6

Bayesian Nonparametric Mixture Models

Peter Müller
UT Austin, USA

CONTENTS

We review the use of Bayesian nonparametric priors for inference in mixtures. Interpreting a mixture model as an expectation with respect to a mixing measure, it becomes natural to complete the model with a prior probability model on the unknown mixing measure. Prior models on random probability measures, like the mixing measure here, are known as Bayesian nonparametric models. We review some commonly used models, including in particular the Dirichlet process prior, normalized random measures with independent increments, and the determinantal point process and variations. Many applications of such models include inference on the implied partition of the experimental units, that is, a clustering of the data. This gives rise to predictive distributions that again take the form of a mixture model.

6.1 Introduction

Inference for mixture models and closely related hierarchical models is one of the big success stories of Bayesian inference. This is particularly true for applications in biostatistics. For example, popular Bayesian models for inference on patient subpopulations are variations of

the following model. Let y_i denote a response for the ith patient. We assume

$$y_i \sim \sum_{g=1}^{G} \eta_g p(y_i \mid \mu_g), \tag{6.1}$$

$i = 1, \ldots, n$, including possibly $G = \infty$. The component-specific model $p(y_i \mid \mu_g)$ could be, for example, a survival model with parameters μ_g, possibly including a regression on patient covariates. Let $\eta = (\eta_1, \ldots, \eta_G)$ denote the weights. The model is completed with a prior probability model $p(\eta)$ and $p(\mu_1, \ldots, \mu_G)$, typically assuming independence of the μ_g, perhaps conditional on some hyperparameters. Inference in this model becomes a problem of Bayesian nonparametric inference when (6.1) is interpreted as an expectation with respect to a random mixing measure, as follows.

Random discrete mixing measure

Let $H = \sum \eta_g \delta_{\mu_g}$ denote a (discrete) random probability measure with atoms at μ_g. The mixture model (6.1) can then be written as $y_i \mid H \sim \int p(y_i \mid \theta)\, dH(\theta)$. Equivalently, we can introduce latent variables θ_i and replace the integral with respect to H by a hierarchical model

$$y_i \mid \theta \sim p(y_i \mid \theta_i), \quad \theta_i \mid H \sim H. \tag{6.2}$$

In model (6.2) it becomes natural to consider a Bayesian nonparametric (BNP) prior $p(H)$ on H. BNP refers to prior probability models on infinite-dimensional parameter spaces, for example, random distributions such as H in (6.2). See, for example, Hjort et al. (2010), Phadia (2013), Ghoshal & van der Vaart (2017), and Müller et al. (2015) for recent reviews of Bayesian nonparametric inference. The most widely used BNP prior $p(H)$ is the Dirichlet process (DP) prior (Ferguson, 1973). We briefly introduce the DP model to serve as a running example in this introduction, but defer more details to Section 6.2. A DP prior is indexed with two parameters, $\alpha > 0$ and a probability distribution H_0, and defines a prior for $H = \sum \eta_g \delta_{\mu_g}$ by assuming

$$\eta_g = v_g \prod_{\ell<g} (1 - v_\ell), \quad \text{with } v_g \sim \mathcal{B}e(1, \alpha), \text{ i.i.d. and } \mu_g \sim H_0, \text{ i.i.d.} \tag{6.3}$$

This is known as the stick-breaking representation (Sethuraman, 1994). We write $H \sim \mathcal{DP}(\alpha, H_0)$. Model (6.2) with a DP prior $p(H) = \mathcal{DP}(\alpha, H_0)$ is known as a DP mixture. It is perhaps the most widely used BNP model; see also Chapter 17 for applications in finance. It was first introduced in Lo (1984), Escobar (1988, 1994), and Escobar & West (1995).

Latent partitions

Closely related to inference for a discrete mixing measure H is the problem of inference on a random partition of the experimental units $[n] = \{1, \ldots, n\}$. The nature of (6.1) as a random partition of $[n]$ is evident if we rewrite (6.1) as a hierarchical model,

$$y_i \mid \theta_i \sim p(y_i \mid \theta_i), \quad \mathrm{P}(\theta_i = \mu_g \mid \eta) = \eta_g, \tag{6.4}$$

with latent variables $\theta = (\theta_1, \ldots, \theta_n)$. Let $\{\theta_1^\star, \ldots, \theta_{G_+}^\star\} \subseteq \{\mu_g,\ g = 1, \ldots, G\}$ denote the $G_+ \leq \min(n, G)$ distinct values among the θ_i. Then $C_k = \{i :\ \theta_i = \theta_k^\star\}$, $k = 1, \ldots, G_+$, defines a partition $\mathcal{C}_n = \{C_1, \ldots, C_{G_+}\}$ of $[n]$ into G_+ clusters. We use $n_k = |C_k|$ to denote the cardinality of the clusters. When n is not understood from the context we add a subindex $_n$ and write $G_{+,n}$ and n_{nk}. Sometimes it will be convenient to introduce cluster membership indicators $z_i = k$ if $i \in C_k$ and use $\mathbf{z} = (z_1, \ldots, z_n)$ to denote the partition. Note that

