11 Model comparison

We used Bayesian model comparison in section 3.2 to interpret the positive/negative result of a medical test. In the present chapter we take this approach a step further and use it to compare parametrized models via the marginal likelihood (also called the evidence). I illustrate this using the coin tossing problem (binomial likelihood) and the line fitting problem (Gaussian likelihood). I will make some comparisons with frequentist hypothesis testing, and we will learn why maximum likelihood (and χ^2) may not be used for model selection. I conclude by highlighting some drawbacks of the Bayesian evidence, and will mention some other metrics that can be used to compare models.

11.1 Bayesian model comparison

Imagine we have a data set D and a model M that we think might describe the data. How good is the model? An interpretation of this question is: how probable is the model in light of the data, $P(M\,|\,D)$? Take the example of drawing playing cards from a deck. The chance of picking out the ace of spades (data D) from a normal deck of 52 cards (model M) is $P(D\,|\,M) = 1/52$. This information alone tells us nothing about whether the deck of cards is normal – it does not tell us $P(M\,|\,D)$ – because if M is true, each and every card has the same probability of being drawn, namely $1/52$. Our information doesn't change fundamentally even if we drew the same card k times (with replacement and shuffling), because every sequence of k cards has the same probability. Yet I am sure you would get suspicious if you drew the ace of spades two or three times in a row. Why? The reason is that you are implicitly thinking of another model – namely an abnormal deck with many aces of spades – under which the data you just got are more likely than under the normal deck model. Without considering the option of such an alternative model you cannot cast doubt on your multiple draws of the ace of spades, because it's as likely as any other sequence under model M. Without an alternative model, the data *must* have come from the one and only model you permit to exist, no matter how unlikely these data are.

We learned back in section 3.2 that in order to find $P(M\,|\,D)$ we need to know not only $P(D\,|\,M)$ but also $P(D\,|\,M')$ and $P(M)$ (see, for example, equation 3.3). This was illustrated in the breast cancer example, which showed that even though $P(D\,|\,M)$ can be very large – even equal to 1.0 – $P(M\,|\,D)$ can still be very small. This is because $P(M\,|\,D)$ depends also on the prior probability of the alternative models for the data. If you draw the ace of spades several times and are suspicious, it is not *only* because of the data. It's also because your prior probability for the abnormal deck, $P(M')$, is not extremely small. If it were vanishingly small, or even zero, then no amount of data could convince you the deck was abnormal.

In the breast cancer example we could calculate $P(M \,|\, D)$ because we had only two models which were complementary, i.e. $P(M) + P(M') = 1$. In many situations we will have more than two models, and often they will not be an exhaustive set. In that case we cannot determine $P(M \,|\, D)$ (as was seen in section 3.2.1). But we saw that we could determine the ratio $P(M_1 \,|\, D)/P(M_2 \,|\, D)$ for any two models M_1 and M_2 to get the *posterior odds ratio*

$$R = \frac{P(D \,|\, M_1) P(M_1)}{P(D \,|\, M_2) P(M_2)}. \tag{11.1}$$

The term $P(D \,|\, M)$ is called the *evidence* for model M. It plays a key role in model comparison. If we can't decide a priori between the two models, then we set $P(M_1) = P(M_2)$. These cancel to leave the ratio of the evidences, which is called the *Bayes factor*

$$BF_{12} = \frac{P(D \,|\, M_1)}{P(D \,|\, M_2)} \tag{11.2}$$

of model 1 to 2. Note that $BF_{21} = BF_{12}^{-1}$. The usual way of doing Bayesian model comparison in practice is with Bayes factors, rather than determining the posterior model probabilities.

How much the odds ratio or Bayes factor must differ from unity before you call one of the models significantly better than the other is a personal choice. The statistics only does the calculations; the decisions you have to make yourself (but see section 3.4). Typically one would not make any claim about one model over another until the factor exceeds 10 or is less than 0.1. Anything within this range is normally considered as indistinguishable. Conservatively one only starts to make claims of real significance if the factor (or its inverse) is more extreme, say 50 or even more.[1]

There are two key aspects to Bayesian model assessment. First, it considers the probability of the data *conditional* on the model, $P(D \,|\, M)$. Since we've now obtained the data D, their probability per se is not relevant. The issue is rather which model is more likely to have produced them. The second key aspect is that it uses the *ratio* of probabilities (the posterior odds ratio or the Bayes factor). As we have seen in the deck of cards example, the individual, absolute values of the evidence are irrelevant. That's because the probability of any particular data set is usually very small, even under the true model (see section 5.2). This aspect makes Bayesian model assessment fundamentally different from frequentist hypothesis testing. As explained in section 10.1, because the frequentist approach only looks at the probability of one hypothesis, and because the probability of any particular data set is vanishingly small (zero, in fact, for real numbers), that approach must resort to the problematic concept of "more extreme data". The ratio of probabilities (or probability densities), in contrast, is well defined without having to resort to imaginary data.

[1] If you are not interested in the models themselves, but only in predictions from them, then you can avoid choosing a model by using instead a posterior-weighted combination of models.

11.1.1 The evidence as a marginal likelihood

The model M may contain one or more parameters θ. We have seen how, using the likelihood and prior, we can infer the posterior $P(\theta \mid D, M)$ for these parameters using Bayes equation,

$$P(\theta \mid D, M) = \frac{P(D \mid \theta, M) P(\theta \mid M)}{P(D \mid M)}. \tag{11.3}$$

The denominator, which we previously treated as just a normalization constant for the posterior, is the evidence. It follows from the rules of probability that

$$P(D \mid M) = \int \underbrace{P(D \mid \theta, M)}_{\text{likelihood}} \underbrace{P(\theta \mid M)}_{\text{prior}} \, d\theta. \tag{11.4}$$

The evidence is the integral of the likelihood over the prior, and for this reason is often called the *marginal likelihood*. It tells us how probable the data are under the model, independent of specific values of θ. Consider two models for a continuous-valued data set: one is a quadratic, the other is a sinusoid. The task of model comparison is to decide which model explains the data best, but without reference to a specific set of fitted parameters, because a specific set is not representative of the model as a whole. In particular, we should not first fit the models parameters by, say, maximum likelihood, and then compare the resulting fitted models. This is because no matter how complex or large the data set, I can always find a sinusoidal model that will fit every point exactly (I just need to make the frequency large enough). So it will always have the higher likelihood. More generally, by comparing maximum likelihood solutions we invariably favour the more flexible (more complex) model. We should instead find the model that gives the largest *average* likelihood, where the averaging is done over the prior. This is the evidence in equation 11.4. It can be thought of as an average of the predictive ability (likelihood) of the model over the plausible values of the model parameters (prior).

Note that when computing the evidence we must use a normalized likelihood, because its normalization constant is, in general, a function of the parameters θ. The prior must be normalized too, because the evidence is an integral over it. If it were not normalized then it would involve an unknown multiplicative constant a, i.e. $P(\theta \mid M) = aP^*(\theta \mid M)$, so the evidence would also involve this unknown constant. This would also not cancel in the Bayes factor, because the other model would have a different prior.

11.2 Example of an analytic evidence calculation: is a coin fair?

In section 5.1 we investigated whether a coin is fair ($p = 1/2$) or not ($p \neq 1/2$) by inferring the posterior PDF $P(p \mid n, r, \alpha, \beta)$, where n is the number of coin tosses, r the number of heads, and α and β are the parameters of the beta prior. A distribution peaked near $1/2$ suggests the coin is fair. In contrast, a PDF with essentially zero probability around this (e.g. the $r = 18$ panel of figure 5.2) would suggest the coin is not fair. Yet even armed with the posterior, the probability of p being *exactly* $1/2$ (or indeed any other number) is zero (see equation 10.1), so we cannot use the posterior PDF alone to say something like "the probability of this coin being fair is x". This is a bit of a false dichotomy though, as we are unlikely to be interested in whether a coin has exactly $p = 1/2$. Design biases aside, the reality of manufacturing means that no two coins will be *identical*. In practice we would probably be more interested in asking whether p lies within some narrow region, say 0.495–0.505, which for all practical purposes is essentially the same as $p = 0.5$. We could then integrate the normalized posterior PDF over this range to give x and quote this as the probability that the coin is more or less fair. The problem here is that we have made an additional assumption of what "essentially the same" is. Furthermore, for a given posterior, the narrower we choose to make the range, the smaller the probability.

A better way to answer this question is to use the evidence to compare two models. Here I will consider the following two models:
M_1: $p = \phi$, which corresponds to a prior $P(p\,|\,M_1) = \delta(p - \phi)$;
M_2: p is unknown, described by a beta distribution prior.
The likelihood is the binomial distribution

$$P_{\text{bin}}(r\,|\,p,n) = \frac{n!}{r!(n-r)!}\,p^r(1-p)^{n-r}. \tag{11.5}$$

The prior for M_2 is

$$P(p\,|\,M_2) = \frac{1}{B(\alpha,\beta)}\,p^{\alpha-1}(1-p)^{\beta-1} \quad \text{for} \quad \alpha > 0,\ \beta > 0, \tag{11.6}$$

where

$$B(\alpha,\beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\,dp \tag{11.7}$$

is the beta function which, when α and β are integers, is

$$B(\alpha,\beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}. \tag{11.8}$$

The evidence (equation 11.4) for either one of the models M is

$$P(D\,|\,M) = P(r\,|\,n,M) = \int_0^1 P_{\text{bin}}(r\,|\,p,n)P(p\,|\,M)\,dp. \tag{11.9}$$

For M_1 the prior is a delta function, so the evidence is simply

$$P(D\,|\,M_1) = \frac{n!}{r!(n-r)!}\phi^r(1-\phi)^{n-r} = P_{\text{bin}}(r\,|\,\phi,n), \tag{11.10}$$

i.e. a binomial distribution with $p = \phi$. The evidence for M_2 is

$$\begin{aligned} P(D\,|\,M_2) &= \frac{n!}{r!(n-r)!}\frac{1}{B(\alpha,\beta)}\int_0^1 p^{r+\alpha-1}(1-p)^{n-r+\beta-1}\,dp \\ &= \frac{n!}{r!(n-r)!}\frac{B(r+\alpha,n-r+\beta)}{B(\alpha,\beta)} \end{aligned} \tag{11.11}$$

