11

Bayesian Models

Every statistical investigation takes place in a context. Information about what question is to be addressed will suggest what data are needed to give useful answers. Before the data are available, one role for this information is to suggest suitable probability models. There may also be information about the values of unknown parameters, and if this can be expressed as a probability density, an approach to inference based on Bayes' theorem is possible. Many statisticians make the stronger claim that this theorem provides the only entirely consistent basis for inference, and insist on its use.

This chapter outlines some aspects of the Bayesian approach to modelling. We first give an account of basic uses of Bayes' theorem and of the role and construction of prior densities. We then turn to inference, dealing with analogues of confidence intervals, tests, approaches to model criticism, and model uncertainty. Until recently computational difficulties placed realistic Bayesian modelling largely out of reach, but over the last 20 years there has been rapid progress and complex models can now be fitted routinely. Section 11.3 gives an account of Bayesian computation, first of analytical approaches based on integral approximations, and then of Monte Carlo methods. The chapter concludes with brief introductions to hierarchical and empirical Bayesian procedures.

11.1 Introduction

11.1.1 Bayes' theorem

Let $A_1, \ldots, A_k$ be events that partition a sample space, and let B be an arbitrary event on that space for which $\Pr(B) > 0$. Then Bayes' theorem is

$$\Pr(A_j \mid B) = \frac{\Pr(B \mid A_j)\Pr(A_j)}{\sum_{i=1}^{k} \Pr(B \mid A_i)\Pr(A_i)}.$$

This reverses the order of conditioning by expressing $\Pr(A_j \mid B)$ in terms of $\Pr(B \mid A_j)$ and the marginal probability $\Pr(B)$ in the denominator. For continuous random variables Y and Z,

$$f_{Z|Y}(z \mid y) = \frac{f_{Y|Z}(y \mid z) f_Z(z)}{\int f_{Y|Z}(y \mid z) f_Z(z)\, dz}, \tag{11.1}$$

provided the marginal density $f(y) > 0$, with integration replaced by summation for discrete variables.

Inference

To see how Bayes' theorem is used for inference, suppose that there is a probability model $f(y \mid \theta)$ for data y. In earlier chapters we have written $f(y \mid \theta) = f(y; \theta)$, but here we use the conditional notation to emphasize that the probability model is a density for the data given the value of θ. Suppose also that we are able to summarize our beliefs about θ in a *prior density*, $\pi(\theta)$, constructed separately from the data y. This implies that we think of the unknown value θ that underlies our data as the outcome of a random variable whose density is $\pi(\theta)$, just as our probability model is that the data y are the observed value of a random variable Y with density $f(y \mid \theta)$. Once the data have been observed, our beliefs about θ are contained in its conditional density given that $Y = y$,

$$\pi(\theta \mid y) = \frac{\pi(\theta) f(y \mid \theta)}{\int \pi(\theta) f(y \mid \theta)\, d\theta}. \tag{11.2}$$

This is the *posterior density* for θ given y. Note that $f(y \mid \theta)$ is the likelihood for θ based on y, so that in terms of θ, we have *posterior* $\propto$ *prior* $\times$ *likelihood*.

Frequentist inference treats θ as an unknown constant, whereas the Bayesian approach treats it as a random variable. We make this distinction explicit by using π to denote a density for θ, which thus has prior and posterior densities $\pi(\theta)$ and $\pi(\theta \mid y)$, rather than $f(\theta)$ and $f(\theta \mid y)$.

It is useful to note that any quantity that does not depend on θ cancels from the denominator and numerator of (11.2). This implies that if we can recognise which density is proportional to (11.2), regarded solely as a function of θ, we can read off the posterior density of θ. Furthermore, the factorization criterion (4.15) implies that the posterior density depends on the data solely through any minimal sufficient statistic for θ.

Example 11.1 (Bernoulli trials) Suppose that conditional on θ, the data $y_1, \ldots, y_n$ are a random sample from the Bernoulli distribution, for which $\Pr(Y_j = 1) = \theta$ and $\Pr(Y_j = 0) = 1 - \theta$, where $0 < \theta < 1$. The likelihood is

$$L(\theta) = f(y \mid \theta) = \prod_{j=1}^{n} \theta^{y_j} (1 - \theta)^{1 - y_j} = \theta^r (1 - \theta)^{n - r}, \quad 0 < \theta < 1,$$

where $r = \sum y_j$.

A natural prior here is the beta density with parameters a and b,

$$\pi(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1, \quad a, b > 0, \tag{11.3}$$

$\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}\,du$ is the gamma function; see Exercise 2.1.3.

where $B(a,b)$ is the beta function $\Gamma(a)\Gamma(b)/\Gamma(a+b)$. Figure 5.4 shows (11.3) for various values of a and b.

The posterior density of θ conditional on the data is given by (11.2), and is

$$\begin{aligned} \pi(\theta \mid y) &= \frac{\theta^{r+a-1}(1-\theta)^{n-r+b-1}/B(a,b)}{\int_0^1 \theta^{r+a-1}(1-\theta)^{n-r+b-1}\,d\theta/B(a,b)} \\ &\propto \theta^{r+a-1}(1-\theta)^{n-r+b-1}, \quad 0 < \theta < 1. \end{aligned} \tag{11.4}$$

As (11.3) has unit integral for all positive a and b, the constant normalizing (11.4) must be $B(a+r, b+n-r)$. Therefore

$$\pi(\theta \mid y) = \frac{1}{B(a+r, b+n-r)}\theta^{r+a-1}(1-\theta)^{n-r+b-1}, \quad 0 < \theta < 1.$$

Thus the posterior density of θ has the same form as the prior: acquiring data has the effect of updating (a,b) to $(a+r, b+n-r)$. As the mean of the $B(a,b)$ density is $a/(a+b)$, the posterior mean is $(r+a)/(n+a+b)$, and this is roughly r/n in large samples. Hence the prior density inserts information equivalent to having seen a sample of $a+b$ observations, of which a were successes. If we were very sure that $\theta \doteq 1/2$, for example, we might take $a = b$ very large, giving a prior density tightly concentrated around $\theta = 1/2$, whereas taking smaller values of a and b would increase the prior uncertainty.

To illustrate this, suppose that $a = b = 1$, so that the initial density of θ is the uniform prior shown in the upper right panel of Figure 5.4, representing ignorance about θ. Then data with $n = 23$ and $r = \sum y_j = 14$ update the prior density to the posterior density in the lower right panel. ■

The use of the beta density as prior for a model whose likelihood is proportional to $\theta^r(1-\theta)^s$ leads to a posterior density that is also beta. This is an example of a *conjugate prior*, an idea to be discussed in Section 11.1.3.

When the parameter takes one of a finite number of values, labelled $1, \ldots, k$, with prior probabilities $\pi_1, \ldots, \pi_k$, the posterior density is the probability mass function

$$\Pr(\theta = j \mid y) = \frac{\pi_j f(y \mid \theta = j)}{\sum_{i=1}^k \pi_i f(y \mid \theta = i)}. \tag{11.5}$$

Example 11.2 (Diagnostic tests) A disease occurs with prevalence γ in a population, and θ indicates that an individual has the disease. Hence $\Pr(\theta = 1) = \gamma$, $\Pr(\theta = 0) = 1 - \gamma$. A diagnostic test gives a result Y, whose distribution is $F_1(y)$ for a diseased individual and $F_0(y)$ otherwise.

The commonest type of test declares that a person is diseased if $Y > y_0$, say, where y_0 is fixed on the basis of past data. The probability that a person is diseased, given a positive test result, is

$$\Pr(\theta = 1 \mid Y > y_0) = \frac{\gamma\{1 - F_1(y_0)\}}{\gamma\{1 - F_1(y_0)\} + (1-\gamma)\{1 - F_0(y_0)\}};$$

this is sometimes called the *positive predictive value* of the test. Its *sensitivity* and *specificity* are $1 - F_1(y_0)$ and $F_0(y_0)$. These are the probabilities of correct classification of diseased and non-diseased persons, while the false negative and false positive ratios are $F_1(y_0)$ and $1 - F_0(y_0)$. One aims to construct tests whose sensitivity and specificity are as high as possible. ■

Prediction

Prediction of the value of a future random variable, Z, is straightforward when there is a prior density for the parameters. The joint density of Z and the data Y may be written

$$f(y, z) = \int f(z \mid y, \theta) f(y \mid \theta) \pi(\theta)\, d\theta,$$

and hence once Y has taken the value y, inference for Z is based on its *posterior predictive density*,

$$f(z \mid y) = \int f(z \mid y, \theta) \pi(\theta \mid y)\, d\theta = \frac{\int f(z \mid y, \theta) f(y \mid \theta) \pi(\theta)\, d\theta}{\int f(y \mid \theta) \pi(\theta)\, d\theta}. \qquad (11.6)$$

This is (11.1) expanded to make explicit the integration over the posterior density of θ.

Example 11.3 (Bernoulli trials) Heads occurs r times among the first n tosses in a sequence of independent throws of a coin. What is the probability of a head on the next throw?

Let θ be the unknown probability of a head and let $Z = 1$ indicate the event that the next toss yields a head. Conditional on θ, $\Pr(Z = 1 \mid y, \theta) = \theta$ independent of the data y so far. If the prior density for θ is beta with parameters a and b, then

$$\begin{aligned} \Pr(Z = 1 \mid y) &= \int_0^1 \Pr(Z = 1 \mid \theta, y) \pi(\theta \mid y)\, d\theta \\ &= \int_0^1 \theta\, \frac{\theta^{a+r-1}(1-\theta)^{b+n-r-1}}{B(a+r, b+n-r)}\, d\theta \\ &= \frac{B(a+r+1, b+n-r)}{B(a+r, b+n-r)} = \frac{a+r}{a+b+n}, \end{aligned}$$

on using results for beta functions; see Example 11.1 and Exercise 2.1.3. As $n, r \to \infty$, this tends to the sample proportion of heads r/n, so the prior information is drowned by the sample. ■

11.1.2 Likelihood principle

There have been many attempts to justify the use of Bayes' theorem as a basis for inference. One line of argument rests on axioms that individuals can use to make optimal decisions in the face of uncertain events, and leads to the view that probability is a measure of personal belief about the world, to be updated by additional knowledge using Bayes' theorem. An account of this would take us too far afield, and instead we outline another argument, which centres on principles intended to guide inference. The force of this is that two basic principles — the sufficiency and conditionality principles — together imply a third — the likelihood principle — which is difficult to apply except through Bayes' theorem. Many statisticians do subscribe to the first two, at least implicitly, thus setting them on the path to Bayesian inference.

We begin by introducing the notion of an experiment E, which yields data y, on which we wish to base inference about θ through the *evidence* $\mathrm{Ev}(E, y)$. The form of this function need not be specified; we merely suppose that it exists and contains all the information about θ based on E and y.

Sufficiency and conditionality principles

The form of the sufficiency principle we shall use is that if an experiment E could give rise to y_1 and y_2, but that there is a statistic $s(\cdot)$ sufficient for θ such that $s(y_1) = s(y_2)$, then any inference for θ should be the same whether y_1 or y_2 is observed, that is $\mathrm{Ev}(E, y_1) = \mathrm{Ev}(E, y_2)$. This is widely accepted, as the factorization criterion (4.15) implies that given the sufficient statistic, the data contain no further information about θ.

A second principle can be motivated by the following classic example.

Example 11.4 (Measuring machines) Suppose that a physical quantity θ can be measured by two machines, both giving normal measurements Y with mean θ. A measurement from the first machine has unit variance, but one from the second has variance 100. The more precise machine is often busy, while the second is used only if the first is unavailable; the upshot is that each is equally likely to be used. Thus if A takes value 1 or 2 depending on the machine used, $\Pr(A = 1) = \Pr(A = 2) = \frac{1}{2}$.

Suppose that an observation obtained is from machine 1. Then clearly any inference about θ should not take into account that machine 2 might have been used, when it is known that it was not. Mathematically this is expressed by saying that the revelant distribution for inference about θ is the conditional distribution of Y given A, rather than the unconditional distribution of Y. For example, the conditional 95% confidence interval for θ given that $A = 1$ is $y \pm 1.96$, whereas the unconditional interval is $y \pm 16.45$, which is clearly much too long if it is known that y came from the $N(\theta, 1)$ distribution. ■

The lesson of this is formalized as follows. Suppose that an experiment

E can be thought of as arising in two stages. In the first stage we observe that a random variable A with known distribution independent of θ takes value a, and in the second stage we observe y_a from a component experiment E_a. This is a mixture experiment, for which the data are (a, y_a). Then one form of the *conditionality principle* says that $\mathrm{Ev}\{E, (a, y_a)\} = \mathrm{Ev}(E_a, y_a)$: the evidence concerning θ based on the compound experiment E is equal to the evidence from the component experiment E_a *actually performed*, the results of other possible components being irrelevant. The key point is that since the distribution of A does not depend on θ, conditioning on A does not lead to a loss of information about θ, but selects the relevant component of the mixture experiment. This principle is widely, even if sometimes unconsciously, accepted; we discuss its implications in more detail in Chapter 12.

Likelihood principle

Suppose that two experiments relating to θ, E_1 and E_2, give rise to data y_1 and y_2 such that the corresponding likelihoods are proportional, that is, for all θ,

$$L(\theta; y_1, E_1) = cL(\theta; y_2, E_2).$$

Then according to one expression of the *likelihood principle*, $\mathrm{Ev}(E_1, y_1) = \mathrm{Ev}(E_2, y_2)$: inference should be based on the observed likelihood alone. Full acceptance of this means rejecting frequentist tools such as significance tests, as the following example shows.

Example 11.5 (Bernoulli trials) Suppose that E_1 consists of observing the number y_1 of successes in a fixed number n_1 of independent Bernoulli trials. The likelihood is then

$$L_1(\theta) = \binom{n_1}{y_1}\theta^{y_1}(1-\theta)^{n_1-y_1}, \quad 0 < \theta < 1,$$

corresponding to the binomial number of successful trials.

Experiment E_2 consists of conducting Bernoulli trials independently until y_2 successes occur, at which point there have been n_2 trials. Here the likelihood,

$$L_2(\theta) = \binom{n_2-1}{y_2-1}\theta^{y_2}(1-\theta)^{n_2-y_2}, \quad 0 < \theta < 1,$$

corresponds to the negative binomial number of trials up to y_2 successes.

Now suppose that it happens that $n_1 = n_2 = n$ and $y_1 = y_2 = y$, giving $L_1(\theta) \propto L_2(\theta)$. Then according to the likelihood principle, inferences based on the two experiments should be the same. But consider testing the hypothesis $H_0 : \theta = \frac{1}{2}$ against the alternative that $\theta < \frac{1}{2}$. In E_1, the test statistic would

be the random number of successes, Y, and the P-value would be

$$\Pr(Y \leq y \mid \theta = \tfrac{1}{2}) = \sum_{r=0}^{y} \binom{n}{r} 2^{-n}, \tag{11.7}$$

while in E_2 the test statistic would be the total number of trials, N, with P-value

$$\Pr(N \geq n \mid \theta = \tfrac{1}{2}) = \sum_{m=n}^{\infty} \binom{m-1}{y-1} 2^{-m}. \tag{11.8}$$

The catch is that (11.7) and (11.8) need not be equal. For example, if $y = 3$ and $n = 12$, the P-values are respectively 0.073 and 0.033, conveying different evidence against H_0. In particular, use of the fixed significance level 0.05 would lead to acceptance or rejection of H_0 depending on the experiment performed. The reason for this is that (11.7) and (11.8) involve summation over portions of two different sample spaces. This conflicts with the likelihood principle, according to which only the data actually observed should contribute to the inference. ■

Construction of tail probabilities such as (11.7) or (11.8), or of confidence intervals, involves consideration of data not actually observed, and thereby disobeys the likelihood principle. This poses a problem for frequentist procedures, because a rational statistician who rejects the likelihood principle should also reject one of the apparently reasonable sufficiency and conditionality principles, which together entail the likelihood principle.

To see this, suppose that we accept the sufficiency and conditionality principles, and that experiments E_1 and E_2 have yielded data y_1 and y_2 such that $L(\theta; y_1, E_1) = cL(\theta; y_2, E_2)$ for some $c > 0$ and all θ. Consider the mixture experiment E that consists of observing (E_a, y_a), where a is the observed value of the binary random variable such that

$$\Pr(A = 1) = \frac{1}{c+1}, \quad \Pr(A = 2) = \frac{c}{c+1};$$

the distribution of A is independent of θ. The outcomes for E are (E_1, y_1) and (E_2, y_2), and the decomposition $\Pr(E_a, y_a; \theta) = \Pr(y_a \mid E_a; \theta)\Pr(E_a)$ shows that the corresponding likelihoods,

$$\frac{1}{c+1} L(\theta; y_1, E_1), \quad \frac{c}{c+1} L(\theta; y_2, E_2),$$

are equal for all θ. Since the likelihood function is itself a minimal sufficient statistic for θ (Exercise 4.2.11), the sufficiency principle implies

$$\mathrm{Ev}\{E, (E_1, y_1)\} = \mathrm{Ev}\{E, (E_2, y_2)\}. \tag{11.9}$$

But the conditionality principle implies

$$\mathrm{Ev}\{E,(E_1,y_1)\} = \mathrm{Ev}(E_1,y_1), \quad \mathrm{Ev}\{E,(E_2,y_2)\} = \mathrm{Ev}(E_2,y_2),$$

and combined with (11.9) we get $\mathrm{Ev}(E_1,y_1) = \mathrm{Ev}(E_2,y_2)$. Thus acceptance of the sufficiency and conditionality principles implies acceptance of the likelihood principle. The converse is also true (Problem 11.6). In fact it can be shown that a stronger version of the conditionality principle on its own implies the likelihood principle.

Statisticians attempting to weaken the force of this argument have criticized its central notions of evidence and mixture experiments, or have insisted that the sufficiency and conditionality principles apply only in a more limited way. They can then accept some form of these principles but not the conclusion of the argument, and continue to use such tools as confidence intervals and P-values. Others deny the validity of the argument on the grounds that it applies only to models known to be true, and this is rare in practice.

Statisticians who embrace the likelihood principle find themselves in an awkward position: their inference should be based on the observed likelihood, $L(\theta)$, but how should it be expressed? In particular, what can be inferred about a scalar component of vector θ? The obvious solution of profiling over the other components of θ can go badly awry, as we shall see in Chapter 12, and the alternative of integrating them out does not give a unique answer (Problem 11.7). Thus the idea of multiplying $L(\theta)$ by a prior density and applying the simple recipe of Bayes' theorem starts to appear very attactive. Moreover, we see from (11.2) that given a particular prior $\pi(\theta)$, Bayesian inference for θ does conform to the likelihood principle, because any constants in $f(y \mid \theta)$ do not appear in the posterior density.

11.1.3 Prior information

Despite its conformity to the likelihood principle, inference based on Bayes' theorem has often been seen as controversial. This is not due to the result itself, which simply states mathematically how the probability density of one random variable changes when another has been observed, but because its use in statistical inference for θ requires the investigator to treat θ as a random variable, and to specify a prior density $\pi(\theta)$ separate from the data. A key issue is the interpretation and choice of π.

In some circumstances it is uncontroversial to treat θ as random. At one extreme the data at hand may be the latest in a stream of similar datasets, each having an underlying parameter that may be supposed to be drawn from a distribution. For example, an accountant may wish to estimate the level of errors in a company's books, θ, based on a sample of transactions that reveals y errors. It will be sensible to treat θ as randomly chosen from a density $\pi(\theta)$

of error rates based on experience with previous firms. Then inference on θ will use both y and $\pi(\theta)$. An example in the use of forensic evidence is when there is a close match between DNA profile data from the scene of a crime and a suspect. Then a database of prior profiles may help to establish whether DNA found at the scene of the crime could plausibly have come from someone else. In these applications the prior information has a frequentist basis, so new issues of interpretation do not arise.

Despite this, the London Court of Appeal (Regina vs. Adams, 1996, 1997) ruled that 'introducing Bayes' theorem ... into a criminal trial plunges the jury into inappropriate and unnecessary realms of complexity, deflecting them from their proper task'.

At the other end of the range of possibilities is the situation where the data are to be used to make subjective decisions such as 'should I bet on the outcome of this race?' Although likely to depend on how facts such as 'Flatfoot has not won a race this season' are viewed, both model and prior information here reflect a personal judgement. Here Bayes' theorem provides the mechanism for updating prior beliefs in the light of whatever data is available, but the inference is a personal assessment of the evidence and has no claim to objective force.

The debate arises when the prior information does not have a frequency interpretation, but the inference required is not purely personal. Many statisticians regard the information in data as being qualitatively different from their prior beliefs about model parameters, and hence find it unacceptable to use Bayes' theorem to combine the two. They argue that although the choice of probability model is usually a matter of individual judgement, that judgement can be checked by comparing the data and fitted model, while by definition prior information cannot be checked directly. To which a Bayesian might reply that the epistemological distinction between data, model, and prior is unclear, because collection of any data must be based on some prior belief, which will often include information about possible models and the likely values of their parameters. Furthermore Bayes' theorem provides a single recipe for inference about unknowns, while frequentist notions such as confidence intervals can violate what seem reasonable principles of inference. Much has been written on this, but we shall avoid getting embroiled, simply noting that in many situations the Bayesian approach is simpler and more direct than frequentist alternatives, and that when they can be compared, the inferences produced by Bayesian and good frequentist procedures are often rather similar, so that the practical consequences of choosing between them are usually not critical. When a frequentist inference differs strongly from any conceivable Bayesian one, it seems wise to pause and reflect awhile.

Whatever its interpretation, a prior must be specified in order for Bayesian analysis to proceed. We now consider aspects of this.

Conjugate densities

In Example 11.1 the combination of a beta prior density for a probability and the likelihood for several Bernoulli trials led to a beta posterior density.

Although too inflexible to encompass the range of prior knowledge that arises in applications, such conjugate combinations of prior and likelihood are useful because of their simple closed forms. They are closely tied to exponential family models.

Example 11.6 (Exponential family) Suppose that $y_1, \ldots, y_n$ is a random sample from the exponential family (5.12)

$$f(y \mid \omega) = \exp\{s(y)^{\mathrm{T}}\theta(\omega) - b(\omega)\} f_0(y),$$

so that in terms of $s = \sum s(y_j)$, the likelihood is proportional to

$$\exp\{s^{\mathrm{T}}\theta(\omega) - nb(\omega)\}. \tag{11.10}$$

If the prior density for ω depends on the quantities ξ and ν and has form

$$\pi(\omega) = \exp\{\xi^{\mathrm{T}}\theta(\omega) - \nu b(\omega) + c(\xi, \nu)\},$$

then the posterior density is proportional to

$$\exp\{(\xi + s)^{\mathrm{T}}\theta(\omega) - (\nu + n)b(\omega)\}.$$

Provided this is integrable the posterior density therefore must be

$$\pi(\omega \mid y) = \exp\{(\xi + s)^{\mathrm{T}}\theta(\omega) - (\nu + n)b(\omega) + c(\xi + s, \nu + n)\}.$$

Thus the prior parameters (ξ, ν) are updated to $(\xi + s, \nu + n)$ by the data. One interpretation of the *hyperparameters* ξ and ν is that the prior information is equivalent to ν prior observations summing to ξ.

For example, the Poisson density with mean ω has kernel $\exp(y \log \omega - \omega)$, so the conjugate prior must have kernel $\exp(\xi \log \omega - \nu\omega)$. For $\xi, \nu > 0$, this is proportional to the gamma density with mean ξ/ν, whose density is

$$\pi(\omega) = \frac{\nu^{\xi}\omega^{\xi-1}}{\Gamma(\xi)} e^{-\nu\omega}, \quad \omega > 0,$$

and which is therefore the conjugate prior for the Poisson mean. As the data update (ξ, ν) to $(\xi + s, \nu + n)$, the posterior density

$$\pi(\omega \mid y) = \frac{(\nu + n)^{\xi+s}\omega^{\xi+s-1}}{\Gamma(\xi + s)} e^{-(\nu+n)\omega}, \quad \omega > 0,$$

also has gamma form. ∎

Example 11.7 (Normal distribution) Let $y_1, \ldots, y_n$ be a normal random sample with mean μ and known variance σ^2. The likelihood is

$\overline{y}$ is the sample average $n^{-1}\sum y_j$.

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right\} \propto \exp\left(\mu\frac{n\overline{y}}{\sigma^2} - \frac{n}{\sigma^2}\tfrac{1}{2}\mu^2\right),$$

which is of form (11.10) with $s = n\overline{y}/\sigma^2$, $k = n/\sigma^2$, $a(\mu) = \mu$, and $\kappa(\mu) = \frac{1}{2}\mu^2$. Therefore the conjugate prior is proportional to

$$\exp\left(\mu\frac{\mu_0}{\tau^2} - \frac{1}{\tau^2}\tfrac{1}{2}\mu^2\right),$$

and must be the normal density with mean μ_0 and variance τ^2. The effect of the data is to update $(\mu_0\tau^{-2}, \tau^{-2})$ to $(\mu_0\tau^{-2} + s\sigma^{-2}, \tau^{-2} + n\sigma^{-2})$, so the posterior density for μ is normal with mean and variance

$$\frac{n\overline{y}/\sigma^2 + \mu_0/\tau^2}{n/\sigma^2 + 1/\tau^2}, \quad \frac{1}{n/\sigma^2 + 1/\tau^2}. \tag{11.11}$$

On writing the mean in (11.11) as

$$\frac{n\overline{y} + (\sigma^2/\tau^2)\mu_0}{n + \sigma^2/\tau^2},$$

we see that the prior injects information equivalent to σ^2/τ^2 observations with mean μ_0, and shrinks the sample average, $\overline{y}$, towards the prior mean by an amount that depends on the ratio of τ^2 to σ^2/n. As $n \to \infty$ or $\tau^2 \to \infty$, corresponding to increasing information in the data relative to the prior, the posterior density becomes normal with mean $\overline{y}$ and variance σ^2/n, so the effect of the prior withers away. As $\tau^2 \to 0$, corresponding to more definite prior knowledge, the posterior approaches the normal density with mean μ_0 and variance τ^2, which is the prior. ■

Conjugate priors are often too restrictive for expression of realistic prior information, but it is straightforward to establish that mixtures of conjugate densities are also conjugate, and this considerably broadens the class of priors with closed-form posterior densities (Problem 11.3).

Ignorance

Sometimes the prior density must express prior ignorance about a parameter. One reason for this may be the need for a 'baseline' analysis as a basis for discussion. Another is the belief that a non-informative prior will allow the data 'to speak for themselves', though it seems optimistic to think that they will spill their secrets without careful interrogation. Nevertheless it is important to weigh how much an inference depends on the prior compared to the data. One way to do this is to contrast inferences from a minimally informative prior with those from the prior actually used.

When θ has bounded support, as in Example 11.1, a uniform prior density, with $\pi(\theta) \propto 1$, seems an obvious choice. When the support of θ is unbounded, such a prior has infinite integral and so is *improper*. An improper prior may nevertheless lead to a proper posterior density. In Example 11.7, for example, we can represent complete ignorance about the prior value of μ by letting

$\tau^2 \to \infty$, in which case the prior is $\pi(\mu) \propto 1$ with support on the entire real line, and the posterior density of μ is normal with mean $\overline{y}$ and variance σ^2/n, which is proper. Prior ignorance about σ in models where the density of the data is of form $\sigma^{-1}g(u/\sigma)$, $u > 0$, $\sigma > 0$, is usually represented by the improper prior $\pi(\sigma) \propto \sigma^{-1}$, $\sigma > 0$. Non-informative priors of this sort exist for more general situations, but there is a fundamental difficulty in representing ignorance in a way that is independent both of the data to be collected and the parametrization of the model (Problem 11.4). The key question is: ignorance about what? The following classic example illustrates this.

Example 11.8 (Bernoulli probability) The probability of success in a Bernoulli trial lies in the interval $[0, 1]$, so if we are completely ignorant of its true value, the obvious prior to use is uniform on the unit interval: $\pi(\theta) = 1$, $0 \le \theta \le 1$. But if we are completely ignorant of θ, we are also completely ignorant of $\psi = \log\{\theta/(1-\theta)\}$, which takes values in the real line. The density implied for ψ by the uniform prior for θ is

$$\pi(\psi) = \pi\{\psi(\theta)\} \times \left|\frac{d\theta}{d\psi}\right| = \frac{e^{\psi}}{(1+e^{\psi})^2}, \quad -\infty < \psi < \infty :$$

the standard logistic density. Far from expressing ignorance about ψ, this density asserts that the prior probability of $|\psi| < 3$ is about 0.9. ■

Jeffreys priors

Apparent paradoxes like that of Example 11.8 led to a widespread rejection of Bayesian inference in the early twentieth century. The key difficulty is that the representation of ignorance is not invariant under reparametrization. A solution to this is to seek invariant priors. For scalar θ the best-known of these is the *Jeffreys prior*

$$\pi(\theta) \propto |i(\theta)|^{1/2}, \tag{11.12}$$

where $i(\theta) = -\mathrm{E}\{d^2\ell(\theta)/d\theta^2\}$ is the expected information for θ based on the log likelihood $\ell(\theta)$; $i(\theta)$ is positive in a regular statistical model. For a smooth reparametrization $\theta = \theta(\psi)$ in terms of ψ, the expected information for ψ is

$$i(\psi) = -\mathrm{E}\left[\frac{d^2\ell\{\theta(\psi)\}}{d\psi^2}\right] = -\mathrm{E}\left\{\frac{d^2\ell(\theta)}{d\theta^2}\right\} \times \left|\frac{d\theta}{d\psi}\right|^2 = i(\theta) \times \left|\frac{d\theta}{d\psi}\right|^2.$$

Consequently $|i(\theta)|^{1/2}d\theta = |i(\psi)|^{1/2}d\psi$: with the choice (11.12), prior information does behave consistently under reparametrization; furthermore such priors give widely-accepted solutions in some standard problems. When θ is vector, $|i(\theta)|$ is taken to be the determinant of $i(\theta)$.

This prior was initially proposed with the aim of giving an 'objective' basis for inference, but after further paradoxes emerged its use was suggested

Sir Harold Jeffreys (1891–1989) studied first in Newcastle and then in Cambridge, where he remained for the rest of his life, becoming Plumian Professor of Astronomy. During World War I he worked in the Cavendish Laboratory, and thereafter studied and taught hydrodynamics and geophysics, being the first to claim that the core of the earth is liquid. In an important series of books he championed objective Bayesian inference long before it became popular (Jeffreys, 1961), and also wrote important works on geophysics and mathematical physics. His unassuming character inspired deep affection.

for convenience, a matter of scientific convention rather than as a logically unassailable expression of ignorance about the parameter.

Example 11.9 (Bernoulli probability) The log likelihood for a single Bernoulli trial with success probability θ is $y \log \theta + (1-y)\log(1-\theta)$, and the Fisher information is $i(\theta) = \theta^{-1}(1-\theta)^{-1}$. Thus the Jeffreys prior is proportional to $\theta^{-1/2}(1-\theta)^{-1/2}$, and so equals the beta density (11.3) shown in the top left panel of Figure 5.4, which while proper does not look uninformative. It can be interpreted as carrying information equivalent to one trial, in which one-half of a success was observed. As the prior information for n independent trials is $ni(\theta)$, the Jeffreys prior is the same because the constant of proportionality is independent of θ. ■

Example 11.10 (Location-scale model) Suppose that $y_1, \ldots, y_n$ is a random sample from a location model $f(y;\eta) = g(y-\eta)$, for real y and η. Then the log likelihood is $\ell(\eta) = \sum \log g(y_j - \eta)$, so

$$i(\eta) = -n \int_{-\infty}^{\infty} \frac{d^2 \log g(y-\eta)}{d\eta^2} g(y-\eta)\, dy.$$

The substitution $u = y - \eta$ shows that $i(\eta)$ is independent of η, and therefore the Jeffreys prior is the constant non-informative prior $\pi(\eta) \propto 1$ for all η.

A modification of this argument (Problem 11.2) shows that the Jeffreys prior for $f(y;\tau) = \tau^{-1} g(y/\tau)$, $y, \tau > 0$, is $\pi(\tau) \propto \tau^{-1}$, which is also widely accepted as non-informative. Both $\pi(\tau)$ and $\pi(\eta)$ are improper.

A difficulty with this approach appears when we consider the location-scale model $f(y;\eta,\tau) = \tau^{-1} g\{(y-\eta)/\tau\}$. Its information matrix has form $i(\eta,\tau) = n\tau^{-2}A$, where the 2×2 matrix A is free of parameters, so $\pi(\eta,\tau) = |i(\eta,\tau)|^{1/2} \propto \tau^{-2}$. This does not equal the prior τ^{-1} arising from taking independent Jeffreys priors for η and τ separately.

The approach is here unsatisfactory because the prior τ^{-2} is not widely accepted as a non-informative statement of uncertainty about τ. More generally this example shows that a non-informative inference for a parameter of interest, η, say, may depend on the model in which η is embedded, in the sense that the inference may depend on the prior chosen for nuisance parameters, even when these are *a priori* independent of η. ■

Jeffreys' general solution to the difficulty raised in Example 11.10 was to treat location parameters as fixed when computing $i(\theta)$. Let $\theta = (\mu_1, \ldots, \mu_p, \psi)$, where the μ_r are location parameters and ψ contains all other parameters in the problem. Then the prior he recommended is

$$\pi(\mu_1, \ldots, \mu_p, \psi) \propto \left| \mathrm{E}\left\{ -\frac{\partial^2 \ell(\mu_1, \ldots, \mu_p, \psi)}{\partial \psi \partial \psi^{\mathrm{T}}} \right\} \right|^{1/2},$$

which produces $\pi(\theta) \propto \tau^{-1}$ in the location-scale model.

Numerous other approaches to representing prior ignorance have been proposed, based for example on notions of invariance, of minimal information, or of matching the coverage of Bayesian and frequentist confidence intervals. To a large extent these are regarded as useful to the extent that they yield Jeffreys priors, and we shall not consider them in detail. To be more explicit about links with the frequentist approach, however, note that if a uniform prior is taken in (11.11), corresponding to $\tau \to \infty$, and we define $\mathcal{A}_y$ to be the interval with limits $\overline{y} \pm z_\alpha n^{-1/2}\sigma$, then the posterior probability $\Pr(\theta \in \mathcal{A}_y \mid y) = 1 - 2\alpha$. Thus $\mathcal{A}_y$ has posterior coverage $(1 - 2\alpha)$. But A_y also has the same coverage for any fixed θ unconditional on y, so the uniform prior yields an interval justifiable from both Bayesian and frequentist viewpoints. Exact results such as this are unobtainable in more general settings, but nonetheless it can be helpful to consider the extent to which Bayesian and frequentist procedures agree.

Some further aspects of Jeffreys priors are outlined in Problem 11.4.

Exercises 11.1

1 In Example 11.3, calculate the predictive probability for k future heads out of m tosses based on r heads observed in n tosses, using a beta prior density.

