- enrichment of school diet with ricebran
- mouthrinse with 0.2% NaF-solution
- oral hygiene

The response is dmfte and the impact of initial dental status modelled via a variable log(dmftb + 0.5). A Be(1,1) prior is assumed on ω and $N(0, 1000)$ priors on the regression coefficients.

Iterations 501–5000 of a two-chain run show a mean probability ω of 0.05. Treatments 1, 2 and 5 have entirely negative 95% credible intervals (i.e. reduces tooth decay), namely -0.23 $(-0.39, -0.05)$, $-0.32(-0.52, -0.12)$ and $-0.23(-0.39, -0.07)$. Böhning *et al.* (1999, p. 202) consider modelling the mixture weights for strata defined by school. Thus ω becomes a vector of six probabilities.

6.6 DISCRETE MIXTURES COMBINED WITH PARAMETRIC RANDOM EFFECTS

Discrete mixture models may identify subpopulations or outlying clusters of cases, whereas the random effects models of Chapter 5 often remove overdispersion. To fully model multimodality, isolated outliers, as well as overdispersion, one may consider discrete mixtures of the conjugate normal–normal, poisson–gamma or beta–binomial models (Moore *et al.*, 2001) or discrete mixtures of poisson–lognormal or binomial–logitnormal models. That is, a discrete mixture strategy is combined with parametric random effects, rather than replacing it. Lenk and Desarbo (2000) advocate such a strategy for nested data models involving repeated observations over time or within clusters; they argue that an excessive number of classes C will be used if allowance is not made for (parametric) heterogeneity within classes.

For an illustration with binomial data, let $y_i \sim \text{Bin}(n_i, \kappa_i)$, where

$$\kappa_i \sim \sum_{j=1}^{C} \pi_j \text{Beta}(\alpha_{ij}, \beta_{ij}).$$

A reparameterisation of the Beta in terms of $\alpha_{ij} = \rho_{ij}\gamma_j$ and $\beta_{ij} = (1 - \rho_{ij})\gamma_j$ facilitates regression modelling (e.g. a logit regression for predicting the mean probabilities ρ_{ij} using predictors X_i). It also permits simple identifiability constraints (e.g. $\rho_1 > \rho_2 > \cdots > \rho_C$). When predictors are not used, one has $\alpha_j = \rho_j\gamma_j$, $\beta_j = (1 - \rho_j)\gamma_j$.

Such a mixture strategy also characterises a class of outlier detection models (e.g. Albert, 1999). Consider a conjugate Poisson-gamma mixture model, with $y_i \sim \text{Po}(\nu_i)$ and $\nu_i \sim \text{Ga}(\alpha, \alpha/\mu_i)$, where $\mu_i = \exp(X_i\beta)$. The parameter α is a precision parameter – as $\alpha \to \infty$ the Poisson is approached. For outlier resistance one may assume the discrete mixture

$$\nu_i \sim \pi \text{Ga}(K\alpha, K\alpha/\mu_i) + (1 - \pi)\text{Ga}(\alpha, \alpha/\mu_i),$$

where π is small (e.g. $\pi = 0.05$) and $0 < K < 1$ (e.g. $K = 0.25$). The first component is 'precision deflated'. In a non-conjugate Poisson–lognormal mixture model with $y_i \sim \text{Po}(\mu_i)$ and $\log(\mu_i) = \beta X_i + u_i$, one might similarly take

$$u_i \sim \pi N(0, K\varsigma) + (1 - \pi)N(0, \varsigma),$$

where $K > 1$ (e.g. $K = 5$ or $K = 10$).

Example 6.6 Heart transplant mortality Albert (1999) considers variations in heart transplant mortality across 94 hospitals using Poisson–gamma mixture models, $y_i \sim \text{Po}(e_i \nu_i)$, where e_i are expected deaths. A single-component gamma-mixing model with $\nu_i \sim \text{Ga}(\alpha, \alpha/\mu)$ is compared with a two-component model allowing for possible outliers. Thus

$$\nu_i \sim \pi \text{Ga}(K\alpha, K\alpha/\mu) + (1 - \pi)\text{Ga}(\alpha, \alpha/\mu)$$

with prior outlier probability $\Pr(H_i = 1) = \pi = 0.1$ and with $K = 0.2$. Iterations 1001–5000 of a two-chain run show the highest outlier probabilities, $\Pr(H_i = 1|y)$ are for hospitals 85 and 63, namely 0.144 and 0.129 compared to the prior probability of 0.10. These hospitals have zero deaths, despite expected deaths of 5.8 and 3.8, respectively.

6.7 NON-PARAMETRIC MIXTURE MODELLING VIA DIRICHLET PROCESS PRIORS

In applications of hierarchical models, including parametric mixture models, there are questions of sensitivity of inferences to the assumed forms (e.g. normal, gamma) for the higher stage priors. The distributions of parameters, including higher stage hyperparameters for random effects, are often uncertain, and not acknowledging this uncertainty may unwarrantedly raise the precision attached to posterior inferences. Alternatively inferences may be distorted by outlying points or by multimodality in random effects or regression errors (i.e. by inconsistencies with the assumed higher level prior). Instead of assuming a known higher stage prior density for random effects θ_i (e.g. MVN or gamma), the DP approach lets the form of the higher stage density G itself be uncertain (West *et al.*, 1994).

The DP strategy involves a baseline density G_0, the prior expectation of G, and a precision parameter α governing the concentration of the prior for G about the mean G_0. As α becomes larger, the concentration around the baseline prior increases, whereas small α (e.g. under 5) tends to result in relatively large departures from the form assumed by G_0. The case $\alpha \to \infty$ means the DPP prior becomes equivalent to a parametric model with G_0 known. For any partition $B_1, \ldots, B_M$ on the support of G_0 the vector of probabilities $\{G(B_1), \ldots, G(B_M)\}$ follows a Dirichlet distribution with parameter vector $\{\alpha G_0(B_1), \ldots, \alpha G_0(B_M)\}$.

Let $y_i, i = 1, \ldots, n$ be drawn from a distribution with unknown parameters θ_i, φ_i

$$f(y_i|\theta_i, \varphi_i)$$

and suppose there is greater uncertainty about the prior for parameters θ_i than for parameters ϕ_i (Escobar and West, 1998). One may adopt a DPP for the θ_i, but a conventional parametric prior for ϕ_i. Under a DPP, a baseline prior G_0 is assumed from which candidate values for θ_i are drawn. So instead of a prior $\theta_i \sim G(\theta_i|\gamma)$ with G a known density and γ a hyperparameter, the uncertainty about the form of the prior is represented by introducing an extra step in the hierarchical specification

$$\theta_i|G \sim G$$
$$G|\alpha, \gamma \sim \text{DP}(\alpha, G_0),$$

where G_0 has hyperparameters γ.

There are several ways to implement a DPP. Following Sethuraman (1994), one way to generate the DPP is to regard the θ_i as iid with density function $q(\ )$ which is an infinite mixture of point masses or continuous densities (Hirano, 1998; Ohlssen *et al.*, in press). This is also known as the 'constructive definition' of the Dirichlet process (Walker *et al.*, 1999). If G_0 consists of a continuous density f, then the DP forms a mixture of continuous densities

$$q(\theta_i) = \sum_{j=1}^{\infty} p_j f(\theta_i|\gamma).$$

This structure is known as a mixed Dirichlet process (Walker *et al.*, 1999, p. 489) and overcomes certain limitations of the original DPP of Ferguson (1973). For example, a DP mixture with normal base densities would be

$$q(\theta_i) = \sum_{j=1}^{\infty} p_j N(\theta_i|\mu_j, \phi_j).$$

Ishwaran and Zarepour (2000) and Ishwaran and James (2002) suggest that this may be truncated at M components with

$$q(\theta_i) = \sum_{j=1}^{M} p_j N(\theta_i|\mu_j, \phi_j)$$

and $\Sigma_{j=1}^{M} p_j = 1$. This leads to an approximate or truncated DP which may be denoted

$$\theta_i \mid G \sim G$$
$$G \mid M, \alpha, \gamma \sim \text{TDP}(\alpha, G_0).$$

Ishwaran and James (2002, pp. 5–6) detail the usually close accuracy of this approximation to the infinite DP for typical α and M values.

The most appropriate value θ_m^* for case i is then selected using a Dirichlet vector of length M with probabilities p_m for each value determined by the precision parameter α. The mixture weights p_j are constructed by 'stick-breaking' (Ishwaran and Zarepour, 2000, p. 384). Thus set $V_M = 1$ and draw $M - 1$ beta variables

$$V_j \sim \text{Be}(1, \alpha) \qquad j = 1, \ldots, M$$

and set

$$\begin{aligned} p_1 &= V_1 \\ p_2 &= V_2(1 - V_1) \\ p_3 &= V_3(1 - V_2)(1 - V_1) \\ &\vdots \\ p_M &= V_M(1 - V_{M-1})(1 - V_{M-2}) \cdots (1 - V_1). \end{aligned}$$

Alternative versions of the stick-breaking prior are discussed by Ishwaran and James (2001) and Ishwaran and Zarepour (2000). For example, one possible alternative (the Poisson-Dirichl

process) has two parameters and assumes

$$V_j \sim \text{Beta}(1-a, b+ja),$$

where $0 \leq a < 1$ and $b > -a$.

If the TDP approach is adopted, one may use the prior on the concentration parameter α to decide the maximum number of potential clusters. Ohlssen *et al.* (in press) present an approximation based on the size of the probability ε of the final mass point p_M, $\varepsilon = E(p_M)$. Then

$$M \approx 1 + \log(\varepsilon)/\log[\alpha/(1+\alpha)]$$

and the choice of the prior on α determines (or should be consistent with) the choice of M. For example, taking $\varepsilon = 0.01$ and $\alpha \sim \text{Unif}(0.5, 10)$ implies M between 5.2 and 49.3, so M might be taken as 50.

A sensible M will also reflect the nature of the data. Suppose in a data smoothing context without predictors (e.g. ranking hospital death rates) that θ_i denote unknown means for each case $i = 1, \ldots, n$. Then a degree of clustering is anticipated in these values so that the data for similar groups of cases suggest that the same value of θ_i would be appropriate for them. In certain cases such as the eye-tracking anomaly data considered earlier, the maximum number of clusters is likely to be considerably less than the number of distinct observations. In that example, there were only 19 distinct values of the count of anomalies, even though there were 104 observations. In other cases heterogeneity in the data might be such that every single case might potentially be a cluster. Thus if every y_i were distinct in value, or even though some y_i were matching they had different predictors, then the maximum number of clusters could be n.

