8

Bayesian Assessment of Hypotheses and Models

8.1 Introduction

The three preceding chapters gave an overview of how Bayesian probability models are constructed. Once a prior distribution is elicited and the form of the likelihood function is agreed upon, Bayesian analysis is conceptually straightforward: nuisance parameters are eliminated via integration and parameters of interest are inferred from their marginal posterior distributions. Further, yet-to-be-observed random variables can be predicted from the corresponding predictive distributions. In all cases, probability is the sole measure of uncertainty at each and everyone of the stages of Bayesian learning.

Inferences are expected to be satisfactory if the entire probability model (the prior, the likelihood, and all assumptions made) is a "good one", in some sense. In practice, however, agreement about the model to be used is more the exception than the rule, unless there is some well-established theory or mechanism underlying the problem. For example, a researcher may be uncertain about which hypothesis or theory holds. Further, almost always, there are alternative choices about the distributional form to be adopted or about the explanatory variables that should enter into a regression equation, say. Hence, it is important to take into account uncertainties about the model-building process. This is perfectly feasible in Bayesian analysis and new concepts do not need to be introduced in this respect. If there is a set of competing models in a certain class, each of the models in the set can be viewed as a different state of a random variable.

The prior distribution of this variable (the model) is updated using the information contained in the data, to arrive at the posterior distribution of the possible states of the model. Then inferences are drawn, either from the most probable model, a posteriori, or from the entire posterior distribution of the models, in a technique called Bayesian model averaging.

In this chapter, several concepts and techniques for the Bayesian evaluation of hypotheses and models are presented. Some of the approaches described are well founded theoretically; others are of a more exploratory nature. The next section defines the posterior probability of a model and an intimately related concept: the Bayes factor. Subsequently, the issue of "testing hypotheses" is presented from a Bayesian perspective. Approximations to the Bayes factor and extensions to the concept are suggested. A third section presents some methods for calculating the Bayes factor, including Monte Carlo procedures, since it is seldom the case that one can arrive at the desired quantities by analytical methods. The fourth and fifth sections present techniques for evaluating goodness of fit and the predictive ability of a model. The final section provides an introduction to Bayesian model averaging, with emphasis on highlighting its theoretical appeal from the point of view of predicting future observations.

8.2 Bayes Factors

8.2.1 Definition

Suppose there are several competing theories, hypotheses, or models about some aspect of a biological system. For example, consider different theories explaining how a population evolves. These theories are mutually exclusive and exhaustive (at least temporarily). The investigator assigns prior probability $p(H_i)$, $(i = 1, 2, ..., K)$ to hypothesis, or theory i, with $\sum_i p(H_i) = 1$. There is no limit to K and nesting requirements are not involved. After observing data $\mathbf{y}$, the posterior probability of hypothesis i is

$$p(H_i|\mathbf{y}) = \frac{p(H_i)\, p(\mathbf{y}|H_i)}{\sum_{i=1}^{K} p(H_i)\, p(\mathbf{y}|H_i)}, \quad i = 1, 2, \ldots, K, \tag{8.1}$$

where $p(\mathbf{y}|H_i)$ is the probability of the data under hypothesis i. If all hypotheses are equally likely a priori, which is the maximum entropy or reference prior in the discrete case (Bernardo, 1979), then

$$p(H_i|\mathbf{y}) = \frac{p(\mathbf{y}|H_i)}{\sum_{i=1}^{K} p(\mathbf{y}|H_i)}.$$

The posterior odds ratio of hypothesis i relative to hypothesis j takes the form

$$\frac{p\left(H_i|\mathbf{y}\right)}{p\left(H_j|\mathbf{y}\right)} = \frac{p\left(H_i\right)}{p\left(H_j\right)} \frac{p\left(\mathbf{y}|H_i\right)}{p\left(\mathbf{y}|H_j\right)}. \tag{8.2}$$

It follows that the posterior odds ratio is the product of the prior odds ratio and of the ratio between the marginal probabilities of observing the data under each of the hypotheses. The Bayes factor is defined to be

$$B_{ij} = \frac{p\left(\mathbf{y}|H_i\right)}{p\left(\mathbf{y}|H_j\right)} = \frac{\frac{p(H_i|\mathbf{y})}{p(H_j|\mathbf{y})}}{\frac{p(H_i)}{p(H_j)}} = \frac{\text{posterior odds ratio}}{\text{prior odds ratio}}. \tag{8.3}$$

According to Kass and Raftery (1995) this terminology is apparently due to Good (1958). A $B_{ij} > 1$ means that H_i is more plausible than H_j in the light of $\mathbf{y}$. While the priors are not visible in the ratio $p\left(\mathbf{y}|H_i\right)/p\left(\mathbf{y}|H_j\right)$, this does not mean that B_{ij} in general is not affected by prior specifications. This point is discussed below.

It is instructive to contrast this approach with the one employed in standard statistical analysis. In classical hypothesis testing, a null hypothesis $H_0 : \theta \in \theta_0$ and an alternative hypothesis $H_1 : \theta \in \theta_1$ are specified. The choice between these hypotheses is driven by the distribution under H_0 of a test statistic that is a function of the data (it could be the likelihood ratio), $T\left(\mathbf{y}\right)$, and by the so-called p-value. This is defined as

$$\Pr\left[T\left(\mathbf{y}\right) \text{ at least as extreme as the value observed}|\theta, H_0\right]. \tag{8.4}$$

Then H_0 is accepted (or rejected, in which case H_1 is accepted) if the p-value is large (small) enough, or one may just quote the p-value and leave things there. Notice that (8.4) represents the probability of obtaining results larger than the one actually obtained; that is, (8.4) is concerned with events that might have occurred, but have not. Thus, the famous quotation from Jeffreys (1961):

> "What the use of p implies, therefore, is that a hypothesis which may be true may be rejected because it has not predicted observable results which have not occurred. ... On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it."

Often (and incorrectly), the p-value is interpreted as the probability that H_0 holds true. The interpretation in terms of probability of hypotheses, $p\left[H_0|T\left(\mathbf{y}\right) = t\left(\mathbf{y}\right)\right]$, which is the Bayesian formulation of the problem, is conceptually more straightforward than the one associated with (8.4). Despite its conceptual clarity, the Bayesian approach is not free from problems. Perhaps not surprisingly, these arise especially in cases when prior information is supposed to convey vague knowledge.

8.2.2 Interpretation

The appeal of the Bayes factor as formulated in (8.3), is that it provides a measure of whether the data have increased or decreased the odds of H_i relative to H_j. This, however, does not mean that in general, the Bayes factor is driven by the data only. It is only when both H_i and H_j are simple hypotheses, that the prior influence vanishes and the Bayes factor takes the form of a likelihood ratio. In general, however, the Bayes factor depends on prior input, a point to which we will return.

Kass and Raftery (1995) give guidelines for interpreting the evidence against some "null hypothesis", H_0. For example, they suggest that a Bayes factor larger than 100 should be construed as "decisive evidence" against the null. Note that a Bayes factor under 1 means that there is evidence in support of H_0. When working in a logarithmic scale, $2 \log B_{ij}$, for example, the values are often easier to interpret by those who are familiar with likelihood ratio tests. It should be made clear from the onset that the Bayes factor cannot be viewed as a statistic having an asymptotic chi-square distribution under the null hypothesis. Again, B_{ij} is the quantity by which prior odds ratios are increased (or decreased) to become posterior odd ratios.

There are many differences between the Bayes factor and the usual likelihood ratio statistic. First, the intervening $p(\mathbf{y}|H_i)$ is not the classical likelihood, in general. Recall that the Bayesian marginal probability (or density) of the data is arrived at by integrating the joint density of the parameters and of the observations over all values that the parameters can take in their allowable space. For example, if hypothesis or model H_i has parameters $\boldsymbol{\theta}_i$, then for continuous data and continuous valued parameter vector

$$\begin{aligned} p(\mathbf{y}|H_i) &= \int p(\mathbf{y}|\boldsymbol{\theta}_i, H_i)\, p(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i \\ &= E_{\boldsymbol{\theta}_i|H_i}\left[p(\mathbf{y}|\boldsymbol{\theta}_i, H_i)\right]. \end{aligned} \tag{8.5}$$

The marginal density is, therefore, the expected value of all possible likelihoods, where the expectation is taken with respect to the prior distribution of the parameters. In likelihood inference, no such integration takes place unless the "parameters" are random variables having a frequentist interpretation. Since, in turn, these random variables have distributions indexed by parameters, the classical likelihood always depends on some fixed, unknown parameters. In the Bayesian approach, on the other hand, any dependence of the marginal distribution of the data is with respect to any hyperparameters the prior distribution may have, and with respect to the form of the model. In fact, $p(\mathbf{y}|H_i)$ is the prior predictive distribution and gives the density or probability of the data calculated before observation, unconditionally with respect to parameter values.

A second important difference is that the Bayes factor is not explicitly related to any critical value defining a rejection region of a certain size. For example, the usual p-values in classical hypothesis testing cannot be interpreted as the probabilities that either the null or the alternative hypotheses are "true". The p-value arises from the distribution of the test statistic (under the null hypothesis) in conceptual replications of the experiment. In contrast, the Bayes factor and the prior odds contribute directly to forming the posterior probabilities of the hypotheses. In order to illustrate, suppose that two models are equally probable, a priori. Then a Bayes factor $B_{01} = 19$, would indicate that the null hypothesis or model is 19 times more probable than its alternative, and that the posterior probability that the null model is true is 0.95. On the other hand, in a likelihood ratio test, a value of the test statistic generating a p-value of 0.95 as defined by (8.4) cannot be construed as evidence that the null hypothesis has a 95% chance of being true.

8.2.3 The Bayes Factor and Hypothesis Testing

Decision-Theoretic View

In Bayesian analysis, "hypothesis testing" is viewed primarily as a decision problem (e.g., Zellner, 1971). Suppose there are two hypotheses or models: H_0 (null) and H_1 (alternative). If one chooses H_0 when H_1 is "true", then a loss L_{10} is incurred. Similarly, when H_1 is adopted when the null holds, the loss is L_{01}. Otherwise, there are no losses.

The posterior expectation of the decision "accept the null hypothesis" is

$$\begin{aligned} E\left(\text{loss}|\text{accept } H_0, \mathbf{y}\right) &= 0 \times p\left(H_0|\mathbf{y}\right) + L_{10}\, p\left(H_1|\mathbf{y}\right) \\ &= L_{10}\, p\left(H_1|\mathbf{y}\right). \end{aligned}$$

Likewise, the expected posterior loss of the decision "accept the alternative" is

$$\begin{aligned} E\left(\text{loss}|\text{reject } H_0, \mathbf{y}\right) &= 0 \times p\left(H_1|\mathbf{y}\right) + L_{01}\, p\left(H_0|\mathbf{y}\right) \\ &= L_{01}\, p\left(H_0|\mathbf{y}\right). \end{aligned}$$

Naturally, if the expected posterior loss of accepting H_0 is larger than that of rejecting it, one would decide to reject the null hypothesis. Then the decision rule is

$$\text{if } E\left(\text{loss}|\text{reject } H_0, \mathbf{y}\right) < E\left(\text{loss}|\text{accept } H_0, \mathbf{y}\ \right) \rightarrow \text{reject } H_0.$$

The preceding is equivalent to

$$L_{01}\, p\left(H_0|\mathbf{y}\right) < L_{10}\, p\left(H_1|\mathbf{y}\right),$$

or, in terms of (8.3),

$$B_{10} = \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)} > \frac{L_{01}\, p(H_0)}{L_{10}\, p(H_1)}. \tag{8.6}$$

This indicates that the null hypothesis is to be rejected if the Bayes factor (ratio of marginal likelihoods under the two hypotheses or models) for the alternative, relative to the null, exceeds the ratio of the prior expected losses. Note that $L_{01}\, p(H_0)$ is the expected prior loss of rejecting the null when this is true; $L_{10}\, p(H_1)$ is the expected prior loss that results when H_1 is true and one accepts H_0. Then the ratio of prior to posterior expected losses

$$\frac{L_{01}\, p(H_0)}{L_{10}\, p(H_1)}$$

plays the role of the "critical" value in classical hypothesis testing. If one views the Bayes factor as the "test statistic", the critical value is higher when one expects to loose more from rejecting the null than from accepting it. In other words, the larger the prior expected loss from rejecting the null (when this hypothesis is true) relative to the prior expected loss of accepting it (when H_1 holds), the larger the weight of the evidence should be in favor of the alternative, as measured by the Bayes factor.

If the losses are such that $L_{01} = L_{10}$, it follows from (8.6) that the decision rule is simply

$$B_{10} = \frac{p(\mathbf{y}|H_1)}{p(\mathbf{y}|H_0)} > \frac{p(H_0)}{p(H_1)}.$$

This implies that if the two models or hypotheses are equiprobable a priori, then the alternative should be chosen over the null whenever the Bayes factor exceeds 1. Similarly, a "critical value" of 10 should be adopted if it is believed a priori that the null hypothesis is 10 times more likely than the alternative. In all cases, it must be noted that the "accept" or "reject" framework depends nontrivially on the form of the loss function, and that adopting $L_{01} = L_{10}$ may not be realistic in many cases.

The definition in the form of (8.6) highlights the importance, in Bayesian testing, of defining non-zero a priori probabilities. This is so even though the Bayes factor can be calculated without specifying $p(H_0)$ and $p(H_1)$. If H_0 or H_1 are a priori impossible, the observations will not modify this information.

Bayesian Comparisons

Contrasting Two Simple Hypotheses

The definition of the Bayes factor in (8.3) as a ratio of marginal densities does not make explicit the influence of the prior distributions. With one

exception, Bayes factors are affected by prior specifications. The exception occurs when the comparison involves two simple hypotheses. In this case, under H_0, a particular value $\boldsymbol{\theta}_0$ is assigned to the parameter vector, whereas under H_1, another value $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ is posited. There is no uncertainty about the value of the parameter under any of the two competing hypotheses. Then one can express the discrete prior probability of hypothesis i as $p(H_i) = \Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_i)$, and the conditional p.d.f. for $\mathbf{y}$ given H_i as $p(\mathbf{y}|H_i) = p(\mathbf{y}|\boldsymbol{\theta} = \boldsymbol{\theta}_i)$. The Bayes factor for the alternative against the null is then

$$B_{10} = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p(\mathbf{y}|\boldsymbol{\theta} = \boldsymbol{\theta}_1)}{p(\mathbf{y}|\boldsymbol{\theta} = \boldsymbol{\theta}_0)}. \tag{8.7}$$

In this particular situation, the Bayes factor is the odds for H_1 relative to H_0 given by the data only. Expression (8.7) is a ratio of standard likelihoods, where the values of the parameters are completely specified. In general, however, B_{10} depends on prior input. When a hypothesis is not simple, in order to arrive at the form equivalent to (8.7), one must compute the expectation of the likelihood of $\boldsymbol{\theta}_i$ with respect to the prior distribution. For continuously distributed values of the vector $\boldsymbol{\theta}_i$ and prior density $p(\boldsymbol{\theta}_i|H_i)$, one writes

$$p(\mathbf{y}|H_i) = \int p(\mathbf{y}|\boldsymbol{\theta}_i, H_i)\, p(\boldsymbol{\theta}_i|H_i)\, d\boldsymbol{\theta}_i.$$

In contrast to the classical likelihood ratio frequentist test, the Bayes factor does not impose nesting restrictions concerning the form of the likelihood functions, as illustrated in the following example adapted from Bernardo and Smith (1994).

