Chapter 3
Bayesian Inference

3.1 Introduction

In the Bayesian approach to inference, all *unknown* quantities contained in a probability model for the observed data are treated as random variables. This is in contrast to the frequentist view described in Chap. 2 in which parameters are treated as fixed *constants*. Specifically, with respect to the inferential targets of Sect. 2.1, the fixed but unknown parameters and hypotheses are viewed as random variables under the Bayesian approach. Additionally, the unknowns may include missing data, or the true covariate value in an errors-in-variables setting.

The structure of this chapter is as follows. In Sect. 3.2 we describe the constituents of the posterior distribution and its summarization and in Sect. 3.3 consider the asymptotic properties of Bayesian estimators. Section 3.4 examines prior specification, and in Sect. 3.5 issues relating to model misspecification are discussed. Section 3.6 describes one approach to accounting for model uncertainty via Bayesian model averaging. As we see in Sect. 3.2, to implement the Bayesian approach, integration over the parameter space is required, and historically this has proved a significant hurdle to the routine use of Bayesian methods. Consequently, we discuss implementation issues in some detail. In Sect. 3.7, we provide a description of so-called conjugate situations in which the required integrals are analytically tractable, before providing an overview of analytical and numerical integration techniques, importance sampling, and direct sampling from the posterior. One particular technique, Markov chain Monte Carlo (MCMC), has greatly extended the range of models that may be analyzed with Bayesian methods, and Sect. 3.8 is devoted to a description of MCMC. Section 3.9 considers the important topic of *exchangeability*, and in Sect. 3.10 hypothesis testing via so-called Bayes factors is discussed. Section 3.11 considers a hybrid approach to inference in which the likelihood is taken as the sampling distribution of an estimator and is combined with a prior via Bayes theorem. Concluding remarks appears in Sect. 3.12, including a comparison of frequentist and Bayesian approaches, and the chapter ends with bibliographic notes in Sect. 3.13.

J. Wakefield, *Bayesian and Frequentist Regression Methods*, Springer Series in Statistics, DOI 10.1007/978-1-4419-0925-1_3,

3.2 The Posterior Distribution and Its Summarization

Let $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_p]^{\mathrm{T}}$ denote all of the unknowns of the model, which we continue to refer to as parameters, and $\boldsymbol{y} = [y_1, \ldots, y_n]^{\mathrm{T}}$ the vector of observed data. Also let $\mathcal{I}$ represent all relevant information that is currently available to the individual who is carrying out the analysis, in addition to $\boldsymbol{y}$. In the following description, we assume for simplicity that each element of $\boldsymbol{\theta}$ is continuous.

Bayesian inference is based on the *posterior* probability distribution of $\boldsymbol{\theta}$ after observing $\boldsymbol{y}$, which is given by Bayes theorem:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \mathcal{I}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}, \mathcal{I})\pi(\boldsymbol{\theta} \mid \mathcal{I})}{p(\boldsymbol{y} \mid \mathcal{I})}. \tag{3.1}$$

There are two key ingredients: the *likelihood* function $p(\boldsymbol{y} \mid \boldsymbol{\theta}, \mathcal{I})$ and the *prior* distribution $\pi(\boldsymbol{\theta} \mid \mathcal{I})$. The latter represents the probability beliefs for $\boldsymbol{\theta}$ held *before* observing the data $\boldsymbol{y}$. Both are dependent upon the current information $\mathcal{I}$. Different individuals will have different information $\mathcal{I}$, and so in general their prior distributions (and possibly their likelihood functions) may differ. The denominator in (3.1), $p(\boldsymbol{y} \mid \mathcal{I})$, is a normalizing constant which ensures that the right-hand side integrates to one over the parameter space. Though of crucial importance, for notational convenience, from this point onwards we suppress the dependence on $\mathcal{I}$, to give

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y})},$$

where the normalizing constant is

$$p(\boldsymbol{y}) = \int_{\theta} p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{3.2}$$

and is the marginal probability of the observed data given the model, that is, the likelihood and the prior. Ignoring this constant gives

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})$$

or, more colloquially,

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}.$$

The use of the posterior distribution for inference is very intuitively appealing since it probabilistically combines the information on the parameters contained in the data and in the prior.

The manner by which inference is updated from prior to posterior extends naturally to the sequential arrival of data. Suppose first that $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ represent the current totality of data. Then the posterior is

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y}_1, \boldsymbol{y}_2)}. \tag{3.3}$$

Now consider a previous occasion at which only $\boldsymbol{y}_1$ was available. The posterior based on these data only is

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1) = \frac{p(\boldsymbol{y}_1 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y}_1)}.$$

After observing $\boldsymbol{y}_1$ and before observing $\boldsymbol{y}_2$, the "prior" for $\boldsymbol{\theta}$ corresponds to the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y}_1)$, since this distribution represents the current beliefs concerning $\boldsymbol{\theta}$. We then update via

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{y}_1)}{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1)}. \tag{3.4}$$

Factorizing the right-hand side of (3.3) gives

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \boldsymbol{\theta})}{p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1)} \times \frac{p(\boldsymbol{y}_1 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{p(\boldsymbol{y}_1)},$$

which equals the right-hand side of (3.4). Hence, consistent inference based on $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ is reached regardless of whether we produce the posterior in one or two stages. In the case of conditionally independent observations,

$$p(\boldsymbol{y}_1, \boldsymbol{y}_2 \mid \boldsymbol{\theta}) = p(\boldsymbol{y}_1 \mid \boldsymbol{\theta})p(\boldsymbol{y}_2 \mid \boldsymbol{\theta})$$

in (3.3) and

$$p(\boldsymbol{y}_2 \mid \boldsymbol{y}_1, \boldsymbol{\theta}) = p(\boldsymbol{y}_2 \mid \boldsymbol{\theta})$$

in (3.4).

At first sight, the Bayesian approach to inference is deceptively straightforward, but there are a number of important issues that must be considered in practice. The first, clearly vital, issue is prior specification. Second, once prior and likelihood ingredients have been decided upon, we need to summarize the (usually) multivariate posterior distribution, and as we will see, this summarization requires integration over the parameter space, which may be of high dimension. Finally, a Bayesian analysis must address the effect that possible model misspecification has on inference. Prior specification is taken up in Sect. 3.4 and model misspecification in Sect. 3.5. Next, posterior summarization is described.

Typically the posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is multivariate, and marginal distributions for parameters of interest will be needed. The univariate marginal distribution for θ_i is

$$p(\theta_i \mid \boldsymbol{y}) = \int_{\theta_{-i}} p(\boldsymbol{\theta} \mid \boldsymbol{y})\, d\boldsymbol{\theta}_{-i}, \tag{3.5}$$

where $\boldsymbol{\theta}_{-i}$ is the vector $\boldsymbol{\theta}$ excluding θ_i, that is, $\boldsymbol{\theta}_{-i} = [\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p]$. While examining the complete distribution will often be informative, reporting

summaries of this distribution is also useful. To this end moments and quantiles may be calculated. For example, the posterior mean is

$$\mathrm{E}[\theta_i \mid \boldsymbol{y}] = \int_{\theta_i} \theta_i p(\theta_i \mid \boldsymbol{y})\, d\theta_i. \tag{3.6}$$

The $100 \times q\%$ quantile, $\theta_i(q)$, with $0 < q < 1$ is found by solving

$$q = \Pr\left[\, \theta_i \leq \theta_i(q) \,\right] = \int_{-\infty}^{\theta_i(q)} p(\theta_i \mid \boldsymbol{y})\, d\theta_i. \tag{3.7}$$

The posterior median $\theta_i(0.5)$ is often an adequate summary of the location of the posterior marginal distribution.

Formally, the choice between posterior means and medians can be made by viewing point estimation as a decision problem. For simplicity suppose that θ is univariate and the action, a, is to choose a point estimate for θ. Let $L(\theta, a)$ denote the loss associated with choosing action a when θ is the true state of nature. The (posterior) expected loss of an action a is

$$\overline{L}(a) = \int_{\theta} L(\theta, a) p(\theta \mid \boldsymbol{y})\, d\theta \tag{3.8}$$

and the optimal choice is the action that minimizes the expected loss. Different loss functions lead to different estimates (Exercise 3.1). For example, minimizing (3.8) with the quadratic loss $L(\theta, a) = (\theta - a)^2$ leads to reporting the posterior mean, $\widehat{a} = \mathrm{E}[\theta \mid \boldsymbol{y}]$. The linear loss,

$$L(\theta, a) = \begin{cases} c_1(a - \theta) & \theta \leq a \\ c_2(\theta - a) & \theta > a \end{cases},$$

corresponds to a loss which is proportional to c_1 if we overestimate and to c_2 if we underestimate. This function leads to $\widehat{a}$ such that

$$\Pr(\theta \leq \widehat{a} \mid \boldsymbol{y}) = \frac{c_2}{c_1 + c_2} = \frac{c_2/c_1}{1 + c_2/c_1},$$

that is, $\widehat{a} = \theta\left(\frac{c_2}{c_1+c_2}\right)$, so that presenting a quantile is the optimal action. Notice that only the ratio of losses is required. When $c_1 = c_2$, under- and overestimation are deemed equally hazardous, and the median of the posterior should be reported.

A $100 \times p\,\%$ equi-tailed *credible interval* $(0 < p < 1)$ is provided by

$$\left[\, \theta_i\left(\{1-p\}/2\right),\ \theta_i\left(\{1+p\}/2\right) \,\right].$$

This interval is the one that is usually reported in the majority of Bayesian analyses carried out, since it is the easiest to calculate. However, in cases where the posterior

is skewed, one may wish to instead calculate a *highest posterior density* (HPD) interval in which points inside the interval have higher posterior density than those outside the interval. Such an interval is also the shortest credible interval.

Another useful inferential quantity is the *predictive* distribution for unobserved (e.g., future) observations $\boldsymbol{z}$. Under conditional independence, so that $p(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{z} \mid \boldsymbol{\theta})$, this distribution is

$$p(\boldsymbol{z} \mid \boldsymbol{y}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, d\boldsymbol{\theta}. \tag{3.9}$$

This derivation clearly assumes that the likelihood for the original data $\boldsymbol{y}$ is also appropriate for the unobserved observations $\boldsymbol{z}$.

The Bayesian approach therefore provides very natural inferential summaries. However, these summaries require the evaluation of integrals, and for most models, these integrals are analytically intractable. Methods for implementation are considered in Sects. 3.7 and 3.8.

3.3 Asymptotic Properties of Bayesian Estimators

Although Bayesian purists would not be concerned with the frequentist properties of Bayesian procedures, personally I find it reassuring if, for a particular model, a Bayesian estimator can be shown to be, as a minimum, consistent. Efficiency is also an interesting concept to examine.

We informally give a number of results, before referencing more rigorous treatments. We only consider parameter vectors of finite dimension. An important condition that we assume in the following is that the prior distribution is positive in a neighborhood of the true value of the parameter.

The famous Bernstein–von Mises theorem states that, with increasing sample size, the posterior distribution tends to a normal distribution whose mean is the MLE and whose variance–covariance matrix is the inverse of Fisher's information. Let $\boldsymbol{\theta}$ be the true value of a p-dimensional parameter, and suppose we are in the situation in which the data are independent and identically distributed. Denote the posterior mean by $\widetilde{\boldsymbol{\theta}}_n = \widetilde{\boldsymbol{\theta}}_n(\boldsymbol{Y}_n) = \mathrm{E}[\boldsymbol{\theta} \mid \boldsymbol{Y}_n]$ and the MLE by $\widehat{\boldsymbol{\theta}}_n$. Then,

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) + \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$$

and we know that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \to_d \mathrm{N}_p[\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta})^{-1}]$, where $\boldsymbol{I}(\boldsymbol{\theta})$ is the information in a sample of size 1 (Sect. 2.4.1). It can be shown that $\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) \to_p \mathbf{0}$ and so

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \to_d \mathrm{N}_p[\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta})^{-1}].$$

Hence, $\widetilde{\boldsymbol{\theta}}_n$ is $\sqrt{n}$-consistent and asymptotically efficient. It is important to emphasize that the effect of the prior diminishes as $n \to \infty$. As van der Vaart (1998, p. 140) dryly notes, "Apparently, for an increasing number of observations one's prior beliefs are erased (or corrected) by the observations."

The Bernstein–von Mises theorem is so-called because of the papers by Bernstein (1917) and von Mises (1931), though the theorem has been refined by a number of authors. For references and a recent treatment, see van der Vaart (1998, Sect. 10.2). An early paper on consistency of Bayesian estimators is Doob (1948) and again there have been many refinements; see van der Vaart (1998, Sect. 10.4). An important assumption is that the parameter space is finite. Diaconis and Freedman (1986) describe the problems that can arise in the infinite-dimensional case.

3.4 Prior Choice

The specification of the prior distribution is clearly a necessary and crucial aspect of the Bayesian approach. With respect to prior choice, an important first observation is that for all $\boldsymbol{\theta}$ for which $\pi(\boldsymbol{\theta}) = 0$, we necessarily have $p(\boldsymbol{\theta} \mid \boldsymbol{y}) = 0$, regardless of any realization of the observed data, which clearly illustrates that great care should be taken in excluding parts of the parameter space a priori.

We distinguish between two types of prior specification. In the first, which we label as *baseline prior* specification, we presume an analysis is required in which the prior distribution has "minimal impact," so that the information in the likelihood dominates the posterior. An alternative label for such an analysis is *objective Bayes*. For an interesting discussion of the merits of this approach, see Berger (2006). Other labels that have been put forward for such prior specification include reference, non-informative and nonsubjective. Such priors may be used in situations (for example, in a regulatory setting) in which one must be as "objective" as possible. There is a vast literature on the construction of objective Bayesian procedures, with an aim often being to define procedures which have good frequentist properties.

An analysis with a baseline prior may be the only analysis performed or, alternatively, may provide an analysis with which other analyses in which *substantive priors* are specified may be compared. Such substantive priors constitute the second type of specification in which the incorporation of contextual information is required. Once we have a candidate substantive prior, it is often beneficial to simulate hypothetical data sets from the prior and examine these realizations to see if they conform to what is desirable. A popular label for analyses for which the priors are, at least in part, based on subject matter information is *subjective Bayes*.

3.4.1 Baseline Priors

On first consideration it would seem that the specification of a baseline prior is straightforward since one can take

$$\pi(\boldsymbol{\theta}) \propto 1, \tag{3.10}$$

so that the posterior distribution is simply proportional to the likelihood $p(\boldsymbol{y} \mid \boldsymbol{\theta})$. There are two major difficulties with the use of (3.10), however.

The first difficulty is that (3.10) provides an improper specification (i.e. it does not integrate to a positive constant $< \infty$) unless the range of each element of $\boldsymbol{\theta}$ is finite. In some instances this may not be a practical problem if the posterior corresponding to the prior is proper and does not exhibit any aberrant behavior (examples of such behavior are presented shortly). A posterior arising from an improper prior may be justified as a limiting case of proper priors, though some statisticians are philosophically troubled by this argument. Another justification for an improper prior is that such a choice may be thought of as approximating a prior that is "locally uniform" close to regions where the likelihood is non-negligible (so that the likelihood dominates) and decreasing to zero outside of this region. Great care must be taken to ensure that the posterior corresponding to an improper prior choice is proper. For nonlinear models, for example, improper priors should never be used (as an example shortly demonstrates). It is difficult to give general guidelines as to when a proper posterior will result from an improper prior. For example, improper priors for the regression parameters in a generalized linear model (which are considered in detail in Chap. 6) will often, but not always, lead to a proper posterior.

Example: Binomial Model

Suppose $Y \mid p \sim \text{Binomial}(n, p)$, with an improper uniform prior on the logit of p, which we denote $\theta = \log[p/(1-p)]$. Then, $\pi(\theta) \propto 1$ implies a prior on p of

$$\pi(p) \propto [p(1-p)]^{-1},$$

which is, of course, also improper.[1] With this prior an improper posterior results if $y = 0$ (or $y = n$) since the non-integrable spike at $p = 0$ (or $p = 1$) remains in the posterior. Note that this prior results in the MLE being recovered as the posterior mean.

Example: Nonlinear Regression Model

To illustrate the non-propriety in a nonlinear situation, consider the simple model

$$Y_i \mid \theta \sim_{ind} \text{N}\left[\exp(-\theta x_i), \sigma^2\right], \tag{3.11}$$

for $i = 1, \ldots, n$, with $\theta > 0$ and σ^2 assumed known. With an improper uniform prior on θ, $\pi(\theta) = 1$, we label the resulting (unnormalized) "posterior" as

$$q(\theta \mid \boldsymbol{y}) = p(\boldsymbol{y} \mid \theta) \times \pi(\theta) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathrm{e}^{-\theta x_i}\right)^2\right].$$

[1] This prior is sometimes known as *Haldane's prior* (Haldane 1948).

As $\theta \to \infty$,

$$q(\theta \mid \boldsymbol{y}) \to \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2\right], \tag{3.12}$$

a constant, so that the posterior is improper, because the tail is non-integrable, that is,

$$\int_{\theta_c}^{\infty} q(\theta \mid \boldsymbol{y}) = \infty$$

for all $\theta_c > 0$. Intuitively, the problem is that as $\theta \to \infty$ the corresponding nonlinear curve does not move increasingly away from the data, but rather to the asymptote $\mathrm{E}[Y \mid \theta] = 0$. The result is that a finite sum of squares results in (3.12), even in the limit. By contrast, there are no asymptotes in a linear model, and so as the parameters increase or decrease to $\pm\infty$, the fitted line moves increasingly far from the data which results in an infinite sum of squares in the limit, in which case the likelihood, and therefore the posterior, is zero. □

To summarize, it is ill-advised to think of improper priors as a default choice. Rather, improper priors should be used with care, and it is better to assume that they will lead to problems until the contrary can be shown. The safest strategy is clearly to specify proper priors, and this is the approach generally taken in this book.

The second difficulty with (3.10) is that if we reparameterize the model in terms of $\boldsymbol{\phi} = \boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{g}(\cdot)$ is a one-one mapping, then the prior for $\boldsymbol{\phi}$ corresponding to (3.10) is

$$\pi(\boldsymbol{\phi}) = \left|\frac{d\boldsymbol{\theta}}{d\boldsymbol{\phi}}\right|,$$

so that, unless $\boldsymbol{g}$ is a linear transformation, the prior is no longer constant. We have just seen an example of this with the binomial model. As another example, consider a variance σ^2, with prior $\pi(\sigma^2) \propto 1$. This choice implies a prior for the standard deviation, $\pi(\sigma) \propto \sigma$, which is nonconstant. The problem is that we cannot be "flat" on different nonlinear scales. This issue indicates that a desirable property in constructing baseline priors is their invariance to parameterization in order to obtain the same prior regardless of the starting parameterization.

A number of methods have been proposed for the specification of baseline or non-informative priors (we avoid the latter term since it is arguable that priors are ever non-informative). Jeffreys (1961, Sect. 3.10) suggested the use of

$$\pi(\boldsymbol{\theta}) \propto |\boldsymbol{I}(\boldsymbol{\theta})|^{1/2}, \tag{3.13}$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ is Fisher's expected information. This prior has the desirable property of invariance to reparameterization. The invariance holds in general but is obvious in the case of univariate θ. If $\phi = g(\theta)$,

$$I_{\phi}(\phi) = I_{\theta}(\theta) \times \left(\frac{d\theta}{d\phi}\right)^2, \tag{3.14}$$

where the subscripts now emphasize the parameterization. Consequently, if we start with

$$\pi_\phi(\phi) \propto I_\phi(\phi)^{1/2}$$

this implies

$$\pi_\theta(\theta) \propto I_\phi \left[g^{-1}(\phi)\right]^{1/2} \left|\frac{d\phi}{d\theta}\right| = I_\theta(\theta)^{1/2}$$

from (3.14). Hence, prior (3.13) results if we use the prescription of Jeffreys, but begin with ϕ. In the case of $Y \mid p \sim \text{Binomial}(n, p)$ the information is $I(p) = n/[p(1-p)]$ (Sect. 2.4.1). Therefore, Jeffreys prior is $\pi(p) \propto [p(1-p)]^{-1/2}$. This prior has the advantage of producing a proper posterior when $y = 0$ or $y = n$, a property not shared by Haldane's prior.

Unfortunately, the application of the above procedure to multivariate $\boldsymbol{\theta}$ can lead to posterior distributions that have undesirable characteristics. For example, in the Neyman–Scott problem, the use of Jeffreys prior gives, as $n \to \infty$, a limiting posterior mean that is inconsistent, in a frequentist sense (see Exercise 3.3).

A refinement of Jeffreys approach for selecting priors on a more objective basis is provided by *reference priors*. We briefly describe this approach heuristically; more detail can be found in Bernardo (1979) and Berger and Bernardo (1992). For any prior/likelihood distribution, suppose we can calculate the expected information concerning a parameter of interest that will be provided by the data. The more informative the prior, the less information the data will provide. An infinitely large sample would provide all of the missing information about the quantity of interest, and the reference prior is chosen to maximize this missing information.

3.4.2 Substantive Priors

The specification of substantive priors is obviously context specific, but we give a number of general considerations. Specific models will be considered in subsequent chapters. In this section we will discuss some general techniques but will not describe prior elicitation in any great detail; see Kadane and Wolfson (1998), O'Hagan (1998), and Craig et al. (1998) and the ensuing discussion for more on this topic which can be thought of as the measurement of probabilities.

When specifying a substantive prior, it is obvious that we need a clear understanding of the meaning of the parameters of the model for which we are specifying priors, and this can often be achieved by reparameterization.

Example: Linear Regression

Consider the simple linear regression $\mathrm{E}[Y \mid x] = \gamma_0 + \gamma_1 x$. Interpretation is often easier if we reparameterize as

$$\mathrm{E}[Y \mid z] = \beta_0 + \beta_1(z - \overline{z})$$

where $z = c \times x$ and c is chosen so that the units of z are convenient. Under this parameterization, β_0 is the expected response at $z = \overline{z}$. It will often be easier to specify a prior for β_0 than for γ_0, the average response at $x = 0$, which may be meaningless. The slope parameter, β_1, is the change in expected response corresponding to a c-unit increase in x (1-unit increase in z).