In general, one draws $m = 1, \ldots, M$ values potential values θ_m^* for θ_i from the baseline density G_0, where M is the anticipated maximum possible number of clusters. This maximum may be n or considerably less if there are repeat observations and no predictors are involved. In practice, only $M^* \leq M \leq n$ distinct values of the M sampled will be allocated to one or more of the n cases.

Another option is based on the Polya Urn representation of the Dirichlet process. Under this, θ_1 is necessarily drawn from G_0, while θ_2 equals θ_1 with probability p_1 and is from the base density with probability $p_0 = 1 - p_1$. Then θ_3 equals θ_1 with probability p_1, equals θ_2 with probability p_2 and is drawn from the base density with probability $p_0 = 1 - p_1 - p_2$ and so on. Finally θ_N equals each preceding θ_i with probability p_i and is drawn from the base density with probability $p_0 = 1 - (p_1 + \cdots + p_{N-1})$. Conditional on $\theta_{[i]} = \{\theta_j, j \neq i\}$, θ_i is drawn from the mixture

$$p(\theta_i|\theta_{[i]}) \propto \sum_{j \neq i} q_j \delta(\theta_j) + \alpha q_0 f(y_i|\theta_i) g(\theta_i|\gamma),$$

where $\delta(\theta_j)$ are discrete measures concentrated at θ_j, $q_j = f(y_i|\theta_j)$, the sampling density of y, and $p_j (j = 0, \ldots, N-1)$ in the Polya Urn scheme are obtained by normalising the values $q_1, q_2, \ldots, \alpha q_0$. The form of q_0 may be obtained analytically when g, the density associated with G_0, is conjugate with the likelihood $f(y|\theta)$ (Kleinman and Ibrahim, 1998). For example, if G_0 is $N(\mu, \sigma^2)$ then $g(\ )$ is $\phi(|\mu,\sigma^2)$. Some problems with this prior are noted by Ishwaran and Zarepour (2000, p. 373).

Often the goal is to use the clusters to achieve a non-parametric smoothing of the data or random effects. Predictive inferences about the underlying population may then be based on sampling new values which may be drawn from different clusters than the observed data (Turner and West, 1993; West, 1992b). As an example, for an overdispersed Poisson outcome, $y_i \sim \text{Po}(\mu_i)$, $i = 1, \ldots, n$, one option might be

$$\log(\mu_i) = \beta + \varepsilon_i$$

with $\varepsilon_i \sim N(0, \tau)$. To insert a DP stage, $N(0, \tau)$ is taken the baseline prior G_0 and $M \leq n$ candidate values ε_m^* sampled from it. The cases $i = 1, \ldots, n$ are allocated to one of these candidate values according to the probabilities determined by the Dirichlet process. This procedure is repeated at each iteration in an MCMC chain. So if case i is allocated to cluster j (i.e. if the configuration indicator $H_i = j$) with candidate value ε_j^*, then $\varepsilon_i = \varepsilon_j^*$ and $y_i \sim \text{Po}(\mu_j^*)$, where

$$\log\left(\mu_j^*\right) = \beta + \varepsilon_j^*.$$

The posterior average error ε_i will be based on averaging over the candidate values assigned at each iteration in the chain.

Alternatively, DP mixing may be used in regression applications and mixing over errors in general linear models is one approach to modelling overdispersion in exponential regression models. These are defined by

$$f(y_i|\nu_i) = c(y_i)\exp[\nu_i y_i - b(\nu_i)]$$
$$g(\mu_i) = X_i\beta$$

with mean $\mu = b'(\nu)$ and variance $V(\mu) = b''(\nu)$, and where β_1 is the intercept. Set X^* equal to X excluding a constant $x_{i1} = 1$ and introduce errors ε_i

$$g(\mu_i) = \beta_1 + X_i^*\beta + \varepsilon_i$$

then DP mixing over the errors is equivalent to modelling heterogeneity in intecepts $\alpha_i = \beta_1 + \varepsilon_i$. Mukhopadhyay and Gelfand (1997) refer to models that mix over the intercepts in this way as DP mixed GLMs, defined by the density

$$f(y|X^*, \beta, G_0) = \int f(y|X^*, \beta, \alpha)dG_0(\alpha).$$

Note that the DPP procedure has some apparent resemblance to standard discrete mixture analysis. Differences are that the number of clusters is random and the average number of clusters M^* emerging from a particular data set, and the chances that a new observation will be drawn from existing or new cluster, depend crucially on the value or prior assumed or α. For large values of α the allocation will be such that most candidate values will be selected and the actual density of ε will be close to the baseline. Selecting a large α leads to more clusters and may result in 'overfitting' or densities that seem implausibly smoothed in terms of prior beliefs about the appropriate number of subgroups (Hirano, 1998). For small α, the allocation is likely to be concentrated on a small number of the candidate values. In this case the DP model comes to resemble a finite (parametric) mixture model.

Appropriate priors, typically $\alpha \sim \text{Ga}(a, b)$ or $\alpha \sim \text{Ga}(k, k/c)$, where c is the prior mean for α, may be set on the precision parameter α. For example, West and Turner (1994) use

the relatively informative prior $\alpha \sim \text{Ga}(10, 10/c)$. Ishwaran and James (2002) recommend $\alpha \sim \text{Ga}(2, 2)$, as it encourages both small and large values of α, and use the result that under the TDP approximation, α may be updated via Gibbs sampling using the conditional

$$\alpha | V \sim \text{Ga}(M + a - 1, b - \log p_M).$$

Mukhopadhyay and Gelfand (1997) in their analysis of overdispersed binomial regression assume $\alpha \sim \text{Ga}(1, 1)$. A form of data augmentation may also be used to sample α (see Escobar and West, 1998, p. 10). The prior on α in turn induces a prior on the actual number of clusters M^* present at any iteration (Antoniak, 1974), with M^* expected to approximately equal $\alpha \log_e (1 + n/\alpha)$. It may be sufficient, however, to select a few trial values of α and assess the impact on the average number of actual clusters (Ibrahim and Kleinman, 1998; Turner and West, 1993). Some possible problems with the identifiability of this parameter are considered by Leonard (1996), especially in data without any ties in the outcome variable.

Example 6.7 Eye-tracking data Consider again the eye-tracking data and assume a Poisson-gamma mixture to model the heterogeneity. A standard approach to such overdispersed count data assumes Poisson sampling, with $y_i \sim \text{Po}(\theta_i)$ and gamma priors on the Poisson means, $\theta_i \sim \text{Ga}(a, b)$, where a and b are preset or themselves assigned priors. Following Escobar and West (1998), initially choose a baseline gamma prior for the θ_i with a and b having preset values, $a = b = 1$. The insertion of a DPP stage means sampling $M \leq n$ candidate values θ_m^* from the baseline $\text{Ga}(a, b)$ density and then allocating each of the $n = 104$ cases to one of these values. Because there are only 19 distinct count values in the sample, one may take $M = 19$ as the maximum possible number of clusters.

The data augmentation prior for α, as in Escobar and West (1998), is used in the code

```
{ for( i in 1 : n) {theta[i] <- theta.star[H[i]];  y[i] ~ dpois(theta[i])
H[i] ~ dcat(p[]); for (j in 1:M) {SC[i,j] <- equals(j,H[i])}}
# Precision Parameter
eta ~ dbeta(alphs,M); alphs <- alpha+1;
a1 <- a+Mstar; b1 <- b - log(eta); a2 <- a+Mstar-1; b2 <- b1
logit(p.alph) <- log(a2)-log(M)-log(b-log(eta))
alph1 ~ dgamma(a1,b1); alph2 ~ dgamma(a2,b2);
alpha <- p.alph*alph1+(1-p.alph)*alph2
# Constructive prior
    p[1] <- V[1]; V[M] <- 1
    for (j in 2:M) {p[j] <- V[j]*(1-V[j-1])*p[j-1]/V[j-1]}
    for (k in 1:M-1){ V[k] ~ dbeta(1,alpha)}
# theta.star prior, hyperparameters
    A ~ dexp(0.1) B ~ dgamma(0.1,0.1)
    for (m in 1:M){ theta.star[m] ~ dgamma(A,B)}
# total clusters
    Mstar <- sum(CL[]); for (j in 1:M) {CL[j] <- step(sum(SC[,j])-1)}}
```

This example shows the ability of a non-parametric analysis to detect discrepancies between prior and data. A two-chain run of 5000 iterations (500 burn in) produces a bimodal posterior distribution for larger values of y_i because the $G(1, 1)$ prior on cluster effects θ_m^* $(m = 1, \ldots, M)$ is too inflexible to accommodate them. Thus case 92 with $y_i = 12$ has

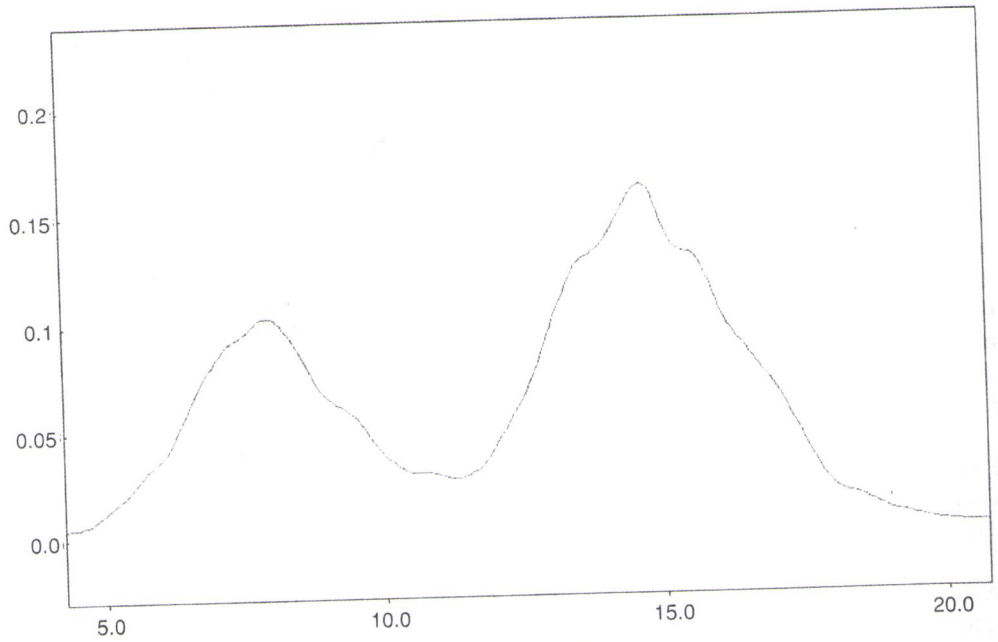

Figure 6.1 Kernel density for θ_{92}.

posterior mean of 12.3 (and relatively large standard deviation 3.8) but the posterior density shows the conflict between prior and data (Figure 6.1). With a Ga(1, 1) prior on the precision parameter α, the average number of clusters chosen is 14.6, and α has posterior mean 6.5.