2 Show that the limits of an unconditional confidence interval of level $(1 - 2\alpha)$ in Example 11.4 involve the solutions to the equation

$$\tfrac{1}{2}\Phi\{(y-\theta)/10\} + \tfrac{1}{2}\Phi(y-\theta) = \alpha, 1-\alpha.$$

Hence justify the approximate 0.95 interval given in the example.

3 (a) Let $y_1, \ldots, y_n$ be a Poisson random sample with mean θ, and suppose that the prior density for θ is gamma,

$$\pi(\theta) = g(\theta; \alpha, \lambda) = \frac{\lambda^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)} \exp(-\lambda\theta), \quad \theta > 0,\ \lambda, \alpha > 0.$$

Show that the posterior density of θ is $g(\theta; \alpha + \sum y_j, \lambda + n)$, and find conditions under which the posterior density remains proper as $\alpha \downarrow 0$ even though the prior density becomes improper in the limit.
(b) Show that $\int \theta g(\theta; \alpha, \lambda)\, d\theta = \alpha/\lambda$. Find the prior and posterior means $\mathrm{E}(\theta)$ and $\mathrm{E}(\theta \mid y)$, and hence give an interpretation of the prior parameters.
(c) Let Z be a new Poisson variable independent of $Y_1, \ldots, Y_n$, also with mean θ. Find its posterior predictive density. To what density does this converge as $n \to \infty$? Does this make sense?

4 How would you express prior ignorance about an angle? About the position of a star in the firmament?

5 If $Y_{ij} \sim N(\mu_i, \sigma^2)$ independently for $i = 1, \ldots, k$ and $j = 1, \ldots, m$, show that the Jeffreys prior for $\mu_1, \ldots, \mu_k, \sigma$ equals $\sigma^{-(k+1)}$. Discuss the form of posterior inferences on σ^2 when $m = 2$. Is this prior reasonable? If not, suggest a better alternative.

Table 11.1
Conjugate prior densities for exponential family samling distributions.

$f(y \mid \theta)$	Parameter	Prior
Binomial	success probability	beta
Poisson	mean	gamma
Exponential	mean	gamma
Normal	mean (known variance)	normal
Normal	variance (known mean)	inverse gamma
Multinomial	probabilities	Dirichlet

6 According to the *principle of insufficient reason* probabilities should be ascribed uniformly to finite sets unless there is some definite reason to do otherwise. Thus the most natural way to express prior ignorance for a parameter θ that inhabits a finite parameter space $\theta_1, \ldots, \theta_k$ is to set $\pi(\theta_1) = \cdots = \pi(\theta_k) = 1/k$. Let $\pi_i = \pi(\theta_i)$.
Consider a parameter space $\{\theta_1, \theta_2\}$, where θ_1 denotes that there is life in orbit around the star Sirius and θ_2 that there is not. Can you see any reason not to take $\pi_1 = \pi_2 = 1/2$?
Now consider the parameter space $\{\omega_1, \omega_2, \omega_3\}$, where ω_1, ω_2, and ω_3 denote the events that there is life around Sirius, that there are planets but no life, and that there are no planets. With this parameter space the principle of insufficient reason gives Pr(life around Sirius) = 1/3.
Discuss this partitioning paradox. What solutions do you see?
(Schafer, 1976, pp. 23–24)

7 Compute the prior and posterior means and variances for exponential family data with the conjugate prior distribution, and discuss their interpretation.

8 Use Example 11.6 to verify the contents of Table 11.1.

9 Let θ be a randomly chosen physical constant. Such constants are measured on an arbitrary scale, so transformations from θ to $\psi = c\theta$ for some constant c should leave the density $\pi(\theta)$ of θ unchanged. Show that this entails $\pi(c\theta) = c^{-1}\pi(\theta)$ for all $c, \theta > 0$, and deduce that $\pi(\theta) \propto \theta^{-1}$.
Let $\tilde{\theta}$ be the first significant digit of θ in some arbitrary units. Show that

$$\Pr(\tilde{\theta} = d) \propto \int_{d10^a}^{(d+1)10^a} u^{-1}\, du, \quad d = 1, \ldots, 9,$$

and hence verify that $\Pr(\tilde{\theta} = d) = \log_{10}(1 + d^{-1})$. Check whether some set of physical 'constants' (e.g. sizes of countries or of lakes) fits this distribution.

11.2 Inference

11.2.1 Posterior summaries

If the information regarding θ is contained in its posterior density given the data y, $\pi(\theta \mid y)$, how do we get at it? In principle this is easy: we simply use the posterior density to calculate the probability of any event of interest. But some summary quantities may be useful. For example, if $\theta = (\psi, \lambda)$ is a

vector, and we are interested in ψ, the *marginal posterior density*

$$\pi(\psi \mid y) = \int \pi(\psi, \lambda \mid y)\, d\lambda,$$

contains the marginal information in the model and prior concerning ψ. It is most useful when ψ has dimension one or two, in which case it can be plotted. It condenses further to moments, quantiles, or the mode of $\pi(\psi \mid y)$.

Normal approximation

One simple approximate summary of a unimodal posterior rests on quadratic series expansion of the log posterior density, analogous to expansion of the log likelihood. In terms of $\tilde{\ell}(\theta) = \log L(\theta) + \log \pi(\theta)$ and the posterior mode $\tilde{\theta}$, we have

$$\begin{aligned} \tilde{\ell}(\theta) &\doteq \tilde{\ell}(\tilde{\theta}) + (\theta - \tilde{\theta})^{\mathrm{T}} \frac{\partial \tilde{\ell}(\tilde{\theta})}{\partial \theta} + \tfrac{1}{2}(\theta - \tilde{\theta})^{\mathrm{T}} \frac{\partial^2 \tilde{\ell}(\tilde{\theta})}{\partial \theta \partial \theta^{\mathrm{T}}} (\theta - \tilde{\theta}) \\ &= \tilde{\ell}(\tilde{\theta}) - \tfrac{1}{2}(\theta - \tilde{\theta})^{\mathrm{T}} \tilde{J}(\tilde{\theta})(\theta - \tilde{\theta}), \end{aligned}$$

provided the mode lies inside the parameter space. Here $\tilde{J}(\theta)$ is the second derivative matrix of $-\tilde{\ell}(\theta)$. This expansion corresponds to a posterior multivariate normal density for θ, with mean $\tilde{\theta}$ and variance matrix $\tilde{J}(\tilde{\theta})^{-1}$, based on which an equitailed $(1 - 2\alpha)$ confidence interval for the rth component θ_r of θ is $\tilde{\theta}_r \pm z_\alpha \tilde{v}_{rr}^{1/2}$, where $\tilde{v}_{rr}$ is the rth diagonal element of $\tilde{J}(\tilde{\theta})^{-1}$.

In large samples the log likelihood contribution is typically much greater than that from the prior, so $\tilde{\theta}$ and $\tilde{J}(\tilde{\theta})$ are essentially indistinguishable from the maximum likelihood estimate $\widehat{\theta}$ and observed information $J(\widehat{\theta})$. Thus likelihood-based confidence intervals may be interpreted as giving approximate Bayesian inferences, if the sample is large. This approximation will usually be better if applied to the marginal posterior of a low-dimensional subset of θ, because of the averaging effect of integration over the other parameters. The same caveats apply when using this approximation as to use of normal approximations for the maximum likelihood estimator; in particular, it may be more suitable for a transformed parameter. We describe a more refined approach in Section 11.3.1.

Other distributions may be used to approximate posterior densities, for example by matching first and second moments.

Posterior confidence sets

The mean and mode of the posterior density are point summaries of $\pi(\theta \mid y)$, but confidence regions or intervals are usually more useful. The Bayesian analogue of a $(1 - 2\alpha)$ confidence interval is a $(1 - 2\alpha)$ *credible set*, defined to be a set, C, of values of θ, whose posterior probability content is at least

Table 11.2
Mortality rates r/m from cardiac surgery in 12 hospitals (Spiegelhalter *et al.*, 1996b, p. 15). Shown are the numbers of deaths r out of m operations.

A	0/47	*B*	18/148	*C*	8/119	*D*	46/810	*E*	8/211	*F*	13/196
G	9/148	*H*	31/215	*I*	14/207	*J*	8/97	*K*	29/256	*L*	24/360

$1 - 2\alpha$. When θ is continuous this is

$$1 - 2\alpha = \Pr(\theta \in C \mid y) = \int_C \pi(\theta \mid y)\, d\theta.$$

When θ is discrete, the integral is replaced by $\sum_{\theta \in C} \pi(\theta \mid y)$. For scalar θ, such a set is equi-tailed if it has form (θ_L, θ_U), where θ_L and θ_U are the posterior α and $1 - \alpha$ quantiles of θ, that is, $\Pr(\theta < \theta_L \mid y) = \Pr(\theta > \theta_U \mid y) = \alpha$.

Often C is chosen so that the posterior density for any θ in C is higher than for any θ not in C. That is, if $\theta \in C$, $\pi(\theta \mid y) \geq \pi(\theta' \mid y)$ for any $\theta' \notin C$. Such a region is called a *highest posterior density credible set*, or more concisely a *HPD credible set*.

Example 11.11 (Cardiac surgery data) Table 11.2 contains data on the mortality levels for cardiac surgery on babies at 12 hospitals. A simple model treats the number of deaths r as binomial with mortality rate θ and denominator m. At hospital A, for example, $m = 47$ and $r = 0$, giving maximum likelihood estimate $\widehat{\theta}_A = 0/47 = 0$, but it seems too optimistic to suppose that θ_A could be so small when the other rates are evidently larger. If we take a beta prior density with $a = b = 1$, the posterior density is beta with parameters $a+r = 1$ and $b+m-r = 48$, as shown in the left panel of Figure 11.1. The 0.95 HPD credible interval is $(0, 6.05)\%$, while the equitailed credible interval uses the 0.025 and 0.975 quantiles of $\pi(\theta_A \mid y)$ and is $(0.05, 7.40)\%$.

The right panel of Figure 11.1 shows the posterior density for the overall mortality rate θ, obtaining by merging all the data, giving $r = 208$ deaths in $m = 2814$ operations. Here the prior parameters a and b have essentially no effect on the posterior, and hence

$$\tilde{\theta} = \frac{a + r - 1}{a + b + m - 2} \doteq \frac{r}{m}, \quad \tilde{J}(\tilde{\theta})^{-1} = \frac{(a + r - 1)(b + m - r - 1)}{(a + b + m - 2)^3} \doteq \frac{r(m - r)}{m^3}.$$

The figure shows the corresponding normal approximation to $\pi(\theta \mid y)$. Evidently inferences from exact and approximate posterior densities will be equivalent for practical purposes.

Both separate and pooled analyses of mortality rates seem unsatisfactory, because although some variation among hospitals is plausible they are likely also to have elements in common. Example 11.26 describes an approach intermediate between those used here. ■

Example 11.12 (Normal distribution) Consider a normal random sample $y_1, \ldots, y_n$ with mean μ and variance σ^2 both unknown. We shall give

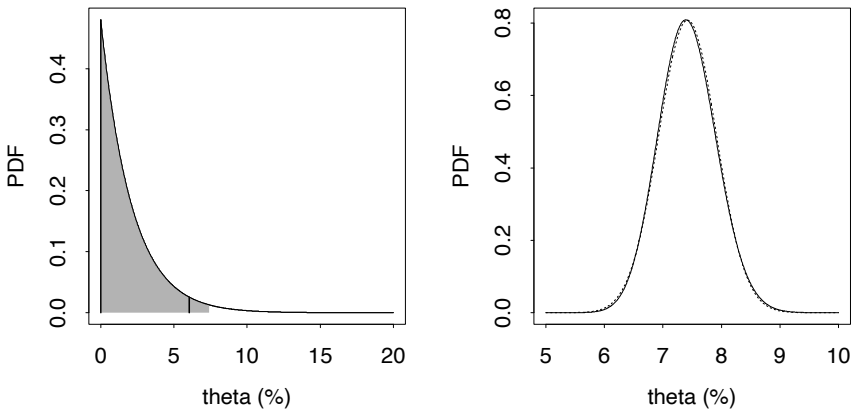


Figure 11.1
Cardiac surgery data. Left panel: posterior density for θ_A showing boundaries of 0.95 highest posterior density interval (vertical lines) and region between posterior 0.025 and 0.975 quantiles of $\pi(\theta_A \mid y)$ (shaded). Right panel: exact posterior beta density for overall mortality rate θ (solid) and normal approximation (dots).

them independent prior densities. As the posterior for (μ, σ^2) depends on y only through the minimal sufficient statistic $(\overline{y}, s^2)$, we have

$\overline{y} = n^{-1} \sum y_j$ and $s^2 = (n-1)^{-1} \sum (y_j - \overline{y})^2$ are the sample average and variance.

$$\begin{aligned} \pi(\mu, \sigma^2 \mid \overline{y}, s^2) &\propto f(\overline{y}, s^2 \mid \mu, \sigma^2)\pi(\mu, \sigma^2) \\ &= f(\overline{y} \mid \mu, \sigma^2) f(s^2 \mid \mu, \sigma^2)\pi(\mu, \sigma^2) \\ &= f(\overline{y} \mid \mu, \sigma^2) f(s^2 \mid \sigma^2)\pi(\mu)\pi(\sigma^2) \\ &\propto \pi(\mu \mid \overline{y}, \sigma^2) f(s^2 \mid \sigma^2)\pi(\sigma^2), \end{aligned} \tag{11.13}$$

where the first step follows from Bayes' theorem, the second from the conditional independence of $\overline{y}$ and σ^2 given μ and σ^2, the third from the prior independence of μ and σ^2 and the independence of s^2 and μ, and the fourth on using Bayes' theorem to get the posterior density for μ conditional on $\overline{y}$ and σ^2. Integration of (11.13) with respect to μ shows that $\pi(\sigma^2 \mid \overline{y}, s^2) \propto f(s^2 \mid \sigma^2)\pi(\sigma^2)$: the marginal posterior density of σ^2 depends only on s^2. However, as σ^2 appears in all three terms, integration of (11.13) with respect to σ^2 shows that the marginal posterior for μ depends on both $\overline{y}$ and s^2.

Let us use the improper priors $\pi(\mu) \propto 1$, $\pi(\sigma^2) \propto \sigma^{-2}$. Example 11.7 shows that the posterior density for μ when σ^2 is known is $N(\overline{y}, \sigma^2/n)$. Conditional on σ^2, the distribution of $(n-1)s^2$ is $\sigma^2 \chi^2_{n-1}$, so our choice of prior gives

$$\begin{aligned} \pi(\sigma^2 \mid s^2) &\propto \pi(\sigma^2) f(s^2 \mid \sigma^2) \\ &\propto (\sigma^2)^{-1} (\sigma^2)^{-(n-1)/2} \exp\left\{-\tfrac{1}{2}(n-1)s^2/\sigma^2\right\}, \quad \sigma^2 > 0. \end{aligned}$$

Thus the marginal posterior density of σ^2 is *inverse gamma*,

$$\frac{\beta^\alpha}{\Gamma(\alpha) x^{\alpha+1}} \exp(-\beta/x), \quad x > 0, \quad \alpha, \beta > 0, \tag{11.14}$$

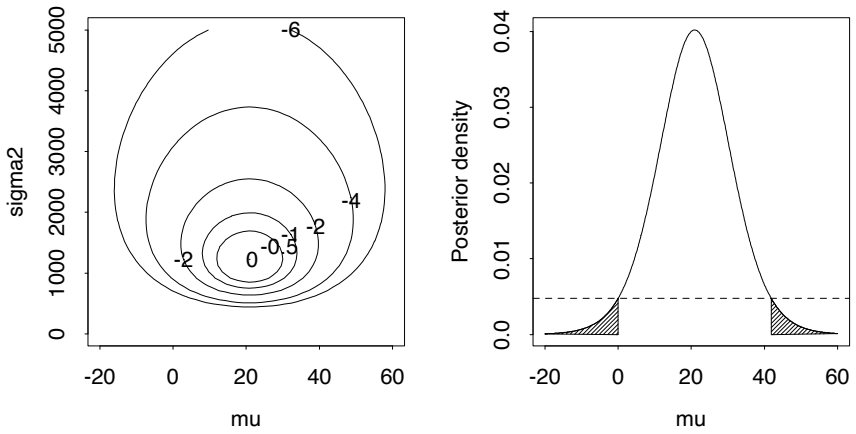


Figure 11.2
Posterior densities of (μ, σ^2) of normal model for maize data. Left: contours of the normalized log joint posterior density. Right: marginal posterior density for μ, showing 95% HPD credible set, which is the set of values of μ whose values of the posterior density $\pi(\mu \mid y)$ lie above the dashed line. The shaded region has area 0.05.

with $x = \sigma^2$, $\alpha = \frac{1}{2}(n-1)$ and $\beta = \frac{1}{2}(n-1)s^2$; a useful shorthand for (11.14) is $IG(\alpha, \beta)$. Its mean and variance are $\beta/(\alpha-1)$ and $\beta^2/\{(\alpha-1)^2(\alpha-2)\}$, provided that $\alpha > 2$. Equivalently, the posterior distribution of σ^2 given s^2 is that of $(n-1)s^2/V$, where $V \sim \chi^2_{n-1}$. The joint posterior density for (μ, σ^2),

$$\pi(\mu, \sigma^2 \mid \overline{y}, s^2) \propto \pi(\mu \mid \overline{y}, \sigma^2)\pi(\sigma^2 \mid s^2).$$

is proportional to

$$(\sigma^2)^{-1/2} \exp\left\{-\frac{n}{2\sigma^2}(\mu - \overline{y})^2\right\} \times (\sigma^2)^{-(n-1)/2-1} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}, \quad (11.15)$$

integration of which over σ^2 yields the marginal posterior density for μ,

$$\pi(\mu \mid \overline{y}, s^2) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \left\{\frac{n}{(n-1)s^2\pi}\right\}^{1/2} \left\{1 + \frac{n(\mu - \overline{y})^2}{(n-1)s^2}\right\}^{-n/2}.$$

Therefore $n^{1/2}(\mu - \overline{y})/s \sim t_{n-1}$ *a posteriori*. The corresponding frequentist result treats $\overline{y}$ and s^2 as random and μ as fixed; here the random variable is μ, with $\overline{y}$ and s^2 regarded as constants.

Figure 11.2 shows posterior densities for μ and σ^2 based on the height differences for the 15 pairs of plants in Table 1.1; here $\overline{y} = 20.93$ and $s^2 = 1424.64$. Evidently the posterior densities are not independent. While the HPD credible set for μ is equi-tailed, that for σ^2 is not. ■

A credible set may contain the same values of θ as a confidence interval, but its interpretation is different. In the Bayesian framework the data are regarded as fixed and the parameter as random, so the endpoints of the credible set are fixed and the probability statement concerns the parameter, regarded

as a random variable. The frequentist approach treats the parameter as an unknown constant and the confidence interval endpoints as random variables; the probability statement concerns their behaviour in repeated sampling from the model.

11.2.2 Bayes factors

The frequentist approach to hypothesis testing compares a null hypothesis H_0 with an alternative H_1 through a test statistic T that tends to be larger under H_1 than under H_0, and rejects H_0 for small values of the significance probability $p_{\text{obs}} = \Pr_0(T \geq t_{\text{obs}})$, where t_{obs} is the value of T actually observed and the probability is computed as if H_0 were true.

The Bayesian approach attaches prior probabilities to the models corresponding to H_0 and H_1 and compares their posterior probabilities

$$\Pr(H_i \mid y) = \frac{\Pr(y \mid H_i)\Pr(H_i)}{\Pr(y \mid H_0)\Pr(H_0) + \Pr(y \mid H_1)\Pr(H_1)}, \quad i = 0, 1.$$

An obvious distinction between this and the frequentist approach is that $\Pr(H_0 \mid y)$ is the probability of H_0 conditional on the data, whereas the P-value may not be interpreted in this way. In Bayesian settings increasing amounts of data may lead to increasing support for one hypothesis relative to the alternatives. This differs from the frequentist approach, where non-rejection of H_0 does not indicate increasing support for it in large samples. A further important difference is that the P-value does not depend on the particular alternative H_1 under discussion. Indeed, whereas frequentist testing does not require H_1 to be fully specified, this is essential for Bayesian testing, which is in this sense more restrictive.

For some purposes it is valuable to use the odds in favour of H_1,

$$\frac{\Pr(H_1 \mid y)}{\Pr(H_0 \mid y)} = \frac{\Pr(y \mid H_1)}{\Pr(y \mid H_0)} \times \frac{\Pr(H_1)}{\Pr(H_0)}. \tag{11.16}$$

The change in prior to posterior odds for H_1 relative to H_0 depends on data only through the *Bayes factor*

$$B_{10} = \frac{\Pr(y \mid H_1)}{\Pr(y \mid H_0)}. \tag{11.17}$$

Thus analogous to the updating rule for inference on θ, we update evidence comparing the models by the rule *posterior odds = Bayes factor × prior odds.*

The simplest situation is when both hypotheses are simple, in which case B_{10} equals the likelihood ratio in favour of H_1. Usually, however, both hypotheses involve parameters, say θ_0 and θ_1, and

$$\Pr(y \mid H_i) = \int f(y \mid H_i, \theta_i)\pi(\theta_i \mid H_i)\, d\theta_i, \quad i = 0, 1,$$

Table 11.3 Interpretation of Bayes factor B_{10} in favour of H_1 over H_0. Since $B_{10} = B_{01}^{-1}$, negating the values of $2 \log B_{10}$ gives the evidence against H_1.

B_{10}	$2 \log B_{10}$	Evidence against H_0
1–3	0–2	Hardly worth a mention
3–20	2–6	Positive
20–150	6–10	Strong
> 150	> 10	Very strong

where $\pi(\theta_i \mid H_i)$ is the prior for θ_i under H_i. In this case the Bayes factor is a ratio of weighted likelihoods. By analogy with the likelihood ratio statistic, the quantity $2 \log B_{10}$ is often used to summarize the evidence for H_1 compared to H_0, with the rough interpretation shown in Table 11.3. This contrasts with the interpretation of a likelihood ratio statistic, whose null χ^2 distribution for nested models would depend on the difference in their degrees of freedom. The log Bayes factor $\log B_{10}$ is sometimes called the *weight of evidence*.

Example 11.13 (HUS data) Example 4.40 introduced data on the numbers of cases of haemolytic uraemic syndrome (HUS) treated at a clinic in Birmingham from 1970 to 1989. The data suggest a sharp rise in incidence around 1980. In that example it was supposed that the annual counts $y_1, \ldots, y_n$ are realizations of independent Poisson variables with means $\mathrm{E}(Y_j) = \lambda_1$ for $j = 1, \ldots, \tau$ and $\mathrm{E}(Y_j) = \lambda_2$ for $j = \tau + 1, \ldots, n$. Here the changepoint τ can take values $1, \ldots, n-1$.

Suppose that our baseline model H_0 is that $\lambda_1 = \lambda_2 = \lambda$, that is, no change, and consider the alternative H_τ of change after year τ. Under H_τ we suppose that λ_1 and λ_2 have independent gamma prior densities with parameters γ and δ. This density has mean γ/δ and variance γ/δ^2. Then $\Pr(y \mid H_\tau)$ equals

$$\int_0^\infty \prod_{j=1}^{\tau} \frac{\lambda_1^{y_j}}{y_j!} e^{-\lambda_1} \times \frac{\delta^\gamma \lambda_1^{\gamma-1}}{\Gamma(\gamma)} e^{-\delta\lambda_1}\, d\lambda_1 \int_0^\infty \prod_{j=\tau+1}^{n} \frac{\lambda_2^{y_j}}{y_j!} e^{-\lambda_2} \times \frac{\delta^\gamma \lambda_2^{\gamma-1}}{\Gamma(\gamma)} e^{-\delta\lambda_2}\, d\lambda_2,$$

or equivalently

$$\frac{\delta^{2\gamma}}{\Gamma(\gamma)^2 \prod_{j=1}^n y_j!} \frac{\Gamma(\gamma + s_\tau)\, \Gamma(\gamma + s_n - s_\tau)}{(\delta + \tau)^{\gamma + s_\tau} (\delta + n - \tau)^{\gamma + s_n - s_\tau}},$$

where $s_\tau = y_1 + \cdots + y_\tau$.

Under H_0 we assume that λ also has the gamma density with parameters γ and δ. Then the Bayes factor for a changepoint in year τ is

$$B_{\tau 0} = \frac{\Gamma(\gamma + s_\tau)\, \Gamma(\gamma + s_n - s_\tau)\, \delta^\gamma (\delta + n)^{\gamma + s_n}}{\Gamma(\gamma)\Gamma(\gamma + s_n)(\delta + \tau)^{\gamma + s_\tau} (\delta + n - \tau)^{\gamma + s_n - s_\tau}}, \quad \tau = 1, \ldots, n-1.$$

For completeness we set $B_{n0} = 1$.

Table 11.4 gives $2 \log B_{\tau 0}$ for $\tau = 1, \ldots, 19$, for values of γ and δ such that

Table 11.4 Bayes factors for comparison of model of change in Poisson parameter after τ years, H_τ, with model of no change H_0 for HUS data y. There is very strong evidence of change in any year from 1976–86.

	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
y	1	5	3	2	2	1	0	0	2	1
$2\log B_{\tau 0}$, $\gamma=\delta=1$	4.9	−0.5	0.6	3.9	7.5	13	24	35	41	51
$2\log B_{\tau 0}$, $\gamma=\delta=0.01$	−1.3	−5.9	−4.5	−1.0	3.0	9.7	20	32	39	51
$2\log B_{\tau 0}$, $\gamma=\delta=0.0001$	−10	−15	−14	−10	−6.1	0.6	11	23	30	42

	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
y	1	7	11	4	7	10	16	16	9	15
$2\log B_{\tau 0}$, $\gamma=\delta=1$	63	55	38	42	40	31	11	−2.9	−5.3	0
$2\log B_{\tau 0}$, $\gamma=\delta=0.01$	64	57	40	47	46	38	18	1.8	1.2	0
$2\log B_{\tau 0}$, $\gamma=\delta=0.0001$	55	48	31	38	37	29	8.8	−7.1	−7.7	0

the prior density for λ has unit mean and variances respectively 1, 10^2, 10^4, corresponding to increasing prior uncertainty. Negative values of $2\log B_{\tau 0}$ correspond to evidence in favour of H_0. There is very strong evidence for change in any year from 1976 to 1986, but the most plausible changepoint is just after 1980. The evidence for change is overwhelming for all the priors chosen. See Practical 11.6. ■

Example 11.14 (Forensic evidence) The following situation can arise when forensic evidence is used in criminal trials: material y found on a suspect is similar to other material, x, at the scene of the crime, and it is desired to know how this affects our view of the case. For simplicity we shall suppose that if x and y come from the same source, the suspect is guilty, an event we shall denote by G. Let E denote any other evidence. Then the odds of guilt, conditional on E and the data, are

$$\begin{aligned}\frac{\Pr(G \mid x,y,E)}{\Pr(\overline{G} \mid x,y,E)} &= \frac{\Pr(x,y \mid G,E)}{\Pr(x,y \mid \overline{G},E)}\frac{\Pr(G \mid E)}{\Pr(\overline{G} \mid E)} \\ &= \frac{\Pr(x,y \mid G)}{\Pr(x \mid \overline{G})\Pr(y \mid \overline{G})} \times \frac{\Pr(G \mid E)}{\Pr(\overline{G} \mid E)}, \qquad (11.18)\end{aligned}$$

where we have supposed that x and y are independent of E, and that they are independent given the event $\overline{G}$ that the suspect is not guilty. The first ratio on the right of (11.18) is the Bayes factor due to the forensic evidence.

Let y and x represent single measurements on the refractive index of glass fragments found on a suspect and at the scene of a burglary. We model the

corresponding random variables as

$$X \mid \theta_1 \sim N(\theta_1, \sigma^2), \quad Y \mid \theta_2 \sim N(\theta_2, \sigma^2),$$

where θ_1 and θ_2 are the true refractive indexes and σ^2 is known. If the suspect is guilty, then $\theta_1 = \theta_2 = \theta$, say. We model natural variation among refractive indexes by supposing that θ is drawn from a population of types of glass whose true refractive indexes are $N(\mu, \tau^2)$, where μ and $\tau^2 \gg \sigma^2$ both known. Thus under G,

$$X, Y \mid \theta \overset{\text{iid}}{\sim} N(\theta, \sigma^2), \quad \theta \sim N(\mu, \tau^2),$$

while under $\overline{G}$, the true indexes θ_1 and θ_2 are independent, giving

$$X \mid \theta_1 \sim N(\theta_1, \sigma^2), \quad Y \mid \theta_2 \sim N(\theta_2, \sigma^2), \quad \theta_1, \theta_2 \overset{\text{iid}}{\sim} N(\mu, \tau^2).$$

It turns out to be easier to work in terms of transformed observations $u = x - y$ and $z = \frac{1}{2}(x + y)$, and to write the corresponding random variables as

$$U = \theta_1 - \theta_2 + \varepsilon_1 - \varepsilon_2, \quad Z = \tfrac{1}{2}(\theta_1 + \theta_2 + \varepsilon_1 + \varepsilon_2), \quad \varepsilon_1, \varepsilon_2 \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Then U and Z are independent and normal both conditionally on θ_1, θ_2 and unconditionally. Under G, $\theta_1 = \theta_2$, so

$$U \sim N(0, 2\sigma^2), \quad Z \sim N(\mu, \tau^2 + \tfrac{1}{2}\sigma^2),$$

while under $\overline{G}$,

$$U \sim N(0, 2\tau^2 + 2\sigma^2), \quad Z \sim N(\mu, \tfrac{1}{2}\tau^2 + \tfrac{1}{2}\sigma^2).$$

As the Jacobian of the transform from (x, y) to (u, z) equals one under both G and $\overline{G}$, and $\tau^2 \gg \sigma^2$, the Bayes factor is roughly

$$\frac{(2\sigma^2)^{-1/2} \exp\{-u^2/(4\sigma^2)\} (\tau^2)^{-1/2} \exp\{-(z-\mu)^2/(4\tau^2)\}}{(2\tau^2)^{-1/2} \exp\{-u^2/(4\tau^2)\} (\tau^2/2)^{-1/2} \exp\{-(z-\mu)^2/\tau^2\}},$$

which equals

$$\left(\frac{\tau^2}{2\sigma^2}\right)^{1/2} \times \exp\left(-\frac{u^2}{4\sigma^2}\right) \times \exp\left\{\frac{(z-\mu)^2}{2\tau^2}\right\}.$$

The interpretation of the second term is that if the difference $u = x - y$ is large relative to its variance $2\sigma^2$, there is strong evidence that θ_1 and θ_2 differ, and this favours $\overline{G}$. The third term measures how typical x and y are. If $z = \frac{1}{2}(x+y)$ is far from its mean, μ, compared to its variance $\frac{1}{2}\tau^2$ under $\overline{G}$, both x and y have similar but unusual refractive indexes, and this strengthens the evidence for G. With $\tau/\sigma = 100$, $u/(2\sigma^2)^{1/2} = 2$, and $(z-\mu)/(\frac{1}{2}\tau^2)^{1/2} = 2$, for example, these factors are respectively 0.135 and 2.718, and the overall Bayes factor is 26.01. Under G a frequentist test for a difference between θ_1 and θ_2 based on u would suggest that $\theta_1 \neq \theta_2$ at the 5% level, but the

Bayes factor gives strong evidence in favour of guilt, as the values of x and y correspond to similar, unusual, types of glass.

A more realistic model would account for non-normality of the distribution of θ. Other forms of evidence, such as DNA fingerprints or cloth samples, require more complex likelihoods in the Bayes factor and use prior information from specially tailored databases. Moreover when the probabilities being modelled are very small, it is important to allow for the possibility of events such as mistakes at the forensic laboratory. ■

We often wish to test nested hypotheses. In a typical example $\theta = (\psi, \lambda)$ for real ψ, and λ varies in an open subset of $\mathbb{R}^p$, with $H_0 : \psi = \psi_0$ and $H_1 : \psi \neq \psi_0$. Then if the same proper continuous prior $\pi(\psi, \lambda)$ is used under both hypotheses, the prior odds in favour of H_1 are infinite because

$$\Pr(H_0) = \int \pi(\psi_0, \lambda)\, d\lambda = 0$$

is an integral over a set of prior probability zero. Thus the posterior odds in favour of H_1 are infinite, whatever the data. This vexation can be eliminated by using different prior densities, weighted according to prior belief in H_0 and H_1, giving overall prior

$\delta(\cdot)$ is the Dirac delta function.

$$\pi(\psi, \lambda) = \delta(\psi - \psi_0)\pi(\psi_0, \lambda \mid H_0)\Pr(H_0) + \pi(\psi, \lambda \mid H_1)\Pr(H_1),$$

where

$$\int \pi(\psi_0, \lambda \mid H_0)\, d\lambda = \int \pi(\psi, \lambda \mid H_1)\, d\psi d\lambda = 1.$$

One result of this is that Bayes factors are more sensitive to the prior than are posterior densities. In particular, improper priors cannot be used, as the Bayes factor depends on the ratio of the two arbitrary constants of proportionality that appear in the priors. One way to remove the arbitrariness is to fix the ratio of these constants using some external argument.

A further difficulty is that when a large number of models must be compared, prior probabilities and proper priors must be assigned to each. This can be hard in practice, and the results may depend strongly on how it is done. This contrasts with frequentist hypothesis testing, where such difficulties do not arise. An apparently even more striking contrast is provided by the following example.

Example 11.15 (Jeffreys–Lindley paradox) Consider testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ based on a normal random sample $y_1, \ldots, y_n$ with mean μ and known variance σ^2. The usual test is based on the normal distribution of $n^{1/2}\overline{Y}/\sigma$ under H_0, and gives P-value $p = \Phi(-n^{1/2}|\overline{y}|/\sigma)$. In the Bayesian framework, we write $\pi_0 = \Pr(H_0)$, and suppose that under H_1, μ is normal

Dennis Victor Lindley (1923–) was educated at Cambridge, and held academic posts there, in Aberwystwyth, and in London. He is a strong advocate of Bayesian statistics. See Smith (1995).

$\overline{Y}$ is the average of the random variables $Y_1, \ldots, Y_n$; its observed value is $\overline{y}$.

