# 3 Statistical models and inference

In this chapter I will introduce the principles of probabilistic inference. We will see how to set up models and will learn about the prior, likelihood, posterior, and evidence. I will show how inference works in practice using two simple examples: model comparison in the context of medical testing, and parameter estimation in astronomy. All of the issues covered in this chapter form the basis for deeper exploration in later chapters.

## 3.1 Introduction to data modelling

We perform experiments or make observations in order to learn about a phenomenon. We may describe the resulting data by calculating statistics and making plots. Such data explorations and summaries are useful – even essential – to get a feel for the data, but they are just a first step. To interpret the data we usually have to model them.

Typically we can only observe a phenomenon in part, and the data we obtain on it are noisy. Inference is the process of making general statements about a phenomenon, via a model, using noisy and incomplete data. The model represents the data in a form that gives us scientific meaning.

To do inference we must describe both the phenomenon itself and the measurement process. Consider modelling the orbit of a planet around its host star. We first define a relevant model $M$. This might describe the orbit as an ellipse (Keplerian orbit), as opposed to an oval or rosette. The model will have some parameters $\theta$ that describe the specific properties of the model. In the elliptical orbit case this would include the size (semi-major axis) and shape (eccentricity) of the ellipse, as well as its orientation in space. But when we observe the motion of a planet about a star, we do not observe directly the shape of the orbit or any of the other parameters. We instead see the planet at different positions (and with different velocities) at different times. The *generative model* (also called the *forward model*) is the theoretical entity that generates (or simulates) the observable data from the model parameters. Normally this is a mathematical equation. In this example it would be a deterministic equation derived from the physical laws of mechanics and gravity: given the model parameters and time of the observation, the position and velocity of the planet can be simulated exactly. Other generative models may be stochastic and only predict a distribution. An example is the kinematic theory of gases, which predicts the distribution of the speeds of gas molecules, but not the speeds of individual molecules.

The generative model predicts the data we would measure (if our model represents reality) in the absence of noise. Yet measurements are always noisy: we won't observe the planet at exactly the position predicted by the model due, for example, to the finite resolution of our instruments or the blurring caused the Earth's atmosphere. We need to know how the measurement process affects our data. For this purpose we must also define a *measurement model*, sometimes also called a *noise model* (discussed already in section 2.2). This describes a probability distribution over possible observations given the ideal (noise-free) data and generally has some parameters too. A widely used example is a Gaussian in the variable $x$ (equation 1.60) in which the mean is the ideal data (predicted by the generative model), $x$ is the data we measure, and the standard deviation is the typical size of the uncertainty in the measurement. This probability distribution is in fact the likelihood, which we shall discuss properly in section 3.3 below.

The key to data modelling is to use the data together with the generative model and measurement model to make consistent, probabilistic inferences. Broadly speaking, there are three different things we will want to do given some data $D$.

(1) **Parameter estimation**. For a specified model $M$ with parameters $\theta$, infer the values of the model parameters, or more specifically, infer the PDF $P(\theta \,|\, D, M)$. This quantity is known as the *parameter posterior PDF*. In terms of our planet orbit example, this means infer the (multidimensional) PDF over the orbit parameters, and perhaps the parameters of the measurement model too.

(2) **Model comparison**. Given a set of different models $\{M_i\}$, find out which one is best supported by the data. Ideally this means finding $P(M_i \,|\, D)$, the *model posterior probability*. But we shall see that ratios of $P(D \,|\, M_i)$ for the various $M_i$ are often adequate. In our example, this corresponds to finding the absolute or relative probabilities of different models for the shape of the orbit (ellipse, oval, rosette, etc.).

(3) **Prediction**. Given a model $M$ (which may have been identified/fitted from the data), predict the data at some new location. Having determined that the planet does have an elliptical orbit, for example, find the PDF over the position of the planet at some future time.

Much of the rest of this book is concerned with showing how we put these ideas into practice. I start in the next section by looking at model comparison for the case of models having no parameters (or rather, we ignore the details of their parametrization). I will then introduce the concept of parameter estimation in section 3.3, and will apply this to a simple one-parameter model in section 3.5. More complicated problems and more sophisticated inference methods will be introduced in the following chapters.

## 3.2 Bayesian model comparison

### 3.2.1 Theory

Medical tests are done to find out whether a patient has a certain disease. Let $M$ denote that a patient has a particular disease and $M'$ that he does not. These two cases are mutually exclusive and exhaustive, so $P(M') = 1 - P(M)$. A medical test produces some data $D$

which might be a simple positive/negative test result, or it could be a more extensive set of blood values, heart rate measurements, etc. We would like to use these data to find out how probable it is that the patient has the disease. That is, we would like to find $P(M\,|\,D)$. As $P(M'\,|\,D) = 1 - P(M\,|\,D)$, we automatically find out the probabilities of both models. We are doing model comparison between two complementary models.

From the fundamental laws of probability (section 1.2.1)

$$\begin{aligned} P(D) &= P(D, M) + P(D, M') \\ &= P(D\,|\,M)P(M) + P(D\,|\,M')P(M') \end{aligned} \tag{3.1}$$

where $P(D)$ is the probability (or probability density)[1] of having observed this particular piece of data at all (under either model). We are interested in $P(M\,|\,D)$. This is related to the above quantities via Bayes' theorem (section 1.2.4)

$$P(M\,|\,D) = \frac{P(D\,|\,M)P(M)}{P(D)}. \tag{3.2}$$

Substituting equation 3.1 into this gives

$$P(M\,|\,D) = \frac{P(D\,|\,M)P(M)}{P(D\,|\,M)P(M) + P(D\,|\,M')P(M')}. \tag{3.3}$$

Dividing by the numerator we can write this as

$$P(M\,|\,D) = \frac{1}{1 + \frac{1}{R}} \tag{3.4}$$

where

$$R = \frac{P(D\,|\,M)P(M)}{P(D\,|\,M')P(M')} \tag{3.5}$$

is the *posterior odds ratio* of the two models.[2] We see from equation 3.2 that this also equals the ratio of the posterior probabilities for the two models, $P(M\,|\,D)/P(M'\,|\,D)$ (because $P(D)$ cancels).

In order to determine the posterior probability that the model is true given the data, $P(M\,|\,D)$, we therefore need three quantities

- $P(D\,|\,M)$, the probability of measuring $D$ when $M$ is true;
- $P(D\,|\,M')$, the probability of measuring $D$ when $M$ is not true;
- $P(M)$, the probability that $M$ is true, independent of the data – the *prior probability*.

This last quantity tells us how probable the model is – how likely the patient has the disease – prior to obtaining the data, i.e. independent of the test result. Typically we have some information about this, such as from medical histories of the disease. But if we really had no information to tell us that $M$ is more or less likely than $M'$, then we could set

[1] $M$ is a discrete proposition, so $P(M)$ and $P(M\,|\,D)$ are actual probabilities, whereas $D$ could be a real variable (scalar or vector), in which case $P(D)$ and $P(D\,|\,M)$ would be probability densities. This mixing is neither a conceptual nor a practical difficulty. The laws of probability hold and the dimensions of the equations agree. As the nature of the variable will determine whether we are talking about a probability or a probability density, for simplicity I will just refer to $P$ as a probability.