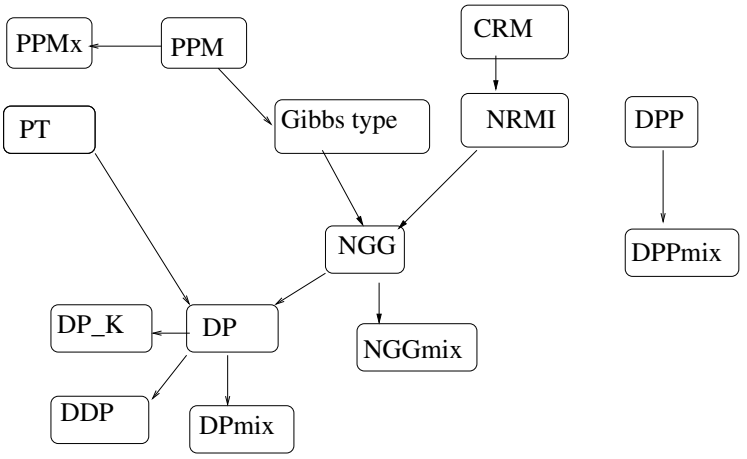


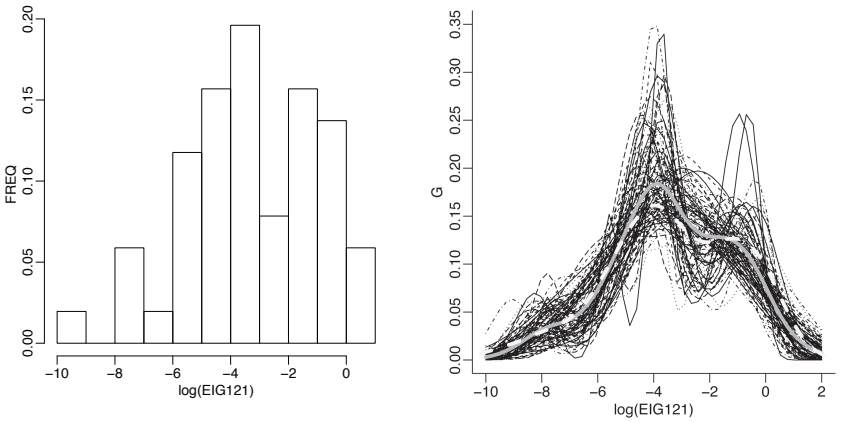
FIGURE 6.1
The diagram shows how BNP models in the upcoming discussion are related to each other. An arrow indicates that one model is a special case of the other (or constructed using the other model). The diagram is far from complete, including only models that are introduced in this chapter.

alternatively we could have introduced clusters $\tilde{C}_g = \{i : \theta_i = \mu_g\}$, $g = 1, \ldots, G$, allowing for possibly empty clusters. However, to avoid ambiguous terminology we will throughout use a convention of indexing only occupied clusters C_k.

Being equivalent model statements, inference on a random partition $\mathcal{C}_n$ is also implied by (6.2). To see this, note that the discrete nature of H implies many ties among the θ_i. If we use the arrangement of ties to define clusters we are back to a prior on the random partition $\mathcal{C}_n$ as in (6.4). That is, let $\theta_k^\star$, $k = 1, \ldots, G_+$, denote the $G_+ \leq n$ unique values and define $C_k = \{i : \theta_i = \theta_k^\star\}$, as before. In other words, any prior $p(H)$ in (6.2) implies a prior $p(\mathcal{C}_n)$. In fact, one can show that any exchangeable random partition $p(\mathcal{C}_n)$ can be introduced this way (Pitman, 2006). Here, exchangeability refers to invariance of $p(\mathcal{C}_n)$ under any permutation of the indices $i = 1, \ldots, n$, for any n. And we assume that the random partition $p(\mathcal{C}_n)$ is consistent across n in the sense that the restriction of a random partition $p(\mathcal{C}_{n+1})$ of $[n+1]$ to $[n]$ gives $p(\mathcal{C}_n)$. That is, $p(\mathcal{C}_n) = \sum_{z_{n+1}} p(\mathcal{C}_{n+1})$.

In this chapter we review some of the popular BNP mixture models, starting with the widely used DP mixture models in Section 6.2 and some generalizations of the DP prior in Section 6.3. We then discuss inference that exploits the posterior distribution on the implied random partition and variations of these models in Section 6.4. Finally, in Section 6.5, we review the use of repulsive priors on the mixing measure, using in particular the determinantal point process (DPP).

In the following discussion we will run into a proliferation of BNP models and acronyms. Figure 6.1 summarizes how these models are related to each other. The diagram is far from complete, including only models that are introduced or at least mentioned in this chapter. For similar diagrams covering more BNP models see, for example, (Phadia, 2013, Figure 1.1) and (Müller et al., 2015, Figure 1).



(a) Histogram of the y_i

(b) Posterior draws of F based on $H \sim p(H \mid \mathbf{y})$

FIGURE 6.2
Example 6.1. (a) Data and (b) posterior inference for F. Panel (b) shows 96 posterior draws of the mixture model F based on $H \sim p(H \mid \mathbf{y})$ (thin black curves) and the posterior mean $\mathrm{E}(F \mid \mathbf{y})$ (thick gray curve). For comparison the figure also shows a kernel density estimate (dashed thick white line).

