Chapter 23

Dirichlet process models

The Dirichlet process is an infinite-dimensional generalization of the Dirichlet distribution that can be used to set a prior on unknown distributions. Furthermore, these unknown densities can be used to extend finite component mixture models to infinite component mixture models.

23.1 Bayesian histograms

We start in this section and the next with the relatively simple setting in which $y_i \overset{iid}{\sim} f$ and the goal is to obtain a Bayes estimate of the density f. The histogram is often (also in this book) used as a simple form of density estimate. In this section we develop a flexible parametric version of the histogram that helps to motivate the fully nonparametric Bayesian density estimation of the following section. The remainder of the chapter shows how the Dirichlet process can be applied beyond density estimation.

Assume we have prespecified knots $\xi = (\xi_0, \xi_1, \ldots, \xi_k)$ to define our histogram estimate, with $\xi_0 < \xi_1 < \cdots < \xi_{k-1} < \xi_k$ and $y_i \in [\xi_0, \xi_k]$. A probability model for the density that is analogous to the histogram is as follows:

$$f(y) = \sum_{h=1}^{k} 1_{\xi_{h-1} < y \leq \xi_h} \frac{\pi_h}{(\xi_h - \xi_{h-1})}, \quad y \in \Re,$$

with $\pi = (\pi_1, \ldots, \pi_k)$ an unknown probability vector. We complete a Bayes specification with a prior distribution for the probabilities. Assume a Dirichlet$(a_1, \ldots, a_k)$ prior distribution for π,

$$p(\pi|a) = \frac{\prod_{h=1}^{k} \Gamma(a_h)}{\Gamma\left(\sum_{h=1}^{k} a_h\right)} \prod_{h=1}^{k} \pi_h^{a_h - 1}.$$

The hyperparameter vector can be re-expressed as $a = \alpha\pi_0$, where

$$\mathrm{E}(\pi|a) = \pi_0 = \left(\frac{a_1}{\sum_h a_h}, \ldots, \frac{a_k}{\sum_h a_h}\right)$$

is the prior mean and α is a scale that is often interpreted as a *prior sample size.*

The posterior distribution of π is then calculated as

$$\begin{aligned} p(\pi|y) &\propto \prod_{h=1}^{k} \pi_h^{a_h - 1} \prod_{i: y_i \in (\xi_{h-1}, \xi_h]} \frac{\pi_h}{\xi_h - \xi_{h-1}} \\ &\propto \prod_{h=1}^{k} \pi_h^{a_h + n_h - 1} \overset{\mathcal{D}}{=} \text{Dirichlet}(a_1 + n_1, \ldots, a_k + n_k), \end{aligned}$$

where $n_h = \sum_i 1_{\xi_{h-1} < y_i \leq \xi_h}$ is the number of observations falling in the hth histogram bin.

To illustrate the Bayesian histogram method, we simulated data from the mixture,

$$f(y) = 0.75\,\text{Beta}(y|1,5) + 0.25\,\text{Beta}(y|20,2),$$

with $n = 100$ samples drawn from this density. Assuming data between [0,1] and choosing 10 equally spaced knots, we applied the Bayesian histogram approach and plotted the true density and simulations from the posterior distribution of the histogram obtained from this procedure.

The Bayesian histogram estimator does an adequate job approximating the true density, but the results are sensitive to the number and locations of knots. However, an appealing property of the Bayesian histogram approach is that implementation is easy since we have conjugacy and the posterior can be calculated analytically. In addition, the approach allows prior information to be included and allows easy production of interval estimates, and hence has some practical advantages over classical histogram estimators. To improve performance a prior can be placed on the numbers and locations of knots, with reversible jump MCMC (see Section 12.3) used for computation, but such an approach is computationally demanding. In addition, even averaging over random knots will tend to introduce artifactual bumps in the density estimate. The Dirichlet prior distribution is perhaps not the best choice due to the lack of smoothing across adjacent bins, but it does have the advantage of conjugacy and simplicity in interpretation of the hyperparameters.

23.2 Dirichlet process prior distributions

Definition and basic properties

Motivated by the simplicity of the Bayesian histogram approach with a Dirichlet prior, one wonders whether we can somehow bypass the need to explicitly specify bins. This would also facilitate extensions to multivariate cases in which there is an explosion of the number of bins that would be needed. With this thought in mind, suppose the sample space is Ω, partitioned into measurable subsets $B_1, \ldots, B_k$. If $\Omega = \Re$, then $B_1, \ldots, B_k$ are simply non-overlapping intervals partitioning the real line into a finite number of bins.

Let P denote the unknown probability measure over $(\Omega, \mathcal{B})$, with $\mathcal{B}$ the collection of all possible subsets of the sample space Ω. The probability measure will assign probabilities to these subsets (bins), with the probabilities allocated to a set of bins $B_1, \ldots, B_k$ partitioning Ω being

$$P(B_1), \ldots, P(B_k) = \int_{B_1} f(y)dy, \ldots, \int_{B_k} f(y)dy.$$

If P is a random probability measure (RPM), then these bin probabilities are random variables. A simple conjugate prior for the bin probabilities corresponds to the Dirichlet distribution. For example, we could let

$$P(B_1), \ldots, P(B_k) \sim \text{Dirichlet}\,(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)), \tag{23.1}$$

where P_0 is a base probability measure providing an initial guess at P, and α is a prior concentration parameter controlling the degree of shrinkage of P toward P_0.

Prior (23.1) is essentially a Bayesian histogram model closely related to that described in the previous section. However, the difference is that (23.1) only specifies that bin B_k is assigned probability $P(B_k)$ and does not specify how probability mass is distributed across the bin B_k. Hence, for a fixed partition $B_1, \ldots, B_k$, (23.1) does not induce a fully specified prior for the random probability measure P. The idea is to eliminate sensitivity to the choice of partition $B_1, \ldots, B_k$ and induce a fully specified prior on P through assuming (23.1) holds for all possible partitions $B_1, \ldots, B_k$ and all k.

For this specification to be coherent, there must exist a random probability measure P such that the probabilities assigned to any measurable partition $B_1, \ldots, B_k$ by P is Dirichlet$(\alpha P_0(B_1), \ldots, \alpha P_0(B_k))$. The existence of such a P can be shown by verifying certain consistency conditions, and the resulting random probability measure P is referred to as a Dirichlet process. Then, as a concise notation to indicate that a probability measure P on $(\Omega, \mathcal{B})$ is assigned a Dirichlet process (DP) prior, let $P \sim \text{DP}(\alpha P_0)$, where $\alpha > 0$ is a scalar precision parameter and P_0 is a baseline probability measure also on $(\Omega, \mathcal{B})$. This baseline P_0 is commonly chosen to correspond to a parametric model such as a Gaussian.