Table 11.5
Dependence of Bayes factor B_{01} on sample size n for a test with significance level 0.01.

n	1	10	100	1000	10^4	10^6	10^8
B_{01}	0.269	0.163	0.376	1.15	3.63	36.2	362

with mean zero and variance τ^2. Then the posterior probabilities are

$$\begin{aligned}\Pr(H_0 \mid y) &= \frac{\pi_0}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n} y_j^2\right),\\ \Pr(H_1 \mid y) &= \frac{1-\pi_0}{(2\pi\sigma^2)^{n/2}(2\pi\tau^2)^{1/2}} \int \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2 - \frac{\mu^2}{2\tau^2}\right\} d\mu,\end{aligned}$$

leading to Bayes factor

$$B_{01} = \left(1 + n\frac{\tau^2}{\sigma^2}\right)^{1/2} \exp\left\{-\frac{n\overline{y}^2}{2\sigma^2(1+n^{-1}\sigma^2/\tau^2)}\right\}$$

in favour of H_0. Now suppose that $n\overline{y}^2/\sigma^2 = z_{\alpha/2}^2$. The significance level of the conventional test is α, but as $n \to \infty$ we see that $B_{01} \doteq n^{1/2}\tau\sigma^{-1}\exp(-z_{\alpha/2}^2/2)$, giving increasingly strong evidence in favour of H_0. Hence the paradox: although with $\overline{y}$ corresponding to $\alpha = 10^{-6}$ we would reject H_0 decisively, the Bayes factor gives increasingly strong support for H_0, because as $n \to \infty$, the weight of the alternative distribution is more and more widely spread compared to the distance from $\overline{y}$ to the null hypothesis value of μ. Table 11.5 gives some values of B_{01} when $\tau^2 = \sigma^2$.

One resolution of this hinges on noticing that a fixed alternative is not appropriate as $n \to \infty$. A test is used when there is doubt as to its outcome — when the data do not evidently contradict the null hypothesis. Mathematically, this means that sensible alternatives are $O(n^{-1/2})$ distant from the null hypothesis. In this case we take $\tau^2 = n^{-1}\delta\sigma^2$, so that as $n \to \infty$ the range of alternatives is fixed relative to the null; sensible values for δ might be in the range 5–20. Then the Bayes factor corresponding to significance level α, $B_{01} = (1+\delta)^{1/2}\exp\{-\frac{1}{2}z_{\alpha/2}^2/(1+\delta^{-1})\}$, does not increase with n. If we take $\delta = 10$ and $\alpha = 0.05$, 0.01, 0.001, and 0.0001, B_{10} equals 1.73, 6.2, 41.4, and 293. According to Table 11.3 these correspond respectively to evidence against H_0 that is hardly worth mentioning, positive, strong, and very strong, broadly agreeing with the usual interpretation of the P-values. ■

11.2.3 Model criticism

The prior density $\pi(\theta)$ introduces further information into the model, with the benefit of directness of inference for θ. The corresponding disbenefit is

the need to assess the appropriateness of $\pi(\theta)$ and the sensitivity of posterior conclusions to the prior, added to the usual concerns about the sampling model $f(y \mid \theta)$. Sensitivity analysis is generally performed simply by comparing posterior inferences based on a range of priors and models. The problems this poses are mainly computational, and we discuss them briefly in Section 11.3.

When just a few parametrized alternative models are in view, the ideas for model comparison outlined in Section 11.2.2 can be applied, supplemented with suitable graphs. In practice, however, consideration of all possible models is usually infeasible, not least because data can spring surprises on the investigator, and so we turn to model-checking when the alternatives are not explicit.

Marginal inference

From a Bayesian viewpoint all information concerning the data and model is contained in the joint density

$$f(y, \theta) = \pi(\theta \mid y) f(y). \tag{11.19}$$

and this suggests that $f(y)$ should be used to check the model. It is relatively clear how to do this when there is a sufficient statistic s and $s = (t, a)$, where a is a function of s whose distribution does not depend on θ; a is an *ancillary statistic*, a notion explored in Section 12.1. Then we can write

$$f(y) = f(y \mid s) f(a) \int f(t \mid a, \theta) \pi(\theta) \, d\theta, \tag{11.20}$$

the first two components of which do not depend on the prior, and hence can be used to give information about the sampling model. The third component of (11.20), $f(t \mid a)$, can be regarded as carrying information about agreement between data and prior. In simple models, consideration of the first two terms can yield standard model-checking tools.

Example 11.16 (Location-scale model) Let $y_1, \ldots, y_n$ be a random sample from the location-scale model $y_j = \eta + \tau \varepsilon_j$, where the ε_j have density g. In general, the order statistics $s = (y_{(1)}, \ldots, y_{(n)})$ form a minimal sufficient statistic for $\theta = (\eta, \tau)$ based on $y_1, \ldots, y_n$. They may be re-expressed as

$$t = \widehat{\theta} = (\widehat{\eta}, \widehat{\tau}), \quad a = \left(\frac{y_{(1)} - \widehat{\eta}}{\widehat{\tau}}, \ldots, \frac{y_{(n)} - \widehat{\eta}}{\widehat{\tau}} \right),$$

where t consists of the maximum likelihood estimators of θ, and the joint distribution of the maximal invariant a is degenerate but independent of η and τ. The suitability of g can be checked by probability plots of a against quantiles of g. Similar ideas extend to regression models.

Given a particular choice of g, agreement between the prior and data would be assessed through the conditional density of $\widehat{\theta}$ given a.

When g is normal, the minimal sufficient statistic is $(\overline{y}, s^2)$ and the assumption of normality is checked using the distribution of y given $\overline{y}$ and s^2. Example 5.14 established that the raw residuals $((y_1 - \overline{y})/s, \ldots, (y_n - \overline{y})/s)$ are independent of $\overline{y}$ and s^2.

The marginal joint distribution of $\overline{y}$ and s^2 enables the prior to be criticized. For instance, suppose that a joint conjugate prior is used for μ and σ^2, with

$$\mu \mid \sigma^2 \;\sim\; N\left(\mu_0, \sigma^2/k_0\right), \quad \sigma^2 \;\sim\; IG\left(\tfrac{1}{2}\nu_0, \tfrac{1}{2}\nu_0\sigma_0^2\right).$$

$IG(\cdot,\cdot)$ denotes the inverse gamma distribution.

Then integration shows that the marginal densities of $\overline{y}$ and s^2 are given by

$$d_1 = \frac{\overline{y} - \mu_0}{\sigma_0\left(n^{-1} + k_0^{-1}\right)^{1/2}} \;\sim\; t_{\nu_0}, \quad d_2 = \frac{s^2}{\sigma_0^2} \;\sim F_{n-1,\nu_0}.$$

Values of d_1 and d_2 that are unusual relative to the distributions of the corresponding random variables D_1 and D_2 can cast doubt on both prior and sampling models. For example, if a probability plot cast no doubt on the assumption of normality, and $d_1 = 100$ nevertheless, the relevance of the prior values μ_0 and σ_0^2 would be called into question. But if the data were not normal but Cauchy, then $\overline{y}$ would have the same distribution as y_1 and very large values of d_1 could arise even if the prior and data agreed about μ.

Consider again the data of Example 11.12, for which the model was normal. Suppose that our prior is that conditional on σ^2, $\mu \sim N(0, \sigma^2)$, and that the prior distribution for σ^2 is $IG(3, 3 \times 100^2)$. Then $d_1 = 0.202$ and $d_2 = 0.1424$. The first is close enough to zero to cast no doubt on the prior mean, but d_2 is rather small relative to the $F_{14,6}$ distribution, and casts some doubt on the prior variance. The corresponding Bayesian P-values are $\Pr(|D_1| > |d_1|) = 0.75$ and $\Pr(D_2 < d_2) = 0.045$; the data are rather more precise than our prior information would suggest. ■

One overall measure of the plausibility of the data under the model is the probability $\Pr\{f(Y_+) \le f(y)\}$, where $f(y)$ is the marginal density of the data actually observed, and Y_+ is a set of data that might have been observed (Problem 11.12). Some controversy surrounds this test and the P-values calculated in the previous example, as they flout the likelihood principle. One view is that the essence of Bayesian inference is to use Bayes' theorem to update prior belief in light of the data. This entails using posterior probabilities or equivalently Bayes factors to compare competing models, and leaves no place for tail probability calculations. A contrary argument is that a Bayes factor measures the relative support for two hypotheses and therefore requires prior specification of each, while some model-checking techniques do not require explicit alternatives: if my prior belief is that $y_1, \ldots, y_{20} \overset{\text{iid}}{\sim} N(0,1)$, I am surprised to learn that the smallest value is -10, even before considering

how this could have arisen. Furthermore, a strict interpretation of the argument for Bayes factors requires the specification of a proper prior distribution over all reasonable alternatives, which seems infeasible in practice. Finally, the argument for the likelihood principle assumes that the model is correct and the case for strict adherence to the principle seems weaker when assessing fit than when performing inference for a parameter.

Prediction diagnostics

Most models do not have a useful reduction in terms of exact minimal sufficient or ancillary statistics, so the ideas outlined above cannot usually be applied. Moreover, $\pi(\theta)$ is often improper in practice and then $f(y)$ is typically improper also, though this need not undercut diagnostic use of $f(y \mid s)f(a)$ if there is a useful sufficient reduction. When $\pi(\theta)$ is improper, posterior predictive distributions can be used to diagnose both problems with individual cases and more general model failures. The idea is to assess the posterior plausibility of suitable functions of the data.

One way to detect single outliers compares observations with their predicted values conditional on the remaining data through the *conditional predictive ordinates* $f(y_j \mid y_{-j})$, where y_{-j} consists of all the data except y_j. Since these quantities may be written in terms of ratios of densities, they depend less on the propriety of priors. There is a close link to cross-validation.

Example 11.17 (Normal linear model) In the normal linear model with known $n \times p$ design matrix X of rank $p < n$, the distribution of the $n \times 1$ response vector y conditional on the $p \times 1$ vector of parameters β and the error variance σ^2 is normal with mean $X\beta$ and covariance matrix $\sigma^2 I_n$, and the least squares estimates and residual estimate of error

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y, \quad s^2 = (n-p)^{-1}y^{\mathrm{T}}\{I - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\}y,$$

are independent and minimal sufficient for β and σ^2.

It would be alarming if the usual standardized residuals r_j had no Bayesian justification. Fortunately they do, as we now see. The simplest argument is that the joint distribution of $a = (r_1, \ldots, r_n)$ is free of the parameters $\theta = (\beta, \sigma^2)$, for which $\widehat{\theta} = (\widehat{\beta}, s^2)$ form a complete minimal sufficient statistic. Basu's theorem (page 724) implies that a is independent of $\widehat{\theta}$, so we infer from (11.20) that the sampling model can be checked by comparing a to its joint distribution. This justifies residual plots and other tricks of the trade.

Concentrationally-challenged readers may want to jump to (11.23).

For a longer more tedious argument for Bayesian use of deletion residuals and hence of the r_j, we compute the conditional predictive ordinate $f(y_j \mid y_{-j})$ under the conjugate prior distribution for β and σ^2,

$$\beta \mid \sigma^2 \;\sim\; N(\gamma, \sigma^2 V), \quad \sigma^2 \;\sim\; IG\left(\tfrac{1}{2}\nu, \tfrac{1}{2}\nu\tau^2\right),$$

where the hyperparameters are the $p \times 1$ vector γ, the $p \times p$ positive definite

symmetric matrix V, and the scalars ν and τ^2; these are all regarded as known. An argument analogous to that leading to (11.13) gives

$$\pi(\beta, \sigma^2 \mid y) \propto \pi(\beta \mid \widehat{\beta}, \sigma^2)\pi(\sigma^2 \mid s^2),$$

so we need only find the posterior distributions of β given $\widehat{\beta}$ and σ^2 and of σ^2 given s^2. As the joint distribution of $(\beta^{\mathrm{T}}, \widehat{\beta}^{\mathrm{T}})^{\mathrm{T}}$ given σ^2 is

$$N_{2p}\left\{\begin{pmatrix}\gamma\\ \gamma\end{pmatrix}, \sigma^2\begin{pmatrix}V & V\\ V & V + (X^{\mathrm{T}}X)^{-1}\end{pmatrix}\right\},$$

(3.21) and Exercise 8.5.2 shows that the posterior distribution of β given $\widehat{\beta}$ and σ^2 is normal with mean and variance matrix

$$\left(X^{\mathrm{T}}X + V^{-1}\right)^{-1}\left(X^{\mathrm{T}}X\widehat{\beta} + V^{-1}\gamma\right), \quad \sigma^2\left(X^{\mathrm{T}}X + V^{-1}\right)^{-1}, \tag{11.21}$$

which generalizes (11.11). As prior uncertainty about γ increases, $V^{-1} \to 0$, and then we see from (11.21) that the posterior mean and variance of β approach $\widehat{\beta}$ and $\sigma^2(X^{\mathrm{T}}X)^{-1}$. Direct calculation shows that the posterior distribution of σ^2 given s^2 is $IG[(\nu + n)/2, \{\nu\tau^2 + (n-p)s^2\}/2]$. If the constant prior $\pi(\beta) \propto 1$ is used, then the posterior mean and variance of β given σ^2 are $\widehat{\beta}$ and $\sigma^2(X^{\mathrm{T}}X)^{-1}$, but the posterior density for σ^2 is $IG[(\nu + n - p)/2, \{\nu\tau^2 + (n-p)s^2\}/2]$; letting $\nu \to 0$ gives the effect of taking $\pi(\beta, \sigma^2) \propto \sigma^{-2}$.

For future reference we note that the distribution of y conditional on σ^2 is normal with mean $X\gamma$ and variance $\sigma^2(I + XVX^{\mathrm{T}})$, and that on integrating over the prior distribution for σ^2, we find that the marginal density $f(y)$ has a multivariate t form

$$\frac{\Gamma\left(\frac{n+\nu}{2}\right)(\nu\tau^2)^{\nu/2}}{\pi^{n/2}\Gamma\left(\frac{\nu}{2}\right)|I + XVX^{\mathrm{T}}|^{1/2}}\left\{\nu\tau^2 + (y - X\gamma)^{\mathrm{T}}\left(I + XVX^{\mathrm{T}}\right)^{-1}(y - X\gamma)\right\}^{-(n+\nu)/2}. \tag{11.22}$$

To find the posterior predictive density of another observation y_+ with $p \times 1$ covariate vector x_+, assumed independent of y conditional on β and σ^2, we write

$$\begin{aligned} f(y_+ \mid y) &= \int f(y_+ \mid \theta)\pi(\theta \mid y)\, d\theta \\ &= \int\int f(y_+ \mid \beta, \sigma^2)\pi(\beta \mid \widehat{\beta}, \sigma^2)\pi(\sigma^2 \mid s^2)\, d\beta\, d\sigma^2 \\ &= \int \pi(\sigma^2 \mid s^2)\int f(y_+ \mid \beta, \sigma^2)\pi(\beta \mid \widehat{\beta}, \sigma^2)\, d\beta\, d\sigma^2. \end{aligned}$$

Now

$$\begin{aligned} y_+ \mid \beta, \sigma^2 &\sim N(x_+^{\mathrm{T}}\beta, \sigma^2), \\ \beta \mid \widehat{\beta}, \sigma^2 &\sim N\left\{(X^{\mathrm{T}}X + V^{-1})^{-1}(X^{\mathrm{T}}X\widehat{\beta} + V^{-1}\gamma), \sigma^2(X^{\mathrm{T}}X + V^{-1})^{-1}\right\}, \end{aligned}$$

from which it follows that conditional on $\widehat{\beta}$ and σ^2, the distribution of y_+ is normal with mean and variance

$$x_+^{\mathrm{T}}(X^{\mathrm{T}}X + V^{-1})^{-1}(X^{\mathrm{T}}X\widehat{\beta} + V^{-1}\gamma), \quad \sigma^2\left\{1 + x_+^{\mathrm{T}}(X^{\mathrm{T}}X + V^{-1})^{-1}x_+\right\}.$$

Integration over the posterior distribution of σ^2 shows that the posterior predictive distribution of y_+ conditional on y is given by

$$\frac{y_+ - x_+^{\mathrm{T}}(X^{\mathrm{T}}X + V^{-1})^{-1}(X^{\mathrm{T}}X\widehat{\beta} + V^{-1}\gamma)}{\left[\left\{\frac{(n-p)s^2+\nu\tau^2}{n+\nu}\right\}\left\{1 + x_+^{\mathrm{T}}(X^{\mathrm{T}}X + V^{-1})^{-1}x_+\right\}\right]^{1/2}} \sim\ t_{n+\nu}. \tag{11.23}$$

For prediction of y_j given the other observations y_{-j}, based on the improper prior $\pi(\beta, \sigma^2) \propto \sigma^{-2}$, we set $V^{-1} = 0$ and $\nu = 0$ and replace y_+ with y_j, x_+ with x_j, $n+\nu$ with $n-p-1$, and $\widehat{\beta}$, s^2 and X with the corresponding quantities $\widehat{\beta}_{-j}$, s^2_{-j} and X_{-j} based on y_{-j}. Then (11.23) becomes

$$\frac{y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j}}{\left[s^2_{-j}\left\{1 + x_j^{\mathrm{T}}(X_{-j}^{\mathrm{T}}X_{-j})^{-1}x_j\right\}\right]^{1/2}} \sim\ t_{n-p-1}.$$

A straightforward calculation reveals that the term in braces in the denominator here is $(1-h_j)^{-1}$, where h_j is the jth leverage based on the full model. Hence prediction of y_j given y_{-j} may be based on the t_{n-p-1} distribution of the deletion residual

$$r_j^* = \frac{(y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j})(1-h_j)^{1/2}}{s_{-j}}.$$

Thus outlier detection based on the conditional predictive ordinate is conducted using the usual deletion residuals r_j^*. As these are monotonic functions of the standardized residuals r_j, this supports Bayesian use of the r_j. ■

More general diagnostics can be based on measures of discrepancy between data and the model, $d = d(y, \theta)$, compared to data Y_+ that might have been generated by the model. Posterior predictive checks are based on comparison of $D_+ = d(Y_+, \theta)$ with its predictive distribution, via

$$\Pr\left\{d(Y_+, \theta) \geq d(y, \theta) \mid y\right\}, \tag{11.24}$$

where the averaging is over both Y_+ and the posterior distribution of θ. Since Y_+ is independent of y given θ, we can write

$$\int \Pr\left\{D_+ \geq d(y, \theta) \mid y, \theta\right\} \pi(\theta \mid y)\, d\theta = \int \Pr\left\{D_+ \geq d(y, \theta) \mid \theta\right\} \pi(\theta \mid y)\, d\theta.$$

Thus a simple way to evaluate (11.24) is to calculate $\Pr\left\{D_+ \geq d(y, \theta) \mid \theta\right\}$ for

fixed θ, and then to average this probability over the posterior density of θ. One omnibus measure of discrepancy is the analogue of Pearson's statistic,

$$d(y,\theta) = \sum_{j=1}^{n} \frac{\{y_j - \mathrm{E}(Y_j \mid \theta)\}^2}{\mathrm{var}(Y_j \mid \theta)},$$

but this may be inappropriate, and typically D_+ is chosen with key aspects of the model in mind. As mentioned above, authors differ over whether (11.24) should be used, though unlike the use of the marginal density of y, inference based on (11.24) does condition on the data.

11.2.4 Prediction and model averaging

In the Bayesian framework prediction is performed through the posterior predictive density (11.6). In practice this is not as simple as it appears, because there may be a number of possible models $M_1, \ldots, M_k$ on which to the base the prediction. Conditional on M_i, the predictive density for z based on y is $f(z \mid y, M_i)$, but this ignores any uncertainty concerning the selection of M_i. This uncertainty can be incorporated by averaging over the posterior distribution of the model selected, to give the *model-averaged prediction*

$$f(z \mid y) = \sum_{i=1}^{k} f(z \mid y, M_i)\mathrm{Pr}(M_i \mid y) \tag{11.25}$$

which is an average of the posterior distributions of z under the different models, weighted according to their posterior probabilities

$$\mathrm{Pr}(M_i \mid y) = \frac{f(y \mid M_i)\mathrm{Pr}(M_i)}{\sum_{l=1}^{k} f(y \mid M_l)\mathrm{Pr}(M_l)}, \tag{11.26}$$

where

$$\begin{aligned} f(y \mid M_i) &= \int f(y \mid \theta_i, M_i)\pi(\theta_i \mid M_i)\, d\theta_i, \\ f(z \mid M_i, y) &= \frac{\int f(z \mid y, \theta_i, M_i) f(y \mid \theta_i, M_i)\pi(\theta_i \mid M_i)\, d\theta_i}{f(y \mid M_i)}. \end{aligned}$$

Here θ_i is the parameter for model M_i, under which the prior is $\pi(\theta_i \mid M_i)$ and the prior probability of M_i is $\mathrm{Pr}(M_i)$. Formally, (11.25) is just a re-expression of (11.6) in which the parameter splits into two parts, one a model indicator, M_i, and the other the parameters conditional on M_i. In using (11.25) it is crucial that z is the same quantity under all models considered, rather than one whose interpretation depends on the model.

In practice the main obstacle to model averaging is computational. For each model, the integrations involved must usually be done numerically using ideas described in Section 11.3. Furthermore there can be many models in

some applications — for example, selecting among 15 covariates in a regression problem gives $2^{15} = 32,768$ models, corresponding to inclusion or exclusion of each covariate separately, without considering outliers, transformations, and so forth. Thus it may be difficult to find the most plausible models, quite apart from the calculations conditional on each model and the difficulties of specifing a prior over model space — giving the same weight to all combinations of covariates will rarely be sensible.

Example 11.18 (Cement data) We fit linear models to the data in Table 8.1 with $n = 13$ observations and four covariates. There are 2^4 possible subsets of the covariates, giving us models $M_1, \ldots, M_{16}$, which for sake of illustration we regard as equally probable *a priori*, though in practice we should hope that a small number of covariates is more likely than a large number. The models are on different parameter spaces, so the discussion in Section 11.2.2 implies that proper, preferably weak, priors should be used. We use the conjugate prior described in Example 11.17, and without loss of generality centre and scale each covariate vector to have average zero and unit variance. We then set V to be the 5×5 matrix with diagonal elements $\phi^2(v, 1, 1, 1, 1)$, where v is the sample variance of y, $\gamma^{\mathrm{T}} = (\overline{y}, 0, 0, 0, 0)$, $\nu = 2.58$, $\tau^2 = 0.28$, and $\phi = 2.85$. This choice implies that the elements of β are independent *a priori*, and should give a weak but proper prior that is consistent between different models and invariant to location and scale changes of the response and explanatory variables.

The marginal density of y under this model is (11.22); for each subset of covariates we use the corresponding submatrix of V. Table 11.6 shows the quantities $2 \log B_{10}$, where $B_{10} = \Pr(y \mid M_1)/\Pr(y \mid M_0)$ is the Bayes factor in favour of a subset of covariates relative to the model with none, the posterior probabilities of each subset, and, for comparison, the residual sums of squares under the usual linear models, which are broadly in line with the probabilities.

Let us try and predict the value of a new response y_+ with covariates $x_+^{\mathrm{T}} = (1, 10, 40, 20, 30)$. Conditional on a particular subset of covariate vectors, the predictive distribution for y_+ is given by (11.23). Figure 11.3 shows these densities for the six models shown in Table 11.6 to have non-negligible support, and the model-averaged predictive density. ■

A different approach to dealing with model uncertainty is to find a plausible model, $f(y \mid \psi)\pi(\psi)$, and then add further parameters λ whose variation allows for the most uncertain aspects of the model, together with a prior that expresses belief about them. This gives an expanded model $f(y \mid \psi, \lambda)\pi(\psi, \lambda)$, to which (11.6) is then applied with $\theta = (\psi, \lambda)$.

Model	RSS	$2 \log B_{10}$	$\Pr(M \mid y)$	a	b
– – – –	2715.8	0.0	0.0000		
1 – – –	1265.7	7.1	0.0000		
– 2 – –	906.3	12.2	0.0000		
– – 3 –	1939.4	0.6	0.0000		
– – – 4	883.9	12.6	0.0000		
1 2 – –	57.9	45.7	0.2027	93.77	2.31
1 – 3 –	1227.1	4.0	0.0000		
1 – – 4	74.8	42.8	0.0480	99.05	2.58
– 2 3 –	415.4	19.3	0.0000		
– 2 – 4	868.9	11.0	0.0000		
– – 3 4	175.7	31.3	0.0002		
1 2 3 –	48.11	43.6	0.0716	95.96	2.80
1 2 – 4	47.97	47.2	0.4344	95.88	2.45
1 – 3 4	50.84	44.2	0.0986	94.66	2.89
– 2 3 4	73.81	33.2	0.0004		
1 2 3 4	47.86	45.0	0.1441	95.20	2.97

Table 11.6
Bayesian prediction using model averaging for the cement data. For each of the 16 possible subsets of covariates, the table shows the log Bayes factor in favour of that subset compared to the model with no covariates and gives the posterior probability of each model. The values of the posterior mean and scale parameters a and b are also shown for the six most plausible models; $(y_+ - a)/b$ has a posterior t density. For comparison, the residual sums of squares are also given.

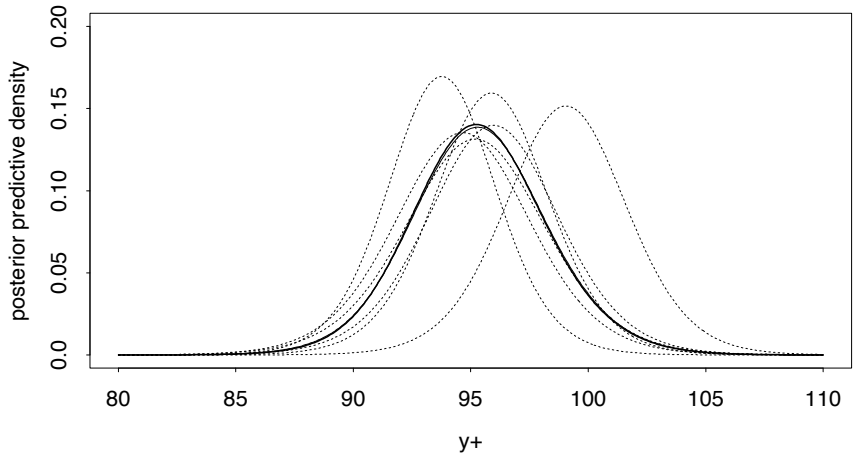


Figure 11.3
Posterior predictive densities for cement data. Predictive densities for y_+ based on individual models are given as dotted curves, and the heavy curve is the averaged prediction from all 16 models.

Exercises 11.2

1 Find elements $\tilde{\theta}$ and $\tilde{J}(\tilde{\theta})$ of the normal approximation to a beta density, and hence check the formulae in Example 11.11. Find also the posterior mean and variance of θ. Give an approximate 0.95 credible interval for θ. How does this differ from a 0.95 confidence interval? Comment.

2 Let $Y_1, \ldots, Y_n$ be a random sample from the uniform distribution on $(0, \theta)$, and

take as prior the Pareto density with parameters β and λ,

$$\pi(\theta) = \beta\lambda^{\beta}\theta^{-\beta-1}, \quad \theta > \lambda, \quad \beta, \lambda > 0.$$

(a) Find the prior distribution function and quantiles for θ, and hence give prior one- and two-sided credible intervals for θ. If $\beta > 1$, find the prior mean of θ.
(b) Show that the posterior density of θ is Pareto with parameters $n + \beta$ and $\max\{Y_1, \ldots, Y_n, \lambda\}$, and hence give posterior credible intervals and the posterior mean for θ.
(c) Interpret λ and β in terms of a prior sample from the uniform density.

3 Check the details of Example 11.7.

4 Two independent samples $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ and $X_1, \ldots, X_m \overset{\text{iid}}{\sim} N(\mu, c\sigma^2)$ are available, where $c > 0$ is known. Find posterior densities for μ and σ based on prior $\pi(\mu, \sigma) \propto 1/\sigma$.

5 Verify (11.21), (11.22), and (11.23). How do (11.21) and (11.22) change when $\text{var}(y_j \mid \beta, \sigma^2) = \sigma^2/w_j$, the w_j being known weights?

6 Travelling in a foreign country, you arrive at midnight in a town you have never heard of. You have no idea of its size. The first thing you see is a bus with the number $y = 100$. What is a reasonable estimate of the total number θ of buses in the town, assuming that they are numbered $1, \ldots, \theta$?
(a) Explain why it is sensible to use the improper prior $\pi(\theta) \propto \theta^{-1}$, $\theta = 1, 2, \ldots$. Assuming that $f(y \mid \theta)$ is uniform on $1, \ldots, \theta$, show that θ has posterior density

$$\pi(\theta \mid y) = \frac{\theta^{-2}}{\sum_{u=y}^{\infty} u^{-2}}, \quad \theta = y, y+1, \ldots.$$

(b) Show that the posterior mean of θ is infinite. Show also that the posterior distribution function is approximately

$$\Pr(\theta \le v \mid y) \doteq \frac{\int_{y-1/2}^{v+1/2} u^{-2}\, du}{\int_{y-1/2}^{\infty} u^{-2}\, du},$$

and that the posterior median is approximately $2y - 3/2$. Give an equi-tailed 95% posterior confidence interval and a 95% HPD interval for θ.
(c) What would you conclude if you saw two buses, numbered 100 and 30?

7 In Example 11.12, calculate the Bayes factor for $H_0 : \mu \le 0$ and $H_1 : \mu > 0$.

8 A forensic laboratory assesses if the DNA profile from a specimen found at a crime scene matches the DNA profile of a suspect. The technology is not perfect, as there is a (small) probability ρ that a match occurs by chance even if the suspect was not present at the scene, and a (larger) probability γ that a match is reported even if the profiles are different; this can arise due to laboratory error such as cross-contamination or accidental switching of profiles.
(a) Let R, S, and M denotes the events that a match is reported, that the specimen does indeed come from the suspect, and that there is a match between the profiles, and suppose that

$$\Pr(R \mid M \cap S) = \Pr(R \mid M \cap \overline{S}) = \Pr(R \mid M) = 1,\ \Pr(\overline{M} \mid S) = 0,\ \Pr(R \mid S) = 1.$$

$\overline{M}$ denotes the complement of M, and $\cap$ means 'and'.

Show that the posterior odds of the profiles matching, given that a match has

been reported, depend on

$$\frac{\Pr(R \mid S)}{\Pr(R \mid \overline{S})} = \frac{\Pr(R \mid M \cap S)\Pr(M \mid S) + \Pr(R \mid \overline{M} \cap S)\Pr(\overline{M} \mid S)}{\Pr(R \mid M \cap \overline{S})\Pr(M \mid \overline{S}) + \Pr(R \mid \overline{M} \cap \overline{S})\Pr(\overline{M} \mid \overline{S})},$$

and establish that this equals $\{\rho + \gamma(1-\rho)\}^{-1}$.
(b) Tabulate $\Pr(R \mid S)/\Pr(R \mid \overline{S})$ when $\rho = 0$, 10^{-9}, 10^{-6}, 10^{-3} and $\gamma = 0$, 10^{-4}, 10^{-3}, 10^{-2}.
(c) At what level of posterior odds would you be willing to convict the suspect, if the only evidence against them was the DNA analysis, and you should only convict if convinced of their guilt 'beyond reasonable doubt'? Would your chosen odds level depend on the likely sentence, if they are found guilty? How does your answer depend on the prior odds of the profiles matching, $\Pr(S)/\Pr(\overline{S})$?

9 One way to set the ratio of arbitrary constants that appears when two models are compared using Bayes factors and improper priors is by *imaginary observations*: we imagine the smallest experiment that would enable the models to be discriminated but maximizes evidence in favour of H_0, and then choose the constants so that the Bayes factor equals one for these data.
Consider data from a Poisson process observed on $[0, t_0]$, and let H_0 and H_1 represent the models with rates $\lambda(t) = \rho$ and $\lambda(t) = \mu\beta^{-1}\{1 - \exp(-\beta t)\}$, where $\rho, \mu, \beta > 0$. Take improper priors $\pi(\rho) = c_0\rho^{-1}$ and $\pi(\mu, \beta) = c_1\mu^{-2}$, with $c_1, c_0 > 0$.
(a) Explain why the smallest experiment that enables the models to be discriminated must have two events, and show that it gives $\Pr(y \mid H_0) = c_0/t_0^2$. Find $\Pr(y \mid H_1)$ and show that it is minimized when both events occur at t_0, with

$$\Pr(y \mid H_1) = c_1 \int_0^\infty \frac{\beta e^{-2\beta t_0}}{1 - e^{-\beta t_0}}\, d\beta = c_1 t_0^{-2} \left(\frac{\pi^2}{6} - 1\right).$$

Deduce that the device of imaginary observations gives $c_0/c_1 = \pi^2/6 - 1$.
(b) Compute the Bayes factor when these two models are compared using the data in Table 6.13. Discuss.
(Section 6.5.1; Raftery, 1988; Spiegelhalter and Smith, 1982)

10 A random sample $y_1, \ldots, y_n$ arises either from a log-normal density, with $\log Y_j \sim N(\mu, \sigma^2)$, or from an exponential density $\rho^{-1}e^{-y/\rho}$. The improper priors chosen are $\pi(\rho) = c_0/\rho$ and $\pi(\mu, \sigma) = c_1/\sigma$, for $\rho, \sigma > 0$ and $c_0, c_1 > 0$. Use imaginary observations to give a value for c_1/c_0.

11.3 Bayesian Computation

11.3.1 Laplace approximation

The goal of Bayesian data analysis is posterior inference for quantities of interest, and this involves integration over one or more of the parameters. Usually the integrals cannot be obtained in closed form and numerical approximations must be used. Deterministic integration procedures such as Gaussian quadrature can sometimes be applied, but they are typically useful only for low-dimensional integrals, and have the drawback of requiring knowledge of

the position and width of any modes of the integrand that is usually unavailable in practice. The most powerful tool for approximate calculation of posterior densities is numerical integration by Monte Carlo simulation, to which we turn after describing an analytical approach known as *Laplace's method*.

Consider the one-dimensional integral

$$I_n = \int_{-\infty}^{\infty} e^{-nh(u)}\, du, \tag{11.27}$$

where $h(u)$ is a smooth convex function with minimum at $u = \tilde{u}$, at which point $dh(\tilde{u})/du = 0$ and $d^2h(\tilde{u})/du^2 > 0$. For compactness of notation we write $h_2 = d^2h(\tilde{u})/du^2$, $h_3 = d^3h(\tilde{u})/du^3$, and so forth. Close to $\tilde{u}$ a Taylor series expansion gives $h(u) \doteq h(\tilde{u}) + \frac{1}{2}h_2(u-\tilde{u})^2$, so

$$\begin{aligned} I_n &\doteq e^{-nh(\tilde{u})} \int_{-\infty}^{\infty} e^{-nh_2(u-\tilde{u})^2/2}\, du \\ &= e^{-nh(\tilde{u})} \int_{-\infty}^{\infty} e^{-z^2/2} \frac{du}{dz}\, dz \\ &= \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})}, \end{aligned}$$

where the first and second equalities use the substitution $z = (nh_2)^{1/2}(u - \tilde{u})$ and the fact that the normal density has unit integral. A more detailed accounting (Exercise 11.3.2) gives

$$I_n = \left(\frac{2\pi}{nh_2}\right)^{1/2} e^{-nh(\tilde{u})} \times \left\{1 + n^{-1}\left(\frac{5h_3^2}{24h_2^3} - \frac{h_4}{8h_2^2}\right) + O\left(n^{-2}\right)\right\}. \tag{11.28}$$

The leading term on the right of (11.28) is known as the *Laplace approximation* to I_n, and we denote it by $\tilde{I}_n$.