Instead, let the baseline gamma prior for the θ_i involve unknown hyperparameters with priors $a \sim E(0.1)$, $b \sim \text{Ga}(0.1, 0.1)$. The posterior means are now $a = 0.4$, $b = 0.08$ from a two-chain run of 5000 iterations. The posterior for θ_{92} is no longer bimodal but still has some skewness. The mean number of clusters is now 15.

Example 6.8 Galaxy velocities To illustrate Normal mixture analysis under a DPP, consider data on velocities (km/sec) for 82 galaxies from Roeder (1990). These are drawn from six well-separated conic sections of the Corona Borealis region. Thus with equal variances across components

$$\begin{aligned} y_i|H_i &\sim N(\mu_{H_i}, \phi) \\ \mu_j &\sim G \\ G|\alpha &\sim \text{DP}(\alpha G_0) \\ G_0 &= N(\mu_0, d\phi). \end{aligned}$$

A Ga(1.5, 1) prior for α is adopted, in line with a prior belief of six clusters when $n = 82$ and the maximum number of clusters taken as $M = 10$. For the parameters ϕ^{-1} and d, gamma priors are used, namely $\phi^{-1} \sim \text{Ga}(1,0.001)$, $d \sim \text{Ga}(2.5, 0.1)$. West (1992b) discusses this model structure and appropriate priors on α, d and ϕ^{-1}.

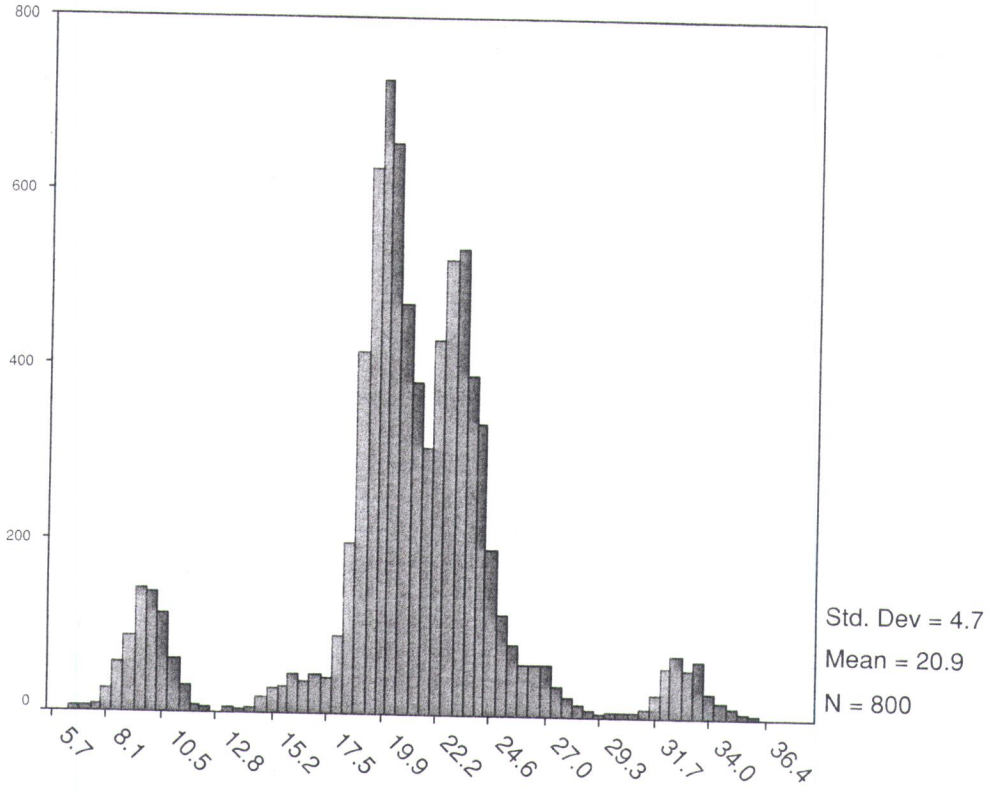

Figure 6.2 Density of y. new.

Predictions from the model are based on sampling a single replicate observation. This involves selecting a new cluster, not necessarily included in the clusters selected for the actual observations (Turner and West, 1993) and then sampling the density of the appropriate cluster mean. This predictive density may be used in various ways, but here it is used to assess whether the predictive velocity exceeds 25 000 km/sec.

A two-chain run of 5000 iterations (convergent at 1000) gives a density for a new value as in Figure 6.2. This shows small subpopulations at approximately 9000 and 33 000 km/sec as are apparent in the original data. The probability that the prediction exceeds 25 000 km/sec is estimated at 0.092 and the parameter d at around 42. The posterior for α has mean 2.7, with the average number of non-empty clusters M^* at 8.7 and 95% of non-empty clusters being between 6 and 10.

6.8 OTHER NON-PARAMETRIC PRIORS

Alternatives to DP priors have been proposed, such as stochastic process priors and partition priors (Walker *et al.*, 1999). The latter include Polya Tree (PT) priors (Hanson *et al.*, 2005,

p. 255; Walker and Mallick, 1997, 1999) and consist of a set of binary tree partitions to allocate a case to its appropriate cluster value selected from a baseline prior G. Consider an unstructured error model for disease counts y_i (and expected cases E_i) for areas $i = 1, \ldots, N$

$$y_i \sim \text{Po}(E_i \mu_i)$$
$$\log(\mu_i) = \beta_0 + \phi e_i$$

and adopt an $N(0, 1)$ density as the baseline density (with distribution function G) for e_i with ϕ an extra unknown. The simplest PT would have one level only and select candidate values e_m^* from two possibilities. The choice would be between candidate values selected from the partition of the real line, either from $B_0 = (-\infty, G^{-1}(0.5))$, or from $B_1 = (G^{-1}(0.5), \infty)$. Thus the partitions of the parameter space at level 1 is based on the 50th percentile of G ensuring that the selected effects are centred (not confounded with the regression intercept). The next binary partition would involve subdivisions of B_0 and B_1 so that $(B_{00}, B_{01}, B_{10}, B_{11})$ are the breaks at level 2. The choice would then be between candidate values selected from the intervals $B_{00} = \{-\infty, G^{-1}(0.25)\}$, $B_{01} = \{G^{-1}(0.25), G^{-1}(0.5)\}$, $B_{10} = \{G^{-1}(0.5), G^{-1}(0.75)\}$ or $B_{11} = (G^{-1}(0.75), \infty)$.

The number of sets, namely ranges of bands from which candidate values (for parameter values or cluster random effects) are chosen, is thus 2^m at level m. Most applications have considered finite Polya partitions to level M (Hanson and Johnson, 2002, p. 1022). Candidate values in the lowest and uppermost bands are selected from truncated densities, with a form defined by G. For intervening bands j, they may be selected from a uniform density with $G^{-1}[(j-1)/2^m]$ and $G^{-1}(j/2^m)$ as the end points.

Walker and Mallick (1997, p. 849) liken the choice of an appropriate candidate value to a cascading particle. The choice between B_0 and B_1 is a Bernoulli choice governed by probabilities C_0 and $1 - C_0$. The probability C_0 may be selected from a prior beta density but Walker and Mallick (1997, p. 851–852) suggest $C_0 = 0.5$ on the basis that the first partition is centred at the median.

In general, if the option B_ε is selected at a particular step, then the particle moves to either $B_{\varepsilon 0}$ or $B_{\varepsilon 1}$ at the next step with respective probabilities $C_{\varepsilon 0}$ and $C_{\varepsilon 1} = 1 - C_{\varepsilon 0}$. These are random beta variables with

$$(C_{\varepsilon 0}, C_{\varepsilon 1}) \sim \text{Beta}(\alpha_{\varepsilon 0}, \alpha_{\varepsilon 1}).$$

The choice of values for $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$ should reflect prior beliefs about the underlying smoothness of F.

For m large, one would set $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1} = c_m$ in such a way that $F(B_{\varepsilon 0})$ and $F(B_{\varepsilon 1})$ are close. This may be done by setting

$$c_m = cm^d \qquad \text{for } c > 0, \quad d > 1, \tag{6.4}$$

so that c_m increases with m (in line with prior expectations that some degree of pooling should be appropriate, based on the smoothness). For example, $c_m = cm^2$ or $c_m = cm^3$ may be used with $c = 0.5$ or $c = 0.1$. Larger values of c mean that the posterior will resemble the baseline prior G more closely (Hanson *et al.*, 2005, p. 256). The DPP corresponds to $c_m = 1/2^m$. Taking

$$c_m = \gamma_1 m^{\gamma_2},$$

one may also set priors on the elements of the beta probabilities, with γ_2 perhaps restricted to small integer values.

The previous small area health example is in fact a mixed PT, analogous to the MDP model (Hanson and Johnson, 2002, p. 1022), since the centering density G is random by virtue of the parameter ϕ. In this example, suppose $M = 4$ is taken as the maximum number of levels. Taking $c_m = 0.5m^2$ and $\tau = 1/\phi$ would lead to the code

```
C <- 0.5; tau2 ~ dgamma(1,1); phi <- 1/sqrt(tau2)
for (m in 2:M) { c[m] <- C*pow(m,2)}
for (i in 1:N){  V[1,i] ~ dbern(0.5)
for (m in 2:M) {p[m,i] ~ dbeta(c[m],c[m])
                V[m,i] ~ dbern(p[m,i])}
# level 1 choice (convert V=0,1 to B=1,2)
        B[1,i] <- V[1,i]+1
# choices at level 2 and above
   for (m in 2:M) { B[m,i] <- sum(Vp[m,i,1:m-1])+V[m,i]+1
   for (j in 1:m-1) { Vp[m,i,j] <- V[m-j,i]*pow(2,j)}}
# select from ordinates of baseline density
  estar[i] <- G.inv[B[M,i]]; y[i] ~ dpois(mu[i]);
  log(mu[i]) <- log(E[i]) + beta0+phi*estar[i] }
```

The options for the baseline density ordinates would then be based on the selected prior G, for example with G an $N(0, 1)$ and $M = 4$, these would be the 6.25th, 12.5th, 18.75th, ..., 93.75th percentiles of G^{-1}.