[2] Be aware that the term "odds ratio" also has other definitions in statistics.

**Table 3.1** The meaning of terms used to describe test results. Columns must sum to 1. Rows generally do not.

| Result $D$ | Model $M$ true? | |
|---|---|---|
| | **yes** | **no** |
| **positive** | true positive<br>$P(D \mid M)$ | false positive<br>$P(D \mid M')$ |
| **negative** | false negative<br>$P(D' \mid M)$ | true negative<br>$P(D' \mid M')$ |

$P(M) = P(M')$. These terms then cancel in the odds ratio, leaving what is known as the *Bayes factor*

$$BF = \frac{P(D \mid M)}{P(D \mid M')}. \tag{3.6}$$

This is the ratio of the probabilities of the data under each model.

If we have many different models $\{M_i\}$ (perhaps these are different diseases) then equation 3.1 has to be replaced by

$$P(D) = \sum_i P(D \mid M_i) P(M_i) \tag{3.7}$$

(this is equation 1.15) where the sum is over *all* models, i.e. $\sum_i P(M_i) = 1$. This requires that the set of models be mutually exclusive (no disease is a subclass of another) and exhaustive (one of the models has to be the absence of all the other diseases). It follows from equation 3.2 that the posterior probability of any one model, say $M_1$, is

$$P(M_1 \mid D) = \frac{P(D \mid M_1) P(M_1)}{\sum_i P(D \mid M_i) P(M_i)}. \tag{3.8}$$

If we do not have a complete set of models then we cannot compute these posterior probabilities. However, we can still compute the odds ratio or Bayes factor between any two models. In practice we would calculate these against a common baseline model (e.g. the absence of all the other diseases).

Let's now put this theory into practice with an example.

### 3.2.2 Does a positive test result mean you have a disease?

A certain test for breast cancer is 90% reliable, which means that if the person has breast cancer she will test positive with a probability of 0.9. The probability of testing positive in absence of the disease (i.e. by mistake) is 0.07. Among women aged 40 to 50 showing no symptoms, 8 in 1000 have breast cancer.[3] A 40-something woman tests positive. What is the probability that she has breast cancer? Make an intuitive guess before reading on.

[3] These numbers have been taken from Gigerenzer (2002).

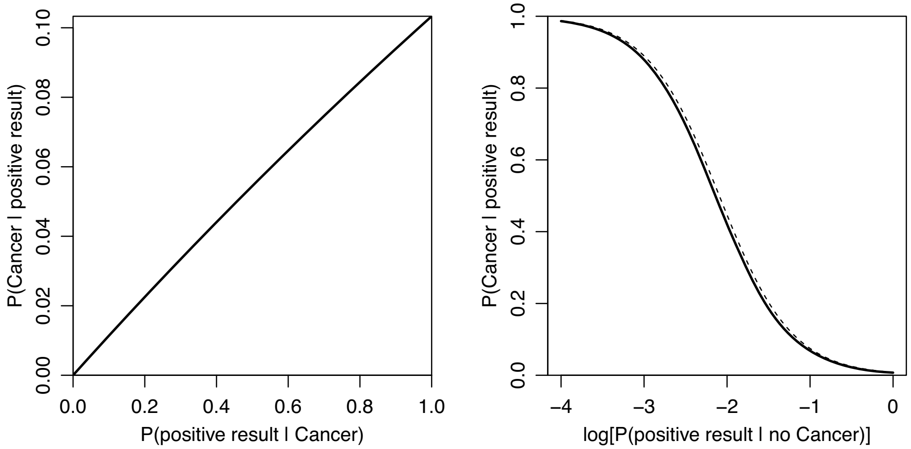

Fig. 3.1 The variation of the probability of having cancer given a positive test result $P(M\,|\,D)$ as a function of (left) the true positive rate $P(D\,|\,M)$ and (right) the false positive rate $P(D\,|\,M')$, in each case keeping the other quantities fixed. The dashed line in the right panel is for $P(D\,|\,M) = 1$. Note the log (base 10) scale on the horizontal axis in the right panel.

We first need to understand the information given in the question. Let $D$ be the test result (data), which is either positive or negative. Let $M$ be the hypothesis (model) that the woman has cancer. We are told that the true positive rate, $P(D\,|\,M)$, is 0.9. The figure of 0.07 is the false positive rate, $P(D\,|\,M')$. (Table 3.1 may be a useful reference for these terms.) The final statement is the *base rate* – the prior probability that $M$ is true – which tells us that $P(M) = 0.008$.

We can now use equation 3.4 to compute $P(M\,|\,D)$, the posterior probability that the woman has cancer given that she has tested positive. The answer is low, just 0.09. Many people are surprised by this. Some people confuse the quantity we want with the reliability $P(D\,|\,M)$, and therefore expect an answer of 0.9. Others incorrectly interpret the false positive rate $P(D\,|\,M')$ as a statement about the disease, rather than a statement about the test. They erroneously equate it with $P(M'\,|\,D)$, and then conclude the probability of having the disease, which is $P(M\,|\,D) = 1 - P(M'\,|\,D)$, to be 0.93. Bayes theorem' shows that it is neither of these. The correct answer is much smaller in this case because it depends on the small base rate. Even though a positive test result is quite probable when the woman has the disease, it is very unlikely that she does have the disease. And although a positive test result in the absence of the disease is quite improbable, it is very likely that she doesn't have the disease in the first place. The first possibility – positive result, disease present – is quantified by $P(D, M) = P(D\,|\,M)P(M)$. The second possibility – positive result, disease absent – is quantified by $P(D, M') = P(D\,|\,M')P(M')$. It is the ratio of these which is decisive. This ratio is $R$ (equation 3.4) and this determines $P(M\,|\,D)$ (equation 3.5).

We can understand this better by investigating how the result varies as a function of the assumed probabilities (you can use the R code at the end of this section to do this). The left panel of figure 3.1 shows how the probability of having cancer given a positive test result, $P(M \,|\, D)$, varies as a function of the true positive rate $P(D \,|\, M)$, keeping the false positive rate and base rate constant. It is virtually linear because the odds ratio $R$ (equation 3.5) is much less than one ($< 0.12$ for all $P(D \,|\, M)$), so it follows from equation 3.4 that $P(M \,|\, D) \simeq R = \text{constant} \times P(D \,|\, M)$. If you redo the plot with a larger base rate you will see nonlinearity.

The right panel of figure 3.1 shows how $P(M \,|\, D)$ varies as a function of the false positive rate $P(D \,|\, M')$, keeping the true positive rate and base rate constant at their original values. It shows that the false positive rate has to be quite low before the test gives reliable results. For example, it must drop below 0.0072 ($\log P(D \,|\, M') = -2.14$) before the probability of having the disease given a positive test result rises above 0.5. To get a correct detection of the disease with a probability of at least 0.9 would require a false positive rate of less than $8 \times 10^{-4}$.