Example: Exponential Regression

It may be easier to specify priors on observable quantities, before transforming back to the parameters. For the nonlinear model (3.11), we might specify a prior for the expected response at $x = \widehat{x}$, $\phi = \exp(-\theta\,\widehat{x})$ to give a prior $\pi_\phi(\phi)$. The prior for θ is

$$\pi_\theta(\theta) = \pi_\phi\left[\exp(-\theta\,\widehat{x})\right] \times \widehat{x}\exp(-\theta\,\widehat{x}),$$

the last term corresponding to the Jacobian of the transformation $\phi \to \theta$. As an example, one might assume a $\mathrm{Be}(a, b)$ prior for ϕ, with a and b chosen to give a 90% interval for ϕ. □

While the axioms of probability are uncontroversial, the interpretation of probability has been contested for centuries. In the frequentist approach of Chap. 2, probability was defined in an objective frequentist sense. If the event A is of interest and an experiment is repeated n times resulting in n_A occasions on which A occurs, then

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n}.$$

In contrast, in the subjective Bayesian worldview, probabilities are viewed as subjective and conditional upon an individual's experiences and knowledge, although one may of course base subjective probabilities upon frequencies. Cox and Hinkley (1974, p. 53) state, with reference to the use of Bayes theorem, "If the prior distribution arises from a physical random mechanism with known properties, this argument is entirely uncontroversial," but continue, "A frequency prior is, however, rarely available. To apply the Bayesian approach more generally a wider concept of probability is required . . . the prior distribution is taken as measuring the investigator's subjective opinion about the parameter from evidence other than the data under analysis."

As alluded to by this last quote, an obvious procedure is to base the prior distribution upon previously collected data. Ideally, preliminary modeling of such data should be carried out to acknowledge sampling error. If one believed that the data-generation mechanism for both sets of data was comparable, then it would be logical to base the posterior on the combined data (and then once again one has to decide on how to pick a prior distribution). Often such comparability is not reasonable, and a conservative approach is to take the prior as the posterior based

on the additional data, but with an inflated variance, to accommodate the additional uncertainty. This approach acknowledges nonsystematic differences, but systematic differences (in particular, biases in one or both studies) may also be present, and this is more difficult to deal with.

Roughly speaking, so long as the prior does not assign zero mass to any region, the likelihood will dominate with increasing sample size (as we saw in Sect. 3.3), so that prior choice becomes decreasingly important. A very difficult problem in prior choice is the specification of the *joint distribution* over multiple parameters. In some contexts one may be able to parameterize the model so that one believes a priori that the components are independent, but in general this will not be possible.

Due to the difficulties of prior specification, a common approach is to carry out a *sensitivity analysis* in which a range of priors are considered and the *robustness* of inference to these choices is examined. An alternative is to *model average* across the different prior models; see Sect. 3.10.

3.4.3 Priors on Meaningful Scales

As we will see in Chaps. 6 and 7, loglinear and linear logistic forms are extremely useful regression models, taking the forms

$$\log \mu = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$
$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

retrospectively, where $\mu = \mathrm{E}[Y]$. Both forms are examples of generalized linear models (GLMs) which are discussed in some detail in Chap. 6.

Often there will be sufficient information in the data for $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^{\mathrm{T}}$ to be analyzed using independent normal priors with large variances (unless, for example, there are many correlated covariates). The use of an improper prior for $\boldsymbol{\beta}$ will often lead to a proper posterior though care should be taken. Chapter 5 discusses prior choice for the linear model and Chap. 6 for GLMs, and Sect. 6.8 provides an example of an improper posterior that arises in the context of a Poisson model with a linear link.

If we wish to use informative priors for $\boldsymbol{\beta}$, we may specify independent normal priors, with the parameters for each component being obtained via specification of two quantiles with associated probabilities. For loglinear and logistic models, these quantiles may be given on the exponentiated scale since these are more interpretable (as the rate ratio and odds ratio, respectively). If θ_1, θ_2 are the quantiles and p_1, p_2 are the associated probabilities, then the parameters of the normal prior are

$$\mu = \frac{z_1\theta_2 - z_2\theta_1}{z_1 - z_2} \tag{3.15}$$

$$\sigma = \frac{\theta_1 - \theta_2}{z_1 - z_2} \tag{3.16}$$

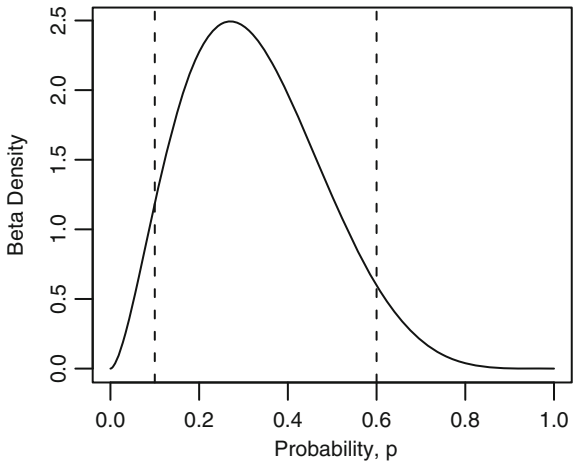

Fig. 3.1 The beta prior, $\text{Be}(2.73, 5.67)$, which gives $\Pr(p < 0.1) = 0.05$ and $\Pr(p < 0.6) = 0.95$

where z_1 and z_2 are the quantiles of a standard normal random variable. For example, in an epidemiological context with a Poisson regression model, we may wish to specify a prior on a relative risk parameter, $\exp(\beta_1)$ which has a median of 1 (corresponding to no association) and a 95% point of 3 (if we think it is unlikely that the relative risk associated with a unit increase in exposure exceeds 3). If we take $\theta_1 = \log(1)$ and $\theta_2 = \log(3)$, along with $p_1 = 0.5$ and $p_2 = 0.95$, then we obtain $\beta_1 \sim \text{N}(0, 0.668^2)$. In general, less care is required in prior choice for intercepts in GLMs since they are very accurately estimated with even small amounts of data.

Many candidate prior distributions contain two parameters. For example, a beta prior may be used for a probability and lognormal or gamma distributions may be used for positive parameters such as measures of scale. A convenient way to choose these parameters is to, as above, specify two quantiles with associated probabilities and then solve for the two parameters. For example, suppose we wish to specify a beta prior, $\text{Be}(a_1, a_2)$, for a probability p, such that the p_1 and p_2 quantiles are q_1 and q_2. Then we may solve

$$[p_1 - \Pr(p < q_1 \mid a_1, a_2)]^2 + [p_2 - \Pr(p < q_2 \mid a_1, a_2)]^2 = 0$$

for a_1, a_2. For example, taking $p_1 = 0.05, p_2 = 0.95, q_1 = 0.1, q_2 = 0.6$ yields $a_1 = 2.73$, $a_2 = 5.67$, and Fig. 3.1 shows the resulting density.

3.4.4 Frequentist Considerations

We briefly give a simple example to illustrate the frequentist bias-variance trade-off of prior specification, by examining the mean squared error (MSE) of a Bayesian estimator. Consider data Y_i, $i = 1, \ldots, n$, with Y_i independently and identically

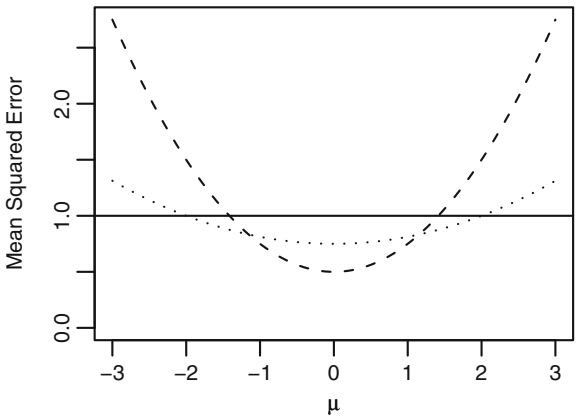

Fig. 3.2 Mean squared error of the posterior mean estimator when $\sqrt{n}(\overline{Y}_n - \mu) \to_d \mathrm{N}(0, \sigma^2)$ with σ^2 known and prior $\mu \sim \mathrm{N}(m, v)$. The *dashed line* represents the case with $v = 1$ and the *dotted line* when $v = 3$, as a function of the parameter μ. The mean squared error of the sample mean is the *solid horizontal line*

distributed with $\mathrm{E}[Y_i \mid \mu] = \mu$ and $\mathrm{var}(Y_i \mid \mu) = \sigma^2$ with σ^2 known. The asymptotic distribution of the sample mean is

$$\sqrt{n}(\overline{Y}_n - \mu) \to_d \mathrm{N}(0, \sigma^2).$$

We treat this distribution as the likelihood and examine a Bayesian analysis with prior

$$\mu \sim \mathrm{N}(m, v).$$

The posterior is

$$\mu \mid \boldsymbol{Y}_n \to_d \mathrm{N}\left(w_n \overline{Y}_n + (1 - w_n)m, w_n \frac{\sigma^2}{n}\right)$$

where

$$w_n = \frac{nv}{nv + \sigma^2}.$$

We first observe that the posterior mean estimator is consistent since $w_n \to 1$ as $n \to \infty$), so long as $v > 0$, but the estimator has finite sample bias if $v^{-1} \neq 0$. The mean squared error of the posterior mean estimator is

$$\begin{aligned} \mathrm{MSE} &= \mathrm{Variance} + \mathrm{Bias}^2 \\ &= w_n \frac{\sigma^2}{n} + [w_n \mu + (1 - w_n)m - \mu]^2 \\ &= w_n \frac{\sigma^2}{n} + (1 - w_n)^2 (m - \mu)^2. \end{aligned}$$

Figure 3.2 illustrates the MSE as a function of μ for two different prior distributions that are both centered at zero but have different variances of $v = 1, 3$. For simplicity

we have chosen $\sigma^2/n = 1$ with $n = 9$ (so that the MSE of the sample mean is 1, and is indicated as the solid horizontal line). The trade-off when specifying the variance of the prior is clear; if the true μ is close to m, then reductions in MSE are achieved with a small v, though the range of μ over which an improved MSE is achieved is narrower than with the wider prior. At values of μ of $m \pm \sqrt{v + \sigma^2/n}$, the MSE of the sample mean and Bayesian estimator are equal. The variance of the estimator is given by the lowest point of the MSE curves, and the bias dominates for large $|\mu|$.

Example: Lung Cancer and Radon

As an example of prior specification, we return to the simple model considered repeatedly in Chap. 2 with likelihood

$$Y_i \mid \boldsymbol{\beta} \sim_{ind} \text{Poisson}\left[\, E_i \exp(\beta_0 + \beta_1 x_i) \,\right],$$

where recall that Y_i are counts of lung cancer incidence in Minnesota in 1998–2002, and x_i is a measure of residential radon in county i, $i = 1, \ldots, n$. The obvious improper prior here is $\pi(\boldsymbol{\beta}) \propto 1$ (and results in a proper posterior for this likelihood).

To specify a substantive prior, we need to have a clear interpretation of the parameters, and β_0 and β_1 are not the most straightforward quantities to contemplate. Hence, we reparameterize the model as

$$Y_i \mid \boldsymbol{\theta} \sim_{ind} \text{Poisson}\left(E_i \theta_0 \theta_1^{x_i - \overline{x}}\right),$$

where $\boldsymbol{\theta} = [\theta_0, \theta_1]^{\mathrm{T}}$ so that

$$\theta_0 = \mathrm{E}[Y/E \mid x = \overline{x}] = \exp(\beta_0 + \beta_1 \overline{x})$$

is the expected standardized mortality ratio in an area with average radon. The standardization that leads to expected numbers E implies we would expect θ_0 to be centered around 1. The parameter $\theta_1 = \exp(\beta_1)$ is the relative risk associated with a one-unit increase in radon. Due to ecological bias, studies often show a negative association between lung cancer incidence and radon (and it is this ecological association we are estimating for this illustration and not the individual-level association). We take lognormal priors for θ_0 and θ_1 and use (3.15) and (3.16) to deduce the lognormal parameters. For θ_0 we take a lognormal prior with 2.5% and 97.5% quantiles of 0.67 and 1.5 to give $\mu = 0, \sigma = 0.21$. For θ_1 we assume the relative risk associated with a one-unit increase in radon is between 0.8 and 1.2 with probability 0.95, to give $\mu = -0.02, \sigma = 0.10$. We return to this example later in the chapter.

3.5 Model Misspecification

The behavior of Bayesian estimators under misspecification of the likelihood has received less attention than frequentist estimators. Recall the result concerning the behavior of the MLE $\widehat{\boldsymbol{\theta}}_n$ under model misspecification summarized in (2.27), which we reproduce here:

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\text{T}}) \;\rightarrow_d\; \text{N}_p\left[\mathbf{0}, \boldsymbol{J}^{-1}\boldsymbol{K}(\boldsymbol{J}^{\text{T}})^{-1}\right]$$

where

$$\boldsymbol{J} = \text{E}_{\text{T}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\text{T}}}\log p(Y \mid \boldsymbol{\theta}_{\text{T}})\right]$$

$$\boldsymbol{K} = \text{E}_{\text{T}}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta}_{\text{T}})\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log p(Y \mid \boldsymbol{\theta}_{\text{T}})\right)^{\text{T}}\right]$$

with $\boldsymbol{\theta}_{\text{T}}$ the true $\boldsymbol{\theta}$ and $p(Y \mid \boldsymbol{\theta})$ the assumed model. Let $\widetilde{\boldsymbol{\theta}}_n = \text{E}[\boldsymbol{\theta} \mid \boldsymbol{Y}_n]$ be the posterior mean which we here view as a function of $\boldsymbol{Y}_n = [Y_1, \ldots, Y_n]^{\text{T}}$. From Sect. 3.3, $\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) \rightarrow_p \mathbf{0}$, and hence

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\text{T}}) \rightarrow_d \text{N}_p\left[\mathbf{0}, \boldsymbol{J}^{-1}\boldsymbol{K}(\boldsymbol{J}^{\text{T}})^{-1}\right].$$

This has important implications since it shows that, asymptotically, the effect of model misspecification on the posterior mean is the same as its effect on the MLE. If the likelihood is of linear exponential family form, correct specification of the mean function leads to consistent estimation of the parameters in the mean model (see Sect. 6.5.1 for details). As with the reported variance of the MLE, the spread of the posterior distribution could be completely inappropriate, however. While sandwich estimation can be used to "correct" the variance estimator for the MLE, there is no such simple solution for the posterior mean, or other Bayesian summaries.

With respect to model misspecification, the emphasis in the Bayesian literature has been on sensitivity analyses, or on embedding a particular likelihood or prior choice within a larger class. Embedding an initial model within a continuous class is a conceptually simple approach. For example, a Poisson model may be easily extended to a negative binomial model.

A difficulty with considering model classes with large numbers of unknown parameters is that uncertainty on parameters of interest will be increased if a simple model is closer to the truth. In particular, model expansion may lead to a decrease in precision, as we now illustrate. As we have seen, as n increases, the prior effect is negligible and the posterior variance is given by the inverse of Fisher's information, (Sect. 3.3). Suppose that we have k parameters in an original model, and we are considering an expanded model with p parameters, and let

$$\begin{bmatrix}\boldsymbol{I}_{11} & \boldsymbol{I}_{12}\\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22}\end{bmatrix},$$

where $\boldsymbol{I}_{11}$ is a $k \times k$ matrix corresponding to the information on the parameters of the simpler model (which includes the parameters of interest), and $\boldsymbol{I}_{22}$ is the $(p-k) \times (p-k)$ information matrix concerning the additional parameters in the enlarged model. In the simpler model, the information on the parameters of interest is $\boldsymbol{I}_{11}$, while for the enlarged model, it is

$$\boldsymbol{I}_{11} - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21},$$

which is never greater than $\boldsymbol{I}_{11}$. This is an oversimplified discussion (as we shall see in Sect. 5.9), but it highlights that there can be a penalty to pay for specifying an overly complex model.

3.6 Bayesian Model Averaging

If a discrete number of models are considered, then model averaging provides an alternative means of assessing model uncertainty. The Bayesian machinery handles multiple models in a very straightforward fashion since essentially the unknown model is treated as an additional discrete parameter. Let $M_1, \ldots, M_J$ denote the J models under consideration and $\boldsymbol{\theta}_j$ the parameters of the jth model. Suppose, for illustration, there is a parameter of interest ϕ (which we assume is univariate) that is well defined for each of the J models under consideration. The posterior for ϕ is a mixture over the J individual model posteriors:

$$p(\phi \mid \boldsymbol{y}) = \sum_{j=1}^{J} p(\phi \mid M_j, \boldsymbol{y}) \Pr(M_j \mid \boldsymbol{y})$$

where

$$\begin{aligned} p(\phi \mid M_j, \boldsymbol{y}) &= \int p(\phi \mid \boldsymbol{\theta}_j, M_j, \boldsymbol{y}) p(\boldsymbol{\theta}_j \mid M_j, \boldsymbol{y})\, d\boldsymbol{\theta}_j \\ &= \frac{1}{p(\boldsymbol{y} \mid M_j)} \int p(\phi \mid \boldsymbol{\theta}_j, M_j, \boldsymbol{y}) p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j) p(\boldsymbol{\theta}_j \mid M_j)\, d\boldsymbol{\theta}_j, \\ \Pr(M_j \mid \boldsymbol{y}) &= \frac{p(\boldsymbol{y} \mid M_j) \Pr(M_j)}{p(\boldsymbol{y})} \\ &= \frac{\int p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j) p(\boldsymbol{\theta}_j \mid M_j)\, d\boldsymbol{\theta}_j \Pr(M_j)}{p(\boldsymbol{y})} \end{aligned}$$

and with $\Pr(M_j)$ the prior belief in model j and $p(\boldsymbol{\theta}_j \mid M_j)$ the prior on the parameters of model M_j. The marginal probabilities of the data under the different models are calculated as

$$p(\boldsymbol{y} \mid M_j) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}_j, M_j) p(\boldsymbol{\theta}_j \mid M_j)\, d\boldsymbol{\theta}_j,$$

with

$$p(\boldsymbol{y}) = \sum_{j=1}^{J} p(\boldsymbol{y} \mid M_j) \Pr(M_j).$$

To summarize the posterior for ϕ, we might report the posterior mean

$$\mathrm{E}[\phi \mid \boldsymbol{y}] = \sum_{j=1}^{J} \mathrm{E}[\phi \mid \boldsymbol{y}, M_j] \times \Pr(M_j \mid \boldsymbol{y}),$$

which is simply the average of the posterior means across models, weighted by the posterior weight received by each model. The posterior variance is

$$\begin{aligned}\mathrm{var}(\phi \mid \boldsymbol{y}) = &\sum_{j=1}^{J} \mathrm{var}(\phi \mid \boldsymbol{y}, M_j) \times \Pr(M_j \mid \boldsymbol{y}) \\ &+ \sum_{j=1}^{J} \{\mathrm{E}[\phi \mid \boldsymbol{y}, M_j] - \mathrm{E}[\phi \mid \boldsymbol{y}]\}^2 \times \Pr(M_j \mid \boldsymbol{y})\end{aligned}$$

which averages the posterior variances concerning ϕ in each model, with the addition of a term that accounts for between-model uncertainty in the mean.

Although model averaging is very appealing in principle, in practice there are many difficult choices, including the choice of the class of models to consider and the priors over both the models and the parameters of the models. Summarization can also be difficult because the parameter of interest may have different interpretations in different models. For example, in a regression setting, suppose we fit the single model

$$\mathrm{E}[Y \mid x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

with β_1 the parameter of interest. The interpretation of β_1 is as the average change in response corresponding to a unit increase in x_1, with x_2 held constant. If we average over this model and the model with x_1 only, then the usual "x_2 held constant" qualifier is not accurate, so a phrase such as "allowing for the possibility of x_2 in the model" may be instead used. Performing model averaging over models which represent different scientific theories is also not appealing if the search for a causal explanation is sought. If prediction is the aim, then model averaging is much more appealing since parameter interpretation is often irrelevant (see Chap. 12). Another disadvantage of model averaging is that it may encourage the user to believe they have accounted for "all" uncertainty in which covariates to include in the model which is a dangerous conclusion to draw.

3.7 Implementation

In this section we provide an overview of methods for evaluating the integrals required for performing Bayesian inference. We begin, in Sect. 3.7.1, by describing so-called conjugate situations in which the prior and likelihood combination is constructed in order for the posterior to be of the same form as the prior. Unfortunately, in a regression setting, conjugate analyses are rarely available beyond the linear model. In Sect. 3.7.2 the analytical Laplace approximation is described. Quadrature methods are considered in Sect. 3.7.3 before we turn to a method that combines Laplace and numerical integration in a very clever way, in Sect. 3.7.4, to give a method known as the *integrated nested Laplace approximation* (INLA). More recently developed sampling-based (Monte Carlo) approaches have transformed the practical application of Bayesian methods, and we therefore describe these approaches in some detail. In Sect. 3.7.5, importance sampling Monte Carlo is considered, and in Sects. 3.7.6 and 3.7.7, direct sampling from the posterior is described. MCMC algorithms are particularly important, and to these we devote Sect. 3.8.

Beyond the crucial importance of integration in Bayesian inference, this material is also relevant in a frequentist context. Specifically, in Part III of this book, we will consider nonlinear and generalized linear mixed effects models for which integration over the random effects is required in order to obtain the likelihood for the fixed effects.

3.7.1 Conjugacy

So-called *conjugate prior* distributions allow analytical evaluation of many of the integrals required for Bayesian inference, at least for certain convenient parameters. A conjugate prior is such that $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and $p(\boldsymbol{\theta})$ belong to the same family. We assume $\dim(\boldsymbol{\theta}) = p$. This definition is not adequate since it will always be true given a suitable definition of the family of distributions. To obtain a more useful class, we first note that if $\boldsymbol{T}(\boldsymbol{Y})$ denotes a *sufficient statistic* for a particular likelihood $p(\cdot \mid \boldsymbol{\theta})$, then

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = p(\boldsymbol{\theta} \mid \boldsymbol{t}) \propto p(\boldsymbol{t} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

This allows a definition of a conjugate family in terms of likelihoods that admit a sufficient statistic of fixed dimension.