Example 8.1 *Two fully specified models: Poisson versus negative binomial process*

Two completely specified models are proposed for counts. A sample of size n with values $y_1, y_2, \ldots, y_n$ is drawn independently from some population. Model P states that the distribution of the observations is Poisson with parameter θ_P. Then the likelihood under this model is

$$p(\mathbf{y}|\theta = \theta_P) = \prod_{i=1}^{n} \left[\frac{\theta_P^{y_i} \exp(-\theta_P)}{y_i!}\right] = \frac{\theta_P^{n\overline{y}}}{\exp(n\theta_P) \prod_{i=1}^{n} y_i!}. \tag{8.8}$$

Model N proposes a negative binomial distribution with parameter θ_P. The corresponding likelihood is

$$p(\mathbf{y}|\theta = \theta_N) = \prod_{i=1}^{n} \left[\theta_N (1-\theta_N)^{y_i}\right] = \theta_N^n (1-\theta_N)^{n\overline{y}}. \tag{8.9}$$

The Bayes factor for Model N relative to Model P is then

$$B_{NP} = \left(\frac{1-\theta_N}{\theta_P}\right)^{n\overline{y}} \frac{\theta_N^n}{\left[\exp(n\theta_P)\prod_{i=1}^{n} y_i!\right]^{-1}},$$

and its logarithm can be expressed as

$$\log B_{NP} = n\left[\overline{y}\log\left(\frac{1-\theta_N}{\theta_P}\right) + \log(\theta_N)\right] + n\theta_P + \sum_{i=1}^{n}\log(y_i!).$$

■

Simple Versus Composite Hypotheses

A second type of comparison is one where one of the models (Model 0 = M_0) postulates a given value of the parameter, whereas the other model (Model 1 = M_1) allows the unknown parameter to take freely any of its values in the allowable parameter space. This is called a simple versus composite test (Bernardo and Smith, 1994), and the Bayes factor in this case takes the form

$$B_{10} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_0)} = \frac{\int p(\mathbf{y}|\boldsymbol{\theta},M_1)\,p(\boldsymbol{\theta}|M_1)\,d\boldsymbol{\theta}}{p(\mathbf{y}|\boldsymbol{\theta}=\boldsymbol{\theta}_0,M_0)},$$

where $p(\boldsymbol{\theta}|M_1)$ is the density of the prior distribution of the parameter vector under the assumptions of Model 1.

There is an interesting relationship between the posterior probability that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and the Bayes factor (Berger, 1985). Denote the prior probability of models 0 and 1 as $p(M_0)$ and $p(M_1)$, respectively, with $p(M_0) + p(M_1) = 1$. The term $p(M_0)$ can also be interpreted as the prior probability that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, the posterior probability that $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is

$$\Pr(\boldsymbol{\theta}=\boldsymbol{\theta}_0|\mathbf{y}) = \frac{p(M_0)\,p(\mathbf{y}|\boldsymbol{\theta}=\boldsymbol{\theta}_0,M_0)}{p(\mathbf{y})}.$$

The constant term in the denominator is given by

$$\begin{aligned} p(\mathbf{y}) &= p(\mathbf{y}|\boldsymbol{\theta}=\boldsymbol{\theta}_0,M_0)\Pr(\boldsymbol{\theta}=\boldsymbol{\theta}_0|M_0)\,p(M_0) \\ &\quad + \int p(\mathbf{y}|\boldsymbol{\theta},M_1)\,p(\boldsymbol{\theta}|M_1)\,p(M_1)\,d\boldsymbol{\theta} \\ &= p(M_0)\,p(\mathbf{y}|\boldsymbol{\theta}=\boldsymbol{\theta}_0,M_0) + p(M_1)\int p(\mathbf{y}|\boldsymbol{\theta},M_1)\,p(\boldsymbol{\theta}|M_1)\,d\boldsymbol{\theta}, \end{aligned}$$

	Genotypes			
	AB/ab	Ab/ab	aB/ab	ab/ab
Phenotype	AB	Ab	aB	ab
Frequency	$\frac{1}{2}(1-r)$	$\frac{1}{2}r$	$\frac{1}{2}r$	$\frac{1}{2}(1-r)$
Observed	a	b	c	d

TABLE 8.1. Genotypic distribution in offspring from a backcross design.

with the equality arising in the last line because $\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|M_0) = 1$. Substituting above yields

$$\Pr(\boldsymbol{\theta} = \boldsymbol{\theta}_0|\mathbf{y}) = \left[1 + \frac{p(M_1)}{p(M_0)} B_{10}\right]^{-1}.$$

It is important to mention that in evaluating a point null hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, say, $\boldsymbol{\theta}_0$ must be assigned a positive probability a priori. The point null hypothesis cannot be tested invoking a continuous prior distribution, since any such prior will give $\boldsymbol{\theta}_0$ prior (and therefore posterior) probability of zero.

In contrast to the traditional likelihood ratio, the test of a parameter value on the boundary of the parameter space using the Bayes factor does not in principle create difficulties. This being so because asymptotic distributions and series expansions do not come into play. Such a test is illustrated in the example below.

Example 8.2 *A point null hypothesis: assessing linkage between two loci* The problem consists of inferring the probability of recombination r between two autosomal loci A and B, each with two alleles. The parameter r is defined in the closed interval $\left[0, \frac{1}{2}\right]$, with the upper value corresponding to the situation where there is no linkage. We wish to derive the posterior distribution of the recombination fraction between the two loci, and to contrast the two models that follow. The null model (M_0) postulates that segregation is independent (that is, $r = \frac{1}{2}$), whereas the alternative model (M_1) claims that the loci are linked (that is, $r < \frac{1}{2}$).

Let the alleles at the corresponding loci be A, a, B, and b, where A and B are dominant alleles. Suppose that a line consisting of coupling heterozygote individuals AB/ab is crossed to homozygotes ab/ab. Hence four offspring classes can be observed: AB/ab, Ab/ab, aB/ab, and ab/ab. Let $n = a + b + c + d$ be the total number of offspring observed. The four possible genotypes resulting from this cross, their phenotypes, the expected frequencies and the observed numbers are shown in Table 8.1. The expected relative frequencies follow from the fact that if the probability of observing a recombinant is r, the individual can be either Ab/ab or aB/ab, with the two classes being equally likely. A similar reasoning applies to the observation of a non-recombinant type.

Suppose that the species under consideration has 22 pairs of autosomal chromosomes. In the absence of prior information about loci A and B, it may be reasonable to assume that the probability that these are located on the same chromosome (and therefore, linked, so that $r < \frac{1}{2}$) is $\frac{1}{22}$. This is so because the probability that 2 randomly picked alleles are in a given chromosome is $\left(\frac{1}{22}\right)^2$, and there are 22 chromosomes in which this can occur. Hence, a priori, $p(M_1) = \frac{1}{22}$ and $p(M_0) = \frac{21}{22}$; the two models are viewed as mutually exclusive and exhaustive.

Next, we must arrive at some reasonable prior distribution for r under M_1. Here, a development by Smith (1959) is followed. First, note that the recombination fraction takes the value $r = \frac{1}{2}$ with prior probability $\frac{21}{22}$. Further, assume a uniform distribution for r otherwise (provided one can view the values $0 < r < \frac{1}{2}$ as "equally likely"). Then the density of this uniform distribution, $p(r|M_1)$ must be such that

$$\Pr\left(r < \frac{1}{2}\right) = \int_0^{\frac{1}{2}} p(u|M_1)\, du = \frac{1}{22}.$$

Solving for the desired uniform density gives

$$p(r|M_1) = \frac{1}{11}.$$

Therefore the prior is the uniform process $p(r|M_1) = \frac{1}{11}$ for $0 < r < \frac{1}{2}$, and the point mass $\frac{21}{22}$ at $r = \frac{1}{2}$. That is, $\Pr\left(r = \frac{1}{2}\right) = p(M_0) = \frac{21}{22}$. Note that given M_0, $\Pr\left(r = \frac{1}{2}|M_0\right) = 1$.

Given the data $\mathbf{y} = (a, b, c, d)'$ in Table 8.1, the conditional distribution of the observations under linkage has the multinomial form

$$\begin{aligned} p(\mathbf{y}|r, M_1) &= \frac{n!}{a!b!c!d!}\left[\frac{1}{2}(1-r)\right]^a\left(\frac{1}{2}r\right)^b\left(\frac{1}{2}r\right)^c\left[\frac{1}{2}(1-r)\right]^d \\ &\propto \left(\frac{1}{2}\right)^n (1-r)^{a+d}\, r^{b+c}. \end{aligned}$$

Under no linkage

$$p\left(\mathbf{y}|r = \frac{1}{2}, M_0\right) = \frac{n!}{a!b!c!d!}\left(\frac{1}{4}\right)^{a+b+c+d} \propto \left(\frac{1}{4}\right)^n.$$

Therefore the posterior odds ratio is given by

$$\begin{aligned} \frac{p(M_1|\mathbf{y})}{p(M_0|\mathbf{y})} &= \frac{p(M_1)}{p(M_0)}\frac{\int_0^{\frac{1}{2}} p(r|M_1)\, p(\mathbf{y}|r, M_1)\, dr}{p\left(\mathbf{y}|r = \frac{1}{2}, M_0\right)} \\ &= \frac{1}{21}\frac{\frac{1}{11}\left(\frac{1}{2}\right)^n \int_0^{\frac{1}{2}} r^{b+c}(1-r)^{a+d}\, dr}{\left(\frac{1}{4}\right)^n}, \end{aligned}$$

where

$$B_{10} = \frac{\int_0^{\frac{1}{2}} p\left(r|M_1\right) p\left(\mathbf{y}|r, M_1\right) dr}{p\left(\mathbf{y}|r = \frac{1}{2}, M_0\right)}.$$

The posterior probability of linkage is

$$p\left(M_1|\mathbf{y}\right) = \frac{p\left(M_1\right) p\left(\mathbf{y}|M_1\right)}{p\left(\mathbf{y}\right)},$$

where

$$p\left(\mathbf{y}|M_1\right) = \int_0^{1/2} p\left(\mathbf{y}|r, M_1\right) p\left(r|M_1\right) dr, \tag{8.10}$$

whereas the posterior probability of no linkage is

$$p\left(M_0|\mathbf{y}\right) = 1 - p\left(M_1|\mathbf{y}\right).$$

The marginal density of the data is equal to

$$p\left(\mathbf{y}\right) = p\left(M_0\right) p\left(\mathbf{y}|r = \frac{1}{2}, M_0\right)$$
$$+p\left(M_1\right) \int_0^{\frac{1}{2}} p\left(r|M_1\right) p\left(\mathbf{y}|r, M_1\right) dr.$$

The integrals in these expressions can easily be evaluated numerically. ■

Example 8.3 *Lindley's paradox*

This problem was brought to light initially by Lindley (1957). The data sampling involves n independent draws from $N\left(\mu, \sigma^2\right)$, with σ^2 known and μ to be inferred. Model 0 corresponds to the simple or sharp hypothesis that $\mu = \mu_0$. Model 1 takes σ^2 as known and μ as unknown, with its prior distribution being $N\left(\mu_1, \sigma_1^2\right)$; the hyperparameters are assumed to be known. This corresponds to a classical setting in which Model 0 is the null hypothesis $\mu = \mu_0$, and Model 1 is the alternative that the parameter can take any value other than $\mu = \mu_0$.

The marginal density of the data under Model 0 is

$$p_0\left(\mathbf{y}|\mu_0, \sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu_0\right)^2\right]$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2\right] \exp\left[-\frac{n}{2\sigma^2}\left(\overline{y} - \mu_0\right)^2\right]. \tag{8.11}$$

The marginal density under Model 1 can be written as

$$p_1\left(\mathbf{y}|\mu_1,\sigma_1^2,\sigma^2\right)=\int p\left(\mathbf{y}|\mu,\sigma^2\right)p\left(\mu|\mu_1,\sigma_1^2\right)d\mu$$
$$=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2\right]\frac{1}{\sqrt{2\pi\sigma_1^2}}$$
$$\int\exp\left[-\frac{n}{2\sigma^2}\left(\overline{y}-\mu\right)^2\right]\exp\left[-\frac{\left(\mu-\mu_1\right)^2}{2\sigma_1^2}\right]d\mu. \tag{8.12}$$

The Bayes factor for Model 0 relative to Model 1 is given by the ratio between (8.11) and (8.12)

$$B_{01}=\frac{\exp\left[-\frac{n}{2\sigma^2}\left(\overline{y}-\mu_0\right)^2\right]}{\frac{1}{\sqrt{2\pi\sigma_1^2}}\int\exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}\left(\mu-\overline{y}\right)^2+\frac{(\mu-\mu_1)^2}{\sigma_1^2}\right]\right\}d\mu}. \tag{8.13}$$

Now the two quadratic forms on μ in the integrand can be combined in the usual manner, leading to

$$\frac{n}{\sigma^2}\left(\mu-\overline{y}\right)^2+\frac{\left(\mu-\mu_1\right)^2}{\sigma_1^2}=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)\left(\mu-\widehat{\mu}\right)^2$$
$$+\frac{n}{\sigma^2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)^{-1}\frac{1}{\sigma_1^2}\left(\overline{y}-\mu_1\right)^2.$$

Above

$$\widehat{\mu}=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)^{-1}\left(\frac{n\overline{y}}{\sigma^2}+\frac{\mu_1}{\sigma_1^2}\right).$$

Carrying out the integration, the Bayes factor becomes, after some algebra,

$$B_{01}=\sqrt{\frac{\sigma_1^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)^{-1}}}\frac{\exp\left[-\frac{n}{2\sigma^2}\left(\overline{y}-\mu_0\right)^2\right]}{\exp\left[-\frac{n}{2\sigma^2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_1^2}\right)^{-1}\frac{1}{\sigma_1^2}\left(\overline{y}-\mu_1\right)^2\right]}. \tag{8.14}$$

Now examine what happens when the prior information becomes more and more diffuse, that is, eventually σ_1^2 is so large that $1/\sigma_1^2$ is near 0. The Bayes factor is, approximately,

$$B_{01}\approx\sqrt{\frac{\sigma_1^2\sigma^2}{n}}\exp\left[-\frac{n}{2\sigma^2}\left(\overline{y}-\mu_0\right)^2\right].$$

For any fixed value of $\overline{y}$, the Bayes factor goes to ∞ when $\sigma_1^2\to\infty$, which implies that $p\left(Model\ 0|\mathbf{y}\right)\to 1$. This means that no matter what the value of $\overline{y}$ is, the null hypothesis would tend to be favored, even for values of

$\left|(\overline{y}-\mu_0)/\sqrt{\sigma^2/n}\right|$ that are large enough to cause rejection of the null at any, arbitrary, "significance" level in classical testing. This result, known as "Lindley's paradox", illustrates that a comparison of models in which one of the hypothesis is "sharp" (simple), strongly depends on the form of the prior distribution. In particular, when the distribution is improper, the Bayes factor leads to acceptance of the null. O'Hagan (1994) concludes that improper priors cannot be used when comparing models. However, the problem is not avoided entirely by adopting vague uniform priors over some large but finite range. This will be discussed later. ■

Comparing Two Composite Hypotheses

Third, the comparison may be a "composite versus composite", that is, one where the two models allow their respective parameters to take any values in the corresponding spaces. Here the Bayes factor is

$$B_{10}=\frac{p\left(\mathbf{y}|M_1\right)}{p\left(\mathbf{y}|M_0\right)}=\frac{\int p\left(\mathbf{y}|\boldsymbol{\theta}_1,M_1\right)p_1\left(\boldsymbol{\theta}_1|M_1\right)d\boldsymbol{\theta}}{\int p\left(\mathbf{y}|\boldsymbol{\theta}_0,M_0\right)p_0\left(\boldsymbol{\theta}_0|M_0\right)d\boldsymbol{\theta}}. \tag{8.15}$$

In general, all constants appearing in $p\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)$ must be included when computing B_{10}.

