beyond four clusters, whereas Ishwaran and James (2002) find at least five to six clusters with a Dirichlet process approach and under an inverse gamma prior for the σ_k^2. They do, however, find only four clusters when a uniform prior $U(0, 20.83)$ is used for σ_k^2 with 20.83 being the observed variance, $V(y)$. Ando (2007) reports six clusters (assuming a monotonic constraint on the μ_k) via several model fit criteria.

Here, K is taken to be 6 with an identifiability constraint on the normal means applied, together with additional data-based features to ensure sensible inferences. Thus, μ_1 is taken to be normal with a minimum of 9.176, namely, to be at least as large as the minimum y value. Then $\mu_k = \mu_{k-1} + \delta_k$, where the increments are distributed $Ga(1,\ 0.001)$ with a maximum of 20. Without the latter constraint, implausibly large μ_6 values were sometimes sampled. Diffuse $Ga(1,\ 0.001)$ priors are assumed on $1/\sigma_k^2$. Initial values for a two chain run were based on an initial single chain run with trial values for parameters.

Using these initial values, convergence[9] is attained by iteration 2,000 in a run of 10,000 iterations and the clusters means (with their posterior standard deviations) are obtained as {9.7 (0.16), 16.1 (0.05), 19.8 (0.17), 22.7 (0.43), 28.2 (2.5), 33.9 (3.0)}, while the mean cluster probabilities are {0.092, 0.034, 0.34, 0.46, 0.034, 0.04}. The mean BIC is 461, obtained using the posterior mean of minus twice the likelihood of the marginal density,

$$p(y|\pi, \mu, \sigma) = \sum_{k=1}^{K} \pi_k \phi(y|\mu_k, \sigma_k),$$

and adding a penalty based on the known parameter total. While an apparently sensible model, replicate samples y_{new} for some data points are biased away from the observations; for example, both $y_6 = 10.23$ and $y_7 = 10.41$ have posterior predictive means and 95% intervals of 9.7 $\{8.7, 10.6\}$, while observations y_{48} to y_{77} vary from 21.8 to 25.6, but have posterior predicted means between 22.5 and 22.8.

To illustrate the logistic-normal approach, the same number of subgroups is assumed, with the same priors on the component group normal parameters $\{\mu_k, \sigma_k^2\}$. $N(0,\ 1000)$ priors are adopted on the ν_k parameters and a Wishart with 5 degrees of freedom and identity scale matrix assumed for Σ_z^{-1}. The fit obtained is very similar to that under a Dirichlet prior. Replicate data, y_{new}, sampled from the model still do not reproduce observations y_6 and y_7 very closely, and show a flat profile for observations 48 to 77, with y_{new} around 22.6.

3.9 Nonparametric Mixing via Dirichlet Process and Polya Tree Priors

In applications of hierarchical models, inferences may depend on the assumed forms (e.g., normal, gamma) for higher stage priors, and will be distorted

if there are unrecognized features such as multiple modes in the underlying second stage effects. Instead of assuming a known prior distribution, G, for second stage latent effects, such as b_i in the normal-normal model of Section 3.3, the Dirichlet process (DP) prior involves a distribution on G itself, so acknowledging uncertainty about its form (Gill and Casella, 2009). The DP prior involves a baseline or base prior G_0, the expectation of G, and a precision or mass parameter, α, governing the concentration of the prior for G about its mean G_0. For any partition $A_1, \ldots, A_M$ on the support of G_0, the vector $\{G(A_1), \ldots, G(A_M)\}$ of probabilities $G(A_m)$ contained in the set $\{A_m, m = 1, \ldots, M\}$ follows a Dirichlet distribution $D(\alpha G_0(A_1), \ldots, \alpha G_M(A_M))$.

Original forms of the DP prior assumed G_0 to be known (fixed). One problem with a DP when G_0 is known is that it assigns a probability of 1 to the space of discrete probability measures (Hanson et al., 2005, 249). An alternative is to take the parameters in G_0 to be unknown, and to follow a set of parametric distributions, with possibly unknown hyperparameters, resulting in a mixture of Dirichlet process or MDP model (Walker et al., 1999, 489). General computational procedures for such models are discussed by Jara (2007) and Ohlssen et al. (2007).