Improving the true positive rate will not help this test. Even when it is 1.0 the curve in the right panel of figure 3.1 hardly changes (the dashed line). The base rate in this example is $P(M) = 0.008$ or $\log P(M) = -2.10$. Inspection of the equations shows that only once the false positive rate drops significantly below the base rate does the test start to become useful. As you probably have no control over the base rate, getting a better test means lowering the false positive rate.[4]

Here is a useful rule of thumb: assuming a test has a high reliability (true positive probability), a positive result only means you are likely to have the disease if the false positive rate is lower than the base rate.

If you test positive for a rare disease, do the above analysis. Better still, do it to help you decide whether or not to take the test in the first place. Given that false positives can have detrimental consequences (such as worry and unnecessary or harmful treatment), it may be that for some people, some tests are not worth doing.

We shall look in more detail at the problem of model comparison in chapter 11, including for the case that the models have free parameters.

R file: `cancer_test.R`

```
##### Apply Bayes theorem to infer cancer probability given test result

# Return P(M|D) given P(D|M), P(D|M’), P(M)
post <- function(p_d_m, p_d_nm, p_m) {
  p_nm  <- 1-p_m
  oddsr <- (p_d_m * p_m) / (p_d_nm * p_nm)
  p_m_d <- 1/(1+1/oddsr)
  return(p_m_d)
}

# Vary reliability of test P(D|M)
```

[4] We are assuming here that we know the base rate exactly. This is often not the case, and an uncertainty in this will propagate into an uncertainty in our value of $P(M \,|\, D)$. In the current case, where $P(M \,|\, D) \propto P(D \,|\, M)$, this propagation is linear: a 10% error in the base rate gives approximately a 10% error in the posterior, for example.

```
p_d_m  <- seq(0.0, 1.0, 0.01) # prob. true positive
p_d_nm <- 0.07                # prob. false positive
p_m    <- 8/1000              # prior probability of m
p_m_d  <- post(p_d_m, p_d_nm, p_m)
pdf("cancer_test_1.pdf", 4, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(p_d_m, p_m_d, type="l", lwd=2, xaxs="i", yaxs="i",
     xlab="P(positive result | Cancer)", ylab="P(Cancer | positive result)")
dev.off()

# Vary false positive probability P(D|!M)
p_d_m  <- 0.9               # prob. true positive
p_d_nm <- 10^seq(-4,0,0.02) # prob. false positive
p_m    <- 8/1000            # prior probability of m
p_m_d  <- post(p_d_m, p_d_nm, p_m)
pdf("cancer_test_2.pdf", 4, 4)
par(mfrow=c(1,1), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(log10(p_d_nm), p_m_d, type="l", lwd=2, ylim=c(0,1), yaxs="i",
     xlab="log[P(positive result | no Cancer)]",
     ylab="P(Cancer | positive result)")
lines(log10(p_d_nm), post(p_d_m=1.0, p_d_nm, p_m), lty=2)
dev.off()
```

### 3.2.3 Thinking in terms of frequencies may help

Probabilities are not frequencies. A frequency refers to how often something occurs when repeated in some way, whereas probability refers more broadly to knowledge and certainty (or lack thereof) gained from data. Nonetheless, it can often help our understanding to use frequencies to represent probabilities.

Let's consider the breast cancer example again, but with a $100\%$ true positive rate, $P(D\,|\,M) = 1$. As before the false positive rate is 0.07 and the base rate is $8/1000$. Imagine a set of 10 000 women in the defined group who undergo the test. Of these, a fraction $8/1000$ have breast cancer, which is 80, and the remaining 9920 do not. Of the 80 with cancer, all will test positive. Of the 9920 without cancer, $0.07 \times 9920 = 694$ will test positive. In total $694 + 80$ have tested positive. Thus the probability of having the disease following a positive test result is

$$\frac{80}{80 + 694} \simeq 0.10. \tag{3.9}$$

This is of course the same calculation as before, but now carried out more intuitively.

## 3.3 Data modelling with parametric models

As discussed earlier in this chapter, the generative model is the theoretical entity that predicts observable data from the model parameters. The model in the previous section had no (explicit) parameters: the model was either true or false. Most models have one or more parameters, which I will denote by $\theta$. A simple example of a generative model with parameters is the straight line

$$f(x; a, b) \;=\; a + bx. \tag{3.10}$$

The parameters of this are the intercept $a$ and the slope $b$. The measurements $y$ will differ from this due to noise, so we write

$$y \;=\; f(x; a, b) + \epsilon \tag{3.11}$$

where $\epsilon$ denotes a number drawn from a random distribution, the noise model (measurement model), which typically also has parameters.

Given a set of data – in the line example a set of measurements $D = \{y\}$ at specified values of $\{x\}$ – we want to infer the values of the parameters of the generative model (and perhaps of the measurement model too). In some cases one might be interested in finding just the single "best" set of parameters that predict the data. But as the data are noisy, there is no unique solution. So more generally we want to know what range of parameters is supported by the data (and its noise). That is, we would like to find the probability distribution over the parameters. The broader the distribution the less well constrained are the parameters and the larger our uncertainty. We can quantify this with the standard deviation, for example. There may also be multiple regions of high probability density, corresponding to multiple solutions. These could never be characterized by a single "best" solution.

The probability density function we are after is $P(\theta\,|\,D, M)$. This is the PDF over the parameters given both the data $D$ and what we know already about the model $M$. An example of such prior knowledge might be that we know on physical grounds that the slope cannot be negative. The PDF $P(\theta\,|\,D, M)$ is called the parameter *posterior* PDF. We use Bayes' theorem to express this in terms of other quantities which, as we will see, are readily available. The difference with respect to the previous section is that we now apply Bayes' theorem to the model parameters, rather than to the model itself, to give

$$P(\theta\,|\,D, M) = \frac{P(D\,|\,\theta, M)P(\theta\,|\,M)}{P(D\,|\,M)}. \tag{3.12}$$

Once we have determined the three terms on the right side of the above equation then we have determined the posterior PDF over the parameters. Recall that all four terms in this equation are probability density functions in whatever quantity appears before the bar "|". (In the previous section we dealt with discrete variables – test positive or negative; model true or false – so all the quantities there were actual probabilities.) Let us now examine these four terms to understand what they mean and how we can determine them. I will then use a simple example to illustrate how we apply this in practice. More advanced examples follow in later chapters.

### 3.3.1 The likelihood

The first term in the numerator of equation 3.12, $P(D\,|\,\theta, M)$, is the *likelihood*. This is the key function in data modelling because it describes both the phenomenon and the measurements. It tells us the probability of getting the data we measured given some value of the parameters. Although this is the probability of $D$ conditional on $\theta$ (and $M$), once the data have been measured they are fixed, so we are more interested in the likelihood's dependence on the parameters. Some authors therefore refer to the "likelihood of the parameters". Yet it remains a PDF in the data with units $D^{-1}$, so we should really refer to the "likelihood of the data" (as "likely" is just a synonym for "probable").