Example 8.4 *Marginal distributions and the Bayes factor in Poisson and negative binomial models*

The setting is as in Example 8.1, but the parameter values are allowed to take any values in their spaces. Following Bernardo and Smith (1994), take as prior distribution for the Poisson parameter

$$\theta_P\sim Ga\left(a_P,b_P\right),$$

and for the parameter of the negative binomial model adopt as prior the Beta distribution

$$\theta_N\sim Be\left(a_N,b_N\right).$$

The a's and b's are known hyperparameters. The marginal distribution of the data under the Poisson model, using (8.8) as likelihood function (suppressing the dependence on hyperparameters in the notation), is obtained as

$$\begin{aligned}p\left(\mathbf{y}|P\right)&=\int\frac{\theta_P^{n\overline{y}}\exp\left(-n\theta_P\right)}{\prod_{i=1}^{n}y_i!}\frac{b_P^{a_P}}{\Gamma\left(a_P\right)}\theta_P^{a_P-1}\exp\left(-b_P\theta_P\right)d\theta_P\\&=\frac{b_P^{a_P}}{\Gamma\left(a_P\right)\prod_{i=1}^{n}y_i!}\int\theta_P^{n\overline{y}+a_P-1}\exp\left[-\left(n+b_P\right)\theta_P\right]d\theta_P.\end{aligned} \tag{8.16}$$

The integrand is the kernel of the density of the

$$\theta_P\sim Ga\left(n\overline{y}+a_P,n+b_P\right)$$

distribution. Hence, the marginal of interest is

$$p(\mathbf{y}|P) = \frac{\Gamma(a_P + n\overline{y})\, b_P^{a_P}}{\Gamma(a_P)\,(n + b_P)^{a_P + n\overline{y}} \prod_{i=1}^{n} y_i!}. \tag{8.17}$$

Similarly, using (8.9), the marginal distribution of the data under the negative binomial model takes the form

$$\begin{aligned} p(\mathbf{y}|N) &= \int \theta_N^n (1-\theta_N)^{n\overline{y}} \frac{\Gamma(a_N + b_N)}{\Gamma(a_N)\Gamma(b_N)} \theta_N^{a_N - 1} (1-\theta_N)^{b_N - 1}\, d\theta_N \\ &= \frac{\Gamma(a_N + b_N)}{\Gamma(a_N)\Gamma(b_N)} \int \theta_N^{a_N + n - 1} (1-\theta_N)^{b_N + n\overline{y} - 1}\, d\theta_N. \end{aligned}$$

The integrand is the kernel of a beta density, so the integral can be evaluated analytically, yielding

$$p(\mathbf{y}|N) = \frac{\Gamma(a_N + b_N)}{\Gamma(a_N)\Gamma(b_N)} \frac{\Gamma(a_N + n)\,\Gamma(b_N + n\overline{y})}{\Gamma(a_N + n + b_N + n\overline{y})}. \tag{8.18}$$

The Bayes factor in favor of the N model relative to the P model is given by the ratio between (8.18) and (8.17). Note that the two marginal densities and the Bayes factor depend only on the data and on the hyperparameters, contrary to the ratio of likelihoods. This is because all unknown parameters are integrated out in the process of finding the marginals. It cannot be overemphasized that all integration constants must be kept when calculating the Bayes factors. In classical likelihood ratio tests, on the other hand, only those parts of the density functions that depend on the parameters are kept. ■

8.2.4 Influence of the Prior Distribution

From its definition, and from Example 8.4, it should be apparent that the Bayes factor depends on the prior distributions adopted for the competing models. The exception is when two simple hypotheses are at play. For the Poisson versus Negative Binomial setting discussed above, Bernardo and Smith (1994) give numerical examples illustrating that minor changes in the values of the hyperparameters produce changes in the direction of the Bayes factors. This dependence is illustrated with a few examples in what follows. Before we do so, note that one can write

$$\begin{aligned} \int p(\mathbf{y}|M_i)\, d\mathbf{y} &= \int\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i d\mathbf{y} \\ &= \int p(\boldsymbol{\theta}_i|M_i) \left[\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, d\mathbf{y} \right] d\boldsymbol{\theta}_i \\ &= \int p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i, \end{aligned}$$

where the last equality follows because $\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, d\mathbf{y} = 1$. The message here is that when $p(\boldsymbol{\theta}_i|M_i)$ is improper, so is $p(\mathbf{y}|M_i)$. In this case, the Bayes factor is not well defined. This is discussed further in Subsection 8.2.5 below.

Example 8.5 *Influence of the bounds of a uniform prior*

Let the sampling model be $y_i|\mu \sim N(\mu, 1)$ and let the prior distribution adopted for μ under Model 1 be uniform over $[-L, L]$. Model 2 postulates the same sampling model but the bounds are $[-\alpha L, \alpha L]$, where α is a known, positive, real number. Suppose n independent samples are drawn, so that the marginal density of the data under Model 2 is

$$p(\mathbf{y}|\alpha, L) = \int_{-\alpha L}^{\alpha L} \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right]\frac{1}{2\alpha L}d\mu$$

$$= \frac{1}{2\alpha L}\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\overline{y})^2\right]\int_{-\alpha L}^{\alpha L}\exp\left[-\frac{n}{2}(\mu-\overline{y})^2\right]d\mu.$$

The integrand is in a normal form and can be evaluated readily, yielding

$$p(\mathbf{y}|\alpha, L) = \frac{1}{2\alpha L}\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(y_i-\overline{y})^2\right]$$
$$\times\left[\Phi\left(\frac{\alpha L-\overline{y}}{\sqrt{\frac{1}{n}}}\right)-\Phi\left(\frac{-\alpha L-\overline{y}}{\sqrt{\frac{1}{n}}}\right)\right]\sqrt{2\pi\frac{1}{n}}.$$

The Bayes factor for Model 2 relative to Model 1 is

$$B_{21} = \frac{p(\mathbf{y}|\alpha, L)}{p(\mathbf{y}|1, L)} = \frac{\Phi\left(\frac{\alpha L-\overline{y}}{\sqrt{\frac{1}{n}}}\right)-\Phi\left(\frac{-\alpha L-\overline{y}}{\sqrt{\frac{1}{n}}}\right)}{\alpha\left[\Phi\left(\frac{L-\overline{y}}{\sqrt{\frac{1}{n}}}\right)-\Phi\left(\frac{-L-\overline{y}}{\sqrt{\frac{1}{n}}}\right)\right]}.$$

This clearly shows that the Bayes factor is sensitive with respect to the value of α. This is relevant in conjunction with the problem outlined in Example 8.3: the difficulties caused by improper priors in model selection via the Bayes factors are not solved satisfactorily by adopting, for example, bounded uniform priors. The Bayes factor depends very strongly on the width of the interval used. ■

Example 8.6 *The Bayes factor for a simple linear model*

Consider the linear model

$$\mathbf{y} = \boldsymbol{\beta} + \mathbf{e},$$

where the variance of the residual distribution, σ^2, is known, so it can be set equal to 1 without loss of generality. Model 1 posits as prior distribution $\boldsymbol{\beta} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_1^2\right)$, whereas for Model 2 the prior distribution is $\boldsymbol{\beta} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_2^2\right)$. Since the sampling model and the priors are both normal, it follows that the marginal distributions of the data are normal as well. The means and variances of these distributions are arrived at directly by taking expectations of the sampling model with respect to the appropriate prior. One gets $\mathbf{y}|\sigma^2, \sigma_1^2 \sim N\left(\mathbf{0}, \mathbf{I}\left(\sigma_1^2 + \sigma^2\right)\right)$ and $\mathbf{y}|\sigma^2, \sigma_2^2 \sim N\left(\mathbf{0}, \mathbf{I}\left(\sigma_2^2 + \sigma^2\right)\right)$ for Models 1 and 2, respectively. The Bayes factor for Model 1 relative to Model 2 is

$$\begin{aligned} B_{12} &= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\left(\sigma_1^2+1\right)}} \exp\left[-\frac{y_i^2}{2\left(\sigma_1^2+1\right)}\right]}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\left(\sigma_2^2+1\right)}} \exp\left[-\frac{y_i^2}{2\left(\sigma_2^2+1\right)}\right]} \\ &= \left(\frac{\sigma_2^2+1}{\sigma_1^2+1}\right)^{\frac{n}{2}} \frac{\exp\left[\frac{\mathbf{y}'\mathbf{y}}{2\left(\sigma_2^2+1\right)}\right]}{\exp\left[\frac{\mathbf{y}'\mathbf{y}}{2\left(\sigma_1^2+1\right)}\right]}. \end{aligned}$$

Taking logarithms and multiplying by 2, to arrive at the same scale as the likelihood ratio statistic, yields

$$2\log\left(B_{12}\right) = n\log\left(\frac{\sigma_2^2+1}{\sigma_1^2+1}\right) + \mathbf{y}'\mathbf{y}\left[\frac{\sigma_1^2-\sigma_2^2}{\left(\sigma_1^2+1\right)\left(\sigma_2^2+1\right)}\right].$$

The first term will contribute toward favoring Model 1 whenever σ_2^2 is larger than σ_1^2, whereas the opposite occurs in the second term. ■

8.2.5 Nested Models

As seen in Chapter 3, a nested model is one that can be viewed as a special case of a more general, larger model, and is typically obtained by fixing or "zeroing in" some parameters in the latter. Following O'Hagan (1994), let the bigger model have parameters (θ, ϕ) and denote it Model 1 whereas in the nested model fix $\phi = \phi_0$, with this value being usually 0. This is Model 0.

Let the prior probability of the larger model (often called the "alternative" one) be π_1, and let the prior density of its parameters be $p_1(\theta, \phi)$. The prior probability of the nested model is $\pi_0 = 1 - \pi_1$, which can be interpreted as the prior probability that $\phi = \phi_0$. This is somewhat perplexing at first sight, since the probability that a continuous parameter takes a given value is 0. However, the fact that consideration is given to the nested model as a plausible model implies that one is assigning some probability to the special situation that $\phi = \phi_0$ holds. In the nested model,

the prior density of the "free parameters" is $p_0(\theta) = p(\theta|\phi = \phi_0)$, that is, the density of the conditional distribution of the theta parameter, given that $\phi = \phi_0$. Now, for the larger model, write

$$p_1(\theta, \phi) = p_1(\theta|\phi)\, p_1(\phi),$$

where $p_1(\theta|\phi)$ is the density of the conditional distribution of θ, given ϕ. In practice, it is reasonable to assume that the conditional density of θ, given ϕ, is continuous at $\phi = \phi_0$ (O'Hagan, 1994).

In order to obtain the marginal density of the data under Model 0 one must integrate the joint density of the observations, and of the free parameters (given $\phi = \phi_0$) with respect to the latter, to obtain

$$\begin{aligned} p(\mathbf{y}|\text{Model 0}) &= \int p(\mathbf{y}|\theta, \phi = \phi_0)\, p(\theta|\phi = \phi_0)\, d\theta \\ &= p(\mathbf{y}|\phi = \phi_0). \end{aligned} \tag{8.19}$$

For the larger model

$$\begin{aligned} p(\mathbf{y}|\text{Model 1}) &= \int \left[\int p(\mathbf{y}|\theta, \phi)\, p(\theta|\phi)\, d\theta \right] p_1(\phi)\, d\phi \\ &= \int p(\mathbf{y}|\phi)\, p_1(\phi)\, d\phi = E_\phi\left[p(\mathbf{y}|\phi)\right]. \end{aligned} \tag{8.20}$$

The expectation above is an average of the sampling model marginal densities (after integrating out θ) taken over all values of ϕ (other than ϕ_0) and with plausibility as conveyed by the prior density $p_1(\phi)$ under the larger model. The posterior probability of the null model is then

$$\begin{aligned} p(\text{Model 0}|\mathbf{y}) &= \frac{p(\mathbf{y}|\text{Model 0})\, \pi_0}{p(\mathbf{y}|\text{Model 0})\, \pi_0 + p(\mathbf{y}|\text{Model 1})\,(1 - \pi_0)} \\ &= \frac{p(\mathbf{y}|\phi = \phi_0)\, \pi_0}{p(\mathbf{y}|\phi = \phi_0)\, \pi_0 + \int p(\mathbf{y}|\phi)\, p_1(\phi)\, d\phi\,(1 - \pi_0)}, \end{aligned} \tag{8.21}$$

and $p(\text{Model 1}|\mathbf{y}) = 1 - p(\text{Model 0}|\mathbf{y})$.

Consider now the case where there is a single parameter, so that Model 0 poses $\phi = \phi_0$ and Model 1 corresponds to the "alternative" hypothesis $\phi \neq \phi_0$ (the problem then consists of one of evaluating the "sharp" null hypothesis $\phi = \phi_0$). Then (8.21) holds as well, with the only difference being that the marginal distributions of the data are calculated directly as

$$p(\mathbf{y}|\text{Model 0}) = p(\mathbf{y}|\phi = \phi_0),$$

and

$$p(\mathbf{y}|\text{Model 1}) = \int p(\mathbf{y}|\phi)\, p_1(\phi)\, d\phi. \tag{8.22}$$

It is instructive to study the consequences of using a vague prior distribution on the Bayes factor. Following O'Hagan (1994), suppose that ϕ is a scalar parameter on $(-\infty, \infty)$. Vague prior knowledge is expressed as the limit of a uniform distribution

$$p_1(\phi) = (2c)^{-1}, \text{ for } -c \leq \phi \leq c,$$

by letting $c \to \infty$, in which case, $p_1(\phi) \to 0$. Then (8.22) is

$$\int p(\mathbf{y}|\phi)\, p_1(\phi)\, d\phi = (2c)^{-1} \int_{-c}^{c} p(\mathbf{y}|\phi)\, d\phi.$$

Often, $p(\mathbf{y}|\phi)$ will tend to zero as ϕ tends to infinity, such that the limit of the integral above is finite. Then as $c \to \infty$, (8.22) tends to zero and the Bayes factor B_{01} tends to infinity. Thus, using a prior with very large spread on ϕ in an attempt to describe vague prior knowledge, forces the Bayes factor to favor Model 0.