The p-parameter exponential family of distributions has the form:

$$p(y_i \mid \boldsymbol{\theta}) = f(y_i) g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\lambda}(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{u}(y_i)\right],$$

where, in general, $\boldsymbol{\lambda}(\boldsymbol{\theta})$ and $\boldsymbol{u}(y_i)$ have the same dimension as $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}(\boldsymbol{\theta})$ is called the *natural parameter* (and in a linear exponential family, we have $\boldsymbol{u}(y_i) = y_i$). For n independent and identically distributed observations from $p(\cdot \mid \boldsymbol{\theta})$,

Table 3.1 Conjugate priors and associated posterior distributions, for various likelihood choices

Prior	Likelihood	Posterior
$\theta \sim \text{N}(m, v)$	$\overline{Y} \mid \theta \sim \text{N}(\theta, \sigma^2/n)$ σ^2 known	$\theta \mid \boldsymbol{y} \sim \text{N}[w\overline{y} + (1-w)m, w\sigma^2/n]$ with $w = v/(v+\sigma^2/n)$
$\theta \sim \text{Be}(a, b)$	$Y \mid \theta \sim \text{Bin}(n, \theta)$	$\theta \mid y \sim \text{Be}(a+y, b+n-y)$
$\theta \sim \text{Ga}(a, b)$	$Y \mid \theta \sim \text{Poisson}(\theta)$	$\theta \mid y \sim \text{Ga}(a+y, b+1)$
$\theta \sim \text{Ga}(a, b)$	$Y \mid \theta \sim \text{Exp}(\theta)$	$\theta \mid y \sim \text{Ga}(a+y, b+1)$

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \left[\prod_{i=1}^{n} f(y_i)\right] g(\boldsymbol{\theta})^n \exp\left[\boldsymbol{\lambda}(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{t}(\boldsymbol{y})\right],$$

where

$$\boldsymbol{t}(\boldsymbol{y}) = \sum_{i=1}^{n} \boldsymbol{u}(y_i).$$

The conjugate prior density is defined as

$$p(\boldsymbol{\theta}) = c(\eta, \boldsymbol{\upsilon}) \times g(\boldsymbol{\theta})^{\eta} \exp\left[\boldsymbol{\lambda}(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{\upsilon}\right],$$

where η and $\boldsymbol{\upsilon}$ are specified, a priori. The resulting posterior distribution is

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = c(\eta + n, \boldsymbol{\upsilon} + \boldsymbol{t}) \times g(\boldsymbol{\theta})^{\eta+n} \exp\left\{\boldsymbol{\lambda}(\boldsymbol{\theta})^{\mathrm{T}} [\boldsymbol{\upsilon} + \boldsymbol{t}(\boldsymbol{y})]\right\},$$

demonstrating conjugacy. Comparison with $p(y_i \mid \boldsymbol{\theta})$ indicates that η may be viewed as a *prior sample size* giving rise to a *sufficient statistic* $\boldsymbol{\upsilon}$.

The above derivations are often not required if one wishes to simply obtain the conjugate distribution for a given likelihood, since it can be determined quickly via inspection of the kernel of the likelihood. The predictive distribution is often more complex to derive, however, but is straightforward under the above formulation. In the case of a conjugate prior, for new observations $\boldsymbol{Z} = [Z_1, ..., Z_m]$ arising as an independent and identically distributed sample from $p(Z \mid \boldsymbol{\theta})$, the *predictive distribution* is

$$p(\boldsymbol{z} \mid \boldsymbol{y}) = \left[\prod_{i=1}^{m} f(z_i)\right] \frac{c[\eta + n, \boldsymbol{\upsilon} + \boldsymbol{t}(\boldsymbol{y})]}{c[\eta + n + m, \boldsymbol{\upsilon} + \boldsymbol{t}(\boldsymbol{y}, \boldsymbol{z})]}.$$

Table 3.1 gives the conjugate choices for a variety of likelihoods.

Beyond the normal linear model, the direct practical use of conjugacy in a regression setting is limited, but as we will see subsequently, the material of this section is very useful when implementing direct sampling or MCMC approaches.

Example: Binomial Likelihood

Suppose we have a single observation from a binomial distribution, $Y \mid \theta \sim$ Binomial(n, θ):

$$p(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

By direct inspection we recognize that the conjugate prior is a beta distribution, but for illustration we follow the more long-winded route. In exponential family form,

$$p(y \mid \theta) = \binom{n}{y} (1-\theta)^n \exp\left[y \log\left(\frac{\theta}{1-\theta}\right)\right],$$

or, in terms of the natural parameter $\lambda = \lambda(\theta) = \log[\theta/(1-\theta)]$,

$$p(y \mid \lambda) = \binom{n}{y} [1+\exp(\lambda)]^{-n} \exp(y\lambda).$$

The conjugate prior for λ is therefore identified as

$$\pi(\lambda) = c(\eta, \upsilon)[1+\exp(\lambda)]^{-\eta} \exp[\upsilon\lambda] \tag{3.17}$$

so that the prior for θ is

$$\begin{aligned}\pi(\theta) &= c(\eta,\upsilon)(1-\theta)^{\eta} \exp\left[\upsilon \log\frac{\theta}{1-\theta}\right] \frac{1}{\theta(1-\theta)} \\ &= \frac{\Gamma(\eta+2)}{\Gamma(\upsilon+1)\Gamma(\eta-\upsilon+1)} \theta^{\upsilon-1}(1-\theta)^{\eta-\upsilon-1},\end{aligned}$$

the Be(a, b) distribution with parameters $a = \upsilon, b = \eta - \upsilon$. An interpretation of these parameters is that a prior sample size $\eta = a + b$ yields the prior sufficient statistic $\upsilon = a$. It follows immediately that the posterior is Be$(a + y, b + n - y)$.

We write

$$\begin{aligned}\mathrm{E}[\theta \mid y] &= \frac{a+y}{a+b+n} \\ &= \frac{y}{n} w + \frac{a}{a+b}(1-w)\end{aligned}$$

where $w = n/(a+b+n)$, so that the posterior mean is a weighted combination of the MLE, $\widehat{\theta} = y/n$, and the prior mean. Similarly,

$$\begin{aligned}\mathrm{mode}[\theta \mid y] &= \frac{a+y-1}{a+b+n-2} \\ &= \frac{y}{n} w^{\star} + \frac{a-1}{a+b-2}(1-w^{\star}),\end{aligned}$$

where $w^\star = n/(a+b+n-2)$, so that the posterior mode is a weighted combination of the prior mode (if it exists) and the MLE. The choice of a uniform distribution, $a = b = 1$, results in the posterior mode equaling the MLE, as expected in this one-dimensional example.

The marginal distribution of the data, given likelihood and prior, is the beta-binomial distribution

$$\Pr(y) = \binom{n}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \times \frac{\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(a+b+n)},$$

for $y = 0, \ldots, n$. If $a = b = 1$, the prior predictive is uniform over the space of outcomes: $p(y) = (n+1)^{-1}$ for $y = 0, 1, \ldots, n$, in line with intuition.

The mean of the prior predictive is

$$\mathrm{E}[Y] = \mathrm{E}_\theta[\mathrm{E}(Y \mid \theta)] = n \times \frac{a}{a+b},$$

with variance

$$\mathrm{var}(Y) = \mathrm{var}_\theta[\mathrm{E}(Y \mid \theta)] + \mathrm{E}_\theta[\mathrm{var}(Y \mid \theta)] = n\mathrm{E}(\theta)[1 - \mathrm{E}(\theta)] \times \frac{a+b+n}{a+b+1},$$

illustrating the overdispersion relative to $\mathrm{var}(Y \mid \theta) = n\theta(1-\theta)$, if $n > 1$. If $n = 1$, there is no overdispersion since we have a single Bernoulli random variable for which the variance is always determined by the mean.

The predictive distribution for a new trial, in which $Z = 0, 1, \ldots, m$ denotes the number of successes and m the number of trials, is

$$p(z \mid y) = \binom{m}{z} \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)} \times \frac{\Gamma(a+b+z)\Gamma(b+n-y+m-z)}{\Gamma(a+b+n+m)},$$

which is another version of the beta-binomial distribution and is an overdispersed binomial for which

$$\mathrm{E}[Z \mid y] = m \times \mathrm{E}[\theta \mid y] = m \times \frac{a+y}{a+b+n},$$

and

$$\mathrm{var}(Z \mid y) = m \times \mathrm{E}(\theta \mid y) \times [1 - \mathrm{E}(\theta \mid y)] \times \frac{a+b+n+m}{a+b+n+1}.$$

As $n \to \infty$, with y/n fixed, the predictive $p(z \mid y)$ approaches the binomial distribution $\mathrm{Bin}(m, y/n)$. This makes sense since, under correct model specification, for large n we effectively know θ, and so binomial variability is the only uncertainty that remains.

3.7.2 Laplace Approximation

In this section let

$$I = \int_{-\infty}^{\infty} \exp[\, nh(\theta)\,]\, d\theta, \tag{3.18}$$

denote a generic integral of interest, and we suppose initially that θ is a scalar. Depending on the form of $h(\cdot)$, (3.18) can correspond to the evaluation of a variety of quantities of interest including $p(\boldsymbol{y})$ and posterior moments. The n appearing in (3.18) is included solely to make the asymptotic arguments more transparent.

Let $\widetilde{\theta}$ denote the mode of $h(\cdot)$. We carry out a Taylor series expansion about $\widetilde{\theta}$, assuming that $h(\cdot)$ is sufficiently well behaved for this operation; in particular we assume that at least two derivatives exist. The expansion is

$$nh(\theta) = n \sum_{k=0}^{\infty} \frac{(\theta - \widetilde{\theta})^k}{k!} h^{(k)}(\widetilde{\theta}),$$

where $h^{(k)}(\widetilde{\theta})$ represents the kth derivative of $h(\cdot)$ evaluated at $\widetilde{\theta}$. Hence,

$$\begin{aligned} I &= \int_{-\infty}^{\infty} \exp\left[n \sum_{k=0}^{\infty} \frac{(\theta - \widetilde{\theta})^k}{k!} h^{(k)}(\widetilde{\theta}) \right] d\theta \\ &\approx \exp\left[nh(\widetilde{\theta}) \right] \int_{-\infty}^{\infty} \exp\left[\frac{nh^{(2)}}{2}(\widetilde{\theta})(\theta - \widetilde{\theta})^2 \right] d\theta, \end{aligned}$$

where we have ignored quadratic terms and above in the Taylor series and exploited $h^{(1)}(\widetilde{\theta}) = 0$. Writing $\widetilde{v} = -1/h^{(2)}(\widetilde{\theta})$ gives the estimate

$$\widehat{I} = \exp\left[nh(\widetilde{\theta}) \right] \left(\frac{2\pi \widetilde{v}}{n} \right)^{1/2}, \tag{3.19}$$

which is known as the *Laplace approximation*. The error is such that

$$\frac{I}{\widehat{I}} = 1 + O(n^{-1}).$$

Suppose we wish to evaluate the posterior expectation of a positive function of interest $\phi(\theta)$, that is,

$$\begin{aligned} \mathrm{E}[\phi(\theta) \mid \boldsymbol{y}] &= \frac{\int \exp[\log \phi(\theta) + \log p(\boldsymbol{y} \mid \theta) + \log \pi(\theta) + \log(d\theta/d\phi)]\, d\theta}{\int \exp[\log p(\boldsymbol{y} \mid \theta) + \log \pi(\theta)]\, d\theta} \\ &= \frac{\int \exp[nh_1(\theta)]\, d\theta}{\int \exp[nh_2(\theta)]\, d\theta}. \end{aligned}$$

where the Jacobian has been included in the numerator of the first line. Application of (3.19) to numerator and denominator gives

$$\widehat{\mathrm{E}}[\phi(\theta) \mid \boldsymbol{y}] = \frac{\widetilde{v}_1}{\widetilde{v}_0} \frac{\exp[nh_1(\widetilde{\theta}_1)]}{\exp[nh_0(\widetilde{\theta}_0)]}$$

where $\widetilde{\theta}_j$ is the mode of $h_j(\cdot)$ and $\widetilde{v}_j = -1/h_j^{(2)}(\widetilde{\theta}_j)$, $j = 0, 1$. Further,

$$\widehat{\mathrm{E}}[\phi(\theta) \mid \boldsymbol{y}] = \mathrm{E}[\phi(\theta) \mid \boldsymbol{y}][1 + O(n^{-2})],$$

since errors in the numerator and denominator cancel (Tierney and Kadane 1986). If ϕ is not positive then a simple solution is to add a large constant to ϕ, apply Laplace's method, and subtract the constant.

Now consider multivariate $\boldsymbol{\theta}$ with $\dim(\boldsymbol{\theta}) = p$ and with required integral

$$I = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp[nh(\boldsymbol{\theta})]\, d\theta_1 \cdots d\theta_p.$$

The above argument may be generalized to give the Laplace approximation

$$\widehat{I} = \exp\left[nh(\widetilde{\boldsymbol{\theta}})\right] \left(\frac{2\pi}{n}\right)^{p/2} \mid \widetilde{\boldsymbol{v}} \mid^{1/2}, \tag{3.20}$$

where $\widetilde{\boldsymbol{\theta}}$ is the maximum of $h(\cdot)$ and $\widetilde{\boldsymbol{v}}$ is the $p \times p$ matrix whose (i, j)th element is

$$-\left.\frac{\partial^2 h}{\partial\theta_i \partial\theta_j}\right|_{\widetilde{\boldsymbol{\theta}}}.$$

An important drawback of analytic approximations is the difficulty in performing error assessment, so that in practice one does not know the accuracy of approximation. The evaluation of derivatives can also be analytically and numerically troublesome. These shortcomings apart, however, we will see that these approximations are useful as components of other approaches, such as the scheme described in Sect. 3.7.4, and for suggesting proposals for importance sampling and MCMC algorithms.

3.7.3 *Quadrature*

We consider numerical integration rules for approximating integrals of the form

$$I = \int f(t)\, dt,$$

via the weighted sum

$$\widehat{I} = \sum_{i=1}^{m} w_i f(t_i),$$

where the points t_i and weights w_i define the integration rule. So-called *Gauss* rules are optimal rules (in a sense we will define shortly) that are constructed to integrate weighted functions of polynomials accurately. Specifically, if $p(t)$ is a polynomial of degree $2m-1$, then the Gauss rule (t_i, w_i) is such that

$$\sum_{i=1}^{m} w_i p(t_i) = \int w(t)p(t)\, dt.$$

It can be shown that no rule has this property for polynomials of degree $2m$, showing the optimality of Gauss rules. Different classes of rule emerge for different choices of weight function. We describe Gauss–Hermite rules that correspond to the weight function

$$w(t) = \exp(-t^2) \tag{3.21}$$

which is of obvious interest in a statistics context. If the integral is of the form

$$I = \int g(t) \exp(-t^2)\, dt$$

and $f(t)$ can be well approximated by a polynomial of degree $2m-1$, we would expect an m-point Gauss–Hermite rule to be accurate.

The points of the Gauss–Hermite rule are the zeroes of the Hermite polynomials $H_m(t)$ with weights

$$w_i = \frac{2^{m-1} m! \sqrt{\pi}}{m^2 [H_{m-1}(t_i)]^2}.$$

In general, the points of the rule need to be located and scaled appropriately. Suppose that μ and σ are the approximate mean and standard deviation of θ, and let $t = (\theta - \mu)/\sqrt{2}\sigma$. The integral of interest is

$$I = \int f(\theta)\, d\theta = \int g(\mu + \sqrt{2}\sigma t)\sqrt{2}\sigma \mathrm{e}^{-t^2}\, dt$$

and applying the transformation yields

$$\widehat{I} = \sum_{i=1}^{m} w_i^\star g(t_i^\star),$$

where $w_i^\star = w_i \sqrt{2}\sigma$ and $t_i^\star = \mu + \sqrt{2}\sigma t_i$.

In practice μ and σ are unknown but may be estimated at the same time as I is evaluated to give an *adaptive Gauss–Hermite* rule (Naylor and Smith 1982).

Suppose $\boldsymbol{\theta}$ is two-dimensional, and we wish to evaluate

$$I = \int f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int \int f(\theta_1, \theta_2)\, d\theta_2\, d\theta_1 = \int f^\star(\theta_1)\, d\theta_1$$

where

$$f^\star(\theta_1) = \int f(\theta_1, \theta_2)\, d\theta_2.$$

We form

$$\widehat{I} = \sum_{i=1}^{m_1} w_i \widehat{f^\star}(\theta_{1i}),$$

with

$$\widehat{f^\star}(\theta_{1i}) = \sum_{j=1}^{m_2} u_j f(\theta_{1i}, \theta_{2j})$$

to give

$$\widehat{I} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_i u_j f(\theta_{1i}, \theta_{2j}),$$

which is known as a *Cartesian Product* rule. Such rules can provide very accurate integration with relatively few points, but the number of points required is prohibitive in high dimensions since for p parameters and m points, a total of m^p points are required. Consequently, these rules tend to be employed when $p \leq 10$.

In common with the Laplace method, quadrature methods do not provide an estimate of the error of the approximation. In practice, consistency of the estimates across increasing grid sizes may be examined.

3.7.4 Integrated Nested Laplace Approximations

We briefly review the INLA computational approach which combines Laplace approximations and numerical integration in a very efficient manner; see Rue et al. (2009) for a more extensive treatment. Consider a model with parameters $\boldsymbol{\theta}_1$ that are assigned normal priors, with the remaining parameters being denoted $\boldsymbol{\theta}_2$ with $G = \dim(\boldsymbol{\theta}_1)$ and $V = \dim(\boldsymbol{\theta}_2)$. Assume for ease of explanation that the normal prior is centered at zero with variance–covariance matrix $\boldsymbol{\Sigma}$, $\mathrm{N}_G(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ depends on elements in $\boldsymbol{\theta}_2$. Many models fall into this class including generalized linear models (Chap. 6) and generalized linear mixed models (Chap. 9). The posterior is

$$\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2)\pi(\boldsymbol{\theta}_2) \prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$
$$\propto \pi(\boldsymbol{\theta}_2) \mid \boldsymbol{\Sigma}(\boldsymbol{\theta}_2) \mid^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{\theta}_1^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\theta}_2)^{-1}\boldsymbol{\theta}_1 + \sum_{i=1}^{n} \log p(\boldsymbol{y}_i \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right]. \tag{3.22}$$

Of particular interest are the posterior univariate marginal distributions $\pi(\theta_{1g} \mid \boldsymbol{y})$, $g = 1, \ldots, G$, and $\pi(\theta_{2v} \mid \boldsymbol{y})$, $v = 1, \ldots, V$. The "normal" parameters $\boldsymbol{\theta}_1$ are dealt with by analytical approximations (as applied to the term in the exponent of (3.22), conditional on specific values of $\boldsymbol{\theta}_2$). Numerical integration techniques are applied to $\boldsymbol{\theta}_2$, so that V should not be too large for accurate inference (Sect. 3.7.3). For elements of $\boldsymbol{\theta}_1$ we write

$$\pi(\theta_{1g} \mid \boldsymbol{y}) = \int \pi(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2, \boldsymbol{y}) \times \pi(\boldsymbol{\theta}_2 \mid \boldsymbol{y})\, d\boldsymbol{\theta}_2$$

which may be evaluated via the approximation

$$\widetilde{\pi}(\theta_{1g} \mid \boldsymbol{y}) = \int \widetilde{\pi}(\theta_{1g} \mid \boldsymbol{\theta}_2, \boldsymbol{y}) \times \widetilde{\pi}(\boldsymbol{\theta}_2 \mid \boldsymbol{y})\, d\boldsymbol{\theta}_2$$
$$\approx \sum_{k=1}^{K} \widetilde{\pi}(\theta_{1g} \mid \boldsymbol{\theta}_2^{(k)}, \boldsymbol{y}) \times \widetilde{\pi}(\boldsymbol{\theta}_2^{(k)} \mid \boldsymbol{y}) \times \Delta_k \tag{3.23}$$

for a set of weights Δ_k, $k = 1, \ldots, K$. Laplace or related analytical approximations are applied to carry out the integration (over $\boldsymbol{\theta}_{1g'}, g' \neq g$) required for evaluation of $\widetilde{\pi}(\theta_{1g} \mid \boldsymbol{\theta}_2, \boldsymbol{y})$. To produce the grid of points $\{\boldsymbol{\theta}_2^{(k)}, k = 1, \ldots, K\}$ over which numerical integration is performed, the mode of $\widetilde{\pi}(\boldsymbol{\theta}_2 \mid \boldsymbol{y})$ is located and the Hessian is approximated, from which the grid of points $\{\boldsymbol{\theta}_2^{(k)}, k = 1, \ldots, K\}$, with associated weights Δ_k, is created and used in (3.23), as was described in Sect. 3.7.3. The output of INLA consists of posterior marginal distributions, which can be summarized via means, variances, and quantiles.

3.7.5 Importance Sampling Monte Carlo

The first sampling-based technique we describe directly estimates the required integrals. To motivate importance sampling Monte Carlo, consider the one-dimensional integral

$$I = \int_0^1 f(\theta)\, d\theta = \mathrm{E}[f(\theta)],$$

where the expectation is with respect to the uniform distribution, $\mathrm{U}(0,1)$. This formulation suggests the obvious estimator

$$\widehat{I}_m = \frac{1}{m}\sum_{t=1}^{m} f(\theta^{(t)}),$$

with $\theta^{(t)} \sim_{iid} \mathrm{U}(0,1)$, $t = 1, \ldots, m$. By the central limit theorem (Appendix G),

$$\sqrt{m}(\widehat{I}_m - I) \;\to_d\; \mathrm{N}[0, \mathrm{var}(f)],$$

where $\mathrm{var}(f) = \mathrm{E}[f(\theta)^2] - I^2$ and we have assumed the latter exists. The form of the variance reveals that the efficiency of the method is determined by how variable the function f is, with respect to the uniform distribution over $[0,1]$. If f were constant, we would have zero variance!

