7 Bayesian model comparison

In the previous chapters we concerned ourselves with parameter estimation, often staying within the context of a single model. In much of cognitive science, however, researchers entertain more than just a single model. Different models often represent competing theories or hypotheses, and the focus of interest is on which substantive theory or hypothesis is more plausible, more useful, and better supported by the data. In order to address these questions we need to move beyond parameter estimation and turn to Bayesian methods for *model comparison*.

7.1 Marginal likelihood

To understand the Bayesian solution to the problem of selecting between competing models, we return to the very first equation of this book: Bayes' rule. We now indicate explicitly that the parameter θ depends on a specific model $\mathcal{M}_1$ that is entertained:

$$\begin{aligned}\text{posterior} = p\left(\theta \mid D, \mathcal{M}_1\right) &= \frac{p\left(D \mid \theta, \mathcal{M}_1\right) p\left(\theta \mid \mathcal{M}_1\right)}{p(D \mid \mathcal{M}_1)} \\ &= \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}.\end{aligned} \tag{7.1}$$

The marginal likelihood $p(D \mid \mathcal{M}_1)$ is a single number that is sometimes called the *evidence*. It indicates the probability of the observed data D in light of the model specification $\mathcal{M}_1$. One interpretation is that the marginal likelihood measures the average quality of the predictions that a model has made for the observed data. The better the predictions, the greater the evidence.

As a simple example, suppose you construct a model $\mathcal{M}_x$ with a single parameter ξ. Furthermore, suppose that this parameter can take on only three values, $\xi_1 = -1$, $\xi_2 = 0$, and $\xi_3 = 1$. You assign these parameter values the following prior probability masses: $p\left(\xi_1\right) = 0.6$, $p\left(\xi_2\right) = 0.3$, and $p\left(\xi_3\right) = 0.1$. These assignments reflect the belief or knowledge that low values of ξ are more likely than high values of ξ. Next, you obtain data D, and you compute the likelihood for all parameter values. For example, assume that $p\left(D \mid \xi_1\right) = 0.001$, $p\left(D \mid \xi_2\right) = 0.002$, and $p\left(D \mid \xi_3\right) = 0.003$.[1]

[1] The likelihood $p\left(D \mid \xi_\star\right)$ quantifies the degree to which the observed data are expected, given a particular parameter value $\xi_\star$. Hence, you can think of the likelihood as a measure of goodness-of-fit.

Then, the marginal likelihood of your model $\mathcal{M}_x$ is given by

$$\begin{aligned} p(D \mid \mathcal{M}_x) &= p(\xi_1)\,p(D \mid \xi_1) + p(\xi_2)\,p(D \mid \xi_2) + p(\xi_3)\,p(D \mid \xi_3) \\ &= 0.6 \times 0.001 + 0.3 \times 0.002 + 0.1 \times 0.003 \\ &= 0.0015. \end{aligned}$$

The marginal likelihood is computed by averaging the likelihood across the parameter space, with prior probabilities acting as averaging weights. Thus, in order to determine how well a model predicted the data, we need to take into account *all* predictions that the model made and weight these by their prior probability. We can restate this mathematically by saying that the marginal likelihood is obtained by averaging out the model parameters in accordance with the *law of total probability*. For a parameter ξ that can take on k discrete values, the marginal likelihood is given by

$$p(D \mid \mathcal{M}_1) = \sum_{i=1}^{k} p(D \mid \xi_i, \mathcal{M}_1)\, p(\xi_i \mid \mathcal{M}_1)\,. \tag{7.2}$$

For a continuously varying parameter θ—such as a binomial rate parameter that can take on any value between 0 and 1—the sum needs to be replaced by an integral, so that

$$p(D \mid \mathcal{M}_1) = \int p(D \mid \theta, \mathcal{M}_1)\, p(\theta \mid \mathcal{M}_1)\, \mathrm{d}\theta. \tag{7.3}$$

Despite the difference in notation between Equations 7.2 and 7.3, the computation is conceptually the same. The likelihood is evaluated for every possible parameter value, weighted by its prior plausibility, and added to the total.

The foregoing shows that in order to obtain firm evidence, a model needs to make a high proportion of good predictions. This is precisely the problem with models that are overly complex. These models are able to make many predictions, but a high proportion of these predictions will turn out to be false. Complex models need to divide their prior predictive probability across all of their predictions, and, in the limit, a model that predicts almost everything has its prior predictive probability spread out thinly: so thinly, in fact, that the occurrence of any particular event cannot substantially increase that model's credibility. This is the Bayesian justification for the adage "a model that predicts everything predicts nothing." As was illustrated above, the marginal likelihood for a model $\mathcal{M}_1$ is calculated by averaging the likelihood $p(D \mid \theta, \mathcal{M}_1)$ over the prior $p(\theta \mid \mathcal{M}_1)$.

The basic principle is that a model is complex when it makes many predictions. In practice, this can come about in a number of ways. The most obvious factor is that the inclusion of more parameters in a model allows it to make more predictions.

