CHAPTER

7

Multimodel Inference

OUTLINE

Inference about ecological processes is almost inevitably model based. No matter how much planning has gone into our investigation prior to collecting data, no matter how carefully we have designed our study, no matter how familiar we are with the system we are studying, there comes a point where we must describe our observations using a mathematical model. The model will have components related to the processes we are studying, and components related to the acquisition of data. Some of these components will be structural, amenable to computation given knowledge of covariates and parameters, other components will be stochastic, describable only as random noise. The model is fully specified, except for unknown parameters estimable from our data.

The process of inference often goes no further than estimating unknown parameters. There is no acknowledgment of model uncertainty, save perhaps a "goodness of fit" test to see whether our observations are consistent with the model.[1] Although we might not always be careful to mention it, we know that our inference is conditional on the model chosen, and might have been different had we posited a different model.

In practice there are usually several, even many plausible models that we could consider for our data. Each has distinct structural features and stochastic components. We are uncertain about which components are necessary. Inference based on the selection of a single model may sweep this uncertainty under the carpet.

The choice of a model is particularly important when candidate models involve complicated structure, and many parameters. We distinguish "parameters of interest" and "nuisance parameters," the former being the objects of our inquiry, the latter being required only so as to avoid a distorted view of the parameters of interest. We do not wish to waste inferential resources by including unnecessary nuisance parameters, nor do we wish to risk being misled by neglecting necessary ones. We also wish to choose the "right" set of parameters of interest, and not some that merely happen to be correlated with the right set.

Instead of conditioning our inference on the choice of a single model, we might wish to include model uncertainty as part of the inferential process. The goal is to produce a composite inference reflecting the uncertainties within and between models. To give a concrete example, suppose that we have analyzed a single data set under K distinct models, obtaining estimated survival rates $\hat{\phi}_k$ and associated standard errors $s(\hat{\phi}_k)$. Suppose further that we can quantify our relative confidence in the models, so as to produce model weights $w_k \geq 0$ satisfying $\sum w_k = 1$. Buckland *et al.* (1997) suggested combining model specific estimates using

$$\tilde{\phi} = \sum_{k=1}^{K} w_k \hat{\phi}_k \tag{7.1}$$

and a composite measure of uncertainty

$$s(\tilde{\phi}) = \sum_{k=1}^{K} w_k \sqrt{s(\hat{\phi}_k)^2 + (\hat{\phi}_k - \tilde{\phi})^2}.$$

Estimates (and standard errors) 0.74(0.13), 0.70(0.12), 0.76(0.10), and 0.65(0.08) with weights 0.4, 0.3, 0.2, and 0.1 are thus combined to $\tilde{\phi} = 0.72$ and $s(\tilde{\phi}) = 0.12$. The idea is that $\tilde{\phi}$ is an estimate based on all of the models, and that $s(\tilde{\phi})$ incorporates uncertainties within and between models.

The two problems of multimodel inference are thus model selection, and model weighting. Model selection is an attempt to choose a best model from a set of candidates; inference is then conditioned on that selection. Model weighting attempts to combine model specific inferences in a way which acknowledges the relative degree of trust we place in models, as well as the uncertainties associated with model specific inferences.

1. Goodness of fit tests are usually conducted with fingers crossed and in fervent hope that the p-value will come out larger than 0.05. In which case, the null hypothesis of model adequacy is treated as having been established – in flat contradiction to the philosophy of hypothesis testing.

Unfortunately, there is currently no consensus on how one ought to acknowledge and account for such uncertainty ... and there probably never will be.[2] There are many competing ideas and methods in the literature, enough to leave the practitioner boggled, astounded, and disheartened. The Bayesian approach to multimodel inference does not, alas, finally settle all of the complicated issues surrounding the topic. In some measure, multimodel inference must remain at the intersection of art and science.

However, Bayesian multimodel inference (BMI) has a strong philosophical appeal; like Bayesian inference generally, it retains the features of simplicity, exactness, and coherency described in Chapter 1. Indeed, BMI is a very natural extension of the basic Bayesian technique: one makes inference about unknown quantities (in this case, *models*) based on their posterior distributions, given data. Posterior model probabilities are used for combining model-specific estimates in the spirit of Eq. (7.1), and to combine model-specific inferences. And as for model selection, if a single model is desired, posterior model probabilities provide an objective basis for choice.

In this chapter we provide an overview of BMI, with comments on model weights, Bayes factors, the Bayesian information criterion (BIC), and the deviance information criterion (DIC). We also discuss computational issues, describing reversible jump Markov chain Monte Carlo (RJMCMC) and simple implementations of BMI in program BUGS.