$M$ specifies both the measurement model and the generative model. In terms of our straight line fitting example, the generative model is the equation for the line $f(x; a, b)$, and the measurement model tells us how the measurements of $y$ at a given $x$ differ from $f(x; a, b)$ on account of the noise. The measurement model describes the distribution of $\epsilon$ in equation 3.11. An example is a Gaussian with standard deviation $\sigma$. Here $\sigma$ is the parameter of the measurement model. The likelihood for any one point $y$ is then

$$P(y \,|\, \theta, M) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{[y - f(x; a, b)]^2}{2\sigma^2}\right]. \tag{3.13}$$

This tells us that the measurement has a Gaussian distribution about the true value. Note that $\theta = (a, b, \sigma)$ is the union of parameters from both the generative model and measurement model.

## 3.3.2 The prior

The other term in the numerator of equation 3.12 is the *prior* $P(\theta \,|\, M)$, which is a PDF over the model parameters. It encapsulates the information we have, independent of the data, about the possible values of the model parameters. In general this covers the parameters of both the generative and measurement models. In the case of the line fitting example the prior may tell us the physically permitted ranges of the slope and intercept, or more generally, what their relative probabilities are. This will be informed by our background knowledge of the problem. It is called the prior because it is the information we have prior to obtaining the data.

Sometimes our prior information is vague and/or it may be difficult to express as a probability distribution. Different people may have different information or different opinions on what prior information is relevant. This is not a weakness of inference. It just reflects the reality that we do not only use our immediate measurements to reach scientific conclusions. I shall say more about priors in section 3.5 and in later chapters – in particular in section 5.3 – so I will defer further discussion until then.

## 3.3.3 The posterior

The posterior $P(\theta \,|\, D, M)$ is the PDF over the model parameters given the data and the background information on the model. Equation 3.12 tells us that the posterior is proportional to the product of the likelihood and the prior. The denominator in that equation, $P(D \,|\, M)$, is independent of $\theta$ so we can think of it as just a normalization constant for now. Thus once we have defined the prior and likelihood, and we've measured some data, we get the posterior probability density to within a normalization constant $Z$, i.e.[5]

$$P(\theta\,|\,D, M) \;=\; \frac{1}{Z}\, P(D\,|\,\theta, M)P(\theta\,|\,M). \tag{3.14}$$

Conceptually, at least, inference really is that straightforward. Following my convention (section 1.2.6) I write the unnormalized posterior using an asterisk

$$P^*(\theta\,|\,D, M) \;=\; P(D\,|\,\theta, M)P(\theta\,|\,M). \tag{3.15}$$

Both the posterior and the prior are PDFs over the model parameters. The difference is that whereas both are conditional on background information, the posterior is conditional on the data as well. This additional dependence is provided by the likelihood. Thus we can see Bayesian inference as a process of improving our knowledge of the parameters by using the data: we update the prior using the likelihood to obtain the posterior. In section 5.1 I will illustrate how this updating works and how it combines the information about the parameter contained in the data and in the prior. We will see another illustration of this in section 9.1.6.

This posterior PDF tells us everything we want to know about the relative probabilities of different parameters. Often we will want to summarize both the typical value of the parameters, for example using the maximum (mode) or mean, and their range, for example using the (co)variance or quantiles. We will discuss summaries of the posterior further in section 5.5.

Sometimes we will adopt a prior that is uniform over the possible range of parameters. In that case the posterior is directly proportional to the likelihood. But we must not forget that the posterior is a PDF over the parameters, not the data. So even if the likelihood is a common function of the data (like a Gaussian), it and the posterior will generally be a different function of the parameters. It follows that even though the likelihood is a normalized function *of the data*, the product of the likelihood with a uniform prior will not generally be a normalized function *of the parameters*.

Certain choices of the likelihood and prior can result in a convenient form for the posterior PDF. In particular, there are cases in which the posterior is in the same "family" of functions as the prior, in which case the prior and posterior are called *conjugate distributions*. For example, if the likelihood is Gaussian (in the data $x$) and the parameter of interest is its mean $\mu$ (the variance is fixed), then the likelihood is also a Gaussian in the mean (inspect equation 1.60). So if the prior on $\mu$ is also chosen to be a Gaussian then the posterior over $\mu$ is likewise Gaussian, because the product of two Gaussians (likelihood and prior) is another Gaussian. Conjugate priors are often used because they make the mathematics easier. We shall encounter various conjugate distributions in chapter 5.

I will have a lot more to say about the parameter posterior. Indeed, chapters 5 to 9 are concerned primarily with how, in practice, we estimate the posterior given data and priors.

[5] The normalization constant of the prior is also independent of $\theta$, so we could absorb it into the proportionality too. The normalization constant of the likelihood, in contrast, is generally dependent on $\theta$. We must therefore use a normalized likelihood when computing the posterior.

### 3.3.4 The evidence

The denominator in equation 3.12 is called the *evidence*. It gives the probability, assuming model $M$ to be true, of observing the data $D$ at all, for any values of $\theta$. I referred to the evidence earlier as "just" a normalization constant for the posterior, because it is the integral of the numerator of equation 3.12 – the product of the likelihood with the prior – over all $\theta$

$$P(D\,|\,M) = \int P(D\,|\,\theta, M) P(\theta\,|\,M)\, d\theta. \tag{3.16}$$

This can be thought of as the integral of the likelihood (a PDF in $D$) over the prior (a PDF in $\theta$), so it is also called the *marginal likelihood*. We shall see in chapter 11 that the evidence plays a key role in model comparison.[6]

The evidence, as a normalization constant, is important if we want to compute certain quantities from the posterior. If we are only interested in the shape of the posterior or the relative probabilities of solutions, or certain statistics like the mode and full-width at half-maximum, then we don't need to normalize the posterior. But if we want to find the mean, standard deviation, quantiles, or anything else that requires us to integrate the posterior, then we need to normalize the posterior.[7]

Sometimes the posterior has a nice functional form, in which case its integral may be analytic or given by a standard integral. This is the case for all the PDFs introduced in section 1.4 and we will encounter more cases in chapter 6. But for many real-world problems we will have to resort to a numerical integration. If the posterior is univariate and non-zero over only a finite range $(x_1, x_2)$, we can do a brute force numerical integration: we evaluate the function on a dense grid and compute the area as the sum of a series of narrow rectangles each of width $\delta x$ and height $x_i$, i.e.

$$\int_{x_1}^{x_2} P(x)\, dx \simeq \sum_{i=1}^{N} P(x_i)\, \delta x \tag{3.17}$$

where $\delta x = (x_2 - x_1)/N$ and the $\{x_i\}$ are at the centers of the rectangles.[8] This is illustrated in figure 3.2 and is known as a Riemann sum.

Better approximations than this rectangle method are the trapezium rule and Simpson's

[6] We already used the evidence for model comparison in the cancer test example in section 3.2.2, where the two models were having cancer and not having cancer. The four probabilities listed in table 3.1 are mathematically equivalent to evidences, but as that problem didn't involve a model with parameters I did not call them evidences.