The definition of the Dirichlet process and properties of the Dirichlet distribution imply,

$$P(B) \sim \text{Beta}\big(\alpha P_0(B), \alpha(1 - P_0(B))\big), \quad \text{for all } B \in \mathcal{B},$$

so that the marginal random probability assigned to any subset B of the support is simply beta distributed. It follows directly that the prior mean has the form

$$\text{E}(P(B)) = P_0(B), \quad \text{for all } B \in \mathcal{B},$$

so that the prior for P is centered on P_0. In addition, the prior variance is

$$\text{var}(P(B)) = \frac{P_0(B)(1 - P_0(B))}{1 + \alpha}, \quad \text{for all } B \in \mathcal{B},$$

so that α is a precision parameter controlling the variance.

Hence, the Dirichlet process is appealing in having a simple specification arising from a model similar to a random histogram but without the dependence on the bins, while also having simple and intuitive forms for the prior mean and variance. The prior can be centered on a parametric model for the distribution of the data through the choice of P_0, while allowing α to control uncertainty in this choice. Moreover, it can be shown that the support of the DP contains all probability measures whose support is contained in the support of the baseline probability measure P_0.

The DP prior distribution also has a conjugacy property which makes inferences straightforward. To demonstrate this, first let $y_i \overset{iid}{\sim} P$, for $i = 1, \ldots, n$ and $P \sim \text{DP}(\alpha P_0)$, where we follow common convention in using P to denote both the probability measure and its corresponding distribution. Then, from (23.1) and conjugacy properties of the finite Dirichlet distribution, for any measurable partition $B_1, \ldots, B_k$, we have

$$P(B_1), \ldots, P(B_k) \,|\, y_1, \ldots, y_n \sim \text{Dirichlet}\left(\alpha P_0(B_1) + \sum_{i=1}^{n} 1_{y_i \in B_1}, \ldots, \alpha P_0(B_k) + \sum_{i=1}^{n} 1_{y_i \in B_k}\right).$$

From this and the above development, it is straightforward to obtain

$$P \,|\, y_1, \ldots, y_n \sim \text{DP}\left(\alpha P_0 + \sum_i \delta_{y_i}\right).$$

The updated precision parameter is $\alpha + n$, so that α is in some sense a prior sample size. The posterior expectation of P is defined as

$$\text{E}(P(B) \,|\, y^n) = \left(\frac{\alpha}{\alpha + n}\right) P_0(B) + \left(\frac{n}{\alpha + n}\right) \sum_{i=1}^{n} \frac{1}{n} \delta_{y_i}. \tag{23.2}$$

Hence, the Bayes estimator of P under squared error loss is the empirical measure with equal masses at the data points shrunk toward the base measure. It is clear that as the sample size increases, the Bayesian estimate of the distribution function under the Dirichlet process prior will converge to the empirical distribution function.

In addition, in the limit as the precision parameter α approaches 0, so that we in some sense have a noninformative prior distribution, the posterior distribution is

$$P \mid y^n \sim \mathrm{DP}\left(\sum_{i=1}^{n} \delta_{y_i}\right).$$

This limiting posterior distribution is sometimes known as the Bayesian bootstrap. Samples from the Bayesian bootstrap correspond to discrete distributions supported at the observed data points with Dirichlet distributed weights. Compared with the classical bootstrap, the Bayesian bootstrap leads to smoothing of the weights.

Even with these many appealing properties, the Dirichlet process prior distribution has some important drawbacks. Firstly, there is a lack of smoothness apparent in (23.2). Ideally, one would not simply take a weighted average of the base measure and the empirical measure with masses at the observed data points, but would allow smooth deviations from the base measure. Smoothness would imply dependence between $P(B_1)$ and $P(B_2)$ for adjacent bins B_1 and B_2. However, the DP actually induces negative correlation between $P(B_1)$ and $P(B_2)$ for any two disjoint sets B_1 and B_2, with no account for the distance between these sets. An even more important concern for density estimation is that realizations from the DP are discrete distributions. Hence, $P \sim \mathrm{DP}(\alpha P_0)$ implies that P will be atomic having nonzero weights only on a set of atoms and will not have a continuous density on the real line.

Despite these drawbacks the DP has been useful in developing flexible models for a wide variety of problems. Before demonstrating some of the applications we introduce an alternative characterization of the Dirichlet process.

Stick-breaking construction

The above specification of the Dirichlet process does not provide an intuition for what realizations $P \sim \mathrm{DP}(\alpha P_0)$ actually look like, since the DP prior was defined indirectly through the marginal probabilities allocated to finite partitions. However, there is a direct constructive representation of the Dirichlet process, which is referred to as the stick-breaking representation, which is useful in obtaining further insight into properties of the DP and as a stepping stone for generalizations.

The stick-breaking representation allows us to induce $P \sim \mathrm{DP}(\alpha P_0)$ by letting

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot), \quad \pi_h = V_h \prod_{l<h}(1 - V_l), \; V_h \sim \mathrm{Beta}(1, \alpha), \; \theta_h \sim P_0,$$

where δ_θ denotes a degenerate distribution with all its mass at θ, the *atoms* $(\theta_h)_{h=1}^{\infty}$ are generated independently from the base distribution P_0, π_h is the probability mass at atom θ_h, and these probability masses are generated from a so-called *stick-breaking process* that guarantees that the weights sum to 1.

To describe the stick-breaking process, we start with a stick of unit length representing the total probability to be allocated to all the atoms. We initially break off a random piece of length V_1, with the length generated from a $\mathrm{Beta}(1, \alpha)$ distribution, and allocate this $\pi_1 = V_1$ probability weight to the randomly generated first atom $\theta_1 \sim P_0$. There is then $1 - V_1$ of the stick remaining to be allocated to the other atoms. We break off a proportion $V_2 \sim \mathrm{Beta}(1, \alpha)$ of the $1 - V_1$ stick and allocate the probability $\pi_2 = V_2(1 - V_1)$ to the second atom $\theta_2 \sim P_0$. As we proceed, the stick gets shorter and shorter so that the lengths allocated to the higher indexed atoms decrease stochastically, with the rate of decrease depending on the DP precision parameter α. Because $\mathrm{E}(V_h) = \frac{1}{1+\alpha}$, values of α close to