To achieve an approximately constant function, we can trivially rewrite the integral as

$$I = \int f(\theta)\, d\theta = \int \frac{f(\theta)}{g(\theta)} g(\theta)\, d\theta = \mathrm{E}_g\left[\frac{f(\theta)}{g(\theta)}\right], \tag{3.24}$$

where we no longer restrict θ to lie in (0, 1). Define the estimator

$$\widehat{I}_m = \frac{1}{m}\sum_{t=1}^{m} \frac{f(\theta^{(t)})}{g(\theta^{(t)})},$$

where $\theta^{(t)} \sim_{iid} g(\cdot)$, with

$$\sqrt{m}(\widehat{I}_m - I) \;\to_d\; \mathrm{N}[0, \mathrm{var}(f/g)],$$

and

$$\mathrm{var}(f/g) = \mathrm{E}_g\left[\left(\frac{f}{g}\right)^2\right] - I^2.$$

The latter may be estimated by

$$\widehat{\mathrm{var}}(f/g) = \frac{1}{m}\sum_{t=1}^{m}\left(\frac{f(\theta^{(t)})}{g(\theta^{(t)})}\right)^2 - \widehat{I}_m^2.$$

Consequently, the aim is to find a density that closely mimics f (up to proportionality), so that the Monte Carlo estimator will have low variance because samples from *important* regions of the parameter space (where the function is large) are being drawn, hence the label *importance sampling Monte Carlo*. A great strength of importance sampling is that it produces not only an estimate of I but a measure of uncertainty also. Specifically, we may construct the 95% confidence interval

$$\left[\widehat{I}_m - 1.96\frac{\sqrt{\widehat{\text{var}}(f/g)}}{\sqrt{m}}, \widehat{I}_m + 1.96\frac{\sqrt{\widehat{\text{var}}(f/g)}}{\sqrt{m}}\right]. \tag{3.25}$$

It may seem strange to be utilizing an asymptotic frequentist interval estimate when evaluating an integral for Bayesian inference, but in this context the "sample size" m is controlled by the user and is large so that an asymptotic interval is uncontroversial (since a flat prior on I would give the same Bayesian interval).

Efficient use of importance sampling critically depends on finding a suitable $g(\cdot)$. From the form of $\text{var}(f/g)$, it is clear that if the support of θ is infinite, $g(\cdot)$ must dominate in the tails; otherwise, the variance will be infinite and the estimate will not be useful in practice (even though the estimator is unbiased). It is also desirable to have a $g(\cdot)$ which is computationally inexpensive to sample from. Student's t, or mixtures of Student's t distributions (West 1993), perhaps with iteration to tune the proposal, are popular.

3.7.6 Direct Sampling Using Conjugacy

The emergence of methods to sample from the posterior distribution have revolutionized the practical applicability of the Bayesian inferential approach. Such methods utilize the duality between samples and densities: Given a sample, we can reconstruct the density and functions of interest, and given an arbitrary density, we can almost always generate a sample, given the range of generic random variate generators available. With respect to the latter, the ability to obtain *direct* samples from a distribution decreases as the dimensionality of the parameter space increases, and MCMC methods provide an attractive alternative. However, as discussed in Sect. 3.8, a major practical disadvantage to the use of MCMC is that the generated samples are dependent which complicates the calculation of Monte Carlo standard errors. Automation of MCMC algorithms is also not straightforward since an assessment of the convergence of the Markov chain is required. Further, it is not straightforward to calculate marginal densities such as (3.5) with MCMC. For problems with small numbers of parameters, direct sampling methods provide a strong competitor to MCMC, primarily because independent samples from the posterior are provided and no assessment of convergence is required.

Suppose we have generated independent samples $\{\boldsymbol{\theta}^{(t)}, t = 1, \ldots, m\}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, with $\boldsymbol{\theta}^{(t)} = [\theta_1^{(t)}, \ldots, \theta_p^{(t)}]$; we describe how such samples may be used for inference. The univariate marginal posterior for $p(\theta_j \mid \boldsymbol{y})$ may be approximated by the histogram constructed from the points $\theta_j^{(t)}$, $t = 1, \ldots, m$. Posterior means $\text{E}[\theta_j \mid \boldsymbol{y}]$ may be approximated by

$$\widehat{\text{E}}[\theta_j \mid \boldsymbol{y}] = \frac{1}{m}\sum_{t=1}^{m} \theta_j^{(t)},$$

with other moments following in an obvious fashion. Coverage probabilities of the form $\Pr(a < \theta_j < b \mid \boldsymbol{y})$ are estimated by

$$\widehat{\Pr}(a < \theta_j < b \mid \boldsymbol{y}) = \frac{1}{m} \sum_{t=1}^{m} I\left(a < \theta_j^{(t)} < b\right),$$

with $I(\cdot)$ representing the indicator function which is 1 if its argument is true and 0 otherwise. The central limit theorem (Appendix G) allows the accuracy of these approximations to be simply determined since the samples are independent.

We discuss how to estimate the standard error associated with the estimate

$$\widehat{\mu}_m = \frac{1}{m} \sum_{t=1}^{m} \theta^{(t)} \tag{3.26}$$

of $\mu = \mathrm{E}[\theta \mid \boldsymbol{y}]$. By the strong law of large numbers, $\widehat{\mu}_m \to_{a.s.} \mu$ as $m \to \infty$, and the central limit theorem (Appendix G) gives

$$\sqrt{m}(\widehat{\mu}_m - \mu) \to_d \mathrm{N}(0, \sigma^2)$$

where $\sigma^2 = \mathrm{var}(\theta \mid \boldsymbol{y})$ (assuming this variance exists). The Monte Carlo standard error is $\sigma/\sqrt{m}$, with consistent estimate of σ:

$$\widehat{\sigma}_m = \sqrt{\frac{1}{m} \sum_{t=1}^{m} (g(\boldsymbol{\theta}^{(t)}) - \widehat{\mu}_m)^2}.$$

By Slutsky's theorem (Appendix G)

$$\frac{\widehat{\mu}_m - \mu}{\widehat{\sigma}_m/\sqrt{m}} \to_d \mathrm{N}(0, 1)$$

as $m \to \infty$. An asymptotic confidence interval for μ is therefore

$$\widehat{\mu}_m \pm 1.96 \times \frac{\widehat{\sigma}_m}{\sqrt{m}}.$$

We may also wish to obtain standard errors for functions that are not simple expectations. For example, consider the posterior variance of a univariate parameter θ:

$$\sigma^2 = \mathrm{var}(\theta \mid \boldsymbol{y}) = \mathrm{E}[(\theta - \mu)^2 \mid \boldsymbol{y}].$$

where $\mu = \mathrm{E}[\theta \mid \boldsymbol{y}]$. An obvious estimator is

$$\widehat{\sigma}_m^2 = \frac{1}{m} \sum_{t=1}^{m} (\theta^{(t)} - \widehat{\mu}_m)^2$$

where $\widehat{\mu}_m$ is given by (3.26). Now,

$$\sqrt{m}\left(\begin{bmatrix}\widehat{\mu}_m\\ \widehat{\sigma}^2_m\end{bmatrix} - \begin{bmatrix}\mu\\ \sigma^2\end{bmatrix}\right) \to_d \mathrm{N}_2\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\sigma^2 & \mu^*_3\\ \mu^*_3 & \mu^*_4 - \sigma^4\end{bmatrix}\right)$$

where $\mu^*_j = \mathrm{E}[(\theta - \mu)^j \mid \boldsymbol{y}]$ is the jth central moment, $j = 3, 4$ (where we assume that these quantities exist). The standard error of $\widehat{\sigma}^2$ is estimated by

$$\sqrt{\frac{\widehat{\mu}^*_{4,m} - \widehat{\sigma}^4_m}{m}} \tag{3.27}$$

where $\widehat{\mu}^*_{4,m} = \frac{1}{m}\sum_{t=1}^{m}(\theta^{(t)} - \widehat{\mu}_m)^4$ which can, unfortunately, be highly unstable. Therefore, accurate interval estimates for σ^2 require larger sample sizes than are needed for accurate estimates for μ.

Once samples from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ are obtained, it is straightforward to convert to samples for a parameter of interest $g(\boldsymbol{\theta})$ via $g(\boldsymbol{\theta}^{(t)})$. This property is important in a conjugate setting since although we have analytical tractability for one set of parameters, we may be interested in functions of interest that are not so convenient. For example, with likelihood $Y \mid \theta \sim \text{Binomial}(n, \theta)$ and prior $\theta \sim \text{Be}(a, b)$, we know that $\theta \mid y \sim \text{Be}(a + y, b + n - y)$. However, suppose we are interested in the odds $g(\theta) = \theta/(1 - \theta)$. Given samples $\theta^{(t)}$ from the beta posterior, we can simply form $g(\theta^{(t)}) = \theta^{(t)}/(1 - \theta^{(t)})$, $t = 1, \ldots, m$. As an aside, in this setting, for a Bayesian analysis with a proper prior, the realizations $Y = 0$ or $Y = n$ do not cause problems, in contrast to the frequentist case in which the MLE for $g(\theta)$ is undefined.

3.7.7 Direct Sampling Using the Rejection Algorithm

The *rejection algorithm* is a generic and widely applicable method for generating samples from arbitrary probability distributions.

Theorem (Rejection Sampling).
Suppose we wish to sample from the distribution

$$f(x) = \frac{f^\star(x)}{\int f^\star(x)\, dx},$$

and we have a proposal distribution $g(\cdot)$ for which

$$M = \sup_x \frac{f^\star(x)}{g(x)} < \infty.$$

Then the algorithm:

1. *Generate $U \sim U(0, 1)$ and, independently, $X \sim g(\cdot)$.*

2. Accept X if

$$U < \frac{f^\star(X)}{Mg(X)},$$

otherwise return to 1,

produces accepted points with distribution $f(x)$, and the acceptance probability is

$$p_a = \frac{\int f^\star(x)\, dx}{M}.$$

Proof. The following is based on Ripley (1987). We have

$$\begin{aligned}
\Pr(X \le x \cap \text{ acceptance }) &= \Pr(X \le x) \Pr(\text{ acceptance } \mid X \le x) \\
&= \int_{-\infty}^{x} g(y) \Pr(\text{ acceptance } \mid y)\, dy \\
&= \int_{-\infty}^{x} g(y) \frac{f^\star(y)}{Mg(y)} dy = \int_{-\infty}^{x} \frac{f^\star(y)}{M} dy.
\end{aligned}$$

The probability of acceptance is

$$\Pr(\text{acceptance}) = \int_{-\infty}^{\infty} \frac{f^\star(y)}{M} dy = p_a.$$

The number of iterations until accepting a point is a geometric random variable with probability p_a. The expected number of iterations until acceptance is p_a^{-1}. It follows that

$$\begin{aligned}
\Pr(X \le x \mid \text{ acceptance}) &= \sum_{i=1}^{\infty} \Pr(\text{ acceptance on the } i\text{th trial }) \\
&= \sum_{i=1}^{\infty} (1-p_a)^{i-1} \int_{-\infty}^{x} \frac{f^\star(y)}{M} dy = \frac{1}{p_a} \int_{-\infty}^{x} \frac{f^\star(y)}{M} dy \\
&= \frac{M}{\int_{-\infty}^{\infty} f^\star(y)} \int_{-\infty}^{x} \frac{f^\star(y)}{M} dy = \int_{-\infty}^{x} f(y) dy,
\end{aligned}$$

as required. □

We describe a rejection algorithm that is convenient for generating samples from the posterior (Smith and Gelfand 1992). Let $\boldsymbol{\theta}$ denote the unknown parameters, and assume that we can evaluate the maximized likelihood

$$M = \sup_{\theta} p(\boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}})$$

where $\widehat{\boldsymbol{\theta}}$ is the MLE. The algorithm then proceeds as follows:

1. Generate $U \sim \mathrm{U}(0,1)$ and, independently, sample from the prior, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.
2. Accept $\boldsymbol{\theta}$ if

$$U < \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})}{M},$$

otherwise return to 1.

The probability that a point is accepted is

$$p_a = \frac{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{M} = \frac{p(\boldsymbol{y})}{M}.$$

This algorithm can be very easy to implement since finding the MLE can often be carried out routinely. We need then only generate points from the prior and evaluate the likelihood at these points. Rejection sampling from the prior is very intuitive; the prior supplies the points which are then "filtered out" via the likelihood.

The empirical rejection rate can be used to derive the normalizing constant as

$$\widetilde{p}(\boldsymbol{y}) = M \times \widehat{p}_a \tag{3.28}$$

which may be useful for model assessment/selection (Sect. 3.10). If we desire m samples from the posterior, the number of generations required from the prior $\pi(\cdot)$ is $m + m^\star$ (where $m^\star$ is the number of rejected points), and $m^\star$ is a negative binomial random variable (Appendix D). The MLE of p_a is $m/(m + m^\star)$.

An alternative importance sampling estimator of the normalizing constant that is more efficient than (3.28) is

$$\widehat{p}(\boldsymbol{y}) = \frac{1}{m + m^\star} \sum_{t=1}^{m+m^\star} p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(t)}), \tag{3.29}$$

where $\boldsymbol{\theta}^{(t)} \sim_{iid} \pi(\cdot)$, $t = 1, \ldots, m + m^\star$. Notice that there is no rejection of points associated with this calculation so that all $m + m^\star$ prior points are used. Although (3.29) is the more efficient estimator, (3.28) provides an alternative estimator as a by-product that is useful for code checking. The estimator (3.28) assumes that all normalizing constants are included in M. If the maximization has been carried out with respect to $M^\star = p^\star(\boldsymbol{y} \mid \boldsymbol{\theta})$ where $p^\star(\boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \boldsymbol{\theta})/c$, then we must instead use the estimate

$$\widetilde{p}(\boldsymbol{y}) = c \times M^\star \times \widehat{p}_a. \tag{3.30}$$

Posterior moments can be estimated directly as averages of the accepted points, or we may implement importance sampling estimators that use all points generated from the prior. For example, the posterior mean

$$\mathrm{E}[\theta \mid \boldsymbol{y}] = \frac{\int \theta p(\boldsymbol{y} \mid \theta)\pi(\theta)\, d\theta}{\int p(\boldsymbol{y} \mid \theta)\pi(\theta)\, d\theta} = \frac{\mathrm{E}\left[\theta p(\boldsymbol{y} \mid \theta)\right]}{\mathrm{E}\left[p(\boldsymbol{y} \mid \theta)\right]}$$

may be estimated by

$$\widehat{\mathrm{E}}[\theta \mid \boldsymbol{y}] = \frac{\frac{1}{m+m^\star}\sum_{t=1}^{m+m^\star} \theta^{(t)} p(\boldsymbol{y} \mid \theta^{(t)})}{\frac{1}{m+m^\star}\sum_{t=1}^{m+m^\star} p(\boldsymbol{y} \mid \theta^{(t)})},$$

where $\theta^{(t)} \sim_{iid} \pi(\cdot)$, $t = 1, \ldots, m + m^\star$.

Clearly we need a proper prior distribution to implement the above algorithm. The efficiency of the algorithm will depend on the correspondence between the likelihood and the prior, as measured through $p(\boldsymbol{y})$. For large n, the algorithm will become less efficient since the likelihood becomes increasingly concentrated, and so prior points are less likely to be accepted (which is another manifestation of the prior becoming less important with increasing sample size, Sect. 3.3).

The rejection algorithm that samples from the prior does not need the functional form of the prior to be available. As an example, Wakefield (1996) used a predictive distribution from a Bayesian analysis as the prior for the analysis of a separate dataset; samples from the predictive distribution could be simply generated, even though no closed form was available for this distribution.

Example: Poisson Likelihood, Lognormal Prior

We illustrate some of the technique described in the previous sections using a Poisson likelihood with data from a geographical cluster investigation carried out in the United Kingdom (Black 1984). The Sellafield nuclear site is located in the northwest of England on the coast of West Cumbria. Initially, the site produced plutonium for defense purposes and subsequently carried out the reprocessing of spent fuel from nuclear power stations in Britain and abroad and stored and discharged to sea low-level radioactive waste. Seascale is a village 3 km to the south of Sellafield and had $y = 4$ cases of lymphoid malignancy among 0–14 year olds during 1968–1982, compared with $E = 0.25$ expected cases (based on the number of children in the region and registration rates for the overall northern region of England). A question here is whether such a large number of cases could have reasonably occurred by chance. There is substantial information available on the incidence of childhood leukemia across the United Kingdom as a whole.

We assume the model $Y \mid \theta \sim \text{Poisson}[E\exp(\theta)]$, where θ is the log relative risk (the ratio of the risk in the study region, to that in the northern region), the MLE of which is $\widehat{\theta} = \log(16) = 2.77$ with asymptotic standard error 0.25. We assume an $\mathrm{N}(\mu, \sigma^2)$ normal prior for θ, which is equivalent to a lognormal prior $\text{LogNorm}(\mu, \sigma^2)$ for $\exp(\theta)$. To choose the prior parameters, we assume, for illustration, that the median relative risk is 1 and the 90% point of the prior is 10, which leads, from (3.15) and (3.16), to $\mu = 0$ and $\sigma^2 = 1.38^2$.

We will estimate

$$\begin{aligned} I_r &= \int_{-\infty}^{\infty} \theta^r \Pr(y \mid \theta)\pi(\theta)\, d\theta \\ &= \frac{E^y (2\pi b^2)^{-1/2}}{y!} \int \exp\left[r \log\theta - E\exp(\theta) + \theta y - \frac{(\theta - a)^2}{2b^2} \right] d\theta \\ &= \frac{E^y (2\pi b^2)^{-1/2}}{y!} \int \exp[h_r(\theta)]\, d\theta \end{aligned}$$

for $r = 0, 1, 2$, to give the normalizing constant and posterior mean and variance as

$$\begin{aligned} p(\boldsymbol{y}) &= I_0 \\ \mathrm{E}[\theta \mid y] &= \frac{I_1}{I_0} \\ \mathrm{var}(\theta \mid y) &= \frac{I_2}{I_0} - \frac{I_1^2}{I_0^2}. \end{aligned}$$

We choose to calculate the posterior variance not because it is a quantity of particular interest but because it provides a summary that is not particularly easy to estimate and so reveals some of the complications of the various methods.

To apply the Laplace method, we first give the first and second derivatives of $h_r(\theta)$:

$$\begin{aligned} h_r^{(1)}(\theta) &= \frac{r}{\theta} - E\exp(\theta) + y - \frac{\theta - a}{b^2} \\ h_r^{(2)}(\theta) &= -\frac{r}{\theta^2} - E\exp(\theta) - \frac{1}{b^2}, \end{aligned}$$

for $r = 0, 1, 2$. The estimates based on the Laplace approximation are shown in Table 3.2. The mean and variance are accurately estimated, but the variance is underestimated for these data. We implemented Gauss–Hermite rules using $m = 5, 10, 15, 20$ points, with the grid centered and scaled by the Laplace approximations of the mean and variance of the posterior. Table 3.2 shows that $\Pr(y)$ and $\mathrm{E}[\theta \mid y]$ are well estimated across all grid sizes, while there is more variability in the estimate of $\mathrm{var}(\theta \mid y)$, though it is more accurately estimated then with the Laplace approximation.

We now turn to importance sampling. We have

$$I_r = \int_{-\infty}^{\infty} f_r(\theta)\, d\theta = \mathrm{E}\left[\frac{f_r(\theta)}{g(\theta)} \right],$$

with $f_r(\theta) = \theta^r \Pr(y \mid \theta)\pi(\theta)$.