Example 8.7 *Normal model: known versus unknown variance*

The setting will be the usual $N(\mu, \sigma^2)$ for each of n independent observations. In the larger model, both the mean and variance are taken as unknown. In the nested model, the variance is assumed to be known, such that $\sigma^2 = \sigma_0^2$. As in O'Hagan (1994), it will be assumed that the conditional prior distribution of the mean is the normal process $\mu|\mu_1, w\sigma^2 \sim N(\mu_1, w\sigma^2)$, where w is a known scalar. This implies that the variance of the prior distribution is proportional to that of the sampling model. Further, it will be assumed that the prior distribution of σ^2 is a scaled inverted chi-square distribution with parameters ν and S^2.

Under the null or nested model (known variance), the prior distribution is then $\mu|\mu_1, w\sigma_0^2 \sim N(\mu_1, w\sigma_0^2)$, and the marginal distribution of the data, following (8.19) and making use of (8.12), is

$$\begin{aligned} p(\mathbf{y}|\text{Model 0}) &= \int p(\mathbf{y}|\mu, \sigma_0^2) p(\mu|\mu_1, w\sigma_0^2)\, d\mu \\ &= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - \overline{y})^2\right] \\ \times \frac{1}{\sqrt{2\pi w\sigma_0^2}} &\int \exp\left[-\frac{n}{2\sigma_0^2}(\overline{y} - \mu)^2\right] \exp\left[-\frac{(\mu - \mu_1)^2}{2w\sigma_0^2}\right] d\mu. \end{aligned}$$

Combining the two quadratics in μ gives

$$p\left(\mathbf{y}|\text{Model 0}\right) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2\right]\frac{1}{\sqrt{2\pi w\sigma_0^2}}$$
$$\exp\left[-\frac{n}{2\sigma_0^2}\left(\frac{n}{\sigma_0^2}+\frac{1}{w\sigma_0^2}\right)^{-1}\frac{1}{w\sigma_0^2}\left(\overline{y}-\mu_1\right)^2\right]$$
$$\int \exp\left[-\frac{1}{2}\left(\frac{n}{\sigma_0^2}+\frac{1}{w\sigma_0^2}\right)\left(\mu-\widehat{\mu}\right)^2\right]d\mu,$$

where $\widehat{\mu}$ has the same form as in Example 8.3. After the integration is carried out, one gets

$$p\left(\mathbf{y}|\text{Model 0}\right) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2\right]\frac{1}{\sqrt{2\pi w\sigma_0^2}}$$
$$\exp\left[-\frac{1}{2}\frac{n}{\sigma_0^2}\left(\frac{n}{\sigma_0^2}+\frac{1}{w\sigma_0^2}\right)^{-1}\frac{1}{w\sigma_0^2}\left(\overline{y}-\mu_1\right)^2\right]\sqrt{2\pi\sigma_0^2\left(n+\frac{1}{w}\right)^{-1}}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \frac{1}{\sqrt{nw+1}}\exp\left[-\frac{Q_y}{2\sigma_0^2}\right] = p\left(\mathbf{y}|\sigma^2{=}\sigma_0^2\right), \tag{8.23}$$

where

$$Q_y = \sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2 + n\left(n+\frac{1}{w}\right)^{-1}\frac{1}{w}\left(\overline{y}-\mu_1\right)^2.$$

In order to obtain the marginal density of the data under Model 1, use is made of (8.20) and of (8.23), although noting that σ^2 is now a free parameter. Then, recalling that the prior distribution of σ^2 is scaled inverted chi-square

$$p\left(\mathbf{y}|\text{Model 1}\right) = \int p\left(\mathbf{y}|\sigma^2\right)p\left(\sigma^2|\nu,S^2\right)d\sigma^2$$
$$= \int \frac{\left(2\pi\sigma^2\right)^{-\frac{n}{2}}}{\left(nw+1\right)^{\frac{1}{2}}}\exp\left[-\frac{Q_y}{2\sigma^2}\right]\frac{\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\left(\sigma^2\right)^{-\left(\frac{\nu+2}{2}\right)}\exp\left(-\frac{\nu S^2}{2\sigma^2}\right)d\sigma^2$$
$$= \frac{\left(2\pi\right)^{-\frac{n}{2}}}{\left(nw+1\right)^{\frac{1}{2}}}\frac{\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)}\int\left(\sigma^2\right)^{-\left(\frac{n+\nu+2}{2}\right)}\exp\left(-\frac{\nu S^2+Q_y}{2\sigma^2}\right)d\sigma^2$$
$$= \frac{\left(2\pi\right)^{-\frac{n}{2}}}{\left(nw+1\right)^{\frac{1}{2}}}\frac{\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}}\Gamma\left(\frac{n+\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\nu S^2+Q_y}{2}\right)^{-\left(\frac{n+\nu}{2}\right)}. \tag{8.24}$$

In order to arrive at the last result, use is made of the gamma integrals (see Chapter 1). The Bayes factor in favor of Model 1 relative to Model 0

is given by the ratio between (8.24) and (8.23) yielding

$$B_{10} = \frac{\left(\frac{\nu S^2}{2}\right)^{\frac{\nu}{2}} \Gamma\left(\frac{n+\nu}{2}\right)}{\left(\sigma_0^2\right)^{-\frac{n}{2}} \exp\left[-\frac{Q_y}{2\sigma_0^2}\right] \Gamma\left(\frac{\nu}{2}\right)} \left(\frac{\nu S^2 + Q_y}{2}\right)^{-\left(\frac{n+\nu}{2}\right)}.$$

■

8.2.6 Approximations to the Bayes Factor

There is extensive literature describing various approximate criteria for Bayesian model selection. Some have been motivated by the desire for suppressing the dependence of the final results on the prior. Ease of computation has also been an important consideration, especially in the pre-MCMC era. Some of these methods are still a useful part of the toolkit for comparing models. This section introduces widely used approximations to the Bayes factor based on asymptotic arguments. The latter are based on regularity conditions which fail when the parameter lies on a boundary of its parameter space (Pauler et al., 1999), a restriction not encountered with the Bayes factor.

The marginal density of the data under Model i, say, is

$$p\left(\mathbf{y}|M_i\right) = \int p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) p\left(\boldsymbol{\theta}_i|M_i\right) d\boldsymbol{\theta}_i, \tag{8.25}$$

where $\boldsymbol{\theta}_i$ is the $p_i \times 1$ vector of parameters under this model. In what follows, it will be assumed that the dimension of the parameter vector does not increase with the number of observations or that, if this occurs, it does so in a manner that, for n being the number of observations, p_i/n goes to 0 as $n \to \infty$. This is important for asymptotic theory to hold. In the context of quantitative genetic applications, there are models in which the number of parameters, e.g., the additive genetic effects, increases as the number of observations increases. For such models the approximations hold provided that these effects are first integrated out, in which case $p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right)$ would be an integrated likelihood. For example, suppose a Gaussian linear model has f location parameters, n additive genetic effects (one for each individual), and two variance components. Then analytical integration of the additive effects (over their prior distribution) would need to be effected before proceeding. On the other hand, if the model is one of repeated measures taken on subjects or clusters (such as a family of half-sibs), it is reasonable to defend the assumption that p_i/n goes to 0 asymptotically.

Using the Posterior Mode

As in Chapter 7, expand the logarithm of the integrand in (8.25) around the posterior mode, $\widetilde{\boldsymbol{\theta}}_i$, using a second-order Taylor series expansion, to

obtain (recall that the gradient vanishes at the maximum value)

$$\begin{aligned}&\log\left[p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) p\left(\boldsymbol{\theta}_i|M_i\right)\right]\\ &\approx \log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right) p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right)\right] - \frac{1}{2}\left(\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\right)' \left(\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}\right)\left(\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\right), \quad (8.26)\end{aligned}$$

where $\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}$ is the corresponding negative Hessian matrix. Then, using this in (8.25),

$$\begin{aligned}p\left(\mathbf{y}|M_i\right) &= \int \exp\left\{\log\left[p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) p\left(\boldsymbol{\theta}_i|M_i\right)\right]\right\} d\boldsymbol{\theta}_i\\ &\approx \exp\left\{\log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right) p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right)\right]\right\}\\ &\times \int \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\right)' \left(\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}\right)\left(\boldsymbol{\theta}_i - \widetilde{\boldsymbol{\theta}}_i\right)\right] d\boldsymbol{\theta}_i.\end{aligned}$$

The integral is in a Gaussian form (this approach to integration is called Laplace's method for integrals), so it can be evaluated readily. Hence

$$p\left(\mathbf{y}|M_i\right) \approx p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right) p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right) (2\pi)^{\frac{p_i}{2}} \left|\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}^{-1}\right|^{\frac{1}{2}}, \quad (8.27)$$

where $\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}^{-1}$ is the variance–covariance matrix of the Gaussian approximation to the posterior distribution. Further

$$\begin{aligned}\log\left[p\left(\mathbf{y}|M_i\right)\right] &\approx \log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right)\right] + \log\left[p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right)\right]\\ &+ \frac{p_i}{2}\log\left(2\pi\right) + \frac{1}{2}\log\left(\left|\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}^{-1}\right|\right). \quad (8.28)\end{aligned}$$

Twice the logarithm of the Bayes factor for Model i relative to Model j, to express the "evidence brought up by the data" in support of Model i relative to j in the same scale as likelihood ratio tests, is then

$$\begin{aligned}2\log\left(B_{ij}\right) &\approx 2\log\left[\frac{p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_i, M_i\right)}{p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_j, M_j\right)}\right] + 2\log\frac{p\left(\widetilde{\boldsymbol{\theta}}_i|M_i\right)}{p\left(\widetilde{\boldsymbol{\theta}}_j|M_j\right)}\\ &+ \left(p_i - p_j\right)\log\left(2\pi\right) + \log\left(\frac{\left|\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_i}^{-1}\right|}{\left|\mathbf{H}_{\widetilde{\boldsymbol{\theta}}_j}^{-1}\right|}\right). \quad (8.29)\end{aligned}$$

Note that the criterion depends on the log-likelihood ratios (evaluated at the posterior modes), on the log-prior ratios (also evaluated at the modes), on the difference between the dimensions of the two competing models, and on a Hessian adjustment.

Using the Maximum Likelihood Estimator

A variant to approximation (8.26) is when the expansion of the logarithm of the product of the prior density and of the conditional distribution of the observations (given the parameters) is about the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_i$, instead of the mode of the posterior distribution (Tierney and Kadane, 1989; O'Hagan, 1994; Kass and Raftery, 1995). Here one obtains in (8.27),

$$p\left(\mathbf{y}|M_i\right) \approx p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right) p\left(\widehat{\boldsymbol{\theta}}_i|M_i\right) (2\pi)^{\frac{p_i}{2}} \left|\mathbf{H}^{-1}_{\widehat{\boldsymbol{\theta}}_i}\right|^{\frac{1}{2}}, \tag{8.30}$$

where $\mathbf{H}_{\widehat{\boldsymbol{\theta}}}$ is the observed information matrix evaluated at the maximum likelihood estimator. In particular, if the observations are i.i.d. one has $\mathbf{H}_{\widehat{\boldsymbol{\theta}}} = n\mathbf{H}_{1,\widehat{\boldsymbol{\theta}}}$, where $\mathbf{H}_{1,\widehat{\boldsymbol{\theta}}}$ is the observed information matrix calculated from a single observation. Then

$$p\left(\mathbf{y}|M_i\right) \approx p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right) p\left(\widehat{\boldsymbol{\theta}}_i|M_i\right) (2\pi)^{\frac{p_i}{2}} (n)^{-\frac{p_i}{2}} \left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}_i}\right|^{\frac{1}{2}}. \tag{8.31}$$

The approximation to twice the logarithm of the Bayes factor becomes

$$\begin{aligned} 2\log\left(B_{ij}\right) \approx 2\log\left[\frac{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right)}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_j, M_j\right)}\right] + 2\log\frac{p\left(\widehat{\boldsymbol{\theta}}_i|M_i\right)}{p\left(\widehat{\boldsymbol{\theta}}_j|M_j\right)} \\ - \left(p_i - p_j\right)\log\frac{n}{2\pi} + \log\frac{\left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}_i}\right|}{\left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}_j}\right|}. \end{aligned} \tag{8.32}$$

It is important to note that even though the asymptotic approximation to the posterior distribution (using the maximum likelihood estimator) does not depend on the prior, the resulting approximation to the Bayes factor does depend on the ratio of priors evaluated at the corresponding maximum likelihood estimators. If the term on the logarithm of the prior densities is excluded, the resulting expression is called the Bayesian information criterion (or BIC) (Schwarz, 1978; Kass and Raftery, 1995; Leonard and Hsu, 1999).

Suppose that the prior conveys some sort of "minimal" information represented by the distribution $\boldsymbol{\theta}_i|M_i \sim N\left(\widehat{\boldsymbol{\theta}}_i, \mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}}\right)$. This is a unit information prior centered at the maximum likelihood estimator and having a precision (inverse of the covariance matrix) equivalent to that brought up

by a sample of size $n = 1$. Using this in (8.31):

$$p(\mathbf{y}|M_i) \approx p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right)(2\pi)^{-\frac{p_i}{2}}\left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}}\right|^{-\frac{1}{2}}$$
$$\times \exp\left[-\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)'\left(\mathbf{H}_{1,\widehat{\boldsymbol{\theta}}1,\widehat{\boldsymbol{\theta}}}\right)\left(\widehat{\boldsymbol{\theta}}-\widehat{\boldsymbol{\theta}}\right)\right](2\pi)^{\frac{p_i}{2}}(n)^{-\frac{p_i}{2}}\left|\mathbf{H}^{-1}_{1,\widehat{\boldsymbol{\theta}}}\right|^{\frac{1}{2}}$$
$$= p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right)(n)^{-\frac{p_i}{2}}. \tag{8.33}$$

Hence

$$2\log(B_{ij}) \approx 2\log\left[\frac{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_i, M_i\right)}{p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_j, M_j\right)}\right] - (p_i - p_j)\log n. \tag{8.34}$$

This is Schwarz (1978) BIC in its most commonly presented form (Kass and Raftery, 1995; O'Hagan, 1994). Some authors (Leonard and Hsu, 1999; Congdon, 2001) use the term BIC to refer just to the approximated marginal densities, e.g., the logarithm of (8.33). At any rate, note that (8.34) is twice the maximized log-likelihood ratio, plus an adjustment that penalizes the model with more parameters. If $n = 1$, there is no penalty. However, the term $(p_i - p_j)$ becomes more important as a sample size increases. When $p_i > p_j$, (8.34) is smaller than twice the log-likelihood ratio, so the adjustment favors parsimony. In contrast, classical testing based on the traditional likelihood ratio tends to favor the more complex models. Contrary to the traditional likelihood ratio, BIC is well defined for nonnested models.

