under the double integral can itself be related to two others which are easier to assign by using Bayes' theorem:

$$\text{prob}(A, B, w, x_o|\{N_k\}, I) \propto \text{prob}(\{N_k\}|A, B, w, x_o, I) \times \text{prob}(A, B, w, x_o|I)\,.$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

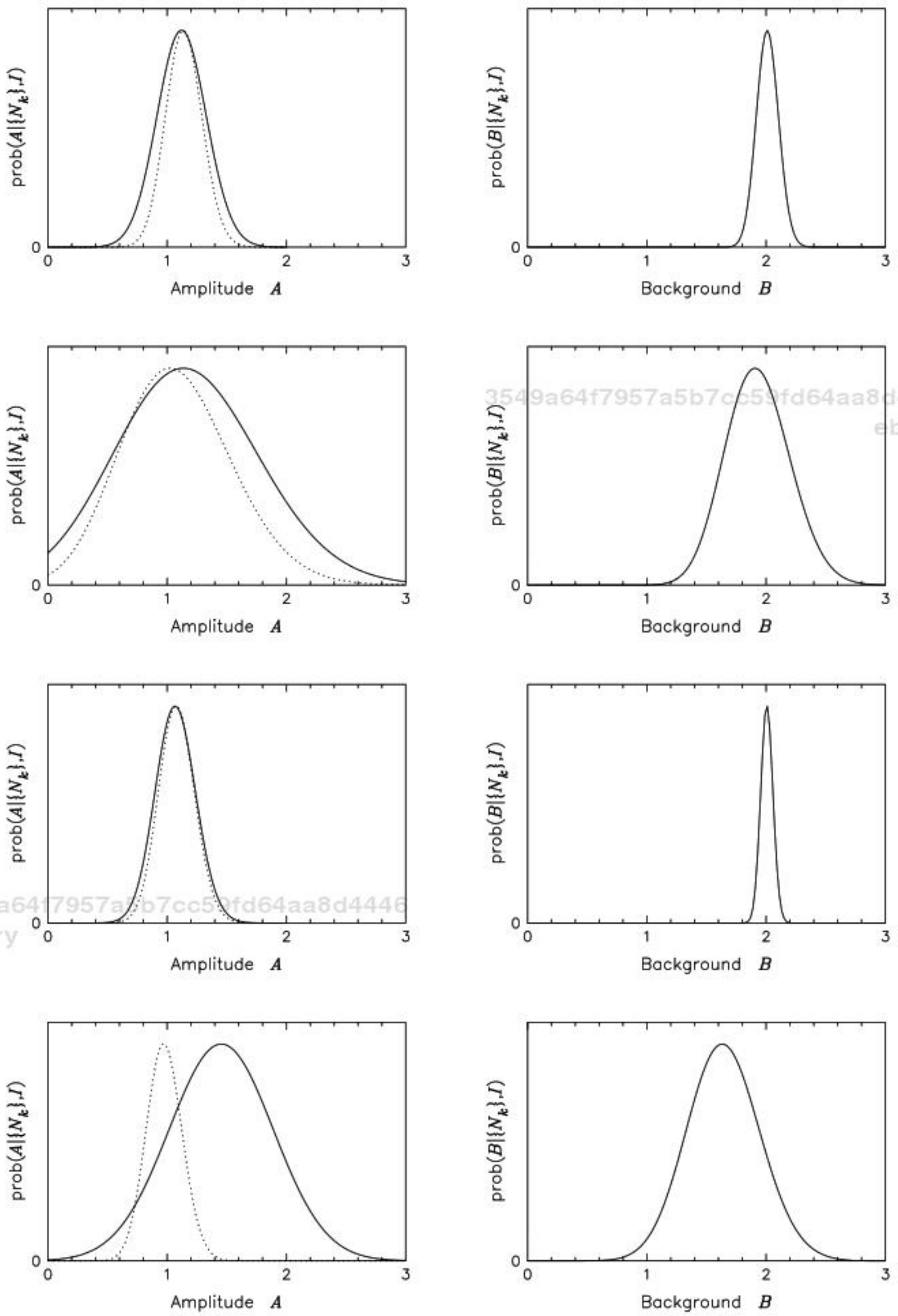


Fig. 3.4 The marginal distributions for the amplitude A and the background B corresponding to the Poisson data, and the resulting posterior pdfs, for the four experimental set-ups in Fig. 3.3. The dotted line shows the posterior pdf for A conditional on knowing the true value of B.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

The first term on the right-hand side is equivalent to the likelihood function of eqns (3.4) and (3.5); the second is the prior pdf for A, B, w and x_o, which can be decomposed as:

$$\text{prob}(A, B, w, x_o|I) = \text{prob}(A, B|I) \times \text{prob}(w, x_o|I) .$$

If we already knew the width and location of the Gaussian peak, then the prior pdf for w and x_o would be very sharp. In the limit of absolute certainty, it would become

$$\text{prob}(w, x_o|I) = \delta(w - 2.12)\, \delta(x_o) ,$$

where the two δ-functions are equal to zero unless $w = 2.12$ (i.e. FWHM$=5$) and $x_o = 0$ respectively; in this case, the integral of eqn (3.11) is very easy to evaluate and yields

$$\text{prob}(A, B|\{N_k\}, I) \propto \text{prob}(\{N_k\}|A, B, w{=}2.12, x_o{=}0, I) \times \text{prob}(A, B|I) .$$

Thus, as expected, the analysis reduces to that of eqn (3.6). If we do not know the values of w and x_o, however, we must assign a suitably broad prior for these parameters (as well as for A and B). The marginalization integral will then require more work, but can either be computed numerically or approximated analytically, as we shall see in Section 3.2. Although probability theory allows us to deal with such experimental uncertainties, we do, of course, expect the marginal distribution $\text{prob}(A, B|\{N_k\}, I)$ to be broader than the conditional pdf $\text{prob}(A, B|\{N_k\}, w, x_o, I)$; this would merely reflect the corresponding lack of relevant information.

3.1.2 Binning the data

When the data were plotted in Fig. 3.3 using histograms, we said that this was because experimental measurements often corresponded to the number of counts detected in channels of finite width. This means that eqn (3.1), for the expected value of the datum D_k, should actually be written as an integral over the kth data-bin:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$D_k = \int_{x_k - \Delta/2}^{x_k + \Delta/2} n_o \left[A\, e^{-(x - x_o)^2/2w^2} + B \right] dx , \tag{3.12}$$

where we have taken all the measurement channels to have the same width Δ. As long as the bin-width is not too large, so that the integrand varies only linearly across it, the integral of eqn (3.12) can be approximated well by the area of a trapezoidal column:

$$D_k \approx n_o \left[A\, e^{-(x_k - x_o)^2/2w^2} + B \right] \Delta . \tag{3.13}$$

Thus, eqn (3.1) is justified because the fixed number Δ can be absorbed into n_o. This new (redefined) constant reflects both the amount of time for which the experimental measurements were made and the size of the 'collecting area'. The bin-width Δ is not always determined by the physical size of the detectors, however, but is often chosen to be large enough so that there are a reasonable number of counts in the resulting composite data-channels. Is there anything to be gained, or lost, in this binning process?

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

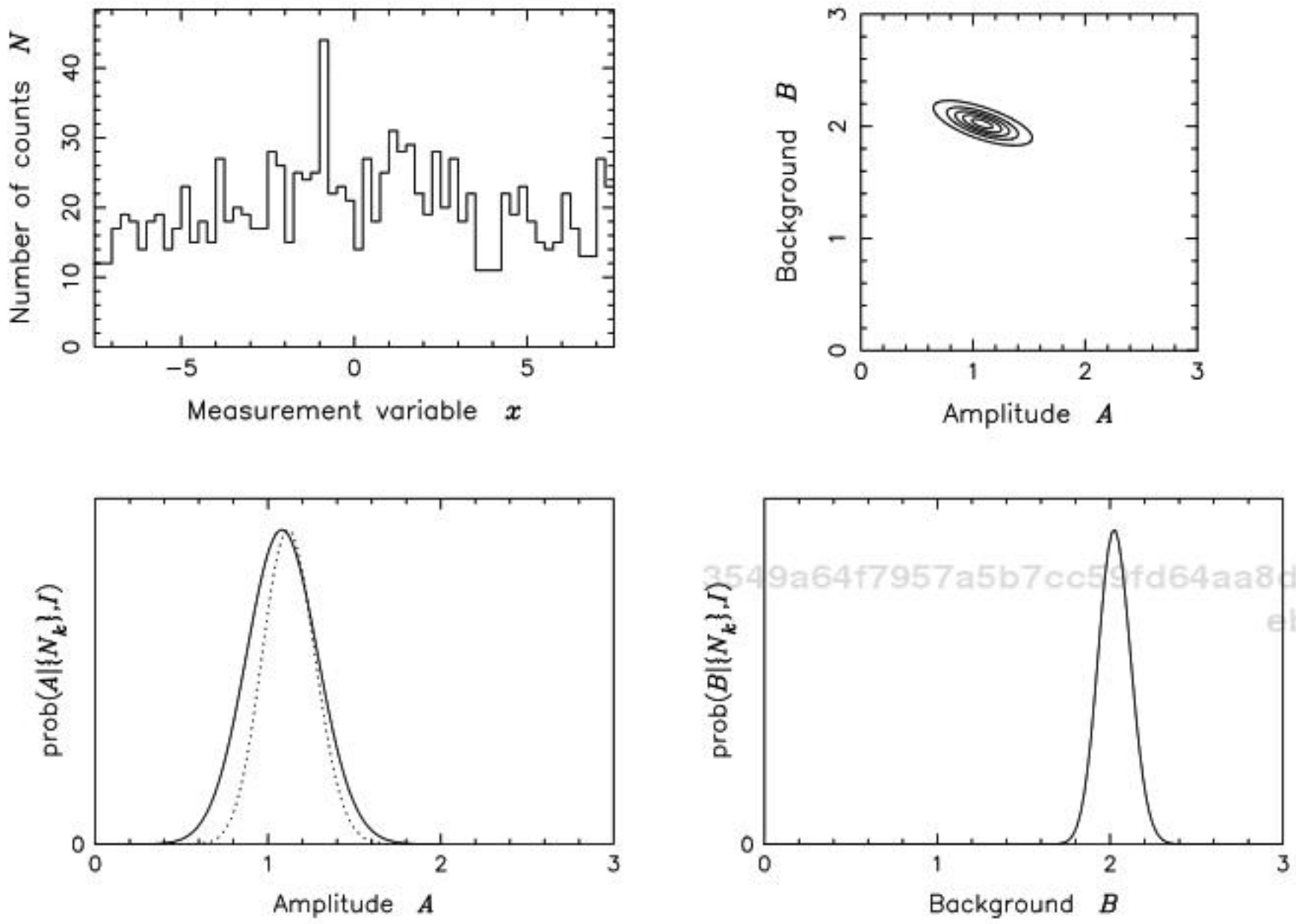


Fig. 3.5 The analysis of data corresponding to the experimental set-up of the first panel in Fig. 3.3, but with bins which are four times narrower.

Figure 3.5 shows the data for the same experimental set-up as in the first panel of Fig. 3.3, but where the bins are four times narrower than before; this data-set looks noisier because, on average, each channel has only a quarter of the previous number of counts. Figure 3.5 also shows the resulting posterior pdf for A and B, and its marginal distributions, given that $w = 2.12$ and $x_o = 0$. Compared with the corresponding pdfs in Figs 3.3 and 3.4, the reliability of the inferred parameters is virtually identical. This illustrates that the perceived improvement in the data quality afforded by the binning procedure is largely cosmetic. While this visual gain is not without value, and there are computational advantages in dealing with fewer 'measurements', we must always be aware that too coarse a binning can destroy useful information in the data. To take an extreme case, if all the counts were added together into a single number, we would completely lose all our ability to distinguish the signal from the background! At a more mundane level, if the data-channels are too wide, we could also get into trouble with the integral approximation of eqn (3.13). Some of the issues related to optimal binning strategies are discussed further in Chapter 7.

3.2 Reliabilities: best estimates, correlations and error-bars

We have now seen an example of an estimation problem involving more than one parameter. Although the posterior pdf is of a higher dimensionality, being a function of several variables, it still encodes our inference about the values of the parameters, given the data and the relevant background information. As before, we often wish to summarize

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

it with just a few numbers: the best estimates and a measure of their reliabilities. Since the probability (density) associated with any particular set of values for the parameters is a measure of how much we believe that they (really) lie in the neighbourhood of those values, our optimal estimate is given by the maximum of the posterior pdf. If we denote the quantities of interest by $\{X_j\}$, with a posterior pdf $P = \text{prob}(\{X_j\}|\{data\}, I)$, then the best estimate of their values, $\{X_{\text{O}j}\}$, is given by the solution to the set of *simultaneous equations*:

$$\left.\frac{\partial P}{\partial X_i}\right|_{\{X_{\text{O}j}\}} = 0\,, \tag{3.14}$$

where $i = 1, 2, \ldots$, up to the number of parameters to be inferred. Strictly speaking, we also need a further test, analogous to eqn (2.9), to ensure that we are dealing with a maximum and not a minimum or a *saddle-point*; we'll say more about that shortly. In writing the *partial derivatives* of P with respect to the X_i's we are, of course, assuming that they can all lie in a continuous range. If some could only take discrete values, our best estimate would still correspond to the one with greatest posterior probability but we couldn't then use the calculus notation of eqn (3.14).

As in Section 2.2, it is more convenient to work with the logarithm of P rather than use the posterior pdf itself:

$$L = \log_{\text{e}}\big[\,\text{prob}(\{X_j\}|\{data\}, I)\,\big]\,. \tag{3.15}$$

Since the logarithm is a monotonic function, the maximum of L occurs at the same place as that of P; hence, eqn (3.14) can be written with L substituted for P. Rather than pursuing this analysis in generality, with several parameters, let us first consider the specific case of just two variables; we will denote them by X and Y, instead of X_1 and X_2, to reduce the multiplicity of subscripts. The pair of simultaneous equations which we must now solve, to obtain our best estimates X_{O} and Y_{O}, are given by

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\left.\frac{\partial L}{\partial X}\right|_{X_{\text{O}},Y_{\text{O}}} = 0 \qquad \text{and} \qquad \left.\frac{\partial L}{\partial Y}\right|_{X_{\text{O}},Y_{\text{O}}} = 0\,, \tag{3.16}$$

where $L = \log_{\text{e}}\big[\,\text{prob}(X, Y|\{data\}, I)\big]$.

To obtain a measure of the reliability of this best estimate, we need to look at the spread of the two-dimensional posterior pdf about the point $(X_{\text{O}}, Y_{\text{O}})$. As in Section 2.2, we can analyse the local behaviour of a (potentially) complicated function by using a Taylor series expansion:

$$L = L(X_{\text{O}}, Y_{\text{O}}) + \frac{1}{2}\left[\left.\frac{\partial^2 L}{\partial X^2}\right|_{X_{\text{O}},Y_{\text{O}}}(X - X_{\text{O}})^2 + \left.\frac{\partial^2 L}{\partial Y^2}\right|_{X_{\text{O}},Y_{\text{O}}}(Y - Y_{\text{O}})^2 + 2\left.\frac{\partial^2 L}{\partial X\,\partial Y}\right|_{X_{\text{O}},Y_{\text{O}}}(X - X_{\text{O}})(Y - Y_{\text{O}})\right] + \cdots\,, \tag{3.17}$$

where X_{O} and Y_{O} are given by the condition of eqn (3.16). Although this expression looks quite horrendous, it is simply the two-dimensional version of eqn (2.11); it's just

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

that there are now four second (partial) derivatives to deal with, instead of only one! We have, in fact, reduced this tally of terms to three by using the equality of the mixed derivatives: $\partial^2 L/\partial X\,\partial Y = \partial^2 L/\partial Y\,\partial X$. The first term in the Taylor series, $L(X_\mathrm{O}, Y_\mathrm{O})$, is a constant and tells us nothing about the shape of the posterior pdf. The two linear terms, which would be proportional to $X - X_\mathrm{O}$ and $Y - Y_\mathrm{O}$, are missing because we are expanding about the maximum (as indicated by eqn 3.16). The three quadratic terms are, therefore, the dominant factors determining the width of the posterior pdf, and play a central rôle in the reliability analysis; let us study them more closely.