More subtly, models also become more complex as prior distributions over parameters become broad. When prior distributions become very broad, relatively low prior probability is assigned to those parts of the parameter space where the likelihood is high (i.e., where the predictions are good). It also means that relatively high prior probability is assigned to the remaining parts of the parameter

Box 7.1 **Ockham's razor**

Ockham's razor is also known as the principle of parsimony, and it embodies a preference for assumptions, theories, and hypotheses that are as simple as possible without being false. The metaphorical razor cuts away all theorizing that is needlessly complex. The razor is named after the English logician and Franciscan friar Father William of Ockham (c.1288–c.1348), who stated "*Numquam ponenda est pluralitas sine necessitate*" (Plurality must never be posited without necessity), and "*Frustra fit per plura quod potest fieri per pauciora*" (It is futile to do with more what can be done with less). However, Ockham's razor appears to be an example of Stigler's law of eponymy, which says that no scientific discovery is named after its original discoverer. Indeed, the principle of parsimony already features in work by Aristotle and Ptolemy. The latter even stated "We consider it a good principle to explain the phenomena by the simplest hypotheses possible." Hence, it may be historically correct to speak not of Ockham's razor, but of Ptolemy's principle of parsimony. Regardless of nomenclature, what is important is that the marginal likelihood acts as an automatic Ockham's razor (Jefferys & Berger, 1992; Myung & Pitt, 1997): models are punished for making predictions that are needlessly flexible with respect to the observed data.

space, those parts where the likelihood is almost zero (i.e., where the predictions are false). These effects combine to lower the average or marginal likelihood. Thus, a rate model that has a prior $\theta \sim \text{Uniform}\big(0.5, 1\big)$ is simpler than a rate model with the prior $\theta \sim \text{Uniform}\big(0, 1\big)$.

A final important factor that influences model complexity is the functional form of the model parameters. Consider for instance two laws of psychophysics that each relate the objective intensity I of a stimulus (e.g., a sound, a flash of light) to its subjective experience $\Psi\left(I\right)$. The first, Fechner's law, states that $\Psi\left(I\right) = k \ln\left(I + \beta\right)$, so that experienced intensity is a negatively accelerating function of stimulus intensity. The second, Stevens' law, states that $\Psi\left(I\right) = kI^{\beta}$, so that experienced intensity can be a negatively or a positively accelerating function of stimulus intensity. Fechner's law and Stevens' law each have two parameters, k and β, but nonetheless Stevens' law is more complex, because it can capture more data patterns and is therefore more difficult to falsify than Fechner's law (Townsend, 1975; Myung & Pitt, 1997).

The marginal likelihood, by assessing the average quality of a model's predictions for the data at hand, automatically takes all of these considerations into account.

Exercises

Exercise 7.1.1 Suppose you construct a second model for the same data D. This model, $\mathcal{M}_y$, has a parameter ζ that can take on two values, ζ_1 and ζ_2. You assign prior probability mass $p(\zeta_1) = 0.3$ and $p(\zeta_2) = 0.7$. For these two values, the likelihoods are 0.002 and 0.003, respectively. Compute the marginal likelihood for $\mathcal{M}_y$.

Exercise 7.1.2 What is the relative support of the data D for $\mathcal{M}_y$ versus $\mathcal{M}_x$?

Exercise 7.1.3 Suppose you construct a third model for the same data D. This model, $\mathcal{M}_z$, has a parameter μ that can take on 5 values, $\mu_1, \mu_2, \ldots, \mu_5$ with equal prior probability. The likelihoods are $p(D \mid \mu_1) = 0.001$, $p(D \mid \mu_2) = 0.001$, $p(D \mid \mu_3) = 0.001$, $p(D \mid \mu_4) = 0.001$, and $p(D \mid \mu_5) = 0.006$. Note that the likelihood for μ_5 is twice as high as the best possible likelihood for $\mathcal{M}_y$ and $\mathcal{M}_x$. Calculate the marginal likelihood for $\mathcal{M}_z$. Do you prefer it over $\mathcal{M}_y$ and $\mathcal{M}_x$? What is the lesson here?

Exercise 7.1.4 Consider Bart and Lisa, who each get 100 euros to bet on the winner of the world cup soccer tournament. Bart decides to divide his money evenly over 10 candidate teams, including those from Brazil and Germany. Lisa divides her money over just two teams, betting 60 euros on the team from Brazil and 40 euros on the team from Germany. Now if either Brazil or Germany turn out to win the 2010 world cup, Lisa wins more money than Bart. Explain in what way this scenario is analogous to the computation of marginal likelihood.

Exercise 7.1.5 Holmes and Watson[2] are involved in a rather simple game of darts, in which, with each dart, the player tries to score as many points as possible. The maximum score per dart is 60, and the minimum score is 0 (when the dart lands outside the board). After 5 darts, Holmes scored {38, 10, 0, 0, 0} and Watson scored {20,20,20,18,16}. How do you determine who is the better player? Consider another game of darts, but now one of the players gets to throw 50 times instead of 5. Explain how this scenario shows the importance of averaging instead of maximizing in order to penalize complexity.

7.2 The Bayes factor

Marginal likelihood is a measure of absolute evidence, in the sense that it is an index of a single model's overall predictive performance. In model selection, however, one is specifically interested in *relative* evidence, that is, the comparison of predictive performance for one model versus another. This comparison is accomplished simply by dividing the marginal likelihoods, yielding a quantity known as the Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995):

$$BF_{12} = \frac{p(D \mid \mathcal{M}_1)}{p(D \mid \mathcal{M}_2)}. \tag{7.4}$$

Here, BF_{12} indicates the extent to which the data support $\mathcal{M}_1$ over $\mathcal{M}_2$, and as such it represents "the standard Bayesian solution to the hypothesis testing and

[2] We thank Wolf Vanpaemel for suggesting this example.

https://doi.org/10.1017/CBO9781139087759.009

model selection problems" (Lewis & Raftery, 1997, p. 648). For example, when $BF_{12} = 5$ the observed data are 5 times more likely to have occurred under $\mathcal{M}_1$ than under $\mathcal{M}_2$, and when $BF_{12} = 0.2$ the observed data are 5 times more likely to have occurred under $\mathcal{M}_2$ than under $\mathcal{M}_1$.[3]

Even though the Bayes factor has an unambiguous and continuous scale, calibrated by betting, it is sometimes useful to summarize the Bayes factor in terms of discrete categories of evidential strength. Jeffreys (1961, Appendix B) proposed the classification scheme shown in Table 7.1.[4] This set of labels facilitates scientific communication but should only be considered an approximate descriptive articulation of different standards of evidence.[5]

Table 7.1 Evidence categories for the Bayes factor BF_{12} (Jeffreys, 1961).