[7] To see this, let $x$ be a real number constrained to lie between 0 and 1, and let $P(x) = 2x$ be a normalized PDF. This has a mean $\int_0^1 2x^2 dx = 2/3$. If we now multiply $P(x)$ by a constant ($\neq 1$) it is no longer normalized, and we would calculate a different (incorrect) mean. The mode, in contrast, is unchanged. My statement about the need to normalize the posterior assumes we are making computations using the functional form of the posterior PDF. If we instead have a set of samples drawn *from* the PDF, then because these are characteristic of the distribution, you can calculate the mean, standard deviation, quantiles, etc., from these samples without evaluating the normalization constant. We will see why this is when we discuss sampling from distributions in chapter 8.

[8] In order to plot a distribution we also just evaluate it on a dense grid and plot it at these as points. If we use some software to plot it as lines, then the software does some kind of interpolation between the point evaluations.

https://doi.org/10.1017/9781108123891.004

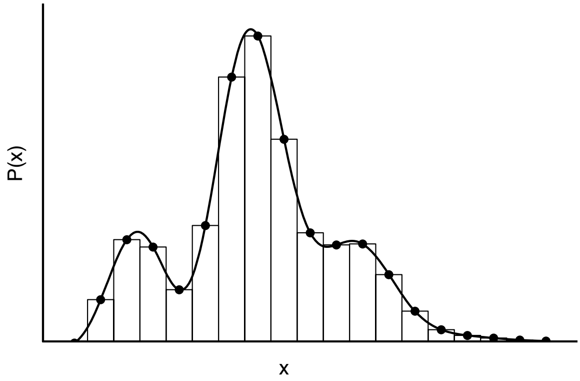

Fig. 3.2 Illustration of the rectangle method for integrating a one-dimensional function.

rule. These use linear and quadratic interpolation (respectively) between neighbouring points – as opposed to constant interpolation in the case of the rectangle method – and so obtain higher accuracy for a given number of evaluations of $P(x)$ (although if $P(x)$ is fast to compute we can just use the rectangle method with narrower spacings). These approaches can be generalized to two dimensions using a double summation, but they become computationally intractable in higher dimensions. This is because the number of function evaluations required to achieve a given precision grows exponentially with the number of dimensions. This problem is known as the *curse of dimensionality* (an unusually evocative name for the field of statistics). For more complex functions in one or two dimensions, more sophisticated methods such as Gaussian quadrature may be required to do integration efficiently. Challenges are presented by higher dimensional functions, and by functions with probability concentrated in a very narrow region (as gridding techniques may miss the peak). For these we can use Monte Carlo methods, which we will turn to in chapter 8.

## 3.4 Making decisions

In section 3.2.2 we computed the posterior probability of having a disease given a positive medical test result. From the point of view of data analysis, this is the end of the problem. But if we need to decide whether or not to have treatment for the disease, we need to go further. To do this we need to know both the net benefit of the treatment for the case that we do have the disease, and the net cost of the treatment for the case that we do not have the disease. In an abstract sense the net benefit of a decision is called its *utility*, denoted $U$. The net benefit of having treatment $T$ when we have the disease $M$ we write as $U(T, M)$. (We assume for simplicity that the treatment is always successful.) The utility of this treatment when we do not have the disease is $U(T, M')$; this will be a negative number due to the negative side-effects of unnecessary treatment. If we test positive – measure data $D$ – and have the treatment, then from the defintion of expectation (section 1.3) the expected utility of the treatment is

$$E[U(T)] = U(T, M)P(M \,|\, D) + U(T, M')P(M' \,|\, D). \tag{3.18}$$

Typically we would elect to have the treatment only if its expected utility is positive. To reach this decision we have to decide what the two stated utilities are; these are likely to vary from person to person depending on how they value the negative side-effects of treatment. We could apply the same analysis to the utility of not having the treatment to find $E[U(T')]$ (and would presumably then decide for $T$ or $T'$ according to which gives the more positive utility).

This approach to decision making can also be applied when our model involves a continuous parameter $\theta$, which the utility depends on, but where we don't know $\theta$ with certainty: our knowledge of it is described by the posterior PDF. To take an example, suppose we want to produce the strongest possible steel alloy by selecting one of a number of possible manufacturing processes $d_k$ ($k = 1 \ldots K$). Each of these processes results in an amount $\theta$ of a solute being added to the alloy. There is a degree of uncertainty in all of these processes: the amount of solute added is described by the PDF $P(\theta \,|\, d_k)$. If too much or too little solute is added then the alloy will be weak; the exact variation depends on the process. The strength of the resulting alloy is our utility, and as it depends on both $d_k$ and $\theta$ we write it as $U(d_k, \theta)$. For a given process $d_k$ the expected utility is

$$E[U(d_k)] = \int U(d_k, \theta)P(\theta \,|\, d_k)\, d\theta. \tag{3.19}$$

If we compute this for each $d_k$ we can then select the process that gives the highest expected utility.

The difference between inference and decision making is that in the latter we are not concerned with the parameter posterior per se, but rather with how this translates into utility for different possible courses of action.

## 3.5 Estimating one parameter from one data point

In chapters 5, 6, and 9 we will look at finding the posterior over one or more model parameters from a set of data using both exact analytic and approximate numerical methods. Here we look at a simpler problem – estimating a single parameter from a single measurement – and use it to gain some insight into inference.[9]

Distances to stars can be measured geometrically via their parallaxes. As the Earth orbits the Sun, the position of a star observed from the Earth will appear to move relative to more distant background stars (see figure 3.3). The nearer the target star, the larger this apparent movement. The size of the angular displacement is known as the *parallax*, $\varpi$. As all stars are very far away compared to the size of the Earth's orbit (150 million km),

[9] This section is based on my more detailed, yet still didactic, analysis of the problem in Bailer-Jones (2015). R code for producing the figures in this section is provided in the file `parallax.R`, available online.

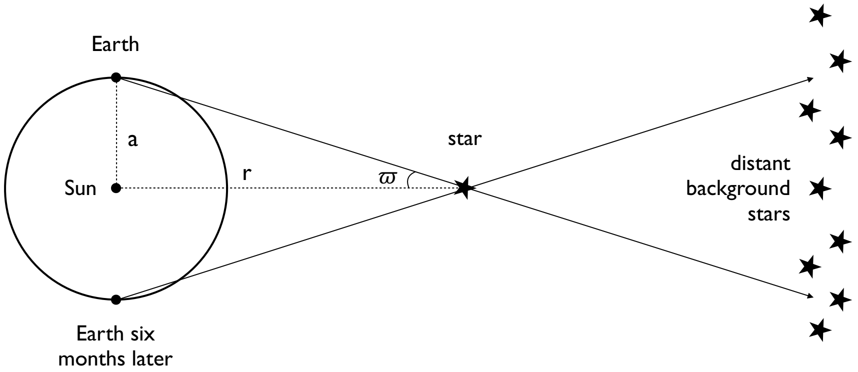

Fig. 3.3 The parallax $\varpi$ of a star is the apparent angular displacement of that star (relative to distant background stars) due to the orbit of the Earth about the Sun. More precisely, the parallax is the angle subtended by the Earth's orbital radius $a$ as seen from the star. As parallaxes are extremely small angles ($\varpi \ll 1$), $\varpi = a/r$ to a very good approximation. When $\varpi$ is 1 arcsecond, $r$ is defined as the *parsec*, which is about $3.1 \times 10^{13}$ km. In this sketch the size of the Earth's orbit has been greatly exaggerated compared to the distance to the star, and the distance to the background stars in reality is orders of magnitude larger again.

their parallaxes are very small. The largest is of order one arcsecond, which is $1/3600$ of a degree. Such a parallax would be produced by a star that is about $3.1 \times 10^{13}$ km away, an amount which is used as a distance unit in astronomy, the *parsec*. From simple geometry (and the fact that parallaxes are very small), if the distance to a star is $r$ parsec, its parallax is $1/r$ arcseconds.[10] These units are implicit in everything that follows.