To aid the generalization of this discussion to the case of several variables a little later, let us rewrite the quadratic part of eqn (3.17) in *matrix* notation; calling the quantity in the square brackets Q, we have

$$Q = \begin{pmatrix} X - X_\mathrm{O} & Y - Y_\mathrm{O} \end{pmatrix} \begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} X - X_\mathrm{O} \\ Y - Y_\mathrm{O} \end{pmatrix}, \tag{3.18}$$

where the components of the 2×2 *symmetric* matrix in the middle are given by the second derivatives of L, evaluated at the maximum $(X_\mathrm{O}, Y_\mathrm{O})$:

$$A = \left.\frac{\partial^2 L}{\partial X^2}\right|_{X_\mathrm{O}, Y_\mathrm{O}}, \quad B = \left.\frac{\partial^2 L}{\partial Y^2}\right|_{X_\mathrm{O}, Y_\mathrm{O}}, \quad C = \left.\frac{\partial^2 L}{\partial X \partial Y}\right|_{X_\mathrm{O}, Y_\mathrm{O}}. \tag{3.19}$$

Figure 3.6 shows a contour of Q in the X–Y plane; within our quadratic approximation, it is also a line along which the posterior pdf is constant. It is an *ellipse*, centred at $(X_\mathrm{O}, Y_\mathrm{O})$, the orientation and *eccentricity* of which are determined by the values of A, B and C; for a given contour-level ($Q = k$), they also govern its size. The directions of the *principal axes* formally correspond to the *eigenvectors* of the second-derivative

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

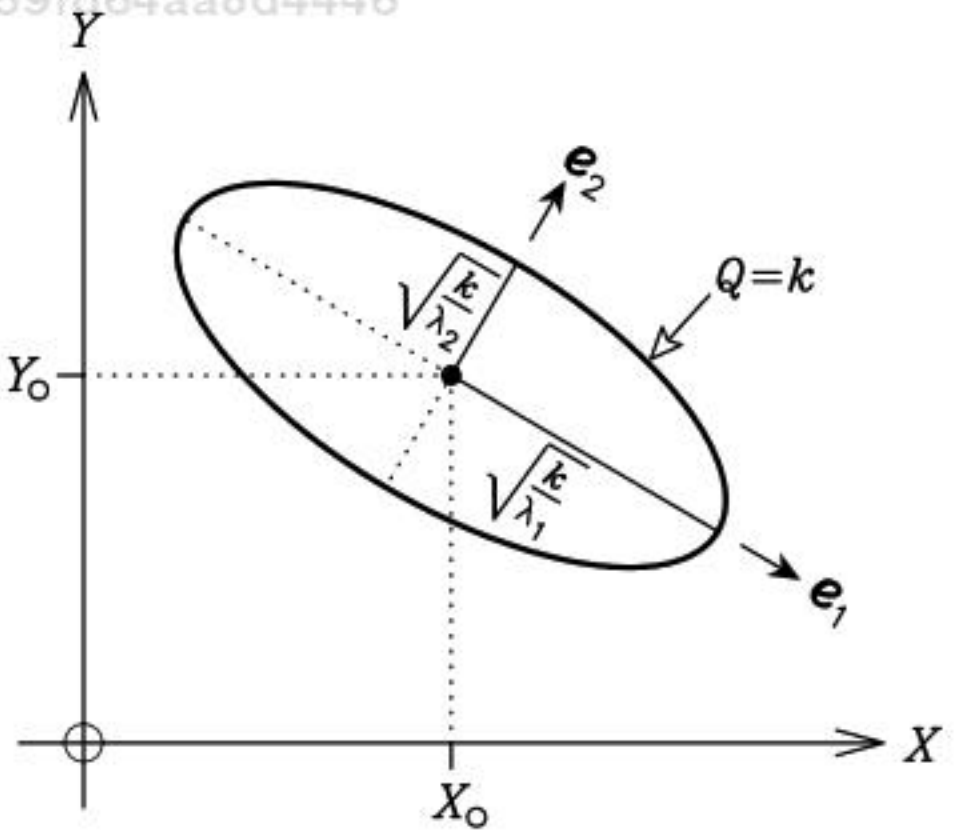


Fig. 3.6 The contour in the X–Y parameter space along which $Q = k$, a constant. It is an ellipse, centred at $(X_\mathrm{O}, Y_\mathrm{O})$, the characteristics of which are determined by the eigenvalues λ and eigenvectors e of the second-derivative matrix defined in eqns (3.18)–(3.20).

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

matrix of eqns (3.18) and (3.19); that is to say, the (x, y) components of $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ in Fig. 3.6 are given by the solutions of the equation

$$\begin{pmatrix} A & C \\ C & B \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \lambda \begin{pmatrix} x \\ y \end{pmatrix} . \tag{3.20}$$

The two *eigenvalues* λ_1 and λ_2 which satisfy eqn (3.20) are, in turn, inversely related to the square of the widths of the ellipse along its principal directions. Furthermore, if the point $(X_\mathrm{O}, Y_\mathrm{O})$ in eqn (3.16) is to be a maximum rather than a minimum or a saddle-point, both λ_1 and λ_2 must be negative; in terms of A, B and C, this requirement takes the form

$$A < 0 , \qquad B < 0 , \qquad AB > C^2 .$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Returning to the question of the reliability of our best estimates, this would be easy to define if the ellipse in Fig. 3.6 was aligned with the X and Y coordinate axes; this is so when $C = 0$. In that case, the error-bars for X_O and Y_O will just be inversely proportional to the square root of (the modulus of) the appropriate eigenvalues (which are then simply A and B). What should we do if the ellipse is skewed?

If we were only interested in knowing the value of X, then we would integrate out Y as a nuisance parameter:

$$\mathrm{prob}(X|\{data\}, I) = \int_{-\infty}^{+\infty} \mathrm{prob}(X, Y|\{data\}, I)\, \mathrm{d}Y .$$

This integral can be done analytically within the quadratic approximation of eqns (3.17)–(3.19): $\mathrm{prob}(X, Y|\{data\}, I) = \exp(L) \propto \exp(Q/2)$. Assuming that the joint posterior pdf is not (significantly) truncated by the bounds of the prior, Appendix A shows the result to be

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\mathrm{prob}(X|\{data\}, I) \propto \exp\left(\frac{1}{2} \left[\frac{AB - C^2}{B} \right] (X - X_\mathrm{O})^2 \right) ,$$

where we have omitted the normalization constant. Comparing this to eqns (2.13) and (2.14), we see that the marginal distribution for X is just a simple one-dimensional Gaussian pdf; our best estimate is still X_O, and its associated error-bar σ_X is given by

$$\sigma_X = \sqrt{\frac{-B}{AB - C^2}} . \tag{3.21}$$

We can obtain an analogous result for Y, yielding a marginal error-bar σ_Y:

$$\sigma_Y = \sqrt{\frac{-A}{AB - C^2}} . \tag{3.22}$$

While the expressions for σ_X and σ_Y above provide a useful measure of the reliability of our best estimates, in general, they paint an incomplete picture. To understand this,

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

we must consider the denominator of eqns (3.21) and (3.22) in a bit more detail. The term $AB-C^2$ is, in fact, the *determinant* of the matrix in eqn (3.20); for such a real symmetric matrix, this is equal to the product of its eigenvalues. Thus if either λ_1 or λ_2 becomes very small, so that the ellipse in Fig. 3.6 is extremely elongated in one of its principal directions, then $AB-C^2 \to 0$; consequently, apart from the special case when $C=0$, both σ_X and σ_Y will be huge. Although this correctly warns us that neither X nor Y can be inferred reliably on the basis of the current data, it fails to tell us that there could still be some joint aspect of the two which can be determined well, because the posterior pdf might be very sharp in one direction while being extremely broad in the other. To see how this information can be conveyed, let us first look at the concept of error-bars again in a slightly different way.

So far, from eqn (2.14), we have thought of the error-bar as representing the width of a Gaussian pdf: FWHM $\approx 2.35\,\sigma$. Another way to think about it is in terms of the *variance* of the posterior, which also gives a measure of its spread. It is formally defined to be the expectation value of the square of the deviations from the mean; assuming a normalized pdf, this is given by

$$\mathrm{Var}(X) = \left\langle (X-\mu)^2 \right\rangle = \int (X-\mu)^2 \,\mathrm{prob}(X|\{data\},I)\,\mathrm{d}X\,, \tag{3.23}$$

where μ is the average value $\langle X \rangle$, as in eqn (2.22). For the one-dimensional normal distribution of eqn (2.14), this integral yields the result

$$\left\langle (X-\mu)^2 \right\rangle = \sigma^2\,. \tag{3.24}$$

The square root of the variance is called the *standard deviation*, or the *root-mean-square* (r.m.s.) error, of X. This definition of the error-bar can be extended to pdfs of more than one variable. Explicitly, for the two-dimensional case we have been considering,

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\sigma_X^2 = \left\langle (X-X_{\mathrm{O}})^2 \right\rangle = \iint (X-X_{\mathrm{O}})^2 \,\mathrm{prob}(X,Y|\{data\},I)\,\mathrm{d}X\,\mathrm{d}Y\,, \tag{3.25}$$

where σ_X is the same as in eqn (3.21) when the double integral is evaluated within the quadratic approximation of eqns (3.17)–(3.19). The corresponding expression for σ_Y is similar, with the X's and Y's swapped around.

The idea of variance can be broadened to consider the simultaneous deviations of both X and Y; this *covariance*, which we will denote as σ_{XY}^2, is given by

$$\begin{aligned}\sigma_{XY}^2 &= \left\langle (X-X_{\mathrm{O}})(Y-Y_{\mathrm{O}}) \right\rangle \\ &= \iint (X-X_{\mathrm{O}})(Y-Y_{\mathrm{O}}) \,\mathrm{prob}(X,Y|\{data\},I)\,\mathrm{d}X\,\mathrm{d}Y\,,\end{aligned} \tag{3.26}$$

and is a measure of the *correlation* between the inferred parameters. If an over-estimate of one usually leads to a larger than average value for the other, then the difference

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$Y - Y_O$ will tend to be positive when $X - X_O$ is positive; if the same is true for under-estimates, so that $Y - Y_O$ is usually negative when $X - X_O$ is as well, the expectation value of the product of the deviations will be positive: the covariance will then be greater than zero. If there is an anti-correlation, so that an over-estimate of one is accompanied by an under-estimation of the other, then the covariance will be negative. When our estimate of one parameter has little, or no, influence on the inferred value of the other, then the magnitude of the covariance will be negligible in comparison to the variance terms; in other words, $|\sigma_{XY}^2| \ll \sqrt{\sigma_X^2\,\sigma_Y^2}$.

When the double integral of eqn (3.26) is evaluated within the quadratic approximation of eqns (3.17)–(3.19), it yields the result

$$\sigma_{XY}^2 = \frac{C}{AB - C^2}. \tag{3.27}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

In conjunction with eqns (3.21) and (3.22), therefore, we see that both the variance and covariance terms are given by (minus) the elements of the *inverse* of the second-derivative matrix of eqn (3.20):

$$\begin{pmatrix} \sigma_X^2 & \sigma_{XY}^2 \\ \sigma_{XY}^2 & \sigma_Y^2 \end{pmatrix} = \frac{1}{AB - C^2}\begin{pmatrix} -B & C \\ C & -A \end{pmatrix} = -\begin{pmatrix} A & C \\ C & B \end{pmatrix}^{-1}. \tag{3.28}$$

This table of 'error-bar products' is called the *covariance matrix*. When $C = 0$, σ_{XY}^2 also equals zero and means that the inferred values of the parameters are uncorrelated. In that case, the principal directions of the corresponding posterior pdf will be parallel to the coordinates axes; this situation is illustrated in Fig. 3.7(a). As the magnitude of C increases, relative to A and B, the posterior pdf becomes more and more skew and elongated; this reflects the growing strength of the correlation between our estimates

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

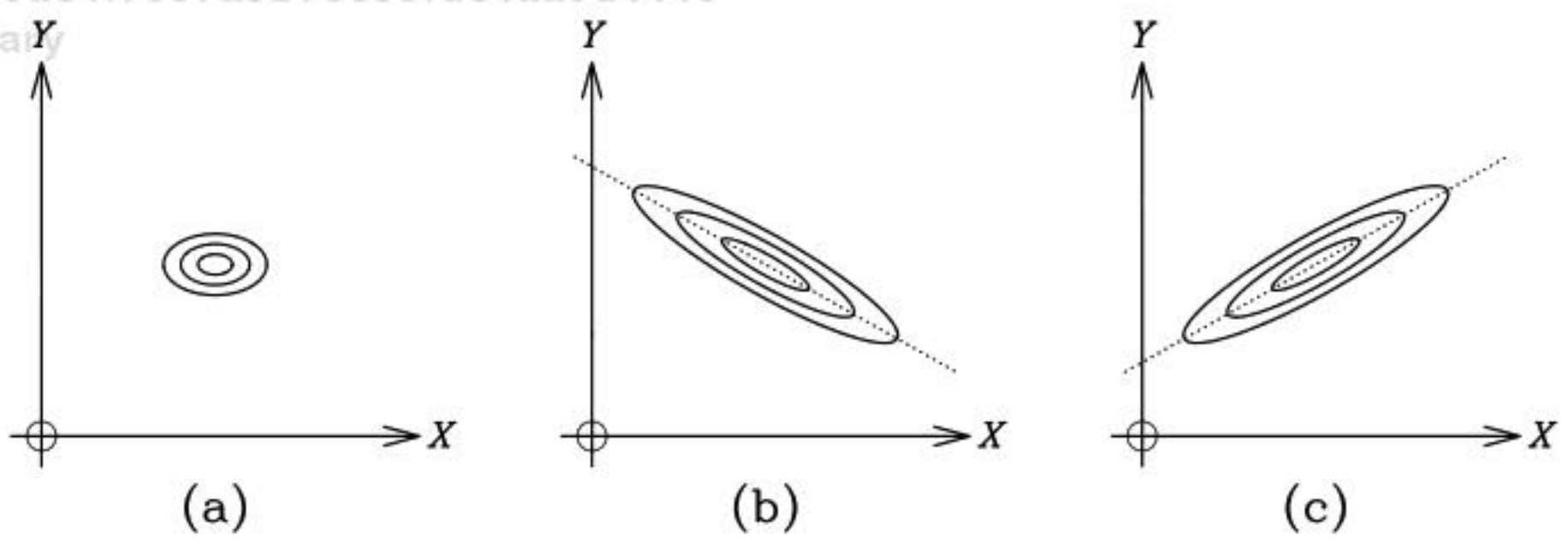


Fig. 3.7 A schematic illustration of covariance and correlation. (a) The contours of a posterior pdf with zero covariance, where the inferred values of X and Y are uncorrelated. (b) The corresponding plot when the covariance is large and negative; $Y + mX =$ constant along the dotted line (where $m > 0$), emphasizing that only this sum of the two parameters can be inferred reliably. (c) The case of positive correlation, where we learn most about the difference $Y - mX$; this is constant along the dotted line.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

of X and Y, and is shown schematically in Figs 3.7(b) and (c). In the extreme case, when $C = \pm\sqrt{AB}$, the elliptical contours will be infinitely wide in one direction (with only the information in the prior preventing this catastrophe!) and oriented at an angle $\pm\tan^{-1}\sqrt{A/B}$ with respect to the X-axis. Although the error-bars σ_X and σ_Y will be huge, saying that our individual estimates of X and Y are completely unreliable, the large off-diagonal elements of the covariance matrix tells us that we can still infer a linear combination of these parameters quite well. If the covariance is negative, then the posterior pdf will be very broad in the direction $Y = -mX$, where $m = \sqrt{A/B}$, but fairly narrow perpendicular to it; this bunching of the probability contours along the lines of '$Y + mX =$ a constant' indicates that the data contain a lot of information about the sum $Y + mX$, but little about the difference $Y - X/m$. Similarly, when the covariance is positive, we can infer the difference $Y - mX$ but not the sum $Y + X/m$.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3.2.1 Generalization of the quadratic approximation

Let us generalize the *bivariate* analysis of the last few pages to the case when there are several variables, and use the opportunity to summarize the main results. The condition for obtaining our best estimate was given in eqn (3.14), and was chosen to be the one which maximized the posterior pdf for the set of M parameters $\{X_j\}$. In terms of its logarithm L, of eqn (3.15), this requirement for the optimal values $\{X_{\mathrm{O}j}\}$, or $\boldsymbol{X}_{\mathrm{O}}$ for simplicity, can be expressed by the simultaneous equations

$$\left.\frac{\partial L}{\partial X_i}\right|_{\boldsymbol{X}_{\mathrm{O}}} = 0\,, \tag{3.29}$$

where $i = 1, 2, \ldots, M$. For the *multivariate* case, the extension of the Taylor series expansion of eqn (3.17) takes the form

$$L = L(\boldsymbol{X}_{\mathrm{O}}) + \frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\left.\frac{\partial^2 L}{\partial X_i \partial X_j}\right|_{\boldsymbol{X}_{\mathrm{O}}}(X_i - X_{\mathrm{O}i})(X_j - X_{\mathrm{O}j}) + \cdots\,. \tag{3.30}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

If the elements of a *column* matrix $(M \times 1)$ are regarded as the components of a vector, and we ignore higher-order terms in the Taylor series expansion, then the exponential of eqn (3.30) yields the following approximation for the posterior pdf:

$$\mathrm{prob}(\boldsymbol{X}|\{data\}, I) \propto \exp\left[\frac{1}{2}\left(\boldsymbol{X} - \boldsymbol{X}_{\mathrm{O}}\right)^{\mathrm{T}} \boldsymbol{\nabla}\boldsymbol{\nabla} L(\boldsymbol{X}_{\mathrm{O}})\left(\boldsymbol{X} - \boldsymbol{X}_{\mathrm{O}}\right)\right], \tag{3.31}$$

where $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ is the symmetric $M \times M$ matrix of second-derivatives, whose ijth element is $\partial^2 L/\partial X_i \partial X_j$, and the *transpose* T of $\boldsymbol{X} - \boldsymbol{X}_{\mathrm{O}}$ is a *row* vector. As such, the quadratic exponent is just a generalization of Q in eqn (3.18) and we can think of Fig. 3.6 as being a two-dimensional slice through it. The pdf of eqn (3.31) is called a *multivariate* Gaussian, since it is a multidimensional version of eqns (2.13) and (2.14). Assuming that the bounds of the prior do not cause a significant truncation of this M-dimensional probability 'bump', Appendix A shows the associated constant of proportionality to be given by $(2\pi)^{-M/2}$ times the square root of the determinant of the $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ matrix; this

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

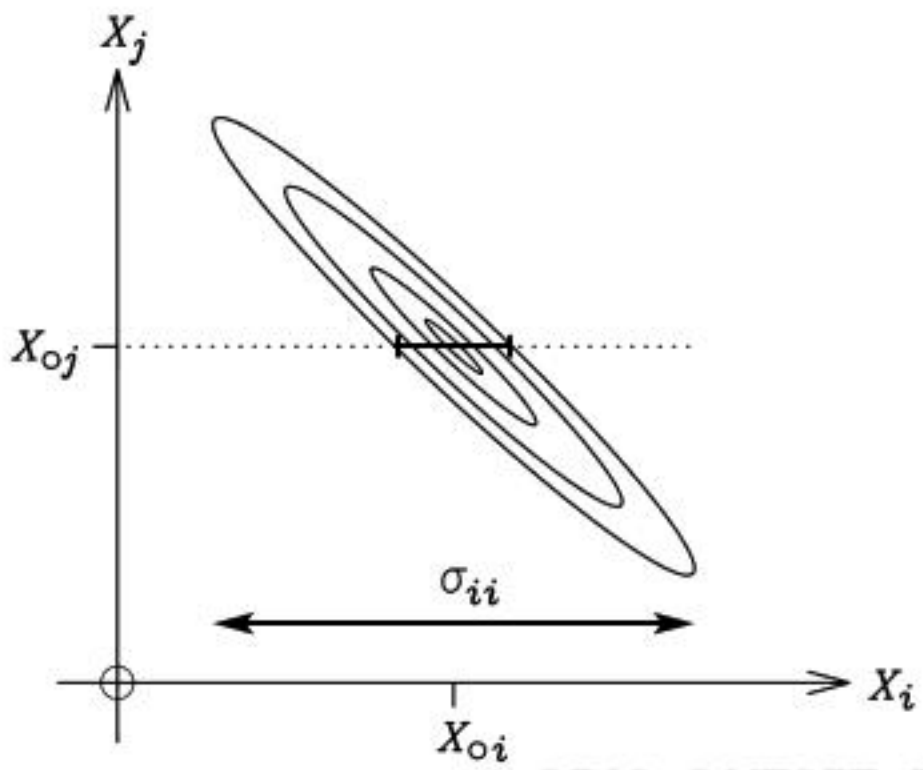


Fig. 3.8 An illustration of the difference between the (correct) marginal error-bar σ_{ii} and its (misleading) 'best-fit' counterpart indicated by the bold line in the middle. The former is given by the ith diagonal element of $(\nabla\nabla L)^{-1}$, while the latter is related to the inverse of $\partial^2 L/\partial X_i{}^2$.

ensures the correct normalization, so that the multidimensional integral of the posterior pdf with respect to all the parameters $\{X_j\}$ is equal to unity.