Example 6.9 Seeds and extracts Walker and Mallick (1997) reanalyse the factorial layout data from Crowder (1978, Table 3). The original model of Crowder proposed variation of expected proportions within cell means

$$y_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij}) \qquad i = 1, \ldots, 4 \quad j = 1, \ldots, n_i$$

with π_{ij} then distributed according to four beta densities $\text{Be}(a_i, b_i)$. The index i corresponds to combinations of two binary factors, seed type (S) and extract type (E). Here the model is reformulated at the level of the $n = 21$ seeds, with $y_k \sim \text{Bin}(\pi_k, n_k), k = 1, \ldots, n$. Walker and Mallick propose a PT non-parametric prior for the overdispersion effects e_k under a logit transform of the π_k as in

$$\text{logit}(\pi_k) = \beta_1 + \beta_2 I(S_k = 2) + \beta_3 I(E_k = 2) + \beta_4 I(S_k = 2, E_k = 2) + e_k,$$

where the base density G for a PT prior with $M = 4$ levels is taken to be $N(0, 1)$. Beta weights are defined including unknowns

$$c_m = \gamma_1 m^{\gamma_2}$$

with priors $\gamma_1 \sim \text{Ga}(0.5,1)$ and $\gamma_2 \sim \text{Po}(2)$ assumed.

The estimated factorial effect parameters, from the second half of a run of 5000 iterations (two chains, convergent from 1000) are similar to those of Walker and Mallick. The means for γ_1 and γ_2 are 0.78 and 1.22, respectively. Only β_3 (the extract effect) is clearly different from

Table 6.2 Seeds and extracts data

Parameter	Mean	2.5%	97.5%
β_1	−0.62	−1.40	0.29
β_2	−0.08	−1.37	1.26
β_3	1.56	0.29	2.91
β_4	−0.91	−2.80	0.93
Germination rates			
π(Extracts = 1, Seeds = 1)	0.38	0.32	0.44
π(Extracts = 2, Seeds = 1)	0.69	0.62	0.75
π(Extracts = 1, Seeds = 2)	0.36	0.28	0.45
π(Extracts = 2, Seeds = 2)	0.49	0.40	0.58

zero (Table 6.2). The probabilities of germination according to levels of each factor are also shown (cf. Crowder, 1978, Table 4).

Example 6.10 Diabetic hospitalisations Diabetic complication rates may be taken as an indicator of the performance of the primary health sector in providing timely and appropriate care. In England, two indicators of diabetes care are regularly monitored, namely (a) the incidence of diabetic ketoacidosis and coma and (b) lower limb amputations. Here, observed and expected cases of both events (for males and females combined over two financial years, 2000–2001 and 2001–2002) are considered for 354 English local authorities. A Poisson regression with log link is assumed. The total of observed and expected cases is the same so the mean of the log response is zero and an intercept is not strictly necessary.

We first consider lower limb amputations alone and contrast a DPP with a PT approach, though the latter actually includes DPP under appropriate settings of c_m in (6.4). Under the DPP (model A), the data are taken as Poisson with

$$y_i|H_i \sim \text{Po}(E_i\nu_i),$$
$$\log(\nu_i) = \phi e_{H_i}$$

with E_i being expected events, and $H_i \sim$ Categorical(**p**), $\mathbf{p} = (p_1, \ldots, p_M)$ with $M = 30$ as the assumed maximum number of clusters and the p_j defined by a stick-breaking prior. The DPP includes a Ga(1,1) prior on α, consistent with an expected prior cluster total of $M^* = 5.9$. The baseline density G_0 is $N(0, 1)$, with ϕ^2 a variance parameter and $1/\phi^2$ is assigned a Ga(0.5,0.5) prior. The relative risks ν_i average 1 at least approximately (here the mean relative risk slightly exceeds 1) and indicate the quality of care; high values indicate lower quality care.

A two-chain run of 5000 iterations (1000 to convergence) is used to make posterior inferences. In particular, the estimated posterior relative risks of amputation over the 354 areas suggest some multimodality as well as outlying areas with very high rates (Figure 6.3). This would not have been so well represented by a unimodal parametric prior. The averages M^* and α are 18.5 and 4.1.

A PT procedure (model B) with 2^6 partitions (i.e. $M = 6$) is then applied with $c_m = cm^2$, with $c = 0.5$ and a $N(0, 1)$ baseline density. There are high correlations between the two sets of posterior risks (DPP vs PT priors) and in the area rankings. Nevertheless the plot of risks

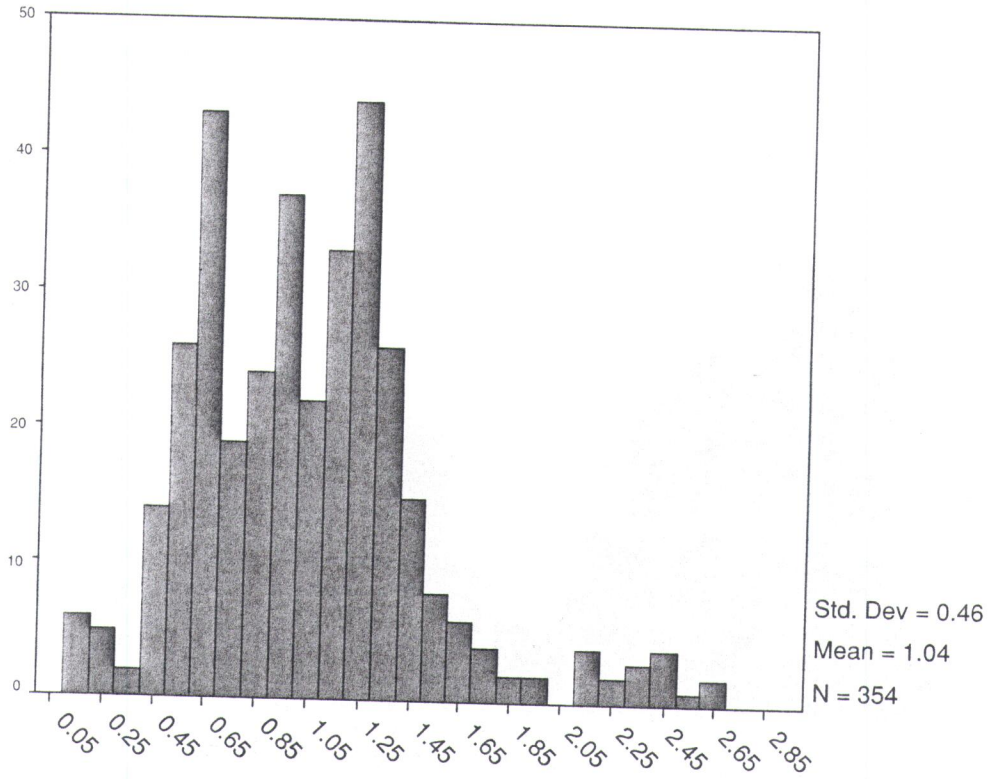

Figure 6.3 Posterior relative risks under DPP.

under the PT prior (Figure 6.4, based on iterations 1001–5000 in a two-chain run) shows less departure from unimodality. This may be an artefact of the restriction to preset parameters in c_m. Reducing c (e.g. to 0.1 or 0.01) leads to a more bimodal plot.

As a final illustration of a non-parametric application, consider deriving an overall index of diabetic care, with higher values indicating less effective care in terms of avoiding undesirable outcomes (a common factor model). Thus with y_1 denoting diabetic amputations and y_2 denoting diabetic ketoacidosis and coma consider the following common factor DP model

$$y_{1i} \sim \text{Po}(E_{1i}\nu_{1i}),\ y_{2i} \sim \text{Po}(E_{2i}\nu_{2i}),$$
$$\log(\nu_{1i}) = \phi_1 e_{H_i}$$
$$\log(\nu_{2i}) = \phi_2 e_{H_i},$$

where H_i are as under model A and the baseline density G_0 is again a standard normal density. The factor loading ϕ_1 is set to 1 for identifiability, while ϕ_2 is free and assigned a normal $N(1, 1)$ prior. The plot of the scores (Figure 6.5) shows some multimodality with three outlying areas (285, 289, 148) having scores approaching 0.5, while a central cluster of areas (109 from 354) have scores between 0 and 0.10.

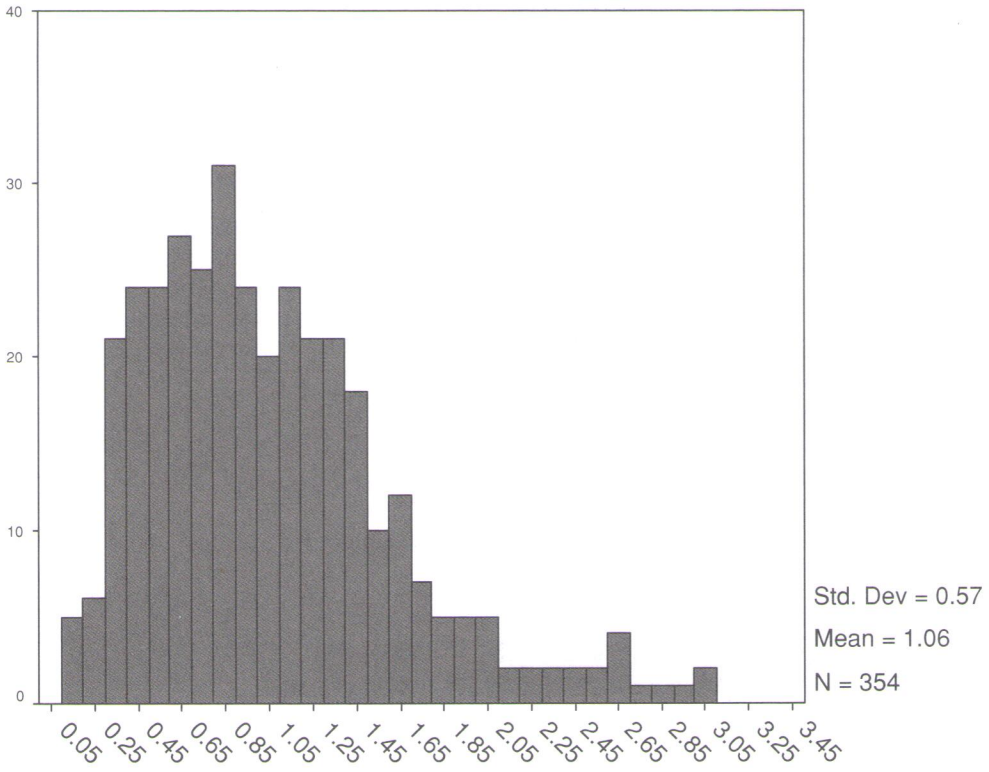

Figure 6.4 Posterior relative risks under PT prior.