Following West et al. (1994), assume conventional first stage sampling densities $y_i \backsim p(y_i|b_i, \psi)$, with distributions $P(y_i|b_i, \psi)$. The uncertainty about the appropriate form of prior arises about the distribution G for the latent effects b_i. Under a DP prior, any set of unit-specific parameters $\{b_1, \ldots, b_n\}$ generated from G lies in a set of $K \leq n$ distinct values $\{\zeta_1, \ldots, \zeta_K\}$, which are sampled from G_0. The concentration parameter α governing the closeness of G to G_0 can be taken as an unknown, or assigned a preset value (e.g., $\alpha = 1$). The number of distinct values or clusters K is stochastic, with an implicit prior determined by α, with limiting mean $\alpha \log(1 + n/\alpha)$. Note that the posterior mean of K is not necessarily a reliable guide to the number of components in the data or effects (e.g., components with substantive meaning), though it can be interpreted as an upper bound on the number of components (Ishwaran and Zarepour, 2000, 381–82). Related algorithms include the Chinese Restaurant process (Ishwaran and James, 2003), a method for randomly assigning objects to groups that results in samples having the equivalent sampling distribution as that obtained by the Dirichlet Process.

Given the realised number of clusters K (at any particular MCMC iteration), b_i are sampled from the set $\{\zeta_1, \ldots, \zeta_K\}$ according to a multinomial distribution. Define cluster indicators $S = \{S_1, \ldots, S_n\}$, where $S_i = k$ if $b_i = \zeta_k$ and denote $N_k = \#\{S_i = k\}$ as the total number of units with $S_i = k$ (i.e., units in the same cluster with a common value ζ_k for the second stage latent effect). If α is taken as unknown, its prior is important in determining the number of clusters. Taking $\alpha \backsim Ga(\eta_1, \eta_2)$ where η_1 and η_2 are relatively large will tend to discourage unduly small or large values for α. Typical values are $\eta_1 = \eta_2 = 1$ or $\eta_1 = \eta_2 = 2$, though taking $\eta_2 > \eta_1$ as in $\{\eta_1 = 2, \eta_2 = 4\}$ tends to encourage repetitions in ζ_k, and can be used to assess the number of components present in the data (Ishwaran and Zarepour, 2000, 377). It is clear that the parameters used in the prior for α may affect the number of

components, but typically there is less concern with this aspect in nonparametric mixture modeling (Leslie et al., 2007).

Consider the assignment of a latent effect b_i to a particular unit, given that the remaining $n-1$ latent effects $b_{[i]} = \{b_1, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n\}$ are already assigned. Also let $S_{[i]}$ be a particular configuration of the remaining $n-1$ effects in $b_{[i]}$ into $K_{[i]}$ distinct values, with $N_{[i]k} = \#\{S_i = k,\ k \neq i\}$ denoting the total of those $n-1$ units having a common value $\zeta_{[i]k}$. Then the conditional prior for b_i follows a Polya urn scheme (Dunson et al., 2007, 165; Hanson et al., 2005, 252; West et al., 1994),

$$\begin{aligned}(b_i|b_{[i]}, S_{[i]}, K_{[i]}, \alpha) &\backsim \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{k\neq i}\delta(b_k) \\ &\backsim \frac{\alpha}{\alpha+n-1}G_0 + \frac{1}{\alpha+n-1}\sum_{k=1}^{K_{[i]}}N_{[i]k}\delta(\zeta_{[i]k}), \qquad (3.5)\end{aligned}$$

where $\delta(u)$ denotes a degenerate distribution having a single value at u. So b_i is distinct from the remaining latent values with probability $(\alpha/\alpha+n-1)$, in which case it is drawn from the base prior G_0. Alternatively it is selected from the existing distinct effects, $\zeta_{[i]k}$, according to a multinomial with probabilities proportional to $(N_{[i]k}/\alpha+n-1)$. This selection scheme extends to the predictive scenario, i.e., to the latent effect for a hypothetical new unit $n+1$, with

$$(b_{n+1}|b, S, K, \alpha) \backsim \frac{\alpha}{\alpha+n}G_0 + \frac{1}{\alpha+n}\sum_{k=1}^{K}N_k\delta(\zeta_k).$$

