Chapter 3

Dirichlet Process

3.1. The Dirichlet Process Prior

3.1.1. Definition

The Dirichlet process (DP) is arguably the most popular BNP model for random probability measures (RPM), and plays a central role in the literature on RPMs, appearing as a special case of a number of other more general models (recall our discussion in Chapter 1). Hence, the DP can be characterized in a number of different ways.

The original definition of the DP is due to Ferguson (1973), who considered a probability space $(\Theta, \mathcal{A}, G)$ and an arbitrary partition $\{A_1, \ldots, A_k\}$ of Θ. A random distribution G is said to follow a Dirichlet process prior with baseline probability measure G_0 and mass parameter M, denoted $G \sim \mathsf{DP}(M, G_0)$, if

$$(G(A_1), \ldots, G(A_1)) \sim \mathsf{Dir}(MG_0(A_1), \ldots, MG_0(A_k)). \tag{3.1}$$

This collection of finite dimensional distributions implies a well defined infinite dimensional model $p(G)$ because they satisfy Kolmogorov's consistency conditions; proving this fact is one of the main focuses of Ferguson's original paper.

An alternative definition of the DP, known as the "stick-breaking" construction, is provided in Sethurman (1994). Let $\delta_\theta(\cdot)$ denote a point mass at θ. An RPM

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\widetilde{\theta}_h}(\cdot) \tag{3.2}$$

has a $\mathsf{DP}(M, G_0)$ prior if $(\widetilde{\theta}_h)$ are i.i.d. samples from G_0 and $w_h = v_h \prod_{k<h}\{1-v_k\}$ with $v_h \sim \mathsf{Beta}(1, M)$, i.i.d. This constructive definition of the DP is extremely useful for extending the model to more complex problems (see for example Chapter 5) and to highlight important properties of the model. Implicit in (3.2) is the fact that G is discrete, even if G_0 is a continuous distribution.

Recall that the DP also induces a species sampling model. In particular, let $\theta_1, \theta_2, \ldots$ be an i.i.d. sequence such that $\theta_i \mid G \sim G$ where $G \sim \mathsf{DP}(M, G_0)$. Since G is almost surely discrete, there will be ties among the θ_is; let k_n be the number of unique values among $\{\theta_1, \ldots, \theta_n\}$, let $\{\theta_1^\star, \ldots, \theta_{k_n}^\star\}$ be these unique values and let n_{nj} be the number of draws among $\{\theta_1, \ldots, \theta_n\}$ that are equal to $\theta_j^\star$. Blackwell and MacQueen (1973) showed that the joint distribution of the θ_is can be characterized in terms of the predictive probability function

$$p(\theta_{n+1} \mid \theta_n, \ldots, \theta_1) \propto \sum_{j=1}^{k_n} n_{nj}\delta_{\theta_j^\star} + MG_0, \tag{3.3}$$

that is, a new θ_i is identical to a previously observed $\theta_j^\star$ with probability proportional to n_{nj} (i.e., how many times that value has been observed) or a new value sampled from the baseline measure with probability proportional to the total mass parameter M. The predictive distribution (3.3) is exactly in the format of (1.1). After integrating G, the observations are exchangeable and have identical marginal distribution G_0, but are not independent.

The allocation process associated with the predictive distribution in (3.3) is also known as the Pólya urn. Consider an urn that initially has M black balls and one colored ball (whose "color" is randomly selected according to G_0). We sequentially draw balls from the urn; if a colored ball is drawn then we returned it to the urn along with another ball of the same color, if a black ball is drawn, we returned it to the urn along with a ball of a new color randomly selected according to G_0. Another metaphor, the Chinese restaurant process (CRP), is popular in the machine learning community and essentially describes the same model.

For another characterizing property, recall that the DP can be characterized as an NRMI. In particular, let μ be a standard Gamma process on Θ with intensity function $\lambda(\cdot) = MG_0(\cdot)$, i.e., $\mu(A) \sim \mathsf{Gamma}(MG_0(A), 1)$ for any $A \subset \Theta$. Then $G(\cdot) \equiv \mu(\cdot)/\mu(\Theta) \sim \mathsf{DP}(M, G_0)$. This follows from (3.1) and the construction of Dirichlet random variables as normalized Gamma random variates (see, for example Robert and Casella, 2005).

Finally, recall that the DP can be characterized as a PPM with cohesion function $c(S_j) = M(n_j - 1)$, as a special case of the PT with $\alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$, and as a NTR process.

3.1.2. Properties

Since the Dirichlet process places a distribution on the random measure G, the quantity $G(A)$ for any $A \subset \Theta$ is a random variable. From Ferguson's definition we have $G(A) \sim \mathsf{Beta}\{MG_0(A), M(1 - G_0(A))\}$. Hence

$$\mathsf{E}\{G(A)\} = G_0(A), \qquad \mathsf{Var}\{G(A)\} = \frac{G_0(A)\{1 - G_0(A)\}}{M + 1}.$$

This means that we can interpret G_0 as the expected shape of the random distribution G, while M controls the variability of the realizations around G_0.

To further clarify this interpretation of the parameters of the DP, we plot in Figure 3.1 realizations from DPs with standard normal baseline measure and different values of M. The random distributions G are discrete with probability one. We therefore use the c.d.f. to display the random distributions. Larger values of M reduce the variability of the realizations of the process, and for small values of M a small number of weights concentrate most of the probability mass, i.e., a few large steps dominate the cdf. Indeed, a priori, the size of the weights decreases geometrically,

$$\mathsf{E}(w_h) = \frac{1}{M + 1}\left(\frac{M}{M + 1}\right)^{h-1}.$$

A particularly appealing property of the Dirichlet process is its conjugacy under i.i.d. sampling. If $\theta_1, \ldots, \theta_n$ is an i.i.d. sample with $\theta_i \mid G \sim G$ and $G \sim \mathsf{DP}(M, G_0)$ then

$$G \mid \theta_1, \ldots, \theta_n \sim \mathsf{DP}\left(M + n, \frac{MG_0 + \sum_{i=1}^n \delta_{\theta_i}}{M + n}\right). \tag{3.4}$$

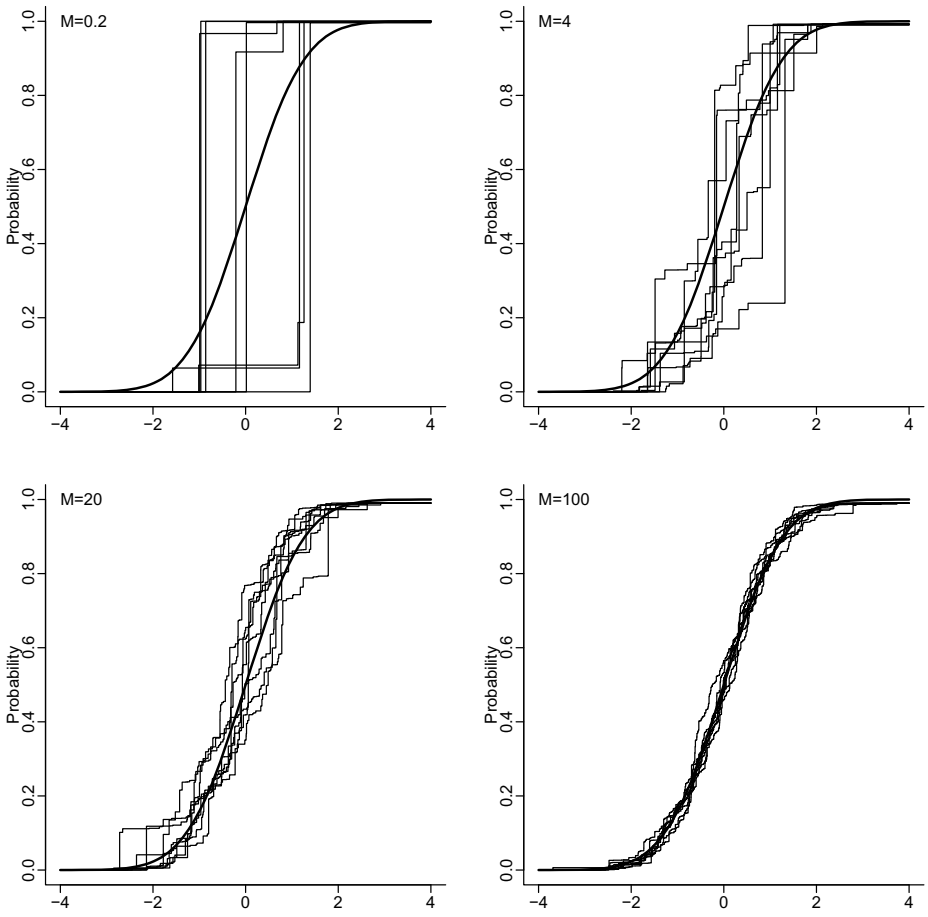

FIG 3.1. *Random distributions generated from a Dirichlet process prior with varying precision parameters M. In all cases, the baseline measure corresponds to a standard normal distribution (thick black curve). Each box contains 8 independent realizations (grey curves) with a common value for M. Note how M controls not only the variability of the realizations around G_0, but also the relative size of the jumps.*

