8.6.6 Alternatives to DIC

For models with discrete parameters, such as finite mixture models, the plug-in deviance required by p_D cannot be calculated since the posterior mean of a discrete parameter is either undefined or not guaranteed to be part of the discrete parameter space. Celeux et al. (2006), and their discussants, investigated ways of defining p_D and DIC for mixture and related models, though found them all problematic. For example, a mixture model could be reformulated by integrating over the discrete component membership parameter, as in Example 11.6.1 — however, the resulting DIC may be sensitive to the constraint chosen to identify the components (§11.6).

Plummer (2008) proposed the *penalised expected deviance* as an alternative model comparison measure. Both this and DIC estimate the ability to predict a replicate dataset, but judge this ability by different loss functions. Whereas DIC estimates the deviance for a replicate dataset *evaluated at the posterior expectations* of parameters θ, Plummer's criterion estimates the *expected* deviance for a replicate dataset. Both criteria incorporate a penalty to adjust for the underestimate in the loss ("optimism") due to using the data twice to both fit and evaluate the model. Since it does not require a "plug-in" estimate such as the posterior mean, this criterion can be used with discrete parameters. It is calculated as $\overline{D}+p_{opt}$, where the optimism p_{opt} is estimated from two parallel MCMC chains using importance sampling, as described by Plummer (2008) and provided in JAGS. The importance sampling method is unstable when there are highly influential observations, otherwise $p_{opt} \approx 2p^*_D$. A similar approximation was derived in the context of variable selection by van der Linde (2005).

The *pseudo-marginal likelihood* was proposed by Geisser and Eddy (1979) as a cross-validatory measure of predictive ability,

$$\prod_i p(y_i|y_{\backslash i}) = \prod_i \int p(y_i|\theta)p(\theta|y_{\backslash i})d\theta,$$

where $y_{\backslash i}$ is all observations excluding y_i. Gelfand and Dey (1994) described an importance sampling method for estimating it based on a single MCMC run, which avoids the need to refit the model with each observation excluded in turn. The full-data posterior density $p(\theta|y)$ is used as a proposal distribution to approximate the leave-one-out posterior $p(\theta|y_{\backslash i})$. Given an MCMC sample $\theta^{(1)},\ldots,\theta^{(T)}$ from the posterior of θ, the importance weights are then $w_{it} = p(\theta^{(t)}|y_{\backslash i})/p(\theta^{(t)}|y) \propto 1/p(y_i|\theta^{(t)})$, and the estimate of $p(y_i|y_{\backslash i})$ is the *harmonic mean* of $p(y_i|\theta^{(t)})$ over the posterior sample:

$$\begin{aligned} p(y_i|y_{\backslash i}) &\approx \sum_t w_{it}p(y_i|y_{\backslash i},\theta^{(t)})/\sum_t w_{it} \\ &= T/\sum_t(1/p(y_i|\theta^{(t)})) \end{aligned}$$

Thus, the quantity $1/p(y_i|\theta^{(t)})$ is monitored during MCMC sampling, and the estimate of $p(y_i|y_{\backslash i})$ is the reciprocal of its posterior mean. The individual $p(y_i|y_{\backslash i})$ are called *conditional predictive ordinates* (CPOs) and may also be used as outlier diagnostics. Again, this method may be unstable, particularly if some of the CPOs are large (common in hierarchical models) and may require a large MCMC sample for a precise estimate. However, unlike DIC, it does not depend on plug-in estimates or on the model parameterisation.

8.7 Bayes factors

Traditional Bayesian comparison of models M_0 and M_1 is based on hypothesis tests using the *Bayes factor*. The posterior odds of model M_0 compared to M_1 are given by

$$\frac{p(M_0|y)}{p(M_1|y)} = \frac{p(M_0)}{p(M_1)}\frac{p(y|M_0)}{p(y|M_1)}$$

where

$$\frac{p(y|M_0)}{p(y|M_1)} = \frac{\int p(y|\theta_0)p(\theta_0)d\theta_0}{\int p(y|\theta_1)p(\theta_1)d\theta_0} = B_{01}$$

is known as the Bayes factor for M_0 compared to M_1. In other words,

$$\text{posterior odds of } M_0 = \text{Bayes factor} \times \text{prior odds of } M_0.$$

The Bayes factor B_{01} quantifies the weight of evidence in favour of the null hypothesis H_0: "M_0 is true." If both models (hypotheses) are equally likely a priori, then their relative prior odds is 1 and B_{01} is the posterior odds in favour of model M_0 (Jeffreys (1939), p. 275, Gelman et al. (2004), p. 185).