Predictions of the first stage response for unit $n+1$ are obtained as,

$$(y_{n+1}|b, S, K, \alpha) \backsim \frac{\alpha}{\alpha+n}P_{n+1}(|\zeta_{n+1}) + \frac{1}{\alpha+n}\sum_{k=1}^{K}N_kP_{n+1}(|\zeta_k),$$

where ζ_{n+1} is an extra draw from G_0. Predictions beyond $n+1$ may be relevant in panel or time series applications (Hirano, 1998).

In terms of Gibbs sampling, Equation 3.5 implies conditional posteriors (Ishwaran and James, 2001, 166; Neal, 2000; West et al., 1994, 367),

$$(b_i|y, b_{[i]}, S_{[i]}, K_{[i]}, \alpha) \backsim \alpha q_{i0}g_0(b_i|y)p(y_i|b_i) + \sum_{k=1}^{K_{[i]}}q_{ik}\delta(\zeta_{[i]k}),$$

where $g_0(b_i|y)$ is the density corresponding to G_0 evaluated at b_i, and where

$$\begin{aligned}q_{i0} &= \int p(y_i|b_i)g_0(b_i)\,db_i \qquad (3.6)\\ q_{ik} &= N_{[i]k}p(y_i|\zeta_{[i]k}) \qquad k>0.\end{aligned}$$

Normalizing the values αq_{i0} and q_{ik} to probabilities $\{r_{i0}, r_{i1}, \ldots, r_{iK_{[i]}}\}$ summing to 1, the conditional posteriors for the subgroup indicators are then,

$$Pr\left(S_i = k | y, b_{[i]}, S_{[i]}, K_{[i]}\right) = r_{ik},$$

where $S_i = 0$ corresponds to drawing a new sample from G_0 under the Polya urn scheme.

3.9.1 Specifying the baseline density

An important aspect of the MDP framework is the specification of G_0. Assume there are p parameters, $(\psi_1, \ldots, \psi_p)$, in G_0, then one has

$$\begin{aligned} y_i | b_i &\backsim p(y_i | b_i), \\ b_1, \ldots, b_n | G, & \\ G | \alpha, G_0 &\backsim DP(\alpha G_0), \\ G_0 = \{p_{01}(\psi_1 | \xi_1), \ldots, p_{0p}(\psi_p | \xi_p)\}, & \end{aligned}$$

where $\psi_1, \ldots, \psi_p$ are unknown, and also possibly some of the defining ξ parameters. Consider a normal mixture with both means and variances possibly differing for each unit (Cao and West, 1996; Hirano, 2002), namely,

$$y_i \backsim N(\mu_i, \sigma_i^2).$$

The appropriate prior G for $b_i = (\mu_i, \sigma_i^2)$ is not certain, therefore,

$$\begin{aligned} (\mu_i, \sigma_i^2) &\backsim G, \\ G &\backsim DP(\alpha G_0), \end{aligned}$$

where G_0 involves the priors $\mu_i \backsim p_{01}(\mu_i | \xi_1), \sigma_i^2 \backsim p_{02}(\sigma_i^2 | \xi_2)$, with ξ_1 and ξ_2 possibly including further unknowns. For example, Hirano (2002) takes,

$$1/\sigma_i^2 \sim \chi^2(s)/(sQ),$$

and

$$\mu_i \sim N\left(m, c\sigma_i^2\right),$$

where s, Q, m, and c are specified but may be varied in a sensitivity analysis.