It will come as no surprise that the maximum of the multivariate Gaussian is defined by the vector $\boldsymbol{X}_{\mathrm{O}}$; the condition for finding its components, in eqn (3.29), can be written compactly as: $\nabla L(\boldsymbol{X}_{\mathrm{O}})=0$. By comparison with the standard one-dimensional Gaussian of eqn (2.14), we see that $\nabla\nabla L$ is analogous to $-1/\sigma^2$; this suggests that the spread (or 'width') of the posterior should be related to the inverse of the second-derivative matrix. Indeed, as shown in Appendix A, the covariance matrix $\boldsymbol{\sigma}^2$ is given by minus the inverse of $\nabla\nabla L$ (evaluated at $\boldsymbol{X}_{\mathrm{O}}$):

$$\left[\boldsymbol{\sigma}^2\right]_{ij} = \left\langle (X_i - X_{\mathrm{O}i})(X_j - X_{\mathrm{O}j}) \right\rangle = -\left[(\nabla\nabla L)^{-1}\right]_{ij}, \tag{3.32}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

and is a generalization of eqns (3.26) and (3.28). The square root of the diagonal elements ($i=j$) corresponds to the (marginal) error-bars for the associated parameters; the off-diagonal components ($i \neq j$) tell us about the correlations between the inferred values of X_i and X_j.

Before concluding our discussion of the quadratic approximation, it is worth emphasizing the fact that the inverse of the diagonal elements of a matrix are not, in general, equal to the diagonal elements of its inverse. Stated in this direct way, few of us would expect to make the mistake. We would be pursuing this folly inadvertently, however, if we were foolishly tempted to estimate the reliability of one parameter in a multivariate problem by holding all the others fixed at their optimal values. The situation is illustrated schematically in Fig. 3.8 and shows that the estimate of the error-bars can be misleadingly small if we try to avoid the marginalization procedure.

3.2.2 Asymmetric and multimodal posterior pdfs

The above analysis, leading to the approximation of the posterior pdf by a multivariate Gaussian, relies on the validity of the quadratic expansion of eqn (3.30). The elliptical

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

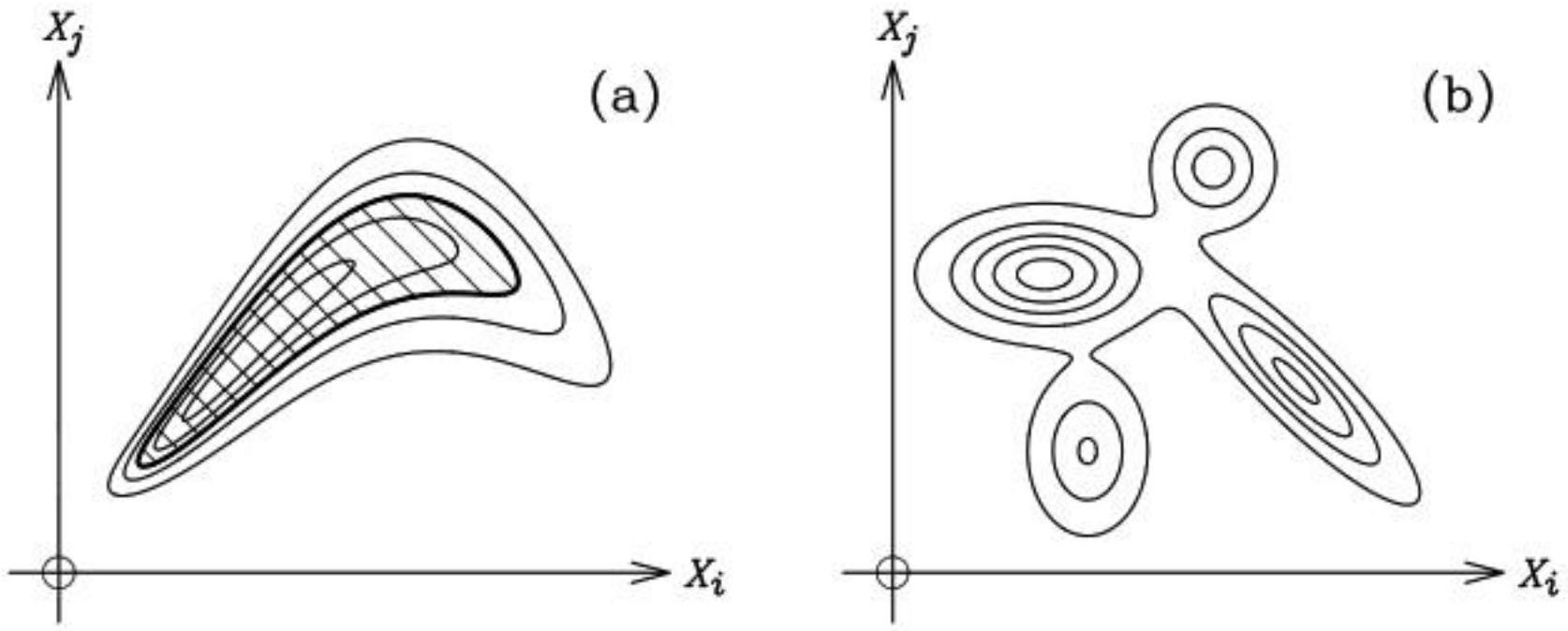


Fig. 3.9 A schematic illustration of two posterior pdfs for which the multivariate Gaussian is not a good approximation: (a) a 'lop-sided' pdf, analogous to the asymmetric case of Section 2.2; (b) a multimodal pdf.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

contours in Fig. 3.3 indicate that this is often a reasonable assumption, although we must be careful about their potentially truncated nature when there are few data. There are situations, however, when the posterior can have a very lop-sided shape; such a pdf is sketched in Fig. 3.9(a). The word 'asymmetric' may not be appropriate here, because there might still be a high degree of symmetry involved, but our difficulty is akin to the case discussed earlier in Section 2.2: namely, the concepts of the error-bar and linear correlation associated with the covariance matrix do not seem suitable; at the very least, the correspondence between the integral definition of covariance and the inverse of $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ will be doubtful. We might think of generalizing the idea of confidence intervals to a multidimensional space, but it will usually be hard to describe the surface of the (smallest) hypervolume containing 90% of the probability (say) in just a few numbers. The prescriptions of Section 2.2 can, of course, be used for the one-dimensional marginal distributions $\mathrm{prob}(X_j|\{data\}, I)$ for each of the parameters; unfortunately, the relevant multiple integrals over the joint pdf will often be difficult to carry out (analytically), and only tell part of the story. Without the convenient properties of the multivariate Gaussian, the maximum of the marginal pdf for X_j may not be the same as the jth component of the vector $\boldsymbol{X}_{\mathrm{O}}$ either; thus we will have to think carefully whether we are interested in the best estimate of all the parameters simultaneously, or just the optimal value of one irrespective of the others.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

The situation may be even worse, because the posterior might have many maxima; such a pdf is shown schematically in Fig. 3.9(b). For some types of problems, this multimodal nature can persist no matter how good the quality of the data. There is no difficulty in defining the best estimate, and expressing its reliability, when one of the probability bumps is much bigger than the others: we simply make a (local) quadratic expansion about the global maximum, and ignore the subsidiary solutions. If several of the maxima are of comparable magnitude, then we could carry out this procedure for all of them and provide a list of the alternative (almost) optimal values and their associated covariance matrices; if the number of significant solutions is very large, however, then we could soon be in trouble. Actually, we tend to have serious difficulties with

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

multimodal pdfs even when there is an obvious single best estimate: the problem is the practical one of how (to write a computer program) to find the global solution in a large multidimensional parameter space when there are many (small) local 'hillocks' to trip us up along the way. We will return to this topic in Section 3.4, when discussing some aspects about the computer implementation of our Bayesian analyses. First, though, let us briefly take another look at the Gaussian noise problem in the light of the multivariate character of this chapter.

3.3 Example 5: Gaussian noise revisited

In Section 2.3, we were concerned with estimating the mean μ of a Gaussian process. Since that chapter was restricted to a consideration of one-parameter examples, we assumed that the magnitude of the error-bar σ for the experimental measurements $\{x_k\}$ was already known. Now we are in a position to be able to relax that condition. What we require is the pdf $\mathrm{prob}(\mu|\{x_k\}, I)$, rather than the corresponding posterior conditional on being given the variance $\mathrm{prob}(\mu|\{x_k\}, \sigma, I)$. As a Gaussian process is defined by both μ and σ, we must integrate out the standard deviation as a nuisance parameter from their joint posterior pdf:

$$\mathrm{prob}(\mu|\{x_k\}, I) = \int_0^\infty \mathrm{prob}(\mu, \sigma|\{x_k\}, I)\, \mathrm{d}\sigma\,, \tag{3.33}$$

where the integrand can be expressed as a product of the likelihood function and the prior, for the two parameters, by using Bayes' theorem:

$$\mathrm{prob}(\mu, \sigma|\{x_k\}, I) \propto \mathrm{prob}(\{x_k\}|\mu, \sigma, I) \times \mathrm{prob}(\mu, \sigma|I)\,. \tag{3.34}$$

As in eqn (2.25), if the data are independent, the likelihood function is simply the product of the probabilities for obtaining the individual measurements; using the definition of a Gaussian distribution stated in eqn (2.23), we obtain

$$\mathrm{prob}(\{x_k\}|\mu, \sigma, I) = \left(\sigma\sqrt{2\pi}\right)^{-N} \exp\left[-\frac{1}{2\,\sigma^2}\sum_{k=1}^{N}(x_k-\mu)^2\right], \tag{3.35}$$

where there are N data.

What about the pdf $\mathrm{prob}(\mu, \sigma|I)$? To express complete prior ignorance about these parameters we should assign a pdf that is uniform with respect to μ and $\log\sigma$; the reason for this peculiar choice will become apparent in Chapter 5, but stems from the fact that the position μ is associated with an additive uncertainty whereas the width σ is a multiplicative scale-factor. For the moment, however, let us pursue our policy of *naïveté* and use a simple flat prior:

$$\mathrm{prob}(\mu, \sigma|I) = \begin{cases} \text{constant} & \text{for } \sigma > 0\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.36}$$

We will indicate how the results would be different if we had made the preferred assignment mentioned above, but our conclusions will be unaffected for all practical purposes.

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Multiplying the prior of eqn (3.36) by the likelihood function of eqn (3.35), as in eqn (3.34), allows us to write the marginal distribution of eqn (3.33) as

$$\text{prob}(\mu|\{x_k\},I) \propto \int_0^\infty t^{N-2} \exp\left[-\frac{t^2}{2}\sum_{k=1}^{N}(x_k-\mu)^2\right] \mathrm{d}t \,, \tag{3.37}$$

where we have made the substitution $\sigma = 1/t$ (so that $\mathrm{d}\sigma = -\mathrm{d}t/t^2$). A rescaling of t through the substitution $\tau = t\sqrt{\Sigma\,(x_k-\mu)^2}$ then reduces eqn (3.37) to

$$\text{prob}(\mu|\{x_k\},I) \propto \left[\sum_{k=1}^{N}(x_k-\mu)^2\right]^{-(N-1)/2} , \tag{3.38}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

because a definite integral involving only τ can be absorbed into the proportionality. The analysis of Section 2.2 shows that we can obtain our best estimate μ_o and a measure of its reliability from the first and second derivatives of the logarithm of this (marginal) posterior pdf. Accordingly,

$$\left.\frac{\mathrm{d}L}{\mathrm{d}\mu}\right|_{\mu_o} = \frac{(N-1)\sum(x_k-\mu_o)}{\sum(x_k-\mu_o)^2} = 0 \,,$$

where $L = \log_e[\text{prob}(\mu|\{x_k\},I)]$ and the summations are from $k=1$ to N; this can only be satisfied if the numerator is equal to zero, and yields the result of eqn (2.28):

$$\mu_o = \frac{1}{N}\sum_{k=1}^{N} x_k \,. \tag{3.39}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

In other words, the optimal estimate of μ is still given by the arithmetic average of the data. Differentiating L for a second time, and evaluating it at the maximum, we have

$$\left.\frac{\mathrm{d}^2L}{\mathrm{d}\mu^2}\right|_{\mu_o} = -\frac{N\,(N-1)}{\sum(x_k-\mu_o)^2} \,.$$

Since eqn (2.15) tells us that the error-bar for the best estimate is given by the inverse of the square root of (minus) the second derivative, we can summarize our inference of the mean by

$$\mu = \mu_o \pm \frac{S}{\sqrt{N}} \,, \quad \text{where } S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k-\mu)^2 \,. \tag{3.40}$$

Comparing this to eqn (2.29), we see that the only difference with the case of Section 2.3 is that the formerly (assumed) known value of σ has been replaced by an estimate derived from the data.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3.3.1 The Student-t and χ^2 distributions

In Section 2.3, we noted that the posterior pdf $\mathrm{prob}(\mu|\{x_k\},\sigma,I)$ was defined completely by the best estimate and its associated error-bar; this was because the quadratic termination of the Taylor series in eqn (2.11) was exact for that case. When σ is not known beforehand, this is no longer true: the summary of eqn (3.40) just represents a useful approximation to the (marginal) posterior $\mathrm{prob}(\mu|\{x_k\},I)$. The actual pdf is given in eqn (3.38); its shape is easier to ascertain if we rewrite the sum $\sum(x_k-\mu)^2$ in terms of two parameters, $\overline{x}$ and V, derived from the data:

$$\sum_{k=1}^{N}(x_k-\mu)^2 \;=\; N\,(\overline{x}-\mu)^2+V\,, \tag{3.41}$$

where $\overline{x}$ is equal to the sample-mean of eqn (3.39) and V is given by

$$V \;=\; \sum_{k=1}^{N}(x_k-\overline{x})^2\,. \tag{3.42}$$

Substituting for the sum from eqn (3.41) into eqn (3.38) yields

$$\mathrm{prob}(\mu|\{x_k\},I) \;\propto\; \Big[N\,(\overline{x}-\mu)^2+V\Big]^{-(N-1)/2}\,, \tag{3.43}$$

and is called a *Student-t* distribution. When $N=3$, this pdf has the same form as the Cauchy distribution of eqn (2.34) which was plotted in Fig. 2.8. It has a maximum at $\mu=\overline{x}$, a FWHM which is proportional to $\sqrt{V}$ and very long tails. As the number of data increases, this type of function is multiplied by itself many times over; as a result, the wide wings of the pdf are soon killed off to leave a more Gaussian-like distribution centred about $\overline{x}$. Hence, the optimal value μ_o is always equal to the sample-average; its error-bar, which is related to V through eqn (3.40), becomes more meaningful as N gets larger ($\geqslant 10$).