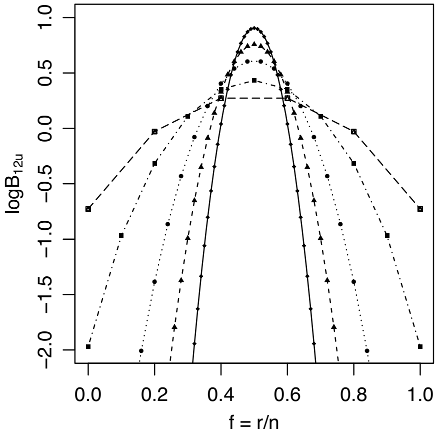

Fig. 11.1 The logarithm (base 10) of the Bayes factor for two models of the coin as a function of the ratio of heads r to total coin tosses n. M_1 assumes $p = 0.5$. M_{2u} assigns a uniform prior to p. The fives lines correspond to $n = 5, 10, 25, 50$, and 100 in order of increasingly narrow distribution.

which follows from the definition of the beta function (equation 11.7).[2] In both cases the evidence is the predictive distribution of the data – the number of heads, r – given the number of tosses and the model (which also encapsulates the corresponding prior). The Bayes factor for these two models is

$$B_{12} \equiv \frac{P(D\,|\,M_1)}{P(D\,|\,M_2)} = \frac{B(\alpha,\beta)}{B(r+\alpha, n-r+\beta)}\,\phi^r(1-\phi)^{n-r}. \tag{11.12}$$

An interesting case is $\alpha = \beta = 1$, the uniform prior, which I'll call M_{2u}. In that case (using equation 11.8)

$$P(D\,|\,M_{2u}) = \frac{n!}{r!(n-r)!}\,\frac{r!(n-r)!}{(n+1)!} = \frac{1}{n+1}\,, \tag{11.13}$$

$$B_{12u} = (n+1)P_{\text{bin}}(r\,|\,\phi, n). \tag{11.14}$$

We may be surprised that the evidence for model M_{2u} is $(n+1)^{-1}$: it does not depend on r at all. It just decreases with an increasing number of coin tosses, regardless of their result. This actually makes sense. The more data we collect, the better we are able to determine the value of p. So the evidence for a model that has *no* preference for any p must decrease as we gather more data. And because it has no preference, it does not actually matter what those data are. This does not imply that the Bayes factor will always favour M_1 as we get more and more data. The evidence for M_1, and therefore the Bayes factor (B_{12u}), does depend on r.

Figure 11.1 plots B_{12u} as a function of the fraction of heads tossed $f = r/n$ for the

[2] This integral of the product of the beta and the binomial over all p gives a PDF in r, and is called the beta-binomial distribution.

case that M_1 represents a fair coin, $\phi = 1/2$. The plot shows this for various sample sizes n. The curves are of course symmetric about $f = 1/2$, because getting r heads has the same probability as getting $n - r$ heads when $p = 1/2$. For $n = 5$ (the broadest curve, with the six values $r = 0 \ldots 5$) the fair coin is only favoured when we get two ($f = 2/5$) or three ($f = 3/5$) heads, and then only marginally: the Bayes factor is just 1.9 in both cases. But even getting five or no heads does not weigh strongly against the coin being fair, as the Bayes factor is 0.19, i.e. M_{2u} is only favoured by a factor of about five. Increasing the amount of data (number of tosses) generally yields a more discriminative Bayes factor, which we see as a narrowing of the curve for increasing n. For example with $n = 50$, 30 heads ($f = 0.6$) gives a Bayes factor of 2.1, but 40 heads ($f = 0.8$) gives a Bayes factor of $1/2150$, which hugely favours M_{2u}. If we instead fix f to 0.4 and then increase the sample size n such that $r/n = 4/10$, $10/25$, $20/50$, and $40/100$, we get Bayes factors of 2.26, 2.53, 2.13, and 1.10 respectively, none of which are extreme enough to discriminate between the two models.[3] If we made $n = 100$ tosses we would need to get as many as 65 heads (or likewise as few as 35 heads) before the Bayes factor drops below 0.1 so that we can claim significant evidence in favour of M_{2u} over M_1. But even with exactly 50 heads in 100 tosses the Bayes factor is only 8.0, so 100 tosses is not enough to ever be confident that the coin is fair (with M_{2u} as the alternative). At least 156 tosses with exactly half heads is required before the Bayes factor exceeds 10.

Following this analysis you might be surprised at how extreme the data have to be in order to favour a model, and you may conclude from this that the evidence approach is conservative. But what this analysis really shows is that the M_{2u} prior is conservative: we are permitting all possible values of p to be equally likely a priori. That's quite an extreme state of ignorance (for a coin picked up at random), so you need a lot of data to overcome it. Values of r/n nearer to 1 or 0 give stronger evidence against M_1 and in favour of M_{2u} (and the Bayes factor varies more rapidly), because the further the data are from $r/n = 0.5$, the less plausible M_1 becomes.

The following code makes figure 11.1. If you want to investigate how the Bayes factor varies when we adopt a more conservative beta prior for model M_2, replace the line defining `BF` with equation 11.12 and insert your chosen values of α and β for the prior.

R file: `coin_evidence.R`

```
##### Calculate the evidence for the coin problem

phi <- 1/2
n   <- c(5,10,25,50,100)
pdf("coin_evidence.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(c(0,1), c(0,1), type="n", xlim=c(0,1), ylim=c(-2,1), xlab="f = r/n",
     ylab=expression(paste(log, B[12][u])))
for(i in 1:length(n)) {
  r  <- 0:n[i]
```

[3] The Bayes factor does not vary monotonically with sample size in this case: 4 out of 10 is (slightly) more discriminative than 40 out of 100 for these particular models (although if we compute further we find that 400 out of 1000 is hugely discriminative, with $B_{12u} = 4.6\times10^{-8}$). For more extreme results, e.g. $f = r/n = 0$, we do get a monotonic variation of the Bayes factor with n, as can be seen in figure 11.1.

```
  f  <- r/n[i]
  BF <- (n[i]+1)*dbinom(x=r, size=n[i], prob=phi)
  points(f, log10(BF), cex=0.5, pch=13+i)
  lines(f, log10(BF), lty=6-i)
}
dev.off()
```

11.3 Example of a numerical evidence calculation: is there evidence for a non-zero gradient?

In the previous section we were able to compute the evidences and therefore the Bayes factor analytically. Here we turn to a problem where we must instead use numerical (Monte Carlo) sampling.

Given a two-dimensional set of points $\{x_i, y_i\}$ with noise only on the y values (the standard deviation of which may or may not be known), we ask whether there is evidence for a linear correlation. We can frame this by asking "is there a significant probability for a non-zero gradient of a fitted line?". We could calculate the posterior probability density over the gradient, as we did in section 9.1. But the probability of any *exact* gradient is zero, and integrating over some narrow range of probability is also problematic (as discussed in section 11.2). A better approach is to ask "what is the Bayes factor $B_{12} = P(D\,|\,M_1)/P(D\,|\,M_2)$, where M_1 is a model with a zero gradient and M_2 is a model with an unknown gradient?". We would have to calculate Bayes factors if we were comparing models that did not share common parameters, such as a polynomial and a sinusoid. The case we are looking at here is actually one of *nested models*, because M_1 is a special case of M_2. In such cases the Bayes factor can be approximated using something called the Savage–Dickey density ratio (e.g. Trotta, 2007). But I will instead compute the marginal likelihood for both models as a means of illustrating the general approach.

As in section 9.1 the generative model is $f(x)$ and we assume an independent Gaussian likelihood for each data point y_i. Writing the prior as $P(\theta\,|\,M)$, the marginal likelihood (equation 11.4) is then

$$P(D\,|\,M) = \int \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{[y_i - f(x_i)]^2}{2\sigma^2}\right] P(\theta\,|\,M)\, d\theta \tag{11.15}$$

where θ includes all the parameters in $f(x)$ as well as σ (which I assume to be the same for all data points). For the models and priors we consider, this integral cannot be performed analytically, so we use a Monte Carlo approximation. As explained in section 8.3 (in particular point 3 at the end of that section), the marginal likelihood can be estimated as

$$P(D\,|\,M) \simeq \frac{1}{N_s} \sum_{l=1}^{l=N_s} P(D\,|\,\theta_l, M) \tag{11.16}$$

where the samples $\{\theta_l\}$ have been drawn from the prior. That is, the evidence is the average of the likelihood over a set of samples drawn from the prior.

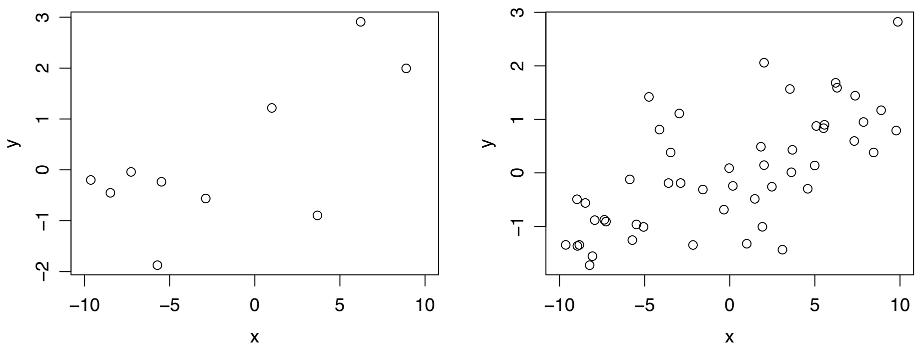

Fig. 11.2 Data drawn from a straight line model $y = b_0 + (\tan\alpha)x$ with $b_0 = 0$, $\tan\alpha = 0.1$, to which zero mean Gaussian noise with standard deviation $\sigma = 1.0$ have been added. Left: 10 data points. Right: 50 data points.