Bayes factor BF_{12}			Interpretation
	>	100	Extreme evidence for $\mathcal{M}_1$
30	–	100	Very strong evidence for $\mathcal{M}_1$
10	–	30	Strong evidence for $\mathcal{M}_1$
3	–	10	Moderate evidence for $\mathcal{M}_1$
1	–	3	Anecdotal evidence for $\mathcal{M}_1$
	1		No evidence
1/3	–	1	Anecdotal evidence for $\mathcal{M}_2$
1/10	–	1/3	Moderate evidence for $\mathcal{M}_2$
1/30	–	1/10	Strong evidence for $\mathcal{M}_2$
1/100	–	1/30	Very strong evidence for $\mathcal{M}_2$
	<	1/100	Extreme evidence for $\mathcal{M}_2$

To illustrate, consider again our binomial example of 9 correct responses out of 10 questions, and the test between two models for performance: guessing (i.e., $\mathcal{M}_1 : \theta = 0.5$) versus not guessing (i.e., $\mathcal{M}_2 : \theta \neq 0.5$). In order to calculate the Bayes factor we need to be explicit about what we mean with "not guessing", which corresponds to defining a prior for θ. Here we use the uniform distribution for θ as a prior, such that $p(\theta \mid \mathcal{M}_2) \sim \text{Uniform}\big(0, 1\big) = \text{Beta}\big(1, 1\big)$.[6] After having specified both $\mathcal{M}_1$ and $\mathcal{M}_2$ we can proceed to calculate the separate marginal likelihoods, and then divide these to obtain the Bayes factor.

The marginal likelihood for $\mathcal{M}_1$ is calculated simply by plugging in the value $\theta = 0.5$ in the binomial equation: $p(D \mid \mathcal{M}_1) = \binom{10}{9}\left(\frac{1}{2}\right)^{10}$. The marginal likelihood

[3] Note that $BF_{12} = 1/BF_{21}$.

[4] We replaced the labels "worth no more than a bare mention" with "anecdotal", "decisive" with "extreme", and "substantial" with "moderate".

[5] The fact that the labels are only approximate is aptly illustrated by Jeffreys himself when he describes a Bayes factor of 5.33 as "odds that would interest a gambler, but would be hardly worth more than a passing mention in a scientific paper" (Jeffreys, 1961, pp. 256-257).

[6] This distribution includes the point $\theta = 0.5$, which may seem odd. However, when we compute the marginal likelihood we integrate over the prior distribution, and the inclusion of any single point is inconsequential, as $\int_a^a f(x)\,\mathrm{d}x = 0$.

Box 7.2 **Bayes versus Fisher**

"The Bayesian method is comparative. It compares the probabilities of the observed event on the null hypothesis and on the alternatives to it. In this respect it is quite different from Fisher's approach which is absolute in the sense that it involves only a single consideration, the null hypothesis. All our uncertainty judgements should be comparative: there are no absolutes here. A striking illustration of this arises in legal trials. When a piece of evidence E is produced in a court investigating the guilt G or innocence I of the defendant, it is not enough merely to consider the probability of E assuming G; one must also contemplate the probability of E supposing I. In fact, the relevant quantity is the ratio of the two probabilities. Generally if evidence is produced to support some thesis, one must also consider the reasonableness of the evidence were the thesis false. Whenever courses of action are contemplated, it is not the merits or demerits of any course that matter, but only the comparison of these qualities with those of other courses." (Lindley, 1993, p. 25)

for model $\mathcal{M}_2$ is more difficult to calculate. As we have seen above, the marginal likelihood is obtained by averaging the likelihood over the prior parameter space, according to Equation 7.3. When we assume $p(\theta \mid \mathcal{M}_2) \sim \text{Beta}(1, 1)$, then Equation 7.3 simplifies to $p(D \mid \mathcal{M}_2) = 1/(n+1)$. Thus, in our binomial example, $BF_{12} = \binom{10}{9}\left(\frac{1}{2}\right)^{10}(n+1) \approx 0.107$. This means that the data are $1/0.107 \approx 9.3$ times more likely under M_2 than they are under M_1.

Exercises

Exercise 7.2.1 Suppose you entertain a set of three models, x, y, and z. Assume you know $BF_{xy} = 4$ and $BF_{xz} = 3$. What is BF_{zy}?

Exercise 7.2.2 Suppose the $BF_{ab} = 1,000,000$, such that the data are one million times more likely to have occurred under $\mathcal{M}_a$ than under $\mathcal{M}_b$. Give two arguments for why you may still believe that $\mathcal{M}_a$ provides an inadequate or incorrect account of the data.

7.3 Posterior model probabilities

The Bayes factor compares the predictive performance of one model versus another, for the data at hand. A complete assessment of relative model preference, however, also requires us to consider how plausible the models are a priori. For example, let $\mathcal{M}_1$ be the hypothesis "neutrinos can travel faster than the speed of light," and let $\mathcal{M}_2$ be the hypothesis "neutrinos cannot travel faster than the speed of light."

The first hypothesis has been described as rather unlikely. Drew Baden, chairman of the physics department at the University of Maryland, compared its plausibility to that of finding a flying carpet. In such cases, even a very large Bayes factor in favor of $\mathcal{M}_1$ may be insufficient to make us believe that $\mathcal{M}_1$ is more likely than $\mathcal{M}_2$.