In our short discussion about the prior earlier, we promised to indicate how the results would differ if we had assigned a pdf which was uniform with respect to $\log\sigma$. We would have been led to the posterior pdf

$$\mathrm{prob}(\mu|\{x_k\},I) \;\propto\; \Big[N\,(\overline{x}-\mu)^2+V\Big]^{-N/2}\,, \tag{3.44}$$

which is identical to eqn (3.43) except that the power of the $N/2$-term has increased by one; that is, we would have obtained a Student-t distribution with $N-1$ *degrees-of-freedom*, instead of $N-2$. Their shapes are very similar, with a maximum at $\overline{x}$, but the pdf of eqn (3.44) is a little narrower than that of eqn (3.43). In terms of eqn (3.40), S^2 would be given by V/N rather than $V/(N-1)$; this difference is negligible if N is moderately large, which is also the régime when the concept of an error-bar is most appropriate. Thus the simple-minded assignment of eqn (3.36) gives us a slightly more conservative estimate of the reliability of the optimal value μ_o, when there are few data, but our conclusions remain essentially unchanged.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Before leaving this section, let us consider what we can learn from the data about the magnitude of the expected error in the measurements. Our inference about σ is described by the (marginal) posterior pdf $\mathrm{prob}(\sigma|\{data\}, I)$:

$$\mathrm{prob}(\sigma|\{x_k\}, I) = \int_{-\infty}^{+\infty} \mathrm{prob}(\mu, \sigma|\{x_k\}, I)\, \mathrm{d}\mu\,. \tag{3.45}$$

Using the likelihood function and prior of eqns (3.35) and (3.36), along with the substitution of eqn (3.41), Bayes' theorem allows to write this as

$$\mathrm{prob}(\sigma|\{x_k\}, I) \propto \sigma^{-N} \exp\left(-\frac{V}{2\,\sigma^2}\right) \int_{-\infty}^{+\infty} \exp\left[-\frac{N\,(\overline{x}-\mu)^2}{2\,\sigma^2}\right] \mathrm{d}\mu\,, \tag{3.46}$$

where we have taken all the terms not involving μ outside the integral, and $\sigma > 0$. Since the area under the one-dimensional normal distribution on the right-hand side is proportional to σ, we obtain the result

$$\mathrm{prob}(\sigma|\{x_k\}, I) \propto \sigma^{1-N} \exp\left(-\frac{V}{2\,\sigma^2}\right). \tag{3.47}$$

In the formal language of statistics, this is related to the χ^2 distribution (through the substitution $X = V/\sigma^2$). As usual, according to the analysis of Section 2.2, the best estimate and its error-bar can be derived from the first and second derivatives of the logarithm of eqn (3.47). This allows us to summarize our inference about σ by

$$\sigma = \sigma_{\mathrm{o}} \pm \frac{\sigma_{\mathrm{o}}}{\sqrt{2(N-1)}}\,, \tag{3.48}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

where the optimal value is $\sigma_{\mathrm{o}} = \sqrt{V/(N-1)}$. Although eqn (3.48) permits negative values of σ, this simply indicates that the quadratic expansion of eqn (2.11) gives a poor approximation to the asymmetric posterior pdf when the number of data is small; it would then be better to give an estimate and error-bar for $\log\sigma$ instead, which we will learn how to do in Section 3.6.

3.4 Algorithms: a numerical interlude

We have now seen several examples of how the rules of probability theory can be used to obtain the posterior pdf for the quantities of interest; in order to summarize the inference, we usually need to find its maximum. Sometimes this optimization can be done analytically, but often we are forced to turn to the computer. The virtue of this electronic machine lies, of course, in its ability to do numerical calculations (and graphical plotting) considerably more quickly, and painlessly, than we could do ourselves by hand. Since computer programming and optimization are large subjects in their own right, our aim here is merely to give a flavour of the range of problems that are likely to

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

be encountered and to give an indication of the sort of procedures that might be used to tackle them. A more advanced account of a novel *Monte Carlo* method for doing Bayesian computation is given in Chapters 9 and 10.

We should point out that the material in this section is to do with *algorithms*, rather than the fundamentals of data analysis. That is to say, the techniques employed to find (as opposed to define) the optimal solution are based purely on practical considerations; the fact that we are using them to solve probabilistic problems is essentially coincidental. Details of the procedures mentioned below can be found in numerous good books; of particular note is *Numerical Recipes*, by Press *et al.* (1986), which includes source-code for computer programs (in FORTRAN, C and Pascal). *Practical Optimization*, by Gill *et al.* (1981), is also excellent, as is the old gem by Acton (1970) entitled *Numerical Methods That (usually) Work*.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3.4.1 Brute force and ignorance

For a one-parameter problem, the most robust way of finding the maximum of the posterior pdf is to plot it out and have a look! On a computer, this brute force and ignorance method entails the division of the horizontal axis into a finite number of grid points, representing the possible values of the parameter, and the evaluation of the posterior probability at each of them. Plotting the latter along the vertical axis, we simply find the biggest value and read off the corresponding optimal estimate of the quantity of interest. As mentioned in Section 2.4, it is numerically better to calculate the logarithm of the posterior probability for all the grid points first; and then subtract the largest number from each and exponentiate. If normalization is required, this pdf can be multiplied by a constant so that the area under the curve is equal to unity.

The greatest advantage of this elementary algorithm is that it gives a complete picture of our inference about the value of the parameter. As long as there are enough grid points (a few hundred usually being quite adequate), which cover a sufficiently large range, this procedure will always work. It doesn't matter whether the posterior pdf is asymmetric, multimodal or whatever; even the question of differentiability is of no concern. This brute force method generalizes readily to two-parameter problems: we just plot the posterior probability vertically against a two-dimensional grid of points, often with contours. This is what was done in Fig. 3.3. As well as all the advantages mentioned above, it is very easy to obtain the marginal distributions: simply add up the probabilities in the 'X' or 'Y' directions, as appropriate. Even sharp cut-offs imposed by the prior pdf cause no difficulty for this numerical integration; this can be seen from the low-count example in Figs 3.3 and 3.4.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Unfortunately, this direct approach, of evaluating the probability everywhere, rapidly becomes impractical after the two-parameter case. Apart from the problem of how to display a function of several variables on a piece of paper, the computational cost of calculating the probabilities is soon prohibitive. If each axis was divided into just 10 discrete points, an M-parameter problem would then require 10^M evaluations. While the calculations for two variables might only take a second, a whole day will be required for seven and the age of the universe for twenty! Super-computers, and ridiculously coarse grids, can buy a little more leeway, but not much. For multivariate analysis, we must

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

look for more efficient algorithms.

3.4.2 The joys of linearity

In Section 3.2, we represented a set of M parameters $\{X_j\}$ by the components of a (column) vector $\boldsymbol{X}$; the condition for our best estimate of their values, denoted by $\boldsymbol{X}_{\mathrm{O}}$, can be written as

$$\boldsymbol{\nabla} L(\boldsymbol{X}_{\mathrm{O}}) = 0\,, \tag{3.49}$$

where the jth element of $\boldsymbol{\nabla} L$ is given by the partial derivative of the logarithm of the posterior pdf $\partial L/\partial X_j$ (evaluated at $\boldsymbol{X} = \boldsymbol{X}_{\mathrm{O}}$). Equation (3.49) is very compact notation for a set of M simultaneous equations; they will be quite difficult to solve, in general, unless they are *linear*. That is to say, if we are fortunate enough to be able to rearrange $\boldsymbol{\nabla} L$ into a form resembling a straight line ('$y = mx + c$'),

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\boldsymbol{\nabla} L = \mathsf{H}\boldsymbol{X} + \boldsymbol{C}\,, \tag{3.50}$$

where the components of the vector $\boldsymbol{C}$, and the square matrix H, are all constant, then the solution to eqn (3.49) is trivial:

$$\boldsymbol{X}_{\mathrm{O}} = -\mathsf{H}^{-1}\boldsymbol{C}\,. \tag{3.51}$$

Differentiating eqn (3.50), we find that the important $\boldsymbol{\nabla}\boldsymbol{\nabla} L$ matrix is invariant with respect to the parameters $\{X_j\}$:

$$\boldsymbol{\nabla}\boldsymbol{\nabla} L = \mathsf{H}\,, \tag{3.52}$$

and so all the higher derivatives are identically zero. The covariance matrix $\boldsymbol{\sigma}^2$, given by minus the inverse of $\boldsymbol{\nabla}\boldsymbol{\nabla} L$, therefore provides a complete description of the shape of the posterior pdf:

$$\left[\boldsymbol{\sigma}^2\right]_{ij} = \left\langle (X_i - X_{\mathrm{O}i})(X_j - X_{\mathrm{O}j}) \right\rangle = -\left[\mathsf{H}^{-1}\right]_{ij}\,,$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

where the $\{X_{\mathrm{O}j}\}$ are the components of the vector $\boldsymbol{X}_{\mathrm{O}}$ in eqn (3.51).

Although it is easy to write down the optimal solution in the form of eqn (3.51), actually evaluating it can still entail a fair amount of computational effort. Luckily, however, there are plenty of good *procedures*, or *subroutines*, available for matrix algebra; algorithmic details can be found in the references mentioned earlier. If we solved the set of simultaneous equations $\mathsf{H}\boldsymbol{X}_{\mathrm{O}} = -\boldsymbol{C}$ directly, with a *Cholesky decomposition* for example, it would avoid the need to invert the matrix H; this should also be somewhat faster numerically for obtaining the best estimate $\boldsymbol{X}_{\mathrm{O}}$. Whichever route we decide to follow, it will be most efficient if the algorithm used exploits the known symmetry of $\boldsymbol{\nabla}\boldsymbol{\nabla} L$ $(\partial^2 L/\partial X_i\,\partial X_j = \partial^2 L/\partial X_j\,\partial X_i)$. The computational time taken for such matrix manipulations tends to scale like the cube of the number of variables M^3; the speed of personal computers is sufficient to tackle most reasonable parameter estimation problems in this way.

We will, of course, have difficulty in solving the equations for the optimal estimate if the determinant of H is zero, or extremely small; in that case, our answers will be

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

highly sensitive to small changes in the data and the corresponding (marginal) error-bars will be huge. In terms of Fig. 3.6, this will happen if the ellipsoid of $Q = k$, where $Q = (\boldsymbol{X}-\boldsymbol{X}_{\mathrm{O}})^{\mathrm{T}} \boldsymbol{\nabla}\boldsymbol{\nabla}L(\boldsymbol{X}_{\mathrm{O}})\,(\boldsymbol{X}-\boldsymbol{X}_{\mathrm{O}})$, is (almost) infinitely long in any of its principal directions. In this situation, it is useful to analyse the eigenvalues and eigenvectors of $\mathbf{H}$: the principal axes indicate which linear combinations of the parameters can be inferred independently of each other and the eigenvalues give the reliability with which they can be estimated. The only real cure for the predicament is to improve the characteristics of the posterior pdf; this can be done by obtaining more relevant data, or by supplementing them with cogent prior information.

3.4.3 Iterative linearization

The linear problem is so convenient, both analytically and computationally, that it is worth trying to make use of it even when $\boldsymbol{\nabla}L$ cannot quite be written in the form of eqn (3.50). To see how we might do this, consider the Taylor series expansion of L about some arbitrary point in parameter space $\boldsymbol{X}_1$:

$$L = L(\boldsymbol{X}_1) + (\boldsymbol{X}-\boldsymbol{X}_1)^{\mathrm{T}} \boldsymbol{\nabla}L(\boldsymbol{X}_1) + \frac{1}{2}(\boldsymbol{X}-\boldsymbol{X}_1)^{\mathrm{T}} \boldsymbol{\nabla}\boldsymbol{\nabla}L(\boldsymbol{X}_1)\,(\boldsymbol{X}-\boldsymbol{X}_1) + \cdots .$$

In the past, the first-derivative term has been missing because we've always expanded about the optimal estimate $\boldsymbol{X}_{\mathrm{O}}$. Differentiating this with respect to the $\{X_j\}$, we obtain

$$\boldsymbol{\nabla}L = \boldsymbol{\nabla}L(\boldsymbol{X}_1) + \boldsymbol{\nabla}\boldsymbol{\nabla}L(\boldsymbol{X}_1)\,(\boldsymbol{X}-\boldsymbol{X}_1) + \cdots , \tag{3.53}$$

which is just the Taylor series for $\boldsymbol{\nabla}L$. If we ignore the higher-order terms on the right-hand side, so that eqn (3.53) can be rearranged into the linear form of eqn (3.50), then the solution to eqn (3.49) is given by

$$\boldsymbol{X}_{\mathrm{O}} \approx \boldsymbol{X}_1 - \Big[\boldsymbol{\nabla}\boldsymbol{\nabla}L(\boldsymbol{X}_1)\Big]^{-1} \boldsymbol{\nabla}L(\boldsymbol{X}_1) . \tag{3.54}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

This relationship will be exact when $\boldsymbol{X}_1 = \boldsymbol{X}_{\mathrm{O}}$, or if $\boldsymbol{\nabla}L$ is truly linear, but it will be a reasonable approximation as long as $\boldsymbol{X}_1$ is close enough to the optimal estimate. As such, it lends itself to suggest an *iterative* algorithm: (i) start with a good guess of the optimal solution $\boldsymbol{X}_1$; (ii) evaluate the gradient-vector $\boldsymbol{\nabla}L$, and the second-derivative matrix $\boldsymbol{\nabla}\boldsymbol{\nabla}L$, at $\boldsymbol{X} = \boldsymbol{X}_1$; (iii) calculate an improved estimate $\boldsymbol{X}_2$, by equating it to the right-hand side of eqn (3.54); (iv) repeat this process until $\boldsymbol{\nabla}L = 0$.

The procedure outlined above is called the *Newton–Raphson* algorithm. It is a generalization of the numerical technique often used to find the *roots* of a function $f(x_{\mathrm{o}}) = 0$; in this case, the function is multivariate: $\boldsymbol{\nabla}L(\boldsymbol{X}_{\mathrm{O}}) = 0$. We can summarize it with the recursion relationship

$$\boldsymbol{X}_{N+1} = \boldsymbol{X}_N - \Big[\boldsymbol{\nabla}\boldsymbol{\nabla}L(\boldsymbol{X}_N)\Big]^{-1} \boldsymbol{\nabla}L(\boldsymbol{X}_N) , \tag{3.55}$$

where $\boldsymbol{X}_N$ is our estimate of the solution after $N-1$ iterations. If $\boldsymbol{\nabla}L$ is really linear, then only one iteration will be required: $\boldsymbol{X}_2$ will equal $\boldsymbol{X}_{\mathrm{O}}$, irrespective of the initial

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

guess $\boldsymbol{X}_1$. In general, the algorithm will rapidly converge to $\boldsymbol{X}_\mathrm{O}$ as long as the starting point is 'close enough' to the optimal solution; we will say more about this shortly, but first let us briefly mention an important practical point.

The stability of iterative procedures can usually be improved by slowing them down a little. This means that it can be advantageous to make a slightly smaller change in going from $\boldsymbol{X}_N$ to $\boldsymbol{X}_{N+1}$ than that recommended by eqn (3.55). Although this can easily be done by multiplying the matrix-vector product on the right of eqn (3.55) by a fractional constant, it tends to be better to achieve a similar effect by adding a small (negative) number c to all the diagonal elements of $\boldsymbol{\nabla}\boldsymbol{\nabla}L$:

$$\boldsymbol{X}_{N+1} = \boldsymbol{X}_N - \left[\boldsymbol{\nabla}\boldsymbol{\nabla}L(\boldsymbol{X}_N) + c\,\mathbf{I}\right]^{-1} \boldsymbol{\nabla}L(\boldsymbol{X}_N)\,, \tag{3.56}$$

where $\mathbf{I}$ is the *identity matrix*, with ones along the diagonal and zeros everywhere else. The reason for this curious suggestion is best understood from the quadratic form in Fig. 3.6, where the nature of a matrix is characterized by its eigenvalues $\{\lambda_j\}$ and eigenvectors $\{\boldsymbol{e}_j\}$; explicitly, for $\boldsymbol{\nabla}\boldsymbol{\nabla}L$, these are the solutions of the equation

$$\left[\boldsymbol{\nabla}\boldsymbol{\nabla}L\right]\boldsymbol{e}_j = \lambda_j\,\boldsymbol{e}_j\,, \tag{3.57}$$

where $j = 1, 2, \ldots, M$. If we add a multiple of the identity matrix to $\boldsymbol{\nabla}\boldsymbol{\nabla}L$, then we find that this new matrix has the same eigenvectors but different eigenvalues:

$$\left[\boldsymbol{\nabla}\boldsymbol{\nabla}L + c\,\mathbf{I}\right]\boldsymbol{e}_j = \left[\lambda_j + c\right]\boldsymbol{e}_j\,, \tag{3.58}$$

Thus the 'beefing-up of the diagonal' has no effect on the orientation of the ellipsoid and leaves the correlations between the parameters in place; with a suitable choice of c, however, it does cause a significant narrowing in those principal directions in which it was very elongated. Since these small eigenvalues are associated with large uncertainties, eqn (3.56) stabilizes the iterative algorithm by selectively reducing their influence. To put it another way, the inverse of a matrix is related to the reciprocal of its determinant; as this is given by the product of the eigenvalues, the small ones will cause $(\boldsymbol{\nabla}\boldsymbol{\nabla}L)^{-1}$ to blow up. By adding a small (negative) multiple of the identity matrix to $\boldsymbol{\nabla}\boldsymbol{\nabla}L$, we can ensure that the magnitude of the determinant is safely greater than zero; this results in a reduced (finite) value for the inverse of the hybrid second-derivative matrix and gives a correspondingly smaller change in our estimate of the solution between iterations.