Denoting S the right hand side of (8.34) divided by 2, as sample size $n \to \infty$, this quantity satisfies:

$$\frac{S - \log B_{ij}}{\log B_{ij}} \to 0,$$

so it is consistent in this sense (Kass and Raftery, 1995). Recent extensions of BIC can be found in Kass (1995).

A related criterion is AIC (or the Akaike's information criterion) (Akaike, 1973), where the penalty is $2(p_i - p_j)$. The argument underlying the AIC is that if two models favor the data equally well, then the more parsimonious one should be favored. The BIC produces an even more drastic penalty, which increases with sample size, as noted.

The differences between the likelihood ratio criterion, the BIC, and the AIC are discussed by O'Hagan (1994) in the context of a nested model. The larger model has parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ and dimension p_2, whereas the "smaller or null" model has a parameter vector $\boldsymbol{\theta}$ with p_1 elements and $p_2 - p_1$ fixed components $\boldsymbol{\phi} = \boldsymbol{\phi}_0$. For a large sample size, the log-likelihood ratio may favor the larger model, yet the penalty, $(p_2 - p_1)\log n$, may be severe enough so that the Bayes factor may end up favoring the null model.

It is instructive to examine the behavior of the approximation to the Bayes factor under repeated sampling from the appropriate model. Consider the BIC as given in (8.32), and take its expected value under the null model, with only the likelihood ratio viewed as a random variable. Recalling that the expected value of twice the log-likelihood ratio statistic under the null hypothesis is equal to the difference in dimension between the competing models, or $(p_2 - p_1)$, one gets

$$E\left[2\log\left(B_{21}\right)\right] \approx 2\log\frac{p\left(\widehat{\boldsymbol{\theta}}_2|M_2\right)}{p\left(\widehat{\boldsymbol{\theta}}_1|M_1\right)} - \left(p_2 - p_1\right)\left[\log\frac{n}{2\pi} - 1\right] + \text{constant}.$$

Hence, as $n \to \infty$, the expected value of the log of the Bayes factor in favor of the larger model goes to $-\infty$. This implies that the posterior probability of the larger model goes to 0 when the null model is true, regardless of the prior odds ratios as conveyed by $p\left(\widehat{\boldsymbol{\theta}}_2|M_2\right)/p\left(\widehat{\boldsymbol{\theta}}_1|M_1\right)$. Conversely, when the larger model is true, the expected value of twice the log-likelihood ratio statistic is approximately equal to $nQ\left(\boldsymbol{\phi}, \boldsymbol{\phi}_0\right)$,where $Q\left(\cdot\right)$ is a quadratic form (O'Hagan, 1994). This is a consequence of the asymptotically normal distribution of the maximum likelihood estimator (see Chapters 3 and 4). Then, under the larger model,

$$E\left[2\log\left(B_{ij}\right)\right] \approx nQ\left(\boldsymbol{\phi}, \boldsymbol{\phi}_0\right) + 2\log\frac{p\left(\widehat{\boldsymbol{\theta}}_i|M_i\right)}{p\left(\widehat{\boldsymbol{\theta}}_j|M_j\right)} - \left(p_2 - p_1\right)\log\frac{n}{2\pi} + \text{constant}.$$

As $n \to \infty$, the logarithm of the Bayes factor in favor of the larger model goes to ∞, since n grows faster than $\log n$. Consequently, the posterior probability of the larger model goes to 1, no matter what the prior odds are. Strictly from a classical point of view, and no matter how large n is, the null model will be rejected with probability equal to the significance level even when the model is true. Hence, more stringent significance levels should be adopted in classical hypothesis testing when sample sizes are large. Classical theory does not give a procedure for modifying the type-1 error as a function of sample size, and the probability of this error is prescribed arbitrarily. As noted by O'Hagan (1994), the Bayesian approach gives an automatic procedure in which in a single formula, such as (8.32), the evidence from the data, the prior odds, the model dimensionality, and the sample size are combined automatically.

8.2.7 Partial and Intrinsic Bayes Factors

The Bayes factor is only defined up to arbitrary constants when prior distributions are improper (i.e., Berger and Pericchi, 1996), as was illustrated at the end of Subsection 8.2.5. Further, when the priors are proper, the

Bayes factor depends on the form of the chosen prior distribution, as seen in connection with (8.32). This dependence does not decrease as sample size increases, contrary to the case of estimation of parameters from posterior distributions. In estimation problems and under regularity conditions, one can obtain an asymptotic approximation centered at the maximum likelihood estimator that does not involve the prior.

Berger and Pericchi (1996) suggested what are called intrinsic Bayes factors, in an attempt to circumvent the dependence on the prior, and to allow for the use of improper prior distributions, such as those based on Jeffreys' rule. Here, a brief overview of one of the several proposed types of Bayes factors (the arithmetic intrinsic Bayes factor) is presented.

Let the data vector of order n be partitioned as

$$\mathbf{y} = \left[\mathbf{y}'_{(1)}, \mathbf{y}'_{(2)}, \ldots, \mathbf{y}'_{(L)}\right]',$$

where $\mathbf{y}_{(l)}$, $(l = 1, 2, ..., L)$ denotes what is called the minimal training sample. This is the minimal number of observations needed for the posterior distribution to be proper. For example, if the minimal size of the training sample is m, there would be C_m^n different possible training samples. The posterior distribution based on the minimal training sample has density $p\left(\theta_i|\mathbf{y}_{(l)}, M_i\right)$. Further, put

$$\mathbf{y} = \left[\mathbf{y}'_{(l)}, \mathbf{y}'_{(-l)}\right]',$$

where $\mathbf{y}_{(-l)}$ is the data vector with $\mathbf{y}_{(l)}$ removed. Then the predictive density of $\mathbf{y}_{(-l)}$ under model i, conditionally on the data of the training sample $\mathbf{y}_{(l)}$, is

$$p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_i\right) = \int p\left(\mathbf{y}_{(-l)}|\theta_i, \mathbf{y}_{(l)}, M_i\right) p\left(\theta_i|\mathbf{y}_{(l)}, M_i\right) d\theta_i.$$

The Bayes factor for model j relative to model i, conditionally on $\mathbf{y}_{(l)}$, or partial Bayes factor (O'Hagan, 1994) is

$$B_{ji}\left(\mathbf{y}_{(l)}\right) = \frac{p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_j\right)}{p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_i\right)}. \tag{8.35}$$

Clearly, the partial Bayes factor depends on the choice of the training sample $\mathbf{y}_{(l)}$. To eliminate this dependence, Berger and Pericchi (1996) propose averaging $B_{ji}\left(\mathbf{y}_{(l)}\right)$ over all $C_m^n = K$ training samples. This yields the arithmetic intrinsic Bayes factor, defined formally as

$$B_{ji}^{AI} = \frac{1}{K}\sum_{l=1}^{K} \frac{p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_j\right)}{p\left(\mathbf{y}_{(-l)}|\mathbf{y}_{(l)}, M_i\right)}. \tag{8.36}$$

This expression can be computed for any pair of models, irrespective of whether these are nested or not. Although the procedure is appealing, some difficulties arise. First, for most realistic hierarchical models it is not possible to determine in advance what the minimum sample size should be in order for the posterior to be proper. Second, and especially in animal breeding, the data sets are very large so, at best, just a few minimal training samples could be processed in practice.

There have been several other attempts to circumvent the need to using proper priors and to restrict the dependence on the prior. These are reviewed in O'Hagan (1994).

8.3 Estimating the Marginal Likelihood from Monte Carlo Samples

Except in highly stylized models, the integration indicated in (8.25) is not feasible by analytical means. An alternative is to use Monte Carlo methods. Here we shall consider the method of importance sampling, which will be encountered again in Chapters 12 and 15, where more details on the technique are given. Suppose samples of $\boldsymbol{\theta}_i$, the parameter vector under Model i, can be obtained from some known distribution that is relatively easy to sample from. This distribution, having the same support as the prior or posterior, is called the importance sampling distribution, and its density will be denoted as $g(\boldsymbol{\theta}_i)$. Then since $\int p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i = 1$, the marginal density of the data under Model i is expressible as

$$\begin{aligned} p(\mathbf{y}|M_i) &= \frac{\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i}{\int p(\boldsymbol{\theta}_i|M_i)\, d\boldsymbol{\theta}_i} \\ &= \frac{\int p(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\, \frac{p(\boldsymbol{\theta}_i|M_i)}{g(\boldsymbol{\theta}_i)} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}{\int \frac{p(\boldsymbol{\theta}_i|M_i)}{g(\boldsymbol{\theta}_i)} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i}. \end{aligned} \tag{8.37}$$

Various Monte Carlo sampling schemes can be derived from (8.37), depending on the importance sampling function adopted. Suppose m samples can be obtained from the distribution with density $g(\boldsymbol{\theta}_i)$; let the samples be $\boldsymbol{\theta}_i^{[j]}$, $(j = 1, 2, \ldots, m)$. Then note that the denominator of (8.37) can be written as

$$\int \frac{p(\boldsymbol{\theta}_i|M_i)}{g(\boldsymbol{\theta}_i)} g(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i = \lim_{m\to\infty} \left[\frac{1}{m} \sum_{j=1}^{m} \frac{p\left(\boldsymbol{\theta}_i^{[j]}|M_i\right)}{g\left(\boldsymbol{\theta}_i^{[j]}\right)} \right],$$

where $p\left(\boldsymbol{\theta}_i^{[j]}|M_i\right)$ is the prior density under Model i evaluated at sampled value j. Likewise, the numerator can be written as

$$\begin{aligned}&\int p\left(\mathbf{y}|\boldsymbol{\theta}_i, M_i\right) \frac{p\left(\boldsymbol{\theta}_i|M_i\right)}{g\left(\boldsymbol{\theta}_i\right)} g\left(\boldsymbol{\theta}_i\right) d\boldsymbol{\theta}_i \\ &= \lim_{m\to\infty}\left[\frac{1}{m}\sum_{j=1}^{m} p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]}, M_i\right) \frac{p\left(\boldsymbol{\theta}_i^{[j]}|M_i\right)}{g\left(\boldsymbol{\theta}_i^{[j]}\right)}\right],\end{aligned}$$

where $p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]}, M_i\right)$ is the density of the sampling model evaluated at the jth sample obtained from the importance distribution. Hence for large m, and putting $w_i^{[j]} = p\left(\boldsymbol{\theta}_i^{[j]}|M_i\right)/g\left(\boldsymbol{\theta}_i^{[j]}\right)$, a consistent estimator of (8.37) is given by the ratio

$$\widehat{p}\left(\mathbf{y}|M_i\right) = \frac{\sum_{j=1}^{m} w_i^{[j]} p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]}, M_i\right)}{\sum_{j=1}^{m} w_i^{[j]}}, \tag{8.38}$$

which is a weighted average of the density of the sampling distribution evaluated at the corresponding sampled values of the parameter vector under the appropriate model.

Sampling from the Prior

If the importance distribution is the prior, each of the weights $w_i^{[j]}$ are equal to 1, and the Monte Carlo estimator (8.38) of the marginal density at the observed value of $\mathbf{y}$ becomes

$$\widehat{p}\left(\mathbf{y}|M_i\right) = \frac{1}{m}\sum_{j=1}^{m} p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]}, M_i\right), \tag{8.39}$$

where the $\boldsymbol{\theta}_i^{[j]}$ are draws from the prior distribution. The procedure is very simple because the joint prior distribution of the parameters is often simple to sample from. However, the estimator is imprecise because, typically, the $\boldsymbol{\theta}_i^{[j]}$ drawn from the prior are conferred little likelihood by the data. There will be just a few draws that will have appreciable likelihood and these will "dominate" the average (Kass and Raftery, 1995). Numerical studies can be found in McCulloch and Rossi (1991).

Sampling from the Posterior

If the importance distribution is the posterior, then

$$w_i = \frac{p\left(\boldsymbol{\theta}_i|M_i\right)}{p\left(\boldsymbol{\theta}_i|\mathbf{y},M_i\right)} = \frac{p\left(\boldsymbol{\theta}_i|M_i\right)}{\frac{p(\mathbf{y}|\boldsymbol{\theta}_i,M_i)p(\boldsymbol{\theta}_i|M_i)}{p(\mathbf{y}|M_i)}} = \frac{p\left(\mathbf{y}|M_i\right)}{p\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)}.$$

Using this in (8.38):

$$\begin{aligned}\widehat{p}\left(\mathbf{y}|M_i\right) &= \frac{\sum_{j=1}^{m} \frac{p(\mathbf{y}|M_i)}{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)} p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)}{\sum_{j=1}^{m} \frac{p(\mathbf{y}|M_i)}{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)}} \\ &= \frac{m}{\sum_{j=1}^{m} \frac{1}{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)}} \\ &= \left[\frac{1}{m}\sum_{j=1}^{m} \frac{1}{p\left(\mathbf{y}|\boldsymbol{\theta}_i^{[j]},M_i\right)}\right]^{-1}. \end{aligned} \tag{8.40}$$

This estimator, the harmonic mean of the likelihood values, was derived by Newton and Raftery (1994), but arguing directly from Bayes theorem. Observe that a rearrangement of the theorem leads to

$$\frac{p\left(\boldsymbol{\theta}_i|M_i\right)}{p\left(\mathbf{y}|M_i\right)} = \frac{p\left(\boldsymbol{\theta}_i|\mathbf{y},M_i\right)}{p\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)}.$$

Then, integrating both sides with respect to $\boldsymbol{\theta}_i$, yields

$$\frac{1}{p\left(\mathbf{y}|M_i\right)} \int p\left(\boldsymbol{\theta}_i|M_i\right) d\boldsymbol{\theta}_i = \int \frac{1}{p\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)} p\left(\boldsymbol{\theta}_i|\mathbf{y},M_i\right) d\boldsymbol{\theta}_i.$$

Since the prior must be proper for the marginal density of the data to be defined, the integral on the left is equal to 1 leading directly to

$$p\left(\mathbf{y}|M_i\right) = \frac{1}{E_{\boldsymbol{\theta}_i|\mathbf{y},M_i}\left[p^{-1}\left(\mathbf{y}|\boldsymbol{\theta}_i,M_i\right)\right]}. \tag{8.41}$$

The Monte Carlo estimator of the reciprocal of the posterior expectation of the reciprocal of the likelihood values is precisely (8.40). An advantage of the harmonic mean estimator is that one does not need to know the form of the posterior distribution. The Markov chain Monte Carlo methods presented in the next part of the book enable one to draw samples from complex, unknown, distributions. The disadvantage, however, is its

numerical instability. The form of (8.40) reveals that values of $\boldsymbol{\theta}_i$ with very small likelihood can have a strong impact on the estimator. An alternative is to form some robust estimator of the harmonic mean (Congdon, 2001) such as a trimmed average. Kass and Raftery (1995) state that, in spite of the lack of stability, the estimator is accurate enough for interpretation on a logarithmic scale.