In section 9.1 we used an MCMC method to sample the posterior over the three parameters of a straight line model: the intercept b_0; the gradient b_1 (or rather the angle $\alpha = \arctan(b_1)$); and the uncertainty σ on the y values. To calculate the evidence we must sample from the prior rather than from the posterior. In order to sample from a distribution it must be proper (normalizable).[4] This is the case for the priors we used on b_0 and α, because they were a Gaussian and a truncated uniform distribution respectively. But our prior for σ was improper; it was the Jeffreys prior with $P(\sigma) \propto 1/\sigma$. To make this a proper distribution I here set it to zero outside some finite range; see equation 5.21.

In contrast to the posterior, the prior is generally a simple function of the model parameters, and so is much easier to draw from. In the present case we have uniform distributions for α and $\log\sigma$, and Gaussian for b_0. Standard methods exist for sampling these; we do not need to use MCMC.

Let us use the evidence to compare two models:

M_2 is a general straight line, $y = b_0 + x\tan\alpha + \epsilon$, where b_0 and α are unknown parameters, and $\epsilon \sim \mathcal{N}(0, \sigma)$, which is the Gaussian random noise used to define the likelihood;

M_1 is the same as M_2, but with zero gradient, $\tan\alpha = 0$.

For the purpose of this demonstration I generate data at random. Ten x values are drawn from $\mathcal{U}(-10, 10)$, and the y value at each is computed from a straight line model with $b_0 = 0$ and $\tan\alpha = 0.1$, to which zero mean Gaussian noise with $\sigma = 1$ is added. M_2 is in principle a better description than M_1 for these data, but which model is favoured depends on the actual data; the noise is large so the data could, just by chance, favour M_1. The data are shown in the left panel of figure 11.2. Given only these data (but neither the model nor the noise level) I don't think it's at all obvious that a sloping line fits better than a horizontal one.[5]

[4] When we sampled the posterior in section 9.1 we only needed to compute the prior, not sample from it, so the fact that it was improper was not a problem. The likelihood ensured that the posterior was a proper distribution.

[5] If I was then told that the noise level was 5, which is larger than the standard deviation in the data, I would go for the horizontal line on the basis that it's simpler.

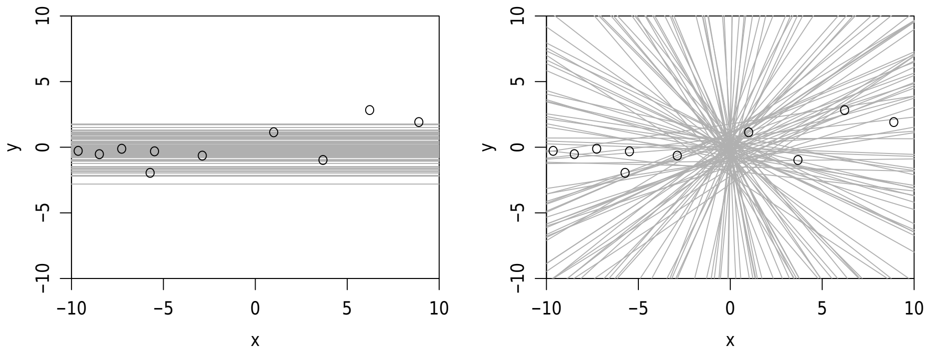

Fig. 11.3 Draws from the model priors shown as functions. Left: draws from M_1, which has $b_0 \sim \mathcal{N}(0,1)$ and $\alpha = 0$. Right: draws from M_2, which has $b_0 \sim \mathcal{N}(0,1)$ and $\alpha \sim \mathcal{U}(0, 2\pi)$. The same scales have been used on both axes and in both panels. The data (open circles) in both panels are those in the left panel of figure 11.2.

The R code to generate the data, sample the prior, calculate the evidence, and produce the plots is given in the following subsection. My parameter priors are

$$P(b_0) = \mathcal{N}(0,1) \tag{11.17}$$

$$P(\alpha) = \mathcal{U}(0, 2\pi) \tag{11.18}$$

$$P(\log \sigma) = \mathcal{U}(\log 0.5, \log 2). \tag{11.19}$$

By sampling the prior we are sampling over the prior set of possible functions, $b_0 + x \tan \alpha$, as well as the noise. It is instructive to plot what functions drawn from these priors look like. The right panel of figure 11.3 shows a random set of 100 functions drawn from the prior for M_2 (the values of $\log \sigma$ are not shown). I have plotted with equal axis scales so we can see the isotropy of the prior. Draws from the prior for M_1 are shown in the left panel.

Having drawn a large number (10^5) of parameter sets from the priors for the two models, we calculate the likelihood for each set using the data, and average these to estimate the evidence for that model (equation 11.16). The results are

$$\log P(D \,|\, M_1) = -8.33 \tag{11.20}$$

$$\log P(D \,|\, M_2) = -8.44 \tag{11.21}$$

$$\log B_{12} = 0.11 \tag{11.22}$$

$$B_{12} = 1.30. \tag{11.23}$$

This tells us that the data are 1.3 times more likely to come from M_1 than M_2. Assuming we have no prior reason to favour one model of the other, we adopt equal model priors $P(M_1) = P(M_2)$, in which case this result means we favour M_1 over M_2 by a factor of 1.3. This is extremely marginal: normally we want the Bayes factor to be larger than 10 (or less than 0.1) in order to claim significant support for one model over the other. Repeating this calculation with a different set of data can give quite different results however, because the true model has a gradient quite close to zero, and the noise is large: five other random sets of ten points give $B_{12} = 1.55, 8.29, 6.73, 0.88$, and 0.37. None of these would allow us to draw a significant conclusion, however.

This result is, of course, dependent on the choice of priors. In fact, the priors should be viewed as an intrinsic part of the model, because two very different choices creates two models with different capabilities. For example, by putting a very narrow prior on α around 0 for model M_2, we essentially end up with M_1. But these extremes aside, it is important to investigate how sensitive our results are to the choice of prior. Let's do this for the data in the left panel of figure 11.2. We first change the standard deviation of the prior on the intercept b_0 in the same way for both M_1 and M_2. The resulting Bayes factor is quite insensitive to this change. For example, decreasing the standard deviation by a factor of ten to 0.1 gives $B_{12} = 2.14$, and increasing it to 10 gives $B_{12} = 1.36$, both of which are insignificant. Other data sets could lead to different conclusions about the prior sensitivity, of course. Allowing σ to have much larger values, by increasing the upper limit on its prior, also has negligible effect on the Bayes factor. But forcing σ to be small does have a significant impact: if the lower and upper limits on $\log \sigma$ are set to 0.1 and 0.5 respectively, we get $B_{12} = 1.6 \times 10^{-6}$. This is because by forcing there to be little noise in the fit, a non-zero gradient can fit the data much better (if still poorly in an absolute sense).

With more data we should be able to distinguish better between competing models (unless neither model has anything to do with the data). Returning to the default values of the priors, I recalculate the evidence and Bayes factors but now using a particular set of 50 randomly drawn data points (shown in the right panel of figure 11.2). The results are

$$\log P(D \,|\, M_1) = -33.87 \tag{11.24}$$

$$\log P(D \,|\, M_2) = -29.37 \tag{11.25}$$

$$\log B_{12} = -4.50 \tag{11.26}$$

$$B_{12} = 3.15 \times 10^{-5}. \tag{11.27}$$

Now M_2 is favoured over M_1 – which is what we would expect – by a factor of about 31 000. Five other random sets of 50 points give $\log B_{12} = -1.66$, -5.95, -3.56, -0.27, and -2.11. The model I am generating the data from is very noisy, so the results are still highly variable. But we nonetheless see that the more data we have, the more stable the result is in the sense that we now tend to favour M_2 by a significant amount (except in one case here).[6] Remember that the Bayes factor (and Bayesian analysis in general) gives a result about the data we actually have, and not about some other data we might have got had we run the experiment again. The fact that there is variance in the data is accommodated by the likelihood. This variance (uncertainty) is incorporated into the width of the posterior, or into getting non-extreme Bayes factors. So while Bayes does not need to consider alternative data sets, it does not ignore the variance in the data.

In this example I have assumed that while the standard deviation of the noise σ is unknown, it is the same for all data points. We could relax this assumption and specify a separate noise parameter σ_i for each data point i. The approach remains the same: we specify a prior for each noise parameter (we could use the same prior for all) and marginal-

[6] I refrain from referring to M_2 as the "true" model because M_2 has an unspecified gradient, whereas the data are drawn from a model with $\tan \alpha = 0.1$.

ize over this larger set of parameters. This gives more parameters than data points, but we saw in section 9.1 that this presents no theoretical difficulties.

11.3.1 R code

The following R script produces everything discussed in the previous section. Most of it is concerned with producing the data and prior plots; the calculations themselves only take up a few lines. The code should be self-explanatory. You should experiment by changing the sampled data (change the seed), changing the amount of data, and changing their standard deviation. Try also changing the priors. Think carefully about what you are doing, and only change one thing at a time. Remember that the model uses only the measured, noisy data: the true model is actually irrelevant.

Be warned that this code can return log evidences of minus infinity! This occurs when all the likelihoods for a model are so small that they are numerically identical to zero, on account of the finite precision of a computer (see section 6.3.3). The evidence is then also zero. This will occur if the data are utterly implausible under one model or the other. Of course, if both models give numerically zero evidence (possible if the priors are chosen poorly for the overall properties of the data), then we still won't know which is better.