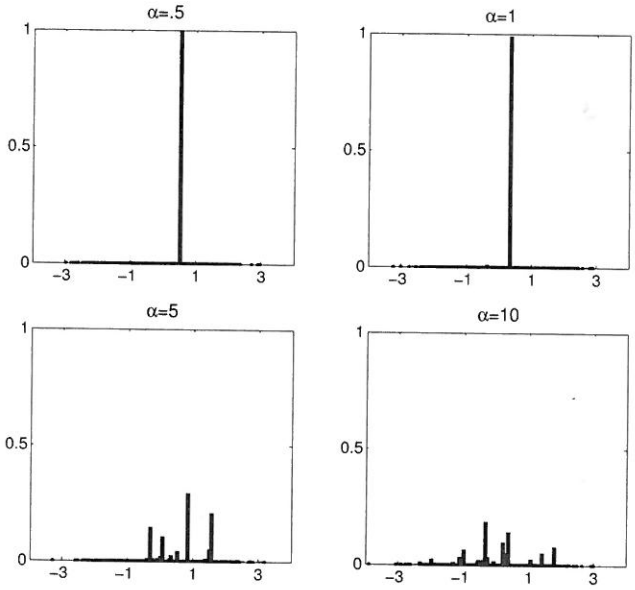

Figure 23.1 *Samples from the stick-breaking representation of the Dirichlet process with different settings of the precision parameter* α.

zero lead to high weight on the first couple atoms, with the remaining atoms being assigned small probabilities.

Figure 23.1 shows realizations of the stick-breaking process for P_0 corresponding to a standard normal distribution and different values of α. From this figure it is apparent that the DP is not appropriate as a direct prior on the distribution of the data, particularly if the data are continuous. For continuous data, each new subject requires a new atom so that a large value of α is required, implying weight close to zero on each atom and hence a small probability of ties in the realizations from P. In the limit as $\alpha \to \infty$, one obtains $y_i \sim P_0$, and hence for large α and no ties in the observations, the DP prior effectively models the data are drawn from the parametric base distribution.

23.3 Dirichlet process mixtures

Specification and Polya urns

The failure of the DP prior distribution as a direct model for the distribution of the data does not imply that it is not useful in applications. Instead, the DP is more appropriately used as a prior for an unknown mixture distribution. Focusing again on the density estimation case for simplicity, a general kernel mixture model for the density can be specified as

$$f(y|P) = \int \mathcal{K}(y|\theta)dP(\theta), \tag{23.3}$$

where $\mathcal{K}(\cdot|\theta)$ is a kernel, with θ including location and possibly scale parameters, and P is a mixing measure. In the special case in which P is treated as discrete with masses at a finite number of k atoms, one obtains a finite mixture model as discussed in Chapter 22.

In a infinite kernel mixture model, one chooses a prior $P \sim \pi_{\mathcal{P}}$ for the unknown mixing measure P, where $\mathcal{P}$ denotes the space of all probability measures on $(\Omega, \mathcal{B})$ and $\pi_{\mathcal{P}}$ denotes a prior over this space. The prior for the mixing measure induces a prior on the density $f(y)$ through the integral mapping in (23.3). If $\pi_{\mathcal{P}}$ is chosen to correspond to a DP prior,

then one obtains a DP mixture model. From (23.2) and (23.3), a DP prior on P leads to

$$f(y) = \sum_{h=1}^{\infty} \pi_h \mathcal{K}(y|\theta_h^*), \tag{23.4}$$

where $\pi =\sim \text{stick}(\alpha)$ is a shorthand notation to denote that the probability weights are sampled from a DP stick-breaking process with parameter α, and with $\theta_h \sim P_0$ independently for $h = 1, \ldots, \infty$.

Expression (23.4) resembles the finite mixture models considered in Chapter 22, but with the important difference that the number of mixture components (latent subpopulations) is set to infinity. However, this does not imply that infinitely many components are *occupied* by the subjects in the sample; rather, the model allows flexibility by introducing new mixture components as subjects are added. Consider the hierarchical specification in which

$$y_i \sim \mathcal{K}(\theta_i), \quad \theta_i \sim P, \quad P \sim \text{DP}(\alpha P_0).$$

This formulation is equivalent to sampling y_i from the infinite mixture model in (23.4). A key question is how to conduct posterior computation under this DP mixture (DPM)? This initially seems problematic in that the mixing measure P is characterized by infinitely many parameters, as is apparent in (23.2), and we no longer have joint conjugacy in which the posterior of P given $y^n = (y_1, \ldots, y_n)$ has a simple form.

A clever way around this problem is to marginalize out P to obtain an induced prior distribution on the subject-specific parameters $\theta^n = (\theta_1, \ldots, \theta_n)$. In particular, marginalizing out P, we obtain the *Polya urn* predictive rule,

$$p(\theta_i|\theta_1, \ldots, \theta_{i-1}) \sim \left(\frac{\alpha}{\alpha + i - 1}\right) P_0(\theta_i) + \sum_{j=1}^{i-1} \left(\frac{1}{\alpha + i - 1}\right) \delta_{\theta_j}. \tag{23.5}$$

This conditional prior distribution consists of a mixture of the base measure P_0 and probability masses at the previous subject's parameter values.

A Chinese restaurant process metaphor is commonly used in describing the Polya urn scheme. Consider a restaurant with infinitely many tables. The first customer sits at a table with dish θ_1^*. The second customer sits at the first table with probability $\frac{\alpha}{1+\alpha}$ or a new table with probability $\frac{1}{1+\alpha}$. This process continues with the ith customer sitting at an occupied table with probability proportional to the number of previous customers at that table and sitting at a new table with probability proportional to α. Each occupied table in the (infinite) restaurant represents a different cluster of subjects, with new clusters added at a rate proportional to $\alpha \log n$ in the asymptotic limit. The number of clusters depends (probabilistically) on the number of subjects n with new clusters introduced as needed as additional subjects are added to the sample. This makes more sense in typical applications than finite mixture models in which k does not depend on n and can be thought of as a formal procedure mimicking the good practice, when fitting a finite mixture model, of manually adding new mixture components as necessary to fit the data.

The simple form for the conditional distribution in (23.5) leads to a useful idea for posterior computation and prediction. From exchangeability of the subjects $i = 1, \ldots, n$, one can obtain the conditional prior distribution for θ_i given $\theta_{(-i)} = (\theta_j, j \neq i)$ as

$$\theta_i|\theta_{-i} \sim \left(\frac{\alpha}{\alpha + n - 1}\right) P_0(\theta_i) + \sum_{h=1}^{k^{(-i)}} \left(\frac{n_h^{(-i)}}{\alpha + n - 1}\right) \delta_{\theta_h^{*(-i)}}, \tag{23.6}$$

where $\theta_h^*, h = 1, \ldots, k^{(-i)}$, are the unique values of $\theta^{(-i)}$, and $n_h^{(-i)} = \sum_{j \neq i} 1_{\theta_j = \theta_h^*}$.