The marginal distribution of y_i (averaged over all possible G) in this case is a mixture of normal distributions, with the number of subgroups, K, randomly varying between 1 and n. The n unit-specific parameter pairs, $b_i = (\mu_i, \sigma_i^2)$, are selected under G from the set of $K_{[i]}$ possible values, $\zeta_k = (\mu_k, \sigma_k^2)$, already drawn from G_0, or by fresh sampling from G_0. The q_{ih} in Equation 3.6 are then obtained as

$$\begin{aligned} q_{i0} &= \int \frac{1}{\sigma_i \sqrt{2\pi}} e^{-(y_i - \mu_i)^2 / 2\sigma_i^2} g_0(\mu_i, \sigma_i^2) d\mu_i d\sigma_i^2, \\ q_{ik} &= N_{[i]k} \frac{1}{\sigma_k \sqrt{2\pi}} e^{-(y_i - \mu_k)^2 / 2\sigma_k^2} \qquad k > 0. \end{aligned}$$

As another example, Kleinman and Ibrahim (1998) consider Gibbs updates in an MDP framework for parameters in general linear mixed models for nested data. For example, let X_i and Z_i be predictors of dimension q and r (possibly overlapping) and consider repeated data, y_{it}, over subjects i, with observation vectors $y_i = (y_{i1}, \ldots, y_{iT})$ and first stage model,

$$y_i \sim N\left(X_i\beta + Z_i b_i, \sigma^2\right),$$

where one may assume conventional normal and inverse gamma priors for β and σ^2. However, for $b_i = (b_{i1}, \ldots, b_{ir})$, greater flexibility is obtained by taking,

$$b_i \backsim G,$$
$$G \backsim DP(\alpha, G_0),$$

where G_0 is the multivariate normal of dimension r, with mean 0 but unknown covariance D. The Wishart distribution in the Gibbs update for D^{-1} is modified for clustering of values among the sampled b_i (Kleinman and Ibrahim, 1998, 94).

3.9.2 Truncated Dirichlet processes and stick-breaking priors

Implementation may be simplified if an alternative way to generate the DP prior is adopted. The basis of this alternative scheme is to regard the density of the unit level effects, b_i, as an infinite mixture of point masses or continuous densities (Hirano, 1998; Ohlssen et al., 2007), with

$$b_i \sim \sum_{k=1}^{\infty} \pi_k h(b_i|\psi_k).$$

This approach is called a Dirichlet process mixture by Hanson et al. (2005, 250) and a dependent Dirichlet process by Dunson et al. (2007, 164). For practical application, Ishwaran and James (2002) and Ishwaran and Zarepour (2000) suggest the infinite representation be approximated by one truncated at $M \leq n$ components with

$$g(b) = \sum_{m=1}^{M} \pi_m h(b|\psi_m),$$

where π_m are sampled by introducing $M-1$ beta distributed random variables,

$$V_m \backsim Be(c_m, d_m),$$

with $V_M = 1$ to ensure the random weights, π_m, sum to 1 (Ishwaran and James, 2001; Sethuraman, 1994). Then $\pi_1 = V_1$ and

$$\pi_m = (1 - V_1)(1 - V_2)\cdots(1 - V_{m-1})V_m \qquad m > 1.$$

This method of generation is known as stick breaking, since at each stage, the procedure randomly breaks what is left of a stick of unit length and assigns the length of the break to the current π_m.

Following Pitman and Yor (1997), the beta parameters $\{c_m, d_m\}$ in the prior for V_m can be written as $c_m = 1 - c, d_m = d + mc$, where $c \in [0, 1)$ and $d > -c$. For an infinite dimensional mixture, the Dirichlet process is obtained by taking $c = 0$ and $d = \alpha$, so that $V_m \backsim Be(1, \alpha)$. When a finite (truncated) mixture is used, setting $c_m = 1 + c/M$ and $d_m = \alpha - m\alpha/M = \alpha(1 - m/M)$ is asymptotically equivalent to the DP process (Ishwaran and James, 2001; Ishwaran and Zarepour, 2002).