6.2 Dirichlet Process Mixtures

6.2.1 The Dirichlet process prior

The original definition of the DP is due to Ferguson (1973). A random distribution H on Θ is said to follow a DP prior with baseline probability measure H_0 and mass parameter α, denoted $H \sim \mathcal{DP}(\alpha, H_0)$, if for any partition $\{A_1, \ldots, A_k\}$ of Θ,

$$(H(A_1), \ldots, H(A_k)) \sim \mathcal{D}(\alpha H_0(A_1), \ldots, \alpha H_0(A_k)).$$

In the introduction we already saw an alternative defining property of the DP, known as the stick-breaking construction. Implicit in the stick-breaking construction is the fact that H is discrete, even if H_0 is a continuous distribution.

Pólya urn

For later reference we state yet another defining property of the DP. Let $\theta_1, \theta_2, \ldots$ be an i.i.d. sequence such that

$$\theta_i \mid H \sim H \quad \text{and} \quad H \sim \mathcal{DP}(\alpha, H_0), \tag{6.5}$$

as in (6.2). As before, let $\{\theta_1^\star, \ldots, \theta_{G_{+,n}}^\star\}$ denote the unique values, let $n_{nk} = \sum_{i=1}^n \mathbb{I}(\theta_i = \theta_k^\star)$ be the number of θ_i, $i \leq n$, equal to $\theta_k^\star$, and let $z_i = k$ if $\theta_i = \theta_k^\star$. Blackwell & MacQueen (1973) showed that the joint distribution of the z_i can be characterized in terms of the predictive probability function,

$$\mathrm{P}(z_i = k \mid z_1, \ldots, z_{i-1}) \propto \begin{cases} n_{i-1,k}, & k = 1, \ldots, G_{+,i-1}, \\ \alpha, & k = G_{+,i-1} + 1. \end{cases} \tag{6.6}$$

This is known as the Pólya urn. For later reference note that (6.6) implies $p(\theta_n \mid \theta_1, \ldots, \theta_{n-1}) \propto \sum_{k=1}^{G_{+,n-1}} n_{n-1,k}\delta_{\theta_k^\star} + \alpha H_0$. Let $\theta_{-i} = (\theta_\ell,\ \ell \neq i)$ and let G_+^{-i} and n_j^{-i} denote the number of unique values and the multiplicities in θ_{-i}. Since (6.5) is symmetric in the indices, the same expression as (6.6) must hold for $p(\theta_i \mid \theta_{-i})$:

$$p(\theta_i \mid \theta_{-i}) \propto \sum_{k=1}^{G_+^{-i}} n_k^{-i}\delta_{\theta_k^\star} + \alpha H_0. \tag{6.7}$$

Posterior distributions

One of the reasons for the wide use of the DP prior is computational ease and, closely related, conjugacy with respect to i.i.d. sampling. If $\theta = (\theta_1, \ldots, \theta_n)$ is an i.i.d. sample with $\theta_i \mid H \sim H$ and $H \sim \mathcal{DP}(\alpha, H_0)$ then

$$H \mid \theta \sim \mathcal{DP}(\alpha + n, H_1)$$

with $H_1 \propto \alpha H_0 + \sum_{i=1}^n \delta_{\theta_i}$. The posterior mean $\mathrm{E}(H \mid \theta) = H_1$ can be interpreted as a weighted average between the baseline measure H_0 and the empirical distribution of the θ_i.

Dirichlet process mixtures

In many applications the discrete nature of H is awkward. This motivates the DP mixture (DPM) model. For example, consider a density estimation problem $y_i \sim F$, i.i.d., $i = 1, \ldots, n$. To proceed with Bayesian inference on the unknown distribution F we complete the model with a prior on F. In most applications it would be unreasonable to directly use a DP prior, because of the implied discrete nature of a DP random measure. Instead we use a convolution of a discrete DP random measure with a continuous kernel, that is, $F(y) = \int p(y \mid \theta)\, dH(\theta)$ and a hyperprior $H \sim \mathcal{DP}(\alpha, H_0)$. This is, of course, exactly the mixture model (6.2) with a DP prior on H,

$$y_i \mid \theta \sim p(y_i \mid \theta_i), \quad \theta_i \mid H \sim H, \quad H \sim \mathcal{DP}(\alpha, H_0). \tag{6.8}$$

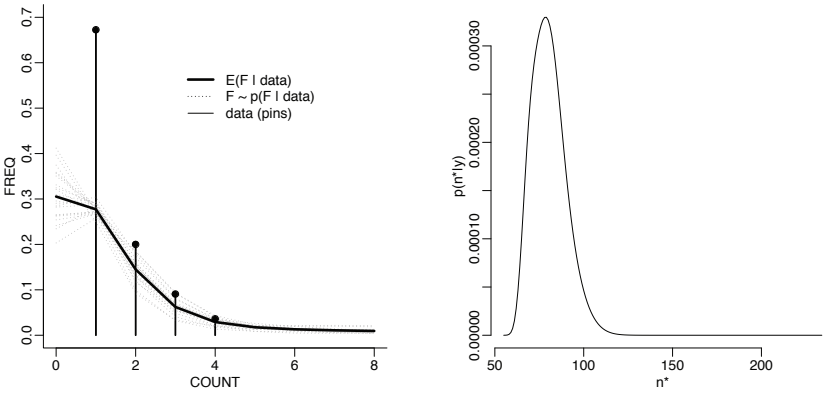
Example 6.1 (Gene expression data) *Figure 6.2a shows measurements y_i corresponding to EIG121 gene expression for $n = 51$ uterine cancer patients. We assume $y_i \sim F$ with a DPM prior* (6.8)*. Figure 6.2b shows posterior inference on F (as a pdf). Inference is based on* 500 *iterations of a Gibbs sampler, using an approximation with a finite DP prior.*