Caution must be exercised in the actual computation of (8.40), to avoid numerical over- or under-flows. A possible strategy could be as follows. Let

$$\begin{aligned} v &= \frac{1}{m}\sum_{j=1}^{m} p^{-1}\left(\mathbf{y}|\boldsymbol{\theta}^{[j]}, M_i\right) \\ &= \frac{1}{m}\sum_{j=1}^{n} S_i^{[j]}, \end{aligned}$$

where $S_i^{[j]} = p^{-1}\left(\mathbf{y}|\boldsymbol{\theta}^{[j]}, M_i\right)$, and store $\log S_i^{[j]}$ in a file for each sampled value. Then, since

$$\exp(x) = \exp(x - c + c) = \exp(x - c)\exp c,$$

one can write v in the form

$$v = \frac{1}{m}\sum_{j=1}^{m} \exp\left(\log S_i^{[j]} - c\right)\exp c,$$

where c is the largest value of $\log S_i^{[j]}$. Taking logarithms yields

$$\log v = \log\left[\frac{1}{m}\sum_{j=1}^{m} \exp\left(\log S_i^{[j]} - c\right)\right] + c.$$

Hence

$$\log\left[\widehat{p}\,(\mathbf{y}|M_i)\right] = -\log v.$$

Chib's Method

Most often, the marginal posterior distributions cannot be identified. However, there are many models where the conditional posterior distributions can be arrived at from inspection of the joint posterior densities. Advantage of this is taken in a Markov chain-based method called the Gibbs sampler, which will be introduced in Chapter 11. Chib (1995) outlined a procedure for estimating the marginal density of the data under a given model when the fully conditional posterior distributions can be identified. These distributions are defined in Section 11.5.1 of Chapter 11. We will

suppress the dependency on the model in the notation, for simplicity. Suppose the parameter vector is partitioned as $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right]'$. The logarithm of the marginal density of the data can be expressed as

$$\begin{aligned}\log p(\mathbf{y}) &= \log\left[p(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right] + \log\left[p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right] - \log\left[p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})\right] \\ &= \log\left[p(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right] + \log\left[p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right] - \log\left[p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})\right] - \log\left[p(\boldsymbol{\theta}_1|\mathbf{y})\right].\end{aligned}$$

Suppose now that samples of $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ have been drawn from the posterior distribution using the Gibbs sampler. Inspection of a large number of samples permits us to calculate, e.g., the posterior mean, mode, or median for each of the elements of the parameter vector, such that one can form, say, the vector of posterior medians $\widetilde{\boldsymbol{\theta}} = \left[\widetilde{\boldsymbol{\theta}}_1', \widetilde{\boldsymbol{\theta}}_2'\right]'$. An estimate of the marginal density of the data can be obtained as

$$\begin{aligned}\log \widehat{p}(\mathbf{y}) = \log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)\right] + \log\left[p\left(\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)\right] - \log\left[p\left(\widetilde{\boldsymbol{\theta}}_2|\widetilde{\boldsymbol{\theta}}_1, \mathbf{y}\right)\right] \\ - \log\left[p\left(\widetilde{\boldsymbol{\theta}}_1|\mathbf{y}\right)\right].\end{aligned} \tag{8.42}$$

If the conditional density $\log\left[p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{y})\right]$ is known, the third term can be evaluated readily. The difficulty resides in the fact that the marginal posterior density may not be known. However, recall that

$$p(\boldsymbol{\theta}_1|\mathbf{y}) = E_{\boldsymbol{\theta}_2|\mathbf{y}}\left[p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \mathbf{y})\right].$$

Hence

$$p\left(\widetilde{\boldsymbol{\theta}}_1|\mathbf{y}\right) = E_{\boldsymbol{\theta}_2|\mathbf{y}}\left[p\left(\widetilde{\boldsymbol{\theta}}_1|\boldsymbol{\theta}_2, \mathbf{y}\right)\right],$$

and an estimate of the marginal posterior density can be obtained as

$$\widehat{p}\left(\widetilde{\boldsymbol{\theta}}_1|\mathbf{y}\right) = \frac{1}{m}\sum_{j=1}^{m} p\left(\widetilde{\boldsymbol{\theta}}_1|\boldsymbol{\theta}_2^{[j]}, \mathbf{y}\right),$$

where $\boldsymbol{\theta}_2^{[j]}$, $(j = 1, 2, \ldots, m)$ are samples from the marginal posterior distribution of $\boldsymbol{\theta}_2$ obtained with the Gibbs sampler. Then, using this in (8.42), the estimated marginal density of the data is arrived at as

$$\begin{aligned}\log \widehat{p}(\mathbf{y}) = \log\left[p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)\right] + \log\left[p\left(\widetilde{\boldsymbol{\theta}}_1, \widetilde{\boldsymbol{\theta}}_2\right)\right] - \log\left[p\left(\widetilde{\boldsymbol{\theta}}_2|\widetilde{\boldsymbol{\theta}}_1, \mathbf{y}\right)\right] \\ - \log\left[\frac{1}{m}\sum_{j=1}^{m} p\left(\widetilde{\boldsymbol{\theta}}_1|\boldsymbol{\theta}_2^{[j]}, \mathbf{y}\right)\right].\end{aligned} \tag{8.43}$$

The procedure is then repeated for each of the models in order to calculate the Bayes factor. However, the fully conditional posterior distribution of one parameter given the other must be identifiable in each of the models. The method can be extended from two to several parameter blocks (Chib, 1995; Han and Carlin, 2001). Additional refinements of the procedure are in Chib and Jeliazkov (2001).

8.4 Goodness of Fit and Model Complexity

In general, as a model becomes increasingly more complex, i.e., by increasing the number of parameters, its fit gets better. For example, it is well known that if one fits n regression coefficients to a data set consisting of n points, the fit is perfect. As seen earlier, the AIC and BIC introduce penalties against more highly parameterized models. A different approach was suggested by Spiegelhalter et al. (2002), and it is based on calculating the expected posterior deviance, i.e., a measure of fit (Dempster, 1974, 1997).

Consider a model with parameter vector $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right]'$. For example, in a mixed linear model $\boldsymbol{\theta}_1$, may be a vector of "fixed" effects such as breed or sex of animal and variance parameters, while $\boldsymbol{\theta}_2$ may be a vector of random effects or missing data. Hence, in some sense, the dimension of $\boldsymbol{\theta}_1$ can be viewed as fixed, as the dimension of the data vector increases, whereas the order of $\boldsymbol{\theta}_2$ may perhaps increase as the dimension of the data vector increases. Clearly, because AIC and BIC are based on asymptotic results, neither can be used in models where the parameters outnumber the observations (Gelfand and Dey, 1994) unless some parameters are integrated out.

There is also the additional problem of interpreting exactly what is the number of parameters, especially in complex hierarchical models. For example, consider a setting with two breeds, individuals within breeds and several observations per individual. Observations within individuals are typically correlated, and their contribution to the total number of parameters is difficult to specify. One could imagine that their "effective contribution" would depend on the degree of correlation; the larger the correlation, the smaller their "effective contribution". The concept of effective number of parameters becomes even more elusive, if individuals are correlated as well due family structure.

Partly to address this problem, Spiegelhalter et al. (2002) suggest an alternative procedure for model comparison that they term deviance information criterion (DIC). For a particular model, it is defined as

$$DIC = 2\overline{D} - D\left(\overline{\boldsymbol{\theta}}\right). \tag{8.44}$$

As we shall see below, the DIC is the result of adding two expressions. The first one represents the fit of the model, summarized by $\overline{D}$. The second one is a measure of the complexity of the model or effective number of parameters, summarized by $\overline{D} - D\left(\overline{\boldsymbol{\theta}}\right)$. In (8.44),

$$\begin{aligned}\overline{D} &= -2\int \left[\log p\left(\mathbf{y}|\boldsymbol{\theta}\right)\right] p\left(\boldsymbol{\theta}|\mathbf{y}\right) d\boldsymbol{\theta} \\ &= E_{\boldsymbol{\theta}|\mathbf{y}}\left[-2\log p\left(\mathbf{y}|\boldsymbol{\theta}\right)\right] \\ &= E_{\boldsymbol{\theta}|\mathbf{y}}\left[D\left(\boldsymbol{\theta}\right)\right],\end{aligned} \tag{8.45}$$

where $D(\boldsymbol{\theta}) = -2\log p(\mathbf{y}|\boldsymbol{\theta})$ is called the deviance (a function of the unknown parameter), $\overline{D}$ is its expected value taken over the posterior distribution of $\boldsymbol{\theta}$, and $D\left(\overline{\boldsymbol{\theta}}\right)$ is the deviance evaluated at the posterior mean of the parameter vector $\boldsymbol{\theta}$. Note that when the deviance is evaluated at the maximum likelihood estimator, one obtains the numerator (or denominator) of the usual likelihood ratio statistic. Thus, one averages out the deviance criterion over values whose plausibilities are dictated by the posterior distribution. The expected deviance is interpreted as a posterior summary of the fit of the model. In general, $\overline{D}$ will need to be computed using Monte Carlo procedures for sampling from the posterior distribution: samples from the posterior are obtained, and then one averages the log-likelihoods evaluated at each of the draws.

As a measure of model complexity (degree of parameterization), Spiegelhalter et al. (2002) suggest using the "effective number of parameters"

$$p_D = \overline{D} - D\left(\overline{\boldsymbol{\theta}}\right). \tag{8.46}$$

In order to motivate this concept, expand the deviance around the posterior mean $\overline{\boldsymbol{\theta}}$, to obtain

$$\begin{aligned} D(\boldsymbol{\theta}) \approx & -2\log p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}\right) - 2\left[\frac{\partial \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]'_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\left(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\right) \\ & -\left(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\right)'\left[\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]\left(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\right). \end{aligned} \tag{8.47}$$

Taking the expectation of (8.47), with respect to the posterior distribution of the parameter vector, gives the expected deviance

$$\begin{aligned} \overline{D} &\approx -2\log p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}\right) + tr\left[-\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} Var(\boldsymbol{\theta}|\mathbf{y}) \\ &= D\left(\overline{\boldsymbol{\theta}}\right) + tr\left\{\left[-\frac{\partial^2 \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} Var(\boldsymbol{\theta}|\mathbf{y})\right\} \\ &= D\left(\overline{\boldsymbol{\theta}}\right) + tr\left\{\left[\mathbf{I}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} Var(\boldsymbol{\theta}|\mathbf{y})\right\}, \end{aligned} \tag{8.48}$$

where $\mathbf{I}(\boldsymbol{\theta})$ is the observed information matrix and $Var(\boldsymbol{\theta}|\mathbf{y})$ is the variance–covariance matrix of the posterior distribution. The trace adjustment in (8.48) is called the "effective number of parameters" and is denoted p_D. Recall from Chapter 7, that an asymptotic approximation to the posterior distribution is given by a normal process having a covariance matrix that is equal to the inverse of the sum of the observed information matrix (evaluated at some mode), plus the negative Hessian of the log-prior density (evaluated at some mode); the latter will be denoted as $\mathbf{P}(\boldsymbol{\theta})_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}$ when evaluated at the posterior mean. Hence, approximately,

$$\begin{aligned} p_D &\approx tr\left\{\left[\mathbf{I}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} Var(\boldsymbol{\theta}|\mathbf{y})\right\} \\ &\approx tr\left\{\left[\mathbf{I}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\left(\left[\mathbf{I}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} + \mathbf{P}(\boldsymbol{\theta})_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\right)^{-1}\right\}. \end{aligned} \tag{8.49}$$

Thus, the effective number of parameters can be interpreted as the information about $\boldsymbol{\theta}$ contained in the likelihood relative to the total information in both the likelihood and the prior. Some additional algebra yields

$$\begin{aligned} p_D &\approx tr\left\{\left([\mathbf{I}(\boldsymbol{\theta})]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} + \mathbf{P}(\boldsymbol{\theta})_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} - \mathbf{P}(\boldsymbol{\theta})_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\right)\right. \\ &\quad \left.\times \left([\mathbf{I}(\boldsymbol{\theta})]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} + \mathbf{P}(\boldsymbol{\theta})_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\right)^{-1}\right\} \\ &= p - tr\left[\mathbf{P}(\boldsymbol{\theta})_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\left([\mathbf{I}(\boldsymbol{\theta})]_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}} + \mathbf{P}(\boldsymbol{\theta})_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}}\right)^{-1}\right]. \end{aligned} \tag{8.50}$$

This representation leads to the interpretation that the effective number of parameters is equal to the number of parameters in $\boldsymbol{\theta}$, minus an adjustment measuring the amount of information in the prior relative to the total information contained in the asymptotic approximation to the posterior. The asymptotic justification of DIC presented here holds in cases where the number of observations grows with respect to the number of parameters in $\boldsymbol{\theta}$.

Spiegelhalter et al. (2002) suggest combining p_D with $\overline{D}$ into the deviance information criterion (DIC) which can also be written as

$$\begin{aligned} DIC &= \overline{D} + p_D \\ &= D(\boldsymbol{\theta}) + 2p_D, \end{aligned} \tag{8.51}$$

with the last expression resulting from (8.46). The value of DIC for a particular model does not convey meaning but rather the difference in DIC across models. Models having a smaller DIC should be favored, as this indicates a better fit and a lower degree of model complexity.

While the DIC is appealing and very easy to calculate using MCMC, a number of issues need to be resolved before it becomes fully accepted as a model comparison tool. For example, in complex multilevel models, different degrees of marginalization are possible, and these lead to different values of DIC for the same model. There is also the problem of determining the Monte Carlo variance of the DIC, which at the moment can only be estimated replicating the MCMC runs. The authors emphasize that they consider DIC to be a preliminary device for screening alternative models.