One of the challenges for BMI, perhaps the most serious, is the selection of vague priors for parameters. When dealing with a single model, given adequate data the choice of vague prior has little influence on inference. Unfortunately this is not the case in the multimodel setting. We describe the problems and some possible solutions.

In recent years, Akaike's information criterion (AIC) has been heavily promoted among wildlife statisticians (Burnham and Anderson, 2002) as a basis for model selection and model weighting. AIC has taken the field by storm, and has been uncritically accepted by many practitioners. Its performance is reasonably evaluated in a Bayesian context; we do so in Sections 7.4.1 and 7.4.3.

7.1 THE BMI MODEL

BMI is really no different than any other sort of Bayesian inference: all quantities are treated as random, the only distinction being between quantities that are known or unknown. Inference is made using posterior distributions of unknown quantities given known quantities.

Thus BMI begins by considering a random variable *Model*, drawn from some collection called the *model set*. We might conceive of Nature blindfolded, making a single draw from a bucket containing K models (Fig. 7.1). If the selected model is fully specified (i.e., there are no unknown parameters in the model specification), Nature generates *Data* according to *Model*; if there are unknown parameters, Nature draws parameters for the particular model from a prior distribution specific to *Model*, and then generates *Data*.

Much of statistical inference is done pretending that we know with certainty which model Nature has drawn. Multimodel inference acknowledges model uncertainty, supposing that *Model* could have been any one in the model set.

2. Nor perhaps *should* there be, for reasons to be explained at the end of this chapter.

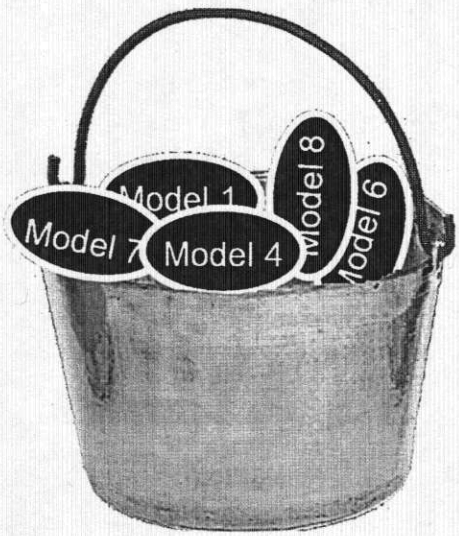

FIGURE 7.1 A bucket of models.

More formally, we conceive of *Model* as a multinomial random variable with index 1 and cell probabilities $\pi_1, \pi_2, \ldots, \pi_K$. If we have no a priori reason for favoring one model over another, we might set $\pi_1 = \pi_2 = \cdots = \pi_K = 1/K$; otherwise, we may choose prior model probabilities reflecting prior beliefs. These beliefs may reflect specific knowledge about the system studied and might also reflect convictions about desirable model features, such as parsimony.

Objections!

We anticipate two objections. First "that's not how things work! *Nature* drawing from a bucket?" But the multinomial draw is merely a model of our uncertainty; a mathematical convenience rather than an exact depiction of reality. The BMI framework is a meta-model, a model about models. Thus like all models, it need not be an exact depiction of reality, merely one that is useful. Subsequently, we describe the process of updating prior model probabilities to posterior model probabilities, based on information provided by data. The usefulness of this meta-model lies in its providing a mathematically formal and coherent system of assessing the support provided by data to various competing models.

Another objection: "But Truth isn't in your bucket!" Much unnecessary ink has been spilled on this topic, with declamations about Science and Truth and Knowledge that can leave our head spinning: is there such a thing as Truth? is it even possible for a Model to be Truth? But entertaining as such philosophical ramblings might be, they have no bearing on the issue at hand: it does not matter whether Truth is in the model set, or not, or whether there even *is* such a thing as Truth. Rather, BMI is conditional inference: we merely condition on Truth being in the model set, without any philosophical baggage. It is completely legitimate to say "I don't believe Truth is in that model set, but if you ask me to pretend it is, I can assign such and such probability to Truth being *this* particular

model, or *that* particular model."[3] Model probabilities are never unconditional, but always conditional on the Model Set. So though we may write $\pi_i = \Pr(\textit{Model } i)$, what we really mean is

$$\pi_i = \Pr(\textit{Model } i | \textit{Model } 1 \text{ or } \textit{Model } 2 \text{ or } \cdots \text{ or } \textit{Model } \mathrm{K}).$$

The bottom line? The bucket of models is *itself* a model, and no more dubious than any other model. It is merely a mathematical convenience to describe our uncertainty.

7.1.1 Example: BMI for Two Fully Specified Models

A geometric random variable has *pdf* $g(y) = p(1-p)^y$ and mean value $(1-p)/p$; a Poisson random variable has *pdf* $f(x) = e^{-\lambda}\lambda^y/y!$ and mean λ. Both take values $y = 0, 1, 2, \ldots$.