There are several points to note about (11.28). First, as $I_n/\tilde{I}_n = 1 + O(n^{-1})$, the error is relative, and $\tilde{I}_n$ is often remarkably accurate. Second, $\tilde{I}_n$ involves only h and its second derivative at $\tilde{u}$, so it is relatively easy to obtain, numerically if necessary. Third, the right-hand side of (11.28) is an asymptotic series for I_n, implying that its partial sums need not converge, and that the approximation may not be improved by including further terms of the series. And fourth, because the bulk of the normal probability integral lies within three standard deviations of its centre, the limits of the integral will not affect $\tilde{I}_n$ provided they lie outside the interval with endpoints $\tilde{u} \pm 3(nh_2)^{-1/2}$ or so.

In the multivariate case, with $h(u)$ again a smooth convex function but u a vector of length p, the same argument but using the multivariate normal density shows that the Laplace approximation to (11.27) is

$$\left(\frac{2\pi}{n}\right)^{p/2} |h_2|^{-1/2} e^{-nh(\tilde{u})}, \tag{11.29}$$

where $\tilde{u}$ solves the $p \times 1$ system of equations $\partial h(u)/\partial u = 0$ and $|h_2|$ is the determinant of the $p \times p$ matrix of second derivatives $\partial^2 h(u)/\partial u \partial u^{\mathrm{T}}$, evaluated at $u = \tilde{u}$, at which point the matrix is positive definite.

In applications an approximation is often required to an integral of form

$$J_n(u_0) = \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{u_0} a(u) e^{-ng(u)} \left\{1 + O(n^{-1})\right\} \, du, \tag{11.30}$$

where u is scalar, $a(u) > 0$, and in addition to possessing the properties of $h(u)$ above, g is such that $g(\tilde{u}) = 0$. The first step in approximating (11.30) is to change the variable of integration from u to $r(u) = \mathrm{sign}(u - \tilde{u})\{2g(u)\}^{1/2}$; that is, $r^2/2 = g(u)$. Then $g'(u) = dg(u)/du$ and $r(u)$ have the same sign, and $r dr/du = g'(u)$, so

$$\begin{aligned} J_n(u_0) &= \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0} a(u) \frac{r}{g'(u)} e^{-nr^2/2} \left\{1 + O(n^{-1})\right\} \, dr \\ &= \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0} e^{-nr^2/2 + \log b(r)} \left\{1 + O(n^{-1})\right\} \, dr, \end{aligned}$$

where the positive quantity $b(r) = a(u) r / g'(u)$ is regarded as a function of r.

We now change variable again, from r to $r^* = r - (rn)^{-1} \log b(r)$, so

$$-nr^{*2} = -nr^2 + 2 \log b(r) - n^{-1} r^{-2} \{\log b(r)\}^2.$$

The Jacobian of the transformation and the third term in $-nr^{*2}$ contribute only to the error of $J_n(u_0)$, so

$$\begin{aligned} J_n(u_0) &= \left(\frac{n}{2\pi}\right)^{1/2} \int_{-\infty}^{r_0^*} e^{-nr^{*2}/2} \left\{1 + O(n^{-1})\right\} \, dr^* \\ &= \Phi(n^{1/2} r_0^*) + O(n^{-1}), \end{aligned} \tag{11.31}$$

where

$$r_0^* = r_0 + (r_0 n)^{-1} \log\left(\frac{v_0}{r_0}\right), \quad r_0 = \mathrm{sign}(u_0 - \tilde{u})\{2g(u_0)\}^{1/2}, \quad v_0 = \frac{g'(u_0)}{a(u_0)}.$$

Variants on this expression play an important role in Chapter 12.

Here is a further approximation for later use. Let $u = (u_1, u_2)$, where u_1 is scalar and u_2 a $p \times 1$ vector, and consider

$$(2\pi)^{-(p+1)/2} c \int_{-\infty}^{u_1^0} du_1 \int du_2 \exp\left\{-nh(u_1, u_2)\right\}, \tag{11.32}$$

where c is constant, the inner integral being over $\mathbb{R}^p$. Here h has its previous smoothness properties, is maximized at $(\tilde{u}_1, \tilde{u}_2)$, and in addition $h(\tilde{u}_1, \tilde{u}_2) = 0$. We fix u_1 and apply Laplace approximation to the inner integral, obtaining

$$(2\pi)^{-1/2} c \int_{-\infty}^{u_1^0} |nh_{22}(u_1, \tilde{u}_{21})|^{-1/2} \exp\left\{-nh(u_1, \tilde{u}_{21})\right\} \left\{1 + O(n^{-1}\right\} \, du_1,$$

where $\tilde{u}_{21} = \tilde{u}_2(u_1)$ maximizes $h(u_1, u_2)$ with respect to u_2 when u_1 is fixed, and $h_{22}(u_1, u_2) = \partial^2 h(u_1, u_2)/\partial u_2 \partial u_2^{\mathrm{T}}$ is the $p \times p$ Hessian matrix of h with respect to u_2. Apart from multiplicative constants, this integral has form (11.30), and so (11.31) may be used to approximate to (11.32), with

$$r_0 = \operatorname{sign}(u_1^0 - \tilde{u}_1)\left\{2h(u_1^0, \tilde{u}_{20})\right\}^{1/2}, \quad v_0 = c^{-1}\frac{\partial h(u_1^0, \tilde{u}_{20})}{\partial u_1}\left|h_{22}(u_1^0, \tilde{u}_{20})\right|^{1/2},$$

where $\tilde{u}_{20}$ is the maximizing value of u_2 when $u_1 = u_1^0$.

Although the formulation of (11.27), (11.30), and (11.32) in terms of n and the $O(1)$ functions h and g simplifies the derivation of (11.29) and (11.31) by clarifying the orders of the various terms, for applications it is equivalent and usually simpler to set $n = 1$ and allow h and g and their derivatives to be $O(n)$.

Inference

One application of Laplace approximation is to the Bayes factor (11.17). For one of the hypotheses we write $\Pr(y) = \int f(y \mid \theta)\pi(\theta)\, d\theta$, with integrand expressed as $\exp\{-h(\theta)\}$, where $h(\theta) = -\ell_m(\theta)$ and

$$\ell_m(\theta) = \log f(y \mid \theta) + \log \pi(\theta)$$

is the log likelihood modified by addition of the log prior. Typically the first term of ℓ_m is $O(n)$, and the second is $O(1)$. The value $\tilde{\theta}$ that minimizes $h(\theta)$ is the maximum *a posteriori* estimate of θ — the value that maximizes the modified log likelihood — and we can apply (11.29). The result is

$$\log \Pr(y) \doteq \log f(y \mid \tilde{\theta}) + \log \pi(\tilde{\theta}) - \tfrac{1}{2}p \log n + \tfrac{1}{2}p \log(2\pi) - \tfrac{1}{2}\log\left|-\frac{\partial^2 \ell_m(\tilde{\theta})}{\partial\theta \partial\theta^{\mathrm{T}}}\right|,$$

where p is the dimension of θ. To further simplify this, note that in large samples the log prior is negligible relative to the log likelihood and $\tilde{\theta}$ is roughly the maximum likelihood estimate $\widehat{\theta}$, and if concerned only with asymptotic properties, we can drop terms that are $O(1)$. This gives the breathtaking approximation

$$-2 \log \Pr(y) \;\doteq\; \mathrm{BIC} = -2 \log f(y \mid \widehat{\theta}) + p \log n.$$

This *Bayes information criterion*, which we met in Section 4.7, is used for rough comparison of competing models.

For a more sophisticated application we write a vector parameter θ as $(\psi, \lambda^{\mathrm{T}})^{\mathrm{T}}$ and approximate the marginal posterior density for the scalar ψ,

$$\pi(\psi \mid y) = \frac{\int f(y \mid \psi, \lambda)\pi(\psi, \lambda)\, d\lambda}{\int f(y \mid \psi, \lambda)\pi(\psi, \lambda)\, d\lambda d\psi}, \tag{11.33}$$

by applying Laplace's method to each integral. The discussion above gives

the approximation to the denominator. For the numerator we take $h_\psi(\lambda) = -\ell_m(\psi, \lambda)$, where the notation emphasises that the approximation is applied only to the integral over λ, for a fixed value of ψ. The resulting approximation may be written as

$$\pi(\psi \mid y) \doteq \left(\frac{n}{2\pi}\right)^{1/2} \left\{ \frac{\left| -\frac{\partial^2 \ell_m(\tilde{\psi}, \tilde{\lambda})}{\partial \theta \partial \theta^{\mathrm{T}}} \right|}{\left| -\frac{\partial^2 \ell_m(\psi, \tilde{\lambda}_\psi)}{\partial \lambda \partial \lambda^{\mathrm{T}}} \right|} \right\}^{1/2} \frac{f(y \mid \psi, \tilde{\lambda}_\psi)\pi(\psi, \tilde{\lambda}_\psi)}{f(y \mid \tilde{\psi}, \tilde{\lambda})\pi(\tilde{\psi}, \tilde{\lambda})}, \qquad (11.34)$$

where $\tilde{\lambda}_\psi$ is the maximum *a posteriori* estimate of λ for fixed ψ and the denominator and numerator determinants are of Hessian matrices of sides $(p-1)$ and p respectively.

The posterior marginal cumulative distribution for ψ may be approximated by applying (11.31) to the integral of (11.34) over the range (∞, ψ_0). We take $u_0 = \psi_0$,

$$g(\psi) = \ell_m(\tilde{\psi}, \tilde{\lambda}) - \ell_m(\psi, \tilde{\lambda}_\psi), \quad a(\psi) = \left\{ \frac{\left| -\frac{\partial^2 \ell_m(\tilde{\psi}, \tilde{\lambda})}{\partial \theta \partial \theta^{\mathrm{T}}} \right|}{\left| -\frac{\partial^2 \ell_m(\psi, \tilde{\lambda}_\psi)}{\partial \lambda \partial \lambda^{\mathrm{T}}} \right|} \right\}^{1/2},$$

and set $r_0^* = r_0 + r_0^{-1} \log(v_0/r_0)$, where

$$\begin{aligned} r_0 &= \operatorname{sign}(\psi_0 - \tilde{\psi}) \left[2 \left\{ \ell_m(\tilde{\psi}, \tilde{\lambda}) - \ell_m(\psi_0, \tilde{\lambda}_{\psi_0}) \right\} \right]^{1/2}, \\ v_0 &= -\frac{\partial \ell_m(\psi_0, \tilde{\lambda}_{\psi_0})}{\partial \psi} \left\{ \frac{\left| -\frac{\partial^2 \ell_m(\psi_0, \tilde{\lambda}_{\psi_0})}{\partial \lambda \partial \lambda^{\mathrm{T}}} \right|}{\left| -\frac{\partial^2 \ell_m(\tilde{\psi}, \tilde{\lambda})}{\partial \theta \partial \theta^{\mathrm{T}}} \right|} \right\}^{1/2}; \end{aligned}$$

here $\tilde{\lambda}_{\psi_0}$ is the maximum *a posteriori* estimate of λ when ψ is fixed at ψ_0. It is often convenient to find the derivatives numerically.

Numerous variant approaches are possible. For example, the ratio of priors in the integral of (11.34) may be included in the function $a(u)$ of (11.30), which case ℓ_m is simply the log likelihood, $\tilde{\theta}$ and $\tilde{\lambda}_\psi$ are maximum likelihood estimates, the Hessians are observed information matrices, and r_0 is the directed likelihood ratio statistic for testing the hypothesis $\psi = \psi_0$. The prior then appears only in v_0. The resulting approximation is generally poorer than that described above, but this idea does suggest a quick way to assess sensitivity to the prior density. The key is to notice that the approximate effect on (11.34) of taking a different prior, $\pi_1(\psi, \lambda)$, say, would be to multiply (11.34) by the ratio $c(\psi) = \{\pi_1(\psi, \tilde{\lambda}_\psi)/\pi(\psi, \tilde{\lambda}_\psi)\}/\{\pi_1(\tilde{\psi}, \tilde{\lambda})/\pi(\tilde{\psi}, \tilde{\lambda})\}$; the effect is approximate because Laplace approximation based on π_1 would not lead to integrals maximized at $\tilde{\lambda}_\psi$ and $(\tilde{\psi}, \tilde{\lambda})$. On the other hand, the effect on these maximizing values of changing the prior is often relatively small. Thus the effect of modifying the prior from π to π_1 may be gauged by changing v_0 to

Case	x	y	Rate estimate ($\times 10^2$)	
			Crude	Empirical Bayes
1	94.320	5	5.3	6.1
2	15.720	1	6.4	10.7
3	62.880	5	8.0	9.1
4	125.760	14	11.1	11.7
5	5.240	3	57.3	58.8
6	31.440	19	60.4	60.6
7	1.048	1	95.4	80.0
8	1.048	1	95.4	80.0
9	2.096	4	190.8	143.7
10	10.480	22	209.9	194.4

Table 11.7 Numbers of failures y of ten pumps in x thousand operating hours, with the crude rate estimate y/x (Gaver and O'Muircheartaigh, 1987). The final column gives empirical Bayes rate estimates discussed in Example 11.28.

$v_0/c(\psi_0)$, and recalculating r_0^* and $\Phi(r_0^*)$. This involves no further maximization or numerical differentation.

Example 11.19 (Pump failure data) Table 11.7 contains the numbers of failures y_j of $n = 10$ pumps in operating periods of x_j thousands of hours. The pumps are from several systems in the nuclear plant Farley 1; pumps 1, 3, 4, and 6 operate continuously, while the rest operate only intermittantly or on standby. For now we suppose that the pumps may be expected to have similar rates of failure, with the jth pump having failure rate λ_j, and that conditional on λ_j, the numbers of failures y_j have independent Poisson distributions with means $\lambda_j x_j$. We further suppose that the λ_j are independent realizations of a gamma variable with parameters α and β, and that β itself has a prior gamma distribution with parameters ν and ϕ. Thus

$$\begin{aligned} f(y \mid \lambda) &= \prod_{j=1}^{n} \frac{(x_j\lambda_j)^{y_j}}{y_j!} e^{-x_j\lambda_j}, \quad \pi(\lambda \mid \beta) = \prod_{j=1}^{n} \frac{\beta^\alpha \lambda_j^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta\lambda_j}, \\ \pi(\beta) &= \frac{\phi^\nu \beta^{\nu-1}}{\Gamma(\nu)} e^{-\phi\beta}, \end{aligned} \tag{11.35}$$

so that the joint density of the data y, the rates λ, and β is

$$f(y \mid \lambda) f(\lambda \mid \beta)\pi(\beta) = c \prod_{j=1}^{n} \left\{ \lambda_j^{y_j+\alpha-1} e^{-\lambda_j(x_j+\beta)} \right\} \times \beta^{n\alpha+\nu-1} e^{-\phi\beta}, \tag{11.36}$$

where c is a constant of proportionality.

To find the conditional density of β, we integrate over the λ_j, to obtain

$$f(y, \beta) = c \prod_{j=1}^{n} \left\{ (x_j + \beta)^{-(y_j+\alpha)} \Gamma(y_j + \alpha) \right\} \times \beta^{n\alpha+\nu-1} e^{-\phi\beta}, \tag{11.37}$$

Table 11.8
Integrals of two approximate posterior densities for β for the pumps data. The first, $\tilde{I}_1$, involves a one-dimensional Laplace approximation to (11.36), while $\tilde{I}_{10}$ involves ten-dimensional Laplace approximation. The table shows how the integral changes when the curvature of the likelihood is increased by a.

a	1	2	3	4	5	10	20
$\tilde{I}_1$	1.022	1.017	1.014	1.012	1.011	1.009	1.007
$\tilde{I}_{10}$	1.782	1.309	1.183	1.127	1.096	1.042	1.019

from which the marginal density of y is obtained by further integration to give

$$f(y) = c \prod_{j=1}^{n} \Gamma(y_j + \alpha) \times \int_0^\infty e^{-h(\beta)}\, d\beta,$$

where $h(\beta) = \phi\beta - (n\alpha + \nu - 1)\log\beta + \sum(y_j + \alpha)\log(x_j + \beta)$; we use I to denote the integral in this expression.

For sake of illustration we take a proper but fairly uninformative prior for β, with $\nu = 0.1$ and $\phi = 1$, and take $\alpha = 1.8$. Application of Laplace's method to I then results in the approximate posterior density for β, $\tilde{\pi}(\beta \mid y) = \tilde{I}^{-1}\exp\{-h(\beta)\}$, which has integral 1.022.

The accuracy of Laplace's method can be tested by taking a different approach, in which we first integrate (11.36) over β, and then apply the multivariate version of Laplace's method to the resulting ten-dimensional integral with respect to the λ_j. In this case the density approximation has integral 1.782, because the ten-dimensional integral approximation, $\tilde{I}_{10}$, is less accurate than $\tilde{I}_1$. To compare the two approaches we recalculate the approximations for data (ax_j, ay_j) and various values of a. This leaves unchanged the failure rates y_j/x_j, but increases by a factor a the Fisher information for each of the λ_j, thereby increasing the curvature of the log likelihood and the accuracy of the approximation. The results in Table 11.8 show that $\tilde{I}_{10}$ rapidly improves as a increases, and that with counts about 4–5 times as large as those observed, Laplace's method gives adequately accurate answers, even in ten dimensions. In practice, of course, $\tilde{I}_1$ would be used.

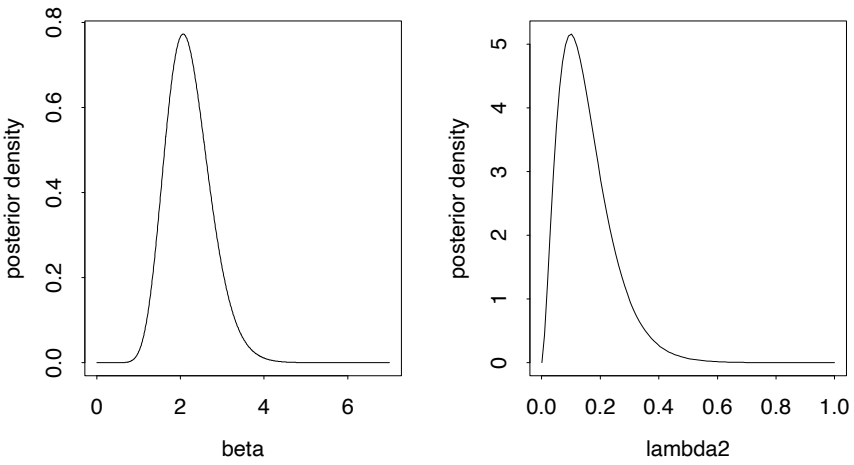
To calculate approximate posterior densities for λ_j, we integrate (11.36) over λ_i, $i \neq j$, and then apply Laplace's method to the numerator and denominator integrals of

$$\pi(\lambda_j \mid y) = \frac{\lambda_j^{y_j+\alpha-1} e^{-\lambda_j x_j} \int_0^\infty e^{-h_j(\beta)}\, d\lambda}{\Gamma(y_j + \alpha)\int_0^\infty e^{-h(\beta)}\, d\beta},$$

where

$$h_j(\beta) = (\phi + \lambda_j)\beta - (n\alpha + \nu - 1)\log\beta + \sum_{i\neq j}(y_i + \alpha)\log(x_i + \beta).$$

The resulting denominator is again $\tilde{I}_1$, while the numerator must be recalculated at each of a range of values of λ_j. Figure 11.4 shows these approximate



gure 11.4
proximate
terior densities
β and λ_2 for the
nps data, based
Laplace
proximation.

densities for β and for λ_2. That for λ_2 has integral 1.0004 and is presumably closer to one because it is based on a ratio of Laplace approximations. ■

The ideal situation for Laplace approximation is when the posterior density is strongly unimodal. When the posterior is multimodal, the approximation can be applied separately to each mode — provided they can all be found. Different approximations apply when the posterior is peaked at the end of its range (Exercise 11.3.5).

11.3.2 Importance sampling

Many Monte Carlo techniques may be applied in Bayesian computation. In this section we discuss ideas based on importance sampling, and in the next section we turn to iterative methods based on simulating Markov chains. Importance sampling gives independent samples, and so measures of uncertainty for estimators are usually fairly readily obtained, but it applies to a limited range of problems. Iterative methods are more widely applicable but it can be difficult to assess their convergence and to give statements of uncertainty for their output.

Suppose we wish to calculate an integral of form

$$\mu = \int m(\theta, y, z)\pi(\theta \mid y)\, d\theta.$$

If we take $m(\theta, y, z) = I(\theta \leq a)$, for example, then $\mu = \Pr(\theta \leq a \mid y)$, while taking $m(\theta, y, z) = f(z \mid y, \theta)$ gives $\mu = f(z \mid y)$, the posterior predictive density for z given the data. Suppose that direct computation of μ is awkward,

but that it is straightforward both to generate a sample $\theta_1, \ldots, \theta_S$ from a density $h(\theta)$ whose support includes that of $\pi(\theta \mid y)$, and to calculate $m(\theta, y, z)$ and $f(y \mid \theta)$. We can then apply importance sampling for estimation of μ, obtaining the unbiased estimator (Section 3.3.2)

$$\widehat{\mu} = S^{-1} \sum_{s=1}^{S} m(\theta_s, y, z) \frac{\pi(\theta_s \mid y)}{h(\theta_s)} = S^{-1} \sum_{s=1}^{S} m(\theta_s, y, z) w(\theta_s), \tag{11.38}$$

say, where $w(\theta) = \pi(\theta \mid y)/h(\theta)$ is an importance sampling weight. An important advantage of $\widehat{\mu}$ over the iterative procedures to be disussed later is that its variance is readily obtained (Exercise 11.3.6).

In practice the importance sampling ratio estimator of μ,

$$\widehat{\mu}_{\text{rat}} = \frac{\sum_{s=1}^{S} m(\theta_s, y, z) w(\theta_s)}{\sum_{s=1}^{S} w(\theta_s)},$$

is more commonly used. This is typically less variable than $\widehat{\mu}$; indeed it performs perfectly if $m(\theta, y, z)$ is constant, as is clear from its variance, given by (Example 2.25)

$$\widehat{\text{var}}(\widehat{\mu}_{\text{rat}}) = \frac{1}{S(S-1)} \sum_{s=1}^{S} \frac{\{m(\theta_s, y, z) - \widehat{\mu}_{\text{rat}}\}^2 w(\theta_s)^2}{\overline{w}^2}, \quad \overline{w} = S^{-1} \sum_{s=1}^{S} w(\theta_s).$$

As usual with importance sampling, a good choice of $h(\theta)$ is crucial if the simulation is to be useful. One possibility is a normal approximation to the posterior density of θ, taking $h(\theta)$ to be $N\left\{\widehat{\theta}, J(\widehat{\theta})^{-1}\right\}$, where $\widehat{\theta}$ and $J(\widehat{\theta})$ are the maximum likelihood estimate and the observed information. Normal approximation may be better if applied to a transformed parameter $\psi = \psi(\theta)$, however, while the light-tailed normal distribution typically gives too few simulations in the tail of the posterior density. Hence it is usually better to generate the θ_s from a shifted and rescaled t_ν density.

Example 11.20 (Challenger data) Table 1.3 gives data on launches of the space shuttle, including the ill-fated Challenger launch. In Examples 1.3, 4.5 and 4.33 we saw how these data may be modelled using a logistic regression model, under which the number of O-rings suffering thermal distress when a launch takes place at temperature x_1°F is binomial with denominator $m = 6$ and probability $\pi(\beta + \beta_1 x_1) = \exp(\beta_0 + \beta_1 x_1)/\{1 + \exp(\beta_0 + \beta_1 x_1)\}$. The likelihood (4.6) for this model is shown in Figure 4.3. Let us represent the data for the 23 successful launches by y, with likelihood $f(y \mid \theta)$; here $\theta = (\beta_0, \beta_1)$.

One aspect of interest when deciding whether to launch the Challenger should have been the number Z of distressed O-rings at its launch temperature of $x_1 = 31^\circ$F. We suppose that, conditional on θ, $f(z \mid \theta)$ is binomial with denominator $m = 6$ and probability $\pi(\beta_0 + 31\beta_1)$, independent of other

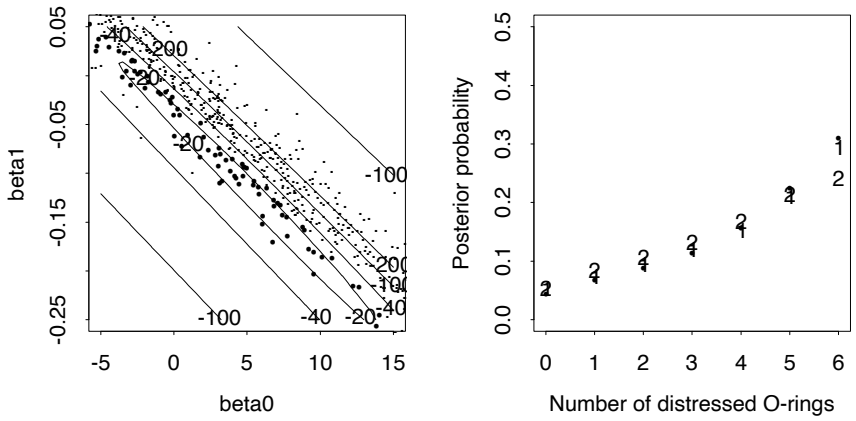


Figure 11.5
Importance sampling applied to shuttle data. Left: pairs (β_0, β_1) simulated from a prior density, with log likelihood contours superimposed. Pairs whose weight w_s exceeds $(100S)^{-1}$ are shown as blobs. The other pairs have very low likelihoods and hence essentially zero posterior probabilities w_s. Right: posterior predictive density for the number of distressed O-rings for a launch at 31°F, using beta prior with $a = b = 0.5$ (blobs), $a = b = 1$ (1) and $a = 1$, $b = 4$ (2), estimated by importance sampling with $S = 10,000$.

launches. Then in the Bayesian framework we should calculate the posterior predictive density for Z,

$$\frac{\int f(z \mid \theta) f(y \mid \theta) \pi(\theta)\, d\theta}{\int f(y \mid \theta) \pi(\theta)\, d\theta},$$

where $\pi(\theta)$ is the prior density on (β_0, β_1).

The parameters β_0 and β_1 are difficult to interpret directly, and instead we consider the probabilities $\pi_1 = \pi(\beta + 60\beta_1)$ and $\pi_2 = \pi(\beta + 80\beta_1)$ that a single O-ring will be distressed at 60 and 80°F. In practice specification of the joint prior density of π_1 and π_2 would require engineering expertise, but in default of this we simply suppose that they have independent beta densities (11.3) with $a = b = 1/2$. For the initial step of the importance sampling algorithm we generate 10,000 independent pairs (π_1, π_2) and then set

$$\beta_1 = \frac{1}{80 - 60} \log \left\{ \frac{\pi_2(1 - \pi_1)}{\pi_2(1 - \pi_1)} \right\}, \quad \beta_0 = \log \left\{ \frac{\pi_1}{1 - \pi_1} \right\} - 60\beta_1.$$

The left panel of Figure 11.5 shows some of the resulting pairs $\theta_s = (\beta_0, \beta_1)$, superimposed on contours of the log likelihood. Pairs whose weight w_s exceeds one-hundredth of its average are shown by blobs. About 30% of the simulated values fall into this category, for which $\sum w_s = 0.9996$, so just 4/10,000ths of the posterior probability is placed on the other 7000 pairs. This occurs both because the prior is much more dispersed than the likelihood, and because they are mismatched, in the sense that the prior value of β_1 for a given β_0 is generally too large — the mode of $f(\beta_1 \mid \beta_0)$ lies to the right of that of $f(y \mid \beta_1, \beta_0)$, considered as a function of β_1 for fixed β_0.

The right panel of Figure 11.5 shows the posterior probabilities of $z =$

$0, \ldots, 6$ distressed rings. There is appreciable probability of damage to most of the rings, as $\Pr(Z \geq 4 \mid y) \doteq 0.65$, with little dependence on the prior. ■

This examples show both the strengths and weaknesses of importance sampling. It is simple to apply, and because $\theta_1, \ldots, \theta_S$ are independent it is easy to obtain a standard error for $\widehat{\mu}$, and then to increase S if necessary. On the other hand the prior is sometimes so overdispersed relative to the likelihood that S must be huge before an appreciable number of the w_s are non-zero, and a better importance sampling distribution must be found. This problem becomes acute when the dimension of θ is large and the curse of dimensionality bites. There are clever ways to improve importance sampling in such situations, but Markov chain methods apply readily to many high-dimensional problems, and to these we now turn.

11.3.3 Markov chain Monte Carlo

The idea of Markov chain Monte Carlo simulation is to construct a Markov chain that will, if run for an infinitely long period, generate samples from a posterior distribution π, specified implicitly and known only up to a normalizing constant. Although it has roots in areas such as statistical physics, its application in mainstream Bayesian statistics is relatively recent and the discussion below is merely a snapshot of a topic in full spate of development. The reader whose memory of Markov chains is hazy may find it useful to review the early pages of Section 6.1.1.

Gibbs sampler

The term Gibbs sampling comes from an analogy with statistical physics, where similar methods are used to generate states from Gibbs distributions. In that context it is called the *heat bath* algorithm.

Let $U = (U_1, \ldots, U_k)$ be a random variable of dimension k whose joint density $\pi(u)$ is unknown. Our goal is to estimate aspects of $\pi(u)$, such as joint or marginal densities and their quantiles, moments such as $\mathrm{E}(U_1)$ and $\mathrm{var}(U_1)$, and so forth. Although $\pi(u)$ itself is unknown, we suppose that we can simulate observations from the *full conditional* densities $\pi(u_i \mid u_{-i})$, where $u_{-i} = (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_k)$. Often in practice the constant normalizing $\pi(u)$ is unknown, but as it does not appear in the $\pi(u_i \mid u_{-i})$, this causes no difficulty. If $\pi(u)$ is proper, then the Hammersley–Clifford theorem implies that under mild conditions $\pi(u)$ is determined by these densities; this does not imply that any set of full conditional densities determines a proper joint density. *Gibbs sampling* is successive simulation from the $\pi(u_i \mid u_{-i})$ according to the algorithm:

1 initialize by taking arbitrary values of $U_1^{(0)}, \ldots, U_k^{(0)}$.
2 Then for $i = 1, \ldots, I$,
  (a) generate $U_1^{(i)}$ from $\pi\left(u_1 \mid u_2 = U_2^{(i-1)}, \ldots, u_k = U_k^{(i-1)}\right)$,
  (b) generate $U_2^{(i)}$ from $\pi\left(u_2 \mid u_1 = U_1^{(i)}, u_3 = U_3^{(i-1)}, \ldots, u_k = U_k^{(i-1)}\right)$,

(c) generate $U_3^{(i)}$ from

$$\pi\left(u_3 \mid u_1 = U_1^{(i)}, u_2 = U_2^{(i)}, u_4 = U_4^{(i-1)}, \ldots, u_k = U_k^{(i-1)}\right),$$

$\vdots$

(k) generate $U_k^{(i)}$ from $\pi\left(u_k \mid u_1 = U_1^{(i)}, \ldots, u_{k-1} = U_{k-1}^{(i)}\right)$.

Here we update each of the U_j in turn, basing each value generated on the $k-1$ previous simulations. This gives a stream of random variables

$$U_1^{(1)}, \ldots, U_k^{(1)}, U_1^{(2)}, \ldots, U_k^{(2)}, \quad \ldots, \quad U_1^{(I-1)}, \ldots, U_k^{(I-1)}, U_1^{(I)}, \ldots, U_k^{(I)},$$

so for the jth component of U we have a sequence $U_j^{(1)}, \ldots, U_j^{(I)}$.

To see why we might hope that $(U_1^{(I)}, \ldots, U_k^{(I)})$ is approximately a sample from $\pi(u)$, suppose that $k = 2$ and that U_1 and U_2 take values in the finite sets $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$. We write their joint and marginal densities as

$$\begin{aligned}
\Pr(U_1 = r, U_2 = s) &= \pi(r, s), \\
\Pr(U_1 = r) &= \pi_1(r) = \sum_{s=1}^{m} \pi(r, s), \quad r = 1, \ldots, n, \\
\Pr(U_2 = s) &= \pi_2(s) = \sum_{r=1}^{n} \pi(r, s), \quad s = 1, \ldots, m,
\end{aligned}$$

with $\pi_1(r), \pi_2(s) > 0$ for all r and s. The conditional densities are

$$p_{sr} = \Pr(U_1 = r \mid U_2 = s) = \frac{\pi(r, s)}{\pi_2(s)}, \quad q_{rs} = \Pr(U_2 = s \mid U_1 = r) = \frac{\pi(r, s)}{\pi_1(r)},$$

which we express as an $m \times n$ matrix P_{21} with (s, r) element p_{sr} and an $n \times m$ matrix P_{12} with (r, s) element q_{rs}. These transition matrices give the probabilities of going from the m possible values of U_2 to the n possible values of U_1 and back again. As they are ratios, p_{rs} and q_{rs} do not involve the normalizing constant for π.

If f_0 is an $m \times 1$ vector containing the distribution of $U_2^{(0)}$, the distributions of $U_1^{(1)}, U_2^{(1)}, U_1^{(2)}, \ldots$, are $f_0^{\mathrm{T}} P_{21}, f_0^{\mathrm{T}} P_{21} P_{12}, f_0^{\mathrm{T}} P_{21} P_{12} P_{21}, \ldots$. Thus each iteration of step 2 of the algorithm corresponds to postmultiplying the current distribution of $U_2^{(i)}$ by the $m \times m$ matrix $H = P_{21} P_{12}$. Hence $U_2^{(I)}$ has distribution $f_0^{\mathrm{T}} H^I$. Conditional on $U_2^{(i)}$, $U_2^{(i+1)}$ is independent of earlier values, so the sequence $U_2^{(1)}, \ldots, U_2^{(I)}$ is a Markov chain with transition matrix H. If the chain is ergodic, then $U_2^{(I)}$ has a unique limiting distribution f as $I \to \infty$, satisfying the equation $f^{\mathrm{T}} H = f^{\mathrm{T}}$. As this limit is unique, we need only show that f is the marginal distribution of U_2 to see that the algorithm ultimately

produces a variable with density π_2. Now the rth element of $\pi_2^{\mathrm{T}} H = \pi_2^{\mathrm{T}} P_{21} P_{12}$ equals

$$\sum_{t=1}^{n}\sum_{s=1}^{m} \pi_2(t) p_{ts} q_{sr} = \sum_{t=1}^{n}\sum_{s=1}^{m} \pi_2(t) \frac{\pi(s,t)}{\pi_2(t)} \frac{\pi(r,s)}{\pi_1(s)} = \pi_2(r),$$

so π_2 is indeed the unique solution to the equation $f^{\mathrm{T}} H = f^{\mathrm{T}}$. By symmetry, $U_1^{(1)}, \ldots, U_1^{(I)}$ is a Markov chain with transition matrix $P_{12}P_{21}$ and limiting distribution π_1. Moreover the fact that $\pi_2^{\mathrm{T}} P_{21} = \pi_1^{\mathrm{T}}$ ensures that the joint distribution of $(U_1^{(I)}, U_2^{(I)})$ converges to $\pi(r,s)$ as $I \to \infty$. Generalization to $k > 2$ works in an obvious way.