Updating the full conditional prior (23.6) with the data, one obtains a conditional posterior distribution for θ_i having the same form but with updated weights on the components and updated P_0, as long as P_0 is conjugate to the kernel $\mathcal{K}$. For example, this occurs when $\mathcal{K}(\cdot|\theta)$ is a normal kernel, with $\theta = (\mu, \phi)$ the mean and precision and P_0 a normal-gamma prior distribution. Potentially, one can update the θ_i's one at a time from these full conditional posterior distributions in implementing Gibbs sampling. However, this approach tends to have poor mixing.

An alternative *marginal Gibbs sampler*, which instead separately updates the allocation of subjects to clusters and the cluster-specific parameters, proceeds as follows. Let $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$ denote the unique values of θ and let $S_i = c$ if $\theta_i = \theta_c^*$ so that S_i denotes allocation of subject i to a cluster. The Gibbs sampler alternates between

1. Update the allocation S by sampling from the multinomial conditional posterior with

$$\Pr(S_i = c \,|\, -) \propto \begin{cases} n_h^{(-i)} \mathcal{K}(y_i|\theta_c^*) & c = 1, \ldots, k^{(-i)} \\ \alpha \int \mathcal{K}(y_i|\theta) dP_0(\theta) & c = k^{(-i)} + 1 \end{cases}$$

   If $S_i = k^{(-i)} + 1$, then subject i is allocated to a singleton cluster.

2. Update the unique values θ^* by sampling from

$$p(\theta_c^* \,|\, -) \propto P_0(\theta_c^*) \prod_{i:S_i=c} \mathcal{K}(y_i|\theta_c^*),$$

   which is simply the posterior distribution under the parametric model that assigns prior distribution P_0 to the parameters θ_h^* and then updates this prior with the likelihood for those subjects in cluster h.

When P_0 is conjugate to the kernel $\mathcal{K}$, the integral in step 1 can be calculated analytically and the conditional posterior in step 2 has the same parametric form as P_0 except with updated parameters. For example, when the kernel is Gaussian, with θ the mean and variance and P_0 a conjugate normal-inverse-gamma prior, the conditional distribution of θ_c^* has the same normal-inverse-gamma form described in Chapter 22 in the finite mixture case. There are modifications available to accommodate nonconjugate cases as well.

In step 1 of the above Gibbs sampler, either the ith subject is allocated to an existing cluster occupied by one of the other subjects in the sample or the subject is allocated to a new cluster. The conditional posterior probability of allocation to a new cluster is proportional to the DP precision parameter α multiplied by the marginal likelihood for the ith subject's data, obtained in integrating the likelihood $\mathcal{K}(y_i|\theta)$ over the prior $\theta \sim P_0$. If α is close to zero or this marginal likelihood is small relative to the likelihoods for the ith subject's data given allocation to one of the occupied clusters, then subject i will tend to be allocated to an existing cluster. Hence, both α and P_0 play important roles in controlling the posterior distribution over clusterings. As α decreases, there is an increasing tendency to cluster subjects, with a parametric model $y_i \sim \mathrm{K}(\theta)$ for a common θ obtained in the limit as $\alpha \to 0$. In practice, it is common to either set $\alpha = 1$ to favor allocation to few clusters or to choose a gamma hyperprior for α to allow greater data-adaptivity, with an additional MCMC step included to update α.

Somewhat more subtle, and often overlooked, is the role of P_0 in controlling clustering behavior. One may naively try a high variance P_0 to express ignorance about the prior distribution of likely locations for the different kernels. However, similarly as discussed in Section 7.4, a flat prior for P_0 can turn out to make strong assumptions, in this case effectively placing a heavy penalty on the introduction of new clusters. This is because as the variance of P_0 becomes high, the marginal likelihood will decrease, since the prior P_0 places small probability in a region of plausible θ values in such cases. In the limit as the variance of $P_0 \to \infty$, the posterior will behave as if the likelihood is $y_i \sim \mathrm{K}(\theta)$ with

a common θ for all individuals. In practice, we recommended constructing an informative P_0 placing high probability on introducing clusters near the support of the data; this can be facilitated by standardizing the data in advance of the analysis. Refer to the relevant discussion in Chapter 22.

This Gibbs sampler for Dirichlet process models closely resembles the Gibbs sampler for finite mixtures, with the main difference being that we marginalize out the weights π on the different clusters and allow the number of clusters to vary across the samples. The number of mixture components k represented in the sample of n subjects is treated as unknown, and we obtain samples from the posterior of k automatically without needing a reversible jump MCMC algorithm. From the Gibbs samples, we can also estimate the predictive density of y_{n+1} using

$$p(y) = \sum_{c=1}^{k} \left(\frac{n_c}{n+\alpha} \right) K(y|\theta_c^*) + \left(\frac{\alpha}{n+\alpha} \right) \int K(y|\theta) dP_0(\theta),$$

averaged over posterior simulations. The simplicity of this Gibbs sampler and the ability to bypass the issue of selection of k by embedding in an infinite mixture model, which automatically introduces new components at a slow rate as needed when additional subjects are added to the sample, are major reasons for the large applied success of Dirichlet process mixture models.

The Gibbs sampler for finite mixture models introduced in Chapter 22 provides an approximation to a DP mixture model with $P \sim DP(\alpha P_0)$ as long as the mixture component-specific parameters are drawn iid *a priori* from P_0 and the prior on the weights is $\pi \sim \text{Dirichlet}(\frac{\alpha}{k}, \ldots, \frac{\alpha}{k})$. The approximation improves with k and in practice one can set k equal to a conservative upper bound on the number of occupied clusters in the sample ($k = 25$ or 50 can work well). Indeed, we refer the reader to the discussions in Chapter 22 pertaining to the issues that arise in finite mixture modeling, as essentially the same issues arise in infinite discrete mixtures, such as DPMs, and the same solutions apply.