We take as proposal, $g(\cdot)$, a normal distribution scaled via the Laplace estimates of location and scale. Table 3.2 shows estimates resulting from the use of $m = 5{,}000$ points and the estimator

Table 3.2 Laplace, Gauss–Hermite, and Monte Carlo approximations for Poisson lognormal model with an observed count of $y =$ and an expected count of $E = 0.25$

	$\Pr(y)$ $(\times 10^2)$	$\mathrm{E}[\theta \mid y]$	$\mathrm{var}(\theta \mid y)$
Truth	1.37	2.27	0.329
Laplace	1.35	2.29	0.304
Gauss–Hermite $m = 5$	1.36	2.27	0.328
Gauss–Hermite $m = 10$	1.37	2.27	0.331
Gauss–Hermite $m = 15$	1.37	2.27	0.331
Gauss–Hermite $m = 20$	1.37	2.27	0.331
Importance sampling	1.37 [1.35,1.38]	2.27 [2.24,2.29]	0.336 [0.310,0.362]
Rejection algorithm	1.37	2.27 [2.25,2.28]	0.332 [0.319,0.346]
Metropolis–Hastings	–	2.27 [2.22,2.32]	0.328 [0.294,0.361]

The importance sampling and rejection algorithms are based on samples of size $m = 5{,}000$. The Metropolis–Hastings algorithm was run for 51,000 iterations, with the first 1,000 discarded as burn-in. 95% confidence intervals for the relevant estimates are displayed (where available) in square brackets in the last three lines of the table

$$\widehat{I}_r = \frac{1}{m} \sum_{t=1}^{m} \frac{f_r(\theta^{(t)})}{g(\theta^{(t)})}$$

where $\theta^{(t)}$ are independent samples from the normal proposal. The variance of the estimator is

$$\mathrm{var}\left(\widehat{I}_r\right) = \frac{\mathrm{var}(f_r/g)}{m}$$

The delta method can be used to produce measures of accuracy for the posterior mean and variance, though these measures are a little cumbersome. The variance of the normalizing constant is

$$\mathrm{var}\left[\widehat{\Pr}(y)\right] = \mathrm{var}(\widehat{I}_0).$$

To evaluate the variances of the posterior mean and posterior variance estimates we need the multivariate delta method. We must also include covariance terms if the same samples are used to evaluate all three integrals. The formulas are:

$$\begin{aligned}
\mathrm{var}\left[\widehat{\mathrm{E}}(\theta \mid y)\right] &= \mathrm{var}\left(\frac{\widehat{I}_1}{\widehat{I}_0}\right) \\
&\approx \frac{\mathrm{var}(\widehat{I}_1)}{\widehat{I}_0^2} + \frac{\widehat{I}_1^2 \mathrm{var}(\widehat{I}_0)}{\widehat{I}_0^4} - \frac{2\widehat{I}_1}{\widehat{I}_0^3}\,\mathrm{cov}(\widehat{I}_0, \widehat{I}_1) \\
\mathrm{var}\left[\widehat{\mathrm{var}}(\theta \mid y)\right] &= \mathrm{var}\left(\frac{\widehat{I}_2}{\widehat{I}_0} - \frac{\widehat{I}_1^2}{\widehat{I}_0^2}\right) \\
&\approx \left(-\frac{\widehat{I}_2}{\widehat{I}_0^2} + \frac{2\widehat{I}_1^2}{\widehat{I}_0^3}\right)^2 \mathrm{var}(\widehat{I}_0) + \left(\frac{-2\widehat{I}_1}{\widehat{I}_0^2}\right)^2 \mathrm{var}(\widehat{I}_1) + \left(\frac{1}{\widehat{I}_0}\right)^2 \mathrm{var}\left(\widehat{I}_2\right)
\end{aligned}$$

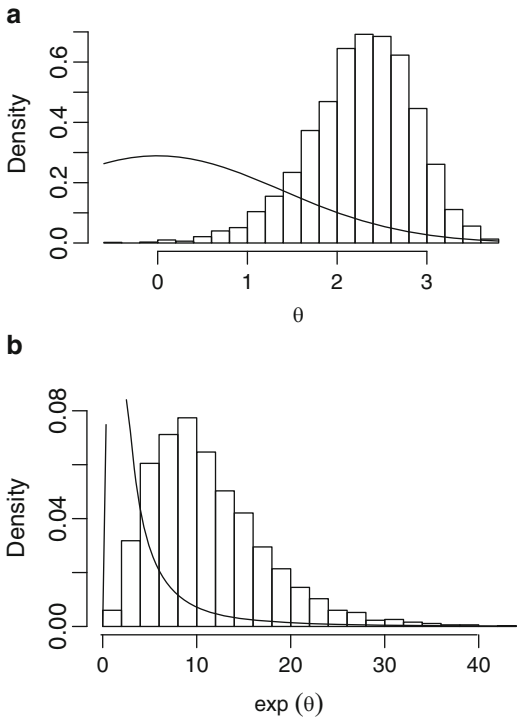

Fig. 3.3 Histogram representations of posterior distributions in the Sellafield example for (**a**) the log relative risk θ and (**b**) the relative risk $\exp(\theta)$, with priors superimposed as *solid lines*. The prior on θ is normal, so that the prior on $\exp(\theta)$ is lognormal

$$+ 2\left(\frac{-\widehat{I}_2}{\widehat{I}_0^2} + \frac{2\widehat{I}_1^2}{\widehat{I}_0^3}\right)\left(\frac{-2\widehat{I}_1}{\widehat{I}_0^2}\right)\mathrm{cov}\left(\widehat{I}_0, \widehat{I}_1\right)$$
$$+ 2\left(\frac{-\widehat{I}_2}{\widehat{I}_0^2} + \frac{2\widehat{I}_1^2}{\widehat{I}_0^3}\right)\left(\frac{1}{\widehat{I}_0}\right)\mathrm{cov}\left(\widehat{I}_0, \widehat{I}_2\right)$$
$$+ 2\left(\frac{-2\widehat{I}_1}{\widehat{I}_0^2}\right)\left(\frac{1}{\widehat{I}_0}\right)\mathrm{cov}\left(\widehat{I}_1, \widehat{I}_2\right).$$

Using these forms we obtain the interval estimates displayed in Table 3.2. The estimates of each of the three summaries are accurate though the interval estimate for the posterior variance is quite wide, because of the inherent instability associated with estimating the standard error.

Finally we implement a rejection algorithm, sampling from the prior distribution and estimating $\Pr(y)$ using the importance sampling estimator, (3.29). The mean and variance of the samples was used to evaluate $\mathrm{E}[\theta \mid y]$ and $\mathrm{var}(\theta \mid y)$, with the standard error of the latter based on (3.27). The acceptance probability was 0.07, the small value being explained by the discrepancy between the prior and the likelihood, which is illustrated in Fig. 3.3(a) which gives a histogram representation, based on 5000 points, of $p(\theta \mid y)$, along with the prior drawn as a solid curve.

Panel (b) displays the marginal posterior distribution of the relative risk, $p(\mathrm{e}^{\theta} \mid y)$, which is of more substantive interest, and is simply produced via exponentiation of the θ samples. The rejection estimates in Table 3.2 have relatively narrow interval estimates.

3.8 Markov Chain Monte Carlo

3.8.1 Markov Chains for Exploring Posterior Distributions

The fundamental idea behind MCMC is to construct a Markov chain over the parameter space, with invariant distribution the posterior distribution of interest. Specifically, consider a random variable $\boldsymbol{x}$ with support $\mathbb{R}^p$ and density $\pi(\cdot)$. We give a short summary of the essence of discrete time Markov chain theory.

A sequence of random variables $\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \ldots$ is called a Markov chain on a state space $\mathbb{R}^p$ if for all t and for all measurable sets A:

$$\Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t-1)}, \ldots, \boldsymbol{X}^{(0)}\right) = \Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)}\right)$$

so that the probability of moving to any set A at time $t+1$ only depends on where we are at time t. Furthermore, for a *homogeneous* Markov chain,

$$\Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)}\right) = \Pr\left(\boldsymbol{X}^{(1)} \in A \mid \boldsymbol{X}^{(0)}\right).$$

If there exists $p(\boldsymbol{x}, \boldsymbol{y})$ such that

$$\Pr(\boldsymbol{X}_1 \in A \mid \boldsymbol{x}) = \int_A p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y},$$

then $p(\boldsymbol{x}, \boldsymbol{y})$ is called the *transition kernel density*. A probability distribution $\pi(\cdot)$ on $\mathbb{R}^p$ is called an *invariant distribution* of a Markov chain with transition kernel density $p(\boldsymbol{x}, \boldsymbol{y})$ if so-called *global balance* holds:

$$\pi(\boldsymbol{y}) = \int_{\mathbb{R}^p} \pi(\boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x}.$$

A Markov chain is called *reversible* if

$$\pi(\boldsymbol{x}) p(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}) p(\boldsymbol{y}, \boldsymbol{x}) \tag{3.31}$$

for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, $\boldsymbol{x} \neq \boldsymbol{y}$. It can shown (Exercise 3.5) that if (3.31) holds, then $\pi(\cdot)$ is the invariant distribution which is useful since (3.31) can be easy to check.

A key idea is that if we have an invariant distribution, then we can evaluate long-term, or ergodic, averages from realizations of the chain. This is crucial for making inference in a Bayesian setting since it means we can estimate quantities of

interest such as posterior means and medians. In Markov chain theory, conditions on the transition kernel under which invariant distributions exist is an important topic. Within an MCMC context, this is not important since the posterior distribution is the invariant distribution and we are concerned with constructing Markov chains (transition kernels) with $\pi(\cdot)$ as invariant distribution. Only very mild conditions are typically required to ensure that $\pi(\cdot)$ is the invariant distribution, typically *aperiodocity* and *irreducibility*. A chain is *periodic* if there are places in the parameter space that can only be reached at certain regularly spaced times; otherwise, it is *aperiodic*. A Markov chain with invariant distribution $\pi(\cdot)$ is *irreducible* if for any starting point, there is positive probability of entering any set to which $\pi(\cdot)$ assigns positive probability.

Suppose that $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ represents the sample path of the Markov chain. Then expectations with respect to the invariant distribution

$$\mu = \mathrm{E}[g(\boldsymbol{x})] = \int g(\boldsymbol{x})\pi(\boldsymbol{x})\, d\boldsymbol{x}$$

may be approximated by $\widehat{\mu}_m = \frac{1}{m}\sum_{t=1}^{m} g(\boldsymbol{x}^{(t)})$. Monte Carlo standard errors are more difficult to obtain than in the independent sampling case. The Markov chain law of large numbers (the ergodic theorem) tells us that

$$\widehat{\mu}_m \to_{a.s.} \mu$$

as $m \to \infty$, and the Markov chain central limit theorem states that

$$\sqrt{m}(\widehat{\mu}_m - \mu) \to_d \mathrm{N}(0, \tau^2)$$

where

$$\tau^2 = \mathrm{var}\left[g(\boldsymbol{x}^{(t)})\right] + 2\sum_{k=1}^{\infty} \mathrm{cov}\left[g(\boldsymbol{x}^{(t)}), g(\boldsymbol{x}^{(t+k)})\right] \tag{3.32}$$

and the summation term accounts for the dependence in the chain. Chan and Geyer (1994) provide assumptions for validity of this form. Section 3.8.6 describes how τ^2 may be estimated in practice. We now describe algorithms that define Markov chains that are well suited to Bayesian computation.

3.8.2 The Metropolis–Hastings Algorithm

The Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970) provides a very flexible method for defining a Markov chain. At iteration t of the Markov chain's evolution, suppose the *current* point is $\boldsymbol{x}^{(t)}$. The following steps provide the new point $\boldsymbol{x}^{(t+1)}$:

1. Sample a point $\boldsymbol{y}$ from a *proposal* distribution $q(\cdot \mid \boldsymbol{x}^{(t)})$.
2. Calculate the acceptance probability:

$$\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = \min\left[\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}^{(\boldsymbol{t})})} \times \frac{q(\boldsymbol{x}^{(t)} \mid \boldsymbol{y})}{q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)})}, 1\right]. \tag{3.33}$$

3. Set

$$\boldsymbol{x}^{(t+1)} = \begin{cases} \boldsymbol{y} & \text{with probability } \alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) \\ \boldsymbol{x}^{(t)} & \text{otherwise.} \end{cases}$$

In a Bayesian context, the term $\pi(\boldsymbol{y})/\pi(\boldsymbol{x}^{(\boldsymbol{t})})$ in (3.33) is the ratio of the posterior densities at the proposed to the current point; since we are taking the ratio, the normalizing constant in the posterior cancels, which is crucial since this is typically unavailable. The second term in (3.33) is the ratio of the density of moving from $\boldsymbol{y} \to \boldsymbol{x}^{(t)}$ to the density of moving from $\boldsymbol{x}^{(t)} \to \boldsymbol{y}$, and it is this term that guarantees global balance and hence that the Markov chain has the correct invariant distribution; see Exercise 3.6. In an *independence* chain, the proposal distribution does not depend on the current point, that is, $q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)})$ is independent of $\boldsymbol{x}^{(t)}$. We now consider a special case of the algorithm that is particularly easy to implement and widely used.

3.8.3 The Metropolis Algorithm

Suppose the proposal distribution is *symmetric* in the sense that

$$g(\boldsymbol{y} \mid \boldsymbol{x}^{(t)}) = g(\boldsymbol{x}^{(t)} \mid \boldsymbol{y}).$$

In this case the product of ratios in (3.33) simplifies to

$$\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = \min\left[\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}^{(\boldsymbol{t})})}, 1\right]$$

so that only the ratio of target posterior densities is required. In the *random walk*, Metropolis algorithm $q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)}) = q(\,|\boldsymbol{y} - \boldsymbol{x}^{(t)}|\,)$, with common choices for $q(\cdot)$ being normal or uniform distributions. In a range of circumstances, an acceptance probability of around 30% is optimal (Roberts et al. 1997), which may be obtained by tuning the proposal density, the variance in a normal proposal, for example. The balancing act is between having high acceptance rates with small movement and having low acceptance rates with large movement.

3.8.4 The Gibbs Sampler

We describe a particularly popular algorithm for simulating from a Markov chain, the Gibbs sampler. We describe two flavors: the sequential Gibbs sampler and the random scan Gibbs sampler. In the following, let $\boldsymbol{x}_{-i}$ represent the vector $\boldsymbol{x}$ with the ith variable removed, that is, $\boldsymbol{x}_{-i} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p]$.

The *sequential scan Gibbs sampling* algorithm starts with some initial value $\boldsymbol{x}^{(0)}$ and then, with current point $\boldsymbol{x}^{(t)} = [x_1^{(t)}, \ldots, x_p^{(t)}]$, undertakes the following p steps to produce a new point $\boldsymbol{x}^{(t+1)} = [x_1^{(t+1)}, \ldots, x_p^{(t+1)}]$:

- Sample $x_1^{(t+1)} \sim \pi_1 \left(x_1 \mid \boldsymbol{x}_{-1}^{(t)} \right)$
- Sample $x_2^{(t+1)} \sim \pi_2 \left(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)} \right)$

$$\vdots$$

- Sample $x_p^{(t+1)} \sim \pi_p \left(x_p \mid \boldsymbol{x}_{-p}^{(t+1)} \right)$.

The beauty of the Gibbs sampler is that the often hard problem of sampling for the full p-dimensional variable $\boldsymbol{x}$ has been broken into sampling for each of the p variables in turn via the *conditional distributions*.

We now illustrate that the Gibbs sampling algorithm produces a transition kernel density that gives the required stationary distribution. We do this by showing that each component is a Metropolis–Hastings step. Consider a single component move in the Gibbs sampler from the current point $\boldsymbol{x}^{(t)}$ to the new point $\boldsymbol{x}^{(t+1)}$, with $\boldsymbol{x}^{(t+1)}$ obtained by replacing the ith component in $\boldsymbol{x}^{(t)}$ with a draw from the full conditional $\pi \left(x_i \mid \boldsymbol{x}_{-i}^{(t)} \right)$. We view this move in light of the Metropolis–Hastings algorithm in which the proposal density is the full conditional itself. Then the Metropolis–Hastings acceptance ratio becomes

$$\begin{aligned} \alpha(\boldsymbol{x}^{(t)}, \boldsymbol{x}^{(t+1)}) &= \min \left[\frac{\pi \left(x_i^{(t+1)}, \boldsymbol{x}_{-i}^{(t)} \right) \pi \left(x_i^{(t)} \mid \boldsymbol{x}_{-i}^{(t+1)} \right)}{\pi \left(x_i^{(t)}, \boldsymbol{x}_{-i}^{(t)} \right) \pi \left(x_i^{(t+1)} \mid \boldsymbol{x}_{-i}^{(t)} \right)}, 1 \right] \\ &= \min \left[\frac{\pi \left(\boldsymbol{x}_{-i}^{(t)} \right)}{\pi \left(\boldsymbol{x}_{-i}^{(t)} \right)}, 1 \right] = 1 \end{aligned}$$

because $\pi \left(\boldsymbol{x}_{-i}^{(t)} \right) = \pi \left(x_i^\star, \boldsymbol{x}_{-i}^{(t)} \right) / \pi \left(x_i^\star \mid \boldsymbol{x}_{-i}^{(t)} \right)$.

Consequently, when we use full conditionals as our proposals in the Metropolis–Hastings step, we always accept. This means that drawing from a full conditional distribution produces a Markov chain with stationary distribution $\pi(\boldsymbol{x})$. Clearly, we cannot keep updating only the ith component, because we will not be able to explore the whole state space this way, that is, we do not have an irreducible Markov chain. Therefore, we can update each component in turn, though this is not the only way to execute Gibbs sampling (though it is the easiest to implement and the most common approach). We can also randomly select an component to update. This is called *random scan* Gibbs sampling:

- Sample a component i by drawing a random variable with probability mass function $[\alpha_1, \ldots, \alpha_p]$ where $\alpha_i > 0$ and $\sum_{i=1}^p \alpha_i = 1$.
- Sample $x_i^{(t+1)} \sim \pi_i \left(x_i \mid \boldsymbol{x}_{-i}^{(t)} \right)$.

Roberts and Sahu (1997) examine the convergence rate of the sequential and random scan Gibbs sampling schemes and show that the sequential scan version has a better rate of convergence in the Gaussian models they examine.

In many cases, conjugacy (Sect. 3.7.1) can be exploited to derive the conditional distributions. Many examples of this are given in Chaps. 5 and 8. It is also common for sampling from a full conditional distribution to not require knowledge of the normalizing constant of the target distribution. For example, we saw in Sect. 3.7.7 that rejection sampling does not require the normalizing constant.

3.8.5 Combining Markov Kernels: Hybrid Schemes

Suppose we can construct m transition kernels, each with invariant distribution $\pi(\cdot)$. There are two simple ways to combine these transition kernels. First, we can construct a Markov chain, where at each step we sequentially generate new states from all kernels in a predetermined order. As long as the new Markov chain is irreducible, then it will have the required invariant distribution, and we can, for example, use the ergodic theorem on the samples from the new Markov chain. Hence, we can combine Gibbs and Metropolis–Hastings steps. One popular form is *Metropolis within Gibbs* in which all components with recognizable conditionals are sampled with Gibbs steps with Metropolis–Hastings for the remainder. In the second method of combining Markov kernels, we first create a probability vector $[\alpha_1, \ldots, \alpha_m]$, then randomly select kernel i with probability α_i, and then use this kernel to move the Markov chain.

In general, one can be creative in the construction of a Markov chain, but care must be taken to ensure the proposed chain is "legal," in the sense of having the required stationary distribution. As an example, a chain with a Metropolis step that keeps proposing points until the kth point, with $k \geq 1$, is accepted does not have the correct invariant distribution.

A final warning is that care is required to ensure that the posterior of interest is proper since there is no built in check when an MCMC scheme is implemented. For example, one may be able to construct a set of proper conditional distributions for Gibbs sampling, even when the joint posterior distribution is not proper; see, for example, Hobert and Casella (1996).

3.8.6 Implementation Details

Although theoretically not required, many users remove an initial number of iterations, the rationale being that inferential summaries should not be influenced by initial points that might be far from the main mass of the posterior distribution. Inference is then based on samples collected subsequent to this "burn-in" period.

In order to obtain valid Monte Carlo standard errors for empirical averages, some estimate for τ^2 in (3.32) is required. Time series methods exist to estimate τ^2, but we describe a simple approach based on batch means (Glynn and Iglehart 1990). The basic idea is to split the output of length m into K batches each of length B, with B chosen to be large enough so that the batch means have low serial correlation; B should not be too large, however, because we want K to be large enough to provide a reliable estimate of τ^2. The mean of the function of interest is then estimated within each of the batches:

$$\widehat{\mu}_k = \frac{1}{B} \sum_{t=(k-1)B+1}^{KB} g(\boldsymbol{x}^{(t)})$$

for $k = 1, \ldots, K$. The combined estimate of the mean is the average of the batch means

$$\widehat{\mu} = \frac{1}{K} \sum_{k=1}^{K} \widehat{\mu}_k.$$

Then $\sqrt{B}(\widehat{\mu}_k - \mu)$, $k = 1, \ldots, K$ are approximately independently distributed as $\mathrm{N}(0, \tau^2)$, and so τ^2 can be estimated by

$$\widehat{\tau}^2 = \frac{B}{K-1} \sum_{k=1}^{K} (\widehat{\mu}_k - \widehat{\mu})^2$$

and

$$\widehat{\mathrm{var}}(\widehat{\mu}) = \frac{\widehat{\tau}^2}{K} = \frac{B}{K(K-1)} \sum_{k=1}^{K} (\widehat{\mu}_k - \widehat{\mu})^2.$$

Normal or Students t confidence intervals can be calculated based on the square root of this quantity. The construction of these intervals has the advantage of being simple, but the output should be viewed with caution as the above derivation contains a number of approximations.

MCMC approaches provide no obvious estimator of the normalizing constant $p(\boldsymbol{y})$, but a number of indirect methods have been proposed (Meng and Wong 1996; DiCiccio et al. 1997)

Aside from directly calculating integrals, we may also form graphical summaries of parameters of interest, essentially using the dependent samples in the same way that we would independent samples. For example, a histogram of $x_i^{(t)}$ provides an estimate of the posterior marginal distribution, $\pi_i(x_i)$, $i = 1, \ldots, p$.

In practice, there are a number of important issues that require thought when implementing MCMC. A crucial question is how large m should be in order to obtain a reliable Monte Carlo estimate. The Markov chain will display better mixing properties if the parameters are approximately independent in the posterior. In an extreme case, if we have independence, then

$$\pi(x_1, \ldots, x_p) = \prod_{i=1}^{p} \pi(x_i)$$

and Gibbs sampling via the conditional distributions $\pi(x_i), i = 1, \ldots, p$, equates to direct sampling from the posterior.

Dependence in the Markov chain may be greatly reduced by sampling simultaneously for variables that are highly depend, a strategy known as *blocking*. Reparameterization may also be helpful in this regard. As the blocks become larger, the acceptance rate (if a Metropolis-Hastings algorithm is used) may be reduced to an unacceptably low level in which case there is a trade-off with respect to the size of blocks to use. Some chains may be very *slow mixing*, and an examination of autocorrelation aids in deciding on the number of iterations required. If storage of samples is an issue, then one may decide to "thin" the chain by only collecting samples at equally spaced intervals.

A number of methods have been proposed for "diagnosing convergence." Trace plots provide a useful method for detecting problems with MCMC convergence and mixing. Ideally, trace plots of unnormalized log posterior and model parameters should look like stationary time series. Slowly mixing Markov chains produce trace plots with high autocorrelation, which can be further visualized by plotting the autocorrelation at different lags. Slow mixing does not imply lack of convergence, however, but that more samples will be required for accurate inference (as can be seen from (3.32)). When examining trace plots and autocorrelations, it is clearer to work with parameters transformed to $\mathbb{R}$. Running multiple chains from different starting points is also very useful since one may compare inference between the different chains. Gelman and Rubin (1992) provide one popular convergence diagnostic based on multiple chains. As with the use of diagnostics in regression modeling, convergence diagnostics may detect evidence of poor behavior, but there is no guarantee of good behavior of the chain, even if all convergence diagnostics appear reasonable.

Example: Poisson Likelihood, Lognormal Prior

Recall the Poisson lognormal example in which $y = 4$ and $E = 0.25$ with a single parameter, the log relative risk θ. Gibbs sampling corresponds to direct sampling from the univariate posterior for θ, which we have already illustrated using the rejection algorithm.

We implement a random walk Metropolis algorithm using a normal kernel and the asymptotic variance of the MLE for θ multiplied by 3 as the variance of the proposal, to achieve a reasonable acceptance probability of 0.32. This multiplier was found by trial and error, based on preliminary runs of the Markov chain. It is important to restart the chain when the proposal is changed based on past realizations to ensure the chain is still Markovian. Table 3.2 gives estimates of the

posterior mean and variance based on a run length of 51,000, with the first 1,000 discarded as a burn-in. The confidence interval for the estimates of the posterior mean and posterior variance is based on the batch means method, with $K = 50$ batches of size $B = 1{,}000$.