R file: `linearmodel_evidence.R`

```
##### Calculate the Bayesian evidence for two linear models of 2D data.
##### M1: alpha=0, M2: P(alpha) ~ 1.

library(gplots) # for plotCI

### Define likelihood

# Return log10(likelihood), a scalar.
# theta is the vector of model parameters, here c(b_0, alpha, log10(ysig)).
# data is the two-column matrix [x,y].
# dnorm(..., log=TRUE) returns log base e, so multiply by 1/ln(10) = 0.434
# to get log base 10
log.like <- function(theta, data) {
  # convert alpha to b_1 and log10(ysig) to ysig
  theta[2] <- tan(theta[2])
  theta[3] <- 10^theta[3]
  # likelihood
  modPred <- drop( theta[1:2] %*% t(cbind(1,data$x)) )
  # Dimensions in mixed vector/matrix products: [Ndat] = [P] %*% [P x Ndat]
  logLike <- (1/log(10))*sum( dnorm(modPred - data$y, mean=0,
                                    sd=theta[3], log=TRUE) )
  return(logLike)
}

### Define true model and simulate experimental data from it

set.seed(75) # 75 gives data and plots in script
Ndat <- 10
x <- sort(runif(Ndat, -10, 10))
sigTrue <- 1
modMat <- c(0,0.1) # 1 x P vector: coefficients, b_p, of sum_{p=0} b_p*x^p
```

```
y <- cbind(1,x) %*% as.matrix(modMat) + rnorm(Ndat, 0, sigTrue) # noisy data
# Dimensions in matrix multiplication:
# [Ndat x 1] = [Ndat x P] %*% [P x 1] + [Ndat]
# cbind does logical thing combining scalar and vector; then vector addition
y <- drop(y) # converts into a vector
pdf("linearmodel_evidence_data_10.pdf", width=5, height=4)
par(mfrow=c(1,1), mar=c(3.5,3.0,0.5,0.5), oma=0.5*c(1,1,1,1),
    mgp=c(2.2,0.8,0), cex=1.2)
plot(x, y, xlim=c(-10,10))
#plotCI(x, y, xlim=c(-10, 10), uiw=sigTrue, gap=0) # data and true error bar
#abline(a=modMat[1], b=modMat[2], col="red") # true model
dev.off()

### Sample from prior

# Sample from prior.
# priorSamp is an array with dimensions (Nsamp, 3) containing the
# samples for b_0, alpha, log10(sigma)
set.seed(100)
Nsamp <- 1e5 # will need to be larger if the priors are broader
priorSamp <- cbind(rnorm(n=Nsamp, mean=0, sd=1),
                   runif(n=Nsamp, min=-pi/2, max=pi/2),
                   runif(n=Nsamp, min=log10(0.5), max=log10(2)))
sel <- sample.int(n=Nsamp, size=100) # 100 of the prior samples

# Plot data and overplot 100 prior models from M2
pdf("linearmodel_evidence_prior_models_M2.pdf", width=5, height=4)
par(mfrow=c(1,1), mar=c(3.5,3.0,0.5,0.5), oma=0.5*c(1,1,1,1),
    mgp=c(2.2,0.8,0), cex=1.2)
plot(x, y, type="n", xlim=c(-10,10), ylim=c(-10,10), xaxs="i", yaxs="i")
for(j in sel) {
  abline(a=priorSamp[j,1], b=tan(priorSamp[j,2]), col="grey")
}
points(x, y)
dev.off()

# Plot data and overplot prior models from M1
pdf("linearmodel_evidence_prior_models_M1.pdf", width=5, height=4)
par(mfrow=c(1,1), mar=c(3.5,3.0,0.5,0.5), oma=0.5*c(1,1,1,1),
    mgp=c(2.2,0.8,0), cex=1.2)
plot(x, y, type="n", xlim=c(-10,10), ylim=c(-10,10), xaxs="i", yaxs="i")
for(j in sel) {
  abline(a=priorSamp[j,1], b=0, col="grey")
}
points(x, y)
dev.off()

### Calculate likelihoods, evidences, and BF

data <- data.frame(cbind(x,y))

# Calculate likelihoods and evidence for model M2
logLikeM2 <- rep(NA, Nsamp)
for(j in 1:Nsamp) {
  logLikeM2[j] <- log.like(theta=priorSamp[j,], data)
}
```

```
logEvM2 <- log10(mean(10^logLikeM2))
# Calculate likelihoods and evidence for model M1
priorM1   <- cbind(priorSamp[,1], rep(0, Nsamp), priorSamp[,3])
logLikeM1 <- rep(NA, Nsamp)
for(j in 1:Nsamp) {
  logLikeM1[j] <- log.like(theta=priorM1[j,], data)
}
logEvM1 <- log10(mean(10^logLikeM1))
# Print results
cat("log10(Ev_1)  = ", logEvM1, "\n")
cat("log10(Ev_2)  = ", logEvM2, "\n")
cat("log10(BF_12) = ", logEvM1 - logEvM2, "\n")
cat("BF_12        = ", 10^(logEvM1 - logEvM2), "\n")
```

11.3.2 A frequentist hypothesis testing approach

We saw in section 10.4 how we can use the t test to perform a frequentist analysis of the significance of a non-zero gradient for the straight line. Applying this to the data generated in the previous section we get the following.

```
summary(lm(y ~ x, data=data))
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
  (Intercept)  0.48161    0.37076   1.299   0.2301
  x            0.14934    0.05713   2.614   0.0309 *
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

  Residual standard error: 1.117 on 8 degrees of freedom
  Multiple R-squared:  0.4607,Adjusted R-squared:  0.3932
  F-statistic: 6.833 on 1 and 8 DF,  p-value: 0.03094
```

The p value is 0.03, so we can reject the hypothesis that the gradient is zero at the 95% confidence level (but not at the 99% level), which suggests that we might need the gradient. This is quite an extreme conclusion compared to the Bayesian analysis that gave $B_{12} = 1.30$, i.e. we could not draw any conclusion. But remember that this t test reaches its conclusion without testing M_2.

The frequentist approach is quite easy, but this should not lure us into relying on it too much.[7] As was discussed in section 10.6 this t test does not tell us which model is better, because it does not explicitly test and compare the models. It only tells us how often we expect to get a gradient this far from zero among a large number of data sets drawn from the null hypothesis which, remember, involves fixing the intercept.

Of course, if we really believed our null hypothesis, then our best fit would be a horizontal line through the mean of the data, which is not equal to the intercept found by fitting a straight line. Thus we should really do a linear regression on the data after we have centred them, i.e. offset them to have zero mean in x and y. We then keep the intercept fixed to

[7] Remember that behind `lm` and the t test there is at least as much computer code as there is in the evidence calculation code (section 11.3.1). It is true, however, that the linear least squares fit and t test involve much less computation, because the evidence calculation requires a large number of likelihood computations.

zero and do a t test on the resulting gradient. If we do this on the data above, the result is as follows.

```
xc <- data$x - mean(data$x)
yc <- data$y - mean(data$y)
summary(lm(yc ~ xc - 1))
  Coefficients:
     Estimate Std. Error t value Pr(>|t|)
  xc  0.14934    0.05386   2.773   0.0217 *
  ---
  Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

  Residual standard error: 1.053 on 9 degrees of freedom
  Multiple R-squared:  0.4607,Adjusted R-squared:  0.4007
  F-statistic: 7.687 on 1 and 9 DF,  p-value: 0.02166
```

The p value is now smaller due to the extra degree of freedom. If we had not kept the gradient fixed in this example, the p value would have been the same as before, because translating the data in x and/or y does not change the best fitting straight line.

11.4 Comparing Gaussians (or other distributions)

We can use the approach of section 11.3 for essentially any model comparison problem. We saw in section 10.2 how we could use frequentist hypothesis testing to determine whether data drawn from a Gaussian distribution had a mean significantly different from a specified value μ. We did this both when the standard deviation was known (the z test) and unknown (t test). To do this with Bayesian model comparison we must define at least two models. The first, M_1, specifies that the data are drawn from a model with known mean μ. The second, M_2, specifies that the data are drawn from a model with unknown mean μ', which is assigned a prior distribution $P(\mu')$. By definition of the problem, the likelihood for one data point (in both cases) is a Gaussian with mean specified by the model, and standard deviation σ that is either known, or is unknown, in which case it must be assigned a prior $P(\sigma)$. The likelihood for N data points $D = \{x\}$ is

$$P(D \,|\, \theta, M) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i - m)^2}{2\sigma^2}\right] \tag{11.28}$$

where for M_1 we have $m = \mu$ and parameters $\theta = \sigma$, and for M_2 we have $m = \mu'$ and parameters $\theta = (\sigma, \mu')$. If instead σ were known then M_1 would have no free parameters and M_2 just one. In the general case of σ being unknown, the evidence for the two models is

$$P(D \,|\, M_1) = \int P(D \,|\, \sigma, M_1) P(\sigma)\; d\sigma \tag{11.29}$$

$$P(D \,|\, M_2) = \iint P(D \,|\, \sigma, \mu', M_2) P(\sigma) P(\mu')\; d\sigma \, d\mu'. \tag{11.30}$$

In general we will have to do these integrals numerically, as we did in section 11.3. If we adopt a Gaussian prior for μ' then the integral over μ' for M_2 will be analytic. If σ is known then this is equivalent to $P(\sigma)$ being a delta function, in which case we can just drop the integral over σ.

These expressions for the evidence are not specific to a Gaussian likelihood, so this method will work with other distributions too (just the meaning and the number of parameters may change). We can also generalize this approach in order to compare two samples of data, rather than one sample to a fixed mean. The procedure is always to define the model with its parameters and priors, and then to define the likelihood. Writing down the evidence, we will usually then have to solve the integrals numerically (by drawing from the priors, or by using one of the other methods described in section 11.6). A practical advantage of this over frequentist hypothesis testing is that it is a unified approach: we do not need to go searching for appropriate test statistics and their distributions.

11.5 How the evidence accounts for model complexity

The evidence tells us how well the data are predicted by the model. Suppose we have some data drawn from a process and we want to compare a simple model M_s with a more complex one M_c. The data might be the (x, y) data set like that considered above, in which case M_s might just be a horizontal line and M_c a fourth-order polynomial. Consider first models with fixed parameters. We can obviously imagine some specific data sets, in particular one drawn from the fourth-order polynomial, for which M_c will predict the data much better than M_s ever can, and so will have a much higher likelihood. But for many more data sets, M_c makes such a specific prediction that it will match the data poorly. M_s, in contrast, will explain few of the data sets really well, but it will never do very badly either. The reason for this is that all models (whether their parameters are fixed or not) have a fixed "budget" of predictive power to explain all data sets. This is indicated by the fact that the model evidence $P(D\,|\,M)$ is a *normalized* probability density function of the data D. Improved predictive power (a larger probability density) over some part of the data space must be compensated by a lower one elsewhere, in order for $P(D\,|\,M)$ to integrate to one. This is illustrated in figure 11.4.