The Bayes factors are in some sense similar to a likelihood ratio, except that the likelihood is *integrated* instead of maximised over the parameter space. As with AIC, there is no need for models to be nested, although unlike AIC, the objective is the identification of the 'true' model (Bernardo and Smith, 1994). Jeffreys (1939) provided a table relating the size of the Bayes factor to the "strength of evidence."

$p(y|M_r)$ is the *marginal likelihood* or *prior predictive probability* of the data, and it is important to note that this will depend crucially on the form of the prior distribution. A simple example will show that Bayes factors require informative prior distributions under each model. Consider a scalar θ so that the relevant term for the Bayes factor is $p(y) = \int p(y|\theta)p(\theta)d\theta$. Suppose θ is given a uniform prior, so that $p(\theta) = 1/(2c)$; $\theta \in [-c, c]$. Then $p(y) = \frac{1}{2c}\int_{-c}^{c} p(y|\theta)d\theta \propto \frac{1}{c}$ for large c. Therefore $p(y)$ can be made arbitrarily small by increasing c.

TABLE 8.4
Calibration of Bayes factors provided by Jeffreys.

Bayes factor range	Strength of evidence in favour of H_0 and against H_1
> 100	Decisive
32 to 100	Very strong
10 to 32	Strong
3.2 to 10	Substantial
1 to 3.2	"Not worth more than a bare mention"
	Strength of evidence against H_0 and in favour of H_1
1 to 1/3.2	"Not worth more than a bare mention"
1/3.2 to 1/10	Substantial
1/10 to 1/32	Strong
1/32 to 1/100	Very strong
$< 1/100$	Decisive

Suppose we are comparing models with weak prior information. Schwarz's Bayesian Information Criterion (BIC) is:

$$\text{BIC} = -2\log p(y|\hat{\theta}) + p\log n,$$

where $\hat{\theta}$ is the maximum likelihood estimate. The difference $\text{BIC}_0 - \text{BIC}_1$ gives an approximation to $-2\log B_{01}$. Kass and Wasserman (1995) show that this approximation has error $O_p(n^{-1/2})$ under a prior distribution which carries information equivalent to a single observation — the *unit-information* prior.

Alternatively, the posterior probability of model r, among a set of models indexed by k, is approximated by

$$p(M_r|y) = \exp(-0.5\text{BIC}_r)/\sum_k \exp(-0.5\text{BIC}_k) \qquad (8.7)$$

Example 8.7.1. *Paul the psychic octopus*
In the 2010 football World Cup competition, Paul "the psychic octopus" made 8 predictions of the winners of football matches and got all $y = 8$ right. Our analysis will ignore the possibilities of draws, assume there was no bias or manipulation in the experiment, and ignore selection effects arising from Paul only becoming famous due to the first correct predictions, in the face of competition from numerous other wildlife. We assume a binomial model with probability θ of a correct prediction.

Rather naively, we could set up two simple hypotheses: H_0 representing that the predictions are just chance, so that $\theta = 0.5$; H_1 representing Paul having 100% predictive ability, so that $\theta = 1$. Since these are simple hypotheses with no unknown parameters, the Bayes factor is just the likelihood ratio $p(y|H_0)/p(y|H_1) = 1/2^8 = 1/256$, which from Table 8.4 represents "decisive" evidence against H_0 by Jeffreys criteria. However, the posterior odds against Paul being psychic also depend on the prior odds $p(H_0)/p(H_1)$ of Paul not having any psychic abilities (or knowledge of football), which it is reasonable to assume are so huge that this likelihood ratio makes little impact!

It may be more sensible to compare H_0 with an alternative hypothesis H_1 that Paul has *some* psychic ability, represented by a prior distribution on $\theta|H_1$. Naively this would be uniform on 0.5 to 1, but we introduce some scepticism by restricting it to be less than 0.55. So we are both sceptical of any effect existing at all, and even if it did exist, sceptical of a large effect. The code then essentially follows that of the biased coin example in §5.4.

```
q[1] <- 0.5; q[2] <- 0.5                # prior assumptions
r <- 8; n <- 8                          # data
r          ~ dbin(theta[pick], n)       # likelihood
pick       ~ dcat(q[])
theta[1] <- 0.5                         # if random (assumption 1)
theta[2]   ~ dunif(0.5, 0.55)           # if psychic
psychic  <- pick - 1                    # 1 if psychic, 0 otherwise
```

```
node    mean   sd     MC error 2.5% median 97.5% start sample
psychic 0.6012 0.4896 0.001601 0.0  1.0    1.0   1     100000
```

The posterior probability of psychic abilities is now 0.6, corresponding to posterior odds $p(H_0|y)/p(H_1|y) = 0.4/0.6 = 0.66$. If the prior odds are taken as 1, this means that the Bayes factor is 0.66 in favour of psychic abilities, but again the prior odds against psychic abilities should realistically be much larger.