Suppose that we have a sample $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_5\}$ of values which either come from a geometric distribution ($M = M_1$) or a Poisson distribution ($M = M_2$). Here, M is a categorical random variable describing Nature's multinomial choice of model. Then the probability of the data is either

$$\Pr(\boldsymbol{Y}|M_1, p) = \prod_{i=1}^{5} g(Y_i) = p^5(1-p)^{5\bar{Y}}. \tag{7.2}$$

or

$$\Pr(\boldsymbol{Y}|M_2, \lambda) = \prod_{i=1}^{5} f(Y_i) = \frac{\exp(-5\lambda)\lambda^{5\bar{Y}}}{\prod_{i=1}^{5} Y_i!}. \tag{7.3}$$

Strictly speaking, we should write the probabilities as conditional on $M = M_i$, as for example $\Pr(\boldsymbol{Y}|M = M_1, p)$, emphasizing that we are treating M as a random variable, with M_1 being an outcome; ease of notation and convention favor the simpler notation.

To make things easy, suppose that we know the population mean is 3 (i.e., either $\lambda = 3$ if the data come from a Poisson distribution, or $p = 1/4$, if the data come from a geometric distribution). The two densities are displayed in Fig. 7.2. Substituting $\lambda = 3$ and $p = 1/4$ in (7.2) and (7.3), we obtain $\Pr(\boldsymbol{Y}|M_i)$ for $i = 1, 2$. Note that the models are fully specified, hence we do not need to include p or λ in the conditional description.

Given prior probability $\pi = \Pr(M_1)$, straightforward application of Bayes' theorem yields posterior probabilities

$$\Pr(M_1|\boldsymbol{Y}) = \frac{\pi \Pr(\boldsymbol{Y}|M_1)}{\pi \Pr(\boldsymbol{Y}|M_1) + (1-\pi)\Pr(\boldsymbol{Y}|M_2)}. \tag{7.4}$$

Consider the data set $\boldsymbol{Y} = \{0, 1, 2, 3, 8\}$. The sample mean is 2.8, consistent with both models, but the sample variance is $s^2 = 9.7$. Given that the Poisson mean and variance are equal, the data would seem to favor the geometric model, which allows for greater variability relative to the mean (compare Fig. 7.2.) If we had no a priori reason to favor the Poisson over the geometric model, we would probably set $\pi = 0.5$ in Eq. (7.4) and obtain posterior probability

3. To illustrate: imagine an urn with 99,999,997 red marbles, 2 blue marbles, and 1 white marble. If I were to draw a marble at random, and report that it was not red, you might not believe me, but could still assign odds to whether it were blue or white, conditional on my claim.

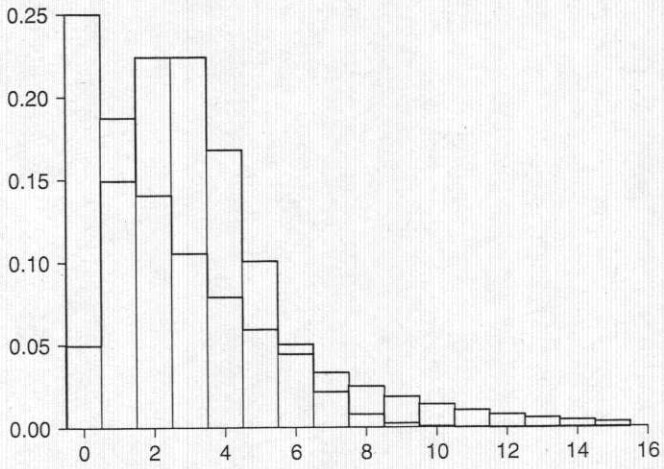

FIGURE 7.2 Poisson (blue) and geometric (red) densities, means = 3.

of 0.852 on the geometric model: the data have shifted the odds from being even (0.5:0.5) or (1:1) to odds of (0.852:0.148) or (5.75:1) in favor of the geometric model.

This change in odds is a very useful summary in BMI. Note that Eq. (7.4) implies

$$\Pr(M_2|\boldsymbol{Y}) = \frac{(1-\pi)\Pr(\boldsymbol{Y}|M_2)}{\pi\Pr(\boldsymbol{Y}|M_1)+(1-\pi)\Pr(\boldsymbol{Y}|M_2)}. \tag{7.5}$$

Thus, the ratio of the left-hand sides of Eqs. (7.4) and (7.5) equals the ratio of their right-hand sides, namely

$$\frac{\Pr(M_1|\boldsymbol{Y})}{\Pr(M_2|\boldsymbol{Y})} = \left(\frac{\pi}{1-\pi}\right) \times \frac{\Pr(\boldsymbol{Y}|M_1)}{\Pr(\boldsymbol{Y}|M_2)}. \tag{7.6}$$

Equation (7.6) relates the posterior model odds (left-hand side) to the prior model odds $\pi/(1-\pi)$. Prior model odds are scaled by the relative probabilities of the data, under the two models.