Most of the densities $\pi(u)$ met in applications are continuous, so this argument is not directly applicable. However any continuous density can be closely approximated by one with countable support, for which essentially the same results hold, so it is not surprising that the ideas apply more widely, and from now on we shall assume that they are applicable to our problems.

Such a simulation will only be useful if convergence to the stationary distribution is not too slow. In discrete cases like that above, the convergence rate is determined by the modulus of the second largest eigenvalue l_2 of H, where $1 = l_1 \geq |l_2| \geq \cdots$. If $|l_2| < 1$, then convergence is geometrically ergodic; see (6.4). In the continuous case it can occur that $|l_2| = 1$ or that l_2 does not exist, either of which will spell trouble. A reversible chain has real eigenvalues and satisfies the detailed balance condition (6.5). Hence it can be useful to make the chain reversible, for example by generating variables in order $1, \ldots, k, k-1, \ldots, 2, \ldots$ or by choosing the next update at random. Either involves modifying step 2 of the algorithm.

Output analysis

The only sure way to know how long a Markov chain simulation algorithm should be run is by theoretical analysis to determine its rate of convergence. This requires knowledge of the stationary distribution being estimated, however, and is possible only in very special cases. A more pragmatic approach is to declare that the algorithm has converged when its output satisfies tests of some sort. Such *convergence diagnostics* can at best detect nonconvergence, however; they cannot guarantee that the output will be useful. Both empirically- and theoretically-based diagnostics have been proposed, and references to them are given in the bibliographic notes. Empirical approaches include contrasting output from the start and the end of a run, and comparing results from parallel independent runs whose initial values have been chosen to be overdispersed relative to the target distribution. Theoretical approaches generally assess whether the output satisfies known properties of stationary chains. In practice it is sensible to use several diagnostics but also to scrutinize time series plots of the output. As different parameters may converge at

different rates, it is important to examine all parameters of interest and also global quantities such as the current log likelihood, prior, and posterior.

If stationarity seems to have been attained, then it is useful to examine correlograms and partial correlograms of output. If the autocorrelations are high, then the statistical efficiency of the algorithm will be low. A chain with low correlations will yield estimators with smaller variance, and is more likely to visit all regions of significant probability mass. The algorithm may need modification to reduce high autocorrelations, for example by reparametrization; see Example 11.24.

Multimodal target densities are awkward because it can be hard to know if all significant modes have been visited. Use of widely separated starting values may then be useful, and so too may be occasional insertion of large random jumps into the algorithm, so that it effectively restarts from a location unrelated to its previous position.

Suppose that the chain seems to have converged after B iterations and is run for a total of $I \gg B$ iterations. In general discussion below we suppose that I is so much larger than B that inference can safely be based on all I iterations, but in practice we use only output from iterations $B+1, \ldots, I$. Let the quantity of interest be $\mu = \int m(u)\pi(u)\,du$, where $\int |m(u)|\pi(u)\,du < \infty$. Unless there is qualitative knowledge about $\pi(u)$ this may involve an act of faith. For example, taking $m(u) = u_1$ gives $\mu = \mathrm{E}(U_1)$, which could be infinite although $\pi(u)$ is proper. Hence unless properties of the posterior density are known it is safer to base inferences on density and quantile estimates than on moments. If μ is finite then it can be estimated by the ergodic average

$$\widehat{\mu} = I^{-1} \sum_{i=1}^{I} m(U^{(i)}), \tag{11.39}$$

where $U^{(i)}$ denotes $(U_1^{(i)}, \ldots, U_k^{(i)})$. The ergodic theorem (6.2) implies that $\widehat{\mu}$ converges almost surely to μ as $I \to \infty$, and under further conditions

$$I^{1/2}\,(\widehat{\mu} - \mu) \quad \xrightarrow{D} \quad N(0, \sigma_m^2), \quad \text{where} \quad 0 < \sigma_m^2 < \infty, \tag{11.40}$$

so $\widehat{\mu}$ is approximately normal for large I. In that case

$$I \times \mathrm{var}(\widehat{\mu}) = I^{-1} \sum_{i=-I+1}^{I-1} (I - |i|)\,\gamma_i \sim \sigma_m^2 = \sum_{i=-\infty}^{\infty} \gamma_i = \gamma_0 \sum_{i=-\infty}^{\infty} \rho_i,$$

where $\gamma_i = \mathrm{cov}\left\{m(U^{(0)}), m(U^{(i)})\right\}$ depends on π and on the construction of the chain, and $\rho_i = \gamma_i/\gamma_0$ is the ith autocorrelation. The marginal variance of $m(U)$ is $\gamma_0 = \mathrm{var}_\pi\{m(U)\}$, which depends only on m and π. The effect of using correlated output is to inflate $\mathrm{var}(\widehat{\mu})$ by a factor $\tau = \sum_{-\infty}^{\infty} \rho_i$ relative to an independent sample of size I, so an estimate $\widehat{\tau}$ from a pilot run may suggest how large I should be. The obvious estimator of τ based on the correlogram

is inconsistent, but better ones exist. One simple possibility is $\widehat{\tau} = \sum_{i=-M}^{M} \widehat{\rho}_i$, where $M = \lfloor 3\widehat{\tau} \rfloor$ is found by iteration.

$\lfloor x \rfloor$ is the smallest integer greater than or equal to x.

Another approach splits the output into b blocks of k successive iterations, with k taken so large that the block averages of the $m(U^{(i)})$ have correlations lower than 0.05, say, and gives the standard error for $\widehat{\mu}$ as if the block averages were a simple random sample.

The density of U_1 at u_1 may be estimated by a kernel method (Section 7.1.2), or by the unbiased estimator (7.12), written in this context as

$$I^{-1} \sum_{i=1}^{I} \pi(u_1 \mid U_{-1}^{(i)}). \tag{11.41}$$

The discussion above presupposes a single long run of the chain. An alternative is S independent parallel runs of length I, leading ultimately to S independent values $U^{(I)}$ from $\pi(u)$. An estimate based on these may be less variable than one based on SI dependent samples from a single chain, and its variance is more easily estimated. Roughly SB iterations must be disregarded, however, compared to B when there is only one chain. From this viewpoint a single run is preferable, but it is then harder to detect lack of convergence.

Example 11.21 (Bivariate normal density) If (U_1, U_2) are bivariate normal with means zero, variances one and correlation ρ, then

ϕ denotes the standard normal density.

$$\pi(u_1 \mid u_2) = \frac{1}{(1-\rho^2)^{1/2}} \phi\left\{ \frac{u_1 - \rho u_2}{(1-\rho^2)^{1/2}} \right\},$$

with a symmetric result for $\pi(u_2 \mid u_1)$, and we can use the marginal standard normal densities of U_1 and U_2 to assess convergence. The upper left panel of Figure 11.6 shows the contours of the joint density when $\rho = 0.75$, together with a sample path of the process starting from an initial value generated uniformly on the square $(-4, 4) \times (-4, 4)$. The updating scheme forces the sample path to consist of steps parallel to the coordinate axes. The upper right panel shows that the sample paths of the Markov chains appear to converge rapidly to their limit distributions, as the calculations in Problem 11.20 show will be the case. This is confirmed by the estimated variance inflation factor $\widehat{\tau} \doteq 3$. The lower left panel shows rapid convergence of the kernel density estimates to their target, based on $S = 100$ parallel chains. The lower right panel illustrates the variability of (11.41), which here performs better than the kernel estimator. ■

Bayesian application

The essence of Bayesian inference is to treat all unknowns as random variables, and to compute their posterior distributions given the data y. The Gibbs sampler is applied by taking $U_1, \ldots, U_k$ to be the unknowns, usually parameters,

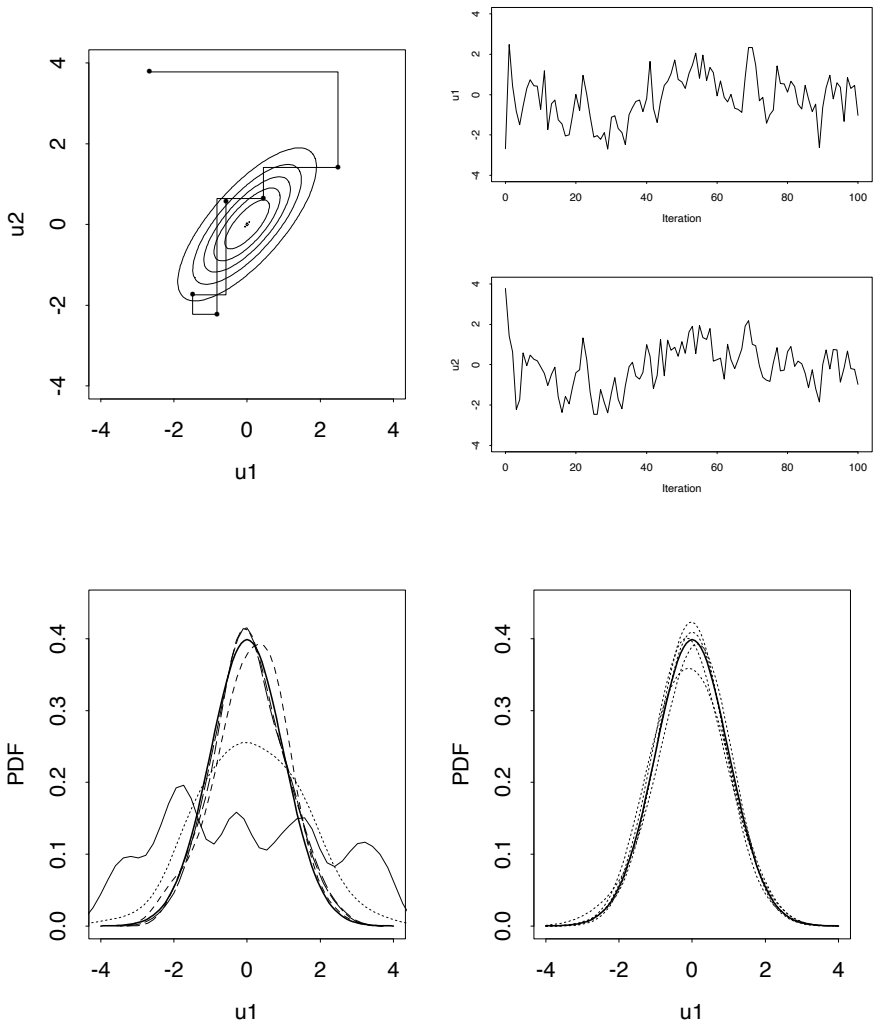


Figure 11.6 Gibbs sampler for bivariate normal density. Top left: contours of the bivariate normal density with $\rho = 0.75$, with the first five iterations of a Gibbs sampler; the blobs are at $(u_1^{(i)}, u_2^{(i)})$, for $i = 0, \ldots, 5$, starting from the top left of the panel. Top right: sample paths of $U_1^{(i)}$ and $U_2^{(i)}$ for $i = 1, \ldots, 100$. Bottom left: kernel density estimates of $\pi_1(u_1)$ (heavy solid) based on 100 parallel chains after I iterations, with $I = 0$ (solid), 2 (dots), 5 (dashes), 10 (large dashes), and 100 (largest dashes); the bandwidth is chosen by uniform cross-validation. Bottom right: estimates (dots) of $\pi_1(u_1)$ (heavy solid) after 100 iterations of 5 replicate chains, based on (11.41).

and simulating conditional on y. The full conditional densities $\pi(u_i \mid u_{-i})$ are typically of form $\pi(\theta_i \mid \theta_{-i}, y)$ and must be obtained before the algorithm can be applied. Fortunately this is often possible for 'nice' models, where the full conditional densities have conjugate forms.

Example 11.22 (Random effects model) The sampling model in the simplest normal one-way layout is

$$y_{tr} = \theta_t + \varepsilon_{tr}, \quad t = 1, \ldots, T, \ r = 1, \ldots, R,$$

where $\theta_1, \ldots, \theta_T \overset{\text{iid}}{\sim} N(\nu, \sigma_\theta^2)$ and independent of this $\varepsilon_{tr} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The focus of interest is usually σ^2 and σ_θ^2.

Bayesian analysis requires prior information, which we suppose to be expressed through the conjugate densities

$$\mu \sim N(\mu_0, \tau^2), \quad \sigma^2 \sim IG(\alpha, \beta), \quad \sigma_\theta^2 \sim IG(\alpha_\theta, \beta_\theta).$$

The full posterior density is then

$$\pi(\mu, \theta, \sigma^2, \sigma_\theta^2 \mid y) \propto f(y \mid \theta, \sigma^2) f(\theta \mid \mu, \sigma_\theta^2) \pi(\mu) \pi(\sigma^2) \pi(\sigma_\theta^2). \tag{11.42}$$

We now take $(U_1, U_2, U_3, U_4) = (\sigma_\theta^2, \sigma^2, \mu, \theta)$, and calculate the full conditional densities needed for Gibbs sampling, always treating the data y as fixed. Each calculation requires integration over just one parameter. For example,

$$\begin{aligned} \pi(\sigma_\theta^2 \mid \sigma^2, \mu, \theta, y) &= \frac{f(y \mid \theta, \sigma^2) f(\theta \mid \mu, \sigma_\theta^2) \pi(\mu) \pi(\sigma^2) \pi(\sigma_\theta^2)}{\int f(y \mid \theta, \sigma^2) f(\theta \mid \mu, \sigma_\theta^2) \pi(\mu) \pi(\sigma^2) \pi(\sigma_\theta^2)\, d\sigma_\theta^2} \\ &= \frac{f(\theta \mid \mu, \sigma_\theta^2) \pi(\mu) \pi(\sigma_\theta^2)}{\int f(\theta \mid \mu, \sigma_\theta^2) \pi(\mu) \pi(\sigma_\theta^2)\, d\sigma_\theta^2} \\ &= \pi(\sigma_\theta^2 \mid \mu, \theta). \end{aligned}$$

Similar calculations reveal that $\pi(\theta \mid \sigma_\theta^2, \sigma^2, \mu, y)$ does not simplify, but that

$$\pi(\sigma^2 \mid \sigma_\theta^2, \mu, \theta, y) = \pi(\sigma^2 \mid \theta, y), \quad \pi(\mu \mid \sigma_\theta^2, \sigma^2, \theta, y) = \pi(\mu \mid \sigma_\theta^2, \theta). \tag{11.43}$$

Arguments paralleling those in Example 11.12 lead to

$$\sigma_\theta^2 \mid \mu, \theta \sim IG\left(\alpha_\theta + \tfrac{1}{2}T, \beta_\theta + \tfrac{1}{2}\sum_{t=1}^{\mathrm{T}} (\theta_t - \mu)^2\right), \tag{11.44}$$

$$\sigma^2 \mid \theta, y \sim IG\left(\alpha + \tfrac{1}{2}TR, \beta + \tfrac{1}{2}\sum_{t=1}^{\mathrm{T}} \sum_{r=1}^{R} (y_{tr} - \theta_t)^2\right), \tag{11.45}$$

$$\mu \mid \sigma_\theta^2, \theta \sim N\left(\frac{\sigma_\theta^2 \mu_0 + \tau^2 \sum_{t=1}^{\mathrm{T}} \theta_t}{\sigma_\theta^2 + T\tau^2}, \frac{\sigma_\theta^2 \tau^2}{\sigma_\theta^2 + T\tau^2}\right). \tag{11.46}$$

The conditional density $\pi(\theta \mid \sigma_\theta^2, \sigma^2, \mu, y)$ is most readily calculated by noting that given μ, σ_θ^2 and σ^2, the statistic $\overline{y}_t$ is sufficient for θ_t, with distribution $N(\theta_t, \sigma^2/R)$, while the prior density for θ_t given σ_θ^2, σ^2, and μ is $N(\mu, \sigma_\theta^2)$. Hence the posterior density for θ_t is

$$\theta_t \mid \sigma_\theta^2, \sigma^2, \mu, y \sim N\left(\frac{R\sigma_\theta^2 \overline{y}_t + \sigma^2 \mu}{R\sigma_\theta^2 + \sigma^2}, \frac{\sigma_\theta^2 \sigma^2}{R\sigma_\theta^2 + \sigma^2}\right), \quad t = 1, \ldots, T, \tag{11.47}$$

and the θ_t are conditionally independent.

Expressions (11.44)–(11.47) give the steps required for an iteration of the Gibbs sampler. As the T updates in (11.47) are independent, they may all be performed at once, if the programming language used permits simultaneous generation of several non-identically-distributed normal variates.

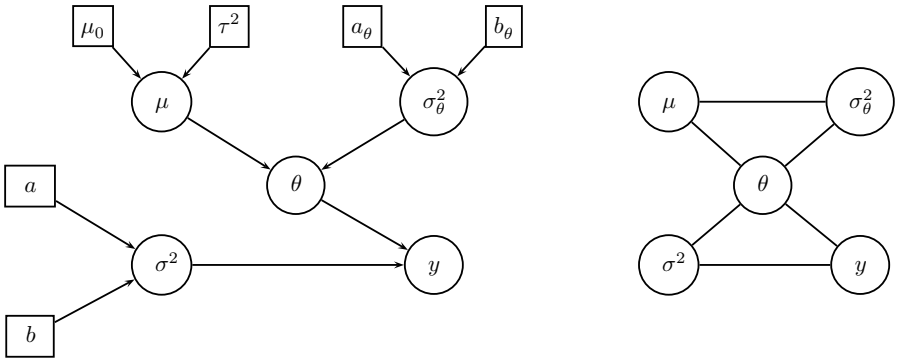


Figure 11.7 Graphs for random effects model of Example 11.22. Left: directed acyclic graph showing dependence of random variables (circles) on themselves and on fixed quantities (rectangles). Right: additional dependence graph, formed by moralizing the directed acyclic graph, that is, joining parents and dropping arrowheads.

	σ^2_θ	σ^2	μ	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
Estimate	23.8	126.4	41.9	53.9	43.0	34.9	39.9	41.3	38.6
Posterior mean	17.1	138.0	41.9	45.8	42.3	39.6	41.2	41.7	40.8
Posterior SD	30.3	33.8	2.4	4.1	2.9	3.4	2.9	2.9	3.0

Table 11.9 Estimated posterior means and standard deviations for the model fitted to the blood data, and simple frequentist estimates from analysis of variance.

Ideas from Section 6.2.2 render the structure of the full conditional densities more intelligible. Figure 11.7 shows the directed acyclic graph and the corresponding conditional independence graph for the present model. Each of μ, σ^2_θ, and σ^2 has two hyperparameters, considered fixed, and μ and σ^2_θ are parents of $\theta_1, \ldots, \theta_T$. Each iteration of the Gibbs sampler traverses the parameter nodes in the conditional independence graph, simulating from the full conditional distribution corresponding to each node with remaining parameters set at their current values. The data y are held fixed throughout.

We applied this algorithm to the data in Table 9.22 on the stickiness of blood. For illustration we took $\alpha = \alpha_\theta = 0.5$, $\beta = \beta_\theta = 1$, $\mu = 0$, and $\tau^2 = 1000$, and generated starting-values for the parameters from the uniform distribution on $(0, 100)$. We ran 25 independent chains with $I = 1000$.

Figure 11.8 shows simulated series for three parameters and estimates of their posterior densities. The burn-in period seems to last for about $B = 100$ iterations, after which the chains seem stable. The chain for σ^2_θ makes some large positive excursions, but the others seem fairly homogeneous, though they both show fairly strong autocorrelations. Estimated variance inflation factors are about 10 for σ^2_θ and μ, but only 1–2.5 for the other parameters, consistent with the top left panels of the figure.

Table 11.9 shows the posterior means and standard deviations for the parameters, with their frequentist estimates. The posterior mean for μ is essentially equal to the overall average $\overline{y}$, but the posterior densities of the θ_t

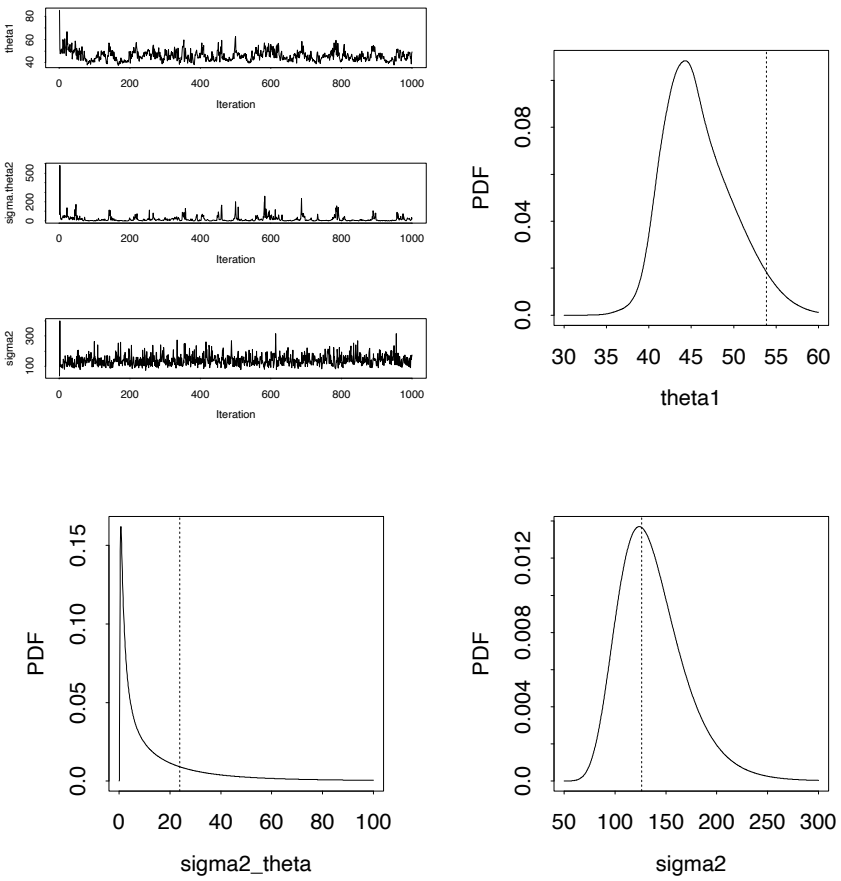


Figure 11.8 Gibbs sampler for normal components of variance model and blood data. Top left: time plots of θ_1, σ_θ^2, and σ^2. The other panels show estimated posterior densities for these parameters, based on applying analogues of (11.41) to the last 200 estimates from each of 25 parallel chains of length 1000. Frequentist estimates are shown as the dotted vertical lines.

are strongly shrunk towards it, because there is evidence that σ_θ^2 is small; its posterior 0.1, 0.5, and 0.9 quantiles are 0.46, 7.1, and 42.1. The variability mostly comes from measurement error, not inter-subject variation. ■

Metropolis–Hastings algorithm

The Gibbs sampler is easy to program, but if the full conditional densities it involves are unavailable or too nasty then a more general algorithm may be needed. A powerful approach known as the *Metropolis–Hastings algorithm* works as follows. In order to update the current value u of a Markov chain, a new value u' is generated using a proposal density $q(u' \mid u)$. Any density q can be used provided $q(u' \mid u) > 0$ if and only if $q(u \mid u') > 0$ and the resulting chain has the properties desired. Having generated u', a move from u to u' is

accepted with probability

$$a(u,u') = \min\left\{1, \frac{\pi(u')q(u \mid u')}{\pi(u)q(u' \mid u)}\right\},$$

but otherwise the chain remains at u. Hence the probability density for a move to u', given that the chain has current value u, is

δ denotes the Dirac delta function.

$$p(u' \mid u) = q(u' \mid u)a(u,u') + r(u)\delta(u-u'),$$

where

$$r(u) = 1 - \int q(v \mid u)a(u,v)\,dv.$$

The first and second terms of $p(u' \mid u)$ are the probability density for a move from u to u' being proposed and accepted, and the probability that a move away from u is rejected.

The Metropolis–Hastings update step satisfies the detailed balance condition (6.5), because

$$\begin{aligned}
\pi(u)p(u' \mid u) &= \pi(u)q(u' \mid u)\min\left\{1, \frac{\pi(u')q(u \mid u')}{\pi(u)q(u' \mid u)}\right\} + \pi(u)r(u)\delta(u-u') \\
&= \pi(u')q(u \mid u')\min\left\{\frac{\pi(u)q(u' \mid u)}{\pi(u')q(u \mid u')}, 1\right\} + \pi(u')r(u')\delta(u'-u) \\
&= \pi(u')p(u \mid u').
\end{aligned}$$

Hence the corresponding Markov chain is reversible with equilibrium distribution π, provided it is irreducible and aperiodic. As π appears only in a ratio $\pi(u')/\pi(u)$ in the acceptance probability $a(u,u')$, the algorithm requires no knowledge of the constant that normalizes π.

If $q(u' \mid u) = q(u \mid u')$, the kernel is called symmetric, and $a(u,u') = \min\{1, \pi(u')/\pi(u)\}$. This occurs in particular if $u' = u + \varepsilon$, where ε is symmetric with density g; then $q(u' \mid u) = g(u'-u) = g(u-u') = q(u \mid u')$. This is called *random walk Metropolis* sampling. It is often applied to transformations of u, or to subsets of its elements, using a different proposal distribution for each subset.

The Gibbs sampler is a form of Metropolis–Hastings algorithm, the proposal density at the ith step of an iteration being

$$q(u' \mid u) = \begin{cases} \pi(u_i' \mid u_{-i}), & u'_{-i} = u_{-i}, \\ 0, & \text{otherwise.} \end{cases}$$

It then follows that

$$\frac{\pi(u')q(u \mid u')}{\pi(u)q(u' \mid u)} = \frac{\pi(u')/\pi(u_i' \mid u_{-i})}{\pi(u)/\pi(u_i \mid u'_{-i})} = \frac{\pi(u')/\pi(u_i' \mid u'_{-i})}{\pi(u)/\pi(u_i \mid u_{-i})} = \frac{\pi(u'_{-i})}{\pi(u_{-i})} = 1,$$

because $u'_{-i} = u_{-i}$. Here the proposals always have $u'_{-i} = u_{-i}$ and are always accepted, because $a(u,u') = \min[1, \pi(u')q(u \mid u')/\{\pi(u)q(u' \mid u)\}] = 1$.

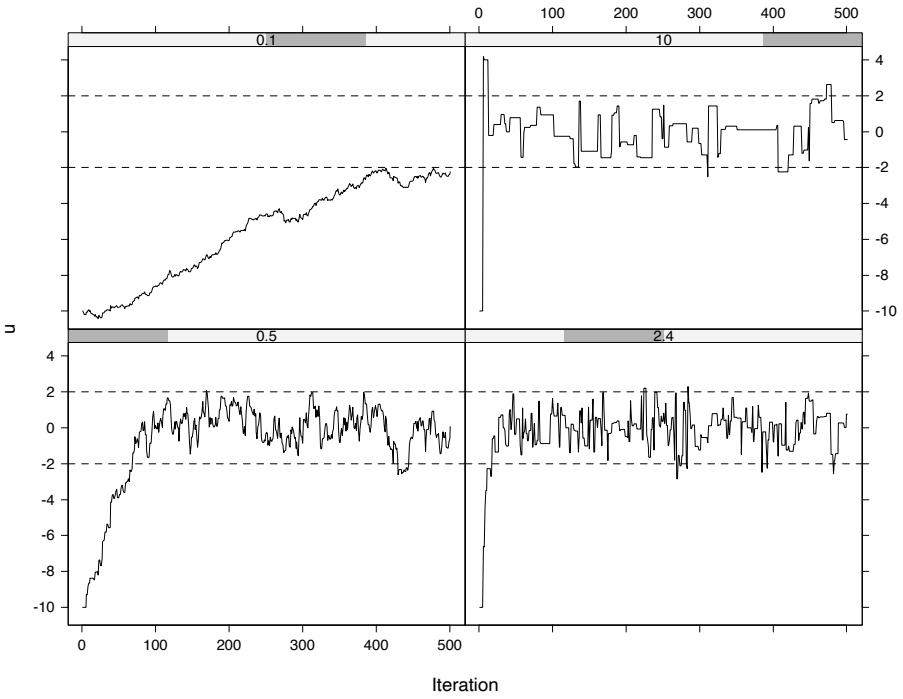


Figure 11.9
Sample paths for Metropolis–Hastings algorithm. The stationary density is standard normal and the proposal density $q(u' \mid u)$ is $N(u, \sigma^2)$, with $\sigma = 0.1$, 0.5, 2.4 and 10. The initial value is $u_0 = -10$ and the same seed is used for the random number generator in each case. Note the dependence of the acceptance rate and convergence to stationarity on σ. The horizontal dashed lines show the 'usual' range for u.

Although there are few theoretical restrictions on the choice of q, practical constraints intervene. For example, if $q(u' \mid u)$ is so chosen that the acceptance probability $a(u, u')$ is essentially zero, the chain will spend long periods without moving and its output will be useless, and if the acceptance probability is close to one at each step but the chain barely moves, the state space will be traversed too slowly. Hence it is important to balance a reasonably high acceptance probability $a(u, u')$ with a chain that moves around its state space quickly enough. This can demand creativity and patience from the programmer.

Example 11.23 (Normal density) For illustration we take the toy problem of using the Metropolis–Hastings algorithm to simulate from the standard normal density $\phi(u) = \pi(u)$. The proposal density, $q(u' \mid u) = \sigma^{-1}\phi\{(u' - u)/\sigma\}$, depends on σ. We take initial value $u_0 = -10$ far from the centre of the stationary distribution. As $q(u' \mid u) = q(u \mid u')$, the acceptance probability is $a(u, u') = \min\{1, \phi(u')/\phi(u)\}$.

Figure 11.9 shows sample paths $u_0, \ldots, u_{500}$ for four values of σ. When $\sigma = 0.1$, only small steps occur but they are accepted with high probability because $\phi(u')/\phi(u) \doteq 1$. Although u changes at almost every step, it moves so little that the chain has not reached equilibrium after 500 iterations. When $\sigma = 0.5$ it takes 100 or so iterations to reach convergence and the chain then

ble 11.10
torette data
elson and Hahn,
72). Censored
ure times are
oted by +.

x (° F)	Failure time (hours)									
150	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+	8064+
170	1764	2772	3444	3542	3780	4860	5196	5448+	5448+	5448+
190	408	408	1344	1344	1440	1680+	1680+	1680+	1680+	1680+
220	408	408	504	504	504	528+	528+	528+	528+	528+

appears to mix fairly fast. When $\sigma = 2.4$ convergence is almost immediate but as the acceptance probability is lower the chain tends to get stuck for slightly longer. When $\sigma = 10$ the acceptance probability is low and although the chain jumps to its stationary range almost at once, it spends long periods without moving.

For comparison the experiment above was repeated 50 times, and the estimated means of $\pi(u)$ were compared. The estimator was the average of the last half of $u_0, \ldots, u_I$, with $I = 500$ iterations; that is, (11.39) with $m(u) = u$ and $B = 250$. Each of the 50 replicates used the same seed and initial value u_0 for each σ; the values of u_0 were generated from the t_5 density. The estimated values of σ_m^2 in (11.40) were 170, 17.7, 6.2, and 8.0 for $\sigma = 0.1$, 0.5, 2.4, and 10; the larger values of σ are preferable, but there is a large efficiency loss relative to the value $\sigma_m^2 = 1$ for independent sampling. This is because of the serial correlations of $u_{B+1}, \ldots, u_I$, which were roughly 0.97, 0.89, 0.62, and 0.83 for $\sigma = 0.1$, 0.5, 2.4, and 10.

Exercise 11.3.11 sheds more light on this example. ■

Example 11.24 (Motorette data) Table 11.10 contains failure times y_{ij} from an accelerated life trial in which ten motorettes were tested at each of four temperatures, with the objective of predicting lifetime at 130°F. We analyse these data using a Weibull model with

$$\Pr(Y_{ij} \le y; x_i) = 1 - \exp\{(y/\theta_i)^\gamma\}, \quad \theta_i = \exp(\beta_0 + \beta_1 x_i), \tag{11.48}$$

for $i = 1, \ldots, 4$, $j = 1, \ldots, 10$, where failure time is taken in units of hundreds of hours and x_i is log(temperature/100).

Here we describe a simple Bayesian analysis using the Metropolis–Hastings algorithm. For illustration we take independent priors on the parameters, $N(0, 100)$ on β_0 and β_1 and exponential with mean 2 on γ. Then the log posterior is

$$\ell_m(\beta_0, \beta_1, \gamma) \equiv -(\beta_0^2+\beta_1^2)/200 - \gamma/2 + \sum_{i=1}^{4}\sum_{j=1}^{10} d_{ij}\{\log\gamma + \gamma\log(y_{ij}/\theta_i)\} - (y_{ij}/\theta_i)^\gamma,$$

where $d_{ij} = 0$ for uncensored y_{ij}.