8.7.1 Lindley–Bartlett paradox in model selection

We have already seen that the Bayes factor depends crucially on the prior distribution within competing models. We now use a simple example to show how this can lead to an apparent conflict between using tail areas to criticise assumptions, and using Bayes factors — a conflict that has become known as the *Lindley–Bartlett paradox*.

Suppose we assume $Y_i \sim \text{N}(\theta, 1)$; we want to test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. Then the sufficient statistic is $\overline{Y}$ with distribution $\overline{Y} \sim \text{N}(\theta, 1/n)$. For H_0, $p(\overline{y}|H_0) = \sqrt{\frac{n}{2\pi}}\exp[-n\overline{y}^2/2]$. For H_1, assume $p(\theta) = 1/(2c)$; $\theta \in [-c, c]$, $\theta \neq 0$, then

$$p(\overline{y}|H_1) = \frac{1}{2c}\int_{-c}^{c}\sqrt{\frac{n}{2\pi}}\exp[-n(\overline{y}-\theta)^2/2]\,d\theta \approx \frac{1}{2c}.$$

Hence the Bayes factor is

$$B_{01} = \frac{p(\overline{y}|H_0)}{p(\overline{y}|H_1)} = \sqrt{\frac{n}{2\pi}}\exp[-n\overline{y}^2/2] \times 2c.$$

From a classical hypothesis-testing perspective, we would declare a "significant" result if $\overline{y} > 1.96/\sqrt{n}$. At this critical value, the Bayes factor is $\sqrt{\frac{n}{2\pi}}\exp[-1.96^2/2] \times 2c$. Hence

- For fixed n, we can give H_0 very strong support by increasing c
- For fixed c, we can give H_0 very strong support by increasing n

So data that would just *reject* H_0 using a classical test will tend to *favour* H_0 for (a) diffuse priors under H_1 and (b) large sample sizes.

8.7.2 Computing marginal likelihoods

Computing Bayes factors for a generic model is challenging outside simple conjugate situations, as reviewed by Han and Carlin (2001) and Ntzoufras (2009). There is no easy method which works for all models specified in BUGS. BIC gives an approximation, as described above, though this essentially implies a default "unit information" prior for the parameters and does not allow user-specified priors. Other methods are based either on

- directly computing the marginal likelihood for each model, or
- considering the model choice as a discrete parameter and jointly sampling from the model and parameter space (§8.8.2).

For computing the marginal likelihood $p(y)$ for a particular model M, *harmonic mean* and related estimators are also sometimes used:

$$p(y) \approx \left(\frac{1}{T}\sum_{t=1}^{T}\left\{\frac{g(\theta^{(t)})}{p(y|\theta^{(t)})p(\theta^{(t)})}\right\}\right)^{-1}$$

where $g()$ is an importance sampling density chosen to approximate the posterior. Although this is temptingly easy to program in BUGS by monitoring the term inside the braces and taking the reciprocal of its posterior mean, it is impractical in all but the simplest of models, since $p(y|\theta^{(t)})$ will frequently be very small; thus unfeasibly long runs would be required to stably estimate the posterior mean — indeed it may never converge (Neal, 2008). *Bridge sampling* or *path sampling* estimators (Gelman and Meng, 1998) are more effective, though usually require problem-specific tuning. Similarly, methods by Chib (1995) and Chib and Jeliazkov (2001) were shown to be effective by Han and Carlin (2001), but to implement these in BUGS would need substantial problem-specific programming, including access to the underlying source code (Ntzoufras, 2009).

Jointly sampling from the model and parameter space is a generally more reliable method of obtaining posterior probabilities of models in a BUGS context, particularly for comparing models with different sets of predictor variables, and techniques to do this in BUGS are reviewed in §8.8.2.

8.8 Model uncertainty

Neglecting uncertainty about the choice of model has been called a "quiet scandal" in statistical practice (Breiman, 1992) – see, for example, Draper (1995) and Burnham and Anderson (2002) for discussions. Drawing conclusions on the basis of a single selected model can conceal the possibility that other plausible models would give different results. Sensitivity analysis is recommended as a minimum, and this section discusses methods to formally incorporate model uncertainty in conclusions.