Given this definition, we might think that if we measure a parallax we can just invert it to determine the distance. But this is only the case if we have no measurement errors in $\varpi$. As we always have measurement errors, determining the distance given a parallax becomes an inference problem. Here we will investigate how to do this and see why inverting a noisy parallax is not the right thing to do.

For a star at true distance $r$, its true but unknown parallax is $1/r$. The measured parallax $\varpi$ is a noisy measurement of $1/r$. Let us assume[11] that $\varpi$ has a Gaussian distribution with unknown mean $1/r$ and known standard deviation $\sigma_\varpi$. That is, we assume $\varpi$ has been drawn from the distribution

$$P(\varpi|r) = \frac{1}{\sigma_\varpi\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_\varpi^2}\left(\varpi - \frac{1}{r}\right)^2\right] \quad \text{where} \quad \sigma_\varpi > 0, \tag{3.20}$$

which is Gaussian in $\varpi$, but of course not in $r$. This is the likelihood: the probability density

[10] Some readers will realize that because the distant background stars are not infinitely far away, measuring a parallax in this way would only give a relative parallax. More sophisticated procedures involving observing stars all over the sky can be used in order to obtain absolute parallaxes.

[11] This model is actually used in practice, with the standard deviation estimated from a noise model for the instrument.

of the data $\varpi$ given the parameter $r$. Equation 3.20 has a finite probability for negative parallaxes, and this probability gets larger with increasing *fractional parallax uncertainty*, $f = \sigma_\varpi/\varpi$. Negative parallaxes arise because the angular measurements used to derive a parallax are noisy: it is possible that the measured displacement of the star is in the opposite direction of that expected from the movement of the observer along the baseline. A negative parallax does *not* correspond to a negative distance, because by definition $r \geq 0$. Negative parallaxes instead suggest that the star is probably quite distant, so the true parallax is small and noise made it negative. The measurement of a negative parallax therefore contains information.

### 3.5.1 Misleading intuition

Suppose we have a parallax measurement $\varpi \pm \sigma_\varpi$ and want to estimate the distance and its uncertainty. The intuitive approach is to report $1/\varpi$ as the distance estimate and to use a first-order Taylor expansion (section 2.8) to give $\sigma_\varpi/\varpi^2$ as the uncertainty. From the definition of the Gaussian likelihood (equation 3.20), the two-sigma intervals are

$$1/r = [\varpi - 2\sigma_\varpi, \varpi] \quad \text{and} \quad 1/r = [\varpi, \varpi + 2\sigma_\varpi], \tag{3.21}$$

each of which includes a fraction $0.954/2 = 0.477$ of the total probability of the distribution $P(\varpi\,|\,r)$. The transformation from $1/r$ to $r$ is monotonic and so preserves the (integrated) probability. Thus the intervals

$$r = [1/(\varpi - 2\sigma_\varpi), 1/\varpi] \quad \text{and} \quad r = [1/\varpi, 1/(\varpi + 2\sigma_\varpi)] \tag{3.22}$$

each also contain a fraction 0.477 of the total probability over the distance. But whereas these intervals are equal-sized in $1/r$ (the Gaussian is symmetric), they are not equal-sized in $r$. The uncertainties do not transform symmetrically because of the nonlinear transformation from $1/r$ to $r$. For example, with $\varpi = 0.1$ and $\sigma_\varpi = 0.02$, the above intervals are $r = [16.67, 10]$ and $r = [10, 7.14]$ respectively. This is shown in figure 3.4. The black line in the left panel plots the likelihood as a function of $1/r$ for this case. As the likelihood is invariant when swapping $\varpi$ and $1/r$, this plot is a Gaussian. The solid black line in the right panel shows the same likelihood, but now plotted against $r$. The asymmetry of the tranformation is clear. The first-order Taylor expansion suggested above corresponds to estimating the distribution in distance as a Gaussian with mean $1/\varpi$ and standard deviation $\sigma_\varpi/\varpi^2$. This is plotted as the black dashed line in the right panel. We see it is quite a poor approximation, even for this relatively small fractional uncertainty of $f = 1/5$.

What if the uncertainties are larger, say $f = 1/2$, so $\varpi = 0.1$ and $\sigma_\varpi = 0.05$? This likelihood is shown as the grey line in figure 3.4. The upper distance interval is now $r = [\infty, 1/\varpi]$. If $f$ is even larger then this interval becomes undefined and we seem to "lose" some of the probability. (Remember, $r$ cannot be negative.) As the Gaussian distribution has infinite support for all values of $\varpi$ and $\sigma_\varpi > 0$, some finite amount of probability in the likelihood function will always correspond to an undefined distance, which is obvious for the grey line in the left panel of figure 3.4.

The problem here is that we are trying to make a probability statment about $r$ using just