As a mathematical expression, $\Pr(\boldsymbol{Y}|M)$ is a function of the data, $\boldsymbol{Y}$, and of the model, M. Restricting our attention to a fixed set of data, $\Pr(\boldsymbol{Y}|M)$ is a function of M alone, the likelihood function for the model. The ratio of model likelihoods in Eq. (7.6) is called the *Bayes factor*. We may thus put Eq. (7.6) into words:

$$\text{Posterior model odds} = \text{Prior model odds} \times \text{Bayes factor}.$$

For the example considered, prior odds of 1:1 were converted to posterior odds of 5.75:1, hence the Bayes factor is 5.75.

7.1.2 Example: BMI with Unknown Parameters

Suppose that instead of comparing a geometric distribution with known $p = 1/4$ to a Poisson distribution with known $\lambda = 3$, the choice was between an unknown geometric distribution

and an unknown Poisson distribution. To calculate the model likelihoods, we would need prior distributions for p and λ.

That is, instead of substituting specific values of p and λ in Eqs. (7.2) and (7.3), we must calculate the average probability of the data under each model, weighted by the prior distributions.

Given a prior distribution $g(p)$ for the parameter p of the geometric distribution, we would calculate

$$\Pr(\boldsymbol{Y}|M_1) = \int \Pr(\boldsymbol{Y}|M_1, p) g(p)\, dp = \int p^5 (1-p)^{5\bar{Y}} g(p)\, dp; \tag{7.7}$$

given a prior distribution $h(\lambda)$ for the mean λ of the Poisson distribution, we would calculate

$$\Pr(\boldsymbol{Y}|M_2) = \int \Pr(\boldsymbol{Y}|M_2, \lambda) h(\lambda)\, d\lambda = \int \frac{\exp(-5\lambda)\lambda^{5\bar{Y}}}{\prod_{i=1}^{5} Y_i!} h(\lambda)\, d\lambda. \tag{7.8}$$

Choice of prior distributions for parameters in BMI is a ticklish business, which we discuss subsequently. Suppose that we choose a $U(0,T)$ prior for the mean λ of the Poisson distribution. It seems reasonable then to choose a prior on p for the geometric distribution such that the mean $(1-p)/p$ also has a $U(0,T)$ distribution. It can be shown using the change of variables theorem (2.2.4) that we require a prior

$$g(p) = \frac{1}{Tp^2},$$

for $1/(T+1) < p < 1$.

For the data set $\boldsymbol{Y} = \{0, 1, 2, 3, 8\}$, it can be shown that as $T \to \infty$, the resulting Bayes factors in favor of the geometric model approach 13.84. It is natural to ask why the evidence in favor of the geometric model appears so much stronger when the mean is unknown, than in our previous analysis when the mean was known (mean = 3 implies BF favoring geometric model is 5.75).

Some intuition is gained by considering Fig. 7.3. It turns out that if the mean value is known, the evidence favoring the geometric model over the Poisson model is minimized when the

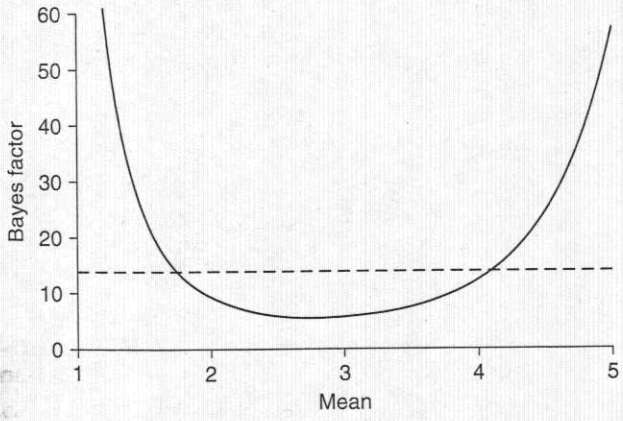

FIGURE 7.3 Bayes factors favoring geometric model over Poisson data as a function of known mean, for data $\boldsymbol{Y} = \{0,1,2,3,8\}$. Dashed line at 13.84 is Bayes factor for analysis with vague prior on mean.

true mean and the sample mean coincide, at $\bar{Y} = 2.80$. There, BF = 5.61, favoring the geometric model. If the mean were known to be a very small or very large value, the data would provide very strong evidence favoring the geometric model over the Poisson. Incorporating prior uncertainty about the mean essentially averages the evidence favoring the geometric model over values of the mean consistent with the observed data.