Hence, the assessment of the relative plausibility of two models after having seen the data requires that we combine information from the models' predictive performance for the data under consideration with the models' a priori plausibility. Expressed more formally,

$$\frac{p(\mathcal{M}_1 \mid D)}{p(\mathcal{M}_2 \mid D)} = \frac{p(D \mid \mathcal{M}_1)}{p(D \mid \mathcal{M}_2)} \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}. \tag{7.5}$$

Or, in words,

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}. \tag{7.6}$$

This equation also yields another interpretation of the Bayes factor, namely as the change from prior odds $p(\mathcal{M}_1)/p(\mathcal{M}_2)$ to posterior odds $p(\mathcal{M}_1 \mid D)/p(\mathcal{M}_2 \mid D)$ that is brought about by the data. When the prior odds are 1, such that $\mathcal{M}_1$ and $\mathcal{M}_2$ are equally likely a priori, the Bayes factors can be converted to posterior probabilities $p(\mathcal{M}_1 \mid D) = BF_{12}/(BF_{12} + 1)$. This means that, for example, $BF_{12} = 2$ translates to $p(\mathcal{M}_1 \mid D) = 2/3$.

Exercises

Exercise 7.3.1 In this book, you have now encountered two qualitatively different kinds of priors. Briefly describe what they are.

Exercise 7.3.2 Consider one model, $\mathcal{M}_1$, that predicts a post-surgery survival rate by gender, age, weight, and history of smoking. A second model, $\mathcal{M}_2$, includes two additional predictors, namely body-mass index and fitness. We compute posterior model probabilities and find that $p(\mathcal{M}_1 \mid D) = .6$ and consequently $p(\mathcal{M}_2 \mid D) = .4$. For a patient Bob, $\mathcal{M}_1$ predicts a survival rate of 90%, and $\mathcal{M}_2$ predicts a survival rate of 80%. What is your prediction for Bob's probability of survival?

7.4 Advantages of the Bayesian approach

Bayesian hypothesis tests (i.e., Bayes factors and posterior model probabilities) implement an automatic Ockham's razor, describe the relative support or preference for a set of two (or more) candidate models, and can be used for model-averaged predictions. Here we highlight two additional advantages of Bayesian hypothesis tests that are of key importance to cognitive science.

First, Bayes factors can be used to obtain evidence in favor of the null hypothesis. Because theories and models often predict the absence of an effect, it is important

Box 7.3 Extraordinary claims require extraordinary evidence

This is quite possibly the single most underestimated maxim in current-day cognitive science. It was stated most eloquently by Scottish philosopher David Hume (1711–1776): "...no testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous, than the fact, which it endeavors to establish; and even in that case there is a mutual destruction of arguments, and the superior only gives us an assurance suitable to that degree of force, which remains, after deducting the inferior." The first real Bayesian, Pierre-Simon Laplace (1749–1827), formulated the same sentiment more concisely: "The weight of evidence for an extraordinary claim must be proportioned to its strangeness." American astronomer Carl Sagan (1934–1996) coined the exact phrase "extraordinary claims require extraordinary evidence." The maxim is incorporated in the Bayesian computation where prior odds are combined with the Bayes factor to yield posterior odds. In many studies in cognitive science, attention is focused almost exclusively on the evidence that the observed data provide for or against a hypothesis. However, even strong evidence may fail to make an implausible claim acceptable. The fact that prior plausibility is difficult to quantify "objectively" is a poor excuse for ignoring it altogether. Nevertheless, most Bayesian statisticians are content when they provide only the Bayes factor; each researcher is then free to multiply that Bayes factor by his or her own prior odds in order to arrive at the posterior estimate of relative model plausibility.

to be able to quantify evidence in support of such predictions (e.g., Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009). In the field of visual word recognition, for instance, the entry-opening theory (Forster, Mohan, & Hector, 2003) predicts that masked priming is absent for items that do not have a lexical representation. Another example from that literature concerns the work by Bowers, Vigliocco, and Haan (1998), who hypothesized that priming depends on abstract letter identities—hence, priming should be equally effective for words that look the same in lower case and upper case (e.g., kiss/KISS) or different (e.g., edge/EDGE). A final example comes from the field of recognition memory, where Dennis and Humphreys' Bind Cue Decide model of episodic MEMory (BCDMEM) predicts the absence of a list-length effect and the absence of a list-strength effect (Dennis & Humphreys, 2001). In contrast to p-value hypothesis testing, Bayesian statistics assigns no special status to the null hypothesis and this means that Bayes factors can be used to quantify evidence for the null hypothesis just as for any other hypothesis.

A second advantage of Bayes factors is that they allow one to monitor the evidence as the data come in (Berger & Berry, 1988). In Bayesian hypothesis testing, "the

Box 7.4 Problems with p-values

This is a Bayesian book, and its focus is not on the deficiencies of p-values. But we will say that p-values are often misinterpreted, that they cannot quantify evidence in favor of the null hypothesis, that they depend on the (possibly unknown) intention with which the researcher collected the data, and that they focus only on what is expected under the null hypothesis, thus ignoring altogether what is expected under the alternative hypothesis. For scholarly details, see Berger and Wolpert (1988); Dennis et al. (2008); Dienes (2011); Edwards et al. (1963); Lindley (1993); Sellke et al. (2001); Wagenmakers (2007); Wagenmakers et al. (2008). For slogans, we like the following:

"The most important conclusion is that, for testing 'precise' hypotheses, p values should not be used directly, because they are too easily misinterpreted. The standard approach in teaching—of stressing the formal definition of a p value while warning against its misinterpretation—has simply been an abysmal failure." (Sellke et al., 2001, p. 71)