However, using an approximate DP scheme with $V_m \backsim Be(1, \alpha)$ and M large is equivalent to the infinite DP process for practical purposes (Ishwaran and James, 2002; Ishwaran and Zarepour, 2000, 383). If a $Ga(\eta_1, \eta_2)$ prior is used for α, its full conditional is $\alpha \backsim Ga(M + \eta_1 - 1, \eta_2 - \log(\pi_M))$ (Ishwaran and Zarepour, 2000, 387). The realized number of clusters is $K \leq M$ as above, and Ishwaran and James (2002) suggest AIC and BIC penalties based on K that can be used for model selection.

Taking $V_m \backsim Be(\alpha, 1)$ rather than $V_m \backsim Be(1, \alpha)$ in the truncated stick-breaking scheme means that larger values of α now imply greater clustering into a few subpopulations. This is an example of the beta process priors considered by Ishwaran and Zarepour (2000). Other truncated mixture sampling schemes that start with a prior on α to give an implicit prior on a stochastic K are available. For example, Ishwaran and Zarepour (2000, 376) consider taking α as an unknown in,

$$(\pi_1, \ldots, \pi_M) \backsim D\left(\frac{\alpha}{M}, \frac{\alpha}{M}, \ldots, \frac{\alpha}{M}\right).$$

Alternatively, Green and Richardson (2001, 357) start off with a prior on K and then select the cluster indicators from a multinomial vector with probabilities $p(S_i = k) = \pi_i$, where $(\pi_1, \ldots, \pi_K)$ follow a Dirichlet density $D(\delta, \ldots, \delta)$. They refer to this as an explicit allocation prior and show how the DP prior is obtained as $K \to \infty$ and $\delta \to 0$ in such a way that $K\delta \to \alpha > 0$.

3.9.3 Polya Tree priors

The Polya Tree (PT) is a more general class than the Dirichlet process and has the benefit that it can place probability 1 on the space of continuous densities (Hanson et al., 2005; Walker et al., 1999). In essence, if the support of a parameter ω is denoted Γ then the PT prior chooses the most appropriate value for ω by successive binary partitioning of Γ. The first partition splits Γ into two disjoint sets $\{B_0, B_1\}$; the probabilities of moving into B_0 and B_1 are C_{00} and $C_{01} = 1 - C_{00}$, with C_{00} set to 0.5. At the second partition, B_0 is split into $\{B_{00}, B_{01}\}$ and B_1 is split into $\{B_{10}, B_{11}\}$, so there are 2^2 sets. At the third partition, B_{00} is split into $\{B_{000}, B_{001}\}$, B_{01} into $\{B_{010}, B_{011}\}$, B_{10} into $\{B_{100}, B_{101}\}$, and B_{11} into $\{B_{110}, B_{111}\}$, so there are 2^3 sets. Generally, the number of sets at the mth partition is 2^m.

The partition probabilities at second and subsequent stages are unknown. Let ε denote a sequence of 0s and 1s. For example, suppose B_1 is selected at step 1 and B_{11} is selected at step 2, then $\varepsilon = [1, 1]$. The choice at the next stage between sets $B_{\varepsilon 0}$ and $B_{\varepsilon 1}$ (i.e., between B_{110} and B_{111}) is governed by probabilities $(C_{\varepsilon 0}, C_{\varepsilon 1})$, with a beta prior for $C_{\varepsilon 0}$, and $C_{\varepsilon 1} = 1 - C_{\varepsilon 0}$. The canonical form for the prior on the partition probabilities at partition m is

$$C_{\varepsilon 0} \backsim Be(c_m, c_m),$$
$$c_m = dm^2,$$

where d may be taken as an extra unknown. The Dirichlet process occurs when $c_m = d/2^m$, so that $c_m \to 0$ as $m \to \infty$, whereas $c_m \to \infty$ as $m \to \infty$ is appropriate if the underlying distribution G is expected to be continuous.