In Example 6.1 we used a DPM model for density estimation for an unknown distribution F in an application that called for a continuous distribution. In some cases the DPM model is useful also for discrete distributions, when inference includes extrapolation beyond the support of the data, or, more generally, borrowing of strength across different parts of the sample space. For example, in the following problem $y_i = 0$ is censored, by the nature of the experiment, and we use a DPM of Poisson distributions for inference on F, including $F(0)$. An honest description of uncertainties on $F(0)$ is critical for this example.

Example 6.2 (T-cell receptors) *Guindani et al. (2014) consider data on counts of distinct T-cell receptors. The diversity of T-cell receptor types is an important characteristic of the immune system. A common summary of the diversity is the clonal-size distribution. The clonal-size distribution is the table of frequencies $\widehat{F}_y$ of counts $y = 1, 2, \ldots, n$. For example, $\widehat{F}_2 = 11$ means that there were* 11 *distinct T-cell receptors that were observed twice, etc. Table 6.1 shows the observed frequencies for one of the experiments considered in Guindani et al. (2014).*

TABLE 6.1
Example 6.2. Frequencies $\widehat{F}_y$ of counts $y_i = 1, 2, \ldots$; $\widehat{F}_0$ is censored

counts y	0	1	2	3	4	≥ 5
frequencies $\widehat{F}_y$	–	37	11	5	2	0



(a) posterior mean $\mathrm{E}(F \mid \mathbf{y})$ and samples $F \sim p(F \mid \mathbf{y})$

(b) $p(n^\star \mid \mathbf{y})$

FIGURE 6.3
Example 6.2. (a) Posterior mean $\mathrm{E}(F \mid \mathbf{y})$ (solid line) of the clonal size distribution F, and posterior draws $F \sim p(F \mid \mathbf{y})$ (dotted thin lines). Only the probabilities at the integer values $y = 1, 2, \ldots$ are meaningful. The points are connected only for a better display. The pins show the empirical frequencies $\widehat{F}_y$ (standardized for $y \geq 1$; whereas F is standardized for $y \geq 0$). (b) The implied posterior $p(n^\star \mid \mathbf{y})$ for the total number of distinct T-cell receptors.

Consider a model $y_i \sim F$, with Poisson kernels, $y_i \sim \mathcal{P}(\theta_i)$ and $\theta_i \sim H$ with $H \sim \mathcal{DP}(\alpha, H_0)$. Importantly, we do not include the constraint $y_i \geq 1$. Instead we assume that T-cell receptor counts are generated as $y_i \sim F$, $i = 1, \ldots, n^\star$, for $n^\star \geq n$, with $y_i = 0$ for $i = n+1, \ldots, n^\star$. Without loss of generality we assume that the observed non-zero counts are the first n counts. The last $n^\star - n$ counts are censored. The total number of T-cell receptors $n^\star$ becomes another model parameter. Posterior inference is implemented using a Gibbs sampler, with one of the transition probabilities sampling $n^\star$ conditional on currently imputed θ_i, $i = 1, \ldots, n$. Figure 6.3a shows inference on the estimated distribution $F(y) = \int \mathcal{P}(y \mid \theta)\, dH(\theta)$ together with posterior draws $F \sim p(F \mid \mathbf{y})$. Figure Figure 6.3b shows the implied posterior distribution $p(n^\star \mid \mathbf{y})$.

6.2.2 Posterior simulation in Dirichlet process mixture models

Posterior inference is usually implemented as posterior MCMC simulation, which becomes particularly easy when the sampling model $p(y_i \mid \theta_i)$ and the base measure H_0 of the DP prior are conjugate. Recall the definition of the cluster membership indicators $z_i = k$ if $i \in C_k$ and the unique values $\{\theta_1^\star, \ldots, \theta_{G_+}^\star\}$, and let $\mathbf{y}_k^\star = (y_i;\ i \in C_k)$ denote the data arranged by clusters. Also, let θ_{-i} denote θ with the ith element removed, and similarly for $\mathbf{z}_{-i}$ and G_+^{-i}.

Posterior MCMC simulations

Escobar (1988) proposed the first posterior Gibbs sampler for the DPM model (6.2), based on transition probabilities that update θ_i by draws from the complete conditional posterior $p(\theta_i \mid \theta_{-i}, \mathbf{y})$. However, this Gibbs sampler can be argued to suffer from a slowly mixing Markov chain. Bush & MacEachern (1996) proposed a variation using two types of transition probabilities. One updates z_i by draws from the complete conditional posterior probability $p(z_i \mid \mathbf{z}_{-i}, \mathbf{y})$, after marginalizing with respect to θ. The other type of transition probability generates $\theta_k^\star$ from $p(\theta_k^\star \mid \mathbf{z}, \mathbf{y})$. We first discuss the latter. This will help us to establish notation that we will need for the other transition probability.