EXERCISES

1. In Example 6.1 use a likelihood calculation and derive the posterior mean of the likelihood and deviance. Use the AIC and BIC criteria to compare solutions $C = 1, 2, 3, 4$.
2. In Example 6.1 obtain the posterior probabilities under $C = 3$ that individual cases belong to different groups. These are averages over iterations of indicator variables.
3. In Example 6.2, extend the comparisons to $C = 4$.
4. For the data of Example 6.2 apply the splitting prior of (6.2) for the cases $C = 2$ and $C = 3$.
5. In Example 6.3, code the basic ZIP model using the individual data approach (as per Model B in Example 6.3). Sample new data (predictions y_{new}) and derive the EPDs for the basic ZIP model and the three group ZIP model as already described in Example 6.3. The BICs for both models can also be obtained since the number of parameters is known.
6. In Example 6.5 (DMFT response), extend the model to allow the ω_j to be specific to school ($j = 1, \ldots, J, J = 6$).

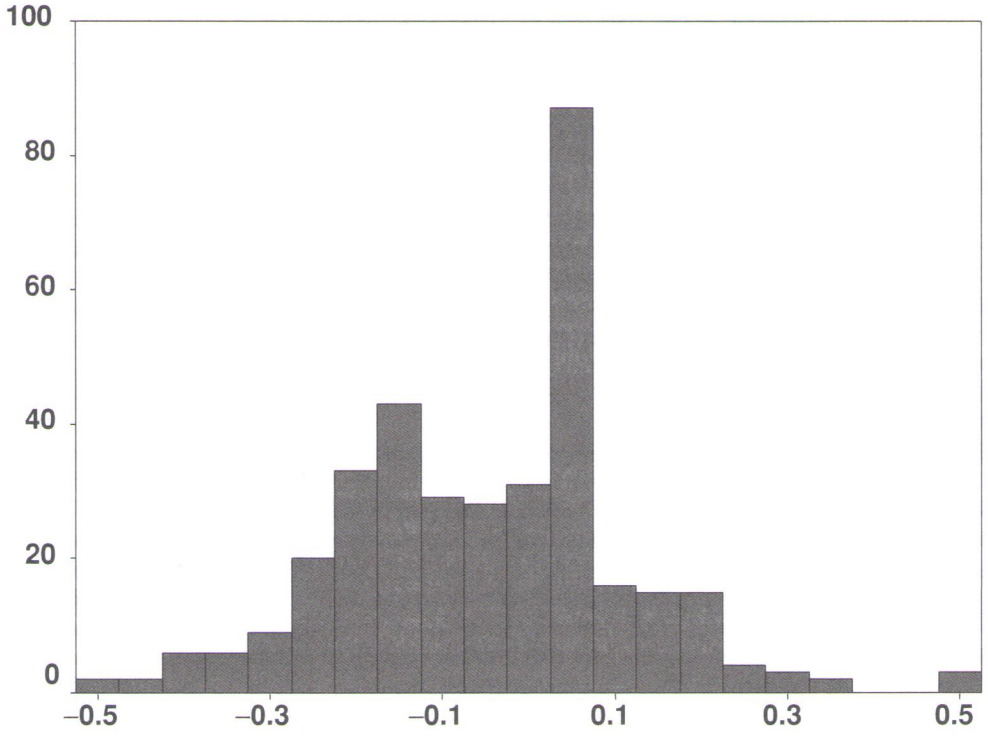

Figure 6.5 Diabetic care index.

7. In Example 6.7 (DP analysis of eye-tracking anomalies) try monitoring the θ_i to obtain posterior means ($\overline{\theta}_i$ for each of the 101 subjects and so obtain the DIC using the definition $\overline{D} - D(\overline{\theta})$. In BUGS this will also require including code to obtain the deviance at each iteration. Assume the hyperparameters of the gamma-mixing density are free. Does adopting a DPP (with α a free parameter) improve over the standard Poisson-gamma mixture (Chapter 5)? Is this conclusion affected by setting α at particular values, e.g. $\alpha = 1$ and $\alpha = 5$ rather than letting it be a free parameter?

8. In Example 6.7 (DPP analysis of eye-tracking anomalies) try a Ga(0.01,0.01) prior on the concentration parameter α. Does this affect the posterior mean for clusters?

9. In Example 6.8 (galaxy clusters) consider the ratio of posterior mean of M^* to its prior mean, as defined by the prior on the DPP concentration parameter α. What is the impact on this ratio of increasing M (the maximum clusters under a truncated DPP) to 20 and what is the impact of combining $M = 20$ with a $G(3, 1)$ prior on α, consistent with a prior mean $M^* = 10$?

10. Use the data from Gelfand *et al.* (1990) relating to growth for $n = 30$ rats at five ages and add a DPP as in West *et al.* (1994, p. 373) and Escobar and West (1998, p. 16). See also the birats example on the WINBUGS site. Thus the bivariate normal model for varying

Example 8.12 Fetal lamb movements An example of the HMM is provide by a time series of lamb fetal movement counts y_t from Leroux and Puterman (1992), where the presence in the mixture of more than one component leads to Poisson overdispersion. Suppose a two-class Markov mixture is applied, with shifts between two Poisson means determined by a Markov chain (i.e. $m = 2$). Dirichlet priors for the elements in each row are assumed, namely

$$p_{i,1:m} \sim \text{Dir}(1, 1, \ldots, 1),$$

although a beta prior can also be used for $m = 2$. The same prior is used for the multinomial vector governing the choice of initial state. For the two Poisson means Ga(1, 1) priors are stipulated, with an identifiability constraint that one is larger – an initial unconstrained run justified such a constraint, showing the means to be widely separated.

With this model, a two-chain run of 5000 iterations (1000 burn-in) shows the state occupied most of the periods (about 220 from 240) to have a low average fetal movement rate (around 0.23), and a minority state with a much higher rate, around 2.2–2.3. The majority state has a high retention rate (reflected in the transition parameter p_{22} around 0.96) while movement out of the minority state is much more frequent.

The actual number of movements y_t is predicted closely, though Leroux and Puterman show that using $m = 3$ components leads to even more accurate prediction of actual counts. The model with $m = 2$ shows relatively small CPOs for the movements at times 85 and 193 (counts of 7 and 4 respectively).

For comparison, and since the outcome is a count, model B consists of an INAR1-type model for the conditional mean. The 'innovation' process is governed by Bernoulli switching between means λ_1 and λ_2 (with $\lambda_2 > \lambda_1$ to guarantee identifiability). Thus

$$\begin{aligned} y_t &\sim \text{Po}(\mu_t), \\ \mu_t &= \pi \circ y_{t-1} + \lambda_1 \delta_t + \lambda_2(1 - \delta_t) \qquad t > 1, \end{aligned}$$

with the first observation having mean

$$\mu_1 = \lambda_1 \delta_1 + \lambda_2(1 - \delta_1).$$

The switching indicators have prior $\delta_t \sim \text{Bern}(\eta)$ with η itself assigned a beta prior. This model also identifies a subpopulation of periods with a much higher movement rate, around 4.5, than the main set of periods. It has a very similar marginal likelihood to the two-state Markov switching model (-180 vs -179).

8.11 OTHER NONLINEAR MODELS

Some of the above models are often characterised as nonlinear, such as the threshold autoregressive approaches. Here some other nonlinear methods are mentioned that bring greater flexibility in modelling certain time series features (e.g. changing volatility, discontinuities in level) but possibly at the cost of computing complexity or heavy parameterisation (Koop and Potter, 1999, p. 260). For instance, for large datasets a flexible but highly parameterised generalisation of the stochastic unit root model is the time-varying autoregression (TVAR)

model (Godsill *et al.*, 2004, p. 160), with

$$y_t = \rho_{1t} y_{t-1} + \rho_{2t} y_{t-2} + \cdots + \rho_{pt} y_{t-p} + \varepsilon_t,$$

where $\varepsilon_t \sim N(0, \sigma^2)$, and where each of the ρ_{kt} follow random walk prior or autoregressive priors, e.g. $\rho_{kt} \sim N(\alpha_k \rho_{k,t-1}, \omega_k^2)$. If the ρ coefficients are to be stationary then RW or AR priors are applied to partial correlation coefficients with transformation back to ρ coefficients as discussed in Section 8.2. An extension allows σ^2 to vary over time also (Godsill *et al.*, 2004, p. 161).

Discrete mixture nonlinear models also seek to represent time series discontinuities. Wong and Li (2000) mention mixture autoregressions with K components differing in lag order p_k and with prior probabilities π_k, so that

$$P(y_t | D_{t-1}) = \sum_{k=1}^{K} \pi_k \Phi\left(\frac{y_t - \rho_{ok} - \rho_{1k} y_{t-1} - \cdots - \rho_{p_k k} y_{t-p_k}}{\sigma_k} \right),$$

where D_{t-1} is all data up to $t-1$. They denote these as MAR(K, p_1, p_2, ..., p_K) models and discuss their ability to represent changing conditional variances. Mueller *et al.* (1997) describe a discrete mixture model for nonlinear AR models that is similar to a TVAR model. For example, suppose there are K possible AR1 models, each with their own intercept and lag coefficient on y_{t-1} and each with their own variance. If $G_t \sim$ Categorical($q_{t,1:K}$) and $G_t = k$, then

$$y_t | G_t \sim N(\rho_{0k} + \rho_{1k} y_{t-1}, 1/\tau_k).$$

The category selector G_t is obtained using a time-varying Gaussian kernel prior, with

$$\Pr(G_t = k) = q_{tk} \propto \exp(-0.5(y_{t-1} - \mu_k)^2 / V),$$

with V an additional variance parameter. Parameters $\theta_k = \{\rho_{0k}, \rho_{1k}, \tau_k\}$ are selected from candidate values $\theta_k^* = \{\beta_{0k}^*, \beta_{1k}^*, \tau_k^*\}$ using a Dirichlet process (DP) prior with concentration parameter κ, thus allowing for greater robustness when there are jumps in series or multimodality. A particular application is to harmonic process models (West, 1995) whereby periods $\lambda_k = 2\pi/[\text{acos}(0.5\rho_k)]$ are estimated from the model

$$y_t | G_t = k \sim N(\rho_k y_{t-1} - y_{t-2}, 1/\tau_k).$$

For stationarity, the constraint $|\rho_k| < 2$ applies. The kernel prior is now multivariate with

$$q_{tk} \propto \exp(-0.5(x_t - \mu_k)' V^{-1} (x_t - \mu_k)),$$

where $x_t = (y_{t-1}, y_{t-2})$ and $\mu_k = (\mu_{1k}, \mu_{2k})$, and V is a covariance matrix.