While, theoretically, the completely continuous case corresponds to $m \to \infty$, in practice the partitioning is truncated at a finite value M. Hanson and Johnson recommend $M = \log_2(n)$, where n is the sample size. The partitions can be taken to coincide with percentiles of G_0, so for example,

$$B_0 = \left(-\infty, G_0^{-1}(0.5)\right], \qquad B_1 = \left[G_0^{-1}(0.5), \infty\right);$$
$$B_{00} = \left(-\infty, G_0^{-1}(0.25)\right], \qquad B_{01} = \left[G_0^{-1}(0.25), G_0^{-1}(0.5)\right],$$
$$B_{10} = \left[G_0^{-1}(0.5), G_0^{-1}(0.75)\right], \qquad B_{11} = \left[G_0^{-1}(0.75), \infty\right);$$

and so on. Let d_{ki} at partition k and option i be a re-expression of the B_ε (e.g., for $k = 3$, $d_{31} = B_{000}$, $d_{32} = B_{001}$, $d_{33} = B_{010}$, $d_{34} = B_{011}$, $d_{35} = B_{100}$, $d_{36} = B_{101}$, $d_{37} = B_{110}$, $d_{38} = B_{111}$). Then at partition k, for $i = 1, \ldots, 2^k$, the interval boundaries are

$$d_{ki} = \left[G_0^{-1}\left(\frac{i-1}{2^k}\right), G_0^{-1}\left(\frac{i}{2^k}\right)\right],$$

with appropriate modifications for the extreme tails.

For example, consider a PT prior on unstructured errors in a Poisson lognormal mixture, with

$$y_i \backsim Po(\mu_i),$$
$$\log(\mu_i) = \beta + \sigma b_i.$$

Then G_0 for $v_i = \sigma b_i$ is a $N(0, \sigma^2)$ density, with G_0 for b_i being a $N(0, 1)$ density. So with $M = 3$ levels, the relevant ordinates from G_0 for defining the eight intervals are $(-1.15, -0.67, -0.32, 0, 0.32, 0.67, 1.15)$.

input as data

Example 3.7. Nicotine Replacement Therapy Xia et al. (2005) analyzed the NRT trials data of Example 3.1 and detected $K = 2$ subpopulations using a discrete mixture of normals model with $y_i \sim N(b_i, s_i^2)$, with s_i^2 known, and

$$b_i \sim \sum_{k=1}^{K} \pi_k N\left(\mu_k, \tau_k^2\right).$$

They found one subgroup to have a nonsignificant treatment effect, with its μ parameter straddling zero. Here we consider both a conventional discrete normal mixture with $K = 2$, and a DP-based mixture. In the former, a monotonicity constraint is applied to the μ_k. The mean likelihood of this model was improved by taking relatively informative $N(0, 1)$ priors on μ_k and $Ga(1, 1)$ priors on $1/\tau_k^2$; these may be judged reasonable in terms of the likely range of treatment-control log odds ratios typically encountered in trials.

With inferences based on the second half of a two chain run of 10,000 iterations, the average likelihood stands at -56 compared to -62 for a standard normal-normal model with $K = 1$. Posterior means for μ_k (and 95% credible intervals) are found to be 0.38 (-0.83, 0.76) and 0.93 (0.55, 2.19), with respective component probabilities 0.48 and 0.52. Sampling replicate data, $y_{\text{rep},i}$, shows the observations to be reproduced effectively with no posterior probabilities, $Pr(y_{\text{rep},i} > y_i|y)$, exceeding 0.05 or 0.95, and in fact varying between 0.11 and 0.88.

In the DP model, conventional priors for $\{b_i, \mu_k, \tau_k^2\}$ are replaced by a DPM structure (Section 3.9) with $b_i = \zeta_k$ when $S_i = k$ and where the realised number of clusters is $K \leq M$, where a maximum of $M = 20$ possible clusters is assumed. The M potential values $\{\mu_m, \tau_m^2\}$ are sampled from normal densities with means $\mu_m \backsim N(m_\mu, 1)$, where m_μ is itself unknown, and with $1/\tau_m^2 \backsim G(1, 1)$. A $Ga(3, 3)$ prior is assumed on the Dirichlet concentration parameter α.