For proposal distribution we update all three parameters simultaneously,

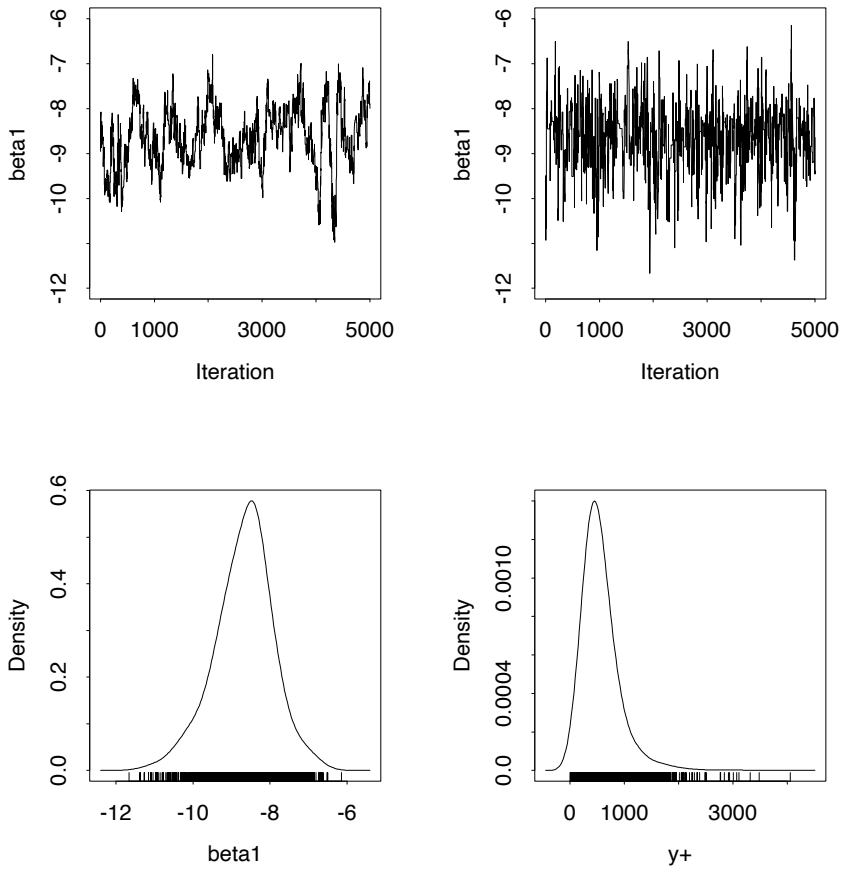


Figure 11.10
Bayesian analysis of motorette data using Metropolis–Hastings algorithm. Upper panels: sample paths for β_1 using two parametrizations, the right one more nearly orthogonal. Lower left: kernel density estimates of $\pi(\beta_1 \mid y)$ and of $\pi(Y_+ \mid y)$, where Y_+ is failure time predicted for 130°F.

by taking $(\beta_0', \beta_1', \log\gamma') = (\beta_0, \beta_1, \log\gamma) + c(s_1 Z_1, s_2 Z_2, s_3 Z_3)$, where the s_r are the standard errors of the corresponding maximum likelihood estimates, $Z_r \overset{\text{iid}}{\sim} N(0,1)$, and c can be chosen to balance the acceptance probability and the size of the move. The ratio $q(u \mid u')/q(u' \mid u)$ reduces to γ'/γ, so the acceptance probability equals

$$a\{(\beta_0', \beta_1', \gamma'), (\beta_0, \beta_1, \gamma)\} = \min\left[1, \exp\{\ell_m(\beta_0, \beta_1, \gamma) - \ell_m(\beta_0', \beta_1', \gamma')\}\,\gamma'/\gamma\right].$$

The chain is clearly irreducible and aperiodic, so the ergodic theorem applies.

We take initial values near the maximum likelihood estimates, and run the chain for 5000 iterations with $c = 0.5$. The sample path for β_1 in the upper left panel of Figure 11.10 shows that despite its acceptance probability of about 0.3, the chain is not moving well over the parameter space. This is confirmed by the correlogram and partial correlogram for successive values of β_1, which

suggest that the chain is essentially an AR(1) process with $\rho_1 \doteq 0.99$. In this case the variance inflation factor is $\widehat{\tau} = 199$, so 5000 successive observations from the chain are worth about 25 independent observations. Sample paths for the other parameters are similar, and varying c does not improve matters. One reason for this is that β_0 and β_1 have correlation about -0.97 *a posteriori*, and the proposal distribution does not respect this. It is better to reduce this correlation by replacing x by $x - \overline{x}$, after which $\operatorname{corr}(\beta_0, \beta_1 \mid y) \doteq -0.4$. The sample path for β_1 from a run of the algorithm starting near the new maximum likelihood estimates, with the new s_r and with $c = 2$, is shown in the upper right panel of Figure 11.10. This chain mixes much better, though its acceptance probability is about 0.2. The usual plots suggest that β_1 follows an AR(1) process with $\rho \doteq 0.9$, and likewise for the other parameters, whose chains show similar good behaviour. Here $\widehat{\tau}$ has the more acceptable value 19, though 5000 iterations would remain too small in practice.

The lower panels of the figure show kernel density estimates of the posterior densities for β_1 and for a predicted failure time Y_+ for temperature 130°F. Once convergence has been verified, it is easy to obtain values for Y_+, simply by simulating a Weibull variable from (11.48) using the current parameter values at each iteration. Quantiles of the simulated distributions may be used to obtain posterior confidence intervals for the corresponding quantities.

The Metropolis–Hastings update described above changes all three parameters on each iteration, or none of them. Alternatively we may attempt to update one parameter, chosen at random. The resulting chain is also ergodic, but it does not improve on the second approach described above. ■

Metropolis–Hastings updates using an appropriate proposal distribution can be used when the full conditional densities needed for particular steps of the Gibbs sampler are not available. Generalizations can be constructed to jump between spaces of differing dimensions, and these are valuable in applications where averaging over various spaces or choosing among them is important. More details are given in the bibliographic notes.

Exercises 11.3

1 Show that Laplace approximation to the gamma function

$$I_{\alpha+1} = \Gamma(\alpha + 1) = \int_0^\infty u^\alpha e^{-u}\, du$$

gives *Stirling's formula*, $\Gamma(\alpha + 1) \doteq \tilde{I}_{\alpha+1} = (2\pi)^{1/2}\alpha^{\alpha+1/2}e^{-\alpha}$, and verify that the $O(\alpha^{-1})$ term in (11.28) is $(12\alpha)^{-1}$. Show that this can be incorporated by modifying $\tilde{I}_{\alpha+1}$ to $\tilde{I}'_{\alpha+1} = (2\pi)^{1/2}(\alpha + \frac{1}{6})^{1/2}\alpha^\alpha e^{-\alpha}$, and check some of the numbers in Table 11.11.

2 Use the facts that if Z is a standard normal variable, $\mathrm{E}(Z^4) = 3$ and $\mathrm{E}(Z^6) = 15$,

Table 11.11 Accuracy of Stirling's formula and related approximations.

α	0.5	1	2	3	4	5
$I_{\alpha+1}$	0.8862	1	2	6	24	120
$\tilde{I}_{\alpha+1}/I_{\alpha+1}$	0.8578	0.9221	0.9595	0.9727	0.9794	0.9834
$\tilde{I}'_{\alpha+1}/I_{\alpha+1}$	0.9905	0.9960	0.9987	0.9994	0.9996	0.9998

to check (11.28). Use properties of normal moments to explain why (11.28) is an expansion with terms in increasing powers of n^{-1} rather than $n^{-1/2}$.

3 Let $f(y;\theta)$ be a unimodal density with mode at $\tilde{y}_\theta$. Show that $\int_{-\infty}^{y} f(u;\theta)\,du$ may be approximated by (11.31), with

$$g(u) = \log f(\tilde{y}_\theta;\theta) - \log f(u;\theta), \quad a(u) = (2\pi)^{1/2} f(\tilde{y}_\theta;\theta),$$

and verify that the approximation is exact for the $N(\theta,\sigma^2)$ density. Investigate its accuracy numerically for the gamma density with shape parameter $\theta > 1$, and for the t_ν density.

4 Consider predicting the outcome of a future random variable Z on the basis of a random sample $Y_1,\ldots,Y_n$ from density $\lambda^{-1}e^{-u/\lambda}$, $u>0$, $\lambda>0$. Show that $\pi(\lambda)\propto\lambda^{-1}$ gives posterior predictive density

$$f(z \mid y) = \frac{\int f(z,y \mid \lambda)\pi(\lambda)\,d\lambda}{\int f(y \mid \lambda)\pi(\lambda)\,d\lambda} = ns^n/(s+z)^{n+1}, \quad z>0,$$

where $s = y_1+\cdots+y_n$.
Show that when Laplace's method is applied to each integral in the predictive density the result is proportional to the exact answer, and assess how close the approximation is to a density when $n=5$.

5 Consider the integral

$$I_n = \int_{u_1}^{u_2} e^{-nh(u)}\,du,$$

where $h(u)$ is a smooth increasing function with minimum at u_1, at which point its derivatives are $h_1 = h'(u_1) > 0$, $h_2 = h''(u_1)$ and so forth. Show that

$$I_n = \frac{1}{nh_1} e^{-nh(u_1)} \left\{1 - e^{-nh_1(u_2-u_1)} + O(n^{-1})\right\},$$

and deduce that

$$\int_{u_1}^{u_2} e^{-nh(u)}\,du \Big/ \int_{u_1}^{\infty} e^{-nh(u)}\,du \doteq 1 - e^{-nh_1(u_2-u_1)}.$$

A posterior density has form $\pi(\theta \mid y)\propto\theta^{-m-1}$, for $\theta>\theta_1$ (Exercise 11.2.2). Find the approximate and exact posterior density and distribution functions of θ, and compare them numerically when $m=5,10,20$ and $\theta_1=1$. Discuss. Investigate how the approximation will change if $h_1=0$.

6 Give an approximate variance for the importance sampling estimator (11.38), and verify the formula for $\mathrm{var}(\widehat{\mu}_{\mathrm{rat}})$.

7 *Sampling-importance resampling* (SIR) works as follows: instead of using (11.38) as an estimator of μ, an independent sample $\theta_1^*, \dots \theta_Q^*$ of size $Q \ll S$ is taken from $\theta_1, \dots, \theta_S$ with probabilities proportional to $w(\theta_1), \dots, w(\theta_S)$. The estimator of μ is $\widehat{\mu}^* = Q^{-1} \sum \theta_q^*$.
(a) Discuss SIR critically when the initial sample is taken from the prior $\pi(\theta)$; this is sometimes called the *Bayesian bootstrap*. Give an explicit discussion in the case of an exponential family model and conjugate prior.
(b) Show that $\mathrm{E}^*(\widehat{\mu}^*) = \widehat{\mu}_{\mathrm{rat}}$, and find its variance. Use the Rao–Blackwell theorem to show that the variance of $\widehat{\mu}^*$ exceeds that of $\widehat{\mu}_{\mathrm{rat}}$.
Under what circumstances would it be sensible to use SIR anyway?
(Rubin, 1987; Smith and Gelfand, 1992; Ross, 1996)

8 Show that the Gibbs sampler with $k > 2$ components updated in order

$$1, \dots, k, 1, \dots, k, 1, \dots, k, \dots$$

is not reversible. Are samplers updated in order $1, \dots, k, k-1, \dots, 1, 2, \dots$, or in a random order reversible?

9 Show that the acceptance probability for a move from u to u' when random walk Metropolis sampling is applied to a transformation $v = v(u)$ of u is

$$\min\left\{1, \frac{\pi(u')|dv/du|}{\pi(u)|dv'/du'|}\right\}.$$

Hence verify the form of $q(u \mid u')/q(u' \mid u)$ given in Example 11.24.
Find the acceptance probability when a component of u takes values in (a, b), and a random walk is proposed for $v = \log\{(u-a)/(b-u)\}$.

10 Suppose that $Y_1, \dots, Y_n$ are taken from an AR(1) process with innovation variance σ^2 and correlation parameter ρ such that $|\rho| < 1$. Show that

$$\mathrm{var}(\overline{Y}) = \frac{\sigma^2}{n^2(1-\rho^2)}\left\{n + 2\sum_{j=1}^{n-1}(n-j)\rho^j\right\},$$

and deduce that as $n \to \infty$ for any fixed ρ, $n\mathrm{var}(\overline{Y}) \to \sigma^2/(1-\rho)^2$.
What happens when $|\rho| = 1$?
Discuss estimation of $\mathrm{var}(\overline{Y})$ based on $(n-1)^{-1}\sum(Y_j - \overline{Y})^2$ and an estimate $\widehat{\rho}$.

11 In Example 11.23, show that the probability of acceptance of a move starting from $u > 0$ equals

$$\tfrac{1}{2} + (1+\sigma^2)^{-1/2}\exp\left(a^2/2\right)\{\Phi(a) + \Phi(b)\} - \Phi(-2u/\sigma),$$

where

$$a = -\frac{\sigma u}{\sqrt{1+\sigma^2}}, \quad b = \frac{-(2+\sigma^2)u}{\sqrt{\sigma^2(1+\sigma^2)}}.$$

Show that the expected move size may be written as

$$\exp\left(\frac{a^2}{2}\right)\left[\frac{\sigma}{1+\sigma^2}\{\phi(a) - \phi(b)\} - \frac{\sigma^2 u}{(1+\sigma^2)^{3/2}}\{\Phi(a) + \Phi(b)\}\right] + \sigma\left\{\phi\left(\frac{-2u}{\sigma}\right) - \phi(0)\right\}.$$

Plot these functions over the range $0 \le u \le 15$ for $\sigma = 0.1, 1, 2.4, 10$, and also with $0 \le \sigma \le 10$ for $u = 0, 1, 2, 3, 10$. What light do these plots cast on the behaviour of the chains in Figure 11.9?

11.4 Bayesian Hierarchical Models

Hierarchical models are useful when data have layers of variation. The incidence of a disease may vary from region to region of a country, for instance, while within regions there is variation due to differences in poverty, pollution, or other factors. If the regional and local incidence rates are regarded as random, we can imagine a hierarchy in which the numbers of diseased persons depend on random local rates, which themselves depend on random regional rates. Such models were discussed briefly from a frequentist viewpoint in Section 9.4. Here we outline the Bayesian approach, using the notion of exchangeability.

The random variables $U_1, \ldots, U_n$ are called *finitely exchangeable* if their density has the property

$$f(u_1, \ldots, u_n) = f\left(u_{\xi(1)}, \ldots, u_{\xi(n)}\right)$$

for any permutation ξ of the set $\{1, \ldots, n\}$. Then the density is completely symmetric in its arguments and in probabilistic terms the $U_1, \ldots, U_n$ are indistinguishable; this does not mean that they are independent. An infinite sequence $U_1, U_2, \ldots$, is called *infinitely exchangeable* if every finite subset of it is finitely exchangeable.

Bruno de Finetti (1906–1985) was born in Innsbruck and studied in Milan and Rome, where he eventually became professor, after working in Trieste as an actuary and at the University of Padova. His main contribution to statistics was to develop personalistic probability, teaching that 'probability does not exist'. (You may think this should have been made clear on page 1 of the book!) He argued that probability distributions express a person's view of the world, with no objective force. His ideas have strongly influenced Bayesian thought.

A key result in this context is *de Finetti's theorem*, whose simplest form says that if $U_1, U_2, \ldots$, is an infinitely exchangeable sequence of binary variables, taking values $u_j = 0, 1$, then for any n there is a distribution G such that

$$f(u_1, \ldots, u_n) = \int_0^1 \prod_{j=1}^{n} \theta^{u_j} (1-\theta)^{1-u_j} \, dG(\theta) \tag{11.49}$$

where

$$G(\theta) = \lim_{m\to\infty} \Pr\left\{m^{-1}(U_1 + \cdots + U_m) \le \theta\right\}, \quad \theta = \lim_{m\to\infty} m^{-1}(U_1 + \cdots + U_m).$$

This is justified at the end of this section. It implies that any set of exchangeable binary variables $U_1, \ldots, U_n$ may be modelled as if they were independent Bernoulli variables, conditional on their success probability θ, this having distribution G and being interpretable as the long-run proportion of successes. More general versions of (11.49) hold for real U_j, for example. The upshot is that a judgement that certain quantities are exchangeable implies that they may be represented as a random sample conditional on a variable that itself has a distribution. This provides the basis of a case in favour of Bayesian inference, because it implies that the conditional density $\Pr(U_{n+1} \mid U_1, \ldots, U_n)$ for

a future variable U_{n+1} given the outcomes of $U_1, \ldots, U_n$, may be represented as a ratio of two integrals of form (11.49), and this is formally equivalent to Bayesian prediction using a prior density on θ.

The essence of hierarchical modelling is to treat not data but particular sets of parameters as exchangeable. For if our model contains parameters $\theta_1, \ldots, \theta_n$, and if we believe *a priori* that these are to be treated completely symmetrically, then they are exchangeable and may be thought of as a random sample from a distribution that is itself unknown. In principle that distribution might be anything, but in practice a tractable one is often chosen.

Example 11.25 (Normal hierarchical model) A prototypical case is the normal model under which $y_1, \ldots, y_n$ satisfy

$$y_j \mid \theta_j \overset{\text{ind}}{\sim} N(\theta_j, v_j), \quad \theta_1, \ldots, \theta_n \mid \mu \overset{\text{iid}}{\sim} N(\mu, \sigma^2), \quad \mu \sim N(\mu_0, \tau^2),$$

where $v_1, \ldots, v_n$, σ^2, μ_0 and τ^2 are known; the last two are *hyperparameters* that control the uncertainty injected at the top level of the hierarchy. The y_j have different variances, but their means θ_j are supposed indistinguishable and hence are modelled as exchangeable, being normal with unknown mean μ. As the joint density of $(\mu, \theta^{\mathrm{T}}, y^{\mathrm{T}})^{\mathrm{T}}$ is multivariate normal of dimension $2n+1$, with mean vector and covariance matrix

$$\mu_0 1_{2n+1}, \quad \begin{pmatrix} \tau^2 & \tau^2 1_n^{\mathrm{T}} & \tau^2 1_n^{\mathrm{T}} \\ \tau^2 1_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n \\ \tau^2 1_n & \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n & V + \tau^2 1_n 1_n^{\mathrm{T}} + \sigma^2 I_n \end{pmatrix}, \tag{11.50}$$

where $V = \operatorname{diag}(v_1, \ldots, v_n)$, the posterior density of $(\mu, \theta^{\mathrm{T}})^{\mathrm{T}}$ given y is also normal. Unenlightening matrix calculations give

$$\mathrm{E}(\mu \mid y) = \frac{\mu_0/\tau^2 + \sum y_j/(\sigma^2 + v_j)}{1/\tau^2 + \sum 1/(\sigma^2 + v_j)}, \ \mathrm{var}(\mu \mid y) = \frac{1}{1/\tau^2 + \sum 1/(\sigma^2 + v_j)},$$

and

$$\mathrm{E}(\theta_j \mid y) = \mathrm{E}(\mu \mid y) + \frac{\sigma^2}{\sigma^2 + v_j} \{y_j - \mathrm{E}(\mu \mid y)\}.$$

The posterior mean of μ is a weighted average of its prior mean μ_0 and of the y_j, weighted according to their precisions conditional on μ. Typically τ^2 is very large, and then $\mathrm{E}(\mu \mid y)$ is essentially a weighted average of the data. Even when $v_j \to 0$ for all j there is still posterior uncertainty about μ, whose variance is σ^2/n because $y_1, \ldots, y_n$ is then a random sample from $N(\mu, \sigma^2)$.

The posterior mean of θ_j is a weighted average of y_j and $\mathrm{E}(\mu \mid y)$, showing shrinkage of y_j towards $\mathrm{E}(\mu \mid y)$ by an amount that depends on v_j. As $v_j \to 0$, $\mathrm{E}(\theta_j \mid y) \to y_j$, while as $v_j \to \infty$, $\mathrm{E}(\theta_j \mid y) \to \mathrm{E}(\mu \mid y)$. This is a characteristic feature of hierarchical models, in which there is a 'borrowing of strength' whereby all the data combine to estimate common parameters such as μ, while estimates of individual parameters such as the θ_j are shrunk towards

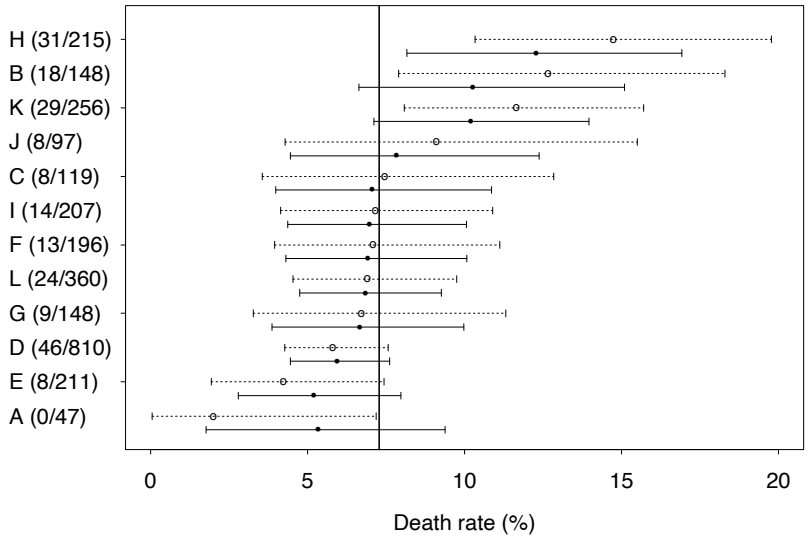


Figure 11.11
Posterior summaries for mortality rates for cardiac surgery data. Posterior means and 0.95 equitailed credible intervals for separate analyses for each hospital are shown by hollow circles and dotted lines, while blobs and solid lines show the corresponding quantities for a hierarchical model. Note the shrinkage of the estimates for the hierarchical model towards the overall posterior mean rate, shown as the solid vertical line; the hierarchical intervals are slightly shorter than those for the simpler model.

common values by amounts that depend on the precision of the corresponding observations, here represented by the v_j. ■

Example 11.26 (Cardiac surgery data) Table 11.2 contains data on mortality of babies undergoing cardiac surgery at 12 hospitals. Although the numbers of operations and the death rates vary, we have no further knowledge of the hospitals and hence no basis for treating them other than entirely symmetrically, suggesting the hierarchical model

$$r_j \mid \theta_j \overset{\text{ind}}{\sim} B(m_j, \theta_j), \quad j = A, \ldots, L, \quad \theta_A, \ldots, \theta_L \mid \zeta \overset{\text{iid}}{\sim} f(\theta \mid \zeta), \quad \zeta \sim \pi(\zeta).$$

Conditional on θ_j, the number of deaths r_j at hospital j is binomial with probability θ_j and denominator m_j, the number of operations, which plays the same role as v_j^{-1} in Example 11.25: when m_j is large then a death rate is relatively precisely known. Conditional on ζ, the θ_j are a random sample from a distribution $f(\theta \mid \zeta)$, and ζ itself has a prior distribution that depends on fixed hyperparameters.

One simple formulation is to let $\beta_j = \log\{\theta_j/(1-\theta_j)\} \sim N(\mu, \sigma^2)$, conditional on $\zeta = (\mu, \sigma^2)$, thereby supposing that the log odds of death have a normal distribution, and to take $\mu \sim N(0, c^2)$ and $\sigma^2 \sim IG(a, b)$, where a, b, and c express proper but vague prior information. For sake of illustration we let $a = b = 10^{-3}$, so σ^2 has prior mean one but variance 10^3, and $c = 10^3$,

giving μ prior variance 10^6. The joint density then has form

$$\prod_j \binom{m_j}{r_j} \frac{e^{r_j\beta_j}}{(1+e^{\beta_j})^{m_j}} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\beta_j-\mu)^2\right\} \times \pi(\mu)\pi(\sigma^2),$$

so the full conditional densities for μ and σ^2 are normal and inverse gamma. Apart from a constant, the full conditional density for β_j has logarithm

$$r_j\beta_j - m_j \log\left(1+e^{\beta_j}\right) - \frac{(\beta_j-\mu)^2}{2\sigma^2},$$

and as this is a sum of two functions concave in β_j, adaptive rejection sampling may be used to simulate β_j given μ, σ^2, and the data; see Example 3.22.

This model was fitted using the Gibbs sampler with 5500 iterations, of which the first 500 were discarded. Convergence appeared rapid.

Figure 11.11 compares results for the hierarchical model with the effect of treating each hospital separately using uniform prior densities for the θ_j. Shrinkage due to the hierarchical fit is strong, particularly for the smaller hospitals; the posterior mean of θ_A, for example, has changed from about 2% to over 5%. Likewise the posterior means of θ_H and θ_B have decreased considerably towards the overall mean. By contrast, the posterior mean of θ_D barely changes because of the large value of m_D. Posterior credible intervals for the hierarchical model are only slightly shorter but they are centred quite differently. The posterior mean rate is about 7.3%, with 0.95 credible interval $(5.3, 9.4)$%. ■

In some cases the hierarchical element is merely a component of a more complex model, as the following example illustrates.

Example 11.27 (Spring barley data) Table 10.21 contains data on a field trial intended to compare the yields of 75 varieties of spring barley allocated randomly to plots in three long narrow blocks. The data were analysed in Example 10.35 using a generalized additive model to accommodate the strong fertility trends over the blocks. In the absence of detailed knowledge about the varieties it seems natural to treat them as exchangeable, and we outline a Bayesian hierarchical approach. We also show how the fertility patterns may be modelled using a simple Markov random field.

Let $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ denote the yields in the $n = 225$ plots and let ψ_j denote the unknown fertility of plot j. Let X denote the $n \times p$ design matrix that shows which of the $p = 75$ variety parameters $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ have been allocated to the plots. Then a normal linear model for the yields is

$$y \mid \beta, \psi, \lambda_y \;\sim\; N_n(\psi + X\beta, I_n/\lambda_y), \tag{11.51}$$

where ψ is the $n \times 1$ vector containing the fertilities and λ_y is the unknown precision of the ys.

We take the prior density of λ_y to be gamma with shape and scale parameters a and b, $G(a, b)$, so that its prior mean and variance are a/b and a/b^2, where a and b are specified. As there is no special treatment structure, we take for the β_r the exchangeable prior $\beta \sim N_p(0, I_p/\lambda_\beta^{-1})$, with $\lambda_\beta \sim G(c, d)$ and c, d specified. For the fertilities we take the normal Markov chain of Example 6.13, for which

$$\pi(\psi \mid \lambda_\psi) \propto \lambda_\psi^{n/2} \exp\left\{-\tfrac{1}{2}\lambda_\psi \sum_{i\sim j}(\psi_i - \psi_j)^2\right\}, \quad \lambda_\psi > 0, \tag{11.52}$$

the summation being over pairs of neighbouring plots and λ_ψ^{-1} being the variance of differences between fertilities. Each ψ_j occurs in n_j terms, where $n_j = 1$ or 2 is the number of plots adjacent to plot j. The sum in (11.52) equals $\psi^{\mathrm{T}} W \psi$, where W is the $n \times n$ tridiagonal matrix with elements

$$w_{ij} = \begin{cases} n_i, & i = j, \\ -1, & i \sim j, \\ 0, & \text{otherwise.} \end{cases}$$

Thus W is block diagonal, with three blocks like the matrix V in Example 6.13 with $\tau = 0$, corresponding to the three physical blocks of the experiment. We take $\lambda_\psi \sim G(g, h)$, with g and h specified.

With these conjugate prior densities, the joint posterior density is

$$\begin{aligned} \pi(\beta, \psi, \lambda) \quad \propto \quad & \lambda_y^{n/2} \exp\left\{-\tfrac{1}{2}\lambda_y (y - \psi - X\beta)^{\mathrm{T}}(y - \psi - X\beta)\right\} \\ & \times \lambda_\beta^{p/2} \exp\left(-\tfrac{1}{2}\lambda_\beta \beta^{\mathrm{T}}\beta\right) \times \lambda_\psi^{p/2} \exp\left(-\tfrac{1}{2}\lambda_\psi \psi^{\mathrm{T}} W \psi\right) \\ & \times \lambda_y^{a-1} \exp\left(-b\lambda_y\right) \times \lambda_\beta^{c-1} \exp\left(-c\lambda_\beta\right) \times \lambda_\psi^{g-1} \exp\left(-h\lambda_\psi\right), \end{aligned}$$

where $\lambda = (\lambda_y, \lambda_\beta, \lambda_\psi)^{\mathrm{T}}$. The full conditional densities turn out to be

$$\beta \mid \psi, \lambda, y \quad \sim \quad N\left\{\lambda_y Q_\beta^{-1} X^{\mathrm{T}}(y - \psi), Q_\beta^{-1}\right\}, \tag{11.53}$$

$$\psi \mid \beta, \lambda, y \quad \sim \quad N\left\{\lambda_y Q_\psi^{-1}(y - X\beta), Q_\psi^{-1}\right\}, \tag{11.54}$$

$$\lambda_y \mid \psi, \beta, y \quad \sim \quad G\left\{a + n/2, b + (y - X\beta - \psi)^{\mathrm{T}}(y - X\beta - \psi)/2\right\}, \tag{11.55}$$

$$\lambda_\beta \mid \psi, \beta, y \quad \sim \quad G\left(c + p/2, d + \beta^{\mathrm{T}}\beta/2\right), \tag{11.56}$$

$$\lambda_\psi \mid \psi, \beta, y \quad \sim \quad G\left(g + n/2, h + \psi^{\mathrm{T}} W \psi/2\right), \tag{11.57}$$

where

$$Q_\beta = \lambda_y X^{\mathrm{T}} X + \lambda_\beta I_p, \quad Q_\psi = \lambda_y I_n + \lambda_\psi W.$$

The elements of λ are independent conditional on the remaining variables. The relatively simple form of the densities in (11.53)–(11.57) suggests using a time-reversible Gibbs sampler, in which β, ψ, and λ are updated in a random order at each iteration. The most direct approach to simulation in (11.53)

and (11.54) is through Cholesky decomposition of Q_β and Q_ψ: in (11.53), for example, we find the lower triangular matrix L such that $LL^{\mathrm{T}} = Q_\beta^{-1}$, generate $\varepsilon \sim N_p(0, I_p)$, and let $\beta = \lambda_y Q_\beta^{-1} X^{\mathrm{T}}(y-\psi)+L\varepsilon$. The block diagonal structure of W means that the ψs for different blocks can be updated separately, so the largest Cholesky decomposition needed is that of a 75×75 matrix. An alternative is to update individual ψ_js in a random order, but although the computational burden is smaller, the algorithm then converges more slowly than with direct use of (11.54).

Note the strong resemblance of (11.53) and (11.54) to the steps of the backfitting algorithm for the corresponding generalized additive model.

The missing response in block 3 is simply a further unknown whose value may be simulated using the relevant marginal density of (11.51). This adds a fourth component to the simulation in random order of β, ψ, and λ at each iteration; there are no other changes to the algorithm.

If the matrix $X^{\mathrm{T}}X$ is diagonal, then the full conditional density for the rth variety effect has form

$$\beta_r \mid \psi, \lambda, y \sim N\left(\frac{\lambda_y m_r \overline{z}_r}{\lambda_\beta + \lambda_y m_r}, \frac{1}{\lambda_\beta + \lambda_y m_r}\right),$$

where $\overline{z}_r$ is the current average of $y_j - \psi_j$ for the m_r plots receiving variety r. Thus the β_r are shrunk towards zero by an amount that depends on the ratio λ_β/λ_y; with $\lambda_\beta = 0$ the mean for β in (11.53) is the least squares estimate computed by regressing $y-\psi$ on the columns of X. Unlike in Example 11.25, however, the normal distributions of the β_r are here averaged over the posterior densities of ψ, λ_y and λ_β.

The algorithm described above was run with random initial values for 10,500 iterations. Time series plots of the parameters and log likelihood suggested that it had converged after 500 iterations, and inferences below are based on the final 10,000 iterations. The variance inflation factors $\widehat{\tau}$ were less than 4 for ψ and β, about 44, 6 and 30 for λ_y, λ_τ and λ_ψ, and about 6 for y_{187}. Thus estimation for λ_y is least reliable, being based on a sample equivalent to about 220 independent observations. A longer run of the algorithm would seem wise in practice. Based on this run, the posterior 0.9 credible intervals for λ_y, λ_ψ and λ_β were $(5.2, 12.4)$, $(5.0, 11.5)$ and $(2.7, 5.7)$ respectively, and differences of two variety effects have posterior densities very close to normal with typical standard deviation of 0.35. The corresponding standard error for the generalized additive model was 0.41, so use of a hierarchical model and injection of prior information has increased the precision of these comparisons.

Figure 11.12 shows some simulated values of ψ and pointwise 0.90 credible envelopes for the true ψ. These envelopes are constructed by joining the 0.05 quantiles of the fertilities simulated from the posterior density, for each location, and likewise with the 0.95 quantiles. By contrast with the analysis in

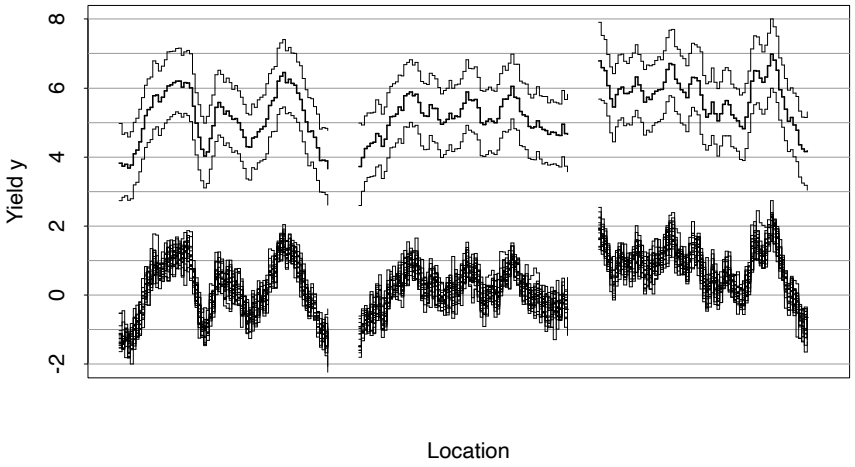


Figure 11.12
Posterior summaries for fertility trend ψ for the three blocks of spring barley data, shown from left to right. Above: median trend (heavy) and overall 0.9 posterior credible bands. Below: 20 simulated trends from Gibbs sampler output.

Table 11.12
Posterior probabilities that a variety is ranked among the best r varieties, estimated from 10,000 iterations of Gibbs sampler.

r	Variety									
	56	35	72	31	55	47	54	18	38	40
1	0.327	0.182	0.149	0.129	0.075	0.055	0.019	0.015	0.012	0.006
2	0.518	0.357	0.299	0.270	0.174	0.136	0.050	0.042	0.035	0.020
5	0.814	0.690	0.643	0.621	0.486	0.416	0.234	0.183	0.153	0.106
10	0.959	0.908	0.887	0.871	0.795	0.743	0.560	0.497	0.429	0.344

Example 10.35, the effective degrees of freedom for ψ, controlled by λ_ψ, are here equal for each block, leading to apparent overfitting of the fertilities for block 2 compared to the generalized additive model. A difference between the models is that the current model corresponds to first differences of ψ being a normal random sample, while in the earlier model the second differences are a normal random sample, giving a smoother fit.

The posterior probabilities that certain varieties rank among the r best are given in Table 11.12. The ordering is somewhat different from that in Example 10.35, perhaps due to the slightly different treatment of fertility effects. As mentioned previously, no single variety strongly outperforms the rest, and future field experiments would have to include several of those included in this trial. This type of information is difficult to obtain using frequentist procedures, but is readily found by manipulating the output of the simulation algorithm described above.

This analysis is relatively easily modified when elements of the model are

changed. Indeed the priors and other components chosen largely for convenience should be varied in order to assess the sensitivity of the conclusions to them; see Exercise 11.3.6. Metropolis–Hastings steps would then typically replace the Gibbs updates in the algorithm. ■

As mentioned above, more complicated hierarchies involve several layers of nested variation. Such models are widely used in certain applications, but their assessment and comparison can be difficult. For instance, shrinkage makes it unclear just how many parameters a hierarchical model has. Hierarchical modelling is an active area of current research.