7.2 BAYES FACTORS

The Bayes factor provides a way of comparing pairs of competing models. The models need not be nested; neither need be a special case of the other, as seen in the example of Section 7.1.2. In this section, we describe some features of Bayes factors.

7.2.1 Bayes Factors and Likelihood Ratio Statistics

The Bayes factor is a likelihood ratio for models. Suppose that we have models M_k with unknown parameters θ_k, for $i = 1, 2, \ldots, K$. Using the bracket notation described in Section 2.2.1, we have data distributions $[\mathbf{Y}|M_k, \theta_k]$ for data and prior distributions $[\theta_k|M_k]$ for parameters. For a fixed data set $\mathbf{Y}$, $[\mathbf{Y}|M, \theta]$ is a joint likelihood for model and parameter. We obtain a marginal likelihood for M by integrating the joint likelihood against the prior for the parameter. That is,

$$[\mathbf{Y}|M] = \int [\mathbf{Y}, \theta|M] d\theta = \int [\mathbf{Y}|\theta, M][\theta|M] d\theta. \tag{7.9}$$

The Bayes factor for comparing M_i to M_j is calculated as the ratio of marginal likelihoods, averaged across the parameters:

$$\mathrm{BF}_{i,j} = \frac{[\mathbf{Y}|M_i]}{[\mathbf{Y}|M_j]}.$$

As such, the Bayes factor is a Bayesian analog of the frequentist likelihood ratio test statistic. The difference is that the likelihood ratio test statistic is the ratio of maximum likelihoods. That is, instead of averaging against a distribution for the unknown parameter [as in Eq. (7.9)], one substitutes a specific value, the MLE $\hat{\theta}_M$:

$$\mathrm{LR}_{ij} = \frac{[\mathbf{Y}|\hat{\theta}_i, M_i]}{[\mathbf{Y}|\hat{\theta}_j, M_j]}.$$

If the models are nested (i.e., model M_i is a special case of model M_j), and if M_i is true, -2 times the natural logarithm of the likelihood ratio statistic can often be treated as having an asymptotic chi-squared distribution.

7.2.2 Bayes Factors are Multipliers of Odds

Given a set of K models, prior model probabilities π_k, $k=1,2,\ldots,K$, and data $\boldsymbol{Y}$, Bayes' theorem yields

$$\Pr(M_i|\boldsymbol{Y}) = \frac{[\boldsymbol{Y}|M_i]\,\pi_i}{\sum_{k=1}^{K}[\boldsymbol{Y}|M_k]\,\pi_k} \tag{7.10}$$

from which it follows that

$$\begin{aligned}\frac{\Pr(M_i|\boldsymbol{Y})}{\Pr(M_j|\boldsymbol{Y})} &= \frac{[\boldsymbol{Y}|M_i]}{[\boldsymbol{Y}|M_j]} \times \left(\frac{\pi_i}{\pi_j}\right) \\ &= \mathrm{BF}_{i,j} \times \left(\frac{\pi_i}{\pi_j}\right),\end{aligned}$$

generalizing Eq. (7.6). That is, the Bayes factor is simply a multiplier for changing prior odds into posterior odds. In the special case of prior odds equal to 1, the Bayes factor itself is the posterior odds.

Bayes factors do not depend on the prior model probabilities; they also do not depend on the model set, being nothing more than pairwise comparisons of models. There is, however, a relation among Bayes factors within a model set. It is clear from their definition that

$$\mathrm{BF}_{1,3} = \mathrm{BF}_{1,2}\mathrm{BF}_{2,3}. \tag{7.11}$$

Similarly, if we know the Bayes factor for model 1 against model 2, $\mathrm{BF}_{1,2}$, the Bayes factor for model 2 against model 1 is $\mathrm{BF}_{2,1} = 1/\mathrm{BF}_{1,2}$.

Equation (7.10) is sometimes rewritten in terms of Bayes factors. Dividing the numerator and denominator of the right-hand side by $[\boldsymbol{Y}|M_1]$, we obtain

$$\Pr(M_i|\boldsymbol{Y}) = \frac{\mathrm{BF}_{i,1}\,\pi_i}{\sum_{k=1}^{K}\mathrm{BF}_{k,1}\,\pi_k}. \tag{7.12}$$

7.2.3 Updating Bayes Factors

A very appealing feature of Bayes factors is that they update naturally as more data are collected. Let $\mathrm{BF}_{i,j}(\boldsymbol{Y}_1)$ denote the Bayes factor based on data set $\boldsymbol{Y}_1$, and let $\mathrm{BF}_{i,j}(\boldsymbol{Y}_2|\boldsymbol{Y}_1)$ denote the Bayes factor based on $\boldsymbol{Y}_2$ having used $\boldsymbol{Y}_1$ to inform priors on unknown parameters.