Example: Lung Cancer and Radon

We return to the lung cancer and radon example, first introduced in Sect. 1.3.3, to demonstrate the use of the Metropolis random walk algorithm in a situation with more than one parameter. For direct comparison with methods applied in Chap. 2, we assume an improper flat prior on $\boldsymbol{\beta} = [\beta_0, \beta_1]$ so that the posterior $p(\boldsymbol{\beta} \mid \boldsymbol{y})$ is proportional to the likelihood.

We begin by implementing a Metropolis random walk algorithm based on a pair of univariate normal distributions. In this example, the Gibbs sampler is less appealing since the required conditional distributions do not assume known forms. The first step is to initialize $\beta_0^{(0)} = \widehat{\beta}_j$, where $\widehat{\beta}_j$, $j = 0, 1$, are the MLEs. We then iterate, at iteration t, between:

1. Generate $\beta_0^\star \sim \mathrm{N}(\beta_0^{(t)}, c_0 \widehat{V}_0)$, where $\widehat{V}_0$ is the asymptotic variance of $\widehat{\beta}_0$. Calculate the acceptance probability:

$$\alpha_0(\beta_0^\star, \beta_0^{(t)}) = \min\left[\frac{p(\beta_0^\star, \beta_1^{(t)} \mid \boldsymbol{y})}{p(\beta_0^{(t)}, \beta_1^{(t)} \mid \boldsymbol{y})}, 1\right]$$

   and set

$$\beta_0^{(t+1)} = \begin{cases} \beta_0^\star & \text{with probability } \alpha_0(\beta_0^\star, \beta_0^{(t)}), \\ \beta_0^{(t)} & \text{otherwise.} \end{cases}$$

2. Generate $\beta_1^\star \sim \mathrm{N}(\beta_1^{(t)}, c_1 \widehat{V}_1)$, where $\widehat{V}_1$ is the asymptotic variance of $\widehat{\beta}_1$. Calculate the acceptance probability:

$$\alpha_1(\beta_1^\star, \beta_1^{(t)}) = \min\left[\frac{p(\beta_0^{(t+1)}, \beta_1^\star \mid \boldsymbol{y})}{p(\beta_0^{(t+1)}, \beta_1^{(t)} \mid \boldsymbol{y})}, 1\right]$$

   and set

$$\beta_1^{(t+1)} = \begin{cases} \beta_1^\star & \text{with probability } \alpha_1(\beta_1^\star, \beta_1^{(t)}), \\ \beta_1^{(t)} & \text{otherwise.} \end{cases}$$

The constants c_0 and c_1 are chosen to provide a trade-off between gaining a high proportion of acceptances and moving around the support of the parameter space; this is illustrated in Fig. 3.4 where the realized parameters from the first 1,000 iterations of two Markov chains are plotted. In panels (a) and (d), we chose

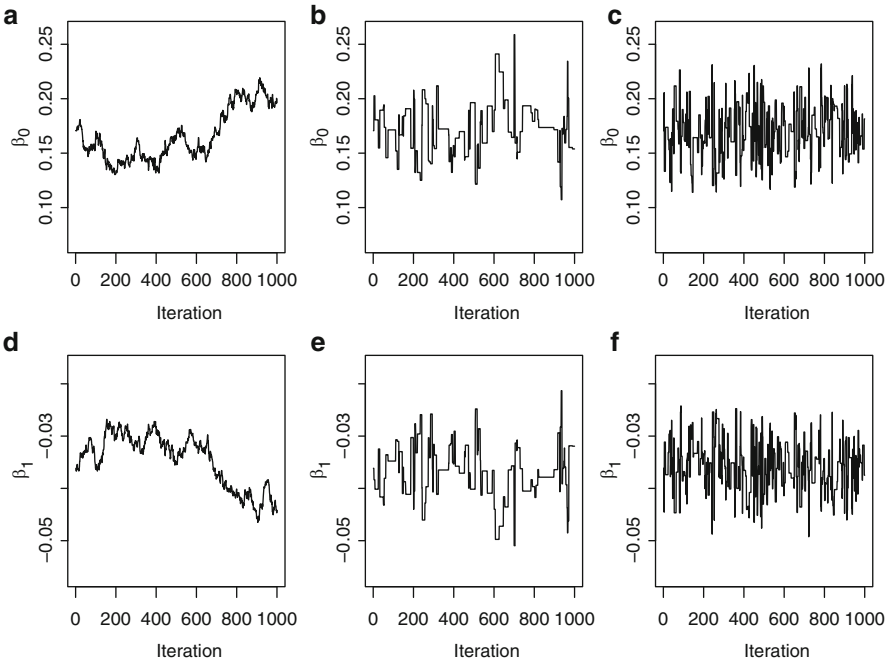

Fig. 3.4 Sample paths from Metropolis–Hastings algorithms for β_0 (*top row*) and β_1 (*bottom row*) for the lung cancer and radon data. In the *left column* the proposal random walk has small variance; in the *center column* large variance and in the *right column*, we use a bivariate proposal

$c_0 = c_1 = c = 0.1$ and in panels (b) and (e) $c_0 = c_1 = c = 2$. For $c = 0.1$ the acceptance rate is 0.90, but movement around the space is slow, as indicated by the meandering nature of the chain, while for $c = 2$ the moves tend to be larger, but the chain sticks at certain values, as seen by the horizontal runs of points (the acceptance rate is 0.14).

Figure 3.6a shows a scatterplot representation of the joint distribution $p(\beta_0, \beta_1 \,|\, \boldsymbol{y})$ and clearly shows the strong negative dependence; the asymptotic correlation between the MLEs $\widehat{\beta}_0$ and $\widehat{\beta}_1$ is -0.90, and the posterior correlation between β_0 and β_1 is -0.90 also (the correspondence between these correlations is not surprising since the sample size is large and the prior is flat). The strong negative dependence is evident in each of the first two columns of Fig. 3.4. Figure 3.5 shows the autocorrelations between sampled parameters at lags of between 1 and 40. The top row is for β_0, and the bottom is for β_1. In panels (a) and (d), the autocorrelations are high because of the small movements of the chain.

The dependence in the chain may be reduced via reparameterization or by generation from a bivariate proposal. We implement the latter with variance–covariance matrix equal to $c \times \text{var}(\widehat{\boldsymbol{\beta}})$. The acceptance rate for the bivariate proposal with $c = 2$ is 0.29, which is reasonable. We then iterate the following:

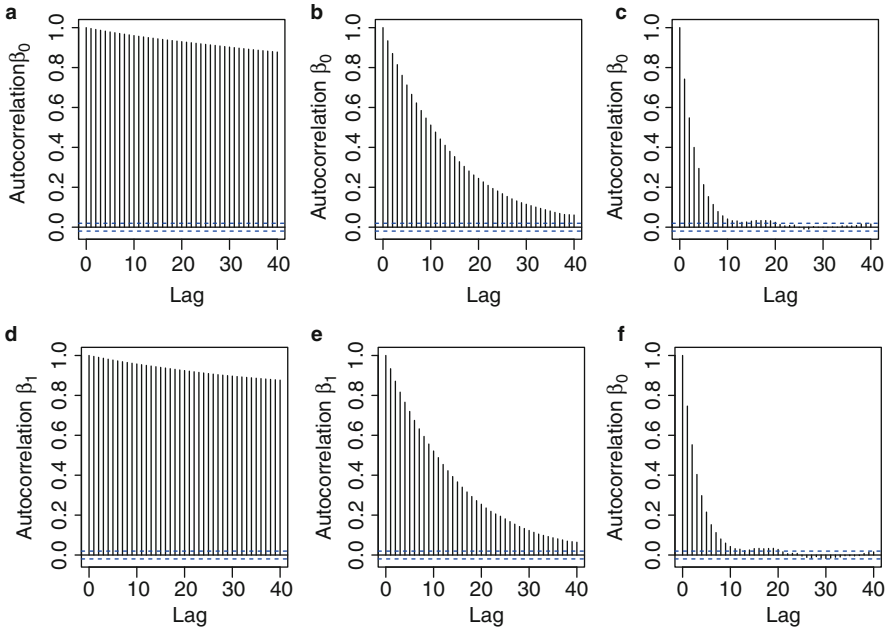

Fig. 3.5 Autocorrelation functions for β_0 (*top row*) and β_1 (*bottom row*) for the lung cancer and radon data. *First column*: univariate random walk, $c = 0.1$, *second column*: univariate random walk, $c = 2$, *third column*: bivariate random walk, $c = 2$

1. Generate $\boldsymbol{\beta}^\star \sim \mathrm{N}_2(\boldsymbol{\beta}^{(t)}, c\widehat{\boldsymbol{V}})$, where $\widehat{\boldsymbol{V}}$ is the asymptotic variance of the MLE $\widehat{\boldsymbol{\beta}}$.
2. Calculate the acceptance probability

$$\alpha(\boldsymbol{\beta}^\star, \boldsymbol{\beta}^{(t)}) = \min\left[\frac{p(\boldsymbol{\beta}^\star \mid \boldsymbol{y})}{p(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{y})}, 1\right]$$

and set

$$\boldsymbol{\beta}^{(t+1)} = \begin{cases} \boldsymbol{\beta}^\star & \text{with probability } \alpha(\boldsymbol{\beta}^\star, \boldsymbol{\beta}^{(t)}), \\ \boldsymbol{\beta}^{(t)} & \text{otherwise.} \end{cases}$$

Note that the choice of c and the dependence in the chain do not jeopardize the invariant distribution, but rather the length of chain until practical convergence is reached and the number of points required for summarization. More points are required when there is high positive dependence in successive iterates, which is clear from (3.32). The final column of Fig. 3.4 shows the sample path from the bivariate proposal, with good movement and little dependence between the parameters. Panels (c) and (f) show that the autocorrelation is also greatly reduced.

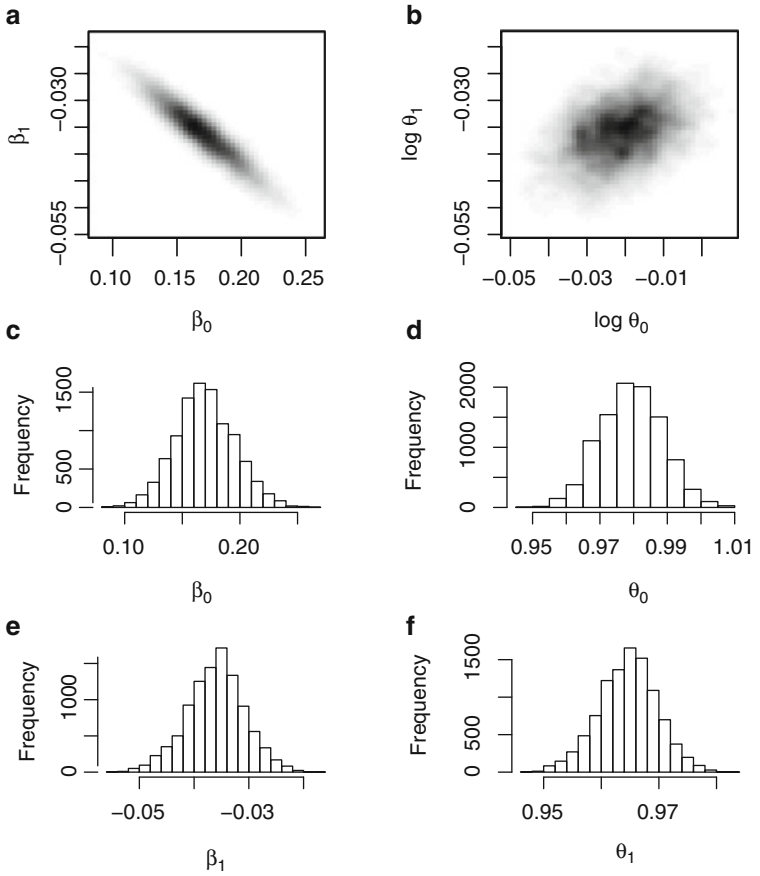

Fig. 3.6 Posterior summaries for the lung cancer and radon data: (**a**) $p(\beta_0, \beta_1 \mid \boldsymbol{y})$, (**b**) $p(\log\ \theta_0, \log\ \theta_1 \mid \boldsymbol{y})$, (**c**) $p(\beta_0 \mid \boldsymbol{y})$, (**d**) $p(\theta_0 \mid \boldsymbol{y})$, (**e**) $p(\beta_1 \mid \boldsymbol{y})$, (**f**) $p(\theta_1 \mid \boldsymbol{y})$

Figure 3.6 shows inference for the reparameterized model

$$Y_i \mid \boldsymbol{\theta} \sim_{ind} \text{Poisson}(E_i \theta_0 \theta_1^{x_i - \overline{x}})$$

where $\theta_0 = \exp(\beta_0 + \beta_1 \overline{x}) > 0$ and $\theta_1 = \exp(\beta_1) > 0$ along with summaries for the β_0, β_1 parameterization. Figure 3.6(b) shows the bivariate posterior for $\log \theta_0, \log \theta_1$ and demonstrates that the parameters are virtually independent (the correlation is -0.03). By comparison there is strong negative dependence between β_0 and β_1 (panel (a)). Panels (d) and (f) show histogram representations of the posteriors of interest $p(\theta_0 \mid \boldsymbol{y})$ and $p(\theta_1 \mid \boldsymbol{y})$.

The posterior median (95% credible interval) for $\exp(\beta_1)$ is $0.965\ [0.954, 0.975]$ which is almost identical to the asymptotic inference under a Poisson model (Table 2.4), which is again not surprising given the large sample size.

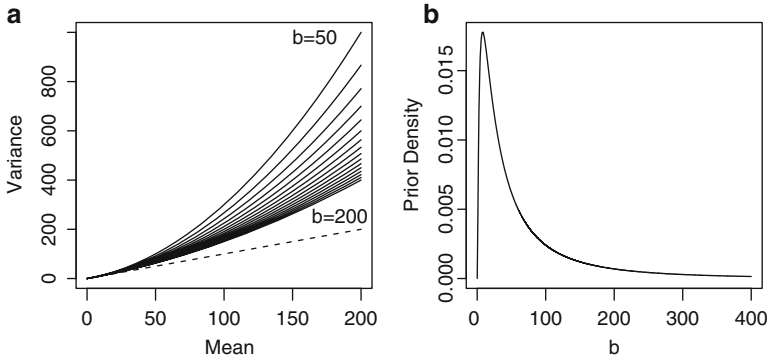

Fig. 3.7 (**a**) Mean–variance relationships, in the negative binomial model, for values of b between 50 and 200, in increments of 10 units. The *dashed line* is the line of equality corresponding to the Poisson model, which is recovered as $b \to \infty$. (**b**) Lognormal prior for b

The Poisson model should be used with caution since the variance is determined by the mean, with no additional parameter to soak up excess-Poisson variability, which is often present in practice. To overcome this shortcoming we provide a Bayesian analysis with a negative binomial likelihood, parameterized so that

$$\mathrm{E}[Y_i \mid \boldsymbol{\beta}, b] = \mu_i(\boldsymbol{\beta}), \quad \mathrm{var}(Y_i \mid \boldsymbol{\beta}, b) = \mu_i(\boldsymbol{\beta})[1 + \mu(\boldsymbol{\beta})/b]. \tag{3.34}$$

We will continue with an improper flat prior for $\boldsymbol{\beta}$, but a prior for b requires more thought. To determine a prior, we plot the mean–variance relationship in Fig. 3.7a, for different values of b. In this example the regression model does not include information on confounders such as smoking. The absence of these variables will certainly lead to bias in the estimate of $\exp(\beta_1)$ due to confounding, but with respect to b, we might expect considerable excess-Poisson variability due to missing variables. The sample average of the observed counts is 158, and we specify a lognormal prior for b by giving two quantiles of the overdispersion, $\mu(1 + \mu/b)$, at $\mu = 158$, and then solve for b. Specifically, we suppose that there is a 50% chance that the overdispersion is less than $1.5 \times \mu$ and a 95% chance that it is less than $5 \times \mu$. Formulas (3.15) and (3.16) give a lognormal prior with parameters 3.68 and 1.26^2 and 5%, 50%, and 95% quantiles of 4.9, 40, and 316, respectively. Figure 3.7(b) gives the resulting lognormal prior density.

A random walk Metropolis algorithm with a normal proposal was constructed for β_0, β_1, b with the variance–covariance matrix taken as 3 times the asymptotic variance–covariance matrix ($\widehat{b}$ is asymptotically independent of $\widehat{\beta}_0$ and $\widehat{\beta}_1$), based on the expected information. The posterior median and 95% credible interval for $\exp(\beta_1)$ are 0.970 [0.955,0.987], and for b the summaries are 57.8 [34.9,105]. The MLE is $\widehat{b} = 61.3$, with asymptotic 95% confidence interval (calculated on the $\log b$ scale and then exponentiated) of [35.4,106]. Therefore, likelihood and Bayesian inference for b are in close agreement for these data. Histograms of samples from

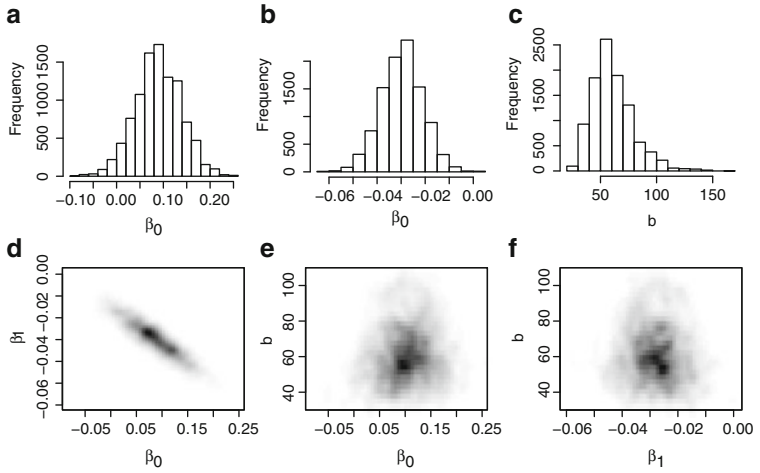

Fig. 3.8 Univariate and bivariate summaries of the posterior $p(\beta_0, \beta_1, b \mid \boldsymbol{y})$, arising from the negative binomial model

the univariate posteriors for β_0, β_1, and b are shown in the first row of Fig. 3.8, while bivariate scatterplots are shown in the second row. The posterior marginals for β_0 and β_1 are very symmetric, while that for b is slightly skewed.

3.8.7 Implementation Summary

While MCMC has revolutionized Bayesian inference in terms of the breadth of applications and complexity of models that can now be considered, other methods may still be preferable in some situations, in particular when the number of parameters is small. Direct sampling from the posterior is particularly appealing since one retains all of the advantages of sample-based inference (e.g., the ability to simply examine generic functions of interest), without the need to worry about the convergence issues associated with MCMC. Quadrature methods are also appealing for low-dimensional problems, since they are highly efficient. The latter is particularly important if the calculation of the likelihood is expensive. Importance sampling Monte Carlo methods are appealing in that error assessment may be carried out; analytical approximations are, in general, poor in this respect.

INLA is very attractive due to its speed of computation, though a reliable measure of accuracy is desirable and there are practical situations in which the method is not accurate. For example, the method is less accurate for binomial data with small denominators (Fong et al. 2010). In exploratory situations, one may always use quick methods such as INLA for initial modeling, with more computationally demanding approaches being used when a set of finals models are honed in upon.

INLA is also useful for performing simulation studies to examine the properties of model summaries. In general, comparing results across different methods is a good idea. When deciding upon a method of implementation, there is often a clear trade-off between efficiency and the time taken to code prospective methods. MCMC methods are often easy to implement, but are not always the most efficient (at least not for basic schemes) and are difficult to automate. For many high-dimensional problems, MCMC may be the only method that is feasible, although INLA may be available if the model is of the required form (a small number of "non-Gaussian" parameters).

An important paper in the history of MCMC is that of Green (1995) in which *reversible jump MCMC* was introduced. This method can be used in situations in which the parameter space is of varying dimension across different models.

3.9 Exchangeability

We now provide a brief discussion of de Finetti's celebrated representation theorem which describes the form of the marginal distribution of a collection of random variables, under certain assumptions. As we will see, this provides one way in which important modeling questions can be framed. We first require the introduction of a very important concept in Bayesian inference, *exchangeability*.

Definition. Let $p(y_1, \ldots, y_n)$ be the joint density of $Y_1, \ldots, Y_n$. If

$$p(y_1, \ldots, y_n) = p(y_{\pi(1)}, \ldots, y_{\pi(n)})$$

for all permutations, π, of $\{1, 2, \ldots, n\}$, then $Y_1, \ldots, Y_n$ are (finitely) *exchangeable*.

This definition essentially says that the labels identifying the individual components are uninformative. Obviously if a collection of n random variables is exchangeable, this implies that the marginal distribution of all single random variables are the same, as are the marginal distributions for all pairs, all triples, etc. A collection of random variables is infinitely exchangeable if every finite subcollection is exchangeable.

As a simple example, consider Bernoulli random variables, Y_i, for $i = 1, 2, 3 = n$. Under exchangeability,

$$\begin{aligned}
\Pr(Y_1 = a, Y_2 = b, Y_3 = c) &= \Pr(Y_1 = a, Y_2 = c, Y_3 = b) \\
&= \Pr(Y_1 = b, Y_2 = a, Y_3 = c) \\
&= \Pr(Y_1 = b, Y_2 = c, Y_3 = a) \\
&= \Pr(Y_1 = c, Y_2 = a, Y_3 = b) \\
&= \Pr(Y_1 = c, Y_2 = b, Y_3 = a)
\end{aligned}$$

for all $a, b, c = 0, 1$.

Result. If $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ and $Y_1, \ldots, Y_n$ are conditionally independent and identically distributed given $\boldsymbol{\theta}$, then $Y_1, \ldots, Y_n$ are exchangeable.

Proof. By definition:

$$\begin{aligned} p(y_1, \ldots, y_n) &= \int p(y_1, \ldots, y_n \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\ &= \int \left[\prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}) \right] \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\ &= \int \left[\prod_{i=1}^{n} p(y_{\pi(i)} \mid \boldsymbol{\theta}) \right] \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \\ &= p(y_{\pi(1)}, \ldots, y_{\pi(n)}) \end{aligned}$$

We now present the converse of this result.