"Bayesian procedures can strengthen a null hypothesis, not only weaken it, whereas classical theory is curiously asymmetric. If the null hypothesis is classically rejected, the alternative hypothesis is willingly embraced, but if the null hypothesis is not rejected, it remains in a kind of limbo of suspended disbelief." (Edwards et al., 1963, p. 235)

"What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure." (Jeffreys, 1961, p. 385)

rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" (Edwards et al., 1963, p. 193). This means that researchers are free to continue data collection in case the evidence (i.e., the Bayes factor) is inconclusive. Likewise, they are free to terminate data collection as soon as the interim evidence is sufficiently compelling. This freedom of "optional stopping" is denied to the researchers who use p-values to test their hypotheses (Wagenmakers, 2007).

7.5 Challenges for the Bayesian approach

Bayesian hypothesis testing comes with two main challenges, one conceptual and one computational. The conceptual challenge arises because the Bayesian hypothesis test is sensitive to the prior distributions for the model parameters (e.g., Bartlett, 1957; Liu & Aitkin, 2008; Vanpaemel, 2010). This occurs because the marginal likelihood is an average taken with respect to the prior. For example, consider a test for the mean μ of a Gaussian distribution with known variance. The null hypothesis $\mathcal{H}_0$ states that μ is zero. Specification of the alternative hypothesis $\mathcal{H}_1$ requires that we assign μ a prior distribution; in other words, we need to quantify $p(\mu \mid \mathcal{H}_1)$, our uncertainty about μ under $\mathcal{H}_1$. One tempting option is to use an "uninformative" prior for μ that does not express much preference for one value of μ over the other. For example, one could use a Gaussian distribution with mean zero and variance 10,000. This approach with uninformative priors often works well for parameter estimation. From a marginal likelihood perspective, however, the use of a low-precision prior effectively creates a model that predicts almost any observed result. When one hedges one's bets to such an extreme degree, the Bayes factor is likely to show a preference for $\mathcal{H}_0$, even when the data appear inconsistent with it.

The problem is not that the Bayesian hypothesis test is sensitive to the prior distribution. This feature merely reflects the workings of the automatic Ockham's razor that is an asset, not a liability, of the Bayesian hypothesis test. The prior distributions are part of the model specification, and low-precision priors correspond to complex models.[7] Instead, the problem is that researchers sometimes only have a vague idea about the vagueness of their prior knowledge, or the relevant available prior information. When the vagueness of the prior is more or less arbitrary, so are the results from the Bayesian hypothesis test.

Several procedures have been proposed to ensure that the results from the Bayesian hypothesis test do not simply reflect the arbitrary precision of the prior distributions. First, one can invest more effort in the *subjective* specification of prior distributions (e.g., Dienes, 2011). This means that the researcher attempts to translate substantive knowledge about the problem at hand into prior probability distributions. Such knowledge may be obtained by eliciting prior beliefs from experts, or by consulting the literature for earlier work on similar problems. Unfortunately, the substantive knowledge that is encoded in the prior distributions does not generalize well to other problems, and consequently each new problem requires its own prior elicitation process. Most researchers have neither the expertise nor the energy to carry out the careful problem-specific elicitation steps that define the subjective approach. In addition, some modeling problems are so large and complex that they defy a careful subjective specification of the prior distribution.

[7] As an aside, model selection methods that are insensitive to prior distributions also have difficulties dealing with order-restricted inference, such as when the complex model has parameters θ_1 and θ_2 free to vary, and the simpler model has the order-restriction $\theta_1 > \theta_2$ (e.g., Hoijtink, Klugkist, & Boelen, 2008).

Box 7.5 Confusion about priors

A persistent confusion when the results from a Bayesian hypothesis test are discussed involves the supposedly arbitrary and profound impact of "the prior". For example, one researcher may conduct an experiment and find that the presence of a big box fails to make people more creative (even though, with a box present, people are perhaps encouraged to "think outside the box"). To quantify the evidence that the data provide in favor of $\mathcal{H}_0$ this researcher may present a Bayes factor, say, $BF_{01} = 15.5$. Invariably, another researcher will object that this Bayesian result depends on the prior plausibility that was assigned to $\mathcal{H}_1$. Obviously, or so the argument goes, when you are skeptical about $\mathcal{H}_1$ you will assign $\mathcal{H}_1$ a low prior plausibility, and hence the end result of the Bayesian test simply reaffirms your initial bias. This argument is false, and reveals a misunderstanding of what it is that the Bayes factor measures. As is clear from Equations 7.4 and 7.5, the Bayes factor does *not* involve the prior probabilities on the models. One researcher may believe that the big-box hypothesis is silly, and another may believe it is entirely plausible, but these different prior opinions about the model's plausibility do not affect the Bayes factor. There is another kind of prior, however, that does influence the Bayes factor. This is not the prior on the models, but the prior distribution for the relevant parameters. This prior distribution reflects our uncertainty about the size of the effect in case $\mathcal{H}_1$ is true. To set this prior one can use a default specification, a subjective specification, and carry out a sensitivity analysis. It is important to understand the difference between the prior on the models and the prior distribution on the models' parameters.

Second, one can try to use formal rules and desiderata to specify prior distributions that yield reasonable results across a wide range of different research contexts (Kass & Wasserman, 1996). Such priors are called *objective* because they do not depend on information specific to the research topic under investigation. For example, one can use a unit-information prior, that is, a prior that contains as much information as a single observation.[8] Similar objective prior distributions have been developed by Jeffreys (1961), Zellner and Siow (1980), Liang, Paulo, Molina, Clyde, and Berger (2008), and others. The results from such objective hypothesis tests may not be definitive, but these tests can nevertheless serve as a good reference analysis that can later be refined, if necessary, by the inclusion of problem-specific information.