If the number of parameters becomes too large, so that the $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ matrix is difficult to store or invert, then the Newton–Raphson procedure can be implemented through a *conjugate-gradient* algorithm; details may be found in *Numerical Recipes* (Press *et al.* 1986), for example. The more serious problem is that the solution to eqn (3.49) will not necessarily represent the position of the maximum, even when the posterior pdf is unimodal. The requirement $\boldsymbol{\nabla}L(\boldsymbol{X}_\mathrm{O}) = 0$ is really a condition for finding a stationary point and can sometimes be satisfied in the extreme tails of the posterior probability. The Newton–Raphson procedure will diverge (to infinity), therefore, if the initial guess $\boldsymbol{X}_1$ is not close enough to the optimal solution. The situation is illustrated schematically in Fig. 3.10.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

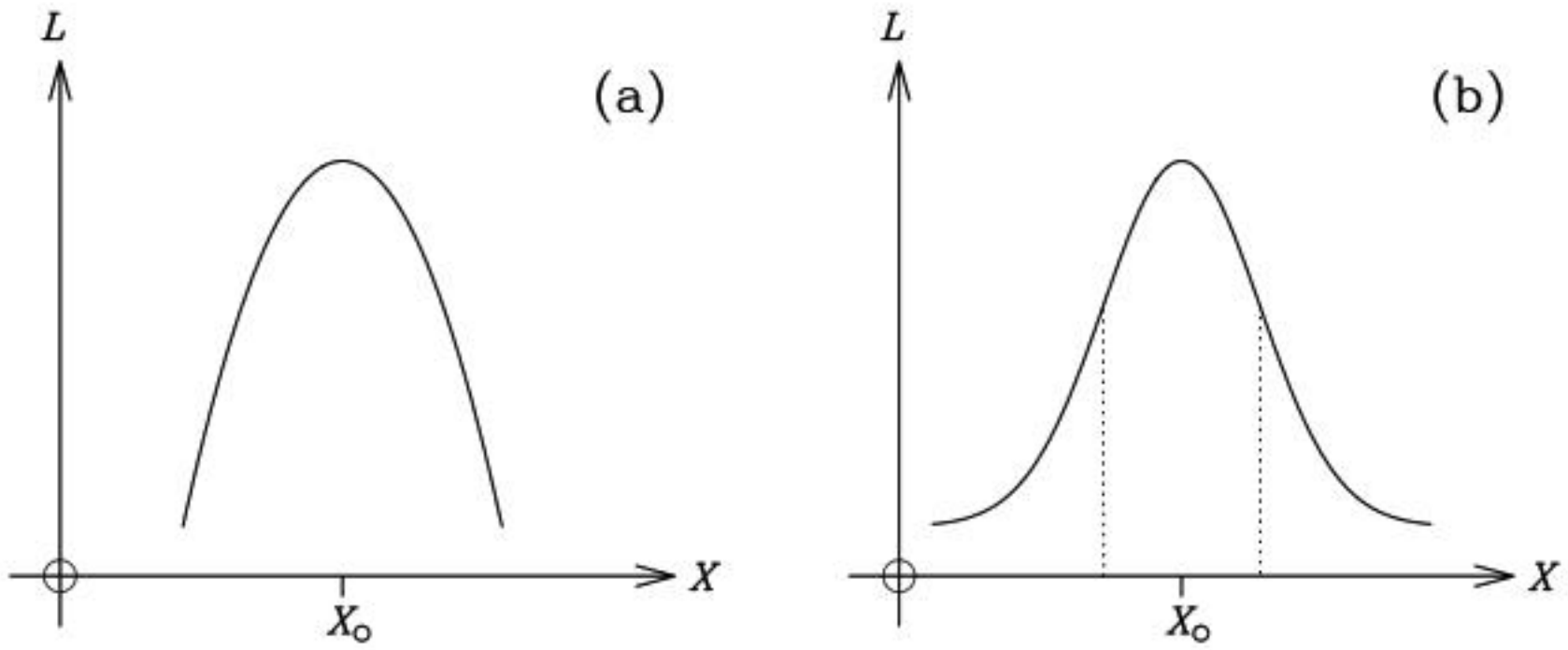


Fig. 3.10 An illustration of the logarithm of two posterior pdfs: (a) a well-behaved (linear-like) problem for which the Newton–Raphson procedure will converge to $\boldsymbol{X}_O$ from any starting point; (b) a unimodal pdf where the Newton–Raphson algorithm will diverge unless the initial guess ($\boldsymbol{X}_1$) lies within the two dotted lines.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

For the case of Fig. 3.10(b), one way to ensure that our computer program will converge towards $\boldsymbol{X}_O$ is to rely predominantly on the direction of the gradient $\boldsymbol{\nabla} L$. A good alternative to such first-derivative methods is provided by the approach of Nelder and Mead (1965): their 'up-hill' *simplex* search algorithm works with the function L directly and is very robust for unimodal pdfs; it is practical to use for up to a couple of dozen parameters. Being gradient-free, it also allows (any) sharp cut-offs imposed by the prior to be encoded in a straightforward manner. Nevertheless, since it lacks the efficiency of Newton–Raphson, it's best used just as a first step to get close to the optimal solution. Besides, we need to evaluate the second-derivative matrix to estimate the error-bars for the parameters; if necessary, the required gradients can even be calculated by *finite-differences* ($\partial L/\partial X_j \approx \left[L(\boldsymbol{X}_{\neq j}, X_j + \delta/2) - L(\boldsymbol{X}_{\neq j}, X_j - \delta/2)\right]/\delta$, etc.).

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3.4.4 Hard problems

The most difficult optimization task occurs for the case when the posterior pdf is multimodal. For up to a couple of parameters, or so, the best thing to do is to use brute force and ignorance: simply 'display' the entire pdf, by evaluating it on a discrete grid, and look for the maximum(s). Unfortunately, as mentioned earlier, this becomes hopelessly impractical for more than a handful of variables. While gradient-based algorithms are the most efficient for multivariate analysis, they cannot help for the multimodal case: a knowledge of the local slope sheds no light on the position of the global maximum. To date, all procedures designed to address this problem entail a large element of inspired trial-and-error. *Simulated annealing* is perhaps the most well-established of these techniques and was proposed by Kirkpatrick *et al.* (1983) on the basis of a *thermodynamic* analogy. The basic scheme involves an exploration of the parameter space by a series of trial random changes in the current estimate of the solution; in terms of eqn (3.55), this could be written as $\boldsymbol{X}_{N+1} = \boldsymbol{X}_N + \Delta\boldsymbol{X}$, where $\Delta\boldsymbol{X}$ is chosen by a random number generator. The proposed update of the solution is always considered advantageous if it yields a higher value of L, but bad moves are also sometimes accepted. This occasional

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

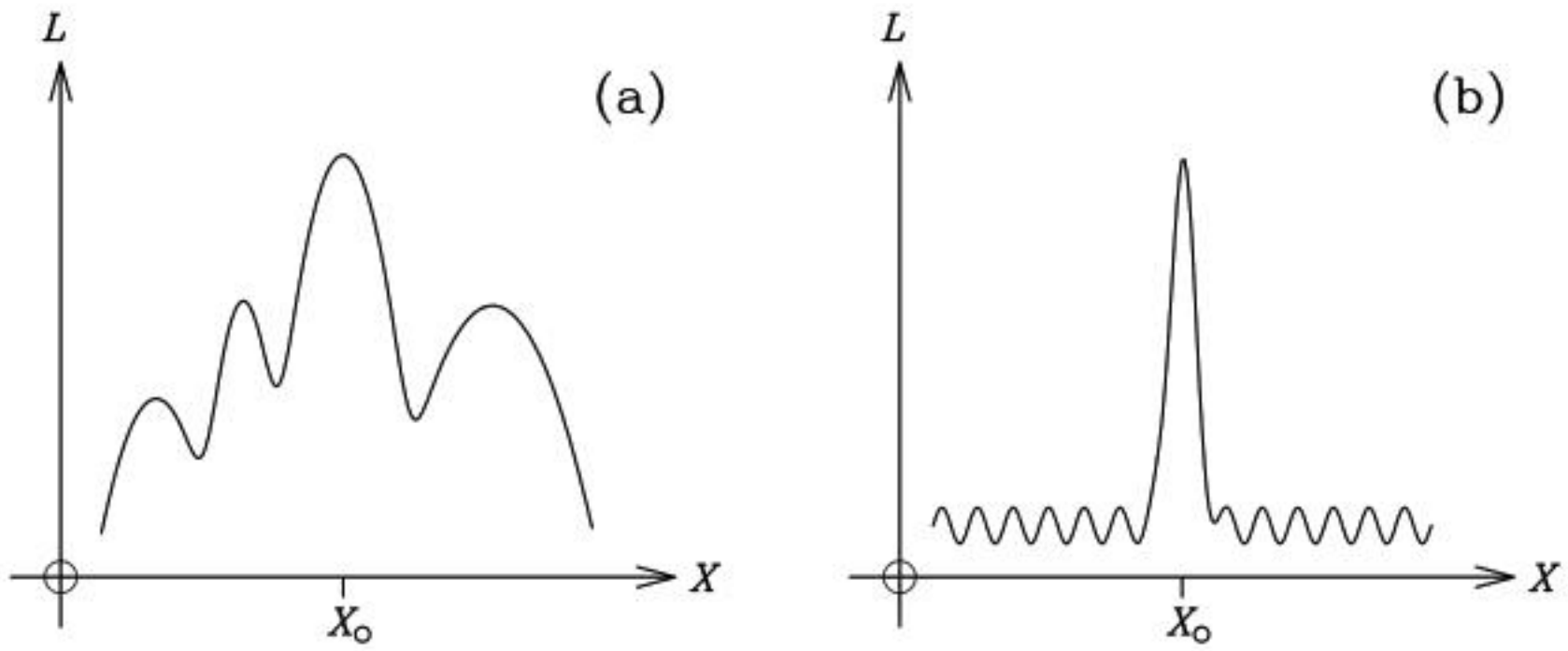


Fig. 3.11 A schematic illustration of the logarithm of two multimodal posterior pdfs. Simulated annealing should work well for case (a), but will not be very helpful for (b).

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

allowance of retrograde steps provides a mechanism for escaping entrapment in local maxima. The shape of L for two multimodal posterior pdfs is shown in Fig. 3.11. Simulated annealing should work like a treat for case (a), where there is an underlying trend towards a global maximum, but is unlikely to be much more efficient than brute force and ignorance for (b).

The real difficulty with the multimodal case is that it is almost impossible to guarantee that the optimal solution has been found, in a finite amount of time, in general. As such, the field remains wide open. Ingenious new schemes, such as *genetic algorithms* (Michalewicz 1992), are always being put forward, but none can claim universal success or applicability. We must also remember that it doesn't really make sense to talk about a single 'best' answer, even in principle, if the posterior pdf has several maxima of comparable magnitude. In these cases, Monte Carlo methods that can simulate exploration of all the important regions of parameter space will have considerable merit; they are the focus of a more detailed discussion in Chapters 9 and 10.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3.5 Approximations: maximum likelihood and least-squares

There has been one common theme to all our discussions, and examples, so far: namely, the use of Bayes' theorem. Its great virtue lies in the fact that it helps us to relate the pdf required to describe our inference, to others which we have a better chance of being able to assign. Using the components of the vectors $\boldsymbol{X}$ and $\boldsymbol{D}$ to denote the set of M parameters of interest and the N measured data, respectively, we can write Bayes' theorem compactly as

$$\text{prob}(\boldsymbol{X}|\boldsymbol{D}, I) \propto \text{prob}(\boldsymbol{D}|\boldsymbol{X}, I) \times \text{prob}(\boldsymbol{X}|I) \,, \tag{3.59}$$

where I represents all the relevant background information. The prior pdf should reflect everything we know about $\boldsymbol{X}$ before the analysis of the current data; if we are largely ignorant, then we might indicate this naïvely with a uniform, or flat, pdf (as a simple limiting case of a very broad one):

$$\text{prob}(\boldsymbol{X}|I) = \text{constant} \,, \tag{3.60}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

for, effectively, all values of $\boldsymbol{X}$. Since this uniform assignment can be absorbed into the (omitted) normalization constant of eqn (3.59), the posterior pdf becomes directly proportional to the likelihood function:

$$\text{prob}(\boldsymbol{X}|\boldsymbol{D}, I) \propto \text{prob}(\boldsymbol{D}|\boldsymbol{X}, I) \ . \tag{3.61}$$

Thus our best estimate $\boldsymbol{X}_{\text{O}}$, given by the maximum of the posterior, is equivalent to the solution which yields the greatest value for the probability of the observed data; in this context, it is usually referred to as the *maximum likelihood* estimate.

We can obtain further simplification by making suitable approximations with regard to the likelihood function itself. If we assume that the data are independent, for example, then their joint pdf $\text{prob}(\boldsymbol{D}|\boldsymbol{X}, I)$ is given by the product of the probabilities for the individual measurements:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(\boldsymbol{D}|\boldsymbol{X}, I) = \prod_{k=1}^{N} \text{prob}(D_k|\boldsymbol{X}, I) \ , \tag{3.62}$$

where D_k is the kth datum. Although we have used this result several times before, it is worth emphasizing that it follows immediately from (the repeated application of) the product rule of eqn (1.2),

$$\text{prob}(D_k, D_l|\boldsymbol{X}, I) = \text{prob}(D_k|D_l, \boldsymbol{X}, I) \times \text{prob}(D_l|\boldsymbol{X}, I) \ ,$$

and the assertion that our knowledge of one datum has no influence on our ability to predict the outcome of another, if we are already given $\boldsymbol{X}$:

$$\text{prob}(D_k|D_l, \boldsymbol{X}, I) = \text{prob}(D_k|\boldsymbol{X}, I) \ .$$

If we also assume that the noise associated with the experimental measurements can reasonably be represented as a Gaussian process, then the probability of an individual datum can be written as:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(D_k|\boldsymbol{X}, I) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left[-\frac{(F_k - D_k)^2}{2\,\sigma_k^2}\right] \ , \tag{3.63}$$

where I implicitly includes a knowledge of both the expected size of the error-bars $\{\sigma_k\}$, and an adequate model of the functional relationship f between the parameters $\boldsymbol{X}$ and the ideal (noiseless) data $\boldsymbol{F}$:

$$F_k = f(\boldsymbol{X}, k) \ . \tag{3.64}$$

Equations (3.62) and (3.63) then allow us to approximate the likelihood function by

$$\text{prob}(\boldsymbol{D}|\boldsymbol{X}, I) \propto \exp\left(-\frac{\chi^2}{2}\right) \ , \tag{3.65}$$

where χ^2 is the sum of the squares of the *normalized residuals* ($R_k = (F_k - D_k)/\sigma_k$):

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\chi^2 = \sum_{k=1}^{N} \left(\frac{F_k - D_k}{\sigma_k} \right)^2 . \tag{3.66}$$

With the uniform prior of eqn (3.60), according to eqn (3.61), the logarithm L of the posterior pdf is simply given by

$$L = \log_e \big[\text{prob}(\boldsymbol{X} | \boldsymbol{D}, I) \big] = \text{constant} - \frac{\chi^2}{2} . \tag{3.67}$$

Since the maximum of the posterior will occur when χ^2 is smallest, the corresponding optimal solution $\boldsymbol{X}_{\text{O}}$ is usually called the *least-squares* estimate.

Maximum likelihood and least-squares are amongst the most frequently used procedures in data analysis; as we have just seen, they can easily be justified through the Bayesian approach. In this context, however, neither of these methods is seen as fundamental or sacrosanct: they are merely what Bayes' theorem reduces to when certain simplifying approximations are deemed to be suitable. If the results from a least-squares analysis seem unsatisfactory, for example, we don't need to throw-up our hands in horror and say 'Least-squares doesn't work!'. Since we understand the assumptions which underlie its use, it's much more fruitful to ponder whether these were appropriate for the problem being considered; if not, we should start again from Bayes' theorem and derive a better statistical prescription for that case. Some examples of the latter are given in Chapter 8.

Perhaps one of the main reasons for the popularity of the least-squares method is that it is simple to apply. This is particularly true if the functional relationship of eqn (3.64) is linear, because then $\boldsymbol{\nabla} L$ can be written in the convenient form of eqn (3.50); as mentioned in Section 3.4, this makes the optimization problem very easy to handle. To see that this is so, let us write the equation for the kth ideal datum as

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$F_k = \sum_{j=1}^{M} T_{kj} X_j + C_k , \tag{3.68}$$

where the values of both T_{kj} and C_k are independent of the parameters $\{X_j\}$; in matrix-vector notation, this would just be $\boldsymbol{F} = \mathsf{T}\boldsymbol{X} + \boldsymbol{C}$. Using the *chain-rule* of differentiation, the jth component of $\boldsymbol{\nabla} L$ is given by

$$\frac{\partial L}{\partial X_j} = -\frac{1}{2} \frac{\partial \chi^2}{\partial X_j} = -\sum_{k=1}^{N} \frac{(F_k - D_k)}{\sigma_k^2} \frac{\partial F_k}{\partial X_j} . \tag{3.69}$$

Although we could substitute for F_k, and $\partial F_k / \partial X_j = T_{kj}$, from eqn (3.68) in eqn (3.69), and rearrange it into the form of eqn (3.50), it's slightly messy to do this algebraically. An easier way to verify the linearity of $\boldsymbol{\nabla} L$ is to differentiate eqn (3.69) with respect to X_i and notice that the elements of the second-derivative matrix are all constant:

$$\frac{\partial^2 L}{\partial X_i \partial X_j} = -\sum_{k=1}^{N} \frac{T_{ki} T_{kj}}{\sigma_k^2} . \tag{3.70}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Since all the higher derivatives of L are identically zero, the posterior pdf is defined completely by the optimal solution $\boldsymbol{X}_{\mathrm{O}}$ and its covariance matrix; the components of the latter are related to twice the inverse of the $\boldsymbol{\nabla}\boldsymbol{\nabla}\chi^2$, or the *Hessian*, matrix:

$$\left\langle (X_i - X_{\mathrm{O}i})(X_j - X_{\mathrm{O}j}) \right\rangle = -\left[(\boldsymbol{\nabla}\boldsymbol{\nabla}L)^{-1} \right]_{ij} = 2\left[(\boldsymbol{\nabla}\boldsymbol{\nabla}\chi^2)^{-1} \right]_{ij} . \qquad (3.71)$$

The useful property that the $\boldsymbol{\nabla}\boldsymbol{\nabla}L$ matrix will be constant if the parameters $\boldsymbol{X}$ are related linearly to the data is not generally true without eqn (3.67); that is why the least-squares approximation is very convenient. For example, the functional model of eqn (3.1), where we are given the shape and position of the signal peak, is linear with respect to the amplitude A and the background B. Nevertheless, the optimal solution $(A_{\mathrm{O}}, B_{\mathrm{O}})$ is difficult to write down analytically because the gradient-vector of the posterior pdf in eqn (3.8) cannot be rearranged into the form of eqn (3.50); this was not a serious impediment, of course, because we just did the problem numerically (with brute force and ignorance). Even for that case of the Poisson likelihood, however, we could have obtained a fairly good estimate by using the least-squares approximation. This is because eqn (3.2) starts to take on Gaussian characteristics if the expected number of counts is more than a handful ($\geqslant 10$); this behaviour can be seen in Fig. 3.2(b), despite the fact that it's a function of a discrete variable. In the limit of large numbers, it can formally be shown that eqn (3.2) is well-represented by a normal distribution:

$$\mathrm{prob}(N|D) = \frac{D^N \mathrm{e}^{-D}}{N!} \propto \exp\left[-\frac{(N-D)^2}{2D} \right] ,$$

In our usual shorthand notation, we could summarize this by saying $N \approx D \pm \sqrt{D}$. As the number of measured counts will be roughly equal to the expected value D, we can make the denominator of the exponent independent of the parameters A and B by replacing the $\sqrt{D}$ error-bar with $\sqrt{N}$; this helps to linearize the optimization problem. With a uniform prior, the logarithm of the posterior pdf is then approximated well by eqn (3.67); in terms of our present notation, the corresponding χ^2 statistic is given by

$$\chi^2 = \sum_{k=1}^{N} \frac{(F_k - D_k)^2}{D_k} ,$$

where D_k is now the number of counts measured in the kth data-channel and F_k is our estimate of their expected number based on the linear relationship of eqn (3.1). This result, where the error-bar in eqn (3.66) is replaced by the square root of the datum ($\sigma_k^2 = D_k$, or more often F_k), tallies with the definition of χ^2 found in many elementary textbooks; it suggests that the assumption of a Poisson process is implicit in its use. This would not be appropriate, for example, if the uncertainties related to an experiment were dominated by thermal (Johnson) noise in a piece of electronic equipment; in that case, $\sigma_k =$ a constant (proportional to the temperature) would be more suitable.