Example 8.8 *Deviance information criterion in the mixed linear model* Consider a hierarchical model with structure

$$\mathbf{y} = \mathbf{W}\boldsymbol{\theta} + \mathbf{e},$$

where $\mathbf{y}|\boldsymbol{\theta}, \mathbf{R} \sim N(\mathbf{W}\boldsymbol{\theta}, \mathbf{R})$. This model has been discussed several times, especially in Chapter 6. In animal breeding $\boldsymbol{\theta} = \left[\boldsymbol{\beta}', \mathbf{u}'\right]'$ is typically a vector of "fixed" and "random" effects, and the corresponding known incidence matrix is then $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$. Suppose that the dimension of $\boldsymbol{\beta}$ (p_β) does not

increase with the number of observations, and that the vector $\mathbf{u}$ (having order p_u) contains the effects of clusters, e.g., half-sib families. Hence, one can conceptually let the number of observations per cluster go to infinity (or, equivalently, think that the number of observations increases more rapidly than the number of clusters). Under these conditions, one can employ the asymptotic approximations discussed earlier. The second level of the hierarchy poses

$$\boldsymbol{\theta}|\boldsymbol{\mu}_{\boldsymbol{\beta}},\boldsymbol{\mu}_{\mathbf{u}},\mathbf{V}_{\beta},\sigma^2_{\beta},\mathbf{G}_u \sim N\left(\left[\begin{array}{c}\boldsymbol{\mu}_{\beta}\\ \boldsymbol{\mu}_u\end{array}\right],\left[\begin{array}{cc}\mathbf{V}_{\beta}\sigma^2_{\beta} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_u\end{array}\right]\right).$$

The dispersion parameters $\mathbf{R},\mathbf{V}_{\beta},\sigma^2_{\beta},\mathbf{G}_u$, and the location vectors $\boldsymbol{\mu}_{\beta}$ and $\boldsymbol{\mu}_u$ are assumed known. As mentioned in Chapter 1, Example 1.18, and shown in Chapter 6, the posterior distribution of $\boldsymbol{\theta}$ is normal, with mean vector

$$\overline{\boldsymbol{\theta}}=\left[\begin{array}{c}\overline{\boldsymbol{\beta}}\\ \overline{\mathbf{u}}\end{array}\right]=\left[\begin{array}{cc}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\frac{\mathbf{V}^{-1}_{\beta}}{\sigma^2_{\beta}} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}_u\end{array}\right]^{-1}$$
$$\times\left[\begin{array}{c}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}+\frac{\mathbf{V}^{-1}_{\beta}}{\sigma^2_{\beta}}\boldsymbol{\mu}_{\beta}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}+\mathbf{G}^{-1}_{\mathbf{u}}\boldsymbol{\mu}_u\end{array}\right],$$

and variance–covariance matrix

$$\mathbf{C}^{-1}=\left[\begin{array}{cc}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}+\frac{1}{\sigma^2_{\beta}}\mathbf{V}^{-1}_{\beta} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}_u\end{array}\right]^{-1}.$$

The deviance is

$$\begin{aligned}D(\boldsymbol{\theta}) &= -2\log p(\mathbf{y}|\boldsymbol{\theta},\mathbf{R})\\ &= N\log(2\pi)+\log|\mathbf{R}|+(\mathbf{y}-\mathbf{W}\boldsymbol{\theta})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{W}\boldsymbol{\theta}).\end{aligned}$$

Then,

$$D\left(\overline{\boldsymbol{\theta}}\right)=N\log(2\pi)+\log|\mathbf{R}|+\left(\mathbf{y}-\mathbf{W}\overline{\boldsymbol{\theta}}\right)'\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{W}\overline{\boldsymbol{\theta}}\right),$$

and the expected deviance becomes

$$\begin{aligned}\overline{D}=N\log(2\pi)+\log|\mathbf{R}|+\left(\mathbf{y}-\mathbf{W}\overline{\boldsymbol{\theta}}\right)'\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{W}\overline{\boldsymbol{\theta}}\right)\\ +tr\left(\mathbf{R}^{-1}\mathbf{W}\mathbf{C}^{-1}\mathbf{W}'\right).\end{aligned}$$

Employing (8.46), the effective number of parameters is

$$\begin{aligned}p_D &= \overline{D}-D\left(\overline{\boldsymbol{\theta}}\right)\\ &= tr\left(\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W}\right).\end{aligned}$$

For example, let $\mathbf{R} = \mathbf{I}\sigma_e^2$ and $\mathbf{G}_u = \mathbf{I}\sigma_u^2$, which results in a variance component model. Further, let $\sigma_\beta^2 \to \infty$, to make prior information about $\boldsymbol{\beta}$ vague. Here

$$\begin{aligned}\mathbf{C}^{-1} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_u^2}\mathbf{I} \end{bmatrix}^{-1} \sigma_e^2 \\ &= \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \sigma_e^2,\end{aligned}$$

and

$$\begin{aligned}\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W} &= \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_u^2}\mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} \end{bmatrix} \\ &= \mathbf{I}_{p_\beta+p_u} - \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}_{p_u} \end{bmatrix}.\end{aligned}$$

Hence

$$\begin{aligned}p_D &= tr\left(\mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{W}\right) \\ &= tr\left(\mathbf{I}_{p_\beta+p_u}\right) - tr\left\{ \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta u} \\ \mathbf{C}^{u\beta} & \mathbf{C}^{uu} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_e^2}{\sigma_u^2}\mathbf{I}_{p_u} \end{bmatrix} \right\} \\ &= p_\beta + p_u - \frac{\sigma_e^2}{\sigma_u^2} tr \begin{bmatrix} \mathbf{0} & \mathbf{C}^{\beta u} \\ \mathbf{0} & \mathbf{C}^{uu} \end{bmatrix} \\ &= p_\beta + p_u - \frac{\sigma_e^2}{\sigma_u^2} tr\left[\mathbf{C}^{uu}\right].\end{aligned}$$

Note that the prior information about the $\mathbf{u}$ vector results in that the effective number of parameters is smaller than the dimension of $\boldsymbol{\theta}$.

■

8.5 Goodness of Fit and Predictive Ability of a Model

The posterior probability of a model and the Bayes factors can be viewed as global measures of model relative plausibility. However, one often needs to go further than that. For example, a model can be the most plausible within a set of competing models and, yet, either be unable to predict the data at hand well or to give reasonable predictions of future observations. Here we will provide just a sketch of some of the procedures that can be used for gauging the quality of fit and predictive performance of a model.

8.5.1 Analysis of Residuals

A comprehensive account of techniques for examination of residuals is given by Barnett and Lewis (1995). In order to illustrate some of the basic ideas, consider, for example, a linear regression analysis. One of the most widely used techniques for assessing fit is to carry out a residual analysis (e.g., Draper and Smith, 1981). In the context of classical regression, one calculates the predicted value of an observation, $\widehat{y}$, and forms the Studentized fitted residual

$$\frac{y-\widehat{y}}{\sqrt{\widehat{\sigma}_e^2}},$$

where $\widehat{\sigma}_e^2$ is typically the unbiased estimator of the residual variance. If the absolute value of the Studentized residual exceeds a certain critical value of the t or normal distributions, then the observation is viewed as suspicious and regarded as a potential outlier. This may be construed as an indication that the model does not fit well.

The Bayesian counterpart of this classical regression analysis consists of examining the posterior distribution of the unobserved standardized quantity

$$r_i = \frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sqrt{\sigma_e^2}},$$

where the row vector $\mathbf{x}_i'$ contains known explanatory variables linking the unknown regression vector $\boldsymbol{\beta}$ to y_i. Using the standard normality assumptions with independent and identically distributed errors, the distribution of r_i under the sampling model is $r_i \sim N(0,1)$, provided σ_e^2 is known. If $\boldsymbol{\beta}$ has the prior distribution $\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{V}_\beta \sim N(\boldsymbol{\alpha}, \mathbf{V}_\beta)$, where the hyperparameters are also known, one obtains as prior (or predictive) distribution of the residual above, given σ_e^2,

$$r_i|\boldsymbol{\alpha}, \sigma_e^2, \mathbf{V}_\beta \sim N\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\alpha}}{\sqrt{\sigma_e^2}}, \frac{\mathbf{x}_i'\mathbf{V}_\beta\mathbf{x}_i}{\sigma_e^2}\right).$$

The unconditional (with respect to σ_e^2) prior distribution of the standardized residual will depend on the prior adopted for σ_e^2. Then one could carry out an analysis of the residuals prior to proceeding with Bayesian learning about the parameters. More commonly, however, the residual analysis will be undertaken based on the joint posterior distribution of $\boldsymbol{\beta}$ and σ_e^2. As seen in Chapter 6, given σ_e^2, the posterior distribution of $\boldsymbol{\beta}$ is the normal process

$$\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{V}_\beta, \sigma_e^2, \mathbf{y} \sim N\left[\widetilde{\boldsymbol{\beta}}, \left(\frac{\mathbf{X}'\mathbf{X}}{\sigma_e^2} + \mathbf{V}_\beta^{-1}\right)^{-1}\right],$$

where

$$\widetilde{\boldsymbol{\beta}} = \left(\frac{\mathbf{X}'\mathbf{X}}{\sigma_e^2} + \mathbf{V}_\beta^{-1}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{y}}{\sigma_e^2} + \mathbf{V}_{\boldsymbol{\beta}}^{-\mathbf{1}}\boldsymbol{\alpha}\right).$$

Further, given σ_e^2, the posterior distribution of the Studentized residual will have the form

$$r_i|\boldsymbol{\alpha}, \mathbf{V}_\beta, \sigma_e^2, \mathbf{y} \sim N\left[\frac{y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}}{\sqrt{\sigma_e^2}}, \frac{\mathbf{x}_i'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma_e^2} + \mathbf{V}_\beta^{-1}\right)^{-1}\mathbf{x}_i}{\sigma_e^2}\right].$$

The unconditional (with respect to σ_e^2) posterior distribution will depend on the form of the marginal posterior distribution of the residual variance, and its density is obtained as

$$p\left(r_i|\boldsymbol{\alpha}, \mathbf{V}_\beta, \sigma_e^2, \mathbf{y}\right) = \int p\left(r_i|\boldsymbol{\alpha}, \mathbf{V}_\beta, \sigma_e^2, \mathbf{y}\right) p\left(\sigma_e^2|\mathbf{y}\right) d\sigma_e^2,$$

where $p\left(\sigma_e^2|\mathbf{y}\right)$ is the marginal posterior density of the residual variance. Unless standard conjugate priors are adopted, the marginal posterior distribution of the Studentized residual cannot be arrived at in closed form. In such a situation, one can adopt the sampling techniques described in the third part of the book and obtain draws from the posterior distribution of the standardized residual. This is done simply by drawing from the posterior distribution of the model parameters. Then, for observation i, one forms samples

$$r_i^{[j]} = \frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}^{[j]}}{\sqrt{\sigma_e^{2[j]}}}, \quad j = 1, 2, \ldots, m,$$

where $\boldsymbol{\beta}^{[j]}$ and $\sigma_e^{2[j]}$ are samples from the joint posterior distribution of the regression vector and of the residual variance. Thus, one obtains an entire distribution for each Studentized residual, which can be used to decide whether or not the observation is in reasonable agreement with what the model predicts. If the value 0 appears at high density in the posterior distribution, this can be construed as an indication that the observation is in conformity with the model.

This simple idea extends naturally to other models in which residuals are well defined. For example, for binary $(0, 1)$ responses analyzed with a probit model, Albert and Chib (1993, 1995) define the Bayesian residual $r_i = y_i - \Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$, which is real valued on the interval $[y_i - 1, y_i]$. If samples are taken from the posterior distribution of $\boldsymbol{\beta}$, one can form corresponding draws from the posterior distribution of each residual. Since $\Phi\left(\mathbf{x}_i'\boldsymbol{\beta}\right)$ takes values between 0 and 1, an observation $y_i = 0$ will be outlying if the posterior distribution of r_i is concentrated towards the endpoint -1, and an observation $y_i = 1$ is suspect if the posterior of r_i is concentrated towards the value 1. A value of 0 appearing at high density in the posterior distribution of the residuals can be interpreted as an indication of reasonable fit. Albert and Chib (1995) propose an alternative residual defined at the level of a latent variable called the liability (see Chapter 14 for a definition of this concept). The reader is referred to their paper for details.

8.5.2 Predictive Ability and Predictive Cross-Validation

Predictive ability and goodness of fit are distinct features of a model. A certain model may explain and predict adequately the observations used for model building. However, it may yield poor predictions of future observations or of data points that are outside the range represented in the data employed for model building. A number of techniques is available for gauging the predictive ability of a Bayesian model. Even though some attention is paid to foundational issues, the approaches here are often eclectic and explorative. They constitute an important set of tools for understanding the predictive ability of a model.

Cross-validation methods involve constructing the posterior distribution of the parameters but leaving some observations out. Then the predictive distributions of the observations that have been removed are derived to examine whether or not the actual data points fall in regions of reasonably high density. Partition the data as $\mathbf{y}' = \left[y_{\text{out}}, \mathbf{y}'_{-\text{out}}\right]$, where y_{out} is the observation to be removed, and $\mathbf{y}_{-\text{out}}$ is the vector of the remaining observations. The density of the posterior predictive distribution can be written as

$$p\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}, M\right) = \int p\left(y_{\text{out}}|\boldsymbol{\theta}, \mathbf{y}_{-\text{out}}, M\right) p\left(\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}, M\right) d\boldsymbol{\theta}, \tag{8.52}$$

where $p\left(\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}, M\right)$ is the density of the posterior distribution built from $\mathbf{y}_{-\text{out}}$ and model M. In hierarchical modeling, one typically writes the sampling distribution of the data such that conditional independence can be exploited. Thus, given the parameters, y_{out} is independent of $\mathbf{y}_{-\text{out}}$, and one can write (suppressing the notation denoting model M)

$$p\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right) = \int p\left(y_{\text{out}}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}\right) d\boldsymbol{\theta}. \tag{8.53}$$

Since, in general, the form of the posterior density is unknown or analytically intractable, the predictive density will be calculated via Monte Carlo methods (Gelfand et al., 1992; Gelfand, 1996). For example, if m draws from the posterior distribution can be made via MCMC procedures, the form of (8.53) suggests the estimator

$$\widehat{p}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right) = \frac{1}{m}\sum_{j=1}^{m} p\left(y_{\text{out}}|\boldsymbol{\theta}^{[j]}\right),$$

where $\boldsymbol{\theta}^{[j]}$ is a draw from $\left[\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}\right]$. The mean and variance of the predictive distribution can also be computed by Monte Carlo procedures. Since the expected value of the sampling model can almost always be deduced readily, e.g., in regression $E\left(y_{\text{out}}|\boldsymbol{\theta}\right) = \mathbf{x}'_{\text{out}}\boldsymbol{\beta}$, the mean of the predictive distribution can be estimated as

$$\widehat{E}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right) = \frac{1}{m}\sum_{j=1}^{m} E\left(y_{\text{out}}|\boldsymbol{\theta}^{[j]}\right). \tag{8.54}$$

Similarly, a Monte Carlo estimate of the variance of the predictive distribution can be obtained as

$$\widehat{Var}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right)=\widehat{E}_{[\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}]}\left[Var\left(y_{\text{out}}|\boldsymbol{\theta}\right)\right]+\widehat{Var}\left[E\left(y_{\text{out}}|\boldsymbol{\theta}\right)\right]. \tag{8.55}$$

This can be illustrated with a regression model, although in this situation there is an analytical solution under the standard assumptions. For example, if the regression model postulates $y_{\text{out}}|\boldsymbol{\beta},\sigma_e^2\sim N\left(x'_{out}\boldsymbol{\beta},\sigma_e^2\right)$, then

$$\widehat{E}_{[\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}]}\left[Var\left(y_{\text{out}}|\boldsymbol{\theta}\right)\right]=\frac{1}{m}\sum_{j=1}^{m}\sigma_e^{2[j]},$$

and

$$\begin{aligned}&\widehat{Var}_{[\boldsymbol{\theta}|\mathbf{y}_{-\text{out}}]}\left[E\left(y_{\text{out}}|\boldsymbol{\theta}\right)\right]=\widehat{Var}\left[x'_{out}\boldsymbol{\beta}\right]\\&=\frac{1}{m}\sum_{j=1}^{m}\left(x'_{out}\boldsymbol{\beta}^{[j]}\right)^2-\left(\frac{1}{m}\sum_{j=1}^{m}x'_{out}\boldsymbol{\beta}^{[j]}\right)^2.\end{aligned}$$

Subsequently, the following composite statistic can be used to evaluate the overall predictive ability of the model (Congdon, 2001):

$$D^2=\sum_{out=1}^{n}\left[\frac{y_{\text{out}}-\widehat{E}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right)}{\sqrt{\widehat{Var}\left(y_{\text{out}}|\mathbf{y}_{-\text{out}}\right)}}\right]^2. \tag{8.56}$$

Models having a smaller value of D^2 would be viewed as having a better predictive ability. Clearly, if n is very large, the computations may be taxing, since n posterior and predictive distributions need to be computed. Other statistics are described in Gelfand et al. (1992) and in Gelfand (1996).