8.8.1 Bayesian model averaging

Posterior model probabilities $p(M_r|y)$ can be used to do "model averaging" to obtain predictions which account for model uncertainty. If we need to predict $\tilde{Y}$, and the predictive distribution assuming M_r is $p(\tilde{y}|y, M_r)$, then the "model-averaged" prediction is

$$p(\tilde{y}|y) = \sum_i p(\tilde{y}|y, M_r)p(M_r|y)$$

where

$$p(M_r|y) = p(M_r)p(y|M_r)/\sum_k\{p(M_k)p(y|M_k)\}$$

However, as discussed in §8.7.2, the marginal likelihood $p(y|M_r)$ involved in this definition is not, in general, straightforward to calculate in BUGS. We now describe techniques to accomplish model-averaged predictions without needing to calculate marginal likelihoods.

8.8.2 MCMC sampling over a space of models

We could consider the model choice as an additional parameter: specify prior probabilities for the model choice m and model-specific parameters θ_m, and sample from their joint posterior distribution $p(m, \theta_m|y)$, thus computing the posterior model probabilities. Any predictions are automatically averaged over the competing models.

Reversible jump MCMC However, if we are choosing between models with different numbers of parameters, then the *dimension* of the space changes as the currently chosen model changes. The *reversible jump* MCMC algorithm was devised by Green (1995) to allow sampling over a space of varying dimension. The Jump add-on to WinBUGS (Lunn et al., 2009c)[†] performs reversible

[†] This is under development for OpenBUGS.

jump for variable selection in linear and binary regression and selecting among polynomial splines with different numbers of knots. See the manual included with Jump for further details and worked examples. It could be extended in the future to select within other classes of models, such as mixture models, for which specialised programming is currently required to implement reversible jump MCMC.

Product space search In reversible jump MCMC, a value for θ_m is only sampled if the sampler is currently visiting model m. Carlin and Chib (1995) described an alternative MCMC method for sampling from $p(m, \theta_m|y)$, where values of θ_m are sampled for all m, whatever the currently chosen model. This requires a *pseudoprior* to be specified for each θ_m conditionally on the model *not* being m. While this is less efficient than reversible jump, it enables standard MCMC algorithms, available in BUGS, to be used. It can suffer from poor mixing unless the pseudopriors and priors on the models are chosen carefully. In practice, each model can be fitted independently and the resulting posteriors used to choose pseudopriors for a joint model. See the `Pines` example for BUGS (available from the BUGS web site or distributed with OpenBUGS) or Carlin and Louis (2008) for further details.

Variable selection priors There are several methods specifically for variable selection in regression models, including stochastic search variable selection (George and McCulloch, 1993) and Gibbs variable selection (Dellaportas et al., 2002). The general idea is that there is a vector of covariate effects β and a vector I of the same length containing 0/1 indicators for each covariate being included in the model. β is then given a "spike and slab" prior (Mitchell and Beauchamp, 1988). This is a mixture of a probability mass $p(\beta_j|I_j = 0)$ concentrated around zero, representing exclusion from the model, and a relatively flat prior $p(\beta_j|I_j = 1)$ given that the variable is included:

$$p(\beta_j) = p(I_j = 1)p(\beta_j|I_j = 1) + p(I_j = 0)p(\beta_j|I_j = 0).$$

In BUGS, an example is

```
beta[j] <- b[pick[j]]       # effect of jth covariate
b[1]     ~ dnorm(0, tau)    # "spike": tau is large, b[1] <- 0
b[2]     ~ dnorm(0, eps)    # "slab": precision eps is small
pick[j] <- I[j] + 1
I[j]     ~ dbern(p[j])
```

where `p[j]` is the prior probability that the jth covariate is included, assuming these probabilities are independent. Thus the posterior probabilities of including each covariate arise naturally as the posterior mean of each `I[j]`. The methods differ in how exactly they define the priors. For more details on these methods and their implementation in BUGS, see O'Hara and Sillanpää (2009) and Ntzoufras (2009) — while programming is generally straightforward, their efficiency and accuracy can depend on the choice of prior and parameterisation.

8.8.3 Model averaging when all models are wrong

Bayesian model averaging involves choosing and computing prior and posterior probabilities on models, interpreted as beliefs in their truth. Bernardo and Smith (1994) showed decision-theoretically that this provides optimal prediction or estimation under an "$\mathcal{M}$-closed" situation — in which the true process which generated the data is among the list of candidate models.

However, often one does not believe any of the models are true — an "$\mathcal{M}$-open" situation. Typically the truth is thought to be more complex than any model being considered. Model averaging is more difficult in this case, though some suggestions have been made. For example, substituting AIC or DIC for BIC in Equation (8.7) gives "Akaike weights" (Akaike, 1979) or DIC weights for averaging models, which measure their predictive ability, rather than their probability of being true. Using DIC in this way is attractively simple, though this method has not been formally assessed or justified. The resulting probabilities are difficult to interpret, though Burnham and Anderson (2002) suggest they represent posterior model probabilities under an implicit prior which favours more complex models at larger sample sizes.