Despite the practical benefits of least-squares, we should emphasize that the real justification for its use hinges on the assignments of a Gaussian likelihood function and

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

uniform prior. Otherwise, what reason do we have for using the sum of the squares of the residuals as a misfit statistic as opposed to anything else? Indeed, there is something to be said for the robustness of the l_1-*norm*:

$$l_1\text{-norm} = \sum_{k=1}^{N} \left| \frac{F_k - D_k}{\sigma_k} \right| . \tag{3.72}$$

This sum of the *moduli* of the (normalized) residuals has the advantage that it is far less susceptible to the effects of 'freak data' than is χ^2 (which is called the l_2-*norm*, in this terminology). Although the least-squares prescription was obtained by adopting the traditional assumptions of independent, additive, Gaussian noise in this section, we will see an alternative derivation in Chapter 5; there we will use the maximum entropy principle to assign the relevant uniform, Gaussian and exponential pdfs needed to justify the maximum likelihood, least-squares and l_1-norm estimates.

Before giving a simple example of the use of least-squares, we should warn against the unfortunate connotations conjured up by the words 'maximum likelihood estimate'. In terms of everyday parlance, it suggests that we have obtained the most probable values for the parameters of interest; without qualification, this is not so! By maximizing the likelihood function, we have found the values which make the measured data most probable; although we expect this to be relevant to our real question, 'What are the most probable values of the parameters, given the data?', it is not the same thing. To put it more plainly: $\text{prob}(A|B) \neq \text{prob}(B|A)$, in general. For example, the probability that it will rain given that there are clouds overhead is quite different from the probability of there being clouds overhead given that it's raining. The required relationship between these distinct conceptual entities is, of course, provided by Bayes' theorem. At this juncture, the reader may be forgiven for thinking that this is purely a philosophical point; after all, we have already noted, in eqn (3.61), that the posterior pdf becomes directly proportional to the likelihood function if we assign a uniform prior (which we often do). This belabouring is important, however, because a recognition of the qualification can have practical consequences. Even with a flat prior, a due consideration of the range over which it is valid can lead to a better estimate of the parameters involved and enable us to tackle problems which would otherwise be inaccessible; we will see examples of this in both Section 3.6 and in Chapter 4.

3.5.1 Fitting a straight line

One of the most frequently encountered problems in data analysis is the fitting of a straight line to graphical data. Suppose we are given a set of N data $\{Y_k\}$, with associated error-bars $\{\sigma_k\}$, measured at 'positions' $\{x_k\}$. What is the best estimate of the two parameters of a straight line which describes them?

Let us do this problem as a simple exercise in the use of least-squares. The situation is illustrated schematically in Fig. 3.12. For the straight-line model, the kth ideal datum y_k is given by

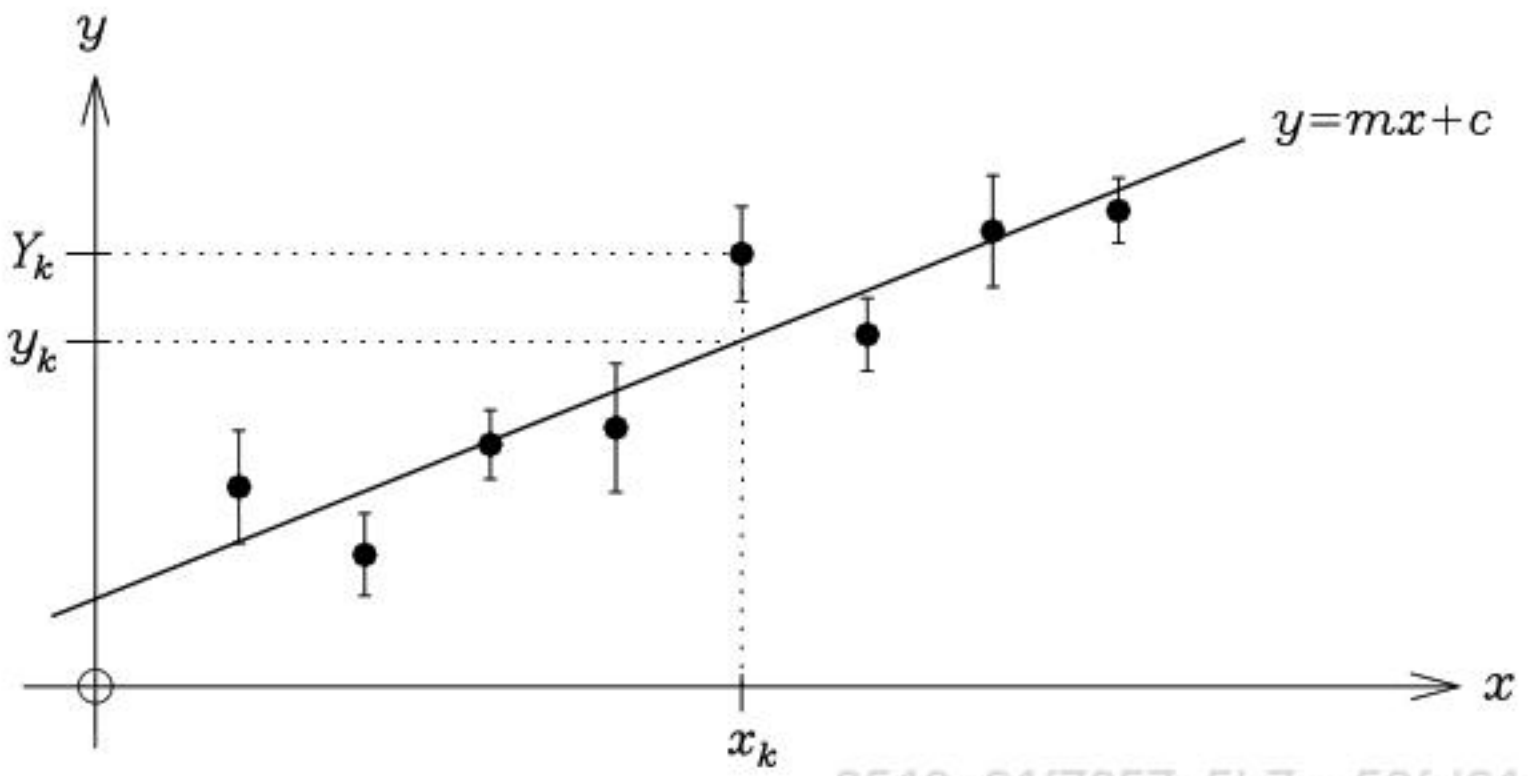
$$y_k = m x_k + c ,$$

where m is the slope and c is the intercept. Substituting $F_k = y_k$, and $D_k = Y_k$, into eqn (3.66), we obtain

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary



3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Fig. 3.12 Fitting a straight line to noisy graphical data.

$$\chi^2 = \sum_{k=1}^{N} \frac{(m x_k + c - Y_k)^2}{\sigma_k^2} .$$

According to eqn (3.69), the components of ∇L are given by ($-1/2$ times) the partial derivatives of χ^2:

$$\frac{\partial \chi^2}{\partial m} = \sum_{k=1}^{N} \frac{2\,(m x_k + c - Y_k)\, x_k}{\sigma_k^2} \qquad \text{and} \qquad \frac{\partial \chi^2}{\partial c} = \sum_{k=1}^{N} \frac{2\,(m x_k + c - Y_k)}{\sigma_k^2} .$$

To make the subsequent algebraic manipulations easier, let us rewrite $\nabla\nabla\chi^2$ in the form of eqn (3.50):

$$\nabla\chi^2 = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} \begin{pmatrix} m \\ c \end{pmatrix} - \begin{pmatrix} p \\ q \end{pmatrix} , \tag{3.73}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

where the constants α, β, γ, p and q are related to the data through

$$\alpha = \sum w_k x_k^2 , \quad \beta = \sum w_k , \quad \gamma = \sum w_k x_k , \quad p = \sum w_k x_k Y_k , \quad q = \sum w_k Y_k ,$$

in which $w_k = 2/\sigma_k^2$ and the summations are from $k = 1$ to N. The best estimate (m_o, c_o), given by the minimum of χ^2, is defined by the requirement that $\nabla\chi^2 = 0$; the pair of linear simultaneous equations resulting from equating eqn (3.73) to zero can be solved directly, or by matrix inversion as in eqn (3.51), and yield

$$m_o = \frac{\beta p - \gamma q}{\alpha\beta - \gamma^2} \qquad \text{and} \qquad c_o = \frac{\alpha q - \gamma p}{\alpha\beta - \gamma^2} . \tag{3.74}$$

From eqn (3.71), the corresponding covariance matrix for these parameters is given by twice the inverse of $\nabla\nabla\chi^2$. By explicit differentiation, or comparison with eqns (3.50) and (3.52), the latter is the same as the 2×2 matrix in eqn (3.73); thus

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\begin{pmatrix} \sigma_{mm}^2 & \sigma_{mc}^2 \\ \sigma_{mc}^2 & \sigma_{cc}^2 \end{pmatrix} = 2\begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}^{-1} = \frac{2}{\alpha\beta - \gamma^2}\begin{pmatrix} \beta & -\gamma \\ -\gamma & \alpha \end{pmatrix}, \tag{3.75}$$

where the square root of the diagonal elements gives the (marginal) error-bars for the inferred values of m and c, and the γ-term describes how they are correlated.

If the error-bars for the data are not known, then we might do the analysis by assuming that they are all of the same size: $\sigma_k = \sigma$. By setting $w_k =$ a constant $= 2/\sigma^2$ in the definitions of α, β, γ, p and q above, we would find that the values of m_o and c_o were independent of the magnitude of σ; however, this would not be true of our estimate of the reliability of the optimal solution. As in Section 3.3, we must take the uncertainty about σ into account by marginalizing over it as a nuisance parameter:

$$\begin{aligned} \text{prob}(m, c|\{Y_k\}, I) &= \int_0^\infty \text{prob}(m, c, \sigma|\{Y_k\}, I)\, \mathrm{d}\sigma \\ &\propto \int_0^\infty \text{prob}(\{Y_k\}|m, c, \sigma, I) \times \text{prob}(m, c, \sigma|I)\, \mathrm{d}\sigma\,, \end{aligned}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

where we have used Bayes' theorem to express the joint posterior pdf on the right as a product of the likelihood function and a prior pdf. With the assumptions of independent, additive, Gaussian noise, the likelihood function is still of the form $\exp(-\chi^2/2)$:

$$\text{prob}(\{Y_k\}|m, c, \sigma, I) \propto \sigma^{-N} \exp\left[-\frac{1}{2\sigma^2}\sum_{k=1}^{N}(m x_k + c - Y_k)^2\right],$$

but we now have to be careful to explicitly retain all the factors which depend on σ. If we combine this with the assignment of a uniform prior, then the integral over σ can be evaluated with the same algebra as in Section 3.3; this results in a Student-t distribution, very similar to that of eqn (3.38):

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(m, c|\{Y_k\}, I) \propto \left[\sum_{k=1}^{N}(m x_k + c - Y_k)^2\right]^{-(N-1)/2}.$$

As usual, the best estimate of m and c, and their related covariance matrix, is given by the partial derivatives of the logarithm of this (marginal) posterior pdf. This leads to the recovery of the formulae of eqns (3.74) and (3.75), with the unknown value of the error-bar σ being replaced by an estimate S derived from the data:

$$S^2 = \frac{1}{N-1}\sum_{k=1}^{N}(m_o x_k + c_o - Y_k)^2\,.$$

In all the preceding analysis, we have implicitly assumed (in I) that the positions $\{x_k\}$ are known exactly. The case in which there is also a significant uncertainty with regard to the x-coordinate is more challenging, and we refer the avid reader to Gull (1989b) for a discussion of the subtleties involved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

3.6 Error-propagation: changing variables

The last topic in this chapter is often called the propagation of errors; despite the somewhat alarming title, suggestive of an unfortunate amplification of mistakes, it's quite innocuous. Here we are concerned with the question of how uncertainties in our estimate of a set of parameters translate into reliabilities of quantities derived from them. For example, suppose we are told that $X = 10 \pm 3$ and $Y = 7 \pm 2$; what can we say about the difference $X - Y$, or the ratio X/Y, or the sum of their squares $X^2 + Y^2$, and so on? In essence, the problem is nothing more than an exercise in the change of variables: given the pdf $\mathrm{prob}(X, Y|I)$, where the relevant information I includes a knowledge of the data if dealing with the posterior, we need the corresponding pdf $\mathrm{prob}(Z|I)$, where $Z = X - Y$, or $Z = X/Y$, or whatever, as appropriate.

Let us start with the easiest type of transformation; namely, one that involves a single variable and some function of it. Given that $Y = f(X)$, how is the pdf $\mathrm{prob}(X|I)$ related to $\mathrm{prob}(Y|I)$? Imagine taking a very small interval δX about some arbitrary point $X = X^*$; the probability that X lies in the range $X^* - \delta X/2$ to $X^* + \delta X/2$ is given by

$$\mathrm{prob}\left(X^* - \frac{\delta X}{2} \leqslant X < X^* + \frac{\delta X}{2} \,\middle|\, I\right) \approx \mathrm{prob}(X = X^*|I)\,\delta X\,, \tag{3.76}$$

where the equality becomes exact in the limit $\delta X \to 0$. Now suppose that we view this pdf as a function of another quantity Y, related (monotonically) to X by $Y = f(X)$; then, f will map the point $X = X^*$ (uniquely) to $Y = Y^* = f(X^*)$ and the interval δX to the corresponding region δY. The situation is shown schematically in Fig. 3.13. Since the range of Y values spanned by $Y^* \pm \delta Y/2$ is equivalent to a variation in X between $X^* \pm \delta X/2$, the area under the pdf $\mathrm{prob}(Y|I)$ should equal the probability represented by eqn (3.76). Therefore, we require that

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\mathrm{prob}(X = X^*|I)\,\delta X \;=\; \mathrm{prob}(Y = Y^*|I)\,\delta Y\,.$$

As this must be true for any point in X-space then, in the limit of infinitesimally small intervals, we obtain the relationship

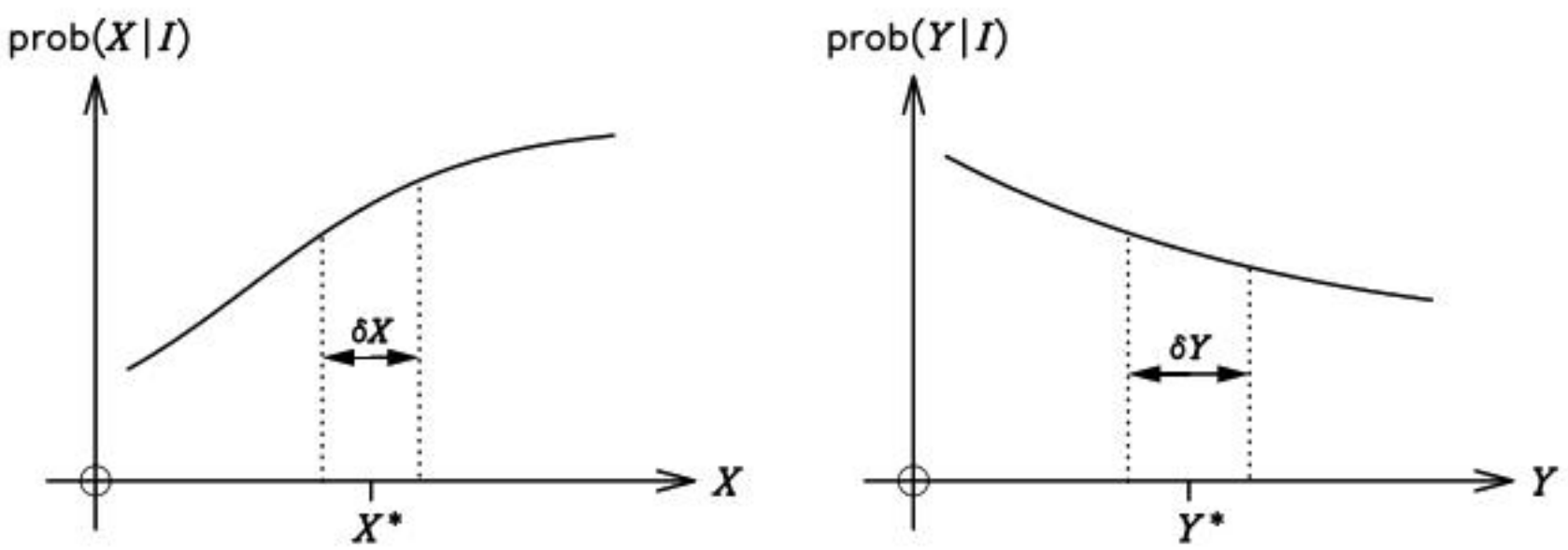


Fig. 3.13 Changing variables in one dimension: the function f maps the point X^* to $Y^* = f(X^*)$ and the small interval δX to the corresponding region δY.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(X|I) \;=\; \text{prob}(Y|I) \times \left|\frac{\mathrm{d}Y}{\mathrm{d}X}\right| , \tag{3.77}$$

where the term on the far right, given by the modulus of $\mathrm{d}f/\mathrm{d}X$, is called the *Jacobian*. The reason for taking the magnitude of the derivative is to ensure that it represents a ratio of lengths even when positive increments in X map to negative ones for Y. If $Y = f(X)$ is not a one-to-one transformation, then eqn (3.77) has to be extended to a summation over all the Y values which correspond to a given X.