The posterior mean,

$$\mathsf{E}(G \mid \theta_1, \ldots, \theta_n) = \frac{MG_0 + \sum_{i=1}^n \delta_{\theta_i}}{M+n},$$

can be interpreted a weighted average between the baseline measure G_0 and the empirical distribution $\frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$. In addition, since the empirical cdf is a consistent estimator if the θ_is are indeed i.i.d. from some true distribution G_T, it is easy to show from (3.4) that, as $n \to \infty$, we have $G(A) \mid \theta_1, \ldots, \theta_n \xrightarrow{P} G_T(A)$ for any measurable set A.

Example 9 (DP Nonparametric density estimation) *We carry out a simulation study with $\theta_i \sim G$, $i = 1, \ldots, n$, independently. We generate two datasets, with $n = 8$ and $n = 50$ observations, respectively, from the true model $G = \mathsf{N}(2, 4)$.*

In both cases, we pretend that G is unknown and carry out density estimation under a BNP prior, $G \sim \mathsf{DP}(M, G_0)$ with $G_0 = \mathsf{N}(0, 1)$ and total mass parameter $M = 5$.

Figure (3.2) shows the simulation truth, the empirical distribution, and the posterior mean $E(G \mid \boldsymbol{\theta})$ under the Dirichlet process prior. We see how the posterior mean is a weighted average of the prior mean and the empirical distribution of the observed data. Note that the posterior distribution converges relatively quickly to the empirical cdf as the sample size grows.

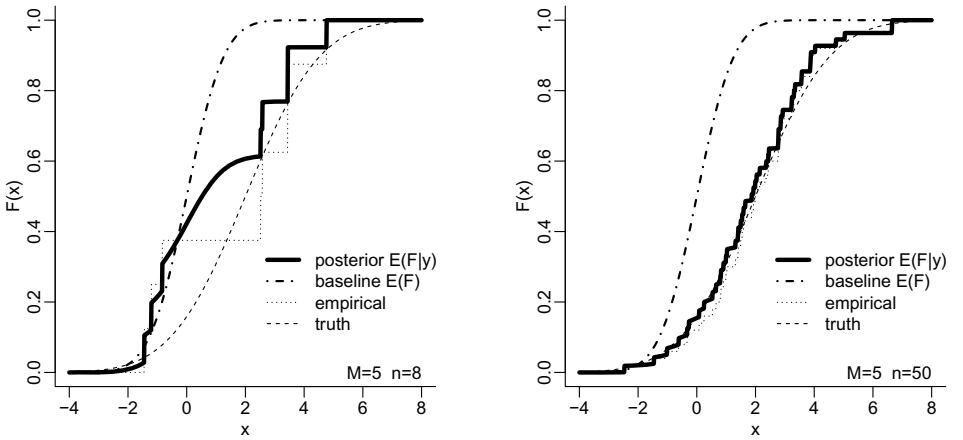

FIG 3.2. *An example of nonparametric density estimation using Dirichlet process priors. Two independent samples with sizes $n = 8$ and $n = 50$ where generated from a normal $\mathsf{N}(2, 4)$ distribution (whose c.d.f. is shown as the dashed black line "truth"). In both cases, the prior precision parameter is $M = 5$, while the baseline measure is a standard normal distribution (dashed dotted, "baseline"). The empirical CDF (dotted step function) and posterior mean (thick black line) are also shown.*

Finally, we discuss some properties of a random sample from the DP that follow from the Pólya urn representation of the process. As mentioned in Chapter 1, the predictive probability function in (3.3) implies a probability model for any partition of the experimental units into clusters $S_j = \{i : \theta_i = \theta_j^\star\}$, i.e., into clusters defined by the ties among the draws θ_i. Recall that we used $\boldsymbol{n} = (n_1, \ldots, n_k)$ for the cluster sizes for a partition of n experimental units into clusters S_j, $j = 1, \ldots, k$. The probability model for $(k, \boldsymbol{n})$ implied by (3.3) can easily be determined as

$$p(k, n_1, \ldots, n_k) = \frac{\Gamma(M)}{\Gamma(M+n)} M^k \prod_{j=1}^{k} \Gamma(n_j). \tag{3.5}$$

The model $p(k, \boldsymbol{n})$ for a random partition is known as the exchangeable product partition function (EPPF). From this EPPF we can obtain the probability mass function for the number of unique values k_n (Antoniak, 1974),

$$p(k_n) = S_{n,k} n! M^{k_n} \frac{\Gamma(M)}{\Gamma(M+n)}, \tag{3.6}$$

where $S_{n,k}$ is the unsigned Stirling number of the first kind. Using a conditional

expectation argument we find

$$\mathsf{E}(k) = \sum_{i=1}^{n} \frac{M}{M+i-1} \approx M \log\left(\frac{M+n}{M}\right)$$

for large n. Another consequence of (3.6) is that the partitions favored by the DP are very uneven, i.e., the DP favors partitions with a small number of large clusters and a large number of smallish ones. This feature of the model is often inappropriate in applications, which has motivated many of the generalizations that we discuss in later chapters.

3.2. DP Mixtures

The discrete nature of the DP random measures is awkward when the unknown distribution is known to be continuous. Even worse, for some hierarchical models the Dirichlet process prior can lead to inconsistent estimators if the true distribution is continuous (for examples, see Diaconis and Freedman, 1986a,b). One way to mitigate this limitation of the DP is to add to the discrete distribution G a convolution with a continous kernel. This is similar in spirit to kernel density estimators, where the empirical distribution is smoothed by convoluting it with an appropriate kernel.

Let $y_1, y_2, \ldots$ be an i.i.d. sample with unknown distribution F. A Dirichlet process mixture prior (DPM) on F posits that

$$y_i \sim F(y_i) = \int p(y_i \mid \theta) G(\mathrm{d}\theta), \qquad G \sim \mathsf{DP}(M, G_0), \tag{3.7}$$

where $p(y_i \mid \theta)$ is a parametric distribution (often referred to as the kernel of the mixture), which is indexed by a finite dimensional parameter θ. For example, in a DP location mixture of normals we have

$$y_i \mid G \sim \int \mathsf{N}(y_i \mid \mu, \sigma^2) G(\mathrm{d}\mu), \qquad G \sim \mathsf{DP}(M, G_0).$$

Figure 2.1 illustrates a DP mixture of normal model.