Blocked Gibbs sampler

By marginalizing out the random probability measure P, we give up the ability to conduct inferences on P. By having approaches that avoid marginalization, we open the door to generalizations of DPMs for which marginalization is not possible analytically. One approach for avoiding marginalization is to rely on the construction in (23.4). Because the stick-breaking construction orders the mixture components so that the weights are stochastically decreasing in the index h, for a sufficiently high index N, we will have that $\sum_{N+1}^{\infty} \pi_h$ has a distribution concentrated near zero. Hence, we can obtain an accurate approximation by letting $V_N = 1$ in the stick-breaking process so that $\pi_h = 0$ for $h = N+1, \ldots, \infty$, with N chosen to be sufficiently large. In practice, $N = 25$ or 50 is commonly chosen as a default, with N providing an upper bound on the number of clusters in the n subjects in the sample. We have rarely seen a need for more than 10 or 15 clusters to accurately fit the unknown density.

The truncation approximation to the DP leads to a straightforward MCMC algorithm for posterior computation, and represents an alternative to the finite Dirichlet approximation described in Chapter 22. It is not clear which of these approaches leads to more efficient posterior computation, though the exchangeability of the components in the finite Dirichlet approximation conveys some advantages in terms of mixing. Using the stick-breaking truncation, the following blocked Gibbs sampler can be used:

1. Update $S_i \in \{1, \ldots, N\}$ by multinomial sampling with

$$\Pr(S_i = c \,|\, -) = \frac{\pi_c \mathcal{K}(y_i|\theta_c^*)}{\sum_{c'=1}^{N} \pi_{c'} \mathcal{K}(y_i|\theta_{c'}^*)}, \quad c' = 1, \ldots, N,$$

where $S_i = c$ if $\theta_i = \theta_c^*$ denotes that subject i is allocated to cluster c.

2. Update the stick-breaking weight V_c, $c = 1, \ldots, N-1$, from $\text{Beta}\left(1+n_c, \alpha+\sum_{c'=c+1}^{N} n_{c'}\right)$.
3. Update θ_c^*, $c = 1, \ldots, N$, exactly as in the finite mixture model, with the parameters for unoccupied clusters with $n_c = 0$ sampled from the prior P_0.

This algorithm involves simple sampling steps and is straightforward to implement. In order to estimate the density $f(y)$ one can follow the approach of monitoring $f(y) = \sum_{c=1}^{N} \pi_h \mathcal{K}(y|\theta_c^*)$ at each iteration over a dense grid of y values (for example, an equally spaced grid of 100 values ranging from the minimum of the observed y's minus a small buffer to the maximum of the observed y's plus a small buffer). Based on these samples, we can compute posterior inferences.

When running the algorithm, it is good practice to monitor $S_{\max} = \max(S_1, \ldots, S_n)$ to verify that the maximum occupied component index has a low probability of being close to the upper bound of N. Otherwise, the upper bound should be increased. Convergence should be monitored on the sampled $f(y)$ values and not to the mixture component-specific parameters. As discussed in Chapter 22, label ambiguity problems often lead to poor mixing of the component-specific parameters, but this may not impact convergence and mixing of the induced density of interest.

Gibbs sampling algorithms that rely on stick-breaking representations have performed well in our experience. But in some cases, all of the above algorithms can encounter slow mixing that arises due to the multimodal nature of the posterior in which it can be difficult to move rapidly between different clusterings. This mixing problem can be partly addressed by incorporating label switching moves and there is also a literature on split-merge algorithms designed for more rapid exploration of the distribution of cluster allocations.

Hyperprior distribution

The DP precision parameter α plays a key role in controlling the prior on the number of clusters, and there are a number of possible strategies in terms of specifying α. One can fix α at a small value to favor allocation to few clusters relative to the sample size, with a commonly used default value corresponding to $\alpha = 1$. This implies that, in the prior distribution, two randomly selected individuals have a 50-50 chance of belonging to the same cluster. Alternatively, one can allow the data to inform about the appropriate value of α by choosing a hyperprior, such as $\alpha \sim \text{Gamma}(a_\alpha, b_\alpha)$, and then updating α during the MCMC analysis. For the blocked Gibbs sampler, the gamma hyperprior is conditionally conjugate with the resulting conditional posterior being

$$\alpha|- \sim \text{Gamma}\left(a_\alpha + N - 1, b_\alpha - \sum_{h=1}^{N-1} \log(1 - V_h)\right).$$

Hence, sampling from this conditional can be included as an additional step in the algorithm described in the previous subsection.

In our experience, the data tend to be informative about the precision parameter α of the Dirichlet process, and hence there is substantial Bayesian learning, with a high variance prior often resulting in a concentrated, low-variance posterior. It may seem counterintuitive that the data can inform strongly about the number of clusters through the hyperparameter α given that maximum likelihood estimation leads to overfitting, with more clusters resulting in a higher maximized likelihood. However, the Bayesian approach favors clusterings and values of α that result in a high *marginal* likelihood. If individuals are allocated to many different clusters, the effective number of parameters in the likelihood is increased, and we then integrate across a larger space in calculating the marginal likelihood. This induces an

intrinsic penalty that favors allocation to few clusters that are really needed to fit the data; there is no tendency for overfitting.

The more difficult and subtle issue is choice of the base measure P_0. Often the base measure is chosen for computational convenience to be conjugate. However, even in conjugate parametric families such as normal-gamma, we can potentially improve flexibility by placing hyperparameters on the parameters in P_0. P_0 can be thought of as inducing the prior for the cluster locations. If these locations are too spread out, because P_0 has high variance, then the penalty in the marginal likelihood for allocating individuals to different clusters will be large, and the tendency will be to overly favor allocation to a single cluster.

It is crucial to consider the measurement scale of the data in choosing P_0. The variance of P_0 is only meaningful relative to the scale of the data. A common approach is to standardize the data y and then choose P_0 to be centered at zero with close to unit variance. If we set unit variance and do not standardize, then how flat P_0 is depends on the unit of measurement in the data—if we change from inches to miles, we may get completely different results.

Example. A toxicology application

As an illustrative application, we consider data from a developmental toxicology study of ethylene glycol in mice conducted by the National Toxicology Program. In particular, y_i is the number of implantations in the ith pregnant mouse, with mice assigned to dose groups of 0, 750, 1500, or 3000 mg/kg/day. Group sizes were 25, 24, 23, and 23, respectively. Scientific interest focuses on studying a dose response trend in the number of implants, and we initially focus on separately estimating the distribution within each group. As in many biological applications in which there are constraints on the range of the counts, the data are underdispersed: the mean is 12.5 and the variance is 6.8 in the control group. Figure 23.2 presents a histogram of the raw data for the control group (25 mice), along with a series of estimates of the posterior probabilities $\Pr(y = j)$ assuming $y_i \sim P$ with $P \sim \mathrm{DP}(\alpha P_0)$, $\alpha = 1$ or 5, and P_0 set to Poisson($\overline{y}$) for simplicity.