Usually the parameters θ of a model are not fixed, and can be fitted to the data. In this case a more complex model is more flexible and so will generally fit a wider range of data sets better than a simpler model, in the sense of producing a larger likelihood. So when comparing our two models, perhaps we should first fit each model by finding the parameters that maximize its likelihood $P(D\,|\,\theta, M)$ and see which model gives the higher value of this. This is wrong, because it will invariably favour the more complex model simply because it is more flexible. Consider a set of ten data points in a two-dimensional plane. A fifth-order polynomial will fit these data better (give a larger maximum likelihood) than a quadratic. And a ninth-order polynomial will fit them perfectly, noise included! But

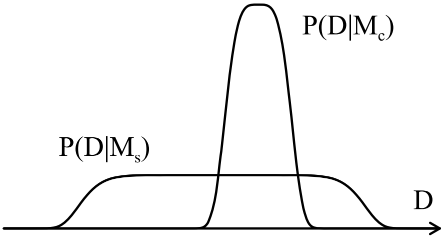

Fig. 11.4 Schematic illustration of the evidence for a simpler (M_s) and a more complex (M_c) model. The distributions show the probability of the data under the two models as a function of the measured data D.

this is not a "good" model in any useful sense of the term. Using the maximum likelihood to select a model would drive us to using ever increasingly complex models until we fit the data perfectly. (A model that fits the data "too well" is called an overfit.) Do not use maximum likelihood ratio tests for model selection!

In section 10.5 I explained how χ^2 is sometimes used to determine the goodness of fit of a given model, so we might think that the next step in that process is to find the model which has the smallest χ^2. But this is equally wrong, because χ^2 is just the negative log likelihood when we have a Gaussian error model; minimizing χ^2 is equivalent to maximizing the likelihood.

Maximum likelihood is the wrong thing to do for model comparison, because it ignores the different flexibilities of the models.

A closer analysis shows that the evidence takes into account the flexibility – or complexity – of the model, and so is an appropriate metric for model comparison. Taking the logarithm of equation 11.3 and rearranging we get

$$\begin{aligned} \log P(D\,|\,M) &= \log P(D\,|\,\theta, M) + [\log P(\theta\,|\,M) - \log P(\theta\,|\,D, M)] \qquad (11.31)\\ \log(\text{evidence}) &= \log(\text{likelihood}) + [\log(\text{prior}) - \log(\text{posterior})]. \end{aligned}$$

A complex model will generally have a posterior that is larger than the prior, and more so than a simple model. This is because (considering linear models) a more complex model has more parameters, and so has to spread its prior probability (which must integrate to unity) over a larger volume of parameter space than a model with fewer parameters. A typical value of its prior probability density will therefore be smaller. Furthermore, the more complex model can fit the data better, so its posterior probability density is typically larger. Both of these effects contribute to the term in square brackets in the above equation being more negative for a complex model than a simple model. This term acts as a penalty which reduces the evidence *for a given likelihood*. Thus for a more complex model to achieve a high evidence, it has to achieve a high enough likelihood in order to overcome this penalty. The evidence can therefore be seen as a combination of the fitting quality and a complexity penalty, and the best model will achieve the best trade-off between these two.

While the evidence involves a complexity penalty, it is not complexity per se that is penalized. What counts is how the plausibility of the model is changed in light of the data.

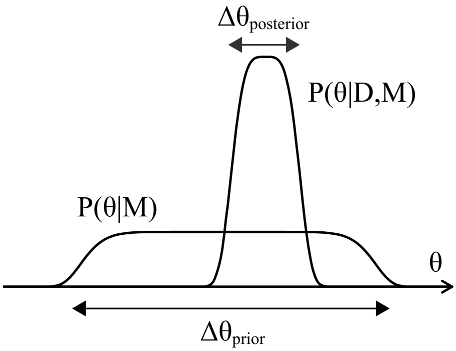
$\Delta\theta_{\text{posterior}}$
P(θ|D,M)
P(θ|M)
θ
$\Delta\theta_{\text{prior}}$

Fig. 11.5 Schematic illustration of the concept of the Occam factor. The prior $P(\theta\,|\,M)$ shrinks to the posterior $P(\theta\,|\,D,M)$ under the action of obtaining the data D.

This can be understood with the concept of the *Occam factor*. If the likelihood is dominated by a single peak at $\hat{\theta}$ over which the prior is more or less constant, then we can approximate the evidence as

$$\begin{aligned} P(D\,|\,M) &= \int P(D\,|\,\theta,M) \times P(\theta\,|\,M)\ d\theta \\ &\simeq P(D\,|\,\hat{\theta},M)\Delta\theta_{\text{posterior}} \times \frac{1}{\Delta\theta_{\text{prior}}} \\ &\simeq \underbrace{L(\hat{\theta})}_{\text{best fit likelihood}} \times \underbrace{\frac{\Delta\theta_{\text{posterior}}}{\Delta\theta_{\text{prior}}}}_{\text{Occam factor}} \end{aligned} \tag{11.32}$$

where $L(\hat{\theta}) = P(D\,|\,\hat{\theta},M)$ is the likelihood at the best fit solution, $\Delta\theta_{\text{prior}}$ is the prior parameter range, and $\Delta\theta_{\text{posterior}}$ is the posterior parameter range (the width of the posterior). This is illustrated in figure 11.5. The Occam factor (which is always less than or equal to one) measures the amount by which the plausible parameter volume shrinks on account of the data. For given $L(\hat{\theta})$, a simple model will fit over a large part of the parameter space, so $\Delta\theta_{\text{posterior}} \simeq \Delta\theta_{\text{prior}}$ and the Occam factor is not significantly less than one. In contrast a more complex model, or one which has to be more finely tuned to fit the data, will have a larger shrinkage, so $\Delta\theta_{\text{posterior}} \ll \Delta\theta_{\text{prior}}$. Now the Occam factor is small and the evidence is reduced. Of course, if the fit is good enough then $L(\hat{\theta})$ will be large, perhaps large enough to dominate the Occam factor and to give the model a large evidence.

The Occam factor is related to the idea of *Occam's razor*, which essentially says that we should chose simpler models where possible. This does not mean that we should always favour the simpler model: some phenomena and data sets are complex, and these will need complex models to describe them well. Bayesian model comparison balances complexity and accuracy via the Occam factor. Depending on the prior adopted – and the prior plays a major role in determining how complex a model is – this approach certainly can identify a more complex model as being best.

The Bayesian evidence naturally accommodates model complexity when doing model comparison. There is no natural parallel to this in frequentist hypothesis testing, where complexity control is either ignored or is imposed in an ad hoc fashion. I stress again that one should not use the ratio of maximum likelihoods to choose between models, because this just favours the more complex model. It is equivalent to compressing the prior parameter range to an arbitrarily small range around the best fit parameters, thereby artificially elevating the Occam factor for that model.

The evidence also encapsulates the concept of bias-variance trade-off mentioned in section 4.8. The bias of a model is the difference between the expected (average) predictions and the true values. The variance is the variability of the predictions around the expected ones. A more complex model is one that can be made to fit the data more precisely, thereby achieving a low bias. But the fit is more sensitive to the exact data: a small change in the data produces a big change in the fitted model. It has a larger variance. A simpler model, in contrast, is less influenced by the data (e.g. no high order terms in the polynomial), so has lower variance. But it cannot be made to fit the data as precisely because it lacks those more flexible terms, resulting a larger bias. Ideally we want a model with low bias and low variance, but in general increasing the model complexity will decrease the bias at the cost of increased variance.

11.6 Other ways of computing the evidence

The direct way of computing the evidence is to sample the prior and to calculate the average likelihood at these samples, as we did in section 11.3. While priors are normally simple and easy to sample from (without MCMC), for complicated problems we may need a very large number of samples to get an accurate measure of the evidence. This will occur if the likelihood (seen as a function of the parameters) is much more peaked than the prior, in which case only a very small fraction of the samples will contribute significantly to the evidence. This problem gets exponentially worse (literally) as the number of parameters increases. We are therefore interested in numerically more stable measures.

There are a number of other ways of numerically approximating the evidence. One is based on the idea of importance sampling (section 8.4.2). Newton & Raftery (1994) show that given a set of samples $\{\theta_l\}$ drawn from the posterior (using MCMC), the evidence can be approximated as

$$P(D\,|\,\sigma, M) \simeq \left(\frac{1}{N_s} \sum_{l=1}^{N_s} P(D\,|\,\sigma, \theta_l, M)^{-1} \right)^{-1} \tag{11.33}$$

which is the harmonic mean of the likelihood at these samples. However, this method is widely criticized in the literature as inaccurate and unstable, and only converges to the true evidence with an impractically large number of samples in most cases. It is not recommended.

One of the attractions of the harmonic mean method is that is uses samples drawn from the posterior. A single sampling procedure can therefore be used both for estimating parameters and for approximating the evidence. Better methods than the harmonic mean exist which do this. One example is nested sampling (e.g. Skilling, 2004; Sivia & Skilling, 2006). Another is thermodynamic integration (e.g. Gregory, 2005; Lartillot & Philippe, 2006; Friel & Pettitt, 2008). This latter method requires running parallel MCMC chains, each of which involves the likelihood raised to a power between 0 and 1, a process known as parallel tempering. The integral of the mean log likelihood values over the chains gives the log evidence.