The model in (3.9) can be represented in a number of alternative ways. Exploiting the stick-breaking construction of the Dirichlet process we can write

$$y_i \mid (w_h), (\widetilde{\theta}_h) \sim \underbrace{\sum_{h=1}^{\infty} w_h p(y_i \mid \widetilde{\theta}_h)}_{F(y_i)}, \tag{3.8}$$

where

$$\widetilde{\theta}_h \sim G_0, \qquad w_h = v_h \prod_{k<h} \{1 - v_k\}, \qquad v_h \sim \mathsf{Beta}(1, M).$$

This representation highlights the nature of the DP mixture model as a discrete mixture. DP mixtures are countable mixtures with an infinite number of components and a specific prior on the weights and the component-specific parameters. Working with an infinite number of components is particularly appealing because it ensures that, for appropriate choices of the kernel $p(y_i \mid \theta)$, the DPM model has

support on a large classes of distributions. For example, Lo (1984) showed that a DP location-scale mixture of normals,

$$y_i \mid G \sim \int \mathsf{N}(y_i \mid \mu, \sigma^2) G(\mathrm{d}\mu, \mathrm{d}\sigma^2), \qquad G \sim \mathsf{DP}(M, G_0),$$

has full support on the space of absolutely continuous distributions. Similarly, a mixture of uniform distributions

$$y_i \sim \int \mathsf{Uni}(y_i \mid -\theta, \theta) G(\mathrm{d}\theta), \qquad G \sim \mathsf{DP}(M, G_0),$$

where $\mathsf{Uni}(x \mid a, b)$ indicates a random variable x with a uniform distribution on $[a, b]$, has full support on the space of all unimodal symmetric distributions.

Another consequence of (3.8) is that the DPM induces clustering among the observations, with M controlling the a priori expected number clusters in the sample. In particular, note that if $M \to 0$, the model reduces to a single component mixture where all observations are i.i.d. from $p(y \mid \theta)$ and $\theta \sim G_0$, i.e., a fully parametric model. On the other hand for $M \to \infty$ each observation is assigned its own singleton cluster and we have $y_i \sim \int p(y_i \mid \theta) G_0(\mathrm{d}\theta)$, i.i.d. Nothing is unknown about the sampling model for y_i.

An alternative representation for (3.9) introduces latent random effects (θ_i) to replace the mixture by a hierarchical model

$$y_i \mid \theta_i \sim p(y_i \mid \theta_i), \qquad \theta_i \mid G \sim G, \qquad G \sim \mathsf{DP}(M, G_0). \tag{3.9}$$

The hierarchical model (3.9) also highlights the nature of clusters generated by ties among the θ_i that arise under sampling from the discrete probability measure G.

Or, integrating out the random measure G,

$$y_i \mid \theta_i \sim p(y_i \mid \theta_i), \qquad (\theta_1, \ldots, \theta_n) \sim p(\theta_1, \ldots, \theta_n),$$

where the joint distribution $p(\theta_1, \ldots, \theta_n)$ is implicitly defined by the sequence of predictive distributions in (3.3). As before, denote by $(\theta_j^\star)$ the unique values among $\theta_1, \ldots, \theta_n$ and introduce indicator variables (s_i) such that $\theta_i = \theta_{s_i}^\star$. Then we can further rewrite the model as

$$y_i \mid s_i, (\theta_j^\star) \sim p(y_i \mid \theta_{s_i}^\star), \quad \theta_j^\star \sim G_0, \quad p(s_1, \ldots, s_n) = \frac{\Gamma(M)}{\Gamma(M+n)} M^k \prod_{j=1}^{k} \Gamma(n_j),$$

where k is the number of distinct values among $s_1, \ldots, s_n$ and $n_j = \sum_i I(s_i = j)$ is the number of s_is that are equal to j. By creating the implied clusters the DPM places a prior distribution on all possible partitions of the data into at most n groups. This is precisely the probability model stated in (3.5).

The last two representations marginalize with respect to the infinite dimensional G. Hence, they are particularly useful for the development of computational tools for the DP (see §3.3.1). Finally, we note that although the mixture in (3.8) has infinitely many terms, for any finite sample size n, at most n distinct $\widetilde{\theta}$ are sampled as $\theta_j^\star$.

3.3. Posterior Simulation for DP Mixture Models

One of the attractive features of the Dirichlet process mixture model is that a number of simulation-based algorithms are available for posterior inference. In this

section we review some of the most commonly used algorithms. Many can be extended to other nonparametric models with just minor modifications. Throughout this section we assume the model

$$y_i \mid \theta_i \sim p(y_i \mid \theta_i), \qquad \theta_i \mid G \sim G(\theta_i), \qquad G \sim \mathsf{DP}(M, G_0). \tag{3.10}$$

That is, a DP mixture model with kernel $p(y_i \mid \theta_i)$ and unknown mixing measure G which follows a Dirichlet process prior.

3.3.1. Collapsed Gibbs Samplers

Conjugate models

Collapsed Gibbs samplers exploit the representation of the DP as a SSM that was discussed in §1.2.1 and §3.1. The first version of this algorithm was developed in Escobar (1988), well before Gibbs samplers were widely used in the statistics literature. Recall the notation from §3.1.1 with $\theta_j^\star$, $j = 1, \ldots, k_{n-1}$ denoting the unique values among $\{\theta_1, \ldots, \theta_{n-1}\}$, $n_{n-1,j}$ denoting the number of θ_i equal to $\theta_j^\star$, and (s_i) denoting the cluster membership indicators with $s_i = j$ if $\theta_i = \theta_j^\star$. Then

$$\theta_n \mid \theta_{n-1}, \ldots, \theta_1 \sim \sum_{j=1}^{k_{n-1}} \frac{n_{n-1,j}}{M+n-1}\delta_{\theta_j^\star} + \frac{M}{M+n-1}G_0.$$

Since sequences generated by a species sampling model are exchangeable, this expression gives us the form of the full conditional prior distribution for any θ_i given $\boldsymbol{\theta}_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$. To see this, just permute the order of the observations so that θ_i becomes the last observation in the sequence. Multiplying by the likelihood $p(y_i \mid \theta_i)$ we find the full conditional posterior distribution for θ_i

$$\begin{aligned}\theta_i \mid \boldsymbol{\theta}_{-i}, \mathbf{y} &\propto \sum_{j=1}^{k^-} n_j^- p(y_i \mid \theta_j^{\star -})\delta_{\theta_j^{\star -}} + M\, p(y_i \mid \theta_i)G_0(\theta_i) \\ &= \sum_{j=1}^{k^-} \left\{n_j^- p(y_i \mid \theta_j^{\star -})\right\}\delta_{\theta_j^{\star -}} + \left\{M \int p(y_i \mid \theta_i) dG_0(\theta_i)\right\} p(\theta_i \mid y_i, G_0),\end{aligned} \tag{3.11}$$

where the superscript $^-$ represents the appropriate quantity with θ_i excluded from the sample. In the last term, $p(\theta_i \mid y_i, G_0) = p(y_i \mid \theta_i)dG_0(\theta_i)/\int p(y_i \mid \theta_i)dG_0(\theta_i)$ is the posterior on θ_i in a singleton cluster, and $\int p(y_i \mid \theta_i)dG_0(\theta_i)$ is the (prior) marginal distribution for y_i under G_0.

The previous results lead to a Gibbs sampler for (θ_i) that proceeds by iteratively sampling each θ_i, which is either equal to one of the unique $\theta_j^\star$s with probability proportional to $n_j^- p(y_i \mid \theta_j^\star)$, or sampled from the posterior distribution based solely on y_i with probability $M \int p(y_i \mid \theta_i)dG_0(\theta_i)$.

The described algorithm tends to mix very slowly when the mixture components are well separated. A faster mixing Markov chain is achieved by including an additional transition probability. Noting that sampling θ_i implies a new value for s_i too, the complete conditional posterior probability (3.11) can be characterized as $p(\theta_i, s_i \mid \mathbf{s}^-, \boldsymbol{\theta}^{\star -}, \mathbf{y})$. A more efficient sampler proceeds by first sampling the indicators from $p(s_i \mid \mathbf{s}^-, \mathbf{y})$ sequentially, and then sampling each $\theta_j^\star$ from $p(\theta_j^\star \mid \mathbf{y}, \mathbf{s})$.