This approach places a Dirichlet process prior directly on the distribution of the count data instead of using a Dirichlet process mixture. Counts are discrete so this seems like a reasonable initial approach. In addition, when a Dirichlet process is used directly for the distribution of the data, one can rely on the conjugacy property to avoid MCMC. In particular, assuming $y_i \overset{iid}{\sim} P$ for $i = 1, \ldots, 25$ (focusing only on the mice in the control group to start), and $P \sim \mathrm{DP}(\alpha P_0)$, we have

$$P|y_1, \ldots, y_n \sim \mathrm{DP}\Big(\alpha P_0 + \sum_{i=1}^{n} \delta_{y_i}\Big),$$

so that the posterior mean probability of $y = j$ is simply

$$\Pr(y = j|y_1, \ldots, y_n) = \left(\frac{\alpha}{\alpha + n}\right) P_0(j) + \left(\frac{1}{\alpha + n}\right) \sum_{i=1}^{n} 1_{y_i = j},$$

where $P_0(j)$ is the probability of $y = j$ under P_0 in the prior distribution. This expression is simply the weighted average of the prior mean and the proportion of cases where $y = j$ in the observed data, with the weight on the prior being α and the weight on the data being n.

To illustrate the behavior as the sample size increases, we take a random subsample of the data of size 10. As Figures 23.2 and 23.3 illustrate, the lack of smoothing in the nonparametric Bayes estimate under a Dirichlet process prior is unappealing in not allowing borrowing of information about local deviations from P_0. In particular, for small sample sizes as in Figure 23.3, the posterior mean probability mass function

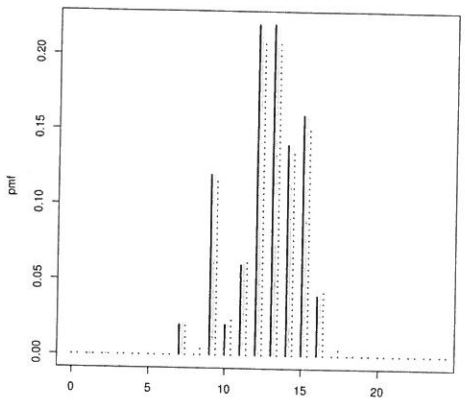

Figure 23.2 *Histogram of the number of implantations per pregnant mouse in the control group (black line) and posterior mean of* $\Pr(y = j)$ *assuming a Dirichlet process prior on the distribution of the number of implants with* $\alpha = 1, 5$ *(gray and black dotted lines, respectively) and base measure* $P_0 = Poisson(\overline{y})$.

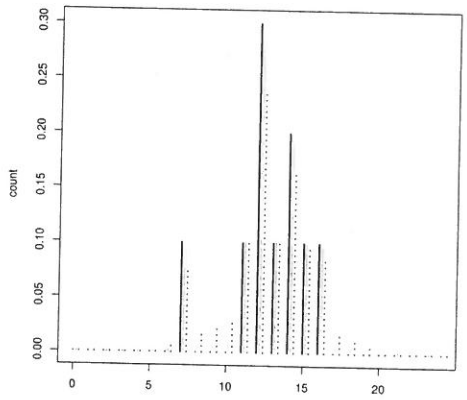

Figure 23.3 *Histogram of a subsample of size* 10 *from the control group on implantation in mice (black line) and posterior mean of* $\Pr(y = j)$ *assuming a DP prior on the distribution of the number of implants with* $\alpha = 1, 5$ *(gray and black dotted lines, respectively) and base measure* $P_0 = Poisson(\overline{y})$.

corresponds to the base measure with high peaks on the observed y. As the sample size increases, the empirical probability mass function increasingly dominates the base. Potentially, by using a Dirichlet process mixture (DPM) instead of a DP directly, one may obtain better performance in practice. For count data, it seems natural to use a Poisson kernel $K(\cdot)$ with a gamma base measure P_0, so that

$$y_i \sim \text{Poisson}(\theta_i), \quad \theta_i \sim P, \quad P \sim \text{DP}(\alpha P_0), \quad P_0 = \text{Gamma}(a, b). \tag{23.7}$$

In this case, we can easily work out the steps involved in the blocked Gibbs sampler.

1. Update $S_i \in \{1, \ldots, N\}$ by multinomial sampling with

$$\Pr(S_i = c \mid -) = \frac{\pi_h \text{Poisson}(y_i|\theta_h^*)}{\sum_{c'=1}^{N} \pi_{c'} \text{Poisson}(y_i|\theta_{c'}^*)}, \quad c = 1, \ldots, N,$$

where $S_i = c$ if $\theta_i = \theta_c^*$ denotes that subject i is allocated to cluster h.

2. Update the stick-breaking weight V_c, $c = 1, \ldots, N-1$, from

$$\text{Beta}\Big(1 + n_c,\ \alpha + \sum_{c'=c+1}^{N} n_{c'}\Big).$$

3. Update θ_h^*, $h = 1, \ldots, N$, from its conditional posterior,

$$\text{Gamma}\Big(a + \sum_{i:S_i=h} y_i,\ b + n_h\Big),$$

with $n_c = \sum_{i=1}^{n} 1_{S_i=c}$, the size of the c-th cluster.

A conservative upper bound of $N = n$ can be used for the truncation level.

Although a Dirichlet process mixture of Poissons is the obvious choice and leads to simple computation, there is a lurking problem with this approach, which is a common issue in hierarchical Poisson models in general. In particular, the Poisson kernel is inflexible in that it restricts the mean and variance to be equal. In using a mixture of Poissons, such as the DPM in (23.7), one can only increase the variance relative to the mean. Hence, mixtures of Poissons are only appropriate for modeling overdispersed count distributions and produce poor results in the toxicology data on implantations. In particular, the estimated dose group-specific distributions of the number of implants under the DPM of Poissons exhibit substantially larger variance than the empirical distributions, suggesting a poor fit.

For continuous data, Gaussian kernels are routinely used and do not have this pitfall in having separate parameters for the mean and variance. Gaussians are easily modified to the count case by relying on rounding. In particular, let

$$y_i = h(y_i^*),\quad y_i^* \sim \text{N}(\mu_i, \tau_i^{-1}),\quad (\mu_i, \tau_i) \sim P,\ i = 1, \ldots, n,\quad P \sim \text{DP}(\alpha P_0), \qquad (23.8)$$

with $h(\cdot)$ a rounding function that has $h(y^*) = j$ if $y^* \in (a_j, a_{j+1}]$ for $j = 0, 1, 2, \ldots, \infty$, $a_0 = -\infty$, $a_j = j - 1$, $j = 1, \ldots, \infty$, and $P_0(\mu, \tau) = N(\mu|\mu_0, \kappa\tau^{-1})\text{Gamma}(\tau|a_\tau, b_\tau)$. For this rounded Gaussian kernel, we can derive a simple blocked Gibbs sampler, which has the following steps.