A two chain run[10] of 5000 iterations shows convergence in α, K, and the latent effects b after around 1000 iterations. The posterior mean and median of K are, respectively, 3.5 and 3, supporting a relatively small number of components in the second stage prior of NRT effects; α has a posterior mean of 0.78. A plot of the posterior means of the b_i does not show sharply distinct subgroups (Figure 3.2), though outliers can be seen, such as trial 36, a trial with a relatively large number of subjects in the "women, long term follow-up category" in Cepeda-Benito et al. (2004). However, the effects show more peakedness than under a normal density (superimposed plot).

Sampling replicate data, $y_{\text{rep},i}$, shows the observations to be reproduced effectively with no posterior probabilities, $Pr(y_{\text{rep},i} > y_i|y)$, exceeding 0.05 or 0.95, and in fact varying between 0.09 and 0.93. The latter value for $Pr(y_{\text{rep},i} > y_i|y)$ is for trial 66 where r_{T} is only 2 (from $N_{\text{T}} = 86$ in the treatment group).

Example 3.8. Eye-Tracking Data Escobar and West (1998) present count data on eye-tracking anomalies in schizophrenic patients ($n = 101$). The data are overdispersed and the first analysis assumes a DP Poisson-gamma mixture with G_0 being a gamma density with unknown shape and scale parameters. So,

$$
\begin{aligned}
y_i &\backsim Po(b_i), \\
b_i &\backsim G, \\
G &\backsim DP(\alpha G_0), \\
G_0 &= Ga(c_g, d_g).
\end{aligned}
$$

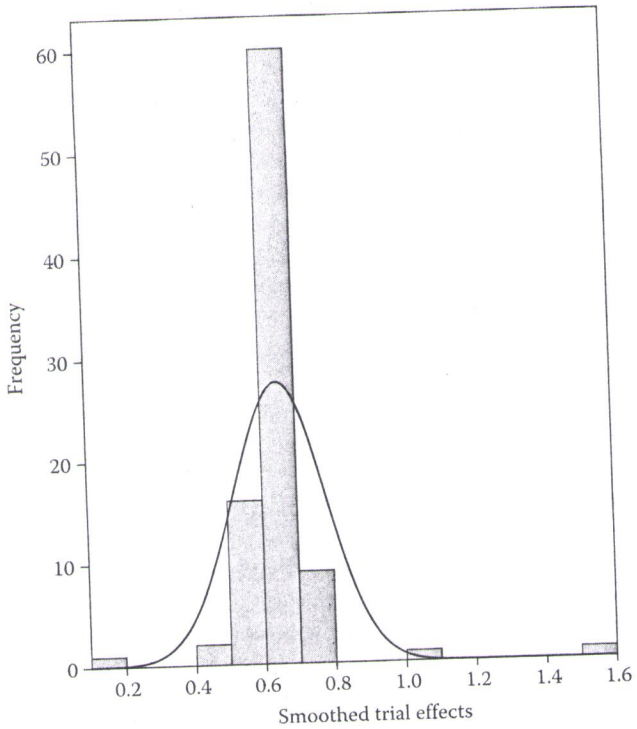

FIGURE 3.2
Histogram of smoothed trial effects.

Taking c_g and d_g to be unknowns results in an MDP prior, which is implemented using the Polya urn prior (Equation 3.5). A $Ga(1,1)$ prior is assumed on α, and exponential $E(1)$ priors on the parameters (c_g, d_g) with a minimum of 0.5 on c_g for numerical stability.

In line with Marshall and Spiegelhalter (2003), the observed y_i are compared with replicates sampled from the predictive distribution $p(y_{\text{rep}}|y)$ to see if y_i are at odds with the model. Discrepancies could be due to genuine outlier status, or to model failures. For discrete data, the relevant p-value is

$$Pr(y_{\text{rep},i} < y_i) + 0.5Pr(y_{\text{rep},i} = y_i).$$

A related check is whether the 95% intervals for $y_{\text{rep},i}$ include y_i (Gelfand, 1996).