To find $p(s_i \mid \mathbf{s}^-, \mathbf{y})$ note first that (3.11) can be written as a hierarchical model with

$$p(s_i = j \mid \mathbf{s}^-, \boldsymbol{\theta}^{\star -}, \mathbf{y}) \propto \begin{cases} n_j^- p(y_i \mid \theta_j^{\star -}) & j = 1, \ldots, k^- \\ M \int p(y_i \mid \theta_i) dG_0(\theta_i) & j = k^- + 1 \end{cases}$$

and

$$p(\theta_i \mid s_i = j, \mathbf{s}^-, \boldsymbol{\theta}^{\star -}, \mathbf{y}) = \begin{cases} \delta_{\theta_j^{\star -}} & j = 1, \ldots, k^- \\ p(\theta_i \mid y_i, G_0) & j = k^- + 1. \end{cases} \tag{3.12}$$

Marginalizing w.r.t. θ_i, but still conditioning on $\boldsymbol{\theta}^{\star -}$, we simply drop the last line. Let $\boldsymbol{y}_j^{\star -} = (y_\ell;\ s_\ell = j \text{ and } \ell \neq i)$ denote the observations in the j-th cluster without y_i. Finally, we remove $\theta_j^{\star -}$ from the conditioning set by integrating with respect to $p(\theta_j^{\star -} \mid \mathbf{s}^-, \mathbf{y}) = p(\theta_j^{\star -} \mid \boldsymbol{y}_j^{\star -})$ and get:

$$p(s_i = j \mid \mathbf{s}^-, \mathbf{y}) \propto \begin{cases} n_j^- \int p(y_i \mid \theta_j^{\star -}) dp(\theta_j^{\star -} \mid \boldsymbol{y}_j^{\star -}) & j \leq k^- \\ M \int p(y_i \mid \theta_i) dG(\theta_i) & j = k^- + 1. \end{cases} \tag{3.13}$$

The full conditional posterior for $\theta_j^\star$ is proportional to

$$p(\theta_j^\star \mid \mathbf{s}, \mathbf{y}) \propto G_0(\theta_j^\star) \prod_{\{i : s_i = j\}} p(y_i \mid \theta_j^\star). \tag{3.14}$$

When $G_0(\theta)$ is conjugate to $p(y_i \mid \theta)$, all of $\int p(y_i \mid \theta_j^{\star -}) dp(\theta_j^{\star -} \mid \boldsymbol{y}_j^{\star -})$, $\int p(y_i \mid \theta_i) dG_0$, and $p(\theta_j^\star \mid \mathbf{s}, \mathbf{y})$ are usually available in closed form and implementation of the algorithm is straightforward.

Example 10 (DPM with Gaussian kernels) *Consider a location mixture of Gaussian kernels, with $p(y_i \mid \theta_i) = \mathsf{N}(\theta_i, \sigma^2)$, and a conjugate baseline measure $G_0 = \mathsf{N}(m, B)$. In that case,*

$$\int p(y_i \mid \theta_i) dG_0(\theta_i) = \mathsf{N}(y_i \mid m, B + \sigma^2),$$

while

$$\int p(y_i \mid \theta_j^{\star -}) dp(\theta_j^{\star -} \mid \boldsymbol{y}_j^{\star -}) = \mathsf{N}(y_i \mid m_j^-, V_j^- + \sigma^2),$$

with $1/V_j^- = 1/B + n_j^-/\sigma^2$ and $m_j^- = V_j^-(m/B + 1/\sigma^2 \sum_{h \in S_j^-} y_h)$. Here $S_j^- = \{h \neq i : \ s_h = j\}$. Also, $p(\theta_j^\star \mid \mathbf{y}) = \mathsf{N}(m_j, V_j^2)$ with the expressions for m_j and V_j being the formulas for m_j^- and V_j^-, but now without excluding the i-th observation.

Non-conjugate models

When $p(y \mid \theta)$ and $G_0(\theta)$ are not conjugate, the integral $\int p(y_i \mid \theta_i) dG_0(\theta_i)$ is often analytically intractable. In that case the collapsed Gibbs samplers require that the integral be approximated numerically, making the implementation inefficient, particularly in high dimensional problems. To overcome this issue, a number of alternative collapsed samplers have been devised to accommodate non-conjugate models; a common feature of most of these methods is that they replace the predictive $\int p(y_i \mid \theta_i) dG_0(\theta_i)$ by $p(y_i \mid \theta_{k^-+1}^\star)$, where $\theta_{k^-+1}^\star$ is a random draw from G_0.

This can be justified on the basis of on an auxiliary probability model that includes the values of the parameters $\theta^\star_{k^-+1}, \ldots, \theta^\star_n$ for hypothetical empty clusters.

We start by describing the "no-gaps" algorithm introduced by MacEachern and Müller (1998). The name derives from the fact that the description of the algorithm relies on the no gaps convention, i.e., occupied clusters are consecutively labeled from 1 to k. As before, the algorithm proceeds by first sampling the indicators (s_i) conditional on $(\theta^\star_j)$, and then sampling the component-specific $(\theta^\star_j)$ conditional on (s_i).

To sample s_i we consider two cases. If in the currently imputed state $n_{s_i} > 1$ then we sample s_i from

$$p(s_i = j \mid \theta^\star_1, \ldots, \theta^\star_n, \mathbf{y}) \propto \begin{cases} n_j^- p(y_i \mid \theta^\star_j) & j = 1, \ldots, k^- \\ \frac{M}{k^-+1} p(y_i \mid \theta^\star_j) & j = k^- + 1. \end{cases} \tag{3.15}$$

On the other hand, if $n_{s_i} = 1$, i.e., y_i is currently forming a singleton cluster on its own, then with probability $(k^- - 1)/k^-$ we leave s_i unchanged, and with probability $1/k^-$ we resample s_i according to (3.15). See MacEachern and Müller (1998) for a justification.

Given the indicators (s_i), the component specific parameters $\theta^\star_1, \ldots \theta^\star_n$ are conditionally independent and can be sampled from the full conditional in (3.14). Since the model is not conjugate, this might require Metropolis-Hastings steps to sample $p(\theta^\star_j \mid \mathbf{s}, \mathbf{y})$. Cluster-specific $\theta^\star_{k+1}, \ldots,$ for hypothetical future clusters are sampled from G_0. However in actual implementation, when G_0 is a distribution for which a direct sampler is available, then we do not need to store the values $\theta^\star_{k+1}, \ldots, \theta^\star_n$, as they can be generated when and as needed to evaluate (3.15). However, one detail in the described MCMC is the following implication. Updating s_i in (3.15) might create new empty clusters when s_i is moved from a current singleton cluster, say $s_i = j_0$, to another existing cluster, $s_i = j \neq j_0$. The currently imputed $\theta^\star_{j_0}$ for the now empty cluster j_0 remains unchanged. In particular, it is not replaced by a draw from G_0.

The "no gaps" algorithm is easy to implement, but mixes slowly due to the reduced probability of opening new clusters. More general algorithms were proposed by Neal (2000), who noted that the joint posterior of the any set of parameters (s_i) and $(\theta^\star_j)$ can be evaluated as

$$p((s_i), (\theta^\star_j) \mid \mathbf{y}) \propto p(s_1, \ldots, s_n) \prod_{j=1}^{k} G_0(\theta^\star_j) \prod_{i=1}^{n} p(y_i \mid \theta^\star_{s_i}), \tag{3.16}$$

where $p(s_1, \ldots, s_n) \propto M^{k-1} \prod_{j=1}^{k} (n_j - 1)!$ (recall equation (3.5)). In principle, this joint distribution can be combined with any reversible proposal to develop Metropolis-Hastings transition probabilities for DP mixture models. As one example, consider making proposals $\widetilde{\theta}_i$ for θ_i (and thus implicitely for s_i) by a draw from the prior conditional:

$$p(\widetilde{\theta}_i) \propto \sum_j n_j^- \delta_{\theta^\star_j -}(\theta_i) + M\, G_0(\theta_i).$$

The acceptance probability is

$$\pi = \min \left\{ 1, \frac{p(y_i \mid \theta_{\widetilde{s}_i})}{p(y_i \mid \theta_{s_i})} \right\}.$$

Then, as in the "no gaps" algorithm, all the of the component parameters $\theta_j^\star$s can be resampled according to (3.14). Other variations of this approach are described in Neal (2000).