One transition probability is defined as a Gibbs step, generating $\theta_k^\star$ from the complete conditional posterior

$$p(\theta_k^\star \mid \mathbf{z}, \mathbf{y}) \propto H_0(\theta_k^\star) \prod_{i \in C_k} p(y_i \mid \theta_k^\star). \tag{6.9}$$

Here we used that *a priori* $p(\theta_k^\star) = H_0(\theta_k^\star)$. This follows from the stick-breaking definition of the DP random measure. The posterior $p(\theta_k^\star \mid \mathbf{z}, \mathbf{y})$ is simply the posterior on $\theta_k^\star$ in a parametric model with prior $H_0(\theta_k^\star)$ and sampling model $p(y_i \mid \theta_k^\star)$, restricted to data y_i, $i \in C_k$.

The other type of transition probabilities updates z_i by draws from the complete conditional posterior distribution $\mathrm{P}(z_i = k \mid \mathbf{z}_{-i}, \mathbf{y})$, which is derived as follows. First consider $p(\theta_i \mid \theta_{-i}, \mathbf{y})$. The prior given in (6.7) is multiplied with the sampling distribution to get

$$p(\theta_i \mid \theta_{-i}, \mathbf{y}) \propto \sum_{k=1}^{G_+^{-i}} n_k^{-i} p(y_i \mid \theta_k^{\star,-i})\, \delta_{\theta_k^{\star,-i}}(\theta_i) + \alpha p(y_i \mid \theta_i) H_0(\theta_i).$$

Recall that G_+^{-i} is the number of unique values after excluding θ_i, and similar for n_k^{-i} and $\theta_k^{\star,-i}$. Recognizing that $\theta_i = \theta_k^{\star,-i}$ implies $z_i = k$, we can write the same conditional as a joint distribution for (θ_i, z_i). Finally, marginalizing with respect to $\theta_k^\star$ using (6.9) replaces $p(y_i \mid \theta_k^{\star,-i})$ by $p(y_i \mid z_i = k, \mathbf{y}_k^{\star,-i}) = \int p(y_i \mid \theta_k^{\star,-i})\, dp(\theta_k^{\star,-i} \mid \mathbf{y}_k^{\star,-i})$, and we eventually find

$$\mathrm{P}(z_i = k \mid \mathbf{z}_{-i}, \mathbf{y}) \propto \begin{cases} n_k^{-i}\, p(y_i \mid z_i = k, \mathbf{y}_k^{\star,-i}), & \text{for } k = 1, \ldots, G_+^{-i}, \\ \alpha\, h_0(y_i), & \text{for } k = G_+^{-i} + 1, \end{cases} \tag{6.10}$$

where $h_0(y_i) = \int p(y_i \mid \theta)\, dH_0(\theta)$. Iterating over draws from (6.9) and (6.10) defines a widely used posterior MCMC scheme for DPM models.

Of course, (6.10) is only of practical use if h_0 is easily evaluated, that is, when $p(y \mid \theta)$ and $H_0(\theta)$ form a conjugate pair. Several alternative posterior simulation methods have been proposed for the more general case. A good discussion appears in Neal (2000). In particular, Algorithm 8 in Neal (2000) provides an easily implemented posterior MCMC scheme for general DPM models. A common characteristic of (6.10) and Algorithm 8 is the use of marginal probabilities $p(z_i \mid \mathbf{z}_{-i}, \mathbf{y})$, marginalizing over the unknown H.

Finite Dirichlet process

Alternatively, Ishwaran & James (2001) introduced an approximate DPM model by truncating the stick-breaking representation (6.3) of the DP after G terms by setting $v_G = 1$. Recall that v_g are the beta-distributed fractions in (6.3). Let $H \sim \mathcal{DP}_G$ denote the finite DP truncated after G terms, let $v = (v_1, \ldots, v_{G-1})$, $\mu = (\mu_1, \ldots, \mu_G)$ and define cluster membership indicators as $z_i = g$ if $\theta_i = \mu_g$ (allowing here for possibly empty clusters with $n_g = 0$). The model is known as the truncated DP. We write $H \sim \mathcal{DP}_G$. It is straightforward to implement a Gibbs sampler for $p(v, \mu, \mathbf{z} \mid \mathbf{y})$.

6.2.3 Dependent mixtures – the dependent Dirichlet process model

An attractive feature of the DPM model is the easy generalization to multiple related mixture models. Such models arise frequently in applications when inference for related populations, cases, etc. is required. Generically, let x_i denote some covariate for the ith observation. For example, $x_i \in \{0, 1\}$ might record standard care versus experimental therapy for patients $i = 1, \ldots, n$ in a clinical trial. For the moment assume that x_i is binary and consider the model

$$F_x(y_i) = p(y_i \mid x_i = x, H_x) = \int p(y_i \mid \theta)\, dH_x(\theta), \tag{6.11}$$

where $\{H_0, H_1\}$ are two mixing measures, for example, with marginal DP prior, $H_x \sim \mathcal{DP}(\alpha, H_0)$. In many applications one would want to complete the model by specifying $p(H_0, H_1)$ such that H_0 and H_1 are dependent. Let