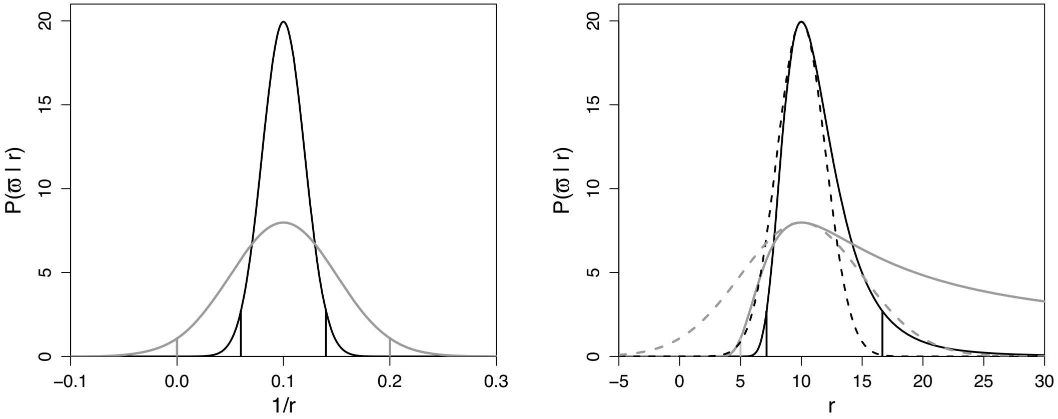

Fig. 3.4 The solid lines show the likelihood (equation 3.20) in the parallax problem for a measured parallax of $\varpi = 0.1$ as a function of $1/r$ in the left panel (a Gaussian) and as a function of $r$ in the right panel. Note that the functions in the right panel are not PDFs over $r$, so they are not normalized. The black line is for $\sigma_\varpi = 0.02$ (so $f \equiv \sigma_\varpi/\varpi = 1/5$) and the grey line is for $\sigma_\varpi = 0.05$ ($f = 1/2$). The vertical lines denote the upper and lower $2\sigma$ limits around $1/\varpi$; the upper limit for the grey curve in the right panel is at $r = \infty$. The dashed lines in the right panel correspond to a Gaussian with mean $1/\varpi$ and standard deviation $\sigma_\varpi/\varpi^2$. Each of these Gaussians has been multiplied by the ratio of its standard deviation to that of the likelihood, $(\sigma_\varpi/\varpi^2)/\sigma_\varpi = 100$, in order to put them on the same vertical scale as the likelihood.

equation 3.20, yet this is not a probability distribution over $r$. The solution is to pose the problem correctly, as an inference problem.[12]

## 3.5.2 The inference problem

We tackle inference problems by writing down Bayes' theorem to give the posterior PDF over the model parameter (here $r$) given the data (here $\varpi$). This is

$$P(r\,|\,\varpi) \;=\; \frac{1}{Z}\,P^*(r\,|\,\varpi) \;=\; \frac{1}{Z}\,P(\varpi\,|\,r)\,P(r) \tag{3.23}$$

where $P^*(r\,|\,\varpi)$ is the unnormalized posterior and $Z$ is the normalization constant. The likelihood is equation 3.20. What do we adopt as the prior? This should embody our knowledge of – or assumptions about – the distance, independent of the parallax we have measured. Some people object to priors on philosophical grounds (*How can science depend on assumptions?*), others on practical grounds (*How can I know the prior if I haven't yet mea-*

[12] We cannot solve this problem through a simple change of variables (section 1.9.1), because that would give a transformation between the true distance and the noisy parallax, i.e. it would assume the true distance is the inverse of the noisy parallax, which it is not.

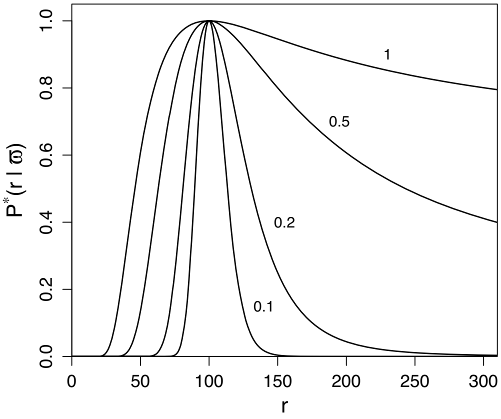

Fig. 3.5 The unnormalized posterior using the improper uniform prior (equation 3.25) for $\varpi = 1/100$ and four values of $f = 0.1, 0.2, 0.5, 1.0$. The unnormalized posteriors have been scaled to all have their mode at $P^*(r \,|\, \varpi) = 1$. Figure reproduced from Bailer-Jones (2015).

*sured any distances?*). The latter is a valid protest and will be discussed in later chapters. Yet without a prior we run into the problems we just saw in the previous section.

### 3.5.3 An improper uniform prior

A common approach is to adopt a uniform prior on the grounds that this does not prefer one value over another. This would not solve the problem, however, as we would end up with a posterior looking like the right panel of figure 3.4. The least we should do is accept physical reality and use

$$P^*(r) = \begin{cases} 1 & \text{if } \; r > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.24}$$

so as to introduce the uncontroversial assumption that distances must be positive. Because it extends to infinity, this prior cannot be normalized. Such priors are referred to as *improper*. In this case the (unnormalized) posterior $P^*(r \,|\, \varpi)$ is just the likelihood but now considered as a function of $r$ rather than $\varpi$, and subject to the additional constraint $r \geq 0$, i.e.

$$P^*(r \,|\, \varpi) = \begin{cases} P(\varpi \,|\, r) & \text{if } \; r > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.25}$$

Examples of this posterior are shown in figure 3.5 for $\varpi = 1/100$ and various values of $f$. Notice how the posterior gets broader for larger uncertainties in the data. It also becomes increasingly skew, as discussed in section 3.5.1. Conversely, the smaller the uncertainty, the more concentrated the posterior becomes around the value predicted purely by the data.

Inspection of equation 3.25/3.20 shows that

$$\lim_{r\to\infty} P^*(r\,|\,\varpi) \;=\; \text{constant}. \tag{3.26}$$

As the posterior does not converge to zero it has an infinite area and so cannot be normalized: it is improper. Consequently it has no mean, no standard deviation, no median, and no quantiles. The only plausible estimator of the distance is the mode of the posterior, which we see from figure 3.5 is well-defined for all values of $f$, and is equal to $1/\varpi$. Yet this estimator is invalid for non-positive parallaxes,[13] and it ignores the measurement uncertainty. Moreover, numerical experiments show that once $f$ grows above about 0.2, this prior for distance estimation gives very large errors in terms of both bias and variance (terms which are defined in a modelling context in section 4.8). Improper posteriors are a bad thing.

### 3.5.4 A proper uniform prior

An obvious improvement on the above is to truncate the prior at some value. The prior, which can then be normalized, is

$$P(r) \;=\; \begin{cases} \dfrac{1}{r_{\rm lim}} & \text{if } \; 0 < r \le r_{\rm lim} \\ 0 & \text{otherwise} \end{cases} \tag{3.27}$$

where $r_{\rm lim}$ is the largest distance we expect for any star in our survey. The unnormalized posterior is the same as in figure 3.5 but set to zero for $r > r_{\rm lim}$, so

$$P^*(r\,|\,\varpi) \;=\; \begin{cases} P(\varpi\,|\,r) & \text{if } \; 0 < r \le r_{\rm lim} \\ 0 & \text{otherwise.} \end{cases} \tag{3.28}$$

This is shown for different values of $f$ in the left panel of figure 3.6. The normalization can easily be done by integrating on a dense grid, as explained in section 3.3.4. The normalized posteriors are shown in the right panel of figure 3.6 (and make quite a different impression). The mean, standard deviation, and quantiles are all defined, and not just the mode. If we nonetheless use the mode as our distance estimator it is

$$r_{\rm mode} \;=\; \begin{cases} \dfrac{1}{\varpi} & \text{if } \; 0 < \dfrac{1}{\varpi} \le r_{\rm lim} \\ r_{\rm lim} & \text{if } \; \dfrac{1}{\varpi} > r_{\rm lim} \\ r_{\rm lim} & \text{if } \; \varpi \le 0. \end{cases} \tag{3.29}$$

Another consequence of using a proper, normalized prior is that negative parallaxes now also produce plausible posterior PDFs (the dashed line in the plot), although this prior has the undesirable property of giving them all the same mode ($r = r_{\rm lim}$). This is nonetheless

[13] The maximum is at infinity. You can get an idea of what the posterior looks like for negative parallaxes by plotting, e.g.

```
r <- seq(from=-10, to=10, by=0.01)
plot(r, exp(-0.5*(-1-1/r)^2), type="l")
```

although at negative values of $r$ the posterior is set to zero by the prior.