Random baseline measures

The DP mixture in (3.10) is often extended by assigning prior distributions for M or for hyperparameters ϕ that index G_0. Since M implicitly controls the number of clusters, a prior on M allows us to introduce prior uncertainty about the distribution of the number of clusters k. Similarly, a prior on G_0 allows us to reflect uncertainty on aspects of the distribution such as the "closeness" or the "size" of the clusters.

Consider a baseline measure that is indexed with hyperparameters ϕ, $G_0(\theta \mid \phi)$, and augment the model with a hyperprior $p(\phi)$ on ϕ. Note that, from the definition of the Dirichlet process, the values of $\theta_1^\star, \ldots, \theta_k^\star$ are independent draws rom G_0. Hence, the full conditional posterior for ϕ is simply

$$p(\phi \mid \ldots) \propto p(\phi) \prod_{j=1}^{k} G_0(\theta_j^\star \mid \phi).$$

When $G_0(\theta \mid \phi)$ and $p(\phi)$ are chosen as a conjugate pair, this posterior reduces to a well known distributions.

Example 11 (DPM with Gaussian kernels, continued) *Consider again the location mixture of Gaussian kernels from Example 10. The base measure G_0 is indexed by $\phi = (m, B)$. It is natural to extend the model with conditionally conjugate priors $p(m) = \mathsf{N}(m_0, D)$ and $p(B) = \mathsf{IGamma}(a, b)$, where $\mathsf{IGamma}(a, b)$ denotes the inverse Gamma distribution with shape parameter a and mean $b/(a-1)$ for $a > 1$. We can interpret m as representing the center of mass for the cluster locations, while B represents the average distance between cluster centers. The full conditional distributions associated with m reduce to*

$$m \mid \ldots \sim \mathsf{N}(m_1, D_1),$$

with $D_1^{-1} = 1/D + k/B$ and $m_1 = D_1(m_0/D + 1/B \sum_{j=1}^{k} \theta_j^\star)$. On the other hand, the full conditional posterior distribution for B is simply

$$B \mid \ldots \sim \mathsf{IGamma}(a_1, b_1),$$

with $a_1 = a + k/2$ and $b_1 = b + \sum_{j=1}^{k} (\theta_j^\star - m)^2/2$.

On the other hand, to estimate the precision parameter M we can use (3.6),

$$p(k \mid M) \propto M^k \frac{\Gamma(M)}{\Gamma(M+n)} = M^k \frac{(M+n)}{M\Gamma(n)} \int_0^1 \eta^M (1-\eta)^{n-1} d\eta.$$

The last equality exploits the normalizing constant of a $\mathsf{Be}(M+1, n)$ beta distribution. Therefore, we can devise a sampler for M by first introducing a latent variable η such that

$$\eta \mid M, k, \ldots \sim \mathsf{Beta}(M+1, n).$$

If, a priori, $M \sim \mathsf{Gamma}(c, d)$, then also

$$M \mid \eta, k, \ldots \sim \frac{c+k-1}{c+k-1+n(d-\log\{\eta\})}\mathsf{Gamma}(c+k, d-\log\{\eta\}) + \frac{n(d-\log\{\eta\})}{c+k-1+n(d-\log\{\eta\})}\mathsf{Gamma}(c+k-1, d-\log(\eta)).$$

This clever auxiliary variable sampler was first introduced in Escobar and West (1995).

3.3.2. Slice Samplers

Slice samplers for DP mixture models were introduced in Walker (2007). Unlike collapsed samplers, slice samplers do not marginalize over G, but use the stick-breaking representation of the process. The discrete nature of $G \sim \mathsf{DP}(M, G_0)$ allows us to write the DPM model as

$$p(y_i \mid (w_h), (\widetilde{\theta}_h)) = \int p(y_i \mid \theta_i) dG(\theta_i) = \sum_{h=1}^{\infty} w_h p(y_i \mid \widetilde{\theta}_h). \tag{3.17}$$

This expression is computationally intractable because of the infinite sum. However, a clever model augmentation with latent variables $u_1, \ldots, u_n$, $0 \le u_i \le 1$, reduces (3.17) to a finite sum. Consider the augmented model

$$p(y_i, u_i \mid (w_h), (\widetilde{\theta}_h)) = \sum_{h=1}^{\infty} I(u_i < w_h) p(y_i \mid \widetilde{\theta}_h), \tag{3.18}$$

where $I(A)$ denotes the indicator function on the set A. Integating w.r.t. u_i reduces the model again to (3.17), as desired. The important trick is that we have $w_h > u_i$ only for a finite number of weights. Hence, conditioning on the latent variables (u_i) has the effect of transforming the infinite mixture into a finite mixture with a fixed number $N_u = \sum_h I(u_i < w_h)$ of components. We agument the model a second time with latent indicators $r_i \in \{1, 2, \ldots\}$ to

$$p(y_i, u_i, r_i \mid (w_h), (\widetilde{\theta}_h)) = I(u_i < w_{r_i}) p(y_i \mid \widetilde{\theta}_{r_i}). \tag{3.19}$$

Marginalizing w.r.t. r_i we immediately get the sum in (3.18), while integrating over r_i and u_i yields (3.17), as desired. The joint distribution of the data, the latent indicators r and u in the extended model is

$$p(\mathbf{y}, \mathbf{u}, \mathbf{r} \mid (w_h), (\widetilde{\theta}_h)) = \prod_{i=1}^{n} I(u_i < w_{r_i}) p(y_i \mid \widetilde{\theta}_{r_i}). \tag{3.20}$$

Note that the indicators r_i in (3.20) are different from the indicators s_i in (3.2). The latter are cluster membership indicators that match θ_i with the unique values $\theta_j^\star$. The earlier are indicators that match θ_i with the point masses $\widetilde{\theta}_h$ in (3.2). However, the two are related because the $\theta_j^\star$ are a sample of $\widetilde{\theta}_h$. With another set of indicators, $t_j = h$ when $\theta_j^\star = \widetilde{\theta}_h$ we would have $r_i = t_{s_i}$.

Working with (3.20) allows for simple updates for all model parameters. In particular, the weights can be updated through the stick-breaking ratios by sampling

$$v_h \mid \ldots, \cancel{(u_i)} \sim \mathsf{Beta}\left(1 + n_h, M + \sum_{k>h} n_k\right),$$

where $n_h = \sum_{i=1}^{n} I(r_i = h)$ is the number of observations such that $r_i = h$. Similarly, the atoms $\widetilde{\theta}_h$ for the occupied components are sampled from

$$p(\widetilde{\theta}_h \mid \ldots) \propto G_0(\widetilde{\theta}_h) \prod_{\{i:r_i=h\}} p(y_i \mid \widetilde{\theta}_h),$$

while the atoms associated with empty components, i.e., $n_h = 0$, can be sampled, on demand, directly from G_0.

Finally, the latent variables u_i are, a posteriori, uniformly distributed

$$u_i \mid \ldots \sim \mathsf{Uni}[0, w_{r_i}],$$

and the indicators are updated from the full conditional

$$\mathsf{Pr}(r_i = h \mid \ldots) \propto I(w_h > u_i) p(y_i \mid \widetilde{\theta}_h).$$

Only a finite number of components satisfy the constrain $w_h > u_i$. Therefore, the normalizing constant for this last full conditional distribution can be computed in closed form. Let $H_i(u_i) = \{h : \ w_h > u_i\}$. Then

$$p(r_i = h \mid \ldots) = \begin{cases} \frac{p(y_i|\widetilde{\theta}_h)}{\sum_{\{k \geq 1 : w_k > u_i\}} p(y_i|\widetilde{\theta}_k)} & h \in H_i(u_i) \\ 0 & \text{otherwise.} \end{cases}$$

3.3.3. Retrospective Samplers

Retrospective samplers for Dirichlet process mixtures were developed by Roberts and Papaspiliopoulos (2008). Like the slice sampler, the retrospective sampler is based on an explicit representation of the mixing distribution G; to avoid the problem of storing an infinite number of weights and atoms, a Metropolis Hastings step is used to allocate observations to components, while the parameters associated with empty components are sampled retrospectively as they become necessary. The same algorithm is developed in Nieto-Barajas *et al.* (2012).