If the likelihood, when viewed as a function of the parameters $\boldsymbol{\theta}$, is dominated by a single narrow peak, then we may be able to approximate it as a multivariate Gaussian about its maximum $\hat{\boldsymbol{\theta}}$. In that case we can use the quadratic approximation of section 7.1 to write the likelihood as

$$P(D\,|\,\boldsymbol{\theta}, M) \;\simeq\; P(D\,|\,\hat{\boldsymbol{\theta}}, M)\exp\left(-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^{\mathsf{T}}\Sigma_{\hat{\boldsymbol{\theta}}}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right). \tag{11.34}$$

The maximum we can find by numerical optimization (see section 12.6). The covariance matrix $\Sigma_{\hat{\boldsymbol{\theta}}}$ is equal to the negative inverse of the matrix of second derivatives of the log likelihood evaluated at $\hat{\boldsymbol{\theta}}$ (equation 7.15). The dependence on the data comes through this and $\hat{\boldsymbol{\theta}}$. Assuming the prior is reasonably constant across this narrow peak, we can approximate it as $P(\hat{\boldsymbol{\theta}}\,|\,M)$. The integration of the likelihood over this prior (equation 11.4) then just involves integrating the exponential term in equation 11.34 over $\boldsymbol{\theta}$. This is the Gaussian integral, the result of which is $(2\pi)^{J/2}|\Sigma_{\hat{\boldsymbol{\theta}}}|$ where J is the dimensionality of the parameter space. The resulting approximation for the evidence is therefore

$$P(D\,|\,M) \;\simeq\; (2\pi)^{J/2}|\Sigma_{\hat{\boldsymbol{\theta}}}|^{1/2}P(D\,|\,\hat{\boldsymbol{\theta}}, M)P(\hat{\boldsymbol{\theta}}\,|\,M). \tag{11.35}$$

This will only be a good approximation in the case of highly informative data, such that the approximations mentioned are valid.

Other methods for computing the evidence are discussed in Kass & Raftery (1995). In the next section I will summarize one other method, which I introduced in Bailer-Jones (2012).

11.6.1 The cross-validation likelihood

Definition

Let $D = (D_1, D_2, \ldots, D_N)$ denote a set of N measurements (they could be scalars or vectors). We group the N measurements into K disjoint partitions $(1 < K \leq N)$. Denote the data in the kth partition as D_k and its complement as D_{-k}, i.e.

$$D \;=\; D_k \cup D_{-k}. \tag{11.36}$$

The principle of the cross-validation (CV) likelihood is to calculate the likelihood of D_k using D_{-k}, without having an additional dependence on a specific choice of model parameters θ. This likelihood is $P(D_k\,|\,D_{-k}, M)$. It tells us how well, under model M, some of the data are predicted by the other data. By marginalization

$$\begin{aligned} L_k \equiv P(D_k \,|\, D_{-k}, M) &= \int P(D_k \,|\, D_{-k}, \theta, M) P(\theta \,|\, D_{-k}, M)\, d\theta \\ &= \int \underbrace{P(D_k \,|\, \theta, M)}_{\text{likelihood}} \underbrace{P(\theta \,|\, D_{-k}, M)}_{\text{posterior}}\, d\theta \end{aligned} \tag{11.37}$$

where D_{-k} drops out of the first term because the model predictions are conditionally independent of these data when θ is specified. I call L_k the *partition likelihood*. It is the likelihood for data D_k marginalized over the posterior computed using all the other data D_{-k}. In practice we estimate the partition likelihood using a Monte Carlo approximation of this integral (cf. equation 8.9),

$$L_k \simeq \frac{1}{N_s} \sum_{l=1}^{N_s} P(D_k \,|\, \theta_l, M), \tag{11.38}$$

where the N_s samples $\{\theta_l\}$ are drawn from the posterior $P(\theta \,|\, D_{-k}, M)$.

Combining the partition likelihoods over all K partitions in some way should give an overall measure of the fit of the model. As L_k is a product of the likelihoods for each data vector within a partition, it scales multiplicatively with the number of data vectors in partition k. This suggests that an appropriate combination is

$$L_{\text{CV}} = \prod_{k=1}^{K} L_k \tag{11.39}$$

which I call the *K-fold CV likelihood*, for $1 < K \leq N$. For $K < N$ its value will depend on the choice of partitions. If $K = N$ there is one data vector per partition, a unique choice. I call this the *leave-one-out CV likelihood*.[8]

The posterior PDF in equation 11.37 is given by Bayes' theorem. As we are sampling, it is sufficient to use the unnormalized posterior, which is

$$P(\theta \,|\, D_{-k}, M) \propto P(D_{-k} \,|\, \theta, M) P(\theta \,|\, M). \tag{11.40}$$

Relation to the evidence

Whereas the evidence involves integrating the likelihood (for D) over the prior (equation 11.4), the partition likelihood involves integrating the likelihood (for D_k) over the posterior (for D_{-k}) (equation 11.37). This is like using D_{-k} to build a new prior from "previous" data. We can use the product rule to write the partition likelihood as

$$L_k \equiv P(D_k \,|\, D_{-k}, M) = \frac{P(D \,|\, M)}{P(D_{-k} \,|\, M)}, \tag{11.41}$$

i.e. the ratio of the evidence calculated over all the data to the evidence calculated on the subset of the data used in the posterior sampling.

[8] In principle we could also compute the K-fold CV likelihood with $K = 1$, in which case we use all of the data both to draw the posterior samples and to calculate the likelihood. This is not a useful measure of goodness-of-fit.

The K-fold CV likelihood is in fact equivalent to the evidence. Using equation 11.39 and 11.41 we can write

$$L_{\mathrm{CV}} = \prod_k \frac{P(D\,|\,M)}{P(D_{-k}\,|\,M)} = \frac{P(D\,|\,M)^K}{\prod_k P(D_{-k}\,|\,M)}. \tag{11.42}$$

The denominator can be written as $P(D\,|\,M)^{K-1}$, as we can see if we consider the case $K = 3$ (dropping the M for brevity)

$$\begin{aligned}\prod_k P(D_{-k}) &= P(D_1, D_2)P(D_2, D_3)P(D_1, D_3) \\ &= P(D_1)^2 P(D_2)^2 P(D_3)^2 \\ &= P(D)^{K-1}\end{aligned} \tag{11.43}$$

where we have used the fact that the data sets are disjoint and (assumed) independent. Thus it follows from equation 11.42 that

$$L_{\mathrm{CV}} = \frac{P(D\,|\,M)^K}{P(D\,|\,M)^{K-1}} = P(D\,|\,M). \tag{11.44}$$

Why use the cross-validation likelihood?

Why would we use the cross-validation likelihood if it is equivalent to the evidence? The main reason is numerical accuracy. The evidence is computed numerically by sampling from the prior (equation 11.16). The prior is often a very broad function compared to the region over which the likelihood is non-negligible. Many of the likelihood evaluations contributing to the Monte Carlo sum are therefore negligible, so this sum could have a large numerical uncertainty. In particular, it is likely to be underestimated. The number of samples we require for a good estimation is unknown (it could be very large), and grows exponentially with the dimensionality of the parameter space. Samples from the posterior (typically obtained by MCMC), in contrast, are more likely to produce a large likelihood, because the posterior is the product of the likelihood with the prior. This is how the partition likelihood is calculated, so it is more likely to be numerically stable. The number of likelihood calculations we will need for the cross-validation likelihood is likely to be far smaller than the number required for the evidence to achieve a stable result.

11.7 Other measures for model comparison: AIC and BIC

The evidence has a simple, solid theoretical basis, but it can be slow to compute. Its dependence on the prior can also make it difficult to interpret when our prior information is hard to represent as a PDF. Other metrics have been developed which may be more convenient for model comparison. Two of these – AIC and BIC – use the maximum likelihood.

The *Akaike information criterion* (AIC) was derived by Akaike (1973, 1974) from concepts in information theory. We have seen that whereas a model with a larger likelihood will fit the data better, selecting models on this basis will just favour more complex models. Often (but not always) a more complex model is one with more parameters. So we could use the maximum likelihood penalised by the number of parameters as a measure of model quality. The AIC does precisely this and is defined as

$$\mathrm{AIC} = -2\ln L_{\mathrm{max}} + 2J \tag{11.45}$$

where J is the number of independently fitted parameters in the model. Like the evidence, the absolute value of AIC is of no relevance. Unlike the (logarithm of the) evidence, the more negative the AIC, the better the model. Note that the AIC is on the scale of twice the natural logarithm of the likelihood, whereas the evidence is linear in the likelihood. To use the AIC for comparing models we simply calculate it for each of the models of interest on the same data, and identify the best model as the one with the most negative AIC. Whether this is *significantly* better than the others is a choice. If we required a model to have ten times the likelihood of another model with the same number of parameters, for example, then we would require a difference in the AIC of at least $2\ln 10 = 4.6$. Note that one parameter has the same relevance as a factor e in likelihood. The number of parameters J in the definition of the AIC counts the presence of variance of the data (whether known or not) as one parameter. Thus if the model were a straight line in one dimension – which has two parameters – fit to noisy data, we would have $J = 3$. This idea of counting parameters is clearly rather simplistic, because a second order polynomial is much less flexible than a sinusoidal function, which also has three parameters.

A variant on AIC is the *Bayesian information criterion* which, as its name suggests, was derived from Bayesian considerations (Schwarz, 1978). It is defined as

$$\mathrm{BIC} = -2\ln L_{\mathrm{max}} + J\ln N \tag{11.46}$$

where N is the number of data points. The BIC generally imposes a larger penalty for the complexity of the model than does the AIC (because $\ln N > 2$ when $N > 8$).

To illustrate the AIC and BIC I calculate it for the two models used in the coin example in section 11.2. Model M_1 has the probability of getting a heads in a single toss fixed at $p = 1/2$. This model has no free parameters so $J = 0$. Model M_2 is the case that p is unknown so has $J = 1$. Actually, in section 11.2 model M_2 is the case that p is unknown *and* is described by a beta prior. But the AIC and BIC do not recognise prior distributions: they only take into account the number of free parameters in the model. Hence they treat all prior distributions over p as being the same model, including a prior that has a very narrow distribution about $p = 1/2$, even though conceptually this is much more like M_1. This shows the inflexibility of the AIC and BIC compared to calculating the evidence. It's the price we pay for their simple (and simplistic) measure of model complexity.