A related idea has been advocated by Gelman et al. (1996). Rather than working with the leave-one-out method in (8.53), they propose generating data $\widetilde{\mathbf{y}}$ from the posterior predictive distribution with density

$$p\left(\widetilde{\mathbf{y}}|\mathbf{y},M\right)=\int p\left(\widetilde{\mathbf{y}}|\boldsymbol{\theta},M\right)p\left(\boldsymbol{\theta}|\mathbf{y},M\right)d\boldsymbol{\theta}. \tag{8.57}$$

One then wishes to study whether the simulated value $\widetilde{\mathbf{y}}$ agrees with the observed data $\mathbf{y}$. Systematic differences between the simulations and the observed data indicate potential failure of model M. Various criteria or test quantities can be used to carry out the comparisons. Examples of these are given in Gelfand (1996). The choice of test quantities should be driven by the aspect of the model whose fit is in question and/or by the purpose with which the model will be used. The method of composition (introduced in Chapter 1), can be used to obtain draws from (8.57), and can be described as follows:

1. Draw $\boldsymbol{\theta}$ from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, M)$. Ways of achieving this are discussed later in this book.

2. Draw $\widetilde{\mathbf{y}}$ from the sampling distribution $p(\widetilde{\mathbf{y}}|\boldsymbol{\theta},M)$. One has now a single realization from the joint distribution $p(\widetilde{\mathbf{y}}, \boldsymbol{\theta}|M)$.

3. Repeat steps 1 and 2 many times.

The set of $\widetilde{\mathbf{y}}'s$ drawn using this algorithm constitutes samples from (8.57). Letting $h(\mathbf{y})$ be a particular test quantity, for example, the average of the top 10 observations, one can then study whether $h(\mathbf{y})$ falls in a region of high posterior probability in the distribution $[h(\widetilde{\mathbf{y}})|\mathbf{y},M]$. This can be repeated for all the models under investigation. Gelman et al. (1996) propose the calculation of Bayesian p-values, p_B, for given test quantities $h(\mathbf{y}, \boldsymbol{\theta})$. The notation emphasizes that, in contrast with classical p-values, the test quantity can depend on both data and parameters. Then,

$$\begin{aligned} p_B &= p[h(\widetilde{\mathbf{y}}, \boldsymbol{\theta}) > h(\mathbf{y}, \boldsymbol{\theta})|\mathbf{y}] \\ &= \int\int I[h(\widetilde{\mathbf{y}}, \boldsymbol{\theta}) > h(\mathbf{y}, \boldsymbol{\theta})]\, p(\widetilde{\mathbf{y}}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} d\widetilde{\mathbf{y}}, \end{aligned} \tag{8.58}$$

gives the probability that the simulated data $\widetilde{\mathbf{y}}$ is more extreme than the observed data $\mathbf{y}$, averaged over the distribution $[\boldsymbol{\theta}|\mathbf{y}]$. A possible test quantity could be

$$h(\mathbf{y}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{[y_i - E(Y_i|\boldsymbol{\theta})]^2}{Var(Y_i|\boldsymbol{\theta})}$$

and

$$h(\widetilde{\mathbf{y}}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\left[\widetilde{y}_i - E\left(\widetilde{Y}_i|\boldsymbol{\theta}\right)\right]^2}{Var\left(\widetilde{Y}_i|\boldsymbol{\theta}\right)}.$$

These are then used for computing (8.58). A cross-validation approach can also be implemented using this idea. An application of these techniques in animal breeding is in Sorensen et al. (2000).

Another way of assessing global predictive ability of a set of models was proposed by Geisser and Eddy (1979) and by Geisser (1993) via the conditional predictive ordinate (CPO). The logarithm of the CPO for Model i is

$$\log[CPO_{\text{Model } i}] = \sum_{out=1}^{n} \log[p(y_{\text{out}}|\mathbf{y}_{-\text{out}}, \text{Model } i)].$$

Gelfand and Dey (1994) describe techniques for calculating the CPO that avoid carrying out the n implementations of the sampling procedure described above. Chapter 12, especially Section 12.4, discusses Monte Carlo implementation of these quantities in more detail.

8.6 Bayesian Model Averaging

8.6.1 General

Consider a survival analysis of sheep or of dairy cows. The information available may consist of covariates such as herd or flock, sire, year-season of birth, molecular markers, and last known survival status, since censoring is pervasive. The objective of the analysis may be to assess the effects of explanatory variables, or to predict the survival time of the future progeny of some of the sires. Hence, one searches for some reasonable survival model (e.g., Gross and Clark, 1975; Collet, 1994) and finds that a proportional hazards model M_1 fits well and that it gives sensible parameter estimates. Then one proceeds to make predictions. However, another proportional hazards model M_2 also fits well, but it gives different estimates and predictions. Which model should be used at the end?

Now imagine a standard regression analysis in which 15 predictor variables are available, and suppose that some "best" model must be sought. Even if second-order and cross-product terms are ignored, there would be 2^{15} different models. For example, suppose that the variables are Y, X_1, X_2. Then, using the standard notation, there are the following four possible models

$$\begin{aligned}
\text{model 1} &: Y = \beta_0 + e, \\
\text{model 2} &: Y = \beta_0 + \beta_1 X_1 + e, \\
\text{model 3} &: Y = \beta_0 + \beta_2 X_2 + e, \\
\text{model 4} &: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e.
\end{aligned}$$

These models may differ little in relative plausibility. Again, which model ought to be used for predictions?

A third example is that of choosing between genetic models to infer parameters, and to predict the genetic merit of future progeny. One specification may be the classical infinitesimal model. A second specification may be a model with a finite number of loci. If so, how many? A third model may pose polygenic variation, plus the effects of some marked QTL.

The preceding three examples illustrate that the problem of model choice is pervasive. Typically, models are chosen in some ad-hoc manner, and inferences are based on the model eventually chosen, as if there were no uncertainty about it. In Bayesian analysis, however, it is possible to view the model as an item subject to uncertainty. Then the "model random variable" is treated as a nuisance, and the posterior distribution of the "model random variable" is used to obtain inferences that automatically take into account the relative plausibility of the models under consideration. This is called Bayesian model averaging, or BMA for short. We will outline the basic ideas, and refer the reader to Madigan and Raftery (1994), Raftery et al. (1997), and Hoeting et al. (1999) for additional details. These authors

argue as follows: since part of the evidence must be used in the process of model selection, ignoring the uncertainty about the model leads to an overstatement of precision in the analysis. In turn, this can lead to declaring "false positives", and the analysis lacks robustness unless, by chance, one stumbles into the "right" model. It will be shown at the end of this section that BMA can be used to enhance the predictive ability of an analysis.

8.6.2 Definitions

Let

$$\begin{aligned} \Delta &= \text{parameter or future data point,} \\ \mathbf{y} &= \text{data,} \\ M &= \{M_1, M_2, \ldots, M_K\} \text{ set of models,} \\ p(M_i) &= \text{prior probability of model } i, \\ p(M_i|\mathbf{y}) &= \text{posterior probability of model } i. \end{aligned}$$

The "usual" Bayesian approach gives, as posterior distribution (or density) of Δ,

$$p(\Delta|\mathbf{y}, M_i) = \frac{p(\mathbf{y}|\Delta, M_i)\, p(\Delta|M_i)}{p(\mathbf{y}|M_i)},$$

and the notation indicates clearly that inferences are conditional on M_i, as if the model were known to be true for sure. In BMA, on the other hand, the idea is to average out over the posterior distribution of the models, leading to

$$\begin{aligned} p(\Delta|\mathbf{y}) &= p(\Delta \text{ and } M_1|\mathbf{y}) + \cdots + p(\Delta \text{ and } M_K|\mathbf{y}) \\ &= \sum_{i=1}^{K} p(\Delta \text{ and } M_i|\mathbf{y}) = \sum_{i=1}^{K} p(\Delta|\mathbf{y}, M_i)\, p(M_i|\mathbf{y}). \end{aligned} \tag{8.59}$$

The preceding expression reveals that, in BMA, the model is treated as a nuisance parameter. Hence, the nuisance is eliminated in the usual manner, by integration or by summing. Then the inferences about a parameter can be viewed as a weighted average of the inferences that would be drawn if each of the models were true, using the posterior probability of the model as a mixing distribution.

In BMA, the posterior expectation and variance are calculated in the usual manner. For example, let the posterior mean of Δ under model k be

$$E(\Delta|M_k, \mathbf{y}) = \int \Delta\; p(\Delta|M_k, \mathbf{y})\, d\Delta = \widehat{\Delta}_k$$

Then, unconditionally with respect to the model, one obtains

$$E(\Delta|\mathbf{y}) = E_{M|\mathbf{y}}\left[E(\Delta|M_k, \mathbf{y})\right] = \sum_{k=1}^{K} \widehat{\Delta}_k\, p(M_k|\mathbf{y}). \tag{8.60}$$

Similarly, one can use the variance decomposition

$$Var\left(\Delta|\mathbf{y}\right) = E_{M|\mathbf{y}}\left[Var\left(\Delta|M,\mathbf{y}\right)\right] + Var_{M|\mathbf{y}}\left[E\left(\Delta|M,\mathbf{y}\right)\right],$$

leading to

$$Var\left(\Delta|\mathbf{y}\right) = \sum_{k=1}^{K} Var\left(\Delta|M_k,\mathbf{y}\right) p\left(M_k|\mathbf{y}\right) + \sum_{k=1}^{K} \left(\widehat{\Delta}_k\right)^2 p\left(M_k|\mathbf{y}\right) - \left[\sum_{k=1}^{K} \widehat{\Delta}_k\, p\left(M_k|\mathbf{y}\right)\right]^2. \tag{8.61}$$

The idea is straightforward, and it makes eminent sense, at least from a Bayesian perspective. The difficulty resides in that there can be many models, as in a regression equation, where there may be at least 2^p (for p being the number of covariates) models, and even more when interactions are included. Hoeting et al. (1999) discusses some of the methods that have been used for reducing the number of terms to be included in the sums appearing in (8.60) and (8.61).

8.6.3 Predictive Ability of BMA

Suppose one partitions the data as

$$\mathbf{y} = \left[\mathbf{y}'_{\text{Build}}, \mathbf{y}'_{\text{Pred}}\right]',$$

where $\mathbf{y}_{\text{Build}}$ is the data used for model building, and $\mathbf{y}_{\text{Pred}}$ includes the data points to be predicted, as in predictive cross-validation. Good (1952) introduced the predictive logscore (PLS) which, for Model k, is

$$PLS_k = -\sum_{y\in\mathbf{y}_{\text{Pred}}} \log p\left(y|M_k,\mathbf{y}_{\text{Build}}\right)$$

$$= -\sum_{y\in\mathbf{y}_{\text{Pred}}} \log \int p\left(y|\boldsymbol{\theta}_k, M_k, \mathbf{y}_{\text{Build}}\right) p\left(\boldsymbol{\theta}_k|M_k,\mathbf{y}_{\text{Build}}\right) d\boldsymbol{\theta}_k, \tag{8.62}$$

where $\boldsymbol{\theta}_k$ is the parameter vector under Model k. It it desirable to have a model with as small a PLS as possible. Under BMA

$$PLS_{BMA} = -\sum_{y\in\mathbf{y}_{\text{Pred}}} \log\left[\sum_{k=1}^{K} p\left(y|M_k,\mathbf{y}_{\text{Build}}\right) p\left(M_k|\mathbf{y}_{\text{Build}}\right)\right]. \tag{8.63}$$

Suppose that the model and the data to be predicted are unknown, which is the usual situation. Now consider the difference

$$PLS_{BMA} - PLS_k = -\sum_{y\in\mathbf{y}_{\text{Pred}}} \log \frac{\sum_{k=1}^{K} p\left(y|M_k,\mathbf{y}_{\text{Build}}\right) p\left(M_k|\mathbf{y}_{\text{Build}}\right)}{p\left(y|M_k,\mathbf{y}_{\text{Build}}\right)}.$$

Next, take expectations of this difference with respect to the predictive distribution under BMA (that is, averaging over all possible models). This distribution has density

$$p\left(\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}\right) = \sum_{k=1}^{K} p\left(\mathbf{y}_{\text{Pred}}|M_k, \mathbf{y}_{\text{Build}}\right) p\left(M_k|\mathbf{y}_{\text{Build}}\right)].$$

Thus,

$$\begin{aligned} &E_{\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}}\left(PLS_{BMA} - PLS_k\right) \\ &= -\sum_{y \in \mathbf{y}_{\text{Pred}}} E\left[\log \frac{\sum_{k=1}^{K} p\left(y|M_k, \mathbf{y}_{\text{Build}}\right) p\left(M_k|\mathbf{y}_{\text{Build}}\right)}{p\left(y|M_k, \mathbf{y}_{\text{Build}}\right)}\right]. \end{aligned}$$

The expected value in the right hand side, taken over the distribution $[\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}]$, is the Kullback–Leibler discrepancy between the predictive distributions of datum y under BMA and under Model k. Since the discrepancy is at least 0, it follows that the right-hand side is at most null. Hence

$$E_{\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}}\left(PLS_{BMA}\right) \leq E_{\mathbf{y}_{\text{Pred}}|\mathbf{y}_{\text{Build}}}\left(PLS_k\right),$$

as in Madigan and Raftery (1994). This implies that under model uncertainty, the predictive performance of BMA (at least in the PLS sense) is expected to be better than that obtained under a single model, even if the latter is the most probable one. Raftery et al. (1997) and Hoeting et al. (1999) present several study cases supporting this theoretical result.

Typically, BMA leads to posterior distributions that are more spread than those under a single model. This illustrates that inferences based on a single model may give an unrealistic statement of precision; this may lead to false positive results.

The reader is now equipped with the foundations on which Bayesian inference rests. As stated before and especially for complex models, it is seldom the case that exact methods of inference can be used. Fortunately, methods for sampling from posterior distributions are available, and these are discussed in Part III of this book.