From the last 4000 iterations of a two chain run[11] of 5000 iterations in OpenBUGS, an average of $K = 12$ distinct clusters is obtained, with posterior mean (sd) for α, c_g, and d_g of 2.86 (1.5), 0.77 (0.27), and 0.13 (0.06). Figure 3.3 shows the prediction y_{new} for a new case, and demonstrates that the main source of overdispersion is skewness in the latent frailties, b_i, rather than multiple modes. The predictive checks based on replicate samples are

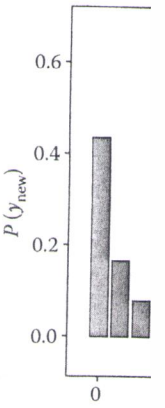

FIGURE 3.3
Prediction for new

satisfactory. Note t
parameters are set,
obtained on some b
to be an extreme ob

A second analysi
namely, $y_i \backsim Po(\mu_i$

where G_0 for $v_i =$
at $M = 4$ and a G
is selected, uniform
defined by G_0, excep

As for the Polya
predictive check ind
95% interval) 2.0 (1
known, then predict
bimodal posteriors f
such that the prior c

Appendix: Con

1. The code for the r

```
model {# predict
ynew ~dnorm(b.
exp(ynew)
```

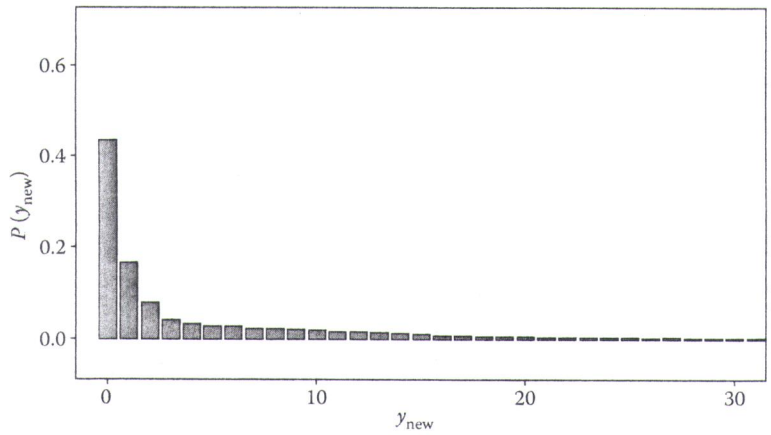

FIGURE 3.3
Prediction for new case.

satisfactory. Note that the same does not apply if the gamma mixing density parameters are set, e.g., $c_g = d_g = 1$. In this case, bimodal posteriors are obtained on some b_i (e.g., b_{92}) and predictive checks for $y_{101} = 34$ suggest it to be an extreme observation.

A second analysis involves a Polya Tree prior in a Poisson-lognormal model, namely, $y_i \backsim Po(\mu_i)$ with

$$\log(\mu_i) = \beta + \sigma b_i,$$

where G_0 for $v_i = \sigma b_i$ is an $N(0, \sigma^2)$ density. The number of stages is set at $M = 4$ and a $Ga(1, 1)$ prior is assumed on $1/\sigma^2$. Once an interval, $B_{\varepsilon m}$, is selected, uniform sampling to generate b_i takes place within the interval defined by G_0, except in the tails where the sampling is from a $N(0, 1)$.

As for the Polya urn model (with the same MCMC details), both types of predictive check indicate no major discrepancies. σ has posterior mean (and 95% interval) 2.0 (1.6, 2.5). If σ is taken to equal 1 so that G_0 is assumed known, then predictive discrepancies do occur. Taking $\sigma = 1$ also leads to bimodal posteriors for individual b_i indicating a clash between prior and data, such that the prior cannot accommodate certain values.

Appendix: Computational Notes

1. The code for the nicotine replacement example is

model {# predictions

ynew ∽dnorm(b.new,inv.s20); b.new ∽dnorm(mu,inv.tau2); OR.new <- exp(ynew)