To formalize the idea of the retrospective sampler, consider simulating observation from the prior model, i.e., simulating an i.i.d. sequence $\theta_1, \ldots, \theta_n$ where $\theta_i \mid G \sim G$ where $G \sim \mathsf{DP}(M, G_0)$. This can be done directly by exploiting the species sampling representation of the process in (3.3). Alternatively, we can first simulate the distribution G using the stick-breaking construction in (3.2), and then sample θ_i conditional on G. Under the second approach we can avoid the difficulties associated with having a countably infinite number of components by utilizing the following scheme.

1. Simulate $w_1 = v_1 \sim \mathsf{Beta}(1, M)$ and $\widetilde{\theta}_1 \sim G_0$, and set $H = 1$, $i = 1$ and $w_0 = 0$.
2. For $i = 1, \ldots, n$
   (a) Simulate $U_i \sim \mathsf{Uni}[0, 1]$.
   (b) If for some $k \leq H$ we have $\sum_{h=0}^{k-1} w_h < U_i \leq \sum_{h=0}^{k-1} w_h + w_k$, then set $\theta_i = \widetilde{\theta}_k$.
   (c) If $U_i > \sum_{h=0}^{H} w_h$, then simulate $v_{H+1} \sim \mathsf{Beta}(1, M)$ and $\widetilde{\theta}_{H+1} \sim G_0$, and set $w_{H+1} = v_{H+1} \prod_{k<H+1}\{1 - v_H\}$ and $H = H + 1$. Go back to step (b).

In words, we generate the weights w_h and point masses only as and when needed. That is all!

We proceed now to describe the posterior sampler for the DPM. As in §3.3.2, we introduce indicator variables (r_i) such that $\theta_i = \widetilde{\theta}_{r_i}$ and define $n_h = \sum_{i=1}^{n} I(r_i = h)$ as the number of observations assigned to component h.

As with the slice sampler, the full conditionals for the v_hs and the $\widetilde{\theta}_h$ are independent and given by

$$v_h \mid \ldots \sim \mathsf{Beta}\left(1 + n_h, M + \sum_{k>h} n_k\right) \tag{3.21}$$

and

$$p(\widetilde{\theta}_h \mid \ldots) \propto G_0(\widetilde{\theta}_h) \prod_{\{i:r_i=h\}} p(y_i \mid \widetilde{\theta}_h). \tag{3.22}$$

On the other hand, the full conditional distribution for the indicator variables is given by

$$\mathsf{Pr}(r_i = h \mid \ldots) \propto w_h p(y_i \mid \widetilde{\theta}_h), \qquad h = 1, 2, \ldots$$

Since computation of the normalizing constant involves a sum over an uncountable number of terms, directly sampling from this distribution is not feasible. To avoid this issue, Roberts and Papaspiliopoulos (2008) describe a Metropolis Hastings algorithm for $\mathbf{s} = (s_1, \ldots, s_n)$. More specifically, for every $i = 1, \ldots, n$ a proposal $\widetilde{\mathbf{s}} = (\widetilde{s}_1, \ldots, \widetilde{s}_n)$ is created by setting $\widetilde{s}_j = s_j$ for $j \neq i$ and generating $\widetilde{s}_i$ according to the proposal distribution

$$p(\widetilde{s}_i = h) \propto \begin{cases} w_h p(y_i \mid \widetilde{\theta}_h) & h \leq \max_i\{r_i\} \\ M_i(\mathbf{s}) w_h & h > \max_i\{r_i\}, \end{cases}$$

where $M_i(\mathbf{s})$ is a user-selected parameter. The corresponding normalizing constant is given in this case by

$$c_i(\mathbf{s}) = \sum_{h=1}^{\max\{r_i\}} w_h p(y_i \mid \widetilde{\theta}_h) + M_i(\mathbf{s}) \left(1 - \sum_{h=1}^{\max\{r_i\}} w_h\right).$$

The proposed value is accepted with probability

$$\alpha_i(\mathbf{s}, \widetilde{\mathbf{s}}) = \begin{cases} 1 & \widetilde{s}_i \leq \max\{r_i\} \text{ and } \max_i\{\widetilde{s}_i\} = \max_i\{r_i\} \\ \min\left\{1, \frac{c_i(\mathbf{s}) M_i(\widetilde{\mathbf{s}})}{c_i(\widetilde{\mathbf{s}}) p(y_i \mid \widetilde{\theta}_{r_i})}\right\} & \widetilde{s}_i \leq \max_i\{r_i\} \text{ and } \max_i\{\widetilde{s}_i\} < \max_i\{r_i\} \\ \min\left\{1, \frac{c_i(\mathbf{s}) p(y_i \mid \widetilde{\theta}_{\widetilde{s}_i})}{c_i(\widetilde{\mathbf{s}}) M_i(\mathbf{s})}\right\} & \widetilde{s}_i > \max_i\{r_i\}. \end{cases}$$

If an observation is allocated to a new component (i.e., a proposal $\widetilde{s}_i > \max\{r_i\}$ is accepted), then the necessary values for the v_hs and $\widetilde{\theta}_h$s are sample retrospectively from (3.21) and (3.22). The constant $M_i(\mathbf{s})$ is selected so that

$$M_i(\mathbf{s}) = \max_{h \leq \max_i\{r_i\}} \left\{p(y_i \mid \widetilde{\theta}_h)\right\}$$

in order to generate a sampler that is geometrically ergodic.

3.3.4. Other Computational Approaches

We have but scratched the surface of possible Markov chain Monte Carlo methods for inference under the Dirichlet process mixture model. For example, Dahl (2003), Jain and Neal (2004) and Jain and Neal (2007) propose split-merge collapsed samplers that provide mechanisms to make global moves in the space of partitions that are induced by the DP prior. Alternatively, MacEachern *et al.* (1999) and Carvalho *et al.* (2010) consider sequential Monte Carlo approaches that are particularly useful for problems where observations are collected sequentially and it is necessary to update the model after each observation is received. Finally, Blei and Jordan (2006) propose a variational algorithm that is computationally efficient for large sample sizes.

3.4. The Finite DP

The Dirichlet process mixture model potentially allows for an infinite number of clusters as $n \to \infty$. However, for any finite sample size n the number k of occupied components cannot be greater than n, and is typically much smaller than that. This suggests that instead of dealing with a countably infinite number of components, we could work with mixtures with a large but finite number of components. This should simplify computation while retaining most of the theoretical advantages of the Dirichlet process.

The first approach we discuss to construct truncated versions of the Dirichlet process is the $\epsilon-$DP of Muliere and Tardella (1998). For any $\epsilon \in (0,1)$, a random distribution G^ϵ is said to follow an ϵ-Dirichlet process if it admits a representation of the form

$$G^\epsilon(\cdot) = \sum_{h=1}^{H_\epsilon} w_h \delta_{\widetilde{\theta}_h}(\cdot) + \left\{1 - \sum_{h=1}^{H_\epsilon} w_h\right\} \delta_{\widetilde{\theta}_{H_\epsilon+1}}(\cdot),$$

where $(\widetilde{\theta}_h)$ is a collection of i.i.d. draws from the baseline measure G_0, $w_h = v_h \prod_{k<h}(1-v_k)$, where (v_h) is another i.i.d. sequence such that $v_h \sim \mathsf{Beta}(1, M)$, and $H_\epsilon = \inf\{m \in \mathbb{N} : \sum_{h=1}^m w_h \geq 1-\epsilon\}$.