$$H_x = \sum_g \eta_g \delta_{\mu_{xg}}. \tag{6.12}$$

The weights are already indexed by a single index g, common across x, in anticipation of the upcoming construction. A very elegant construction to achieve the desired dependence is by defining $p(\mu_{0g}, \mu_{1g})$ as a bivariate dependent distribution, but independent across g. In the general case of $\{H_x;\ x \in X\}$ with arbitrary index set X the bivariate distribution $p(\mu_{0g}, \mu_{1g})$ is replaced by a suitable stochastic process indexed by x; for example, a Gaussian process for $x \in \mathbb{R}$. This is the construction of the dependent DP (DDP) of MacEachern (1999). In one of several variations of the model the weights are common across x, as we already anticipated above in (6.12).

Example 6.3 (Oral cancer) *We use a dependent mixture of normal models as in* (6.11) *and* (6.12) *with a bivariate normal prior on* (μ_{0g}, μ_{1g}) *to model log survival times for* $n = 80$ *patients with cancer of the mouth (Klein & Moeschberger, 2003, Section 1.11). The covariate* $x_i \in \{0, 1\}$ *is an indicator for aneuploid* ($x = 1$, *abnormal number of chromosomes) versus diploid* ($x = 0$) *tumors. Figure 6.4 shows the posterior estimated distributions* $\mathrm{E}(F_x \mid \mathbf{y})$, $x = 0, 1$. *The estimates are plotted as survival functions on the absolute scale of the survival times. The thin lines show posterior simulations* $F_x \sim p(F_x \mid \mathbf{y})$.

Inference under the DP, the DDP, and many variations (and many other BNP models) is implemented in the R package `DPpackage@DPpackage` (Jara, 2007; Jara et al., 2011). Other public domain implementations include functions for some DP models in the R package bayesm (Rossi et al., 2005; Rossi & McCulloch, 2008). Also the Bayesian regression software by Karabatsos (2014) includes a wide variety of BNP regression problems in a menu-driven package.

6.3 Normalized Generalized Gamma Process Mixtures

6.3.1 NRMI construction

Yet another defining property of the DP is the construction as a normalized gamma process (Ferguson, 1973). The gamma process is a particular example of a much wider class of models known as completely random measures (CRMs; Kingman, 1993, Chapter 8). Consider any non-intersecting measurable subsets $A_1, \ldots, A_k$ of the desired sample space X. The defining

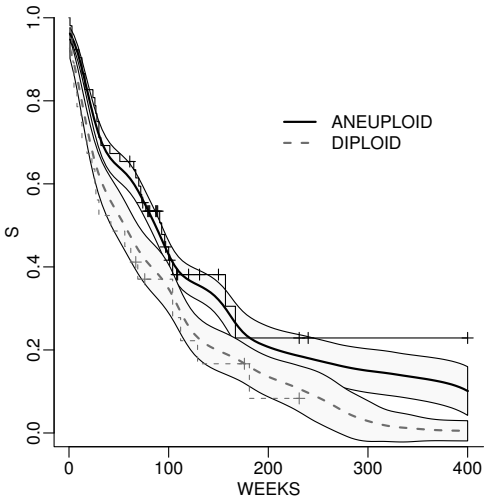


FIGURE 6.4
Estimated survival function $\mathrm{E}(F_x \mid \mathbf{y})$ by tumor type (solid black and dashed curves). The bands around the estimated survival functions show pointwise ± 1.0 posterior standard deviation bounds. The piecewise constant lines plot the Kaplan–Meier estimates.

property of a CRM $\tilde{H}$ is that $\tilde{H}(A_j)$ be mutually independent. The gamma process is a CRM with $\tilde{H}(A_j) \sim \mathcal{G}(\alpha H_0(A_j), 1)$, for a probability measure H_0 and $\alpha > 0$. Normalizing $\tilde{H}$ by $H(A_j) = \tilde{H}(A_j)/\tilde{H}(X)$ defines a DP prior with base measure αH_0.

Completely random measures and the normalized generalized gamma process

Replacing the gamma process by any other CRM defines alternative BNP priors for random probability measures. Such priors are known as normalized random measures with independent increments (NRMI) and were first described in Regazzini et al. (2003); they include a large number of BNP priors. A recent review of NRMIs appears in Lijoi & Prünster (2010). Besides the DP prior, another interesting example is the normalized generalized gamma process (NGG), discussed in Lijoi et al. (2007). We write $H \sim \mathrm{NGG}(\alpha, \kappa, \gamma, H_0)$. The NGG is indexed by a total mass parameter $\alpha > 0$, two more scalar parameters $\kappa \geq 0$ and $\gamma \in [0, 1)$, and a base probability measure H_0. In fact, the DP is a special case of the NGG with $\kappa = 1$ and $\gamma = 0$.