As a concrete example of this procedure, let's return to the lighthouse problem of Section 2.4. There, in eqn (2.32), we had assigned a uniform pdf for the azimuth θ: $\text{prob}(\theta|\alpha,\beta,I) = 1/\pi$, where the angle θ had to lie between $\pm\pi/2$ radians. In eqn (2.33), we also had the connection between the azimuth and the position along the coast x: $\beta\tan\theta = x - \alpha$. Differentiating both sides of eqn (2.33) with respect to x, we have

$$\beta\sec^2\theta \times \frac{\mathrm{d}\theta}{\mathrm{d}x} \;=\; 1 .$$

Using the trigonometric identity $\tan^2\theta + 1 \equiv \sec^2\theta$, and substituting for $\tan\theta$ from eqn (2.33), the Jacobian can be written as

$$\frac{\mathrm{d}\theta}{\mathrm{d}x} \;=\; \left(\beta\Big[1+\tan^2\theta\Big]\right)^{-1} \;=\; \left(\beta\left[1+\left(\frac{x-\alpha}{\beta}\right)^2\right]\right)^{-1} .$$

Finally, we can use eqn (3.77) to transform the pdf for θ in eqn (2.32) to its equivalent form in terms of x; after a little algebraic rearrangement, we obtain (as promised) the Cauchy distribution of eqn (2.34):

$$\text{prob}(x|\alpha,\beta,I) \;=\; \text{prob}(\theta|\alpha,\beta,I) \times \left|\frac{\mathrm{d}\theta}{\mathrm{d}x}\right| \;=\; \frac{\beta}{\pi\Big[\beta^2+(x-\alpha)^2\Big]} .$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

The result in eqn (3.77) can be generalized to the case of several variables. Although it becomes more difficult to give a simple pictorial representation, like the one in Fig. 3.13, the central theme of the argument is the same: if we want to write the pdf for M parameters $\{X_j\}$ in terms of the same number of quantities $\{Y_j\}$ related to them, then we must ensure that

$$\text{prob}(\{X_j\}|I)\;\delta X_1\,\delta X_2\cdots\delta X_M \;=\; \text{prob}(\{Y_j\}|I)\;\delta^M\text{Vol}\big(\{Y_j\}\big) ,$$

where $\delta^M\text{Vol}\big(\{Y_j\}\big)$ is the M-dimensional volume in Y-space mapped out by the small hypercube region $\delta X_1\,\delta X_2\cdots\delta X_M$ in X-space. After some effort, most textbooks on mathematical methods (for scientists) derive the formula

$$\delta^M\text{Vol}\big(\{Y_j\}\big) \;=\; \left|\frac{\partial\,(Y_1\,,Y_2\,,\cdots,Y_M)}{\partial\,(X_1\,,X_2\,,\cdots,X_M)}\right|\;\delta X_1\,\delta X_2\cdots\delta X_M ,$$

where the strange-looking quantity in the modulus sign is the multivariate Jacobian; it is given by the determinant of the $M\times M$ matrix of partial derivatives $\partial Y_i/\partial X_j$. Thus the general form of eqn (3.77) can be written as

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(\{X_j\}|I) \;=\; \text{prob}(\{Y_j\}|I) \times \left|\frac{\partial\,(Y_1\,,Y_2\,,\cdots,Y_M)}{\partial\,(X_1\,,X_2\,,\cdots,X_M)}\right|\,. \tag{3.78}$$

To illustrate the use of eqn (3.78), let's consider the transformation of a pdf defined on a two-dimensional *Cartesian* grid (x, y) to its equivalent form in *polar* coordinates (R, θ). From Fig. 3.14(a), it can be seen that the functional relationship between the two sets of parameters is given by

$$x \;=\; R\cos\theta \qquad \text{and} \qquad y \;=\; R\sin\theta\,.$$

Taking the partial derivatives of x and y with respect to R and θ, we can easily evaluate the determinant of the resulting 2×2 matrix for the Jacobian:

$$\left|\frac{\partial\,(x,y)}{\partial\,(R,\theta)}\right| \;=\; \begin{vmatrix} \cos\theta & -R\sin\theta \\ \sin\theta & R\cos\theta \end{vmatrix} \;=\; R\left[\cos^2\theta + \sin^2\theta\right] \;=\; R\,,$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

where the final simplification is obtained from the identity $\sin^2\theta + \cos^2\theta \equiv 1$. According to eqn (3.78), therefore, the pdf $\text{prob}(R, \theta|I)$ is related to $\text{prob}(x, y|I)$ by

$$\text{prob}(R,\theta|I) \;=\; \text{prob}(x,y|I) \times R\,. \tag{3.79}$$

Thus if the pdf for x and y was an *isotropic*, bivariate, Gaussian,

$$\text{prob}(x,y|I) \;=\; \frac{1}{2\pi\sigma^2}\,\exp\left[-\,\frac{\left(x^2+y^2\right)}{2\,\sigma^2}\right]\,, \tag{3.80}$$

then the corresponding pdf for R and θ would take the form

$$\text{prob}(R,\theta|I) \;=\; \frac{R}{2\pi\sigma^2}\,\exp\left(-\,\frac{R^2}{2\,\sigma^2}\right)\,. \tag{3.81}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Rather than working through the formal Jacobian analysis given above, we could have obtained eqn (3.79) directly from a simple geometrical argument: the probability

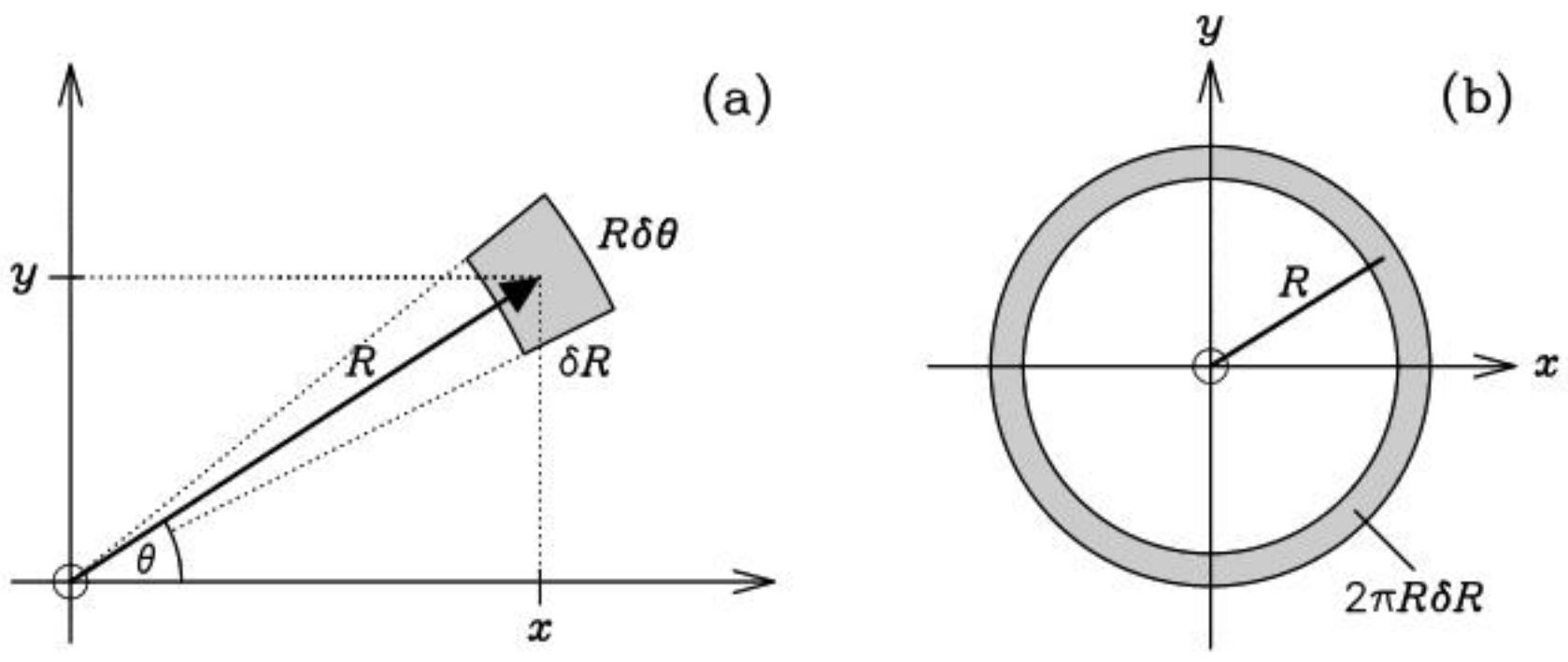


Fig. 3.14 Changing variables from Cartesian to polar coordinates.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

that the polar parameters lie in the small range $R \pm \delta R/2$, and $\theta \pm \delta\theta/2$, is given by the product of the value of the pdf $\mathrm{prob}(x, y|I)$ at the point corresponding to (R, θ) and the shaded element of area $R\,\delta\theta\,\delta R$ in Fig. 3.14(a); by the definition of the pdf for R and θ, of course, this is equal to $\mathrm{prob}(R, \theta|I)\,\delta\theta\,\delta R$. Hence, the desired result follows: $\mathrm{prob}(R, \theta|I)\,\delta R\,\delta\theta = \mathrm{prob}(x, y|I)\,R\,\delta R\,\delta\theta$.

Having transformed $\mathrm{prob}(x, y|I)$ to polar coordinates, it becomes a straightforward task to obtain the pdf for the radius $R = \sqrt{x^2 + y^2}$: we just need to marginalize the joint pdf $\mathrm{prob}(R, \theta|I)$ over θ. For the case of the bivariate Gaussian in eqn (3.80), eqn (3.81) is integrated readily with respect to θ:

$$\mathrm{prob}(R|I) = \int_0^{2\pi} \mathrm{prob}(R, \theta|I)\,\mathrm{d}\theta = \frac{R}{\sigma^2}\exp\left(-\frac{R^2}{2\sigma^2}\right). \tag{3.82}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Again, we could also have arrived at this result pictorially: since the value of the pdf in eqn (3.80) depends only on the distance from the origin, and not the direction, the probability that R lies in a narrow range δR is given by the product of the magnitude of the Gaussian pdf at that radius and the corresponding area of the shaded ring shown in Fig. 3.14(b); hence, $\mathrm{prob}(R|I)\,\delta R = \mathrm{prob}(x, y|I)\,2\pi R\,\delta R$ and eqn (3.82) follows immediately with the substitution of $x^2 + y^2 = R^2$ in eqn (3.80).

A multidimensional generalization of this last argument allows us to derive the χ^2 distribution mentioned towards the end of Section 3.3. The likelihood function of eqns (3.65) and (3.66) is an N-dimensional, isotropic, Gaussian when viewed in terms of the normalized residuals:

$$\mathrm{prob}(\boldsymbol{D}|\boldsymbol{X}, I) \propto \exp\left[-\frac{r_1^2 + r_2^2 + \cdots + r_N^2}{2}\right],$$

where $r_k = (F_k - D_k)/\sigma_k$. As in the case above, the value of the pdf depends only on the distance from the origin ($r_k = 0$); this radius $R = \sqrt{\Sigma r_k^2}$ is, of course, just the square root of χ^2. The probability that R lies in a narrow range δR, $\mathrm{prob}(R|\boldsymbol{X}, I)\,\delta R$, will equal the product of the magnitude of the likelihood function at that radius and the hypervolume of the associated spherical shell; since the latter is proportional to R^{N-1}, the marginal distribution is just an extension of eqn (3.82):

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\mathrm{prob}(R|\boldsymbol{X}, I) \propto R^{N-1}\exp\left(-R^2/2\right).$$

To convert this into a pdf for χ^2, we just need to carry out a one-to-one transformation according to eqn (3.77); using the functional relationship $\chi^2 = R^2$, we obtain

$$\mathrm{prob}\left(\chi^2|\boldsymbol{X}, I\right) \propto \left(\chi^2\right)^{N/2-1}\exp\left(-\chi^2/2\right). \tag{3.83}$$

Technically, this is called a χ^2 distribution with N degrees of freedom. For $N \geqslant 2$, it has a maximum at $N-2$; if the number of data is large, its shape is well-described by a Gaussian pdf summarized by $\chi^2 \approx N \pm \sqrt{2N}$.

We have now seen the basic ingredients required for the propagation of errors: it either involves a transformation in the sense of eqn (3.78), or an integration such as eqn

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

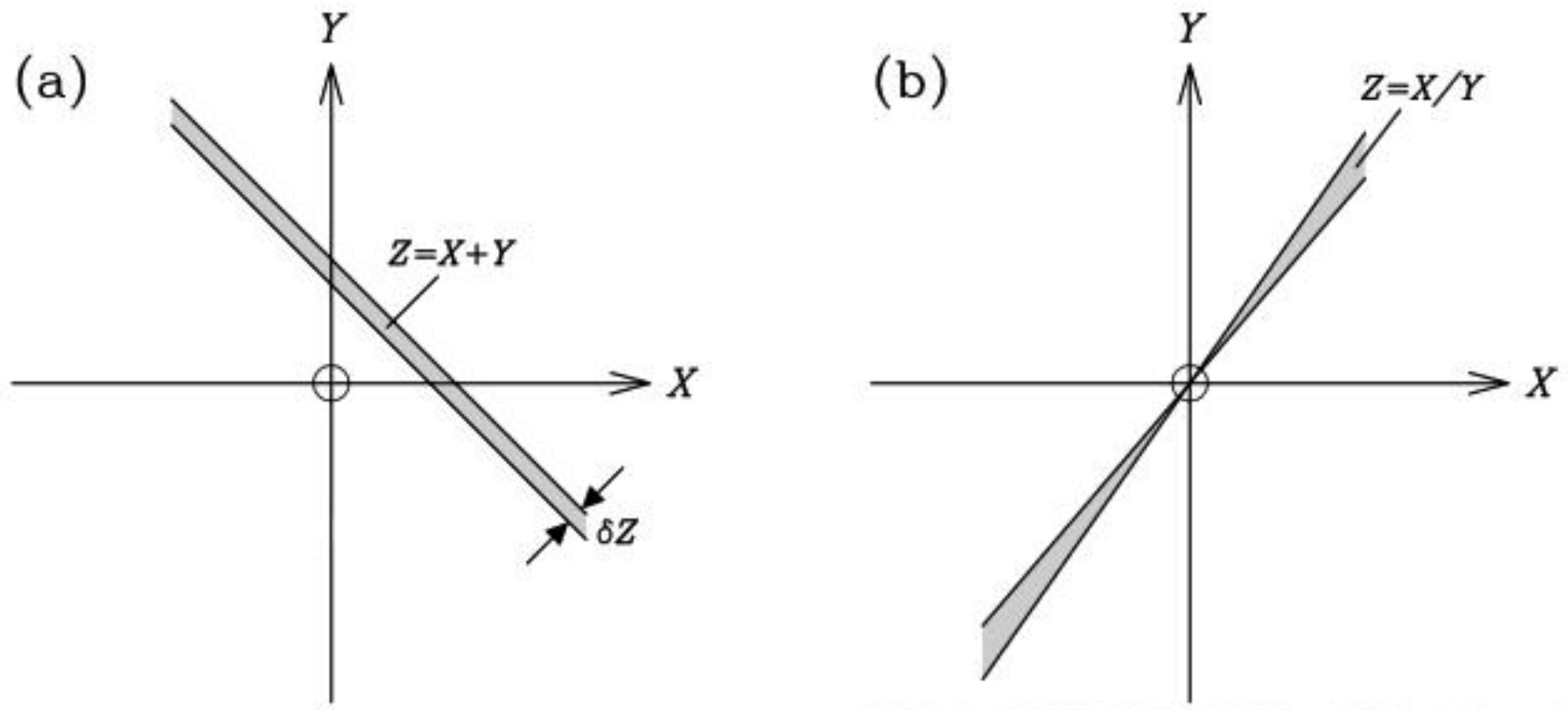


Fig. 3.15 The elements of area over which the joint pdf for the two parameters $\text{prob}(X, Y|I)$ must be integrated to obtain an estimate of: (a) the sum $Z = X + Y$; and (b) the ratio $Z = X/Y$.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

(3.82), or a combination of the two. For the common problem of wanting to estimate the sum $Z = X + Y$, or the ratio $Z = X/Y$, of two parameters X and Y, we simply need to integrate the joint pdf $\text{prob}(X, Y|I)$ along the shaded strips in Fig. 3.15. If this is not intuitively obvious, we can justify it analytically by using marginalization and the product rule:

$$\begin{aligned}\text{prob}(Z|I) &= \iint \text{prob}(Z|X, Y, I)\ \text{prob}(X, Y|I)\ \mathrm{d}X\,\mathrm{d}Y \\ &= \iint \delta\big(Z - f(X, Y)\big)\ \text{prob}(X, Y|I)\ \mathrm{d}X\,\mathrm{d}Y\,, \end{aligned} \tag{3.84}$$

where the Dirac δ-function in the second line is equal to zero unless $Z = f(X, Y)$, and indicates that Z is determined unambiguously (by the function f) when X and Y are known. Let's explicitly work through this procedure for the case $Z = X + Y$.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Following eqn (3.84), the integral along the diagonal strip in Fig. 3.15(a) can be written as

$$\text{prob}(Z|I) = \iint \text{prob}(X, Y|I)\ \delta\big(Z - (X + Y)\big)\ \mathrm{d}X\,\mathrm{d}Y\,. \tag{3.85}$$