The likelihood for both models M_1 and M_2 is the binomial distribution, equation 11.5. Model M_1 has no free parameters, so its maximum likelihood L_{max} is the value of the binomial distribution at $p = 1/2$. The maximum likelihood for model M_2 can be found by differentiation (see section 4.4) to be at $p = r/n$. The number of data points N in the definitions of AIC and BIC is the number of coin tosses, n. The difference in the AIC and

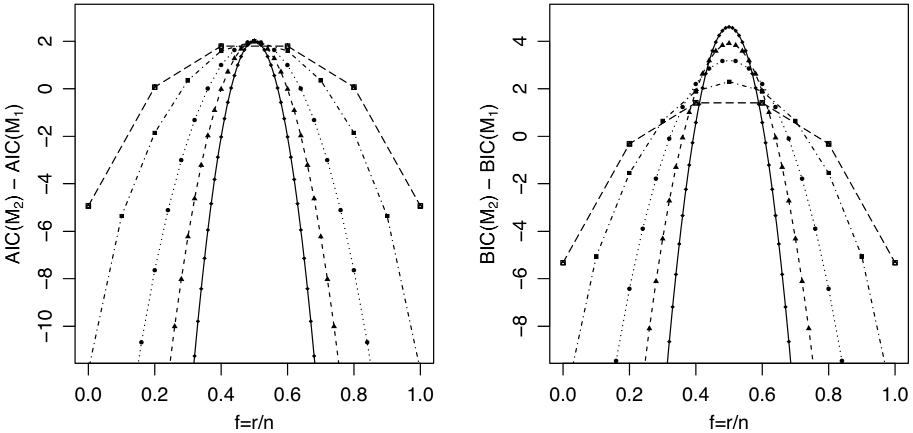

Fig. 11.6 Left: the difference between the AIC for M_2 (which has p as a free parameter) and M_1 ($p = 1/2$) in the coin problem from section 11.2. As in figure 11.1 the five lines are, in order of increasingly narrow distribution, for $n = 5, 10, 25, 50$, and 100. Note that the AIC – and thus the vertical axis here – is proportional to twice the natural logarithm of the likelihood, whereas the vertical axis in figure 11.1 is proportional to the base ten logarithm of the evidence (and therefore also the likelihood). Both plots cover a factor of order 1000 in likelihood (the range on the vertical axis here is about $2 \ln 1000 = 13.8$). Right: as the left panel, but for the BIC.

BIC for the two models is

$$\begin{aligned} \mathrm{AIC}(M_2) - \mathrm{AIC}(M_1) &= 2 \ln P_{\mathrm{bin}}(r \,|\, p = 1/2, n) - 2 \ln P_{\mathrm{bin}}(r \,|\, p = r/n, n) + 2 \\ \mathrm{BIC}(M_2) - \mathrm{BIC}(M_1) &= 2 \ln P_{\mathrm{bin}}(r \,|\, p = 1/2, n) - 2 \ln P_{\mathrm{bin}}(r \,|\, p = r/n, n) + \ln n. \end{aligned} \tag{11.47}$$

We can now calculate these for different experimental setups and results – for different n and different numbers of heads r – and use the results to decide which model is favoured.

The results for the AIC are shown in the left panel of figure 11.6. We see a broadly similar behaviour as that obtained with Bayes factors (figure 11.1), although a closer look shows that the two metrics disagree about the relative performance of the two models as n changes. In this particular example, the variation of the BIC with the data (right panel of figure 11.6) is much more like the Bayes factor. But it must be stressed that this is not always the case. The code below produces the two plots.

R file: `coin_aic_bic.R`

```
##### Demonstration of AIC and BIC on coin modelling comparison problem

phi <- 1/2
n   <- c(5,10,25,50,100)

pdf("coin_aic.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(c(0,1), c(0,1), type="n", xlim=c(0,1), ylim=c(-11,3), xlab="f=r/n",
     ylab=expression(paste("AIC(",M[2],") - AIC(",M[1],")")))
for(i in 1:length(n)) {
  r   <- 0:n[i]
  f   <- r/n[i]
  aicM1 <- -2*dbinom(x=r, size=n[i], prob=1/2,     log=TRUE)
  aicM2 <- -2*dbinom(x=r, size=n[i], prob=r/n[i], log=TRUE) + 2
  points(f, aicM2-aicM1, cex=0.5, pch=13+i)
  lines(f,  aicM2-aicM1, lty=6-i)
}
dev.off()

pdf("coin_bic.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(c(0,1), c(0,1), type="n", xlim=c(0,1), ylim=c(-9,5), xlab="f=r/n",
     ylab=expression(paste("BIC(",M[2],") - BIC(",M[1],")")))
for(i in 1:length(n)) {
  r   <- 0:n[i]
  f   <- r/n[i]
  bicM1 <- -2*dbinom(x=r, size=n[i], prob=1/2,     log=TRUE)
  bicM2 <- -2*dbinom(x=r, size=n[i], prob=r/n[i], log=TRUE) + log(n[i])
  points(f, bicM2-bicM1, cex=0.5, pch=13+i)
  lines(f,  bicM2-bicM1, lty=6-i)
}
dev.off()
```

In the above example the maximum likelihood and therefore the AIC and BIC could be calculated analytically. But so could the evidence for the priors we chose, so the AIC and BIC offer no advantage in this case. In most real-world problems, all of these metrics would need to be computed numerically. Whereas computing the evidence normally involves drawing from the prior (to do a Monte Carlo integration), finding the maximum likelihood is an optimization process (section 12.6). The latter may be faster, and moreover does not require us to select a prior. Yet this is precisely the weak point of the AIC and BIC: they rely on just one value of the likelihood (the maximum) and use simplistic measures of model complexity based on the number of parameters and number of data points. Within a class of models (e.g. all polynomials) in which the models are nested this is not such a problem, because in comparing AIC or BIC we are only interested in the difference in the number of parameters between the models, and this will be clear. But overall, whereas the AIC and BIC are potentially more convenient than the evidence, they are not as general as a metric for model selection. Further discussion can be found in Kass & Raftery (1995).

11.8 The stopping problem

The stopping problem is a well-known issue that illustrates a fundamental problem with hypothesis testing using p values. It has been discussed by many authors in different contexts. This particular example is taken from Gregory (2005).

In an astronomical survey we observe $n = 102$ stars and determine that $r = 5$ of these are white dwarfs. Our hypothesis M is that 10% of stars in such surveys are white dwarfs. Do the data support this hypothesis?

11.8.1 The frequentist approach(es)

From a frequentist hypothesis testing point of view we want to do a two-sided test of our hypothesis that the true fraction of white dwarfs is $p = 0.1$. The test is two-sided because both high deviations and low deviations of the data from this value of p are evidence against M. We assume that all of the observations are independent.

We first give the data and this information to Alice. This is what she does.

Frequentist Alice. Having observed n stars, she says that the probability that r of these are white dwarfs is given by the binomial distribution

$$P(r\,|\,p,n) \;=\; \binom{n}{r}\, p^r (1-p)^{n-r}. \tag{11.48}$$

Thus (twice) the probability of observing five or fewer white dwarfs is

$$P_A \;=\; 2\sum_{r=0}^{r=5} P(r\,|\,p,n) \;=\; 0.102 \tag{11.49}$$

the factor of 2 arising because she is doing a two-sided test. This is Alice's p value. It is more than 0.05, so she cannot reject M at the 95% level (or any other higher level).

We then give the same information and data to Bob, but additionally tell him that before we started observing, we decided to keep observing stars until we had observed five white dwarfs, and then we stopped. This is what he does.

Frequentist Bob. Given this extra information, the random variable is now n, not r. The probability that there were $r-1$ white dwarfs among the first $n-1$ observations, and therefore also $(n-1)-(r-1)$ stars which were not white dwarfs, is given by a binomial distribution. The probability that the last observation was a white dwarf is p. Thus the probability of the data is

$$\begin{aligned} P(n\,|\,p,r) \;&=\; \binom{n-1}{r-1}\, p^{(r-1)}(1-p)^{(n-1)-(r-1)} \times p \\ &=\; \binom{n-1}{r-1}\, p^r (1-p)^{n-r}. \end{aligned} \tag{11.50}$$

This, incidentally, is called the *negative binomial distribution*, and is a probability distribution in n, in contrast to the binomial distribution, which is a probability distribution in r. Thus (twice) the probability of observing 102 stars or more given that we observed five white dwarfs is

$$\begin{aligned} P_B \;&=\; 2\sum_{n=102}^{n=\infty} P(n\,|\,p,r) \\ &=\; 2\left(1 - \sum_{n=5}^{n=101} P(n\,|\,p,r)\right) \;=\; 0.044. \end{aligned} \tag{11.51}$$

where again we have a factor of 2 due to the two-sided test. Bob's p value is less than 0.05, so he rejects M at the 95% confidence level (and any other lower level).

Who is right?

Frequentist hypothesis testing considers both analyses to be correct. The contradiction arises because in order to calculate a p value we must define a reference set of hypothetical data which were not, but which could have been, observed. This requires us to make assumptions about what could have been observed. Alice considers n to be fixed and r to be the random variable, so she takes as a reference set different unobserved values of r. Bob, in contrast, considers r to be fixed and n to be the random variable, so his reference set is all the possible values of n.

To interpret these results as a frequentist we would have to decide in advance (before we started the experiment) which approach we were going to take. Let's assume we decide to keep observing stars until we observe five white dwarfs. What do we do if we get bad weather, or a lack of funds prevents us from reaching this number? We could no longer do our original analysis, but we also know that a different analysis could give us the opposite result. We are in a dilemma and would have to throw away the data. But obviously the data don't suddenly become worthless: they are what they are. The problem is that this approach to the analysis is dependent on irrelevant information, namely which variable we choose to be random.