As for any CRM, a realization from the generalized gamma process (before normalization) can be generated using the following constructive definition (Kingman, 1993, Section 8.2). Assume we wish to generate a random measure $\tilde{H}$ on a measurable space X, for example $\mathbb{R}^d$. We set up a Poisson process over $\mathbb{R}^+ \times X$ with Poisson intensity $\nu(\tilde{\eta}, \mu)$. The choice of $\nu(\cdot)$ determines different CRMs. The arguments are already labeled in anticipation of the next step. For the generalized gamma process we use

$$\nu(\tilde{\eta}, \mu) = \nu_1(\tilde{\eta})\, \alpha H_0(\mu), \quad \text{with } \nu_1(\tilde{\eta}) = e^{-\kappa\tilde{\eta}} \tilde{\eta}^{-(1+\gamma)} / \Gamma(1-\gamma). \tag{6.13}$$

Let $(\tilde{\eta}_g, \mu_g),\ g = 1, \ldots,$ denote a realization of this Poisson process. Then $\tilde{H} \propto \sum \tilde{\eta}_g \delta_{\mu_g}$ is a realization of the desired CRM. Here, $\tilde{H}$ is still the (non-normalized) CRM. The normalized measure H rescales the weights $\tilde{\eta}_g$ to unit total mass.

NGG mixtures

Barrios et al. (2013) discuss mixture models with an NGG prior on the mixing measure, similar to (6.8), but with an NGG prior replacing the DP prior:

$$y_i \mid \theta \sim p(y_i \mid \theta_i), \quad \theta_i \mid H \sim H, \quad H \sim \text{NGG}(\alpha, \kappa, \gamma, H_0). \tag{6.14}$$

The discussion in Barrios et al. (2013) is more general, allowing for any other NRMI, but the NGG is a sufficiently rich model for most purposes. However, in comparison to the DP prior the additional flexibility of the NGG is important for modeling. This is extensively discussed in De Blasi et al. (2014) and Barrios et al. (2013). For example, consider two clusters k and ℓ with cluster sizes $n_k > n_\ell$. As before, let z_i denote a latent cluster membership indicator in an equivalent hierarchical model version of (6.14), let $\mathbf{z}_{-i} = (z_j,\ j \neq i)$, and define $n_k^{-i} = \sum_{j \neq i} \mathbb{I}(z_j = k)$ to be cluster sizes without the ith unit. Then *a priori* $\text{P}(z_i = k \mid \mathbf{z}_{-i})/\text{P}(z_i = \ell \mid \mathbf{z}_{-i}) = (n_k^{-i} - \gamma)/(n_\ell^{-i} - \gamma)$. The implication for data analysis is that cluster sizes under NGG priors with $\gamma > 0$ tend to be more concentrated, with few large clusters including most experimental units. Perhaps more importantly, the implied prior on the number of clusters, G_+, is more flexible under the NGG prior, in the sense that for matching prior means, hyperprior parameters can be chosen to allow for substantially more prior variance for G_+. This allows the number of clusters to be *a posteriori* adjusted as needed for the data. Under the DP prior, $p(G_+)$ is centered around approximately $\alpha \log(n)$. That is, prior centering determines α, leaving no more flexibility to inflate prior variance.

6.3.2 Posterior simulation for normalized generalized gamma process mixtures

Most importantly, posterior inference under (6.14) is still easily implemented. Barrios et al. (2013), Favaro & Teh (2013), and Argiento et al. (2010) outline specific MCMC algorithms. These are based on a representation of the posterior distribution for NRMIs discussed in James et al. (2009), under independent sampling, $\theta_i \sim H$ as in (6.14), and an NRMI prior for H. Details of the general result are not needed for the upcoming algorithm for NGG mixtures. We only outline the setup, and give specific details for the NGG mixture. The representation involves a model augmentation of the posterior $p(\theta, \tilde{H} \mid \mathbf{y})$ under (6.14) with a latent variable, u, using

$$p(u \mid \theta, \tilde{H}, \mathbf{y}) = p(u \mid G_+) \propto u^{n-1}(u + \kappa)^{G_+ \gamma - n} e^{-\alpha(u+\kappa)^\gamma/\gamma}. \tag{6.15}$$

For the following description of the algorithm it is convenient to distinguish atoms of H that are matched with currently imputed θ_i versus unmatched atoms. Also posterior inference is most easily discussed for the random measure $\tilde{H}$, before normalization. As before, let $\{\theta_k^\star,\ k = 1, \ldots, G_+\}$ denote the unique θ_is. Then

$$\tilde{H} = \sum_{k=1}^{G_+} \eta_k^\star \delta_{\theta_k^\star} + \tilde{H}_C, \quad \text{with } \tilde{H}_C = \sum_{g=1}^{\infty} \tilde{\eta}_g \delta_{\mu_g}. \tag{6.16}$$

Note that the split of $\tilde{H}$ in (6.16) implicitly is a function of θ (to identify the unique $\theta_k^\star$) and can only be used when conditioning on θ. In the MCMC implementation we approximate $\tilde{H}_C$ by using the G terms with largest $\tilde{\eta}_g$ only. This is possible since the algorithm for generating $\tilde{H}_C$ samples the $\tilde{\eta}_g$ in decreasing order; see below. We can therefore assume that the weights are indexed by decreasing order, $\tilde{\eta}_g \geq \tilde{\eta}_{g+1}$. Let $\tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_G)$, $\mu = (\mu_1, \ldots, \mu_G)$. Finally, let $\eta^\star = (\eta_1^\star, \ldots, \eta_{G_+}^\star)$ and $\boldsymbol{m} = (n_1, \ldots, n_{G_+})$.