Bootstrapping DIC A more interpretable way of averaging models without invoking a "true model" is to *bootstrap* the model selection process. Assuming independent data points, we resample from the data, choose the best-fitting model according to some criterion, repeat the process a large number of times, and average the resulting predictions. Buckland et al. (1997) used this method with AIC in a frequentist context, and Jackson et al. (2010a) used it with DIC for averaging Bayesian models. The resulting model probabilities $p(M_r|y)$ are the proportion of samples to which model r fits best according to the criterion. These are not Bayesian posterior probabilities, but rather *frequentist* probabilities, under sampling uncertainty, that the model will give the best predictions among those being compared.

Resampling and refitting would often be impractical for Bayesian models fitted by MCMC, which are typically intensive to compute. Therefore Jackson et al. (2010a), following Vehtari and Lampinen (2002), adapted a "Bayesian bootstrap" method which only requires one model fit and no resampling. Instead of sampling with replacement from the data vector y, the Bayesian bootstrap samples *sets of probabilities* q_i that the underlying random variable Y takes the value of each sample point $y_1, \ldots, y_n$. In one bootstrap iteration, samples $q_i^{(rep)}$ of q_i are drawn from a "flat" Dirichlet(1,...,1) distribution. This is the posterior distribution of *the sampling distribution of* Y, which is assumed to be a discrete distribution over the observed values. This posterior is obtained by combining the sample $y_1, \ldots, y_n$ with an improper prior (Rubin, 1981).

The bootstrap replicate of a sample statistic (e.g., the mean) that can be expressed as $\sum_i f(y_i)$ is the weighted sum $n \sum_i q_i^{(rep)} f(y_i)$. Since the DIC can be decomposed into a sum over observations i, $DIC(y|M_r) = \sum_{i=1}^n DIC_i$, where $DIC_i = 2\overline{D(y_i|\theta)} - D(y_i|\hat{\theta})$, the bootstrap replicate of the DIC is

$$DIC(y|M_r)^{(rep)} = n \sum_{i=1}^{n} q_i^{(rep)} DIC_i$$

The sample of replicate DICs for each competing model can be used to give a bootstrap "confidence interval" surrounding the DIC for each model and probabilities that each model is best among the candidates.

Implementing this in BUGS requires the contribution to the posterior deviance from each data point to be monitored explicitly, similar to the method of deviance residuals (§8.3.4). For example, in a normal model:

```
for (i in 1:n) {
  y[i]    ~ dnorm(mu[i], tau)
  dev[i] <- log(2*pi) - log(tau) + pow(y[i] - mu[i], 2)*tau
  ...
}
```

The deviance of the observation y[i] evaluated at the posterior means of mu[i] and tau is subtracted from the posterior mean of dev[i] to produce DIC_i. The replicates can then be computed outside BUGS, using random samples of Dirichlet(1,. . .,1) variables (created, e.g., by BUGS). Note that the resulting model-averaged posterior has no Bayesian interpretation, since two sampling models for the data are used simultaneously — it is best viewed as a computational approximation to resampling.

8.8.4 Model expansion

Instead of averaging over a *discrete* set of models, a more flexible framework for model uncertainty is to work within a single model that encompasses all reasonable possibilities. This is recommended, for example, by Gelman et al. (2004). Model uncertainty is then considered as a choice over a *continuous* space of models. Support for different model choices is assessed by examining posterior distributions of parameters in the larger model, as in § 8.5, and the model is checked to ensure that it gives predictions which agree with observations, as in §8.4.

The class of Bayesian nonparametric models illustrated in § 11.8, for example, can reasonably be thought to "include the truth" in most practical situations. However, these do not naturally represent many model choice situations — a common example is whether to include or exclude a covariate in a regression. The encompassing flexible model would then be the one which includes all covariates being considered. In §8.8.2, we described flexible models of this type, where the prior distributions for the covariate effects had "spikes" at zero representing the possibility that the covariate is not included.

Smooth distributions are often a more realistic expression of prior belief than the mixture priors of this kind implied by discrete model averaging. Giving privilege to an effect of zero would not make sense if all potential predictors are thought to have non-zero, though perhaps inconsequentially small, effects. On the other hand, routinely using very vague priors for all potential effects would often lead to identifiability problems or poor predictive ability. Weakly informative priors might then be used, which typically "shrink" the effect towards zero. Gelman et al. (2008), for example, recommend a default Cauchy prior for logistic regression. For a review and comparison of such "shrinkage" priors for linear regression, see O'Hara and Sillanpää (2009), and for binary and survival regression see Rockova et al. (2012).