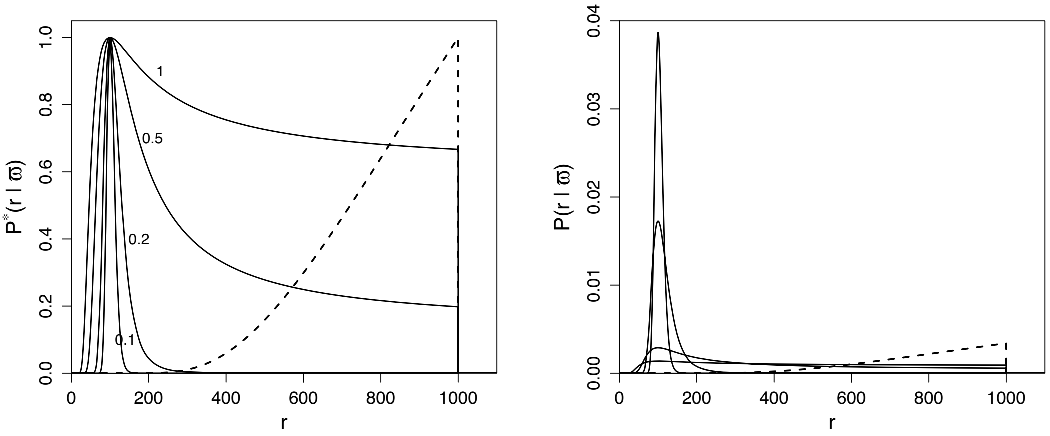

Fig. 3.6 Left: the unnormalized posterior using the truncated uniform prior with $r_{\mathrm{lim}} = 1000$ (equation 3.28). The four solid lines are for $\varpi = 1/100$ and $f = 0.1, 0.2, 0.5, 1.0$. These unnormalized posteriors have been scaled to all have their mode at $P^*(r \,|\, \varpi) = 1$. The dashed line is for a negative parallax, $\varpi = -1/100$, and $|f| = 0.25$. The right panel shows the same posteriors, but now normalized. Figure reproduced from Bailer-Jones (2015).

an improvement over inferring arbitrarily large distances for arbitrarily small, but noisy, parallaxes.

### 3.5.5 What is a good prior?

The uniform distance prior, even when truncated, may seem attractive because it does not prefer one distance over the other. It appears to be “uninformative”. But there is nothing fundamental about a prior that is uniform in $r$. Why not uniform in, say, $\log r$? Is this less “fundamental”? Worse, a prior that is uniform in $r$ corresponds to assuming that the volume density of stars in space varies as $P(V) \propto 1/r^2$, as seen from the observer.[14] Not only does this assign the observer a privileged position in the Galaxy (because the density falls off in all directions from us), but it assumes that the distribution of stars in the Galaxy is scale independent, which it demonstrably is not: the Galaxy has a characteristic length scale. This apparently harmless uniform prior is actually making strong assumptions that violate our astrophysical knowledge. There is in fact no such thing as an “uninformative” prior; just different priors that convey different information.

In the parallax example an improved prior might be

$$P(r) = \begin{cases} \dfrac{3}{r_{\mathrm{lim}}^3} r^2 & \text{if } 0 < r \le r_{\mathrm{lim}} \\ 0 & \text{otherwise} \end{cases} \tag{3.30}$$

[14] $V \propto r^3$, so using $P(V)dV = P(r)dr$ to make a change of variables, we get $P(V) \propto P(r)/r^2$.

which corresponds to a uniform volume density of stars out to distance $r_{\rm lim}$. We could replace the hard cut-off with an exponential decrease in the volume density, $P(V) \propto \exp(-r/L)$, for which the prior is

$$P(r) = \begin{cases} \dfrac{1}{2L^3} r^2 e^{-r/L} & \text{if } \; r > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.31}$$

where $L > 0$ is a length scale which could be chosen based on current knowledge of the size of the Galaxy.[15]

Priors that are physically more acceptable necessarily make more physical assumptions. This may run into objections from those who claim that priors should play no role in inference; yet priors are all about introducing relevant contraints into the inference. By ignoring priors and just selecting the maximum of the likelihood (which here is $1/\varpi$), we can get nonsensical answers, and very poor estimates if the uncertainties are large. One may be tempted to circumvent this problem by discarding all data with large uncertainties, but that may result in discarding much of the hard-won data. Worse, it will bias scientific results on a population of objects, because by discarding poorer data we preferentially discard the more distant and fainter stars.

Opinions vary considerably on how strong priors should be and on what additional information we should use in our inference.[16] In the distance example one could make a case for using the best model we currently have of the distribution of stars in the Galaxy. But others will say that this influences the results too much, and we should do something more minimal. What one person calls a bias, another may call relevant information. Often priors cannot be avoided, even in principle (e.g. section 3.2.2), and inference is impossible without some assumptions. I generally advocate using a minimal prior consistent with the physical constraints, but even this will rarely result in a unique specification. A good approach is to try out different plausible minimal priors, and if these give very different results for the data available, this is just an indication that the data are not very informative. Your data are poor, so your results unavoidably depend more strongly on the prior. If you can't live with this uncertainty, you'll need to get better data.

We'll discuss further the issue of assigning priors, in particular in section 5.3.

## 3.6 An inference story

One of the key messages of this chapter is that inference needs to take into account all relevant information. This may be more than meets the eye, quite literally, as the following story demonstrates.

During World War Two, aircraft undertook bombing raids that were sometimes fatal for

[15] I will use this prior and this problem in section 12.7 to illustrate numerical optimization and bootstrap estimates.

[16] Readers with an astronomical background will recognise that we often have colour or even spectral information on stars. This can give fractional distance uncertainties as small as of order 10%, depending on the type of star. But the issues discussed still apply.

the aircrews. If the planes could be better armoured they would be more likely to survive. But armour is heavy, so it wasn't possible to put more armour everywhere. It was observed that while the returning planes often had damage on the wings and fuselage, they hardly ever had any damage on the cockpit or engines. It was concluded that the planes should receive more armour in these damaged areas in order to decrease aircraft losses. But assuming that planes could be hit anywhere with equal probability, why wasn't damage seen in certain areas of the returning planes? The answer is in the question: the observations are made of the *returning* planes. Presumably those planes that were hit on the cockpit or fuselage rarely returned. These are therefore the more vulnerable areas of the aircraft. Thus the conclusion based on all the data – i.e. also on what damage is *not* seen – is the opposite of the initial conclusion: more armour is required where damage is not seen.[17]

[17] This observation is part of a more detailed analysis undertaken by Abraham Wald (1943).