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

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

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.

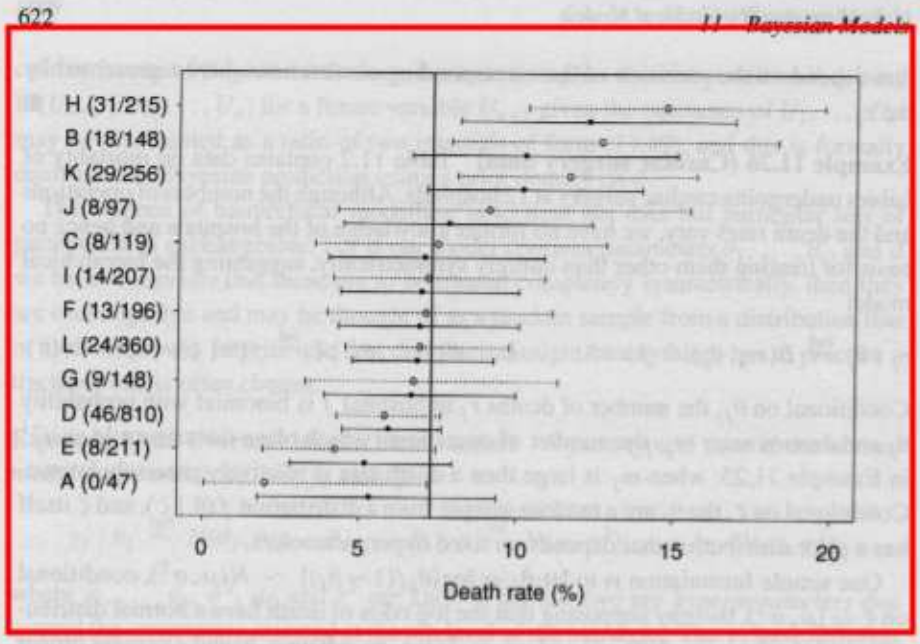

Figure 11.11 Posterior summaries for mortality rates for cardiac surgery data. Posterior means and 0.95 equitailed credible intervals for separate analyses for each hospital are shown by hollow circles and dotted lines, while blobs and solid lines show the corresponding quantities for a hierarchical 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.

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 $y = (y_1, \ldots, y_n)^T$ 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 $\beta = (\beta_1, \ldots, \beta_p)^T$ 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

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

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

Table 11.2 Mortality r/m from cardiac surgery in 12 hospitals (Spiegelhalter *et al.*, 1996, 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

provided the mode lies inside the parameter space. Here $\tilde{J}(\theta)$ is the second derivative matrix of $-\tilde{l}(\theta)$. 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 $1-2\alpha$. 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