[8] This assumption also underlies the popular Bayesian information criterion (BIC: Schwarz, 1978). Masson (2011) provides a tutorial on how to use the BIC for statistical problems such as ANOVA.

Box 7.6 Prior sensitivity

"When different reasonable priors yield substantially different answers, can it be right to state that there is a single answer? Would it not be better to admit that there is scientific uncertainty, with the conclusion depending on prior beliefs?" (Berger, 1985, p. 125)

Third, one can use sophisticated procedures such as the local Bayes factor (A. F. M. Smith & Spiegelhalter, 1980), the intrinsic Bayes factor (Berger & Mortera, 1999; Berger & Pericchi, 1996), the fractional Bayes factor (O'Hagan, 1995), and the partial Bayes factor (O'Hagan, 1995). Gill (2002, Chapter 7) provides a summary of this class of methods. The idea of the partial Bayes factor is to sacrifice a small part of the data to obtain a posterior that is relatively insensitive to the various priors one might entertain. The Bayes factor is then calculated by integrating the likelihood over this posterior instead of over the original prior. Procedures such as these are still undergoing further development and deserve more study.

Fourth, the dependence of the results on the width of the prior can be studied explicitly, by means of a sensitivity analysis. In such an analysis one varies the width of the prior distribution (across a reasonable range) and studies the corresponding fluctuations in the Bayes factor. Whenever these fluctuations cause meaningful, qualitative differences in conclusions one should acknowledge that the interpretation of the data is strongly dependent on prior beliefs, and that additional data may need to be collected before inference is robust across plausible prior beliefs.

The computational challenge for Bayesian hypothesis testing is that the marginal likelihood and the Bayes factor are often quite difficult to calculate. Earlier, we saw that with a uniform prior on the binomial rate parameter θ—$p(\theta \mid \mathcal{M}_1) \sim \text{Beta}\big(1,1\big)$—the marginal likelihood simplifies from $\int p(D \mid \theta, \mathcal{M}_1) p\,(\theta \mid \mathcal{M}_1)\, \mathrm{d}\theta$ to $1/\,(1+n)$.[9] However, in all but a few simple models, such simplifications are impossible. In order to be able to compute the marginal likelihood or the Bayes factor for more complex models, a range of computational methods have been developed. A recent summary lists as many as 15 different methods (Gamerman & Lopes, 2006, Chapter 7).

For example, one method computes the marginal likelihood by means of the *candidates' formula* (Besag, 1989) or the *basic marginal likelihood identity* (Chib, 1995; Chib & Jeliazkov, 2001). One simply exchanges the roles of posterior and marginal likelihood in Equation 7.1 to obtain

$$p(D \mid \mathcal{M}_1) = \frac{p\,(D \mid \theta, \mathcal{M}_1)\, p\,(\theta \mid \mathcal{M}_1)}{p\,(\theta \mid D, \mathcal{M}_1)}, \tag{7.7}$$

[9] This becomes intuitively clear from an inspection of the prior predictive distribution in the lower panel of Figure 3.9.

which holds for any one value of θ. When the posterior is available analytically, one only needs to plug in a single value of θ and obtain the marginal likelihood immediately. This method can however also be applied when the posterior is only available through MCMC output, either from the Gibbs sampler (Chib, 1995) or the Metropolis–Hastings algorithm (Chib & Jeliazkov, 2001).

Another method that computes the marginal likelihood is to sample repeatedly parameter values from the prior, calculate the associated likelihoods, and then take the likelihood average. When the posterior is highly peaked compared to the prior—as will happen with many data or with a medium-sized parameter space—it becomes necessary to employ more efficient sampling methods, with a concomitant increase in computational complexity.

Finally, it is also possible to compute the Bayes factor directly, without first calculating the constituent marginal likelihoods. The basic idea is to generalize the MCMC sampling routines for parameter estimation to incorporate a "model indicator" variable. In the case of two competing models, the model indicator variable z, say, can take on two values. For example, it can take $z = 1$ when the sampler is in model M_1, and $z = 2$ when the sampler is in model M_2. The Bayes factor is then estimated by the relative frequency with which $z = 1$ versus $z = 2$. This MCMC approach to model selection is called transdimensional MCMC (e.g., Sisson, 2005), an approach that encompasses both reversible jump MCMC (P. J. Green, 1995) and the product space technique (Carlin & Chib, 1995; Lodewyckx et al., 2011; Scheibehenne, Rieskamp, & Wagenmakers, 2013).[10]

Almost all of these computational methods suffer from the fact that they become less efficient and more difficult to implement as the underlying models become more complex. We now turn to an alternative method, whose implementation is extremely straightforward. The method's main limitation is that it applies only to *nested* models, a limitation that also holds for p-values.

7.6 The Savage–Dickey method

In the simplest classical hypothesis testing framework, one contemplates two models. One is the null hypothesis that fixes one of its parameters to a pre-specified value of substantive interest, say $\mathcal{H}_0 : \phi = \phi_0$; the other model is the alternative hypothesis, in which that parameter is free to vary, say $\mathcal{H}_1 : \phi \neq \phi_0$. Hence, the null hypothesis is nested under the alternative hypothesis, that is, $\mathcal{H}_0$ can be obtained from H_1 by setting ϕ equal to ϕ_0. Note that in the classical framework, $\mathcal{H}_0$ is generally a sharp null hypothesis, or a "point null". That is, the null hypothesis states that ϕ is exactly equal to ϕ_0.