Thus,

$$\mathrm{BF}_{i,j}(\boldsymbol{Y}_1) = \frac{[\boldsymbol{Y}_1|M_i]}{[\boldsymbol{Y}_1|M_j]} = \frac{\int[\boldsymbol{Y}_1|M_i,\theta]\,[\theta|M_i]\,d\theta}{\int[\boldsymbol{Y}_1|M_j,\theta]\,[\theta|M_j]\,d\theta},$$

and

$$\mathrm{BF}_{i,j}(\boldsymbol{Y}_2|\boldsymbol{Y}_1) = \frac{[\boldsymbol{Y}_2|\boldsymbol{Y}_1,M_i]}{[\boldsymbol{Y}_2|\boldsymbol{Y}_1,M_j]} = \frac{\int[\boldsymbol{Y}_2|\boldsymbol{Y}_1,M_i,\theta]\,[\theta|\boldsymbol{Y}_1,M_i]\,d\theta}{\int[\boldsymbol{Y}_2|\boldsymbol{Y}_1,M_j,\theta]\,[\theta|\boldsymbol{Y}_1,M_j]\,d\theta}.$$

Note that the priors on θ change from the first to the second of these calculations: in the second case, they have been informed by the data $\boldsymbol{Y}_1$.

Consequently

$$\mathrm{BF}_{i,j}(\boldsymbol{Y}_1,\boldsymbol{Y}_2) \equiv \frac{[\boldsymbol{Y}_2,\boldsymbol{Y}_1|M_i]}{[\boldsymbol{Y}_2,\boldsymbol{Y}_1|M_j]} = \frac{[\boldsymbol{Y}_2|\boldsymbol{Y}_1,M_i]}{[\boldsymbol{Y}_2|\boldsymbol{Y}_1,M_j]} \times \frac{[\boldsymbol{Y}_1|M_i]}{[\boldsymbol{Y}_1|M_j]}$$
$$= \mathrm{BF}_{i,j}(\boldsymbol{Y}_2|\boldsymbol{Y}_1) \times \mathrm{BF}_{i,j}(\boldsymbol{Y}_1).$$

It makes sense that there should be a simple mechanism for describing the accumulation of evidence in favor of one model over another, as new data are obtained. This feature is conspicuously absent from sequences of hypothesis tests under the frequentist paradigm, though various ad hoc mechanisms have been proposed.

7.2.4 Bayes Factors as Measures of Relative Support

The larger the value of $\mathrm{BF}_{i,j}$, the greater the support provided by the data to Model M_i relative to Model M_j. But how ought we to interpret the numbers? What does $\mathrm{BF}_{i,j}=5$ mean? On what scale are we operating?

We can get some idea of how to interpret Bayes factors by evaluating their effect on prior model weights in producing posterior model weights. Suppose that the model set consists of two models. From Eq. (7.4) it follows that

$$\Pr(M_1|\boldsymbol{Y}) = \frac{\mathrm{BF}_{1,2}\,\pi}{\mathrm{BF}_{1,2}\,\pi + (1-\pi)}, \tag{7.13}$$

and consequently that

$$\Pr(M_1|\boldsymbol{Y}) \geq p_0 \quad \text{if and only if} \quad \pi \geq \frac{p_0}{p_0 + \mathrm{BF}_{1,2}(1-p_0)}.$$

Thus for example, if BF = 50, any prior probability $\pi \geq 0.16$ will produce a posterior model weight of at least 90%.

Equation (7.13) is plotted for various values of BF in Fig. 7.4. The larger the Bayes factor, the closer the posterior model probability is to 1; if the Bayes factor is large enough, the posterior probability must be nearly 1 unless the prior probability is nearly zero.

Bayes factors are treated as quantitative measures of the strength of evidence in favor of one model relative to another. In this regard, they stand in marked contrast to frequentist hypothesis tests, which measure only the evidence *against* a model. Harold Jeffreys (Jeffreys, 1961) suggested that the strength of evidence can be categorized according to the classification in Table 7.1. Alternative but similar classifications have been proposed by various authors (for example, Kass and Raftery, 1995). We tend to evaluate Bayes factors using Eq. (7.13), with $\pi = 0.50$. Given that $M = M_1$ or $M = M_2$, the posterior probability of model M_1 is

$$\Pr(M_1|M_1 \text{ or } M_2,\boldsymbol{Y}) = \frac{\mathrm{BF}_{1,2}}{1+\mathrm{BF}_{1,2}};$$

this probability, or the description as odds multiplier [Eq. (7.11)] is better than an arbitrary cut-off in the spirit of frequentist α levels. As Kass and Raftery point out, "the interpretation may depend on the context." "Overwhelming evidence" for the superiority of one laundry detergent over another might not suffice in a criminal trial.