1. Update $S_i \in \{1, \ldots, N\}$ by multinomial sampling with

$$\Pr(S_i = c\,|\,-) = \frac{\pi_c p(y_i|\mu_c^*, \tau_c^*)}{\sum_{c'=1}^{N} \pi_{c'} p(y_i|\mu_{c'}^*, \tau_{c'}^*)},\quad c = 1, \ldots, N,$$

where $p(y_i|\mu_c^*, \tau_c^*) = \Phi(a_{j+1}|\mu_c^*, \tau_c^*) - \Phi(a_j|\mu_c^*, \tau_c^*)$, and $\Phi(z|\mu, \tau)$ is the normal cumulative distribution function with location μ and precision τ.

2. Update stick-breaking weight V_h, $h = 1, \ldots, N-1$, from

$$\text{Beta}\Big(1 + n_c, \alpha + \sum_{c'=c+1}^{N} n_{c'}\Big).$$

3. Generate each y_i^* from the full conditional posterior

$$u_i \sim \text{U}\big(\Phi(a_{y_i}|\mu_{S_i}^*, \tau_{S_i}^*), \Phi(a_{y_i+1}|\mu_{S_i}^*, \tau_{S_i}^*)\big),\ y_i^* = \Phi^{-1}(u_i|\mu_{S_i}^*, \tau_{S_i}^*).$$

4. Update $\theta_c^* = (\mu_c^*, \tau_c^*)$, $c = 1, \ldots, N$, from its conditional posterior,

$$\text{N}(\mu_c^*|\hat{\mu}_c, \hat{\kappa}_c\tau_c^{-1})\text{Gamma}(\tau_c|\hat{a}_{\tau_c}, \hat{b}_{\tau_c}),$$

with $\hat{a}_{\tau_c} = a_\tau + \frac{n_c}{2}$, $\hat{b}_{\tau_c} = b_\tau + \frac{1}{2}\big(\sum_{i:S_i=c}(y_i^* - \overline{y}_c^*) + \frac{n_c}{1+\kappa n_c}(\overline{y}_c^* - \mu_0)^2\big)$, $\hat{\kappa}_c = (\kappa^{-1} + n_c)^{-1}$ and $\hat{\mu}_c = \hat{\kappa}_c(\kappa^{-1}\mu_0 + n_h\overline{y}_c^*)$.

Essentially, we just impute the latent y_i^* within the third step of the Gibbs sampler and otherwise proceed as if we were modeling the data using a DPM location-scale mixture of Gaussians. In fact, the above steps can also be used for Bayesian density estimation of continuous densities in which the observed data are y_i^* and we have no need for step 3.

We repeated our analysis of the toxicology data on implantations using the DPM of rounded Gaussians approach, and obtained an excellent fit to the data, improving on the DPM of Poissons result. The empirical cumulative distribution functions are entirely enclosed within pointwise 95% credible intervals. To conduct inferences on changes in the distribution of the number of implants with dose, we estimated summaries of the posterior distributions for changes in each percentile between the control group and each of the exposed groups. Negative changes for an exposed group relative to control suggest an adverse impact of dose. The estimated posterior probabilities of a negative average change across the percentiles was 0.72 in the 750 mg/kg group, 0.99 in the 1500 mg/kg group, and 0.94 in the 3000 mg/kg group. Hence, there was substantial evidence of a stochastic decrease in the number of implants in the higher two dose groups relative to control.

23.4 Beyond density estimation

Nonparametric residual distributions

Density estimation has been used to this point primarily to simplify the exposition of a difficult topic. The real attraction of Dirichlet process mixture (DPM) models is that they can be used much more broadly for relaxing parametric assumptions in hierarchical models. This section is meant to give a flavor of some of the possibilities without being comprehensive. First, consider the linear regression setting with a nonparametric error distribution:

$$y_i = X_i\beta + \epsilon_i, \quad \epsilon_i \sim f, \tag{23.9}$$

where $X_i = (X_{i1}, \ldots, X_{ip})$ is a vector of predictors and ϵ_i is an error term with distribution f. The assumption of linearity in the mean is easily relaxed as discussed earlier. Here, we consider the problem of relaxing the assumption that f, the distribution of errors, has a parametric form.

In Chapter 17 we considered the t model as a way to downweight the influence of outliers. This is easily accomplished computationally by expressing the t_ν distribution as a scale mixture of normals by letting $\epsilon_i \sim \mathrm{N}(0, \phi_i^{-1}\sigma^2)$, with $\phi_i \sim \text{Gamma} \sim (\frac{\nu}{2}, \frac{\nu}{2})$. Although the t distribution may be preferred over the normal due to its heavy tails, it still has a restrictive shape and we could instead model f nonparametrically using a DP scale mixture of normals:

$$\epsilon_i \sim \mathrm{N}(0, \phi_i^{-1}), \quad \phi_i \sim P, \quad P \sim \mathrm{DP}(\alpha P_0),$$

where P_0 is chosen to correspond to $\text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$ to center the prior for f on a t distribution, while allowing more flexibility. The resulting prior for f is flexible but is still restricted to be unimodal and symmetric about zero.

An alternative which removes the unimodality and symmetry constraints is to use a location mixture of Gaussians for f. Removing the intercept from the $X_i\beta$ term and allowing f to have an unknown mean, let

$$\epsilon_i \sim \mathrm{N}(\mu_i, \tau^{-1}), \quad \mu_i \sim P, \quad P \sim \mathrm{DP}(\alpha P_0), \quad \tau \sim \text{Gamma}(a_\tau, b_\tau),$$

with P_0 chosen as $\mathrm{N}(0, \tau^{-1})$. The computations for density estimation can be easily adapted to include steps for updating the regression coefficients β and then replacing y_i with $y_i - X_i\beta$ in the previous steps.