Justification of (11.49)

To establish (11.49), suppose that r lies in $0, \ldots, n$ and that $m > n$. Then exchangeability of $U_1, \ldots, U_m$ implies that the conditional probability

$$\Pr(U_1 + \cdots + U_n = r \mid U_1 + \cdots + U_m = s)$$

equals the probability of seeing r 1's in n draws without replacement from an urn containing s 1's and $m-s$ 0's, which is $\binom{m}{n}^{-1}\binom{s}{r}\binom{m-s}{n-r}$ for $s = r, \ldots, m-(n-r)$ and zero otherwise. Hence

$$\begin{aligned}\Pr(U_1 + \cdots + U_n = r) &= \sum_{s=r}^{m-(n-r)} \binom{m}{n}^{-1}\binom{s}{r}\binom{m-s}{n-r}\Pr(U_1 + \cdots + U_m = s) \\ &= \binom{n}{r}\sum_{s=r}^{m-(n-r)} \frac{s^{(r)}(m-s)^{(n-r)}}{m^{(n)}}\Pr(U_1 + \cdots + U_m = s),\end{aligned}$$

where $s^{(r)} = s(s-1)\cdots(s-r+1)$ and so forth. If $G_m(\theta)$ denotes the distribution putting mass $\Pr(U_1 + \cdots + U_m = s)$ at s/m, for $s = 0, \ldots, m$, then

$$\Pr(U_1 + \cdots + U_n = r) = \binom{n}{r}\int_0^1 \frac{(m\theta)^{(r)}\{m(1-\theta)\}^{(n-r)}}{m^{(n)}}\, dG_m(\theta).$$

As $m \to \infty$,

$$\frac{(m\theta)^{(r)}\{m(1-\theta)\}^{(n-r)}}{m^{(n)}} \to \theta^r(1-\theta)^{n-r},$$

and in fact there is an infinite subsequence of values of m such that G_m converges to a limit G that is a distribution function. To establish (11.49) we simply note that

$$\binom{n}{r} f(u_{\xi(1)}, \ldots, u_{\xi(n)}) = \Pr(U_1 + \cdots + U_n = r)$$

for any permutation ξ of $\{1, \ldots, n\}$ such that $u_{\xi(1)} + \cdots + u_{\xi(n)} = r$, giving

$$f(u_1, \ldots, u_n) = \int_0^1 \theta^r(1-\theta)^{n-r}\, dG(\theta) = \int_0^1 \prod_{j=1}^n \theta^{u_j}(1-\theta)^{1-u_j}\, dG(\theta)$$

as desired.

Exercises 11.4

1 Two balls are drawn successively without replacement from an urn containing three white and two red balls. Are the outcomes of the first and second draws independent? Are they exchangeable?

2 Under what conditions are the Bernoulli random variables Y_1 and $Y_2 = 1 - Y_1$ exchangeable? What about $Y_1, \ldots, Y_n$ given that $Y_1 + \cdots + Y_n = m$?

3 Establish (11.50), and use it and (3.21) to verify the given formulae for the posterior mean and variance for μ.

4 Describe how a Metropolis–Hastings update could be used to avoid adaptive rejection sampling from the full conditional density for β in Example 11.26. Compare and contrast the two approaches.

5 In a variant on the hierarchical Poisson model in Example 11.19, let $Y_1, \ldots, Y_n$ be independent Poisson variables with means $\theta_1, \ldots, \theta_n$, let $\theta_1, \ldots, \theta_n$ be a random sample from the density $\beta e^{-\theta\beta}$, $\theta > 0$, and let the prior density of β be uniform on the positive half-line. Find $\mathrm{E}(\theta_j \mid y, \beta)$, and show that if $n\overline{y} > 1$ then the posterior distribution of $\gamma = 1/(1+\beta)$ is Beta with parameters $n\overline{y} - 1$ and $n+1$. Hence show that the posterior mean of θ_j is $(y_j+1)(n\overline{y}-1)/(n\overline{y}+n)$. Under what condition is this greater than the estimate $\widehat{\theta}_j = y_j$ obtained under the classical model with no link among the θs? Explain.

6 (a) Give the directed acyclic and conditional independence graphs for the model in Example 11.27, and verify (11.53)–(11.57).
(b) What changes to the algorithm are needed if (11.52) is replaced by

$$\pi(\psi \mid \lambda_\psi) \propto \lambda_\psi^{n/2} \exp\left\{ -\tfrac{1}{2}\lambda_\psi \sum_{i \sim j} |\psi_i - \psi_j| \right\}, \quad \lambda_\psi > 0?$$

What changes are needed if (11.51) specifies that the y_j have independent t_ν densities, for some known ν?
(c) How would you allow different degrees of smoothing for the different blocks? (Besag *et al.*, 1995)

11.5 Empirical Bayes Inference

11.5.1 Basic ideas

The borrowing of strength achieved by hierarchical Bayes models increases the precision of parameter estimation at the cost of specifying prior distributions at two levels. This can be bothersome in practice, because priors on hyperparameters are difficult to verify and it is natural to worry about their effect on subsequent inferences. Sensitivity analysis, comparing results from different priors, is valuable, but another possibility in some cases is to estimate the hyperparameters from the data. Many Bayesians deprecate this empirical

Bayes approach as essentially frequentist; we shall skirt this issue and simply sketch the main ideas.

Consider the model

$$y_1, \ldots, y_n \mid \theta_1, \ldots, \theta_n \overset{\text{ind}}{\sim} f(y_1 \mid \theta_1), \ldots, f(y_n \mid \theta_n), \quad \theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} \pi(\theta \mid \gamma).$$

A fully Bayesian specification would add a prior density $\pi(\gamma)$ for γ, with inference for the θ_j based on the marginal posterior densities $\pi(\theta_j \mid y)$. If we do not add this further level of complexity, then the data have marginal density

$$f(y_1, \ldots, y_n \mid \gamma) = \prod_{j=1}^{n} \int f(y_j \mid \theta_j) \pi(\theta_j \mid \gamma) \, d\theta_j$$

from which we might estimate γ. An obvious approach is to use the maximum likelihood estimator $\widehat{\gamma}$ found from this density, and then to base inferences on the posterior densities $\pi(\theta_j \mid y, \widehat{\gamma})$, for example computing posterior moments

$$\mathrm{E}(\theta_j^r \mid y, \widehat{\gamma}) = \left. \frac{\int \theta_j^r f(y_j \mid \theta_j) \pi(\theta_j \mid \gamma) \, d\theta_j}{\int f(y_j \mid \theta_j) \pi(\theta_j \mid \gamma) \, d\theta_j} \right|_{\gamma = \widehat{\gamma}}.$$

Numerical methods are generally needed to evaluate the integrals. Full Bayesian analysis would integrate out γ with respect to its prior density, thereby accounting for uncertainty about γ rather than simply setting it to $\widehat{\gamma}$.

Example 11.28 (Normal distribution) Consider the model

$$y_1, \ldots, y_n \mid \theta_1, \ldots, \theta_n \overset{\text{ind}}{\sim} N(\theta_j, v_j), \quad \theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} N(\mu, \tau^2),$$

where the v_j are known positive constants, and suppose initially that $\tau^2 > 0$ is also known. The conditional distribution of θ_j given y is

$$N\left(\xi_j \mu + (1 - \xi_j) y_j, (1 - \xi_j) v_j\right), \text{ with } \xi_j = \frac{v_j}{v_j + \tau^2}, \quad j = 1, \ldots, n, \quad (11.58)$$

and the y_j are marginally independent with $N(\mu, v_j + \tau^2)$ densities. The maximum likelihood estimate of μ is therefore

$$\widehat{\mu} = \widehat{\mu}(\tau^2) = \frac{\sum_{j=1}^{n} y_j / (v_j + \tau^2)}{\sum_{j=1}^{n} 1 / (v_j + \tau^2)},$$

and the empirical Bayes estimate of θ_j is found by substituting this into $\mathrm{E}(\theta_j \mid y)$, to give

$$\tilde{\theta}_j = \xi_j \widehat{\mu} + (1 - \xi_j) y_j = \widehat{\mu} + (1 - \xi_j)(y_j - \widehat{\mu}). \quad (11.59)$$

When $\xi_j = 0$ then $\tilde{\theta}_j = y_j$ is unbiased for θ_j. Taking $\xi_j > 0$ gives non-zero shrinkage and biased estimation of $\tilde{\theta}_j$, but the hope is that the borrowing of strength induced by shrinkage towards a common mean will reduce overall

mean squared error. The degree of shrinkage towards $\widehat{\mu}$ depends on v_j/τ^2. This is disquieting because the amount of shrinkage bears no relation to the data. Thus if the y_j were very different doubt would be cast on the model, but the formulation pays no heed to this.

When τ^2 is unknown, its profile log likelihood is

$$\ell_{\mathrm{p}}(\tau^2) \equiv -\tfrac{1}{2}\sum_{j=1}^{n}\log(v_j+\tau^2) - \tfrac{1}{2}\sum_{j=1}^{n}\left\{y_j - \widehat{\mu}(\tau^2)\right\}^2/(v_j+\tau^2), \quad \tau^2 \geq 0,$$

from which the maximum likelihood estimate $\widehat{\tau}^2$ can be obtained. If $\widehat{\tau}^2 = 0$ then the data give no evidence of variation in the θ_j, all the y_j have mean μ, and all the $\tilde{\theta}_j$ are shrunk to $\widehat{\mu}$. If $\widehat{\tau}^2 > 0$, then ξ_j is replaced by $v_j/(v_j+\widehat{\tau}^2)$ in (11.59). As $0 \leq v_j/(v_j+\widehat{\tau}^2) \leq 1$, $\tilde{\theta}_j$ lies between y_j and $\widehat{\mu}$.

Confidence intervals for the θ_j may be computed by replacing μ and τ^2 in (11.58) by estimates, but their coverage will be lower than the nominal level because the variability of $\widehat{\mu}$ and $\widehat{\tau}^2$ is unaccounted for. Approaches to overcoming this have been proposed, but we shall not treat them here. ■

Example 11.29 (Toxoplasmosis data) Example 10.29 discusses estimation of levels of toxoplasmosis in 34 cities in El Salvador. For a simple analysis of these data, we let $y_j = \log\{(r_j+1/2)/(m_j-r_j+1/2)\}$ represent empirical logistic transformations of the binomial responses giving the level of toxoplasmosis, with approximate variances $v_j = (r_j+1/2)^{-1} + (m_j-r_j+1/2)^{-1}$ treated as known. We generalize Example 11.28 to encompass regression by taking

$$y_1,\ldots,y_n \mid \theta_1,\ldots,\theta_n \overset{\text{ind}}{\sim} N(\theta_j, v_j), \quad \theta_j \mid \beta \overset{\text{ind}}{\sim} N(x_j^{\mathrm{T}}\beta, v_j'), \quad j = 1,\ldots,n,$$

so that the θ_j vary around means $x_j^{\mathrm{T}}\beta$. Then

$$\theta_j \mid y, \beta, v_j' \overset{\text{ind}}{\sim} N\left\{(1-\xi_j)y_j + \xi_j x_j^{\mathrm{T}}\beta, v_j(1-\xi_j)\right\}, \quad \xi_j = v_j/(v_j+v_j'),$$

and marginally $y_j \overset{\text{ind}}{\sim} N(x_j^{\mathrm{T}}\beta, v_j+v_j')$, for $j = 1,\ldots,n$. Maximum likelihood yields the weighted least squares estimator $\widehat{\beta} = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wy$, where W is the diagonal matrix with elements $w_j = (v_j+v_j')^{-1}$, leading to shrinkage estimators $\tilde{\theta}_j = (1-\xi_j)y_j + \xi_j x_j^{\mathrm{T}}\widehat{\beta}$ of the θ_j, with estimated variances $v_j(1-\xi_j)$.

The v_j' typically depend on unknown parameters that may be estimated from the profile likelihood. Here we take $v_1' = \cdots = v_n' = \tau^2$. If $x^{\mathrm{T}}\beta$ equals a constant, then $\widehat{\tau}^2 = 0.17$, but it is better to let $x^{\mathrm{T}}\beta$ be a cubic function of rainfall, leading to $\widehat{\tau}^2 = 0.1$. Figure 11.13 shows strong shrinkage of the individual estimates y_j towards their regression counterparts $x_j\widehat{\beta}$. The average variance reduces by a factor of almost ten, from $\overline{v} = 0.68$ to $\overline{v(1-\xi)} = 0.07$, and one would expect a large decrease in overall mean squared error.

The empirical Bayes estimates of the toxoplasmosis levels themselves are

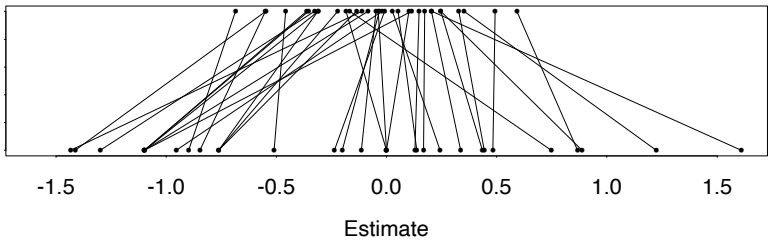


Figure 11.13 Shrinkage of individual estimates (lower blobs) towards regession estimates (upper blobs) for toxoplasmosis data.

r	1	2	3	4	5	6	7	8	9	10	Total
0+	14376	4343	2292	1463	1043	837	638	519	430	364	26305
10+	305	259	242	223	187	181	179	130	127	128	1961
20+	104	105	99	112	93	74	83	76	72	63	881
30+	73	47	56	59	53	45	34	49	45	52	513
40+	49	41	30	35	37	21	41	30	28	19	331
50+	25	19	28	27	31	19	19	22	23	14	227
60+	30	19	21	18	15	10	15	14	11	16	169
70+	13	12	10	16	18	11	8	15	12	7	122
80+	13	12	11	8	10	11	7	12	9	8	101
90+	4	7	6	7	10	10	15	7	7	5	78

Table 11.13 Shakespeare's word type frequencies (Efron and Thisted, 1976; Thisted and Efron, 1987). Entry r is n_r, the number of word types used exactly r times. There are 846 word types which appear more than 100 times, for a total of 31,534 word types.

obtained by inverse logistic transformation, with standard errors from the delta method. A more detailed analysis, or simulation, would be needed to account for the uncertainty in $\widehat{\beta}$ and $\widehat{\tau}^2$. ■

The previous examples illustrate parametric empirical Bayes inference, in which the prior for θ is taken from a parametrized family of distributions. In practice an alternative is to try and estimate the prior nonparametrically. The resulting estimators are generally unstable if the data are not extensive, and some form of smoothing may be needed.

Example 11.30 (Shakespeare's vocabulary data) The canon of Shakespeare's accepted works contains 884,647 words, with 31,534 distinct word types. A word type is a distinguishable arrangement of letters, so 'king' is different from 'kings' and 'alehouse' different from both 'ale' and 'house'. Table 11.13 shows how many word types occurred once, twice, and so on in the canon: 14,376 appear just once, 4343 appear twice, and so forth. If n_r is the number of word types appearing r times, then $\sum_{r=1}^{\infty} n_r = 31{,}534$.

If a new body of work containing $884{,}647t$ words was found, how many new

word types might it contain? Taking $t = 1$ corresponds to finding a new set of works the same size as the canon, while setting $t = \infty$ enables us to estimate Shakespeare's total vocabulary.

Finding a new word type in a body of work is analogous to finding a new species of animal among those caught in a trap. Suppose that there are S species in total, and that after trapping over the period $[-1, 0]$ we have y_s members of species s. We assume that they enter the trap according to a Poisson process of rate λ_s per unit of time, so y_s is Poisson with mean λ_s, and let $n_r = \sum_s I(y_s = r)$ be the number of species observed exactly r times in the trapping period $[-1, 0]$. Let $G(\lambda)$ be the unknown distribution function of $\lambda_1, \ldots, \lambda_S$. Then the expected number of species seen in $(0, t]$ that were seen exactly r times in the previous interval $[-1, 0]$ is

$$\begin{aligned}
\nu_r(t) &= S \int_0^\infty e^{-\lambda} \frac{\lambda^r}{r!} (1 - e^{-\lambda t}) dG(\lambda) \\
&= S \int_0^\infty e^{-\lambda} \frac{\lambda^r}{r!} \left\{ \lambda t - \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^3}{3!} - \cdots \right\} dG(\lambda) \\
&= \sum_{k=1}^{\infty} (-1)^{k+1} \binom{r+k}{k} t^k \eta_{r+k}, \qquad (11.60)
\end{aligned}$$

where

$$\eta_r = \mathrm{E}(n_r) = S \int_0^\infty \frac{\lambda^r}{r!} e^{-\lambda} \, dG(\lambda), \quad r = 1, 2, \ldots .$$

The convergence of (11.60) will depend on t, but if it does converge, then an unbiased nonparametric empirical Bayes estimator $\tilde{\nu}_r(t)$ is obtained by replacing the η_r by estimates $\tilde{\eta}_r = n_r$ obtained from the marginal distribution across the species. If the S Poisson processes are independent, then the n_r will be approximately independent Poisson variables with means η_r. Thus for example,

$$\mathrm{var}\,\{\tilde{\nu}_0(t)\} = \mathrm{var}\left(n_1 t - n_2 t^2 + n_3 t^3 - \cdots\right) \doteq \sum_{r=1}^{\infty} \eta_r t^{2r} \doteq \sum_{r=1}^{\infty} n_r t^{2r}$$

provides a standard error for $\tilde{\nu}_0(t)$.

For the data in Table 11.13, $\tilde{\nu}_0(1) = 11{,}430$ with standard error 178. It turns out not to be possible to give an upper bound for the size of Shakepeare's vocabulary, but a fairly realistic lower bound can be established of about 35,000 word types that he knew but which do not appear in the canon.

Parametric empirical Bayes models employ parametric distributions for G, one candidate being gamma with mean and variance ξ/β and ξ/β^2. Then

$$\eta_r = \eta_1 \frac{\Gamma(r + \xi)}{r! \Gamma(1 + \xi)} \left(\frac{\beta}{1 + \beta} \right)^{r-1}, \quad r = 1, 2, \ldots,$$

proportional to the negative binomial density truncated so that $r > 0$. In the negative binomial case $\xi > 0$, but here any value of $\xi > -1$ is possible; $\xi = 0$ gives the logarithmic series distribution, the first to be fitted to species abundance data. The parameters can be estimated by maximum likelihood fitting of the multinomial distribution of $n_1, \ldots, n_{r_0}$, for some suitable r_0. Taking $r_0 = 40$ yields $\widehat{\eta}_1 = 14{,}376$, $\widehat{\xi} = -0.3954$ and $\widehat{\beta} = 104.3$. The fit to Table 11.13 is then remarkably good, giving $\tilde{\nu}_0(1) \doteq 11{,}483$, very close to the nonparametric empirical Bayes estimate.

In 1985 a previously unknown nine-stanza poem was found in the Bodleian Library in Oxford. It consists of 429 words with 258 word types, of which nine do not appear in the canon. The empirical counts can be compared with the values $\tilde{\nu}_r(t)$ with $t = 429/884{,}647$; for example $\tilde{\nu}_0(t) = 6.97$ is in fair agreement with the observed number of nine new words. Detailed work suggests that at least on the basis of the word counts, the poem might be attributable to Shakespeare. Scholarly debate continues, however, as word usage in the new poem differs from that in the canon. ■

Shrinkage improves estimators in many models. Before discussing an unexpected consequence of this, we outline some key notions of decision theory.

11.5.2 Decision theory

Sometimes data are gathered in order to decide among decisions whose payoffs are known explicitly. The decision chosen will depend on the data y, and the choice is made according to a *decision rule* $\delta(y)$, which takes a value in a decision space $\mathcal{D}$. Thus δ is a mapping from the sample space $\mathcal{Y}$ to $\mathcal{D}$.

The fact that some decisions have better consequences than others is quantified through a *loss function* $l(d, \theta)$, which represents the loss due to making decision d when the true state of nature is θ. A bad decision incurs a big loss, a better decision a smaller one.

At the time a decision is taken its loss is unknown because of uncertainty about θ. Nevertheless, provided we have prior information on θ, we can calculate the posterior expected loss,

$$\mathrm{E}\left\{l(d,\theta) \mid y\right\} = \int l(d,\theta)\pi(\theta \mid y)\, d\theta = \frac{\int l(d,\theta) f(y \mid \theta)\pi(\theta)\, d\theta}{\int f(y \mid \theta)\pi(\theta)\, d\theta}.$$

This is a function of d and y. If we want to make a decision leading to as small a loss as possible, one strategy is to choose the decision d that minimizes the posterior expected loss for the particular y that has been observed. Thus $\delta(y) = d$, where $\mathrm{E}\left\{l(d', \theta) \mid y\right\} \geq \mathrm{E}\left\{l(d, \theta) \mid y\right\}$ for every $d' \in \mathcal{D}$. This is called the *Bayes rule* for loss function l with respect to prior π.

Example 11.31 (Discrimination) Suppose we must decide whether or

not a patient with measurements y has a disease that has prevalence γ in the population. Let $\theta = 1$ indicate the event that he is diseased. Then

$$\Pr(\theta = 1) = \gamma, \quad \Pr(\theta = 0) = 1 - \gamma,$$

and y has densities $f_1(y)$ and $f_0(y)$ according to the unknown value of θ, which represents the state of nature. The possible decisions are

$$d_0 = \text{'patient is not diseased'}, \quad d_1 = \text{'patient is diseased'},$$

and a decision rule $\delta(y)$ is a procedure that chooses one of these.

Let l_{ij} denote the loss made when $\theta = i$ and decision d_j is made. We set $l_{00} = l_{11} = 0$, so there is no loss when a decision is correct, and assume that $l_{10}, l_{01} > 0$. The posterior expected losses associated with d_0 and d_1 are

$$\mathrm{E}\{l(d_0, \theta) \mid y\} = \frac{l_{00}(1-\gamma)f_0(y) + l_{10}\gamma f_1(y)}{(1-\gamma)f_0(y) + \gamma f_1(y)} = \frac{l_{10}\gamma f_1(y)}{(1-\gamma)f_0(y) + \gamma f_1(y)}$$

and

$$\mathrm{E}\{l(d_1, \theta) \mid y\} = \frac{l_{01}(1-\gamma)f_0(y) + l_{11}\gamma f_1(y)}{(1-\gamma)f_0(y) + \gamma f_1(y)} = \frac{l_{01}(1-\gamma)f_0(y)}{(1-\gamma)f_0(y) + \gamma f_1(y)}.$$

The posterior expected loss is minimized by d_0 if $l_{10}\gamma f_1(y) < l_{01}(1-\gamma)f_0(y)$ and otherwise by d_1; we are indifferent if $l_{10}\gamma f_1(y) = l_{01}(1-\gamma)f_0(y)$.

This Bayes rule can be expressed in more familiar terms: choose d_0 if

$$\frac{f_0(y)}{f_1(y)} > \frac{l_{10}\gamma}{l_{01}(1-\gamma)},$$

and otherwise choose d_1. This is reminiscent of the Neyman–Pearson lemma, though here the value determining the decision involves γ and the loss function rather than a null distribution for y. ■

The set-up described thus far applies to decisions to be made once the data are known. But actions must sometimes be taken before any data are available — for example, an experimental design should be chosen to maximize the information in future data. It then seems wise to average the loss incurred over the future data. The expected loss due to using decision rule $\delta(y)$ when the true state of nature is θ is called the *risk function* of δ,

$$R_\delta(\theta) = \int l\{\delta(y), \theta\} f(y \mid \theta)\, dy.$$

If we have prior density $\pi(\theta)$ for θ, the overall expected loss due to using δ is the *Bayes risk*,

$$\begin{aligned} \int R_\delta(\theta)\pi(\theta)\, d\theta &= \int \pi(\theta) \int l\{\delta(y), \theta\} f(y \mid \theta)\, dy\, d\theta \\ &= \int f(y) \int l\{\delta(y), \theta\} \pi(\theta \mid y)\, d\theta\, dy. \end{aligned}$$

For any given y this is minimized by the decision $\delta(y)$ minimizing the inner integral, and this choice of δ is the Bayes rule for the prior $\pi(\theta)$. Thus the Bayes rule minimizes expected loss for both post-data and pre-data decisions.

If we view estimation as a decision problem, then a decision is a choice of the value $\tilde{\theta}$ to be used to estimate θ, and the loss depends on θ and $\tilde{\theta}$. A common choice is squared error loss, $l(\tilde{\theta}, \theta) = (\tilde{\theta} - \theta)^2$. The Bayes rule then uses as estimator the posterior mean of θ,

$$m(y) = \int \theta \pi(\theta \mid y)\, d\theta.$$

To see why, let $\tilde{\theta}(y)$ be any other estimator, and note that as

$$\left\{\tilde{\theta}(y) - \theta\right\}^2 = \left\{\tilde{\theta}(y) - m(y)\right\}^2 + 2\left\{\tilde{\theta}(y) - m(y)\right\}\{m(y) - \theta\} + \{m(y) - \theta\}^2,$$

the posterior expected loss

$$\int \left\{\tilde{\theta}(y) - \theta\right\}^2 \pi(\theta \mid y)\, d\theta = \left\{\tilde{\theta}(y) - m(y)\right\}^2 + \int \{m(y) - \theta\}^2 \pi(\theta \mid y)\, d\theta \tag{11.61}$$

is minimized by choosing $\tilde{\theta}(y) = m(y)$.

Admissible decision rules

We saw above that if a prior density for θ is available, one should choose the decision that minimizes the posterior expected loss with respect to that prior. But if no prior is available then we must attempt to make a good decision whatever the value of θ. We can compare two decision rules δ and δ' through their risk functions. If $R_{\delta'}(\theta) \geq R_{\delta}(\theta)$ for all θ, with strict inequality for some θ, then we say that δ' is *inadmissible* — it is beaten by another rule. If no such rule can be found, δ' is said to be *admissible*. Provided the decision formulation is accepted and considerations such as robustness may be ignored, we should clearly restrict attention to admissible decision rules.

The Bayes rule δ_B corresponding to a proper prior $\pi(\theta)$ is always admissible. For if not, there is a rule δ' such that $R_{\delta'}(\theta) \leq R_{\delta_B}(\theta)$, with strict inequality for some set of values of θ to which π attaches positive probability. The corresponding Bayes risks satisfy

$$\int \pi(\theta) R_{\delta'}(\theta)\, d\theta < \int \pi(\theta) R_{\delta_B}(\theta)\, d\theta,$$

contradicting the fact that δ_B minimizes the Bayes risk with respect to $\pi(\theta)$.

In a particular setting there may be many admissible decision rules. We can choose among them by minimizing $\sup_\theta R_\delta(\theta)$. This generally very conservative choice is called a *minimax rule*. An admissible decision rule δ with constant risk is minimax. For otherwise there exists a rule δ' such that for all

θ,

$$R_{\delta'}(\theta) \leq \sup_{\theta} R_{\delta'}(\theta) < \sup_{\theta} R_{\delta}(\theta).$$

But if δ has constant risk, then the right-hand side of this expression is constant, and δ must be inadmissible, which is a contradiction.

Example 11.32 (Normal distribution) Suppose that $Y_1, \ldots, Y_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution with known σ^2 and that we wish to choose an estimator $\tilde{\mu}$ of μ among

1 $\delta_1(Y) = \overline{Y}$, the sample average;
2 $\delta_2(Y)$ is the median of $Y_1, \ldots, Y_n$; and
3 $\delta_3(Y) = (n\overline{Y}/\sigma^2 + \mu_0/\tau^2)/(n/\sigma^2 + 1/\tau^2)$, the posterior mean for μ under the prior $N(\mu_0, \tau^2)$; see (11.11).

We take loss function $(\tilde{\mu} - \mu)^2$, so $\delta(Y)$ has risk $R_\delta(\mu)$ equal to its mean squared error, $\mathrm{E}\left[\{\delta(Y) - \mu\}^2\right]$, the expectation being over Y for fixed μ.

The average $\delta_1(Y)$ has mean and variance μ and σ^2/n, while the median $\delta_2(Y)$ has approximate mean and variance μ and $\pi\sigma^2/(2n)$. Their risks are

$$R_{\delta_1}(\mu) = \sigma^2/n, \quad R_{\delta_2}(\mu) \doteq \pi\sigma^2/(2n).$$

The posterior mean $\delta_3(Y)$ has bias and variance

$$\frac{n\mu/\sigma^2 + \mu_0/\tau^2}{n/\sigma^2 + 1/\tau^2} - \mu, \quad \frac{n/\sigma^2}{(n/\sigma^2 + 1/\tau^2)^2},$$

and so

$$R_{\delta_3}(\mu) = \frac{n/\sigma^2 + (\mu - \mu_0)^2/\tau^2}{(n/\sigma^2 + 1/\tau^2)^2}.$$

As $R_{\delta_2}(\mu) > R_{\delta_1}(\mu)$ for all μ, δ_2 is inadmissible. It can be shown that δ_1 is admissible, and as it has constant risk it is minimax. The rule δ_3 is Bayes and hence admissible. If τ^2 is small, δ_3 will be greatly preferable to δ_1 for values of μ close to the prior mean μ_0. Contrariwise if τ^2 is large, corresponding to weak prior information, then $R_{\delta_3}(\mu) < R_{\delta_1}(\mu)$ over a wide range, but the improvement is small. When $\tau \to \infty$, we see that $\delta_3 \to \delta_1$. ■

Shrinkage and squared error loss

Having set up machinery for the comparison of estimators using risk, we investigate the gains due to shrinkage when using empirical Bayes estimation.

Let $Y_1, \ldots, Y_n$ be independent normal variables with means $\theta_1, \ldots, \theta_n$ and unit variance. We consider estimation of $\theta_1, \ldots, \theta_n$ by $\tilde{\theta}_1, \ldots, \tilde{\theta}_n$ using as risk function the sum of squared errors

$$R_{\tilde{\theta}}(\theta) = \mathrm{E}\left\{\sum_{j=1}^{n} (\tilde{\theta}_j - \theta_j)^2\right\}, \tag{11.62}$$

the expectation being over Y with θ fixed. At first sight this formulation seems highly artificial, but in fact it is paradigmatic of many situations, one being the semiparametric models discussed in Section 10.7. The maximum likelihood estimators arise when $\tilde{\theta}_j = Y_j$ and have risk $R_{\tilde{\theta}}(\theta) = n$. Are better estimators available?

One possibility stems from taking (11.59) when $v_1 = \cdots = v_n$. Then $\widehat{\mu} = \overline{Y}$ does not depend on τ^2, whose maximum likelihood estimator is given by

$$\widehat{\tau}_+^2 = \max\left(n^{-1}W - 1, 0\right), \quad W = \sum_{j=1}^{n}(Y_j - \overline{Y})^2.$$

The eventual conclusion is unchanged but the computations below simplify if we replace $\widehat{\tau}_+^2$ by W/b, where we choose b to minimize the risk. Substitution into (11.59) gives the shrinkage estimators

$$\tilde{\theta}_j = \overline{Y} + (1 - b/W)(Y_j - \overline{Y}), \quad j = 1, \ldots, n. \tag{11.63}$$

These are more appealing than (11.59), because the degree of shrinkage depends on the data, being small if the Y_j are widely separated and W is large. 'Overshrinkage' occurs if $b/W > 1$, so in practice one would use a non-negative estimator such as $\widehat{\tau}_+$.

We show below that the risk of (11.63) using squared error loss is

$$R_{\tilde{\theta}}(\theta) = n + b\left\{b - 2(n-3)\right\}\mathrm{E}(W^{-1}). \tag{11.64}$$

This has minimum value $n-(n-3)^2\mathrm{E}(W^{-1})$ when $b = n-3$, and as $\mathrm{E}(W^{-1}) > 0$ this risk is uniformly less than n when $n > 3$. That is, when means of four or more normal variables are estimated simultaneously using (11.63) and squared error loss, the maximum likelihood estimator is inadmissible: the paragon of point estimation should not be used. This risk improvement is often called the *Stein effect* after its chief discoverer.

Charles Stein (1920–) studied at Chicago and Columbia universities and since 1953 has worked at Stanford University. He has made important contributions to mathematical statistics. See DeGroot (1986b).

This striking result rests on the cumulation of risk across observations; the chosen risk function would not be sensible if interest focused on a single θ_j. The extent to which shrinkage reduces the risk depends on the distribution of W, which is non-central chi-squared with non-centrality parameter $\rho = \sum(\theta_j - \overline{\theta})^2$. If $\rho = 0$, that is, all the θ_j are equal, then $\mathrm{E}(W^{-1}) = (n-3)^{-1}$ and $R_{\tilde{\theta}}(\theta) = 3$ independent of n. In this case shrinkage yields a dramatically improved estimator. If ρ is large, then the means of the Y_j are widely separated and $\mathrm{E}(W^{-1})$ is small, so $R_{\tilde{\theta}}(\theta)$ is only slightly less than n: the gain from shrinkage is then small. When $\overline{Y}$ in (11.63) and in W is replaced by a fixed prior value μ, then essentially the same result applies, with the maximum likelihood estimator then inadmissible when $n > 2$. The amount of shrinkage then depends on the distance from θ to the prior mean μ, and is large if this distance is small.

Similar results apply more generally, for example to regression and to multivariate situations. The broad lesson is that frequentist estimation of related quantities may be improved by using shrinkage procedures.

Derivation of (11.64)
Note first that with $\tilde{\theta}_j$ given in (11.63), $\sum(\tilde{\theta}_j - \theta_j)^2$ equals

$$\sum_{j=1}^{n} \left\{ \overline{Y} + (1 - b/W)(Y_j - \overline{Y}) - \theta_j \right\}^2 = \sum_{j=1}^{n} \left\{ Y_j - \theta_j - b(Y_j - \overline{Y})/W \right\}^2$$

and this equals

$$\sum_{j=1}^{n} (Y_j - \theta_j)^2 - 2bW^{-1} \sum_{j=1}^{n} (Y_j - \theta_j)(Y_j - \overline{Y}) + b^2 W^{-1}. \tag{11.65}$$

The first term has expectation n and the last appears in (11.64), so we must deal with the middle term.

Consider $\mathrm{E}\{(Y_j - \theta_j)h_j(Y)\}$, where $h_j(y)$ is a sufficiently well-behaved function. Integration by parts, recalling that $Y_j \overset{\text{ind}}{\sim} N(\theta_j, 1)$, and that $d\phi(z)/dz = -z\phi(z)$, implies that $\mathrm{E}\{(Y_j - \theta_j)h_j(Y)\} = \mathrm{E}\{\partial h_j(Y)/\partial Y_j\}$. Setting

$$h_j(Y) = \frac{Y_j - \overline{Y}}{W} = \frac{Y_j - \overline{Y}}{\sum_i (Y_i - \overline{Y})^2}$$

yields

$$\frac{\partial h_j(Y)}{\partial Y_j} = \frac{1 - n^{-1}}{W} - 2\frac{(Y_j - \overline{Y})^2}{W^2},$$

and a little algebra establishes that the central term in (11.65) has expectation $-2b(n-3)\mathrm{E}(W^{-1})$. Expression (11.64) follows directly.