Example 8.13 Lynx data, AR mixtures Wong and Li (2000) consider the well-known lynx data ($T = 114$) and detect a two-component mixture ($K = 2$) with lags in y at $t-1$ and $t-2$ in each component, namely a MAR(2, 2, 2) model. The analysis conditions on the first two data points. A two-component model is applied here with constraints on $\tau_k = 1/\sigma_k^2$ for identifiability. Another possibility might be a constraint on π_k. The lag parameters $\{\rho_{0k}, \rho_{1k}, \rho_{2k}\}$, $k = 1, \ldots, K$, are assigned $N(0, 1)$ priors.

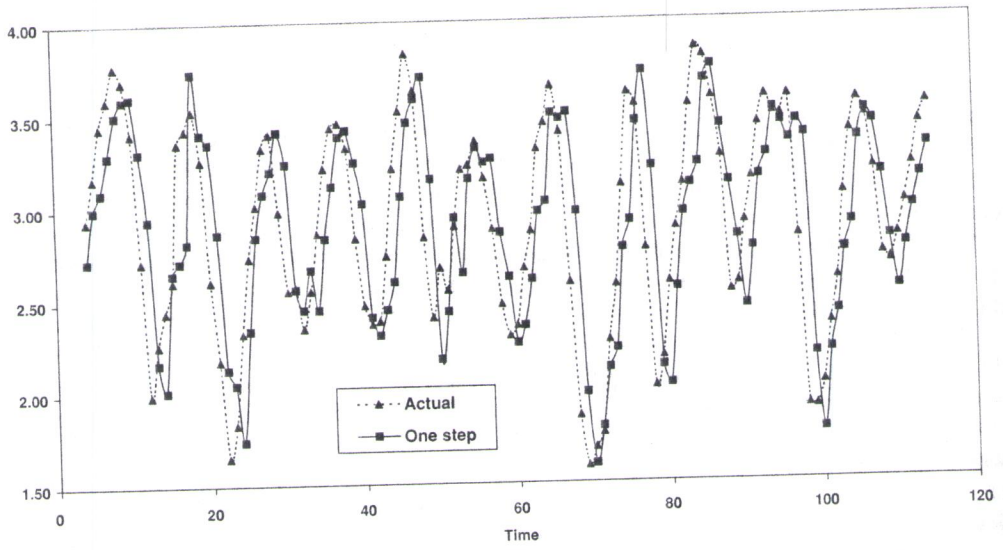

Figure 8.7 One-step predictions (log10 lynx trappings) under discrete mixture AR.

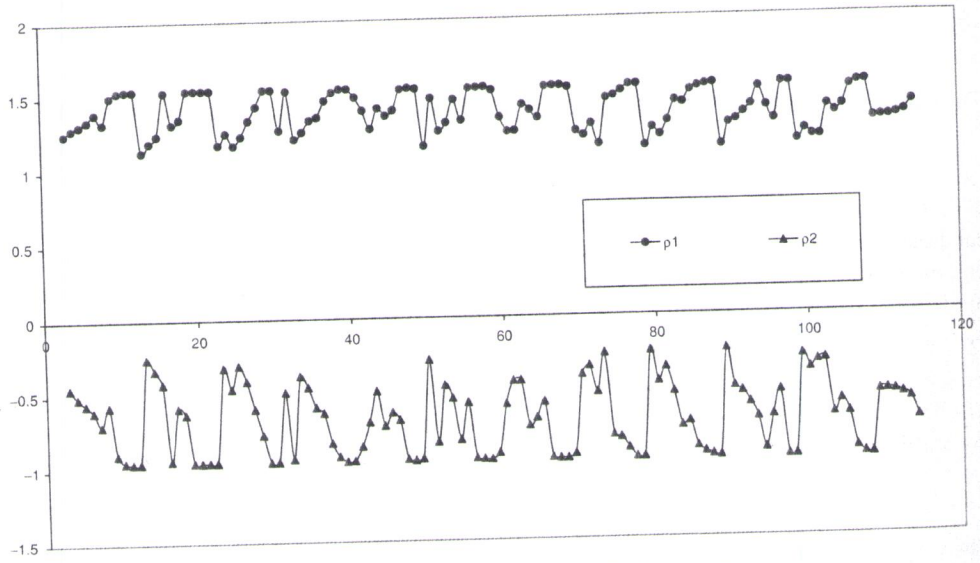

Figure 8.8 Varying first- and second-order lag coefficients

The second half of a two-chain run of 10 000 iterations shows a smaller component ($\pi_1 = 0.30$) with $\sigma_1 = (1/\tau_1)^{0.5} = 0.09$. Means (sd) for the lag parameters are $\rho_{01} = 0.72(0.26)$, $\rho_{11} = 1.07$ (0.16) and $\rho_{21} = -0.27(0.15)$. For the larger component these parameters have means (sd) of 1.01 (0.16), 1.49 (0.10) and −0.86 (0.10), respectively. One-step predictions are made and have an MSE of 0.228 (see Figure 8.7), while the concurrent

predictive mean error sum of squares is 0.073. (This is the sum of squared differences between y_t and $y_{\text{new},t}$ divided by 112).

To apply a DP stage on the possible parameters in the components of the AR2 model, the maximum number of possible components is set at $M = 5$. A Ga(0.1, 0.1) prior is assumed for the Dirichlet concentration parameter with small values excluded. It is assumed that V is a correlation matrix with off-diagonal element ρ, while the μ_k are uniform within the minimum and maximum of the observed data. To obtain the posterior density of the realised number of components K, one can monitor a selected parameter and then via postprocessing obtain the number of distinct values obtained at each iteration.

A two-chain run of 10 000 iterations (with the second half for inferences) shows a posterior mean for κ of 0.79, with a mean (95% interval) for ρ of $-0.22(-0.80, 0.54)$. The predictive MSE is 0.223, a slight improvement over the standard discrete mixture, with mean ESS of 0.070. Figure 8.8 plots the 112 posterior means of ρ_{1t} and ρ_{2t} (time-varying lags on y_{t-1} and y_{t-2}) over times $t = 3, \ldots, T$ obtained by monitoring the category G_t selected at each iteration for time t.

EXERCISES

1. In Example 8.2 (Real GNP series) apply the stochastic unit root model $y_t = \rho_t y_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim N(0, \sigma^2)$, and $\exp(\alpha_t) = \rho_t$. With $p = 1$ and $p = 2$ in the AR model for the α_t series, assess the probability μ_α is below 0.

2. In Example 8.3 (the trapped lynx series), try using priors on the AR and MA coefficients based on the maximum likelihood solution but with the precision downweighted by 10. The maximum likelihood estimates from SPSS are

	Mean	s.e.
ρ_1	2.07	0.126
ρ_2	−1.77	0.200
ρ_3	0.49	0.123
θ_1	0.90	0.121
θ_2	−0.09	0.141
θ_3	−0.49	0.100
ν	2.90	0.064

   Also consider estimation with the priors as in the worked example but conditioning on the first three data points. Finally consider the model as in the worked example, including modelling of latent pre-series values, but introduce an error outlier mechanism such that with probability 0.05, some ε_t have variance 10 times σ_ε^2. How do these options affect parameter estimates and one-step-ahead predictions?

3. In Example 8.5 (Consumption and income), try including binary predictor selection indicators in the VAR4 model (e.g. in an SSVS prior) and compare inferences on lag effects to a model without any form of predictor selection.

multiplicative frailty for multivariate nested data with the hazard for patient i, hospital j and outcome k. A typical model for this type of data might be

$$h_{jk}(t_{ijk}|X_{ij}, \zeta_j, w_{ik}) = h_{0jk}(t_{ijk}|\theta_{jk}) \exp(X_{ij}\beta_{jk}) w_{ik}\zeta_j,$$

where the β_j model hospital effects on each outcome, ζ_j are gamma hospital frailties and w_{ik} are patient frailties specific to outcomes.

A semiparametric form of the AFT model provides opportunities for modelling frailty. Consider the AFT model

$$t_i = \exp(X_i\beta)V_i,$$

or in the log scale

$$\log(t_i) = X_i\beta + \varepsilon_i.$$

Instead of standard assumptions regarding V or ε, one may model their density non-parametrically, for example via a Dirichlet process or Polya tree prior (Walker and Mallick, 1999). This amounts to semiparametric intercept variation. A discrete mixture with known small number of groups is also possible, with a two-group mixture representing high- and low-frailty subjects.

Example 13.8 Veterans lung cancer survival To allow for heterogeneity in survival in the data from Example 13.1, a discrete mixture of parametric hazards (with known number of components) is one possible approach. This allows ready extension to include mixing on the hazard and regression parameters, as well as just the level, whereas a continuous mixture is most flexible for intercept variation only. Here only the intercept (i.e. the overall level of frailty) is allowed to vary between groups, and a two-group mixture is adopted. Extension to varying Weibull slopes is left as an exercise. A Dirichlet prior on the mixing proportions π_1 and π_2 is used with equal prior weights of 1 on each group.

The last 9000 of a two-chain run of 10 000 iterations leads to estimates of $\pi_1 = 0.26$ and $\pi_2 = 0.74$ with means $\beta_{01} = -6.39$ and $\beta_{02} = -4.36$ (Table 13.5). So a small low-mortality group is distinguished. The Weibull parameter becomes more clearly above 1, with an average of 1.51 and 95% interval from 1.12 to 1.76. Among the covariate effects, the impact of the Karnofsky score in particular is enhanced.