8.9 Discussion on model comparison

Broadly, there are two rather different approaches to Bayesian model comparison — one based on Bayes factors (or BIC) and the other on DIC, AIC, or similar measures. We can contrast the approaches under the following headings:

- *Assumptions.* The Bayes factor approach attempts to identify the "correct" model and implicitly assumes that such a thing exists and is in the families of distributions being considered. Posterior probabilities of models rely on it being meaningful to place probabilities on models. DIC/AIC makes no such assumption and only seeks short-term predictive ability.

- *Prior distributions.* Bayes factors require proper prior distributions (although these could be unit-information priors, as in BIC), which are not required for DIC/AIC.

- *Computation.* Bayes factors are notoriously difficult to compute in MCMC, requiring problem-specific programming or approximation; computation of DIC is generally straightforward.

- *Model uncertainty.* Model averaging to account for uncertainty about model choice is natural within a Bayes factor approach, provided one is willing to specify and interpret prior and posterior probabilities on models. Otherwise, DIC or Akaike weights, or bootstrapping, could be used for model averaging, though the theoretical justification is weak.

Working within an expanded model is a more flexible approach to model uncertainty which does not require averaging over a discrete set of models.

- *The "focus" of the analysis.* When dealing with hierarchical models, different model comparison methods can be related to which aspect of the model is of primary interest (§10.8.1).

The first issue is the most important: the situations in which it is reasonable to assume that any particular model is "true" appear very limited. We would therefore argue that the use of Bayes factors is restricted to domains where competing models correspond to clear, identifiable hypotheses that could in principle be proven to be "correct." Examples might include genetics applications where an individual is assumed to either carry some gene or not, or studies of supernatural abilities (as in Example 8.7.1) where the existence of any ability, however small, would be remarkable.

In either approach, we would recommend thorough criticism of model assumptions, as described in the first half of this chapter, and if there is model uncertainty, addressing it either formally or through clear sensitivity analyses.

8.10 Prior-data conflict

Bayesian analysis has traditionally focused on "turning the Bayesian handle," combining a prior distribution with a likelihood to produce a posterior distribution. But what if the likelihood and the prior seem in conflict, in the sense that they support almost non-intersecting areas of the parameter space? A naive approach would just charge on regardless, but this can lead to absurd results: for example, if we assume a normal observation $y \sim \mathrm{N}(\theta, 1)$ with standard normal prior $\theta \sim \mathrm{N}(0,1)$, then an observation $y = 10$ will lead to a posterior distribution $\theta \sim \mathrm{N}(5, 0.5)$, which is tightly situated around $\theta = 5$, a value supported neither by prior nor data (we note that if we instead assumed Student's t distributions we would obtain a bimodal posterior distribution, as in Example 8.6.3). This has been nicely ridiculed by Stephen Senn's characterisation of a Bayesian as someone who, suspecting a donkey and catching a glimpse of what looks like a horse, strongly concludes he has seen a mule (Senn (1997), p. 46, Spiegelhalter et al. (2004), p. 63).

There are two broad approaches to handling conflict: "identification" and "accommodation." Throughout this discussion we generally assume that the data are given priority and, in the case of conflict, it is the prior distribution that is called into question and discarded if necessary. However, the techniques can be easily adapted to give priority to the prior and discard divergent data, essentially adapting techniques previously used for identifying outliers.

8.10.1 Identification of prior-data conflict

This approach considers the current prior as a null hypothesis and checks whether the data fit the prior model or not. It is essentially a p-value argument as described for model checking (§8.4), in which an observed summary statistic t_0 is compared to a predictive distribution $p_0(t) = \int p(t|\theta)\, p(\theta) d\theta$, but using predictions arising from the prior rather than from the posterior distribution. The aim is to identify conflict and then one can decide whether to question the prior or the data.

As a simple example, assume the prior is $\theta \sim \mathrm{N}(\mu, \sigma^2/n_0)$ and the sampling distribution is $Y_m \sim \mathrm{N}(\theta, \sigma^2/m)$. Then the predictive distribution is $Y_m \sim \mathrm{N}(\mu, \sigma^2/m + \sigma^2/n_0)$ and so the predictive p-value is

$$\Pr(Y_m < y_m) = \Phi\left(\frac{y_m - \mu}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{m}}} \right).$$

We note that this is also the tail area associated with a standardised test statistic contrasting the likelihood and the prior: i.e., suppose we assumed $Y_m \sim \mathrm{N}(\theta_1, \sigma^2/m)$ and interpret the prior distribution as resulting from an observation $\mu \sim \mathrm{N}(\theta_2, \sigma^2/n_0)$, then a classical test of the null hypothesis $H_0 : \theta_1 = \theta_2$ would be based on

$$z_m = \frac{y_m - \mu}{\sigma\sqrt{\frac{1}{n_0} + \frac{1}{m}}},$$

a measure of *conflict* between data and prior.