Nonparametric models for parameters that vary by group

In Chapter 15 we considered hierarchical linear models with varying coefficients. Uncertainty about the distribution of the coefficients can be taken into account by placing DP or DPM priors on their distributions. As a simple illustration, consider the one-factor Anova model,

$$y_{ij} = \mu_i + \epsilon_{ij}, \quad \mu_i \sim f, \quad \epsilon_{ij} \sim g,$$

with $y_i = (y_{i1}, \ldots, y_{i,n_i})$ a vector of repeated measurements for item i, μ_i a subject-specific mean, and ϵ_{ij} an observation-specific residual. Typical parametric models would let f correspond to a $\mathrm{N}(\mu, \psi^{-1})$ density, while letting $g \equiv \mathrm{N}(0, \sigma^2)$.

To allow more flexibility in characterizing variability among subjects, we can instead let

$$\mu_i \sim P, \quad P \sim \mathrm{DP}(\alpha P_0), \tag{23.10}$$

where P is the unknown distribution of the varying parameters and for simplicity we model the residual density g as $\mathrm{N}(0, \sigma^2)$. Placing a DP prior on the distribution induces a *latent class* model in which the subjects are grouped into an unknown number of clusters, with

$$\mu_i = \mu^*_{S_i}, \quad \Pr(S_i = h) = \pi_h, \quad h = 1, 2, \ldots,$$

where $S_i \in \{1, \ldots, \infty\}$ is a latent class index, and π_h is the probability of allocation to latent class h, with these probabilities following the stick-breaking form as in (23.2). As for finite latent class models, this formulation assumes that the distribution of the varying parameters is discrete so that different subjects can have identical values of the parameters. This may be useful as a simplifying assumption and the posterior means will be different for every subject, since the clustering is *soft* and probabilistic, with the posterior means of μ_i obtained averaging across the posterior distribution on the cluster allocation.

There are some practical questions that arise in considering nonparametric hierarchical models such as (23.10). The first is whether the data contain information to allow nonparametric estimation of P given that the modeled parameters μ_i are not observed directly for any of the subjects. The answer to this question and the interpretation of the resulting estimate depends on the number of observations per subject. Suppose initially that $n_i = 1$ for all subjects. In this case, we do not have any information in the data to distinguish variability among subjects from variability among measurements within a subject. However, under the assumption of normality of the residual density g, we still have substantial information in the data about P in that P accommodates lack of fit of the normal residual distribution. In the general case in which $n_i \geq 1$ and normal g is assumed, P has a dual role in allowing for lack of fit of the normal distribution for the residuals and systematic variability among subjects. When there are many observations per subject, that later role dominates, but when n_i is small interpretation of P needs to take into account the dual roles.

One natural possibility for removing this confounding is to also model g using a Dirichlet process mixture of Gaussians. In this case, the data contain less information about the distribution, and accurate estimation may require a dataset with many observations per item and many items. In the case in which the distribution of the parameters and the residual distribution are both unknown, an identifiability issue does arise in that it is difficult in nonparametric Bayes models to restrict the mean of the distribution to be zero. However, one can run the MCMC analysis for an overparameterized model without restrictions on the means and then post-process to estimate the overall mean and mean-centered parameters and residual densities.

Functional data analysis

In Chapter 21 we discussed Gaussian processes for functional data analysis, where responses and predictors for a subject are not modeled as scalar or vector-valued random variables but instead as random functions defined at infinitely many points. Here we consider a basis function expansion related to the approaches considered in Chapter 20.

Let $y_i = (y_{i1}, \ldots, y_{in_i})$ denote the observations on function f_i for subject i, where y_{ij} is an observation at point t_{ij}, with $t_{ij} \in \mathcal{T}$. Allowing for measurement errors in observations of a smooth trajectory, let

$$y_{ij} \sim \mathrm{N}(f_i(t_{ij}), \sigma^2). \tag{23.11}$$

and

$$f_i(t) = \sum_{h=1}^{H} \theta_{ih} b_h(t), \quad \theta_i = (\theta_{i1}, \ldots, \theta_{iH}),$$

where $b = \{b_l\}_{h=1}^{H}$ is a collection of basis functions and θ_i is a vector of subject-specific basis coefficients. Here, we are assuming a common collection of potential basis functions, but by allowing elements of the θ_i coefficient vectors to be zero or close to zero, we can discard unnecessary basis functions and even accommodate subject-specific basis function selection. In many applications, it is necessary to allow different subjects to have a different basis for sufficient flexibility. By using a common dictionary of bases across subjects, we allow a common backbone from which a hierarchical model for borrowing information can be built.

To borrow information across subjects and model the variability in the individual functions, let

$$\theta_i \sim P,$$

where the H-dimensional distribution P must be specified or modeled. Potentially, we can consider a parametric family in which $P = \mathrm{N}_H(\theta, \Omega)$, with the resulting mean function then corresponding to $\overline{f}(t) = b(t)\theta$, where $b(t) = (b_1(t), \ldots, b_H(t))$. This mean function provides a population-averaged curve. In addition, the hierarchical covariance matrix Ω characterizes heterogeneity among the subjects in their functions. There are several practical issues that arise with this parametric hierarchical model. Firstly, the number of basis functions p is typically moderate to large, and hence Ω will have many parameters and it can be difficult to reliably estimate all these parameters. In addition, there is no allowance for basis function selection through shrinking the basis coefficients to zero. Although one could potentially choose a prior for Ω that allows diagonal elements close to zero, this would discard the corresponding basis functions for all subjects and does not accommodate subject-specific selection. Finally, the normality assumption for the varying parameters implies a restrictive type of variability across subjects; for example, it cannot accommodate sub-populations of subjects having different functions and outlying functions.

An alternative is to use a Dirichlet process prior, $P \sim \mathrm{DP}(\alpha P_0)$. This will induce functional clustering with

$$f_i(t) = f^*_{S_i}(t), \quad f^*_h(t) = b(t)\theta^*_c, \quad \Pr(S_i = c) = \pi_c, \quad \theta^*_c \sim P_0,$$

All individuals within cluster c will have $f_i(t) = f^*_c(t)$, with the basis coefficients characterizing the cluster c function being $\theta^*_c = (\theta^*_{c1}, \ldots, \theta^*_{cH})$. By choosing an appropriate base measure P_0 within the *functional DP*, one can allow the basis functions to differ across the clusters and hence allow individual-specific basis selection through cluster-specific basis selection.

There are two good possibilities in this regard. Firstly, one can let $P_0 = \bigotimes_{h=1}^{H} P_{0h}$, with P_{0h} specified to have a variable selection-type mixture form:

$$P_{0h}(\cdot) = \pi_{0h}\delta_0(\cdot) + (1 - \pi_{0h})\mathrm{N}(\cdot|0, \psi_h^{-1}),$$