The definition of the ϵ-Dirichlet process is analogous to that of the (regular) Dirichlet process, but the sum stops after a random number of draws, $H_\epsilon \sim \mathsf{Poi}(-M\log\epsilon)$, and the remaining mass (which, by construction, must be no larger than ϵ) is assigned to the last atom. By bounding the probability associated with this last atom, the definition ensures that the total variation distance between the finite and the infinite versions of the process is no larger than ϵ. Indeed, let $(\widetilde{\theta}_h)$ and (v_h) be two i.i.d. sequences such that $\widetilde{\theta}_h \sim G_0$ and $v_h \sim \mathsf{Beta}(1, M)$, and define

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\widetilde{\theta}_h}, \qquad G^\epsilon(\cdot) = \sum_{h=1}^{H_\epsilon} w_h \delta_{\widetilde{\theta}_h}(\cdot) + \left\{1 - \sum_{h=1}^{H_\epsilon} w_h\right\} \delta_{\widetilde{\theta}_{H_\epsilon+1}}(\cdot),$$

where $H_\epsilon = \inf\{m \in \mathbb{N} : \sum_{h=1}^m w_h \geq 1-\epsilon\}$. Then

$$\sup_B \{|G(B) - G^\epsilon(B)|\} \leq \epsilon.$$

Naturally, as $\epsilon \to 0$, draws from the ϵ-DP converge (in total variation norm) to the draws from a regular DP. Hence, the class of ϵ-DPs is dense, in the sense that it

is rich enough to approximate arbitrarily well any distribution on the underlying probability space.

An alternative definition of a truncated Dirichlet process was introduced in Ishwaran and James (2001, 2002). Rather than using a random stopping rule for the number of point masses that ensures a bound on the total variation norm, they argue instead for using a fixed number of atoms H and study the behavior of the random distribution as the number of atoms grows. In particular, Ishwaran and James (2001) show that the marginal distributions for a sample $(x_1, \ldots, x_n)$ under $G = \sum_{h=1}^{\infty} w_h \delta_{\widetilde{\theta}_h}$ and $G^H = \sum_{h=1}^{H} w_h \delta_{\widetilde{\theta}_h}$ are almost indistinguishable and the L_1 distance between these marginal distribution is bounded by $4ne^{-(H-1)/M}$. This results suggest that, as long as M is not very large, small H already obtain a good approximation, even if n is large.

Hierarchical models based on finite DPs have some important computational advantages over models based on infinite DPs. For example, since the number of atoms is finite, techniques for posterior inference on finite mixture models can be employed to perform estimation on finite DP mixtures. Indeed, the main motivation of Ishwaran and James (2001) and Ishwaran and James (2002) is to develop alternative computational algorithms for stick-breaking priors, which they call blocked Gibbs samplers.

We introduce latent indicators $r_1, \ldots, r_n$ such that $\theta_i = \widetilde{\theta}_{r_i}$. The joint distribution is then given by

$$p\{(r_i), (\widetilde{\theta}_h), (v_h) \mid \mathbf{y}\} \propto \prod_{i=1}^{n} p(y_i \mid \widetilde{\theta}_{r_i}) \prod_{i=1}^{n} p\{r_i \mid (v_h)\} \prod_{h=1}^{H} dG_0(\widetilde{\theta}_h) \prod_{h=1}^{H-1} p(v_h \mid M).$$

From this we find

$$p(\widetilde{\theta}_h \mid \ldots) \propto G_0(\widetilde{\theta}_h) \prod_{\{i:r_i=h\}} p(y_i \mid \widetilde{\theta}_h). \tag{3.23}$$

Similarly, the stick-breaking weights (v_h) are conditionally independent with

$$v_h \mid \ldots \sim \mathsf{Beta}\left(1 + n_h, M + \sum_{k>h} n_k\right), \tag{3.24}$$

where $n_h = \sum_{i=1}^{n} I(r_i = h)$ is the number of observations such that $r_i = h$. Finally, the posterior full conditional distribution for r_i is

$$\mathsf{Pr}(r_i = h \mid \ldots) \propto w_h p(y_i \mid \widetilde{\theta}_h), \tag{3.25}$$

where $w_h = v_h \prod_{k<h}(1 - v_h)$ for $h < H$ and $w_H = \prod_{k \leq H}(1 - v_h)$.

Notice the similarities between this algorithm and the slice sampler we described in §3.3.2. The structure of both algorithms is almost identical, with the main distinction being that in the blocked Gibbs sampler the number of components H is predetermined before running the algorithm, while in the slice sampler the number of components being used for computation (which is typically larger than the number of occupied components) is determined dynamically as part of the algorithm.

Another advantage of working with a truncated version of the Dirichlet process is that computation for functionals of the random distribution G is greatly simplified because G can be explicitly evaluated. For example, the predictive distribution for

a new observation y_{n+1} can be easily evaluated by noting that, under the truncated model,

$$p(y_{n+1} \mid y_1, \ldots, y_n) = \mathsf{E}\left\{ \sum_{h=1}^{H} w_h p(y_{n+1} \mid \widetilde{\theta}_h) \mid y_1, \ldots, y_n \right\}.$$

Other functionals of G^H can be easily computed in the same way (see also §3.6).

3.5. Mixtures of DP

In §3.3.1 we discussed the possibility of making the baseline measure random. In this section we formalize this construction through the introduction of mixtures of DPs (MDP). MDP's where first introduced by Antoniak (1974) as a generalization of the Dirichlet process. In contrast to the DPM, where the DP is the prior model for the mixing measure in a mixture of parametric distributions, the MDP arises when the baseline measure G_0 and/or the concentration parameter M are random. I.e., $G_0 = G_{0,\eta}$ and/or $M = M_\eta$ are indexed by η and we define a random probability measure by

$$G \mid \eta \sim \mathsf{DP}(M_\eta, G_{0,\eta}), \qquad \eta \sim H(\eta).$$

Then, G is said to follow a mixture of Dirichlet processes with precision M_η, baseline measure $G_{0,\eta}$ and mixing distribution H, or simply $G \sim \int \mathsf{DP}(M_\eta, G_{0,\eta}) dH(\eta)$.

Note that the MDP and the DPM are entirely different models. The earlier mixes some kernel with respect to a DP random measure. The latter mixes DP random measure with respect to a prior on hyperparameters. Nevertheless, there is a natural connection between both models (see Antoniak, 1974). The posterior law for the mixing distribution G in a DPM model follows a MDP distribution. Consider

$$y_i \mid \theta_i \sim p(y_i \mid \theta_i), \qquad \theta_i \mid G \sim G, \qquad G \sim \mathsf{DP}(M, G_0),$$

then, the law of $G \mid y_1, \ldots, y_n$ can be written as

$$G \mid y_1, \ldots, y_n \sim \int \mathsf{DP}\left(M + n, \frac{M G_0 + \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}^\star_{r_i}}}{M + n} \right) dP(s_1, \ldots, s_n, \theta^\star_1, \ldots, \theta^\star_k \mid y_1, \ldots, y_n). \tag{3.26}$$

The posterior distribution over G induced by a DPM is simply a MDP!

3.6. Functionals of DPs

3.6.1. Inference for Non-linear Functionals of DP

We return to inference for the DPM model (3.9) again,

$$y_i \mid G \sim F = \int p(y_i \mid \theta) G(\mathrm{d}\theta), \qquad G \sim \mathsf{DP}(M, G_0).$$

Sometimes investigators are interested in posterior inference for functionals of the unknown distribution F. For example, in density estimation with an unknown distribution F one might be interested in computing $\mathsf{E}\{f \mid y_1, \ldots, y_n\}$ as a point

estimate of the density f associated with the unknown distribution F. Alternatively, we might be interested in providing posterior distributions for quantiles of F, i.e., $p(F^{-1}(\gamma) \mid y_1, \ldots, y_n)$ for some $\gamma \in (0,1)$ where F^{-1} is the inverse c.d.f. of F and $\gamma \in (0,1)$ is a prespecified percentile.