We use one-sided p-values throughout, identifying both high and low values as "interesting."

Example 8.10.1. *Surgery (continued): assessing prior-data conflict*
In Example 1.1.1 we considered a prior distribution for a mortality rate that could be expressed as a Beta(3,27), which has a mean of 10%. In Example 2.7.1 we then assumed that 20 operations were to take place and obtained the predictive probability of the number of successes. Suppose, however, that after the first five operations there had been two deaths, that is, 40% mortality — is this grounds for deciding that the prior distribution was "wrong"?

We can calculate the predictive distribution for Y either in its beta-binomial closed form or by Monte Carlo simulation. Since this predictive distribution is discrete, we assume a mid-p-value, $\Pr(Y > y) + \frac{1}{2}\Pr(Y = y)$.

```
r.obs <- 2
```

```
theta  ~ dbeta(3, 27)
r      ~ dbin(theta, 5)   # sampling distribution
P     <- step(r-r.obs-0.5) + 0.5*equals(r, r.obs) # mid-p-value
```

The mean of P is 0.054, suggesting some evidence of conflict with the prior distribution.

8.10.2 Accommodation of prior-data conflict

Suppose that instead of simply identifying conflict, we wanted to automatically accommodate it: we assume that in the case of conflict we would want to reject or downweight the prior distribution. A natural way of modelling this is to imagine competing priors, perhaps drawn from disagreeing experts. One prior might represent our current opinion and be given substantial prior weight, while an alternative could represent a weak prior covering a wider range of alternatives: this is a natural application of mixture priors (§5.4) in which the idea is essentially to get the data to "choose" between the alternatives.

Example 8.10.2. *Surgery (continued): mixture of priors*
Our prior for the underlying mortality risk in the previous example was Beta(3,27). But suppose a claim was made that the procedure was much more dangerous than this; in fact the mortality rate could be around 50%. Such a prior opinion might be represented by a Beta(3,3) distribution, which is symmetric with mean 0.5 and standard deviation $=\sqrt{0.5 \times 0.5/7} = 0.19$. Suppose, as above, out of the first five operations there are two deaths — what should we now believe about the true mortality rate? What do we expect to happen over the next 10 operations?

A crucial input is the relative belief in the two competing prior distributions, prior 1: $\theta \sim \text{Beta}(3,27)$ or prior 2: $\theta \sim \text{Beta}(3,3)$. We shall take them as initially equally plausible, corresponding to $q_1 = \Pr(\text{prior } 1) = 0.5$. The code shows how a "pick" formulation is used to select the appropriate parameters for the prior distribution.

```
model {
  theta        ~ dbeta(a[pick], b[pick])
  pick         ~ dcat(q[1:2])
  q[1]        <- 0.50
  q[2]        <- 0.50
  q.post[1] <- equals(pick, 1)   # = 1 if prior 1 picked
  q.post[2] <- equals(pick, 2)   # = 1 if prior 2 picked
  r            ~ dbin(theta, n)    # sampling distribution
  r.pred       ~ dbin(theta, m)    # predictive distribution
}
```

```
node      mean   sd     MC error 2.5%    median 97.5% start sample
q.post[1] 0.2416 0.4281 0.004438 0.0     0.0    1.0   1001  50000
q.post[2] 0.7584 0.4281 0.004438 0.0     1.0    1.0   1001  50000
y.pred    3.789  2.328  0.0177   0.0     4.0    8.0   1001  50000
theta     0.3786 0.1843 0.00164  0.07455 0.3872 0.721 1001  50000
```

Given these early results, there is now a 76% probability that the "sceptic prior is appropriate and that this is a high-risk operation, and we would n expect 4 (95% interval 0 to 8) deaths out of the next 10 operations. Such formulation may naturally lead to a bimodal posterior distribution for θ, as sho in Figure 8.11.

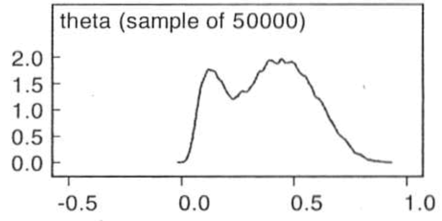

FIGURE 8.11
Posterior distribution for surgery mortality using a mixture of priors.