Example 13.9 Small cell lung cancer Ying *et al.* (1995) consider survival times for 121 small cell lung cancer patients involving a cross-over trial for two drugs (etoposide E and cisplatin C); 62 patients are randomised to arm A (C followed by E), whereas arm B has E followed by C. Apart from treatment ($X_1 = 1$ for arm B patients, $X_1 = 0$ for arm A), patient age at entry to trial (X_2) is another predictor – which a Cox regression suggests significantly enhances mortality (i.e. that age is negatively related to survival time). A Cox regression also shows a negative effect of arm B on survival.

As a baseline for these data, a logistic model is here adopted for natural logs of the survival times (Ying *et al.* consider log10 transformed times), in line with an AFT log-logistic survival

Table 13.5 Veterans cancer data, parameter estimates

	Mean	St devn	2.5%	97.5%
Single group model				
Constant	−4.26	0.55	−5.35	−3.13
Karnofsky score	−0.26	0.06	−0.37	−0.14
Prior therapy (PT)	1.95	0.65	0.65	3.21
Small cell type	0.72	0.25	0.23	1.20
Adeno cell type	1.16	0.29	0.58	1.73
Large cell type	0.30	0.27	−0.24	0.81
PT × Karnofsky	−0.32	0.11	−0.53	−0.10
α (Weibull parameter)	1.11	0.07	0.96	1.25
Two group model				
Probability (group 1)	0.26	0.12	0.10	0.59
Probability (group 2)	0.74	0.12	0.41	0.90
Constant (group 1)	−6.39	1.01	−8.16	−4.07
Constant (group 2)	−4.36	0.77	−5.90	−2.82
Karnofsky score	−0.48	0.08	−0.65	−0.31
Prior therapy (PT)	1.97	0.61	0.83	3.22
Small cell type	0.86	0.35	0.10	1.50
Adeno cell type	1.05	0.42	0.19	1.84
Large cell type	0.14	0.38	−0.64	0.87
PT × Karnofsky	−0.32	0.10	−0.53	−0.14
α (Weibull parameter)	1.51	0.12	1.29	1.76

mechanism (Collett, 1994). So

$$\log(t_i) \sim L(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}, 1/\kappa) I(t_i^*,),$$

with priors $\gamma_j \sim N(0, 1000)$, and $\kappa \sim \text{Ga}(1, 0.001)$, and where t^* represents times at censoring, or 0 when failure times are observed. The median survival formulae are monitored for patients aged 62 (cf. Ying *et al.* who find a median survival time of 603 days in arm A for patients of this age). Iterations 1001–10 000 of a two-chain run show posterior means on $\{\gamma_0, \gamma_1, \gamma_2\}$ of $\{7.47, -0.42, -0.015\}$ with the 95% intervals for treatment and age being $(-0.71, -0.14)$ and $(-0.033, 0.001)$ respectively. So age is strictly not significant in diminishing survival times, but assignment to arm B does significantly reduce survival time. The median survival times under arms A and B (for patients aged 62) are estimated as 686 and 450 days.

A non-parametric frailty effect is first introduced in the form of a two-group discrete mixture for γ_0. A monotonicity constraint $\gamma_{02} > \gamma_{01}$ is used for identification, with the increment $\delta = \gamma_{02} - \gamma_{01}$ assumed to be $N(0, 1)$. A Dirichlet prior on the probabilities π_k of each intercept is assumed, with prior weight of 1 on each probability. The average intercept (required for obtaining median survival times) is estimated at each iteration as $\gamma_0 = \pi_1\gamma_{01} + \pi_2\gamma_{02}$. Age and treatment effects are similar to the first model, with posterior means -0.014 $(-0.030, 0.0006)$ and -0.41 $(-0.67, -0.16)$. The estimated median survival times are, however, increased to 476

(arm B) and 721 (arm A). This method detects a minority population with extended survival ($\pi_2 = 0.26$ and $\gamma_{02} = 8.38$).

A third model draws on the principles of the analysis of these data by Walker and Mallick (1999), who use a Polya tree prior on the errors ε in

$$\log(t_i) = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \varepsilon_i.$$

Here a Dirichlet process prior (DPP) is adopted on varying intercepts rather than the errors directly, with

$$\log(t_i) \sim L(\gamma_{0L_i} + \gamma_1 x_{1i} + \gamma_2 x_{2i}, 1/\kappa) I(t_i^*,),$$
$$L_i \sim \text{Categorical}(p_1, p_2, \ldots, p_M),$$

where $M = 20$, and $p = (p_1, p_2, \ldots, p_M)$ is generated using a stick-breaking prior. With $r_1, r_2, \ldots, r_{M-1}$ being Beta(1, λ) random variables (and $r_M = 1$), this involves setting $p_1 = r_1$, $p_2 = r_2(1 - r_1)$, $p_3 = r_3(1 - r_2)(1 - r_1)$, $\ldots$, $p_M = r_M(1 - r_{M-1})(1 - r_{M-2}) \ldots (1 - r_1)$. λ is assigned a Ga(5, 1) prior but sensitivity analysis to assuming different preset λ values, or other priors on λ can be adopted. The baseline density for the intercepts is

$$\gamma_{0j} \sim N(\mu_g, 1/\tau_g),\ j = 1, \ldots, M,$$

where $\mu_g \sim N(7, 1)$ and $\tau_g \sim \text{Ga}(1, 1)$. The relatively informative prior for μ_g is based on the earlier standard parametric analysis.

The resulting plot of the posterior means of the intercepts, based on iterations 1000–20 000 of a two-chain run, suggests positive skew or even bimodality, namely, some individuals with unusually high survival chances (Figure 13.3). The median number of clusters is 15. The median survival times for the two arms are estimated as 489 (arm B) and 723 (arm A), very close to the estimates under the simpler two-group discrete mixture model.

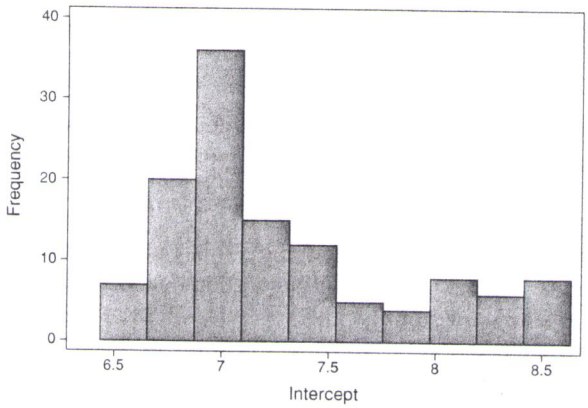

Figure 13.3 Histogram of varying intercepts (DPP model).

Another possibility for a non-parametric approach (analysis left to the reader) involves a DPP on multiplicative factors to produce varying scale (non-parametric scale mixing) with

$$\log(t_i) \sim L(\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}, 1/[\kappa \nu_i])I(t_i^*,),$$
$$\nu_i = \eta[L_i],$$
$$L_i \sim \text{Categorical}(p_1, p_2, \ldots, p_M).$$

The baseline density for the scale-mixing parameters is

$$\eta_j \sim \text{Ga}(\phi, \phi), \qquad j = 1, \ldots, M,$$

where $\phi \sim E(1)$. This approach may be relevant in the case outlier points were suspected.

13.8 DISCRETE TIME SURVIVAL MODELS

Even when events occur in continuous time, many event histories actually record only the nearest month or year (e.g. marital or job histories). Adopting a continuous time analysis in the presence of many tied failure times would give inconsistent estimates (Prentice and Gloeckler, 1978). Sometimes durations may be grouped by definition – for example the number of menstrual cycles to conception after marriage, or number of school years before removal (Muthen and Masyn, 2005).

Suppose the time scale is partitioned into J intervals $(a_{j-1}, a_j]$, $j = 1, \ldots, J$, not necessarily of equal length, with $a_0 = 0$, and a_J equalling the maximum observed time, censored or failure. Censoring (an individual exits in an interval without failure being recorded, e.g. due to dropout) is assumed to occur at the end of intervals. The observed survival times T_i define a discrete value j in the range $\{1, \ldots, J\}$ if $a_{j-1} \leq T < a_j$ (written as $T_i = j$), with failure occurring in the jth interval if $a_{j-1} \leq T_i < a_j$ and $\delta_i = 1$. The actual location of the failure during the interval is usually not known.

Conditional on time constant and time-varying predictors, X_i and Z_{ij} respectively, the discrete hazard of failure in interval j given survival till then is the conditional probability

$$h(T_i = j|X_i, Z_{ij}) = \Pr(T = j|T \geq j, X_i, Z_{ij}) = F(\alpha_j + X_i\beta_j + Z_{ij}\gamma_j) \tag{13.8}$$

where F is a distribution function. A common approach to modelling this probability (Kalbfleisch and Prentice, 1980) assumes an EV distribution function

$$F(\eta) = 1 - \exp\{-\exp(\eta)\},$$

leading to a complementary log–log link for h. This can be obtained from assuming an underlying continuous survival process and proportional hazard effects. Another possibility (Thompson, 1977) is a logit link for h, with

$$F(\eta) = \exp(\eta)/[1 + \exp(\eta)]. \tag{13.9}$$

The impact of time can be modelled flexibly within the regression term η, for example via a random walk (Fahrmier, 1994), via a polynomial function (Efron, 1988) or via any time series prior, for example a hidden Markov chain (Kozumi, 2000). If a distinct intercept or regression

Table 15.5 Klein model I structural parameter estimates

Parameter	Mean	St. devn	2.5%	97.5%
β_1	15.90	2.03	11.92	19.85
β_2	−0.11	0.19	−0.49	0.26
β_3	0.81	0.05	0.71	0.90
β_4	0.39	0.18	0.06	0.74
β_5	24.75	5.18	15.76	34.94
β_6	−0.06	0.09	−0.24	0.13
β_7	0.87	0.10	0.65	1.05
β_8	−0.18	0.03	−0.23	−0.14
β_9	−2.18	1.95	−6.16	1.56
β_{10}	0.40	0.06	0.28	0.51
β_{11}	0.24	0.06	0.12	0.35
β_{12}	0.09	0.04	0.01	0.18

Data on C, Y and I for the United States for 1955–1986 are presented by Griffiths *et al.* (1993, p. 592), and are in billion dollars (divided by 1000 for numerical convenience).