Theorem. *de Finetti's representation theorem for 0/1 random variables.*

If $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of 0/1 random variables, there exists a distribution $\pi(\cdot)$ such that the joint mass function $\Pr(y_1, \ldots, y_n)$ has the form

$$\Pr(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \pi(\theta) \, d\theta,$$

where

$$\int_0^\theta \pi(u) \, du = \lim_{n \to \infty} \Pr\left(\frac{Z_n}{n} \leq \theta \right),$$

with $Z_n = Y_1 + \ldots + Y_n$, and $\theta = \lim_{n \to \infty} Z_n / n$.

Proof. The following is based on Bernardo and Smith (1994). Let $z_n = y_1 + \ldots + y_n$ be the number of 1's (which we label "successes") in the first n observations. Then, due to exchangeability,

$$\Pr(y_1 + \ldots + y_n = z_n) = \binom{n}{z_n} \Pr(Y_{\pi(1)}, \ldots, Y_{\pi(n)}),$$

for all permutations π of $\{1, 2, \ldots, n\}$ such that $y_{\pi(1)} + \ldots + y_{\pi(n)} = z_n$. We can embed the event $y_1 + \ldots + y_n = z_n$ within a longer sequence and

$$\Pr(Y_1 + \ldots + Y_n = z_n) = \sum_{Z_N = z_n}^{N-(n-z_n)} \Pr(z_n, z_N) = \sum_{Z_N = z_n}^{N-(n-z_n)} \Pr(z_n \mid z_N) \Pr(z_N),$$

where $\Pr(z_N)$ is the "prior" belief in the number of successes out of N. To obtain the conditional probability, we observe that it is "as if" we have a population of N

items of which z_N are successes and $N - z_N$ failures, from which we draw n items. The distribution of $z_n \mid z_N$ successes is therefore hypergeometric so that

$$\Pr(y_1 + \ldots + y_n = z_n) = \sum_{z_N = z_n}^{N-(n-z_n)} \frac{\binom{z_N}{z_n}\binom{N - z_N}{n - z_n}}{\binom{N}{n}} \Pr(z_N).$$

We now let $\Pi(\theta)$ be the step function which is 0 for $\theta < 0$ and has jumps of $\Pr(z_N)$ at $\theta = z_N/N$, $z_N = 0, \ldots, N$. We now let $N \to \infty$. Then the hypergeometric distribution tends to a binomial distribution with parameters n and θ and the prior $\Pr(z_N)$ is translated into a prior for θ, which we write as $\pi(\theta)$. Consequently,

$$\Pr(y_1 + \ldots + y_n = z_n) \to \binom{n}{z_n} \int \theta^{z_n}(1-\theta)^{n-z_n}\pi(\theta)\, d\theta,$$

as $N \to \infty$. □

The implications of this theorem are of great significance. By the strong law of large numbers, $\theta = \lim_{n\to\infty} Z_n/n$, so that $\pi(\cdot)$ represents our beliefs about the limiting relative frequency of 1's. Hence, we have an interpretation of θ. Further, we may view the Y_i as conditional independent, Bernoulli random variables, conditional on the random variable θ.

In conventional language, we have a *likelihood function*

$$\Pr(y_1, \ldots, y_n \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i},$$

where the *parameter* θ is assigned a *prior distribution* $\pi(\theta)$.

In general, if $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of random variables, there exists a probability density function $\pi(\cdot)$ such that

$$p(y_1, \ldots, y_n) = \int \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta})\pi(\theta)\, d\boldsymbol{\theta}, \tag{3.35}$$

with $p(Y \mid \boldsymbol{\theta})$ denoting the density function corresponding to the "unknown parameter" $\boldsymbol{\theta}$. A sketch proof of (3.35) may be found in Bernardo and Smith (1994). This result tells us that a conditional independence model can be justified via an exchangeability argument. In this general case, further assumptions on $Y_1, Y_2, \ldots$ are required to identify $p(Y \mid \boldsymbol{\theta})$. Bernardo and Smith (1994) present the assumptions that lead to a number of common modeling choices. For example, suppose that $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of random variables such that $Y_i > 0$, $i = 1, 2, \ldots$. Further, suppose that for any event A in $\mathbb{R} \times \ldots \times \mathbb{R}$, and for all n,

$$\Pr[(y_1, \ldots, y_n) \in A] = \Pr[(y_1, \ldots, y_n) \in A + \boldsymbol{a}]$$

for all $\boldsymbol{a} \in \mathbb{R} \times \ldots \times \mathbb{R}$ such that $\boldsymbol{a}^{\mathrm{T}}\mathbf{1} = 0$ and $A + \boldsymbol{a}$ is an event in $\mathbb{R} \times \ldots \times \mathbb{R}$. Then the joint density for $y_1, \ldots, y_n$ is

$$p(y_1, \ldots, y_n) = \int_0^\infty \prod_{i=1}^{n} \theta^{-1} \exp(-\theta^{-1} y_i) \times \pi(\theta)\, d\theta$$

where $\int_0^\infty \pi(u)\, du = \lim_{n\to\infty} \Pr(\overline{y}_n \leq \theta)$ and $\overline{y}_n = (y_1 + \ldots + y_n)/n$. For a proof, see Diaconis and Ylvisaker (1980). Hence, a belief in exchangeability and a "lack of memory" property leads to the integral of the predictive distribution being the marginal distribution that is constructed from the product of a conditionally independent set of *exponential* random variables and a prior. The parameter is identified as the sample mean from a large number of observations.

This kind of approach is of theoretical interest, but in practice the choice of likelihood will often be based more directly on the context and previous experience with similar data types. Exchangeability is very useful in practice for prior specification, however. Before one uses a particular conditional independence model, one can think about whether all units are deemed exchangeable. If some collection of units are distinguishable, then one should not assume conditional independence for all units, and one may instead separate the units into groups within which exchangeability holds. For further discussion, see Sect. 8.6.

In terms of modeling, if we believe that a sequence of random variables is exchangeable, this allows us to write down a conditional independence model. We emphasize that independence is a very different assumption since it implies that we learn nothing from past observations:

$$p(y_{m+1}, \ldots, y_n \mid y_1, \ldots, y_m) = p(y_{m+1}, \ldots, y_n)$$

In a regression context, the situation is slightly more complicated. Informally, exchangeability within covariate-defined groups gives the usual conditional independence model, where we now condition on parameters and covariates; Bernardo and Smith (1994, Sect. 4.64) contains details.

3.10 Hypothesis Testing with Bayes Factors

We now turn to a description of Bayes factors, which are the conventional Bayesian method for comparison of hypotheses/models. Let the observed data be denoted $\boldsymbol{y} = [y_1, \ldots, y_n]$, and assume two hypotheses of interest, H_0 and H_1. The application of Bayes theorem gives the probability of the hypothesis H_0, given data $\boldsymbol{y}$, as

$$\Pr(H_0 \mid \boldsymbol{y}, H_0 \cup H_1) = \frac{p(\boldsymbol{y} \mid H_0) \Pr(H_0 \mid H_0 \cup H_1)}{p(\boldsymbol{y} \mid H_0 \cup H_1)}$$

Table 3.3 Losses corresponding to the decision δ, when the truth is H and L_{I} and L_{II} are the losses associated with type I and II errors, respectively

		Decision	
$L(\delta, H)$		$\delta = 0$	$\delta = 1$
Truth H	H_0	0	L_{I}
	H_1	L_{II}	0

where

$$p(\boldsymbol{y} \mid H_0 \cup H_1) = p(\boldsymbol{y} \mid H_0) \Pr(H_0 \mid H_0 \cup H_1) + p(\boldsymbol{y} \mid H_1) \Pr(H_1 \mid H_0 \cup H_1)$$

is the probability of the data averaged over H_0 and H_1. The prior probability that H_0 is true, given one of H_0 and H_1 is true, is $\Pr(H_0 \mid H_0 \cup H_1)$, and $\Pr(H_1 \mid H_0 \cup H_1) = 1 - \Pr(H_0 \mid H_0 \cup H_1)$ is the prior on the alternative hypothesis. This simple calculation makes it clear that to evaluate the probability that the null is true, one is actually calculating the probability of the null *given* that H_0 or H_1 is true. Therefore, we are calculating the "relative truth"; H_0 may provide a poor fit to the data, but H_1 may be even worse. Although conditioning on $H_0 \cup H_1$ is crucial to interpretation, we will drop it for compactness of notation.

If we wish to compare models H_0 and H_1, then a natural measure is given by the *posterior odds*

$$\frac{\Pr(H_0 \mid \boldsymbol{y})}{\Pr(H_1 \mid \boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid H_0)}{p(\boldsymbol{y} \mid H_1)} \times \frac{\Pr(H_0)}{\Pr(H_1)}, \tag{3.36}$$

where the *Bayes factor*

$$\text{BF} = \frac{p(\boldsymbol{y} \mid H_0)}{p(\boldsymbol{y} \mid H_1)}$$

is the ratio of the marginal distributions of the data under the two models, and $\Pr(H_0)/\Pr(H_1)$ is the *prior odds*. Care is required in the choice of priors when Bayes factors are calculated; see Sect. 4.3.2 for further discussion.

Depending on the nature of the analysis, we may: simply report the Bayes factor; or we may place priors on the hypotheses and calculate the posterior odds of H_0; or we may go a step further and derive a decision rule. Suppose we pursue the latter and let $\delta = 0/1$ represent the decision to pick H_0/H_1. With respect to Table 3.3, the posterior expected loss associated with decision δ is

$$\text{E}[L(\delta, H)] = L(\delta, H_0) \Pr(H_0 \mid \boldsymbol{y}) + L(\delta, H_1) \Pr(H_1 \mid \boldsymbol{y})$$

so that for the two possible decisions (accept/reject H_0) the expected losses are

$$\begin{aligned}\mathrm{E}[L(\delta = 0, H)] &= 0 \times \Pr(H_0 \mid \boldsymbol{y}) + L_{\text{II}} \times \Pr(H_1 \mid \boldsymbol{y}) \\ \mathrm{E}[L(\delta = 1, H)] &= L_{\text{I}} \times \Pr(H_0 \mid \boldsymbol{y}) + 0 \times \Pr(H_1 \mid \boldsymbol{y}).\end{aligned}$$

To find the decision that minimizes posterior expected cost, let $v = \Pr(H_1 \mid \boldsymbol{y})$ so that

$$\mathrm{E}[L(\delta = 0, H)] = L_{\text{II}} \times v \tag{3.37}$$

$$\mathrm{E}[L(\delta = 1, H)] = L_{\text{I}} \times (1 - v). \tag{3.38}$$

We should choose $\delta = 1$ if $L_{\text{II}} \times v \geq L_{\text{I}}(1 - v)$, that is, if $v/(1 - v) \geq L_{\text{I}}/L_{\text{II}}$, or $v \geq L_{\text{I}}/(L_{\text{I}} + L_{\text{II}})$. Hence, we report H_1 if

$$\Pr(H_1 \mid \boldsymbol{y}) \geq \frac{L_{\text{I}}}{L_{\text{I}} + L_{\text{II}}} = \frac{1}{1 + L_{\text{II}}/L_{\text{I}}},$$

illustrating that we only need to specify the ratio of losses. If incorrect decisions are equally costly, we should therefore report the hypothesis that has the greatest posterior probability, in line with intuition. These calculations can clearly be extended to three or more hypotheses. The models that represent each hypothesis need not be nested as with likelihood ratio tests, though careful prior choice is required so as to not inadvertently favor one model over another. One remedy to this difficulty is described in Sect. 6.16.3.

To evaluate the Bayes factor, we need to calculate the normalizing constants under H_0 and H_1. A generic normalizing constant is

$$I = p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \, d\boldsymbol{\theta}. \tag{3.39}$$

We next derive a popular approximation to the Bayes factor. The integral (3.39) is an integral of the form (3.18) with

$$nh(\boldsymbol{\theta}) = \log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}).$$

Letting $\widetilde{\boldsymbol{\theta}}$ denote the posterior mode, we may apply (3.20) with $nh(\widetilde{\boldsymbol{\theta}}) = \log p(\boldsymbol{y} \mid \widetilde{\boldsymbol{\theta}}) + \log \pi(\widetilde{\boldsymbol{\theta}})$ to give the Laplace approximation

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y} \mid \widetilde{\boldsymbol{\theta}}) + \log \pi(\widetilde{\boldsymbol{\theta}}) + \frac{p}{2} \log 2\pi - \frac{p}{2} \log n + \frac{1}{2} \log \mid \widetilde{\boldsymbol{v}} \mid .$$

As n increases, the prior contribution will become negligible, and the posterior mode will be close to the MLE $\widehat{\boldsymbol{\theta}}$. Dropping terms of $O(1)$, we obtain the crude approximation

$$-2 \log p(\boldsymbol{y}) \approx -2 \log p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}) + p \log n.$$

Let hypothesis H_j be indexed by parameters $\boldsymbol{\theta}_j$ of length p_j and $\widehat{\boldsymbol{\theta}}_j$ denote the MLEs for $j = 0, 1$. Without loss of generality, assume $p_0 \leq p_1$. We may approximate twice the log Bayes factor by

$$\begin{aligned} 2\left[\log p(\boldsymbol{y} \mid H_0) - \log p(\boldsymbol{y} \mid H_1)\right] & \\ = 2\left[\log p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_0) - \log p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}}_1)\right] + (p_1 - p_0)\log n & \\ = 2\left[l(\widehat{\boldsymbol{\theta}}_0) - l(\widehat{\boldsymbol{\theta}}_1)\right] + (p_1 - p_0)\log n & \end{aligned} \tag{3.40}$$

which is the log-likelihood ratio statistic (see Sect. 2.9.5) with the addition of a term that penalizes complexity; (3.40) is known as the Bayesian information criteria (BIC). The Schwarz criterion (Schwarz 1978) is the BIC divided by 2. If the maximized likelihoods are approximately equal, then model H_0 is preferred if $p_0 < p_1$, as it contains fewer parameters. As n increases, the penalty term increases in size showing the difference in behavior with frequentist tests in which significance levels are often kept constant with respect to sample size. A more detailed comparison of Bayesian and frequentist approaches to hypothesis testing will be carried out in Chap. 4.

3.11 Bayesian Inference Based on a Sampling Distribution

We now describe an approach to Bayesian inference which is pragmatic and computationally simple and allows frequentist summaries to be embedded within a Bayesian framework. This is useful in situations in which one would like to examine the impact of prior specification. It is also appealing to examine frequentist procedures with no formal Bayesian justification from a Bayesian slant. Suppose we are in a situation in which the sample size n is sufficiently large for accurate asymptotic inference and suppose we have a parameter $\boldsymbol{\theta}$ of length p. The sampling distribution of the estimator is

$$\widehat{\boldsymbol{\theta}}_n \mid \boldsymbol{\theta} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{V}_n),$$

where $\boldsymbol{V}_n$ is assumed known. The notation here is sloppy; it would be more accurate to state the distribution as

$$\boldsymbol{V}_n^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \sim \mathrm{N}_p(\mathbf{0}, \mathbf{I}).$$

Appealing to conjugacy, it is then convenient to combine this "likelihood" with the prior $\boldsymbol{\theta} \sim \mathrm{N}_p(\boldsymbol{m}, \boldsymbol{W})$ to give the posterior

$$\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}_n \sim \mathrm{N}_p(\boldsymbol{m}_n^\star, \boldsymbol{W}_n^\star) \tag{3.41}$$

where

$$\boldsymbol{W}_n^\star = (\boldsymbol{W}^{-1} + \boldsymbol{V}_n^{-1})^{-1}$$
$$\boldsymbol{m}_n^\star = \boldsymbol{W}_n^\star(\boldsymbol{W}^{-1}\boldsymbol{m} + \boldsymbol{V}_n^{-1}\widehat{\boldsymbol{\theta}}_n)$$

The posterior distribution is therefore easy to determine since we only require a point estimate $\widehat{\boldsymbol{\theta}}_n$, with an associated variance–covariance matrix, and specification of the prior mean and variance–covariance matrix.

An even more straightforward approach, when a single parameter is of interest, is to ignore the remaining nuisance parameters and focus only on this single estimate and standard error. There are a number of advantages to this approach, not least of which is the removal of the need for prior specification over the nuisance parameters. Let θ denote the parameter of interest and $\boldsymbol{\alpha}$ the $(p \times 1)$ vector of nuisance parameters. Following Wakefield (2009a), we give a derivation beginning with the asymptotic distribution (we drop the explicit dependence on n for notational convenience):

$$\begin{bmatrix} \widehat{\boldsymbol{\alpha}} \\ \widehat{\theta} \end{bmatrix} \sim \mathrm{N}_{p+1}\left(\begin{bmatrix} \boldsymbol{\alpha} \\ \theta \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{00} & \boldsymbol{I}_{01} \\ \boldsymbol{I}_{01}^{\mathrm{T}} & I_{11} \end{bmatrix}^{-1} \right) \tag{3.42}$$

where $\boldsymbol{I}_{00}$ is the $p \times p$ expected information matrix for $\boldsymbol{\alpha}$, I_{11} is the information concerning θ, and $\boldsymbol{I}_{01}$ is the $p \times 1$ vector of cross terms. We now reparameterize the model and consider $(\boldsymbol{\alpha}, \theta) \rightarrow (\boldsymbol{\gamma}, \theta)$ where

$$\boldsymbol{\gamma} = \boldsymbol{\alpha} + \frac{\boldsymbol{I}_{01}}{I_{00}}\theta$$

which yields

$$\begin{bmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\theta} \end{bmatrix} \sim \mathrm{N}_{p+1}\left(\begin{bmatrix} \boldsymbol{\gamma} \\ \theta \end{bmatrix}, \begin{bmatrix} \boldsymbol{I}_{00}^\star & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & I_{11} \end{bmatrix}^{-1} \right) \tag{3.43}$$

where $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\alpha}} + (\boldsymbol{I}_{01}/I_{00})\,\widehat{\theta}$ and $\mathbf{0}$ is a $p \times 1$ vector of zeros. Hence, asymptotically, the "likelihood" factors into independent pieces

$$p(\widehat{\boldsymbol{\gamma}}, \widehat{\theta} \mid \boldsymbol{\gamma}, \theta) = p(\widehat{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}) \times p(\widehat{\theta} \mid \theta).$$

We now assume independent priors on $\boldsymbol{\gamma}$ and θ, $\pi(\boldsymbol{\gamma}, \theta) = \pi(\boldsymbol{\gamma})\pi(\theta)$, to give

$$\begin{aligned} p(\boldsymbol{\gamma}, \theta \mid \widehat{\boldsymbol{\gamma}}, \widehat{\theta}) &= p(\widehat{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma})\pi(\boldsymbol{\gamma})p(\widehat{\theta} \mid \theta)\pi(\theta) \\ &= p(\boldsymbol{\gamma} \mid \widehat{\boldsymbol{\gamma}})p(\theta \mid \widehat{\theta}) \end{aligned}$$

so that the posterior factors also and we can concentrate on $p(\theta \mid \widehat{\theta})$ alone. The simple model

$$\begin{aligned} \widehat{\theta} \mid \theta &\sim \mathrm{N}(\theta, V) \\ \theta &\sim \mathrm{N}(m, W) \end{aligned}$$

therefore results in the posterior

$$\theta \mid \widehat{\theta} \sim \mathrm{N}\left[(W^{-1}+V^{-1})^{-1}(W^{-1}m+V^{-1}\widehat{\theta}),(W^{-1}+V^{-1})^{-1}\right]. \tag{3.44}$$

The above approach is similar to the "null orthogonality" reparameterization of Kass and Vaidyanathan (1992). The reparameterization is also that which is used when the linear model

$$Y_i = \alpha + x_i\theta + \epsilon_i$$

is written as

$$Y_i = \gamma + (x_i - \overline{x})\theta + \epsilon_i$$

which, of course, yields uncorrelated least squares estimators $\widehat{\gamma}, \widehat{\theta}$. The reparameterization trick works because of the assumption of independent priors on $\boldsymbol{\gamma}$ and θ which, of course, does not imply independent priors on $\boldsymbol{\alpha}$ and θ. However, we emphasize that we do not need to explicitly specify priors on $\boldsymbol{\gamma}$, because the terms involving $\boldsymbol{\gamma}$ cancel in the calculation.

Bayes factors can also be simply evaluated under either of the approximations, (3.41) or (3.44). To illustrate for the latter, suppose θ is univariate, and we wish to compare the hypotheses

$$H_0 : \theta = 0, \quad H_1 : \theta \neq 0,$$

with the prior under the alternative, $\theta \sim \mathrm{N}(0, W)$. The Bayes factor is

$$\begin{aligned}\mathrm{BF} &= \frac{p(\widehat{\theta} \mid \theta_0)}{\int p(\widehat{\theta} \mid \theta)\pi(\theta)\, d\theta} \\ &= \sqrt{\frac{V+W}{V}} \exp\left[-\frac{1}{2}\frac{\widehat{\theta}^2}{V}\frac{W}{V+W}\right].\end{aligned} \tag{3.45}$$

This approach allows a Bayesian interpretation of published results, since all that is required for calculation of (3.45) is $\widehat{\theta}$ and V, which may be derived from a confidence interval or the estimate with its associated standard error.

More controversially, an advantage of the use of the asymptotic distribution of the MLE only is that the Bayes factor calculation may be based on nonstandard likelihoods or estimating functions which do not have formal Bayesian justifications. For example, the estimate and standard error may arise from conditional or marginal likelihoods (as described in Sect. 2.4.2), or using sandwich estimates of the variance. As discussed in Chap. 2, a strength of modern frequentist methods based on estimating functions is that estimators are produced that are consistent under much milder assumptions than were used to derive the estimators (e.g., the estimator may be based on a score equation, but the variance estimate may not require the likelihood to be correctly specified). The use of a consistent variance estimate with (3.45) allows the benefits of frequentist sandwich estimation and

Bayesian prior specification to be combined. Bayesian hypothesis testing may also be based on frequentist summaries. Exercises 3.10 and 3.11 give further details on the approach described in this section, including the extension to having estimators and standard errors from multiple studies.