First, we consider inference for functionals of F under the collapsed sampler discussed in §3.3.1. For linear functionals of F, we can explicitly marginalize with respect to G and compute point estimators directly from the sampler output. For example, we can compute $\mathsf{E}\{f(y^\star) \mid y_1, \ldots, y_n\}$ by changing the order of integration in

$$\mathsf{E}\{f(y^\star) \mid y_1, \ldots, y_n\} = \mathsf{E}\left\{\int\int p(y^\star \mid \theta) G(\mathrm{d}\theta)\right\} = \int p(y^\star \mid \theta) G^\star(\mathrm{d}\theta),$$

where $G^\star = \mathsf{E}\{G \mid y_1, \ldots, y_n\}$. Since $G \mid y_1, \ldots, y_n$ is a MDP, as we discussed at the end of §3.5 we have

$$\begin{aligned}(3.27)\quad &\mathsf{E}\{f(y^\star) \mid y_1, \ldots, y_n\} \\ &= \int \left\{ \frac{M}{M+n} \int p(y^\star \mid \theta) G_0(\mathrm{d}\theta) + \sum_{j=1}^{k^{(b)}} \frac{n_j}{n+M} p(y^\star \mid \theta_j^\star) \right\} \\ &\qquad\qquad dp(s_1, \ldots, s_n, \theta_1^\star, \ldots, \theta_k^\star \mid y_1, \ldots, y_n).\end{aligned}$$

Given a Monte Carlo posterior sample $(s_1^{(b)}, \ldots, s_n^{(b)}, \theta_1^{\star(b)}, \ldots, \theta_{k^{(b)}}^{\star(b)})_{b=1}^B$ from

$$p(s_1, \ldots, s_n, \theta_1^\star, \ldots, \theta_k^\star \mid y_1, \ldots, y_n),$$

(3.27) can be approximated by the Monte Carlo estimator

$$\begin{aligned}(3.28)\quad &\mathsf{E}\{f(y^\star) \mid y_1, \ldots, y_n\} \\ &= \frac{1}{B} \sum_{i=1}^{B} \left\{ \frac{M}{M+n} \int p(y^\star \mid \theta) G_0(\mathrm{d}\theta) + \sum_{j=1}^{k^{(b)}} \frac{n_j^{(b)}}{n+M} p(y^\star \mid \theta_j^{\star(b)}) \right\}.\end{aligned}$$

Alternatively one could evaluate $E[f(y^\star) \mid \mathbf{y}]$ as the posterior predictive $p(y_{n+1} \mid \mathbf{y})$ in a random sample $y_i \sim F$,

$$p(y_{n+1} \mid \mathbf{y}) = E[p(y_{n+1} \mid F, \mathbf{y}) \mid \mathbf{y}] = E[f(y_{n+1}) \mid \mathbf{y}],$$

which leads to the same MCMC estimate (3.28).

For non-linear functionals, and more generally, if we want to obtain the full posterior distribution of a given functional, we need to deal with the infinite dimensional mixing distribution G. Gelfand and Kottas (2002) exploit the fact that $p(G, s_1, \ldots, s_n, \theta_1^\star, \ldots, \theta_k^\star \mid y_1, \ldots, y_n)$ can be factorized as

$$\begin{aligned}&p(G, s_1, \ldots, s_n, \theta_1^\star, \ldots, \theta_k^\star \mid y_1, \ldots, y_n) \\ &\qquad = p(G \mid s_1, \ldots, s_n, \theta_1^\star, \ldots, \theta_k^\star) p(s_1, \ldots, s_n, \theta_1^\star, \ldots, \theta_k^\star \mid y_1, \ldots, y_n),\end{aligned}$$

where $p(G \mid s_1, \ldots, s_n, \theta_1^\star, \ldots, \theta_k^\star)$ is simply a Dirichlet process with base measure $G_1 \propto MG_0 + \sum_{j=1}^k n_j \delta_{\theta_j^\star}$. Hence, given a realization $(s_1^{(b)}, \ldots, s_n^{(b)}, \theta_1^{\star(b)}, \ldots, \theta_{k^{(b)}}^{\star(b)})$ from the posterior distribution, a realization from the posterior for G can be constructed as

$$G^{(b)}(\cdot) = \sum_{h=1}^{\infty} \varpi_h \delta_{\widetilde{\theta}_h}(\cdot)$$

where $\varpi_h = z_h \prod_{k<h}(1 - z_h)$, $z_h \sim \mathsf{Beta}(1, M + n)$ and $\widetilde{\theta}_h \sim G_1$.

In practice, an ϵ truncation of the DP (recall our discussion from §3.4) is used, so that we generate only a random (but finite!) number of atoms H_ϵ such that $H_\epsilon = \inf\{m \in \mathbb{N} : \sum_{h=1}^{m} \varpi_h > 1 - \epsilon\}$. Given $G^{(b)}$, we can evaluate any functional of F. For example, a sample from $p(F^{-1}(\gamma) \mid y_1, \ldots, y_n)$, can be obtained by computing (for each iteration of the MCMC) the value $q_\gamma^{(b)}$ such that

$$\gamma = \sum_{h=1}^{H_\epsilon} \varpi_h^{(b)} P(q_\gamma^{(b)} \mid \theta_h^{\star(b)}).$$

Here $P(y) = \int_{-\infty} p(y \mid \theta)$ is the c.d.f. of the kernel in (3.9). The values $q_\gamma^{(1)}, \ldots, q_\gamma^{(B)}$ can then be used to perform posterior inference on $F^{-1}(\gamma)$.

Finally, we consider inference for functionals of F under the slice and retrospective samplers that are discussed in §3.3.2 and §3.3.3. Both of these samplers rely on an explicit representation of the mixing distribution G. Therefore inference for functionals is, in principle, straightforward. However, a word of caution is in order. Even though both the slice and the retrospective samplers perform dynamically adaptive truncations of G, the accuracy of these truncations (in terms of how well they approximate the infinite dimensional G) is not predetermined. Hence, in practice we might need to explicitly represent (and simulate) additional mixture components in order to ensure that the posterior inference for the functionals of G is sensible. Note that this is not the case if an almost sure truncation is used because the value of H is predetermined beforehand to ensure a good approximation.

3.6.2. Centering the DP

A particular example of inference for functionals of a DP random probability measure arises in applications to mixed effects models. Recall the discussion of mixed effects models in §2.3. To be specific, assume a linear model

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \tag{3.29}$$

with $\epsilon_{ij} \sim \mathsf{N}(0, \sigma^2)$, i.i.d. For example, y_{ij} might be logarithm of the prostate-specific antigen (log PSA) measurements for the i-th patient at time t_{ij}. Here (β_0, β_1) are fixed effects including intercept and overall growth rate of PSA, and $\mathbf{b}_i = (b_{0i}, b_{1i})$ are patient-specific random effects. The random effects are assumed to arise from a random effects distribution $\mathbf{b}_i \sim G$. When the investigator is not willing or able to assume a parametric model for G then we might complete the model with a BNP prior, for example,

$$\mathbf{b}_i \sim G, \qquad\qquad G \sim \mathsf{DP}(M, G_0),$$

with $G_0(\mathbf{b}) = \mathsf{N}(\mathbf{b} \mid 0, D)$ and a (conditionally) conjugate hyperprior on D. Let $\mu_G = \int x dG(x)$ and $\Sigma_G = \int (x - \mu_G)(x - \mu_G)' \, dG(x)$ denote the first and (centered) second moment of G. For a random probability measure G the moments μ_G and Σ_G become random variables. Even when G_0 is chosen to have a zero mean $\mu_{G_0} =$

0, the random moments μ_G are almost surely not zero. This greatly complicates interpretation of the fixed effects (β_0, β_1). Similarly, the elements of Σ_G are the variance components; reporting D, as is often done, only approximates Σ_G.

Inference on fixed effects in (3.29) is best formalized as inference on $(\boldsymbol{\beta} + \boldsymbol{\mu}_G)$, and inference on variance components requires the posterior distribution of Σ_G. Fortunately both are easily available from standard Monte Carlo output for MCMC posterior simulation under model (3.29). Li *et al.* (2010) give explicit formulas for the first two posterior moments of $(\boldsymbol{\beta} + \boldsymbol{\mu}_G)$ and Σ_G. All can be evaluated by postprocessing with an available Monte Carlo sample.