The formulation above works well when both priors are from the sar family of distributions. Alternatively, we could follow the approach of Exan ple 8.7.1 and model

```
y        ~ dbin(theta[pick], n)
theta[1] ~ dbeta(3, 27)
theta[2] ~ dbeta(3, 3)
```

which generalises easily to different parametric forms for the competing pri distributions.

This idea can be extended from a small set of possible prior distributior to a continuous mixture, and in so doing we can provide a "robust" pric that will have some influence if the data and prior agree, and otherwise wi be overwhelmed. Essentially this can be implemented by adopting a "heavy tailed" prior distribution that supports a wide range of possibilities but ha little influence in the extremes. For example, if we have a normal samplin distribution $Y \sim \text{N}(\mu, 1)$, but the prior distribution is a Student's t distribu tion, then in the case of conflict the prior is essentially ignored (Dawid, 1973 as the long tails of the prior "accommodate" the conflicting observation.

Using the ideas introduced in §8.2 and §8.5, we can express the t distribution as a normal distribution whose unknown precision is drawn from a chi-squared distribution. Specifically, suppose we thought that a reasonable prior distribution was normal with precision 1, but we wished to express some doubt about this assumption. If we take $\mu \sim \mathrm{N}(0, 1/\lambda)$, where $\lambda = X_k^2/k$ and $X_k^2 \sim \chi_k^2$, we are essentially assuming a t_k prior distribution for μ.

Example 8.10.3. *Prior robustness using a t prior distribution*
Suppose we assume $Y \sim \mathrm{N}(\mu, 1)$, a prior mean $E[\mu] = 0$, and we want to build in prior robustness by assuming a t_k distribution for μ. We shall illustrate this with $k = 1, 2, 4, 10, 50, 1000$; $k = 1$ corresponds to a very heavy-tailed Cauchy distribution with no mean or variance, while $k = 1000$ is essentially a normal distribution. We construct these t distributions as scale mixtures of normals, as in Example 8.2.1.

Suppose we then observe a single data point $y = 4$, apparently conflicting with the prior mean of 0.

```
y.obs <- 4
df[1] <-1; df[2] <- 2; df[3] <- 4
df[4] <- 10; df[5] <- 50; df[6] <- 1000
########################################
for (i in 1:6) {
  y[i]          <- y.obs         # replicate data
  y[i]           ~ dnorm(mu[i], 1)
  mu[i]          ~ dnorm(0, lambda[i])
  lambda[i]     <- X[i]/df[i]    # precision is chi-square/df
  X[i]           ~ dchisqr(df[i])
  # compare with prior distributions
  mu.rep[i]      ~ dnorm(0, lambda.rep[i])
  lambda.rep[i] <- X.rep[i]/df[i]
  X.rep[i]       ~ dchisqr(df[i])
}
```

```
node      mean   sd     MC error 2.5%    median 97.5% start sample
lambda[1] 0.2091 0.3384 0.004393 0.00338 0.1054 1.072 1001  10000
lambda[2] 0.3047 0.3576 0.004708 0.0155  0.1954 1.252 1001  10000
lambda[3] 0.4709 0.3992 0.005804 0.05558 0.3619 1.536 1001  10000
lambda[4] 0.6988 0.3487 0.004382 0.2132  0.6339 1.56  1001  10000
lambda[5] 0.9299 0.1895 0.001948 0.5993  0.9165 1.333 1001  10000
lambda[6] 0.9961 0.045  3.477E-4 0.9095  0.9951 1.086 1001  10000
mu[1]     3.449  1.072  0.01359  1.321   3.456  5.482 1001  10000
mu[2]     3.212  1.065  0.01376  1.135   3.198  5.316 1001  10000
mu[3]     2.859  1.035  0.01576  0.9113  2.833  4.97  1001  10000
mu[4]     2.444  0.8902 0.01013  0.7911  2.42   4.277 1001  10000
mu[5]     2.084  0.7478 0.007878 0.6675  2.071  3.575 1001  10000
mu[6]     2.004  0.7031 0.00638  0.6264  1.996  3.401 1001  10000
```

The estimate of μ based on the Cauchy ($k = 1$) is hardly influenced by the pri and a low value for λ is estimated. The normal ($k = 1000$) has the posterior mea mid-way between the data and the prior — an implausible conclusion whichev is true — and estimates λ to be almost 1.

We could think of a prior t distribution as a sensitivity analysis when w are unsure of a reasonable prior variance for a parameter with a normal prio Assuming a t prior leads to the data taking preference if there is apparei "conflict" with the prior mean, since if the data and the prior mean are vei different, this will tend to support the assumption of a large prior varianc and so tends to assume that λ is small.