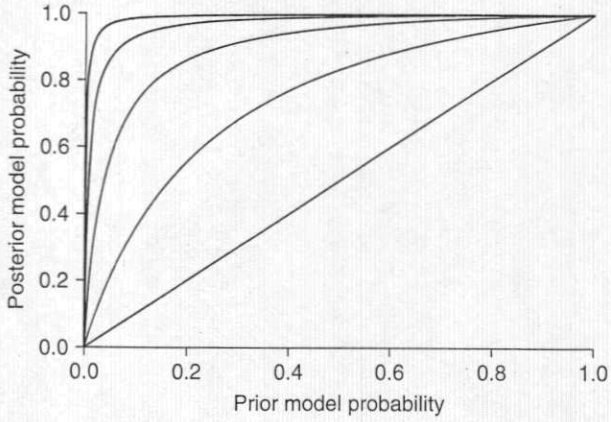

FIGURE 7.4 Posterior model probabilities (y-axis) as a function of prior model probabilities (x-axis) for Bayes factors 1, 5, 25, 125, and 625 (green, red, purple, blue, black).

TABLE 7.1 Bayes factors as weights of evidence.

Result	Interpretation
$1 < B_{1,2} < 3$	There is little support for M_1 over M_2
$3 < B_{1,2} < 12$	There is some support for M_1 over M_2
$12 < B_{1,2} < 150$	M_1 is strongly supported over M_2
$B_{1,2} > 150$	The support for M_1 is overwhelming

7.2.5 Problems with Vague Priors on Parameters

Bayesian analyses often use vague priors on parameters to let the data speak for themselves. For example, if $[X|\mu] = N(\mu, 1)$, the prior distribution $[\mu] = N(\mu_0, \sigma^2)$ results in posterior distribution

$$[\mu|X] = N\left(\frac{1}{1+\sigma^2}\mu_0 + \frac{\sigma^2}{1+\sigma^2}X, \frac{\sigma^2}{1+\sigma^2}\right),$$

which approximates the likelihood function for μ as $\sigma \to \infty$. Thus if we use a large enough value of σ in the prior, representing very limited prior knowledge about μ, inference will be based essentially on the likelihood alone, on the information provided by data rather than prior.

The "prior" with $\sigma = \infty$ is improper: it is not a true distribution function. However, we can use proper priors based on large but finite values of σ, increasingly vague in their specification of prior knowledge, and with diminishing effects on estimation. For instance, suppose $X = 1$; setting $\mu_0 = 0$, inference based on priors with $\sigma = 100$ and $\sigma = 10^6$ will yield nearly identical results, and both approximate the result based on $\sigma = \infty$.

Thus it is often the case that for the purpose of single model inference (estimation), the choice among vague priors and even of improper priors is of little consequence. Unfortunately this boon does not extend to multimodel inference, particularly when the number of

parameters varies among models. Continuing with the same example, suppose that we wish to compare two models under which X is a sample from a normal distribution with variance equal to 1. Under Model 1, $\mu = 0$; under Model 2, μ is unknown. Model 1 has no unknown parameters, so the marginal distribution is $[X|M=1] = N(0,1)$. For Model 2 we choose the prior $[\mu] = N(\mu_0, \sigma^2)$, obtaining marginal distribution $[X|M=2] = N(\mu_0, 1+\sigma^2)$. Thus the Bayes factor is

$$\begin{aligned} \mathrm{BF}_{1,2} &= \frac{\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2} X^2\right)}{\frac{1}{\sqrt{2\pi(1+\sigma^2)}} \exp\left(\frac{-1}{2(1+\sigma^2)} (X-\mu_0)^2\right)} \\ &= \sqrt{1+\sigma^2} \exp\left(-\frac{1}{2}\left\{\frac{X^2\sigma^2 + 2X\mu_0 - \mu_0^2}{1+\sigma^2}\right\}\right). \end{aligned}$$

For any fixed X and μ_0, $\mathrm{BF}_{1,2}$ acts like $\sigma \exp\left(-\frac{1}{2}X^2\right)$ for large values of the prior variance σ^2; that is, $\mathrm{BF}_{1,2} \to \infty$ as $\sigma \to \infty$, regardless of the value X. The vaguer the prior, the greater the prejudice in favor of the simpler model; with the improper prior ($\sigma = \infty$) on μ, we need not even collect data, as the matter will be decided in advance, in favor of Model 1.