[10] We recommend the Lodewyckx et al. article and accompanying software to anyone who wants to learn how to apply transdimensional MCMC techniques in WinBUGS or JAGS. In this book we focus on the Savage–Dickey technique because it is easier to understand and implement.

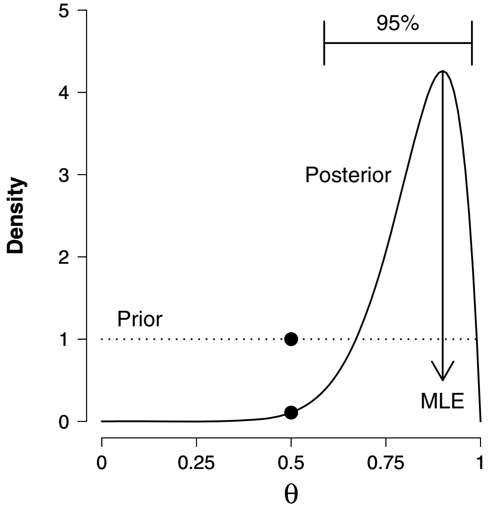

Fig. 7.1 Prior and posterior distributions for binomial rate parameter θ, after observing 9 correct responses and 1 incorrect response. The mode of the posterior distribution for θ is 0.9, equal to the maximum likelihood estimate, and the 95% credible interval extends from 0.59 to 0.98. The height of the distributions at $\theta = 0.5$ is indicated by a black dot; the ratio of these heights quantifies the evidence for $\mathcal{H}_0 : \theta = 0.5$ versus $\mathcal{H}_1 : \theta \sim \text{Beta}(1, 1)$.

For example, in our binomial example you answered 9 out of 10 questions correctly. Were you guessing or not? The Bayesian and the frequentist framework define $\mathcal{H}_0 : \theta = 0.5$ as the null hypothesis for chance performance. The alternative hypothesis under which $\mathcal{H}_0$ is nested could be defined as $\mathcal{H}_1 : \theta \neq .5$, or, more specifically, as $\mathcal{H}_1 : \theta \sim \text{Beta}(1, 1)$, which states that θ is free to vary from 0 to 1, and that it has a uniform prior distribution as shown in Figure 7.1.

For the binomial example, the Bayes factor for $\mathcal{H}_0$ versus $\mathcal{H}_1$ could be obtained by analytically integrating out the model parameter θ. However, the Bayes factor may likewise be obtained by only considering $\mathcal{H}_1$, and dividing the height of the posterior for θ by the height of the prior for θ, at the point of interest. This surprising result was first published by Dickey and Lientz (1970), who attributed it to Leonard J. "Jimmie" Savage. The result is now generally known as the *Savage–Dickey density ratio* (e.g., Dickey, 1971); for extensions and generalizations, see Chen (2005), Verdinelli and Wasserman (1995), and Wetzels, Grasman, and Wagenmakers (2010). Mathematically, the Savage–Dickey density ratio says that

$$BF_{01} = \frac{p(D \mid \mathcal{H}_0)}{p(D \mid \mathcal{H}_1)} = \frac{p(\theta = 0.5 \mid D, \mathcal{H}_1)}{p(\theta = 0.5 \mid \mathcal{H}_1)}. \qquad (7.8)$$

A straightforward mathematical proof is presented in O'Hagan and Forster (2004, pp. 174–177).

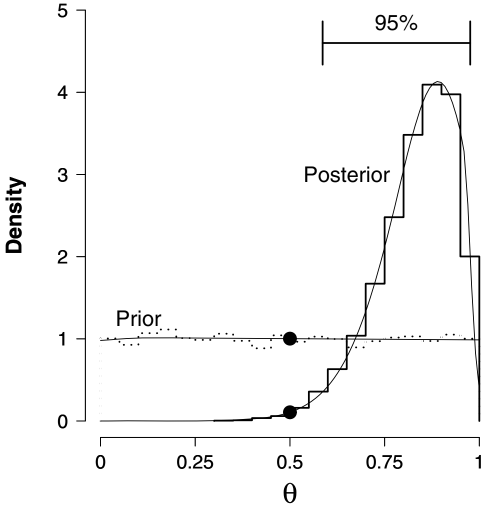

Fig. 7.2 MCMC-based prior and posterior distributions for the binomial rate parameter θ, after observing 9 correct responses and 1 incorrect response. The thin solid lines indicate the fit of a non-parametric density estimator. Based on this density estimator, the mode of the posterior distribution for θ is approximately 0.89, and the 95% credible interval extends from 0.59 to 0.98, closely matching the analytical results from Figure 7.1.

In Figure 7.1, the two thick dots located at $\theta = .5$ provide the required information. It is evident from the figure that after observing 9 out of 10 correct responses, the height of the density at $\theta = 0.5$ has decreased, so that one would expect these data to cast doubt on the null hypothesis and support the alternative hypothesis. Specifically, the height of the prior distribution at $\theta = 0.5$ equals 1, and the height of the posterior distribution at $\theta = 0.5$ equals 0.107. From Equation 7.8 the corresponding Bayes factor is $BF_{01} = 0.107/1 = 0.107$, and this corresponds exactly to the Bayes factor that was calculated by integrating out θ.

It is clear that the same procedure can be followed when the height of the posterior is not available in closed form, but instead has to be estimated from the histogram of MCMC samples. Figure 7.2 shows the estimates for the prior and the posterior densities as obtained from MCMC output (Stone, Hansen, Kooperberg, & Truong, 1997). The estimated height of the prior and posterior distributions at $\theta = 0.5$ equal 1.00 and 0.107, respectively.