Here we regress C_t on $P_{Z_t} X_t$ where $P_{Z_t} = Z_t(Z_t'Z_t)^{-1}Z_t'$ is the projection matrix for $Z_t = (1, I_t)$, and $X_t = (1, Y_t)$ (Bound *et al.*, 1995). The analysis seeks to estimate the investment multiplier $\lambda = 1/(1-\beta)$ as well as the coefficients themselves. A beta prior is used for β reflecting economic expectations. Using the last 9000 of a two-chain run of 10 000 iterations, the posterior mean and median for β (namely 0.876 and 0.882) are similar to those cited by Griffiths *et al.* A point estimate of λ could use either the mean or the median of β, giving multipliers of around 8.3. However, allowing for the uncertainty in β (especially in its upper range) implies a highly skewed density for λ, with mean of 28.7 as against a median of 8.44.

15.5 ENDOGENOUS REGRESSION INVOLVING DISCRETE VARIABLES

For simultaneous and recursive models involving discrete variables, those that have received most attention, including Bayesian treatments, are simultaneous probit and tobit models (e.g. Chib, 2003; Chib and Hamilton, 2002; Li, 1998; Li and Poirier, 2003; Smith *et al.*, 2004). Bayesian estimation improves on two-stage procedures for estimating the simultaneous probit (e.g. Alvarez and Glasgow, 1999; Keshk, 2003) or full information maximum likelihood methods (Stratmann, 1992). Both Li (1998) and Smith *et al.* (2004) focus on a triangular two-equation system, which for both y_1 and y_2 binary is

$$y_{i1}^* = \gamma y_{i2} + X_{i1}\beta_1 + u_{i1},$$
$$y_{i2}^* = X_{i2}\beta_2 + u_{i2},$$

with $y_{1i} = 1$ if $y_{i1}^* > 0$, and $y_{i1} = 0$ otherwise, and similarly for y_{i2}^*. Li (1998) considers the tobit–probit case where $y_{i1} = y_{i1}^*$ if $y_{i1}^* > 0$, and $y_{i1} = 0$ otherwise. With augmentation in this way (Albert and Chib, 1993; Chib, 1992), the system is equivalent to the metric data triangular recursive system of Zellner (1971, p. 252). The bivariate normal for (u_{i1}, u_{i2}) has dispersion

matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}.$$

Li decomposes the joint density as $(u_{i1}|u_{i2})(u_{i2})$, so that

$$y_{i1}^* = \gamma y_{i2} + X_{i1}\beta_1 + \sigma_{12}(y_{i2}^* - X_{i2}\beta_2) + e_i,$$
$$y_{i2}^* = X_{i2}\beta_2 + u_{2i},$$

where u_{i2} is $N(0, 1)$, and $e_i \sim N(0, \sigma_{11} - \sigma_{12}^2)$. Simultaneous logit and simultaneous multinomial models have also been proposed (Schmidt and Strauss, 1975), while Berkhout and Plug (2004) consider a recursive model for Poisson data.

A specific type of recursive model occurs in what are termed *endogenous treatment models*. These involve assessing the causal effect of a categorical treatment or exposure variable (usually binary) on a metric or discrete response such as a health behaviour that it is sought to modify. The treatment variable is non-randomly assigned but subject to selection bias, and is therefore endogenous with the response. This is typically the case in observational situations (rather than experimental trials) where treatment is to some degree self-selected, and may be correlated with unobserved patient factors (e.g. compliance, susceptibility to health messages) that also affect the main response. Although called endogenous treatment models, one may include a variety of analogous applications, examples being wage returns to union membership (the 'treatment') as in Chib and Hamilton (2002), and health utilisation according to whether privately insured (Munkin and Trivedi, 2003).

As an example, let y_i be a count of adverse health behaviours (number of alcoholic drinks in previous week), let $T_i = 1$ (or 0) for participation (non-participation) in a treatment, where 'treatment' might include medical advice to change behaviours, and let X_i and W_i be observed influences on the health behaviour itself and on the allocation to treatment. Then $Y_i \sim \text{Po}(\mu_i)$,

$$\log(\mu_i) = X_i\beta + \delta T_i + u_{i1}, \tag{15.8.1}$$

where u_{i1} represents unobserved influences on the health response. For the treatment allocation, an augmented data model is assumed, based on the equivalence $\Pr(T_i = 1) = \Pr(T_i^* > 0)$, namely

$$T_i^* = W_i\gamma + u_{i2}, \tag{15.8.2}$$

where u_{i2} represents unobserved influences on treatment allocation. The correlation between treatment and response is modelled via a bivariate normal or some other bivariate model for $u_i = (u_{i1}, u_{i2})$. Kozumi (2002) considers bivariate Student t models for u_i involving normal scale mixing with gamma-distributed scaling factors, $\lambda_i \sim \text{Ga}(\nu/2, \nu/2)$, while Jochmann (2003) and Chib and Hamilton (2002) sample the λ_i semiparametrically using a Dirichlet process prior. With multivariate normal errors,

$$(u_{i1}, u_{i2}) \sim N(0, \Sigma_u), \tag{15.9.1}$$

where

$$\Sigma_u = \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix}, \tag{15.9.2}$$

with the variance of u_{i2} set to 1 for identifiability. This model may also be expressed with (15.8.1) as

$$\log(\mu_i) = X_i\beta + \delta T_i + \sigma u_{i1},$$

with $(u_{i1}, u_{i2}) \sim N(0, R_u)$, where R_u is a correlation matrix.

A 'common factor' model is also possible, and again assuming a count response with mean μ_i,

$$\log(\mu_i) = X_i\beta + \delta T_i + \lambda\zeta_i,$$
$$T_i^* = W_i\gamma + \zeta_i + u_i,$$

where $\zeta_i \sim N(0, \phi)$ and $u_i \sim N(0, 1)$, with ϕ a free parameter, and λ interpreted as a factor loading.

Jochmann (2003) and Chib and Hamilton (2002) demonstrate the switching regime version of the endogenous treatment model whereby each subject has a partially latent bivariate observation $\{y_{i0}, y_{i1}\}$, one observed, the other missing according to their observed T_i. If T_i is 1 then $y_{i1} = y_i$ and y_{i0} is missing, while if T_i is 0, then $y_{i0} = y_i$ and y_{i1} is missing. Then for y_i metric and normality assumed

$$y_{i0} = X_i\beta_0 + u_{i0},$$
$$y_{i1} = X_i\beta_1 + u_{i1},$$
$$T_i^* = W_i\gamma + u_{i2},$$

where

$$(u_{i0}, u_{i1}, u_{i2}) \sim N\left(0, \begin{pmatrix} \sigma_0^2 & 0 & \sigma_0\rho_{02} \\ 0 & \sigma_1^2 & \sigma_1\rho_{12} \\ \sigma_0\rho_{02} & \sigma_1\rho_{12} & 1 \end{pmatrix}\right).$$

The difference $y_{i1} - y_{i0}$ is taken as a measure of the impact of the treatment. Recently, Chib (2004) shows how this model can be analysed without involving the joint distribution of the y_{i0} and y_{i1}. This simplifies the model analysis considerably.

Rossi *et al.* (2005) and Manchanda *et al.* (2004) consider a shared factor model for two related longitudinal count responses, with a direct effect of one response on the other also present. The responses are sales y_{it} of prescription drugs to physician i at period t, and 'detailing' totals D_{it} (i.e. numbers of sales calls) made to the same physicians. Physicians vary in their overall prescribing rates and in responsiveness to sales promotion, so with $Y_{it} \sim \text{Po}(\mu_{it})$, one may specify

$$\log(\mu_{it}) = \beta_{i1} + \beta_{i2}D_{it} + \beta_{i3}\log(y_{i,t-1} + d),$$

where $d = 1$, β_{i1} denotes variation in prescribing regardless of detailing levels, β_{i2} measures physician responsiveness to sales promotion and β_{i3} denotes varying lag effects. The random physician effects are possibly related to observed physician attributes W_i (e.g. type of

Table 15.6 Endogenous treatment model, posterior summary

	Mean	2.5%	97.5%
Σ_{11}	4.45	3.89	5.08
Σ_{12}	1.65	1.40	1.92
δ	−2.04	−2.47	−1.62
β_0	2.24	2.06	2.43
β_1	−0.25	−0.43	−0.07
β_2	0.05	−0.21	0.32
γ_0	−0.59	−0.72	−0.43
γ_1	−0.22	−0.33	−0.11
γ_2	0.32	0.17	0.47
γ_3	−0.21	−0.32	−0.09
γ_4	0.22	0.12	0.32
γ_5	0.28	0.16	0.40

physician), so

$$(\beta_{i1}, \beta_{i2}, \beta_{i3}) \sim N_3(W_i\Delta, \Sigma_\beta).$$

Moreover, detailing efforts (e.g. allocations of sales staff or other marketing promotion directed to different physicians) are related to latent physician effects, via a model such as $D_{it} \sim \text{Po}(\lambda_i)$ where

$$\log(\lambda_i) = \gamma_0 + \gamma_1\beta_{i1} + \gamma_2\beta_{i2}.$$

For example, $\gamma_2 < 0$ would mean that less responsive physicians are detailed at higher levels.

Example 15.9 Drinking and physician advice Kenkel and Terza (2001) consider observational data for 2467 hypertensive subjects relating to a count y_i of alcoholic beverages consumed in past fortnight, and physician advice on the medical risks of excess alcohol use (T, binary). The model is as in (15.8)–(15.9),

$$\log(\mu_i) = X_i\beta + \delta T_i + u_{i1},$$
$$T_i^* = W_i\gamma + u_{i2},$$
$$(u_{i1}, u_{i2}) \sim N(0, \Sigma_u)$$
$$\Sigma_u = \begin{bmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{bmatrix},$$

with additional predictors in the Poisson regression X_1 (binary, 1 = education over 12 years, 0 = 12 years or less) and X_2 (binary, 1 for black ethnicity, 0 = non-black). In the treatment regression $W_1 = X_1$, $W_2 = X_2$, W_3 (binary, 1 = has health insurance, 0 = uninsured), W_4 (binary, 1 = receiving registered medical care), and W_5 (binary, 1 = heart condition).

A Ga(1, 0.001) prior is assumed for the unknown variance in Σ and an $N(0, 1)$ prior for the covariance $\rho\sigma$, and N(0, 100) priors for the treatment and other fixed effects. The second half of a two-chain run of 20 000 iterations shows a clear treatment effect that reduces alcohol use (Table 15.6). Alcohol use also falls with longer education, and this variable also reduces