Exercises 11.5

1 In Example 11.29, suppose that $v_j' = \tau^2 v_j$. Show that an unbiased estimator of τ^2 is then $SS/(n-p) - 1$, where SS is the residual sum of squares and p is the dimension of β, and explain why a better estimator is $\max\{SS/(n-p) - 1, 0\}$. Find also the profile log likelihood when $v_j' = \tau^2$.

2 Consider estimating the success probability θ for a binomial variable R with denominator m, using a beta prior distribution with parameters $a, b > 0$.
(a) Show that the marginal probability $\Pr(R = r \mid \mu, \nu)$ has beta-binomial form

$$\frac{\Gamma(\nu)}{\Gamma(\nu\mu)\Gamma\{\nu(1-\mu)\}} \binom{m}{r} \frac{\Gamma(r + \nu\mu)\Gamma\{m - r + \nu(1-\mu)\}}{\Gamma(m + \nu)}, \quad r = 0, \ldots, m,$$

where $\mu = a/(a+b)$ and $\nu = a + b$, and deduce that

$$\mathrm{E}(R/m) = \mu, \quad \mathrm{var}(R/m) = \frac{\mu(1-\mu)}{m}\left(1 + \frac{m-1}{\nu+1}\right).$$

(b) Show that methods of moments estimators based on a random sample $R_1, \ldots, R_n$ all with denominator m are

$$\widehat{\mu} = \overline{R}, \quad \widehat{\nu} = \frac{\widehat{\mu}(1-\widehat{\mu}) - S^2}{S^2 - \widehat{\mu}(1-\widehat{\mu})/m},$$

where $\overline{R}$ and S^2 are the sample average and variance of the R_j.
(c) Find the mean and variance of the conditional distribution of θ given R, and show that the mean can be written as a shrinking of R/m towards μ. Hence give the empirical Bayes estimates of the θ_j.

3 Consider a logistic regression model for Example 11.29. Show that the marginal log likelihood for β, τ^2 may be written as

$$\sum_{j=1}^{n} \log \int \frac{e^{r_j\theta}}{(1+e^{\theta})^{m_j}} \phi\left(\frac{\theta - x_j^{\mathrm{T}}\beta}{\tau}\right) d\theta - \log \tau.$$

Use Laplace approximation to remove the integrals, and outline how you would then estimate β and τ^2. Give also a Laplace approximation for the posterior mean of θ_j given the data, β and τ.

4 Consider the exponential family density $f(y \mid \theta) = \theta^y e^{-\kappa(\theta)} f_0(y)$ for integer y, where $f_0(y)$ is known. If $\pi(\theta)$ is any prior on θ, show that

$$\mathrm{E}(\theta \mid y) = \frac{\int \theta^{y+1} e^{-\kappa(\theta)} \pi(\theta)\, d\theta}{\int \theta^y e^{-\kappa(\theta)} \pi(\theta)\, d\theta} = \frac{\Pr_\pi(Y = y+1) f_0(y)}{\Pr_\pi(Y = y) f_0(y+1)},$$

where $\Pr_\pi(Y = y)$ is the marginal probability that $Y = y$, averaged over π. Given a sample $y_1, \ldots, y_n$ from the corresponding empirical Bayes model, explain why $\mathrm{E}(\theta_j \mid y_j)$ may be estimated by

$$\frac{f_0(y_j) \sum_{i=1}^{n} I(y_i = y_j + 1)}{f_0(y_j + 1) \sum_{i=1}^{n} I(y_i = y_j)}.$$

Do you think this estimator will be numerically stable? Check by simulating some data and trying it out.

5 Let $X_1, \ldots, X_n$ be a Poisson random sample with mean μ. Previous experience suggests prior density

$$\pi(\mu) = \frac{1}{\Gamma(\nu)} \mu^{\nu-1} e^{-\mu}, \quad 0 < \mu < \infty, \nu > 0.$$

If the loss function for an estimator $\tilde{\mu}$ of μ is $(\tilde{\mu} - \mu)^2$, determine an estimator that minimizes the expected loss and compare its bias and variance with those of the maximum likelihood estimator.

6 The proportion θ of defective items from a production process varies because of fluctuations in the the raw material. Records show that the prior density for θ is proportional to $\theta(1-\theta)^4$. A hundred items are inspected from a large batch all made from a homogeneous batch of raw material, and six are found to be defective.
Find the posterior density function for the proportion θ of defectives in the batch. The cost of estimating θ by $\widehat{\theta}$ is $\theta^2(\widehat{\theta} - \theta)^2$. Find also the value of $\widehat{\theta}$ which minimizes the expected cost, and the value of the minimum expected cost.

7 The loss when the success probability θ in Bernoulli trials is estimated by $\tilde{\theta}$ is $(\tilde{\theta}-\theta)^2\theta^{-1}(1-\theta)^{-1}$. Show that if the prior distribution for θ is uniform and m trials result in r successes then the corresponding Bayes estimator for θ is r/m. Hence show that r/m is also a minimax estimator for θ.

8 A population consists of k classes $\theta_1, \ldots, \theta_k$ and it is required to classify an individual on the basis of an observation Y having density $f_i(y \mid \theta_i)$ when the individual belongs to class $i = 1, \ldots, k$. The classes have prior probabilities $\pi_1, \ldots, \pi_k$ and the loss in classifying an individual from class i into class j is l_{ij}.
(a) Find the posterior probability $\pi_i(y) = \Pr(\text{class } i \mid y)$ and the posterior risk of allocating the individual to class i.
(b) Now consider the case of 0–1 loss, that is, $l_{ij} = 0$ if $i = j$ and $l_{ij} = 1$ otherwise. Show that the risk is the probability of misclassification.
(b) Suppose that $k = 3$, that $\pi_1 = \pi_2 = \pi_3 = 1/3$ and that Y is normally distributed with mean i and variance 1 in class i. Find the Bayes rule for classifying an observation. Use it to classify the observation $y = 2.2$.

9 Let $Y_j \overset{\text{ind}}{\sim} N(\theta_j, 1)$, $j = 1, \ldots, n$, let $\mu^{\mathrm{T}} = (\mu_1, \ldots, \mu_n)$ be a constant vector, and consider the estimator of $\theta_1, \ldots, \theta_n$ given by

$$\tilde{\theta}_j = \mu + \left\{1 - b \Big/ \sum (Y_i - \mu_i)^2\right\}(Y_j - \mu), \quad j = 1, \ldots, n.$$

Show that the risk under squared error loss, (11.62), reduces to (11.64) with $n-3$ replaced by $n-2$. Discuss the consequences of this.

11.6 Bibliographic Notes

The Bayesian approach to statistics, then called the inverse probability approach, played a central role in the early and middle parts of the nineteenth century, and was central to Laplace's work. It then fell into disrepute after strong attacks were made on the principle of insufficient reason and remained there for many years. During the 1920s and 1930s R. A. Fisher strongly criticised the use of prior distributions to represent ignorance. The publication in 1939 of the first edition of the influential Jeffreys (1961) marked the start of a resurgence of interest in Bayesian inference, which was consolidated by further important advocacy in the 1950s, particularly after difficulties with frequentist procedures emerged. Interest has mounted especially strongly since serious Bayesian computation became routinely possible.

Introductory books on the Bayesian approach are O'Hagan (1988), Lee (1997), and Robert (2001), while the excellent Carlin and Louis (2000) and Gelman *et al.* (1995) are more oriented towards applications; see also Box and Tiao (1973), and Leonard and Hsu (1999). More advanced accounts are Berger (1985) and Bernardo and Smith (1994), while De Finetti (1974, 1975) is *de rigeur* for the serious reader. The likelihood principle and its relation to the Bayesian approach is discussed at length by Berger and Wolpert (1988).

Bayesian model averaging is described by Hoeting *et al.* (1999), who give other references to the topic.

The role and derivation of prior information has been much debated. For some flavour of this, see Lindley (2000) and its discussion. A valuable review of arguments for non-subjective representations of prior ignorance is given by Kass and Wasserman (1996). The elicitation of priors is extensively discussed by Kadane and Wolfson (1998), O'Hagan (1998), and Craig *et al.* (1998).

Laplace approximation is a standard tool in asymptotics, with close links to saddlepoint approximation. A statistical account is given by Barndorff-Nielsen and Cox (1989), which gives further references. It has been used sporadically in Bayesian contexts at least since the 1960s. Tierney and Kadane (1986) and Tierney *et al.* (1989) raised its profile for modern readers. The same idea can be applied to other distributions; see for example Leonard *et al.* (1994).

Markov chain Monte Carlo methods originated in statistical physics. The original algorithm of Metropolis *et al.* (1953) was broadened to what is now called the Metropolis–Hastings algorithm by Hastings (1970), a paper astonishingly overlooked for two decades, though known to researchers in spatial statistics and image analysis (Geman and Geman, 1984; Ripley, 1987, 1988). The last decade has made up for this oversight, with rapid progress being made in the 1990s following Gelfand and Smith (1990)'s adoption of the Gibbs sampler for mainstream Bayesian application. Valuable books on Bayesian use of such procedures are Gilks *et al.* (1996), Gamerman (1997), and Robert and Casella (1999), while Brooks (1998) and Green (2001) give excellent shorter accounts. Example 11.27 is taken from Besag *et al.* (1995), while further interesting applications are contained in Besag *et al.* (1991) and Besag and Green (1993). Tanner (1996) describes a number of related algorithms, including variants on the EM algorithm and data augmentation. Green (1995) and Stephens (2000) describe procedures that may be applied when the parameter space has varying dimension.

Spiegelhalter *et al.* (1996a) describe software for Bayesian use of Gibbs sampling algorithms, with many examples in the accompanying manuals (Spiegelhalter *et al.*, 1996b,c). Cowles and Carlin (1996) and Brooks and Gelman (1998) review numerous convergence diagnostics for Markov chain Monte Carlo output.

Decision theory is treated by Lindley (1985), Smith (1988), Raiffa and Schlaifer (1961), and Ferguson (1967). Hierarchical modelling is discussed in many of the above references. Carlin and Louis (2000) give a modern account of empirical Bayes methods, while the more theoretical Maritz and Lwin (1989) predates modern computational developments. The discovery of the inadmissibility of the maximum likelihood estimator by Stein (1956) and the effects of shrinkage spurred much work; see Morris (1983) for a review.

11.7 Problems

1 Show that the integration in (11.6) is avoided by rewriting it as

$$f(z \mid y) = \frac{f(z \mid y, \theta)\pi(\theta \mid y)}{\pi(\theta \mid y, z)}.$$

Note that the terms on the right need be calculated only for a single θ.
Use this formula to give a general expression for the density of a future observation in an exponential family with a conjugate prior, and check your result using Example 11.3.
(Besag, 1989)

2 (a) Consider a scale model with density $f(y) = \tau^{-1}g(y/\tau)$, $y > 0$, depending on a positive parameter τ. Show that this can be written as a location model in terms of $\log y$ and $\log \tau$, and infer that the non-informative prior for τ is $\pi(\tau) \propto \tau^{-1}$, for $\tau > 0$.
(b) Verify that the expected information matrix for the location-scale model $f(y; \eta, \tau) = \tau^{-1}g\{(y - \eta)/\tau\}$, for real η and positive τ, has the form given in Example 11.10, and hence check the Jeffreys prior for η and τ given there.

3 Show that if $y_1, \ldots, y_n$ is a random sample from an exponential family with conjugate prior $\pi(\theta \mid \lambda, m)$, any finite mixture of conjugate priors,

$$\sum_{j=1}^{k} p_j \pi(\theta, \lambda_j, m_j), \quad \sum_j p_j = 1, p_j \geq 0,$$

is also conjugate. Check the details when $y_1, \ldots, y_n$ is a random sample from the Bernoulli distribution with probability θ.

4 Inference for a probability θ proceeds either by observing a single Bernoulli trial, X, with probability θ, or by observing the outcome of a geometric random variable, Y, with density $\theta(1-\theta)^{y-1}$, $y = 1, 2, \ldots,$. Show that the corresponding Jeffreys priors are $\theta^{-1/2}(1-\theta)^{-1/2}$ and $\theta^{-1}(1-\theta)^{-1/2}$, and deduce that although the likelihoods for X and Y are equal, subsequent inferences may differ. Does this make sense to you?

5 Let y_1, y_2 be the observed value of a random variable from the bivariate density

$$f(y_1, y_2; \theta) = \pi^{-3/2} \frac{\exp\left\{-(y_1 + y_2 - 2\theta)^2/4\right\}}{1 + (y_1 - y_2)^2}, \quad -\infty < y_1, y_2, \theta < \infty.$$

Show that the likelihood for θ is the same as for two independent observations from the $N(\theta, 1)$ density, but that confidence intervals for θ based the average $\overline{y}$ are not the same under both models, in contravention of the likelihood principle.

6 Show that acceptance of the likelihood principle implies acceptance of the sufficiency and conditionality principles.

7 Consider a likelihood $L(\psi, \lambda)$, and suppose that in order to respect the likelihood principle we base inferences for ψ on the integrated likelihood

$$\int L(\psi, \lambda)\, d\lambda.$$

(a) Compare what happens when X and Y have independent exponential distributions with means (i) λ^{-1} and $(\lambda\psi)^{-1}$, (ii) λ and λ/ψ. Discuss.

0	1	2	3	4	3	4	2	2	1
0	2	0	2	4	2	3	3	4	2
1	1	1	1	4	1	5	2	2	3
4	1	2	5	2	0	3	2	1	1
3	1	4	3	1	0	0	2	7	0

Table 11.14
Counts of of balsam-fir seedlings in five feet square quadrats.

(b) Suppose that the parameters in (i) are given prior density $\pi(\psi, \lambda)$ and that we compute the marginal posterior density for ψ. Establish that if the corresponding prior density is used in the parametrization in (ii), the problems in (a) do not arise.

8 Obtain expressions for the mean, variance, and mode of the inverse gamma density (11.14), and express its quantiles in terms of those of the gamma density. Use your results to summarize the posterior density of σ^2 in Example 11.12. Calculate also 95% HPD and equi-tailed credible sets for σ^2.

9 (a) Let y be Poisson with mean θ and gamma prior $\lambda^\nu \theta^{\nu-1} \exp(-\lambda\theta)/\Gamma(\nu)$, for $\theta > 0$. Show that if $\nu = \frac{1}{2}$ and $y = 0$, the posterior density for θ has mode zero, and that a HPD credible set for θ has form $(0, \theta_U)$.
(b) Show that a HPD credible set for $\phi = \log\theta$ has form(ϕ_L, ϕ_U), with both endpoints finite. How does this compare to the interval transformed from (a)? Why does the difference arise?
(c) Compare the intervals in (a) and (b) with the use of quantiles of $\pi(\theta \mid y)$ to construct an equi-tailed credible set for θ, and with confidence intervals based on the likelihood ratio statistic.

You may like to check that for $b > 0$, the function $g(u) = au - be^u$ is concave with a maximum at a finite u if $a > 0$, but that if $a < 0$, it is monotonic decreasing.

10 Use (11.15) to show that the joint conjugate density for the normal mean and variance has $\mu \sim N(\mu_0, \sigma^2/k)$ conditional on σ^2, with σ^2 having an inverse gamma density. Give interpretations of the hyperparameters, and investigate under what conditions the conjugate prior approaches the improper prior in which $\pi(\mu, \sigma^2) \propto \sigma^{-2}$.
Consider instead replacing the prior variance σ^2/k of μ by a known quantity τ^2. Is the resulting joint prior conjugate?

11 Two competing models for a random sample of count data $y_1, \ldots, y_n$ are that they are independent Poisson variables with mean θ, or independent geometric variables with density $\theta(1-\theta)^{y-1}$, for $y = 0, 1, \ldots$, with $0 < \theta < 1$; this density has mean θ^{-1}. Give the posterior odds and Bayes factor for comparison of these models, using conjugate priors for θ in both cases.
What are your prior mean and variance for the numbers of seedlings per five foot square quadrat in a fir plantation? Use them to deduce the corresponding parameters of the conjugate priors for the Poisson and geometric models. Calculate your prior odds and Bayes factor for comparison of the two models applied to the data in Table 11.14. Investigate their sensitivity to other choices of prior mean and variance.

12 Consider a random sample $y_1, \ldots, y_n$ from the $N(\mu, \sigma^2)$ distribution, with conjugate prior $N(\mu_0, \sigma^2/k)$ for μ; here σ^2 and the hyperparameters μ_0 and k are known. Show that the marginal density of the data

$$f(y) \quad \propto \quad \sigma^{-(n+1)} \left(\sigma^2 n^{-1} + \sigma^2 k^{-1}\right)^{1/2} \exp\left[-\tfrac{1}{2}\left\{\frac{(n-1)s^2}{\sigma^2} + \frac{(\overline{y} - \mu_0)^2}{\sigma^2/n + \sigma^2/k}\right\}\right]$$

$$\propto \quad \exp\left\{-\tfrac{1}{2}d(y)\right\},$$

say. Hence show that if Y_+ is a set of data from this marginal density, $\Pr\{f(Y_+) \leq f(y)\} = \Pr\{\chi^2_n \geq d(y)\}$. Evaluate this for the sample 77, 74, 75, 78, with $\mu_0 = 70$, $\sigma^2 = 1$, and $k_0 = \frac{1}{2}$. What do you conclude about the model?
Do the corresponding development when σ^2 has an inverse gamma prior.
(Box, 1980)

13 Suppose that $y_1, \ldots, y_n$ is a random sample from the Poisson distribution with mean θ, and that the prior information for θ is gamma with scale and shape parameters λ and ν. Show that the marginal density of y is

$$f(y) = \frac{s!}{\prod_{j=1}^n y_j!} n^{-s} \times \frac{\Gamma(s+\nu)}{\Gamma(\nu)s!} \frac{\lambda^\nu n^s}{(\lambda+n)^{\nu+s}}, \quad y_1, \ldots, y_n \geq 0,$$

where $s = \sum_j y_j$, and give an interpretation of it.
Suppose that the data in Table 11.14 are treated as Poisson variables, and that prior information suggests that $\lambda = 1$ and $\nu = \frac{1}{2}$. Is this compatible with the data? Do the data seem Poisson, regardless of the prior?

14 In the usual normal linear regression model, $y = X\beta + \varepsilon$, suppose that σ^2 is known and that β has prior density

$$\pi(\beta) = \frac{1}{|\Omega|^{1/2}(2\pi)^{p/2}} \exp\left\{-(\beta-\beta_0)^{\mathrm{T}}\Omega^{-1}(\beta-\beta_0)/2\right\},$$

where Ω and β_0 are known. Find the posterior density of β.

15 Show that the $(1-2\alpha)$ HPD credible interval for a continuous unimodal posterior density $\pi(\theta \mid y)$ is the shortest credible interval with level $(1-2\alpha)$.

16 An autoregressive process of order one with correlation parameter ρ is stationary only if $|\rho| < 1$. Discuss Bayesian inference for such a process. How might you (a) impose stationarity through the prior, (b) compute the probability that the process underlying data y is non-stationary, (c) compare the models of stationarity and non-stationarity?

17 Study the derivation of BIC for a random sample of size n. Investigate the sizes of the neglected terms for nested normal linear models with known variance. Suggest a better model comparison criterion that is almost equally simple.

18 The lifetime in months, y, of an individual with a certain disease is thought to be exponential with mean $1/(\alpha+\beta x)$, where $\alpha, \beta > 0$ are unknown parameters and x a known covariate. Data (x_j, y_j) are observed for n independent individuals, some of the lifetimes being right-censored. The prior density for α and β is

$$\pi(\alpha, \beta) = ab\exp(-\alpha a - \beta b), \quad \alpha, \beta > 0,$$

where $a, b > 0$ are specified. Show that an approximate predictive density for the uncensored lifetime, z, of a future individual with covariate t is

$$\widehat{f}(z|t, y_1, \ldots, y_n) = (\widehat{\alpha} + \widehat{\beta}t)\exp\{-(\widehat{\alpha}+\widehat{\beta}t)z\}, \quad z > 0,$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ satisfy the equations

$$b + \sum_{j=1}^n x_j y_j = \sum_{j\in\mathcal{U}} \frac{x_j}{\alpha+\beta x_j}, \quad a + \sum_{j=1}^n y_j = \sum_{j\in\mathcal{U}} \frac{1}{\alpha+\beta x_j},$$

and $\mathcal{U}$ denotes the set of uncensored individuals.

19 Suppose that (U_1, U_2) lies in a product space, of form $\mathcal{U}_1 \times \mathcal{U}_2$.
(a) Show that

$$\pi(u_1) = \frac{\pi(u_1 \mid u_2)}{\pi(u_2 \mid u_1)} \pi(u_2), \quad \text{for any } u_1 \in \mathcal{U}_1, u_2 \in \mathcal{U}_2,$$

and deduce that for each $u_2 \in \mathcal{U}_2$ and an arbitrary $u_1' \in \mathcal{U}_1$,

$$\pi(u_2) = \left\{ \int \frac{\pi(u_1 \mid u_2)}{\pi(u_2 \mid u_1)} \, du_1 \right\}^{-1} = \frac{\pi(u_2 \mid u_1')}{\pi(u_1' \mid u_2)} \left\{ \int \frac{\pi(u_2 \mid u_1')}{\pi(u_1' \mid u_2)} \, du_2 \right\}^{-1}.$$

(b) If $U_2^1, \ldots, U_2^S$ is a random sample from $\pi(u_2 \mid u_1')$, show that

$$\widehat{\pi}(u_2) = \frac{\pi(u_2 \mid u_1')}{\pi(u_1' \mid u_2)} \left\{ S^{-1} \sum_{s=1}^{S} \pi(u_1' \mid U_2^s)^{-1} \right\}^{-1} \xrightarrow{P} \pi(u_2) \text{ as } S \to \infty.$$

(c) Verify that the code below applies this approach to the bivariate normal model in Example 11.21.

```
S <- 1000; rho <- 0.75; u1p <- -2  # u1p is u1prime
z <- seq(from=-4,to=4,length=200)
plot(z,dnorm(z),type="l",ylim=c(0,1.5))
for (r in 1:20)          # 20 replicates of the simulation
{ u2.sim <- rnorm(S, rho*u1p, sqrt(1-rho^2))
  if (r==1) rug(u2.sim)   # rug with one of the u2 samples
  const <- mean( 1/dnorm(u1p,rho*u2.sim,sqrt(1-rho^2)) )
  dz <- dnorm(z,rho*u1p,sqrt(1-rho^2))/dnorm(u1p,rho*z,sqrt(1-rho^2))
  lines(z, dz/const) }
```

Does this work well? Why not? Try with $u_1' = -2, -1, 0$.
What lesson does this example suggest for the use of this approach in general?

20 (a) Let (U_1, U_2) have a joint density π, marginal densities π_1 and π_2, and conditional densities $\pi_{1|2}$ and $\pi_{2|1}$. Show that π_1 satisfies the integral equation

$$\pi_1(u) = \int h(u, v) \pi_1(v) \, dv, \quad \text{where} \quad h(u, v) = \int \pi_{1|2}(u \mid w) \pi_{2|1}(w \mid v) \, dw.$$

(b) In Example 11.21, establish that the conditional distributions of $U_2^{(i+1)} \mid U_1^{(i)} = v$, $U_1^{(i+1)} \mid U_2^{(i+1)} = w$, and $U_1^{(i+1)} \mid U_1^{(i)} = v$, $i = 1, \ldots, I-1$, are those of

$$\rho v + (1-\rho^2)^{1/2} \varepsilon_1, \quad \rho w + (1-\rho^2)^{1/2} \varepsilon_2, \quad \rho^2 v + (1-\rho^4)^{1/2} \varepsilon_3,$$

where $\varepsilon_j \overset{\text{iid}}{\sim} N(0,1)$. Hence write down $h(u, v)$ for this problem.
(c) Show by induction that the conditional distribution of $U_1^{(I+1)} \mid U_1^{(1)} = v$ is the same as that of $\rho^{2I} v + (1-\rho^{4I})^{1/2} \varepsilon_4$, and hence show that (i) the Markov chain $U_1^{(1)}, U_1^{(2)}, \ldots$ is in equilibrium when $U_2^{(0)}$ has the standard normal density, and (ii) the chain will reach equilibrium provided $U_2^{(0)}$ may not equal $\pm\infty$.

21 The unmodified Gibbs sampler can be a poor way to generate values from a

posterior density with several widely separated modes. Let $U = (U_1, U_2)^{\mathrm{T}}$ and consider

$$\pi(u) = \gamma\phi(u_1 - \delta)\phi(u_2 - \delta) + (1 - \gamma)\phi(u_1 + \delta)\phi(u_2 + \delta),$$

where $u = (u_1, u_2)^{\mathrm{T}}$, $0 < \gamma < 1$ and $\delta > 0$; this is a mixture of two bivariate normal densities whose separation depends on δ and whose relative sizes depend on γ.
(a) When $\gamma = 1/2$, sketch contours of π and the conditional density of U_1 given $U_2 = u_2$ for $u_2 = -2\delta, \delta, 0, \delta, 2\delta$. Sketch also some sample paths for a Gibbs sampling algorithm. What problem do you foresee if $\delta > 4$, say?
(b) Show that the conditional density of U_1 given $U_2 = u_2$ may be written

$$\alpha(u_2)\phi(u_1 - \delta) + \{1 - \alpha(u_2)\}\phi(u_1 + \delta), \quad \text{where} \quad \alpha(u_2) = \frac{\gamma e^{2\delta u_2}}{1 - \gamma + \gamma e^{2\delta u_2}},$$

and write down a Gibbs sampling algorithm for π.
(c) If $c > 0$ is large enough that $\Phi(-c)$ is negligible, show that the probability that the sampler stays in the same mode during R iterations of the sampler is bounded below by

$$\exp\left\{-2R(\gamma^{-1} - 1)e^{-2\delta c}\right\},$$

and compute this for $\delta = 2$, 3 and some suitable values of c. Comment.
(d) Find the joint distribution of $V = (V_1, V_2)^{\mathrm{T}} = 2^{-1/2}(U_1 + U_2, U_1 - U_2)^{\mathrm{T}}$ and show that if simulation is performed in terms of V, convergence is immediate. Comment on the implications for implementing the Gibbs sampler.

22 Table 5.9 gives data from k clinical trials as 2×2 tables $(R_{Tj}, m_{Tj}; R_{Cj}, m_{Cj})$, where R_{Tj} is the number of deaths in the treatment group of m_{Tj} patients and similarly in the control group, for $j = 1, \ldots, k$. As a model for such data, ignoring publication bias, assume that R_{Cj} and R_{Tj} are independent binomial variables with denominators m_{Cj} and m_{Tj} and probabilities

$$\frac{\exp(\mu_j)}{1 + \exp(\mu_j)}, \quad \frac{\exp(\mu_j + \delta_j)}{1 + \exp(\mu_j + \delta_j)}, \quad j = 1, \ldots, k,$$

where $\delta_j \overset{\text{iid}}{\sim} N(\gamma, \tau^2)$ represent the treatment effects. Suitable prior densities are assumed for $\mu_1, \ldots, \mu_k$, γ and τ^2.
(a) Write down the directed acyclic graph for this model, derive its conditional independence graph, and hence give steps of a Markov chain Monte Carlo algorithm to sample from the posterior density of $\mu_1, \ldots, \mu_k$, γ and τ^2. If any steps require Metropolis–Hastings sampling, suggest how you would implement it and give the acceptance probabilities.
(b) How does your sampler change if one of the R_Cs is missing?
(c) How should your sampler be modified to generate from the posterior predictive density of δ_+, the value of δ for a new trial?
(d) How should your algorithm be modified if an hierarchical model is used for the μ_j?

23 A Poisson process with rate

$$\lambda(t) = \begin{cases} \lambda_0, & 0 < t \le \tau, \\ \lambda_1, & \tau < t \le t_0, \end{cases}$$

where τ is known, is observed on the interval $(0, t_0]$. Let n_0 and n_1 denote the numbers of events seen before and after τ, and suppose that λ_0 and λ_1 are

independent gamma variables with parameters ν and β, where ν is specified and β has a gamma prior density with specified parameters a and b.
(a) Check that the joint density of $n_0, n_1, \lambda_0, \lambda_1$, and β is

$$\frac{(\lambda_0\tau)^{n_0}}{n_0!}e^{-\lambda_0\tau}\frac{\{\lambda_1(t_0-\tau)\}^{n_1}}{n_1!}e^{-\lambda_1(t_0-\tau)}\frac{\lambda_0^{\nu-1}\beta^\nu}{\Gamma(\nu)}e^{-\lambda_0\beta}\frac{\lambda_1^{\nu-1}\beta^\nu}{\Gamma(\nu)}e^{-\lambda_1\beta}\frac{\beta^{a-1}b^a}{\Gamma(a)}e^{-b\beta}.$$

Show that λ_0, λ_1, and β have gamma full conditional densities, and hence give a reversible Gibbs sampler algorithm for simulating from their joint posterior density. Extend this to a process with known multiple change points $\tau_1, \ldots, \tau_k$, for which

$$\lambda(t) = \begin{cases} \lambda_0, & 0 < t \le \tau_1, \\ \lambda_1, & \tau_1 < t \le \tau_2, \\ \cdots & \cdots \\ \lambda_k, & \tau_k < t \le t_0. \end{cases}$$

(b) Now suppose that ν is unknown, with prior gamma density with specified parameters c and d. Show that a random walk Metropolis–Hastings move to update $\log\nu$ to $\log\nu'$ has acceptance probability

$$\min\left[1, \left\{\frac{\Gamma(\nu)}{\Gamma(\nu')}\right\}^{k+1}\left(\frac{\nu'}{\nu}\right)^c\left(e^{-d}\beta^{k+1}\prod\lambda_j\right)^{\nu'-\nu}\right].$$

How would you add this to the algorithm in (a) to retain reversibility?
(c) Now suppose that although k is known, $\tau_1, \ldots, \tau_k$ are not. Show that the joint density of the even order statistics from a random sample of size $2k+1$ from the uniform density on $(0, t_0)$ is proportional to

$$\tau_1(\tau_2-\tau_1)\cdots(\tau_k-\tau_{k-1})(t_0-\tau_l), \quad 0 < \tau_1 < \cdots < \tau_k < t_0.$$

Suppose that this is taken as the prior for the positions of the k changepoints, and that these are updated singly with proposals in which τ_i' is drawn uniformly from (τ_{i-1}, τ_{i+1}), with obvious changes for τ_1 and τ_k. Find the acceptance probabilities for these moves.

24 In a Bayesian formulation of Problem 6.16, we suppose that the computer program is one of many to be debugged, and that the mean number of bugs per program has a Poisson distribution with mean μ/β, where $\mu, \beta > 0$. The actual number of bugs in a particular program is m, and each gives rise to a failure after an exponential time with mean β^{-1}, independent of the others. On failure, the corresponding bug is found and removed at once.
(a) Debugging takes place over the interval $[0, t_0]$ and failures are seen to occur at times $0 < t_1 < \cdots < t_n < t_0$. Show that

$$f(y \mid m, \beta) = \frac{m!}{(m-n)!}\beta^n \exp\{-\beta t_0(m + s/t_0 - n)\}, \quad \beta > 0, m = n, n+1, \ldots,$$

where y represents the failure times and $s = \sum_{j=1}^n t_j$.
(b) We take prior $\pi(\mu, \beta) \propto \mu^{-2}$, $\mu > 0$. Show that

$$\begin{aligned} \pi(y, m) &= \int_0^\infty\int_0^\infty f(y \mid m, \beta) f(m \mid \beta, \mu)\pi(\beta, \mu)\, d\beta d\mu \\ &\propto (m - n + s/t_0)^{-n}\prod_{i=1}^{n-2}(m-n+i), \quad m = n, n+1, \ldots, \end{aligned}$$

and give expressions for the posterior probabilities (i) that the program has been entirely debugged and (ii) that there are no failures in $[t_0, t_0 + u]$.
(c) Use the data in Table 6.13 to give a 95% HPD credible interval for the number of bugs remaining after 31 failures. Compute the probability that the program had been entirely debugged (i) after 31 failures and (ii) after 34 failures. Should the program have been released when it was?
(d) Discuss how the appropriateness of the model might be checked.
(Example 2.28; Raftery, 1988)

25 Let $Y_1, \ldots, Y_n$ be independent normal variables with means $\mu_1, \ldots, \mu_n$ and common variance σ^2. Show that if the prior density for μ_j is

δ is the Dirac delta function.

$$\pi(\mu_j) = \gamma\tau^{-1}\phi(\mu_j/\tau) + (1-\gamma)\delta(\mu_j), \quad \tau > 0, 0 < \gamma < 1,$$

with all the μ_j independent *a priori*, then $\pi(\mu_j \mid y_j)$ is also a mixture of a point mass and a normal density, and give an interpretation of its parameters.
(a) Find the posterior mean and median of μ_j when σ is known, and sketch how they vary as functions of y_j. Which would you prefer if the signal is sparse, that is, many of the μ_j are known *a priori* to equal zero but it is not known which?
(b) How would you find empirical Bayes estimates of τ, γ, and σ?
(c) In applications of the tails of the normal density might be too light to represent the distribution of non-zero μ_j well. How could you modify π to allow for this?

26 Suppose that $y_1, \ldots, y_n$ are independent Poisson variables with means $\lambda_j x_j$, where the x_j are known constants, and that the λ_j are a random sample from the gamma density with mean ξ/ν.
(a) Show that the marginal density of y_j is

$$f(y_j; \xi, \nu) = \frac{\Gamma(y_j + \xi)}{\Gamma(\xi) y_j!} \frac{x_j^{y_j} \nu^{\xi}}{(x_j + \xi)^{y_j + \xi}}, \quad y_j = 0, 1, \ldots, \quad \xi, \nu > 0,$$

and give its mean. Say how you would estimate ξ and ν based on $y_1, \ldots, y_n$.
(b) Establish that

$$\mathrm{E}(\lambda_j \mid y, \xi, \nu) = \frac{y_j + \xi}{x_j + \nu}, \quad \mathrm{var}(\lambda_j \mid y, \xi, \nu) = \frac{y_j + \xi}{(x_j + \nu)^2},$$

and give an interpretation of this.
(c) Check that the code below computes the maximum likelihood estimates $\widehat{\xi}$ and $\widehat{\nu}$, and applies it to the data in Table 11.7. Discuss.

```
x <- c(94.32,15.72,62.88,125.76,5.24,31.44,1.048,1.048,2.096,10.48)
y <- c(5,1,5,14,3,19,1,1,4,22)
L <- function(p, y, x)
        -sum(dnbinom(y, size=p[1], prob=p[2]/(x+p[2]), log=T))
fit <- nlm(L, p=c(1,1), y=y, x=x) # marginal maximum likelihood
xi <- fit$estimate[1]
nu <- fit$estimate[2]
ests <- (y+xi)/(x+nu)
vars <- (y+xi)/(x+nu)^2
cbind(ests,vars)
```