3.12 Concluding Remarks

Bayesian analyses should not be restricted to convenient likelihoods and likelihood/prior combinations; this is especially true with the advent of modern computational approaches. However, one still needs to be careful that the sampling scheme (i.e., the design) is acknowledged by the likelihood specification and that the likelihood/prior combination leads to a proper posterior.

We now follow up on Sect. 1.6 and describe situations in which frequentist and Bayesian methods are likely to agree and when one is preferable over the other. We concentrate on estimation since point and interval estimation are directly comparable under the two paradigms. For model comparison, the objectives of Bayes factors and hypothesis tests are fundamentally different (see, e.g., Berger (2003)), and so comparison is more difficult. Chapter 4 compares and critiques frequentist and Bayesian approaches to hypothesis testing.

On a philosophical level, the Bayesian approach is satisfying since one simply follows the rules of probability as applied to the unknowns whether they be parameters or hypotheses. This is in stark contrast to the frequentist approach in which the parameters are fixed. Consequently, credible intervals are probabilistic and easily interpretable, and posterior distributions on parameters of interest are obtained through marginalization. Another appealing characteristic is that the Bayesian approach to inference may be formally derived via decision theory; see, for example, Bernardo and Smith (1994). A concept that has received a lot of discussion is the *likelihood principle* (Berger and Wolpert 1988; Royall 1997) which states that the likelihood function contains all relevant information. So two sets of data with proportional likelihoods should lead to the same conclusion. The likelihood principle leads one toward a Bayesian approach since all frequentist criteria invalidate this principle, and a true likelihood approach as followed by, for example, Royall (1997) is difficult to calibrate. The likelihood principle is a cornerstone of many Bayesian developments, but in this book we follow a far more pragmatic approach and so do not provide further details on this topic.

In contrast, the frequentist approach is more difficult to justify on philosophical grounds. Instead, much theory has been developed in terms of optimality within a frequentist set of guidelines. For example, as discussed in Sect. 2.8, there is a Gauss–Markov theorem for linear estimating functions (Godambe and Heyde 1987; McCullagh 1983), while Crowder (1987) considers the optimality of quadratic estimating functions.

We have seen that, so long as the prior does not exclude regions of the parameter space, Bayesian estimators have similar frequentist properties to MLEs. The greatest

drawback of the Bayesian approach is the need to specify both a likelihood and a prior distribution. Sensitivity to each of these components can be examined, but carrying out such an endeavor in practice is difficult and one is then faced with the difficulty of how results should be reported. The frequentist approach to model misspecification is quite different, and the use of sandwich estimation to give a consistent standard error is very appealing. There is no Bayesian approach analogous to sandwich estimation, but see Szpiro et al. (2010) for some progress on a Bayesian justification of sandwich estimation.

For small n, Bayesian methods are desirable; in an extreme case if the number of parameters exceeds n, then a Bayesian approach (or some form of penalization, see Chaps. 10–12) must be followed. In this situation there is no way that the likelihood can be checked and inference will be sensitive to both likelihood and prior choices. When the model is very complex, then Bayesian methods are again advantageous since they allow a rigorous treatment of nuisance parameters; MCMC has allowed the consideration of more and more complicated hierarchical models, for example. Spatial models, particularly those that exploit Markov random field second stages, provide a good example of models that are very naturally analyzed using MCMC or INLA, where the conditional independencies may be exploited; see Sect. 9.7 for an illustrative example. Unfortunately, assessments of the effects of model misspecification are difficult for such complex models; instead sensitivity studies are again typically carried out. Consistency results under model misspecification are difficult to come by for complex models (such as those discussed in Chap. 9). Bayesian methods are also appealing in situations in which the maximum likelihood estimator provides a poor summary of the likelihood, for example, in variance components problems.

If n is sufficiently large for asymptotic normality of the sampling distribution to be accurate, then frequentist methods have advantages over Bayesian alternatives. In particular, as just mentioned, sandwich estimation can be used to provide a consistent estimator of the variance–covariance matrix of the estimator. Hence, if the estimator is consistent, reliable confidence coverage will be guaranteed. We stress that n needs to be sufficiently large for the sandwich estimator to be stable. A typical Bayesian approach would be to increase model complexity, often through the introduction of random effects. The difficulty with this is that although more flexibility is achieved, a specific form needs to be assumed for the mean–variance relationship, in contrast to sandwich estimation.

We briefly mention two topics which have not been discussed in this chapter. The *linear Bayesian* method (Goldstein and Wooff 2007) is an appealing approach in which Bayesian inference is carried out on the basis of expectation rather than probability. The appeal comes from the removal of the need to specify complete prior distributions, rather the means and variances of the parameters only require specification. The deviance information criterion (DIC) is a popular approach for comparison of models that was introduced by Spiegelhalter et al. (1998). The method is controversial, however, as the discussion of the aforementioned paper makes clear; see also Plummer (2008).

3.13 Bibliographic Notes

Bayes' original paper was published posthumously as Bayes (1763). The book by Jeffreys was highly influential: the original edition was published in 1939 and the third edition as Jeffreys (1961). Other influential works include Savage (1972) and translations of de Finetti's books, De Finetti (1974, 1975).

Bernardo and Smith (1994) provide a thorough description of the decision-theoretic justification of the Bayesian approach. O'Hagan and Forster (2004) give a good overview of Bayesian methodology and Gelman et al. (2004) and Carlin and Louis (2009) descriptions with a more practical flavor. Robert (2001) provides a decision-theoretic approach. Hoff (2009) is an excellent introductory text.

Approaches to addressing the sensitivity of inference to different prior choices, are described in O'Hagan (1994, Chap. 7). A good overview of methods for integration is provided by Evans and Swartz (2000). Lindley (1980), Tierney and Kadane (1986), and Kass et al. (1990) provide details of the Laplace method in a Bayesian context. Devroye (1986) provides an excellent and detailed overview of random variate generation. Smith and Gelfand (1992) emphasize the duality between samples and densities and illustrate the use of simple rejection algorithms in a Bayesian context. Gamerman and Lopes (2006) provides an introduction to MCMC; an up-to- date summary may be found in Brooks et al. (2011). Computational techniques that have not been discussed include reversible jump Markov chain Monte Carlo (Green 1995) which may be used when the parameter space changes dimension across models, variational approximations (Jordan et al. 1999; Ormerod and Wand 2010), and approximate Bayesian computation (ABC) (Beaumont et al. 2002; Fearnhead and Prangle 2012). Kass and Raftery (1995) give a review of Bayes factors, including a discussion of computation and prior choice. Johnson (2008) discusses the use of Bayes factors based on summary statistics.

3.14 Exercises

3.1 Derive the posterior mean and posterior quantiles as the solution to quadratic and linear loss, respectively, as described in Sect. 3.2.

3.2 Consider a random sample $Y_i \mid \theta \sim_{iid} \mathrm{N}(\theta, \sigma^2)$, $i = 1, \ldots, n$, with θ unknown and σ^2 known.

(a) By writing the likelihood in exponential family form, obtain the conjugate prior and hence the posterior distribution.

(b) Using the conjugate formulation, derive the predictive distribution for a new univariate observation Z from $\mathrm{N}(\theta, \sigma^2)$, assumed conditionally independent of $Y_1, \ldots, Y_n$.

3.3 Consider the Neyman–Scott problem in which $Y_{ij} \mid \mu_i, \sigma^2 \sim_{ind} \mathrm{N}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, 2$.

Table 3.4 Case-control data: $Y = 1$ corresponds to the event of esophageal cancer, and $X = 1$ exposure to greater than 80 g of alcohol per day

	$X = 0$	$X = 1$	
$Y = 1$	104	96	200
$Y = 0$	666	109	775

(a) Show that Jeffreys prior in this case is

$$\pi(\mu_1, \ldots, \mu_n, \sigma^2) \propto \sigma^{-n-2}.$$

(b) Derive the posterior distribution corresponding to this prior and show that

$$\mathrm{E}[\sigma^2 \mid \boldsymbol{y}] = \frac{1}{2(n-1)} \sum_{i=1}^{n} \frac{(Y_{i1} - Y_{i2})^2}{2}.$$

(c) Hence, using Exercise 2.6, show that $\mathrm{E}[\sigma^2 \mid \boldsymbol{y}] \to \sigma^2/2$ as $n \to \infty$, so that the posterior mean is inconsistent.

(d) Examine the posterior distribution corresponding to the prior

$$\pi(\mu_1, \ldots, \mu_n, \sigma^2) \propto \sigma^{-2}.$$

(e) Is the posterior mean for σ^2 consistent in this case?

3.4 Consider the data given in Table 3.4, which are a simplified version of those reported in Breslow and Day (1980). These data arose from a case-control study (Sect. 7.10) that was carried out to investigate the relationship between esophageal cancer and various risk factors. There are 200 cases and 775 controls. Disease status is denoted Y with $Y = 0/1$ corresponding to without/with disease, and alcohol consumption is represented by X with $X = 0/1$ denoting $< 80\,\mathrm{g}/ \geq 80\,\mathrm{g}$ on average per day. Let the probabilities of high alcohol consumption in the cases and controls be denoted

$$p_1 = \Pr(X = 1 \mid Y = 1) \quad \text{and} \quad p_2 = \Pr(X = 1 \mid Y = 0),$$

respectively. Further, let X_1 be the number exposed from n_1 cases and X_2 be the number exposed from n_2 controls. Suppose $X_i \mid p_i \sim \text{Binomial}(n_i, p_i)$ in the case ($i = 1$) and control ($i = 2$) groups.

(a) Of particular interest in studies such as this is the *odds ratio* defined by

$$\theta = \frac{\Pr(Y = 1 \mid X = 1)/\Pr(Y = 0 \mid X = 1)}{\Pr(Y = 1 \mid X = 0)/\Pr(Y = 0 \mid X = 0)}.$$

Show that the odds ratio is equal to

$$\theta = \frac{\Pr(X = 1 \mid Y = 1)/\Pr(X = 0 \mid Y = 1)}{\Pr(X = 1 \mid Y = 0)/\Pr(X = 0 \mid Y = 0)} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

(b) Obtain the MLE and a 90% confidence interval for θ, for the data of Table 3.4.

(c) We now consider a Bayesian analysis. Assume that the prior distribution for p_i is the beta distribution $\text{Be}(a, b)$ for $i = 1, 2$. Show that the posterior distribution $p_i \mid x_i$ is given by the beta distribution $\text{Be}(a+x_i, b+n_i-x_i)$, $i = 1, 2$.

(d) Consider the case $a = b = 1$. Obtain expressions for the posterior mean, mode, and standard deviation. Evaluate these posterior summaries for the data of Table 3.4. Report 90% posterior credible intervals for p_1 and p_2.

(e) Obtain the asymptotic form of the posterior distribution and obtain 90% credible intervals for p_1 and p_2. Compare this interval with the exact calculation of the previous part.

(f) Simulate samples $p_1^{(t)}, p_2^{(t)}$, $t = 1, \ldots, T = 1{,}000$ from the posterior distributions $p_1 \mid x_1$ and $p_2 \mid x_2$. Form histogram representations of the posterior distributions using these samples, and obtain sample-based 90% credible intervals.

(g) Obtain samples from the posterior distribution of $\theta \mid x_1, x_2$ and provide a histogram representation of the posterior. Obtain the posterior median and 90% credible interval for $\theta \mid x_1, x_2$ and compare with the likelihood analysis.

(h) Suppose the rate of esophageal cancer is 17 in 100,000. Describe how this information may be used to evaluate

$$q_1 = \Pr(Y = 1 \mid X = 1) \quad \text{and} \quad q_0 = \Pr(Y = 1 \mid X = 0).$$

3.5 Prove that if global balance, as given by (3.31), holds then $\pi(\cdot)$ is the invariant distribution, that is,

$$\pi(A) = \int_{\mathbb{R}^p} \pi(\boldsymbol{x}) P(\boldsymbol{x}, A)\, d\boldsymbol{x},$$

for all measurable sets A.

3.6 Prove that the Metropolis–Hastings algorithm, defined through (3.33), has invariant distribution $\pi(\cdot)$, by showing that detailed balance (3.31) holds.

3.7 We consider the data described in the example at the end of Sect. 3.7.7 concerning the leukemia count, Y, assumed to follow a Poisson distribution with mean $E \times \delta$. Consider the $y = 4$ observed leukemia cases in Seascale, with expected number of cases $E = 0.25$. Previously in this chapter, a lognormal prior was assumed for δ. In this exercise, a conjugate gamma prior will be used.

(a) Show that with a $\text{Ga}(a, b)$ prior, the posterior distribution for δ is a gamma distribution also. Hence, determine the posterior mean, mode, and variance. Show that the posterior mean can be written as a weighted combination of the MLE and the prior mean. Similarly write the posterior mode as a weighted combination of the MLE and the prior mode.
(b) Determine the form of the prior predictive $\Pr(y)$ and show that it corresponds to a negative binomial distribution.
(c) Obtain the predictive distribution $\Pr(z \mid y)$ for the number of cases z in a future period of time with expected number of cases $E^\star$.
(d) Obtain the posterior distribution under gamma prior distributions with parameters $a = b = 0.1$, $a = b = 1.0$, and $a = b = 10$. Determine the 5%, 50%, and 95% posterior quantiles in each case and comment on the sensitivity to the prior.

3.8 Consider a situation in which the likelihood may be summarized as

$$\sqrt{n}(\overline{Y}_n - \mu) \rightarrow_d \text{N}(0, \sigma^2),$$

where $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$, with σ^2 known, and the prior for μ is the Cauchy distribution with parameters 0 and 1, that is,

$$p(\mu) = \frac{1}{\pi(1+\mu^2)}, \quad -\infty < \mu < \infty.$$

We label this likelihood-prior combination as model M_c.

(a) Describe a rejection algorithm for obtaining samples from the posterior distribution, with the proposal density taken as the prior.
(b) Implement the rejection algorithm for the case in which $\overline{y} = 0.2$, σ^2=2 and $n = 10$. Provide a histogram representation of the posterior, and evaluate the posterior mean and variance. Also obtain an estimate of the normalizing constant, $p(\boldsymbol{y} \mid M_c)$.
(c) Describe an importance sampling algorithm for evaluating $p(\boldsymbol{y} \mid M_c)$, $\text{E}[\mu \mid \boldsymbol{y}, M_c]$, and $\text{var}(\mu \mid \boldsymbol{y}, M_c)$.
(d) For the data of part (b), implement the importance sampling algorithm, and calculate $p(\boldsymbol{y} \mid M_c)$ and $\text{E}[\mu \mid \boldsymbol{y}, M_c]$ and $\text{var}(\mu \mid \boldsymbol{y}, M_c)$.
(e) Now assume that the prior for μ is the normal distribution $\text{N}(0, 0.4)$. Denote this model M_n. Obtain the form of the posterior distribution in this case.
(f) For the data of part (b), obtain the normalizing constant $p(\boldsymbol{y} \mid M_n)$ and the posterior mean and variance. Compare these summaries with those obtained under the Cauchy prior. Interpret the ratio

$$\frac{p(\boldsymbol{y} \mid M_n)}{p(\boldsymbol{y} \mid M_c)},$$

that is, the Bayes factor, for these data.

Table 3.5 Genetic data from an experiment carried out by Mendel that concerned the numbers of peas that were classified by their shape and color

Round yellow	Wrinkled yellow	Round green	Wrinkled green	Total
n_1	n_2	n_3	n_4	n_+
315	101	108	32	556

3.9 The data in Table 3.5 result from one of the famous experiments carried out by Mendel in which pure bred peas with wrinkled green seeds were crossed with pure bred peas with wrinkled green seeds. These data are given on page 15 of the English translation (Mendel 1901) of Mendel (1866). All of the first-generation hybrids had round yellow seeds (since this characteristic is dominant), but when these plants were self-pollinated, four different phenotypes (characteristics) were observed and are displayed in Table 3.5.

A model for these data is provided by the multinomial $M_4(n_+, \boldsymbol{p})$ where $\boldsymbol{p} = [p_1, p_2, p_3, p_4]^{\mathrm{T}}$, and p_j denotes the probability of falling in cell j, $j = 1, \ldots, 4$, that is,

$$\Pr(\boldsymbol{N} = \boldsymbol{n} \mid \boldsymbol{p}) = \frac{n_{+!}}{\prod_{j=1}^{4} n_j!} \prod_{j=1}^{4} p_j^{n_j},$$

where $\boldsymbol{N} = [N_1, \ldots, N_4]^{\mathrm{T}}$ and $\boldsymbol{n} = [n_1, \ldots, n_4]^{\mathrm{T}}$. In this exercise a Bayesian analysis of these data will be carried out using the conjugate Dirichlet prior distribution, $\mathrm{Dir}(a_1, a_2, a_3, a_4)$:

$$p(\boldsymbol{p}) = \frac{\Gamma\left(\sum_{j=1}^{4} a_j\right)}{\prod_{j=1}^{4} \Gamma(a_j)} \prod_{j=1}^{4} p_j^{a_j - 1},$$

where $a_j > 0$, $j = 1, \ldots, 4$, are specified a priori.

(a) Show that the marginal prior distributions for p_j are the beta distributions $\mathrm{Be}(a_j, a - a_j)$, where $a = \sum_{j=1}^{4} a_j$.
(b) Obtain the distributional form, and the associated parameters, of the posterior distribution $p(\boldsymbol{p} \mid \boldsymbol{n})$.
(c) For the genetic data and under a prior for $\boldsymbol{p}$ that is uniform over the simplex (i.e., $a_1 = a_2 = a_3 = a_4 = 1$), evaluate $\mathrm{E}[p_j \mid \boldsymbol{n}]$ and $\mathrm{s.d.}(p_j \mid \boldsymbol{n})$, $j = 1, \ldots, 4$.
(d) Obtain histogram representations and 90% credible intervals for $p_j \mid \boldsymbol{n}$, $j = 1, \ldots, 4$.
(e) Determine the form of the predictive distribution for $[N_1, N_2, N_3, N_4]$ given $n_+ = \sum_j n_j$. Describe how a sample from this predictive distribution could be obtained.

A particular model of interest is that which states that genes are inherited independently of each other, so that the ratio of counts is 9:3:3:1, or

$$H_0 : p_{10} = \frac{9}{16}, p_{20} = \frac{3}{16}, p_{30} = \frac{3}{16}, p_{40} = \frac{1}{16}.$$

The evidence in favor of this model, versus the alternative of H_1 : $\boldsymbol{p}$ unspecified, will now be determined.

(f) For the data in Table 3.5, carry out a likelihood ratio test comparing H_0 and H_1.

(g) Obtain analytical expressions for $\Pr(\boldsymbol{n} \mid H_0)$ and $\Pr(\boldsymbol{n} \mid H_1)$.

(h) Evaluate the Bayes factor $\Pr(\boldsymbol{n} \mid H_0)/\Pr(\boldsymbol{n} \mid H_1)$ for the genetic data. Comment on the evidence for/against H_0 and compare with the conclusion from the likelihood ratio test statistic.

3.10 With respect to Sect. 3.11, consider the "likelihood," $\widehat{\theta} \mid \theta \sim \mathrm{N}(\theta, V)$ and the prior $\theta \sim \mathrm{N}(0, W)$. Show that $\theta \mid \widehat{\theta} \sim \mathrm{N}(r\widehat{\theta}, rV)$ where $r = W/(V + W)$.

3.11 Again consider the situation discussed in Sect. 3.11 in which a Bayesian analysis is carried out based not on the full data but rather on summary statistics.

(a) Suppose data are to be combined from two studies with a common underlying parameter θ. The estimates from the two studies are $\widehat{\theta}_1, \widehat{\theta}_2$ with standard errors $\sqrt{V_1}$ and $\sqrt{V_2}$ (with the two estimators being conditionally independent given θ). Show that the Bayes factor that summarizes the evidence from the two studies, that is,

$$\frac{p(\widehat{\theta}_1, \widehat{\theta}_2 \mid H_0)}{p(\widehat{\theta}_1, \widehat{\theta}_2 \mid H_1)},$$

takes the form

$$\mathrm{BF}(\widehat{\theta}_1, \widehat{\theta}_2) = \sqrt{\frac{W}{RV_1V_2}} \exp\left[-\frac{1}{2}\left(Z_1^2 RV_2 + 2Z_1Z_2R\sqrt{V_1V_2} + Z_2^2 RV_1\right)\right]$$

where $R = W/(V_1W + V_2W + V_1V_2)$ and $Z_1 = \widehat{\theta}_1/\sqrt{V_1}$ and $Z_2 = \widehat{\theta}_2/\sqrt{V_2}$ are the usual Z-statistics.

(b) Suppose now there are K studies with estimates $\widehat{\theta}_k$ and asymptotic variances V_k, $k = 1, \ldots, K$, and again assume a common underlying parameter θ. Show that the Bayes factor

$$\frac{p(\widehat{\theta}_1, \ldots, \widehat{\theta}_K \mid H_0)}{p(\widehat{\theta}_1, \ldots, \widehat{\theta}_K \mid H_1)},$$

takes the form

$$\begin{aligned}
&\mathrm{BF}(\widehat{\theta}_1, \ldots, \widehat{\theta}_K) \\
&= \frac{\prod_{k=1}^{K} (2\pi V_k)^{-1/2} \exp\left(-\frac{\widehat{\theta}_k^2}{2V_k}\right)}{\int \prod_{k=1}^{K} (2\pi V_k)^{-1/2} \exp\left(-\frac{\left(\widehat{\theta}_k^2 - \theta\right)^2}{2V_k}\right) (2\pi W)^{-1/2} \exp\left(-\frac{\theta^2}{2W}\right) d\theta} \\
&= \sqrt{W\left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)} \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{K} \frac{\widehat{\theta}_k}{V_k}\right)^2 \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1}\right].
\end{aligned}$$

Further, show that the posterior summarizing beliefs about θ given the K estimates is

$$\theta \mid \widehat{\theta}_1, \ldots, \widehat{\theta}_K \sim \mathrm{N}(\mu, \sigma^2)$$

where

$$\mu = \left(\sum_{k=1}^{K} \frac{\widehat{\theta}_k}{V_k}\right) \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1}$$

and

$$\sigma^2 = \left(W^{-1} + \sum_{k=1}^{K} V_k^{-1}\right)^{-1}.$$