In most nested model comparisons, $\mathcal{H}_0$ and $\mathcal{H}_1$ have several free parameters in common. These parameters are usually not of direct interest, and they are not the focus of the hypothesis test. Hence, the common parameters are known as *nuisance parameters*. For example, one might want to test whether or not the mean of a Gaussian distribution is zero—$\mathcal{H}_0 : \mu = \mu_0$ versus $\mathcal{H}_1 : \mu \neq \mu_0$—whereas the variance σ^2 is common to both models and not of immediate interest.

In general then, the framework of nested models features a parameter vector $\theta = (\phi, \psi)$, where ϕ denotes the parameter of substantive interest that is subject to test, and ψ denotes the set of nuisance parameters. The null hypothesis $\mathcal{H}_0$ posits that ϕ is constrained to some special value, so that $\phi = \phi_0$. The alternative hypothesis $\mathcal{H}_1$ assumes that ϕ is free to vary. Now consider $\mathcal{H}_1$, and let $\phi \to \phi_0$. This effectively means that $\mathcal{H}_1$ reduces to $\mathcal{H}_0$, and it is therefore reasonable to assume that $p\left(\psi \mid \phi \to \phi_0, \mathcal{H}_1\right) = p\left(\psi \mid \mathcal{H}_0\right)$. In other words, when $\phi \to \phi_0$ the prior for the nuisance parameters under $\mathcal{H}_1$ should equal the prior for the nuisance parameters under $\mathcal{H}_0$. When this condition holds, the nuisance parameters can be ignored, so that again

$$BF_{01} = \frac{p(D \mid \mathcal{H}_0)}{p\left(D \mid \mathcal{H}_1\right)} = \frac{p\left(\phi = \phi_0 \mid D, \mathcal{H}_1\right)}{p\left(\phi = \phi_0 \mid \mathcal{H}_1\right)}, \tag{7.9}$$

which equals the ratio of the heights for the posterior and the prior distribution for ϕ at ϕ_0. Thus, the Savage–Dickey density ratio holds under relatively general conditions. The next chapters use concrete examples to illustrate how cognitive scientists can use the Savage–Dickey density ratio test to their advantage.

Exercises

Exercise 7.6.1 The Bayes factor is relatively sensitive to the width of the prior distributions for the model parameters. Use Equation 7.9 to argue why this is the case.

Exercise 7.6.2 The Bayes factor is relatively sensitive to the width of the prior distributions, but only for the parameters that differ between the models under consideration. Use Equation 7.9 to argue why this is the case.

Exercise 7.6.3 What is the main advantage of the Savage–Dickey procedure?

7.7 Disclaimer and summary

The material covered in this chapter is controversial. Several professional Bayesian statisticians advise against the use of Bayes factors and instead recommend methods based on an assessment of the posterior distribution. And indeed, Bayes factors should be used with care, as their interpretation stands or falls with the plausibility of the models under comparison. For example, is it ever plausible to assume the complete absence of an effect? In other words, can the null hypothesis $\mathcal{H}_0$ ever be exactly true? If one does not believe that such a point null hypothesis can ever be true, even as an approximation, than the entire comparison between $\mathcal{H}_0$ and $\mathcal{H}_1$ may become meaningless (but see Berger & Delampady, 1987). We believe that, at least for experimental studies, point null hypotheses may often be true exactly. Homeopathy does not cure cancer, people cannot look into the future, and we doubt that people are any more creative when they stand next to a big box (but see Leung et al., 2012).

The most prominent argument against Bayes factors is that they depend largely on the specification of the prior distributions. This argument is true, but it presupposes that the specification of a model is somehow separate from the specification of the prior distributions. The Bayes factor perspective on model selection is that, before models can be compared, they need to be specified completely, and this includes the number of parameters, their prior distributions, and their functional form. All of these properties jointly determine an essential characteristic of a model, which is its ability to generate predictions. The evidence that the data provide for and against a model can only be properly assessed when one is able to discount the ability of that model to fit all kinds of other data as well.

In sum, the Bayes factor should be used with care, preferably in combination with other methods and a sensitivity analysis. That said, the Bayes factor has a number of undeniable advantages, some of which we have already discussed and others that will become evident in later chapters. One general advantage is that the Bayes factor directly addresses the key question that cognitive scientists care about: "To what extent do my data support $\mathcal{H}_1$ over $\mathcal{H}_0$?" Another general advantage is that the Bayes factor follows from the basic tenets of probability theory. Not only does this impart to the Bayes factor all kinds of pleasant properties—including consistency, transitivity, and immunity against optional stopping—it also means that alternative methods necessarily violate these basic tenets. Consequently, although alternative methods may work well for some situations, it is always possible to find situations in which these alternative methods fail and the Bayes factor succeeds.

Model selection and hypothesis testing are difficult topics, and it may take you a while to grasp the concepts involved. The next chapters provide concrete examples that show Bayesian model selection in action. From the current chapter, it is important that you understand the following key points:

- Complex models are models that make many predictions. This may happen because they have many parameters, because they have prior parameter distributions that are relatively non-precise and spread out over a wide range, or because they have parameters that have a complicated functional form. Complex models are difficult to falsify.
- The Bayes factor penalizes models for needless complexity and therefore it implements Ptolemy's principle of parsimony.
- The Bayes factor provides a comparative measure of evidence as it pits the predictive adequacy of one model against that of another.
- The Bayes factor requires a careful selection of prior distributions, as these form an integral part of the model specification.
- Extraordinary claims require extraordinary evidence.

Exercise

Exercise 7.7.1 Browse the empirical literature of your subfield of study. Do you find the null hypotheses plausible? That is, could they ever be exactly true?