Aitkin (1991) suggested comparing models using the ratio of posterior mean likelihoods, a quantity which he named the Posterior Bayes factor (PBF). Recall that the marginal distribution used in computing Bayes factors is the average value of the likelihood against the prior $[\theta_M|M]$, viz.,

$$[\boldsymbol{Y}|M] = \mathrm{E}_{[\theta_M|M]}([\boldsymbol{Y}|M,\theta_M]) = \int [\boldsymbol{Y}|M,\theta_M][\theta_M|M]\, d\theta_M;$$

Aitkin's proposal was to replace the prior $[\theta_M|M]$ by the posterior $[\theta_M|M,\boldsymbol{Y}]$ in the calculation, using ratios of

$$\mathrm{E}_{[\theta_M|M,\boldsymbol{Y}]}([\boldsymbol{Y}|M,\theta_M]) = \int [\boldsymbol{Y}|M,\theta_M][\theta_M|M,\boldsymbol{Y}]\, d\theta_M.$$

The PBF thus avoids problems with vague priors on parameters by using an informative prior, one that has been informed by the data.[4]

The PBF has been roundly criticized by Bayesian statisticians on the grounds that it amounts to double dipping, using the data twice, once to estimate parameters (i.e., to obtain the posterior distribution of the parameters) and then again to compute model weights (see Discussion following (Aitkin, 1991; Berger and Pericchi, 1996). Using the posterior, the argument goes, is to overstate the fit of the model to the data, by suggesting that unknown parameters take values consistent with the data. Although the essence of the criticism is legitimate, the PBF may be a useful tool for comparing models, developed in the spirit of BMI, but avoiding problems associated with vague priors.[5] Aitkin ably addresses the criticisms of the PBF both in the concluding comments of his paper, and after the subsequent discussion.

4. The BIC, discussed subsequently in Section 7.3.3, makes implicit use of a similar default prior, though one that is intended to be minimally informative.

5. The PBF can be understood and justified as a measure of fit based on the posterior predictive distribution; see Section 5.1.2.

Another suggestion, offered in the discussion of Aitkin's paper, is to split off some of the data into a training sample, and to use the posterior distributions arising from the training samples as informative priors for analysis of the remaining data; Berger and Pericchi (1996) develop the idea further, describing an "intrinsic Bayes factor" based on priors trained by the smallest sample sizes needed for estimation.

We conclude with this summary: the choice of priors on parameters matters in multimodel inference. There is no easy or automatic choice available. Our view is that priors on parameters should be chosen with the goal of avoiding a priori preference for one model over another; that such preference (e.g., for parsimonious modeling) should be reflected in the prior weights on models. We illustrate this perspective in our multimodel analysis of return rates for tagged trout, Section 7.4.3.

7.3 MULTIMODEL COMPUTATION

BMI can present serious computational challenges. In addition to the usual problem of computing posterior distributions for quantities of interest under each model, we must also be able to compute Bayes factors for comparing models and computing model weights.

Ideally, we would compute marginal likelihoods $[Y|M]$ for each model M, and use these to compute Bayes factors directly. Unfortunately, this approach is rarely feasible.

In this section we review two alternative approaches. The first approach is to use Gibbs sampling with *Model* treated as an unobserved quantity. Bayes factors are then computed as the ratio of (observed) posterior model odds to (specified) prior model odds. We illustrate this in Section 7.3.1 with two examples using program BUGS. In Section 7.3.2, we describe a special implementation of MCMC designed for multimodel inference, reversible jump MCMC (RJMCMC). It is possible, in some cases, to implement RJMCMC in BUGS. Programming and tuning RJMCMC can be challenging, but the basic ideas are fairly straightforward.

The second approach is to approximate $[Y|M]$ using the BIC, which we review in Section 7.3.3.

7.3.1 Multimodel Inference in BUGS

BMI can sometimes be implemented using program BUGS, with Model treated as a categorical random variable. Considerable care is required; analysts should be on the lookout for long autocorrelations in Model. We illustrate the general approach with two examples.

An Example with Nonnested Models

In Section 7.1.2, we compared the fit of geometric and Poisson models to a data set consisting of five observations. There, we calculated the integrals in (7.7) and (7.8) exactly, obtaining $BF_{1,2} = 13.84$ for the given data set.

The analysis can be carried out using Gibbs sampling, with the BUGS code given in Panel 7.1. Several features of this code require comment. First, note that Model is a categorical random variable, taking the values 1 or 2, for the geometric and Poisson models, respectively. We have assigned prior probabilities of 0.10 and 0.90 to the two models, and compute the Bayes factor as the ratio of posterior odds to prior odds. We generated a chain of