11.8.2 The Bayesian resolution

The frequentist approach runs into problems because it is asking the wrong question, namely the probability of getting the data, whereas what we should be doing is an inference of p, i.e. computing its PDF given the data. Let's now take this Bayesian approach in which our model M' is that p is unknown and is described by a prior $P(p\,|\,M')$. Both Alice and Bob want to determine

$$P(p\,|\,D, M') = \frac{1}{Z} P(D\,|\,p, M') P(p\,|\,M'). \tag{11.52}$$

Bayesian Alice. The likelihood is

$$P(D\,|\,p, M') = P(r\,|\,p, n) = \binom{n}{r} p^r (1-p)^{n-r} \tag{11.53}$$

so the posterior is

$$P(p\,|\,D, M') = \frac{p^r (1-p)^{n-r} P(p\,|\,M')}{\int p^r (1-p)^{n-r} P(p\,|\,M')\, dp} \tag{11.54}$$

where the binomial coefficient cancels out because it is independent of p.

Bayesian Bob. The likelihood is

$$P(D\,|\,p, M') = P(n\,|\,p, r) = \binom{n-1}{r-1} p^r (1-p)^{n-r}. \tag{11.55}$$

Taking the same prior as Alice, the posterior is

$$P(p\,|\,D, M') \;=\; \frac{p^r(1-p)^{n-r}P(p\,|\,M')}{\int p^r(1-p)^{n-r}P(p\,|\,M')\,dp} \tag{11.56}$$

and once again the binomial coefficient involving the factorials cancels out.

The Bayesian approach gives the same conclusion about the probability of p and therefore about how well the hypothesis is supported by the data. Alice and Bob now arrive at the same answer because both the binomial and negative binomial distributions have the same dependence on p. The two frequentist approaches differ from each other because they look at the distribution of the "random variable", either r or n, and this has a different dependence in the two cases. In a Bayesian approach the data is not a random variable that could have been something else. It is the data which are given, and the parameter which is unknown.

To test the original hypothesis we can look at the posterior PDF over p. This is shown in figure 11.7 using a uniform prior. The value of $p = 0.1$ is quite far down in the low tail of the distribution, thus casting doubt on the hypothesis.

An alternative approach is to compute the Bayes factor between a model with $p = 0.1$ and the model with a prior $P(p\,|\,M')$. The evidence is the integral of the product of the likelihood and prior. The binomial coefficient, which depends on r and n but not on p, factorizes out of the integral. When Alice forms her Bayes factor, the binomial coefficient cancels in the ratio. The same happens for Bob, leaving them both with the same dependence on p. So the two of them would compute identical Bayes factors also. This approach tells us which of the two models fits better, but not whether the original hypothesis is true or not. But even if we took one of the frequentist approaches we could also not say whether the hypothesis was true or not: we could only reject or fail to reject it at some confidence level (see sections 10.1 and 10.6).

The following R code calculates the frequentist and Bayesian solutions and produces the plot. The function `dnbinom` computes the density of the negative binomial distribution. To plot figure 11.7 I normalize the posterior numerically (so you can easily insert any prior), but as I use a uniform prior $P(p\,|\,M') = 1$, the normalization constant here is $B(r+1, n-r+1)$ (compare equation 11.7 with the denominator in equations 11.54 and 11.56).

R file: `code/stopping_problem.R`

```
##### The stopping problem

p <- 0.1
n <- 102 # no. trials
r <- 5   # no. successes

PA <- 2*sum(dbinom(x=0:r, size=n, prob=p))
# = 2*pbinom(q=r, size=n, prob=p)
PB <- 2*(1 - sum(dnbinom(x=0:(n-r-1), size=r, prob=0.1)))
# = 2*sum(dnbinom(x=(n-r):1e4, size=r, prob=0.1))
# where 1e4 is used as an approximation of infinity.
# Note the definitions adopted in dnbinom():
# x=n-r (no. failures) and size=r. Can check by comparing:
```

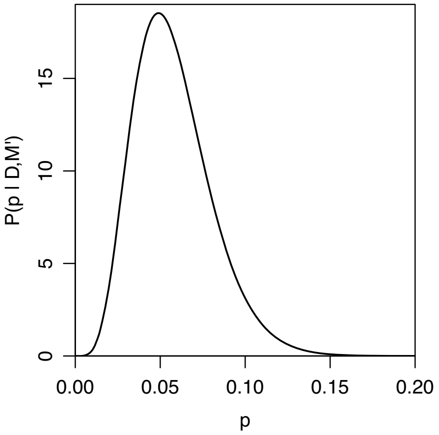

Fig. 11.7 Posterior PDF over p in the stopping problem for $n = 102$, $r = 5$ when using a uniform prior on p.

```
# choose(n-1, r-1)*p^r*(1-p)^(n-r)
# dnbinom(x=n-r, size=r, prob=p)

pdf("stopping_problem.pdf", 4, 4)
par(mfrow=c(1,1), mar=c(3.5,3.5,0.5,1), oma=0.5*c(1,1,1,1),
    mgp=c(2.2,0.8,0))
Nsamp  <- 1e4
deltap <- 1/Nsamp
pgrid  <- seq(from=1/(2*Nsamp), by=1/Nsamp, length.out=Nsamp)
pdense <- pgrid^r * (1-pgrid)^(n-r)   # with uniform prior
pdense <- pdense/(deltap*sum(pdense)) # normalize posterior
# could instead do analytically for a uniform prior: pdense/beta(r+1, n-r+1)
plot(pgrid, pdense, type="l", lwd=1.5, xaxs="i", yaxs="i", xlim=c(0,0.2),
     ylim=c(0,19), xlab="p", ylab="P(p | D,M')")
dev.off()
```

11.9 Issues with Bayesian model comparison

While Bayesian model comparison overcomes many of the problems of frequentist hypothesis testing, it is not without its own issues.

(1) The main issue is the dependence of the evidence on the prior. Conceptually priors are a good thing, as they are a way of formalizing the additional information beyond the data (such as plausibility constraints, invariances, etc.) that one invariably has. Moreover, we have seen in earlier chapters how priors are unavoidable, and can even be essential to get meaningful solutions at all. In parameter estimation, i.e. when computing the posterior PDF $P(\theta \,|\, D, M)$, we have seen how the dependence of the posteriors on the

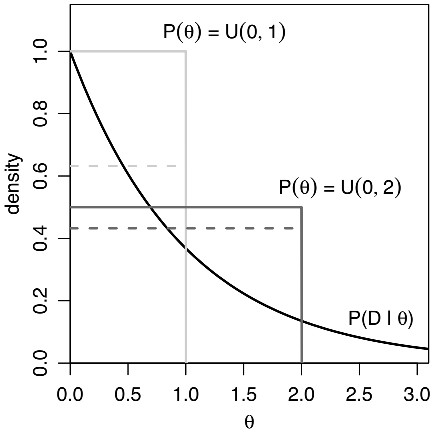

Fig. 11.8 Illustration of the dependence of the evidence on the prior. The black line shows the likelihood as a function of the parameter θ. The light grey and dark grey solid lines show uniform priors with ranges $(0, 1)$ and $(0, 2)$ respectively. The dashed lines show the two corresponding values of the evidences, the evidence being the integral of the likelihood over the prior.

prior becomes ever weaker as the data become more informative (e.g. figure 5.6). The evidence tends to have a stronger dependence on the prior. Consider a one parameter problem in which the likelihood $P(D\,|\,\theta, M)$ has an exponential dependence on the parameter, $e^{-\theta}$ (see figure 11.8). Suppose we want to adopt a uniform prior. As we must sample from the prior to compute the evidence, we need a proper prior, so we need to set limits on it. If we first choose $\mathcal{U}(0, 1)$ (the light grey line in the figure), then the evidence, which is the integral of the likelihood over this, is 0.63 (shown by the dashed line). But if we then decided to use a prior with twice the range, $\mathcal{U}(0, 2)$, the evidence becomes 0.42 (the dark grey lines in the figure). In general, changing the prior range changes the evidence. In principle this is what we want, because the prior is an integral part of the model: changing the prior changes the model. But it can be a problem in practice if we are not sure what prior range we should use. There is not much we can do about this other than do our best to specify plausible prior ranges based on the available data, and to investigate the sensitivity of the Bayes factors to changes in the priors. Clearly one should not use the narrowest priors imaginable for one model, and the broadest imaginable for another. If the likelihood is very peaked (informative data) and all plausible priors are essentially flat over this range, then the evidence is insensitive to the choice of prior. This is exploited by the quadratic approximation of the evidence (equation 11.35).

(2) For equal model priors the expression for the posterior probability of some model M_1 (see equation 3.8) is

$$P(M_1\,|\,D) = \frac{1}{1 + \sum_{k=2}^{K} BF_{k1}} \tag{11.57}$$

where BF_{k1} is the Bayes factor of model M_k with respect to model M_1. These are always positive, so the posterior probability of M_1 increases monotonically as the number of models included in the sum decreases. Thus if we (erroneously) do not include all plausible alternative models (those which don't have $BF_{k1} \simeq 0$) in our analysis, the sum is smaller than it should be, and the posterior probability of M_1 is artificially increased. This could happen if the models are complex, so we cannot be sure that we have an exhaustive set. It's not really a problem in practice, however, because we don't normally try to calculate model posterior probabilities. We normally just calculate posterior odds ratios or Bayes factors of a series of models relative to a baseline model. This allows us to identify the best of a set of given models, even when they are incomplete.

(3) Bayesian model comparison does what's written on the label: it compares models. It cannot be used to assess the "absolute" probability of an isolated model. From a Bayesian perspective this makes no sense, because if we only ever have one model, the data must have come from it. But in practice we often want to get some idea of whether just a single model is a plausible explanation of the data.

(4) Bayesian inference can be computationally intensive. While there are some analytic results (such as those in chapters 5 and 6, and reasonable approximations can sometimes be made to give analytic results), general problems require us to sample and integrate PDFs using numerical methods. The computational bottle-neck is often the calculation of the likelihood. This computational cost is the price we pay for the generality and power of the method.

(5) The entire Bayesian approach depends on the likelihood function. There may be problems where the likelihood is hard to define accurately, making any inferences based on it of limited value. In such cases we may be able to substitute the likelihood for a similarity measure and still obtain reliable results. One such approach is known as *Approximate Bayesian Computation* (ABC).

We will see some practical alternatives to using the evidence to control for model complexity, and therefore to select the most appropriate model, in chapter 12, using cross-validation and regularization together with basis functions.