If the information I only tells us that $X = x_\text{o} \pm \sigma_x$, and $Y = y_\text{o} \pm \sigma_y$, then it is reasonable to assume that these parameters are not correlated; $\text{prob}(X, Y|I)$ is then just the product of the individual pdfs for X and Y, so that eqn (3.85) becomes

$$\text{prob}(Z|I) = \int \mathrm{d}X\ \text{prob}(X|I) \int \text{prob}(Y|I)\ \delta\big(Z - X - Y\big)\ \mathrm{d}Y\,.$$

Since the δ-function is infinitely sharp (but has unit area), the Y integrand is zero unless $Y = Z - X$; the integral with respect to Y is trivial, therefore, and the pdf for the sum reduces to a *convolution* of the pdfs for X and Y:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(Z|I) = \int \text{prob}(X|I)\ \text{prob}(Y = Z - X\,|I)\ \mathrm{d}X\,. \tag{3.86}$$

The nature of the information I also suggests that we should assign Gaussian pdfs for the two parameters, with maxima at x_o and y_o, and widths σ_x and σ_y, respectively; substituting these into eqn (3.86), we have

$$\text{prob}(Z|I) = \frac{1}{2\pi\sigma_x\sigma_y}\int_{-\infty}^{+\infty} \exp\left[-\frac{(X-x_o)^2}{2\,\sigma_x^2}\right]\exp\left[-\frac{(Z-X-y_o)^2}{2\,\sigma_y^2}\right]\mathrm{d}X\,.$$

After some fairly tedious algebra, consisting largely of completing the square for X in the exponent and simplifying the resultant expression, we obtain

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\text{prob}(Z|I) = \frac{1}{\sigma_z\sqrt{2\pi}}\exp\left[-\frac{(Z-z_o)^2}{2\,\sigma_z^2}\right], \tag{3.87a}$$

where

$$z_o = x_o + y_o \qquad \text{and} \qquad \sigma_z^2 = \sigma_x^2 + \sigma_y^2\,. \tag{3.87b}$$

Thus the pdf for the sum is a Gaussian, with a maximum at z_o and width of σ_z; in fact, the pdf for the difference $Z = X - Y$ is the same except that z_o is then given by $x_o - y_o$.

3.6.1 A useful short cut

The calculation above seems rather long-winded for such a simple problem. Could we have got there by a simpler route? After all, we are not usually interested in knowing the fine detail of the shapes of the pdfs; for practical purposes, we're often satisfied to approximate them with Gaussians. Within such limits, the answer is 'Yes, there is an easier method!'

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Intuitively, we might have guessed that the best estimate for the difference of the two parameters was $x_o - y_o$; the corresponding error-bar, given σ_x and σ_y, requires a little more thought. Suppose we perturb $Z = X - Y$:

$$\delta Z = \delta X - \delta Y\,. \tag{3.88}$$

This would then tell us how small changes in the values of X and Y would affect our estimate of their difference. For the error-bars, of course, we are interested in studying the deviations from the optimal solution: $\delta X = X - x_o$, and so on. According to the integral definitions of eqns (3.25) and (3.26), and the information I, we know that

$$\langle \delta X^2 \rangle = \sigma_x^2\,, \qquad \langle \delta Y^2 \rangle = \sigma_y^2\,, \qquad \langle \delta X\,\delta Y \rangle = 0\,. \tag{3.89}$$

Note that by δX^2 we mean $(\delta X)^2$, not $\delta(X^2)$, and so on. Squaring both sides of eqn (3.88) and taking expectation values, we have

$$\langle \delta Z^2 \rangle = \langle \delta X^2 + \delta Y^2 - 2\,\delta X\,\delta Y \rangle = \langle \delta X^2 \rangle + \langle \delta Y^2 \rangle - 2\langle \delta X\,\delta Y \rangle\,,$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

where we have used the linear property that the integral of a sum of terms is equal to the sum of the integrals of those terms, in separating the expectations in the last step. Substituting from eqn (3.89), we obtain the error-bar of the difference $X-Y$:

$$\sigma_z = \sqrt{\langle \delta Z^2 \rangle} = \sqrt{\sigma_x^2 + \sigma_y^2} \, .$$

As promised, this is the same as for the sum $X+Y$ in eqn (3.87b).

For a second example of this procedure, let's work out the formula for the error-bar of the ratio of two parameters. Perturbing $Z = X/Y$, using the *quotient* rule of differentiation, we have

$$\delta Z = \frac{Y\,\delta X - X\,\delta Y}{Y^2} \, .$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

This expression can be simplified slightly by dividing through by the ratio itself:

$$\frac{\delta Z}{Z} = \frac{\delta X}{X} - \frac{\delta Y}{Y} \, .$$

Squaring both sides and taking expectation values, we obtain

$$\frac{\langle \delta Z^2 \rangle}{z_o^2} = \frac{\langle \delta X^2 \rangle}{x_o^2} + \frac{\langle \delta Y^2 \rangle}{y_o^2} - 2\,\frac{\langle \delta X\,\delta Y \rangle}{x_o\,y_o} \, ,$$

where the X, Y and Z in the denominator have been replaced by the constants x_o, y_o and z_o $(=x_o/y_o)$ because we are interested in the deviations from the optimal solution. Finally, substituting from eqn (3.89) and taking square roots gives the desired formula for the error-bar of the ratio:

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

$$\frac{\sigma_z}{z_o} = \sqrt{\left(\frac{\sigma_x}{x_o}\right)^2 + \left(\frac{\sigma_y}{y_o}\right)^2} \, . \tag{3.90}$$

Equivalently, changing variables to logarithms through eqn (3.77) gives us

$$\sigma_{\log Z} = \sqrt{\sigma_{\log X}^2 + \sigma_{\log Y}^2} \, ,$$

which illustrates the variability of $\log Z = \log X - \log Y$. Despite its virtues, let us end our discussion of error-propagation with a salutary warning against the blind use of this nifty short cut.

3.6.2 Taking the square root of a number

Data of the type shown in Fig. 3.3 are often collected in diffraction studies of crystalline materials. With reference to the associated model of Fig. 3.1, the amplitude of the Bragg (or signal) peak is usually estimated using a least-squares fitting program and the results stated as $A = A_O \pm \sigma_A$; any uncertainties about the background and the peak shape do, of course, give rise to a correspondingly larger (marginal) error-bar. Since this

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

amplitude is related to the modulus squared of the *complex structure factor*, $A = |F|^2$, crystallographers need to take square roots. Writing $f = |F|$, therefore, we have

$$A = f^2 . \tag{3.91}$$

For the purposes of this simple example, we are merely interested in obtaining the best estimate f_o and a measure of its reliability σ_f.

This seems like a fairly easy problem for the short cut procedure: the optimal solution is obviously $f_o = \sqrt{A_O}$; to propagate the error, we just differentiate eqn (3.91), square both sides, and take the expectation values:

$$\langle \delta A^2 \rangle = 4 f_o^2 \langle \delta f^2 \rangle = 4 A_O \langle \delta f^2 \rangle ,$$

3549a64f7957a5b7cc59fd64aa8d4446

Substituting $\sigma_A^2 = \langle \delta A^2 \rangle$, and rearranging this equation for the error-bar $\sigma_f = \sqrt{\langle \delta f^2 \rangle}$, we quickly arrive at the result

ebrary

$$f = \sqrt{A_O} \pm \frac{\sigma_A}{2\sqrt{A_O}} . \tag{3.92}$$

Unfortunately, this elementary analysis breaks down when A_O is negative! Such an occurrence is not unusual, particularly for weak and strongly overlapping *reflections* (or signal peaks). So, what has gone wrong?

Well, we have made two mistakes — both are quite common. The first stems from our failure to distinguish clearly between the posterior pdf and the likelihood function. The least-squares fit is just a shorthand way of saying that, as a function of the amplitude, $\text{prob}(\{data\}|A, I)$ is approximately Gaussian, with a maximum at A_O and a width of σ_A:

$$\text{prob}(\{data\}|A, I) \propto \exp\left[-\frac{(A - A_O)^2}{2\sigma_A^2} \right] . \tag{3.93}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Our inference about the value of A, however, is described by $\text{prob}(A|\{data\}, I)$. To relate these two pdfs, we need to use Bayes' theorem:

$$\text{prob}(A|\{data\}, I) \propto \text{prob}(\{data\}|A, I) \times \text{prob}(A|I) . \tag{3.94}$$

Since we know physically that the amplitude must be positive, we should, at least, assign a prior which is zero for negative values of A; a naïve choice which incorporates this information is

$$\text{prob}(A|I) = \begin{cases} \text{constant} & \text{for } A \geqslant 0 , \\ 0 & \text{otherwise} . \end{cases} \tag{3.95}$$

Multiplying eqns (3.93) and (3.95), we see that the best estimate of A will always be positive. This is true even when the amplitude A_O which gives the closest agreement with the data is negative; in that case, the posterior pdf will be a severely truncated Gaussian. The short cut procedure for the error-propagation will be flawed, therefore, because it implicitly relies on an expansion about a centrally located maximum; this

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

was our second mistake. Since there are no such limitations on the formal change of variables, from $\text{prob}(A|\{data\},I)$ to $\text{prob}(f|\{data\},I)$, we should do the proper calculation according to eqn (3.77).

Having obtained the posterior pdf for the amplitude of the Bragg peak, the transformation we require to describe our inference about the modulus of the structure factor is given by

$$\text{prob}(f|\{data\},I) \;=\; \text{prob}(A|\{data\},I) \times \left|\frac{\mathrm{d}A}{\mathrm{d}f}\right|,$$

where the derivative of eqn (3.91) yields the Jacobian $|\mathrm{d}A/\mathrm{d}f| = 2f$, with $f = |F| \geqslant 0$. This allows us to write the posterior pdf for f as

$$\text{prob}(f|\{data\},I) \;\propto\; f \times \exp\left[-\frac{\left(f^2 - A_\mathrm{O}\right)^2}{2\,\sigma_A^2}\right], \tag{3.96}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

for $f \geqslant 0$, and zero otherwise. As usual, the strange-looking expression of eqn (3.96) can be approximated by the more familiar Gaussian pdf:

$$\text{prob}(f|\{data\},I) \;\approx\; \frac{1}{\sigma_f\sqrt{2\pi}}\,\exp\left[-\frac{(f-f_\mathrm{o})^2}{2\,\sigma_f^2}\right], \tag{3.97}$$

where, according to eqns (2.12) and (2.15), the parameters f_o and σ_f are given by the first and second derivatives of the logarithm $L = \log_\mathrm{e}\big[\text{prob}(f|\{data\},I)\big]$:

$$2\,f_\mathrm{o}^2 \;=\; A_\mathrm{O} + \left(A_\mathrm{O}^2 + 2\sigma_A^2\right)^{1/2}, \tag{3.98a}$$

where

$$\sigma_f^{-2} \;=\; \frac{1}{f_\mathrm{o}^2} + \frac{2\left(3\,f_\mathrm{o}^2 - A_\mathrm{O}\right)}{\sigma_A^2}\,. \tag{3.98b}$$

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

It can easily be shown that these formulae for the best estimate of f, and its error-bar, reduce to the form of eqn (3.92) in the limit $A_\mathrm{O} \gg \sigma_A$. For those reflections where the amplitude of the Bragg peak is poorly determined by the data, the answers can be very different. Let us illustrate this graphically with a few examples.

Consider first of all a good reflection, $A = 9 \pm 1$, as depicted in the top panels in Fig. 3.16. It confirms that, in this case, the result of the short cut procedure agrees with the proper solution. Moving on to the next pair of pdfs, which are for a poorly determined amplitude, $A = 1 \pm 9$, we notice an interesting divergence. The solution of eqn (3.96), which is plotted with a solid line, indicates that the magnitude of the structure factor could be vanishingly small but is very unlikely to be larger than 6; the optimal estimate is about 2.6. By contrast, the result of eqn (3.92), shown as a dashed line, gives 1.0 for the best estimate and does not exclude values of up to about 12 as being unreasonable. It can also be seen that the Gaussian pdf of eqns (3.97) and (3.98), plotted with a dotted line, is a respectable approximation to eqn (3.96). Finally, the most striking case occurs when A_O is negative; for example, $A = -20 \pm 9$. The last two panels in Fig. 3.16 show that the proper probabilistic solution is perfectly well-behaved, whereas the short cut

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

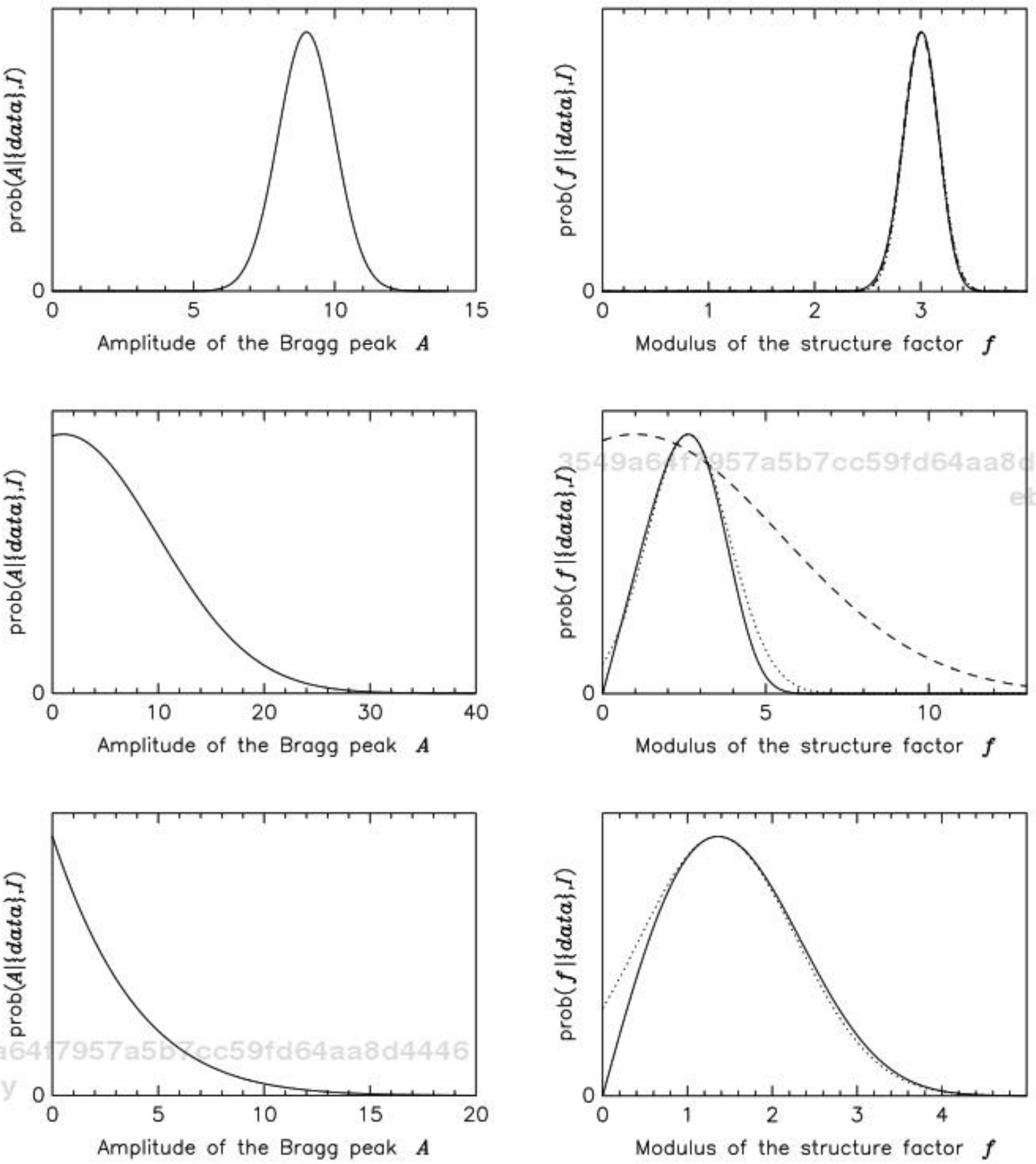


Fig. 3.16 The left-hand panels show the posterior pdf for the amplitude of the Bragg peak, given the naïve prior of eqn (3.95) and the least-squares likelihood function of eqn (3.93), with $A = 9 \pm 1$, 1 ± 9 and -20 ± 9. The right-hand plots are the corresponding pdfs for the modulus of the structure factor ($f = \sqrt{A}$): the solid, dotted and (when appropriate) dashed lines are for eqns (3.96), (3.98) and (3.92), respectively.

procedure breaks down completely because eqn (3.92) does not yield a real value for f. Thus we see how a simple recognition of the relevance of the prior, and strict adherence to the rules of probability theory, can offer a significant advantage over a 'cook-book' approach to least-squares and error-propagation. This analysis can be extended to the case of strongly overlapping peaks, which often occurs in powder diffraction; details can be found in Sivia and David (1994).

3549a64f7957a5b7cc59fd64aa8d4446
ebrary

Sivia, D. S., and Skilling, John. Data Analysis : A Bayesian Tutorial. Oxford, GBR: Oxford University Press, UK, 2006. ProQuest ebrary. Web. 26 June 2015.
Copyright © 2006. Oxford University Press, UK. All rights reserved.