An MCMC scheme for normalized generalized gamma process mixture models

We describe the particular Gibbs sampling implementation that is proposed in Barrios et al. (2013). The following steps define transition probabilities of an MCMC scheme for the posterior distribution $p(\theta, \tilde{H}, u \mid \mathbf{y})$ under (6.14), augmented with (6.15). The algorithm includes five transition probabilities. Let $[a \mid b, c]$ indicate sampling from the conditional distribution of parameter a given b, c. All distributions are complete conditional posterior distributions, with the absence of any variables in the conditioning set indicating conditional indpendence. In some cases dependence on θ is only indirectly through $\mathbf{z}$, or even just G_+ or n_j. The five steps are: (i) $[u \mid G_+]$; (ii) $[\eta_k^\star \mid u, n_k]$; (iii) $[\theta_k^\star \mid \mathbf{z}, \mathbf{y}]$; (iv) $[\eta, \mu \mid u]$; (v) $[\theta_i \mid \tilde{H}]$.

In step (i) we generate u from (6.15). Favaro & Teh (2013) recommend instead sampling $v = \log(u)$, as the complete conditional distribution $p(v \mid G_+)$ turns out to be log concave, allowing for easier random variable generation. In step (ii) we update $\eta_k^\star$, $k = 1, \ldots, G_+$, by generating from the complete conditional posterior which under the NGG simplifies to

$$\eta_k^\star \mid u, n_j \sim \mathcal{G}(n_j - \gamma, \kappa + u).$$

In step (iii) we draw from the complete conditional posterior distribution for $\theta_k^\star$. This step is identical to (6.9).

Step (iv) updates $\tilde{H}_C$. James et al. (2009) show that, conditional on u, the random $\tilde{H}_C$ is again a CRM with Poisson intensity $\nu^\star(\tilde{\eta}, \mu)$, that is, replacing the original $\nu(\cdot, \cdot)$ by an updated Poisson intensity $\nu^\star(\cdot, \cdot)$. In the case of the NGG this simpifies to

$$\tilde{H}_C \sim \mathrm{NGG}(\alpha, \kappa + u, \gamma, H_0).$$

We can use the following easy algorithm to generate $\tilde{H}_C = \sum \tilde{\eta}_g \delta_{\mu_g}$. Ferguson & Klass (1972) introduce a clever scheme to generate the weights $\tilde{\eta}_g$ in decreasing order. The construction requires a function $N(v) = \int_v^\infty \nu_1(\tilde{\eta}) d\tilde{\eta}$, using the factor $\nu_1(\tilde{\eta})$ from definition (6.13), with $\kappa^\star = \kappa + u$ in place of κ. That is, $N(v) = \frac{\alpha}{\Gamma(1-\gamma)} \int_v^\infty e^{-(\kappa+u)\tilde{\eta}} \tilde{\eta}^{-(1+\gamma)} d\tilde{\eta}$. Next let $\xi_1, \xi_2, \ldots$ denote a realization from a unit-rate Poisson process, that is, $\xi_g - \xi_{g-1} \sim \mathcal{E}(1)$ are i.i.d. exponential draws (starting with $\xi_0 = 0$). Then

$$\tilde{\eta}_g = N^{-1}(\xi_g).$$

In words, plot the function $N(v)$ against $v \geq 0$, mark the ξ_g on the vertical axis, and then use $N^{-1}(\cdot)$ to map $\xi_1, \xi_2, \ldots$, to the horizontal v-axis. The construction delivers $\tilde{\eta}_g$, already ordered by decreasing size. The construction of $\tilde{H}_C$ is completed by generating the locations $\mu_g \sim H_0$, i.i.d.

Finally, in step (v) we resample θ by generating $p(\theta_i \mid \tilde{H}, y_i) \propto \tilde{H}(\theta_i) p(y_i \mid \theta_i)$. Write $\tilde{H} = \sum_{\ell=1}^\infty \tilde{w}_\ell \delta_{m_\ell}$, using a single running index ℓ for all terms in (6.16). Then $\mathrm{P}(\theta_i = m_\ell \mid y_i) \propto \tilde{w}_\ell \, p(y_i \mid m_\ell)$.

The described MCMC scheme is implemented in the R package `BNPdensity`, which is available in the CRAN package repository http://cran.r-project.org/.

An alternative MCMC scheme for posterior inference under model (6.14) is described in Favaro & Teh (2013) and also in Argiento et al. (2010). Favaro & Teh (2013) describe what can be characterized as a modified version of the Pólya urn. Recall that the Pólya urn (6.6) defines the marginal distribution of $(\theta_1, \ldots, \theta_n)$ under the DP prior, after marginalizing with respect to H. Similarly, Favaro & Teh (2013) describe a method for sampling $p(\theta_1, \ldots, \theta_n \mid u, \mathbf{y})$, marginalizing with respect to H. Generating $u \mid \theta$ proceeds as in step (i) above. Additionally, they describe the complete conditional posterior distributions for the NGG hyperparameters. This allows model (6.14) to be augmented with a hyperprior on the NGG parameters.

6.4 Bayesian Nonparametric Mixtures with Random Partitions

Recall that we started the discussion by observing that a mixture model (6.1) can naturally be thought of as a mixture with respect to a mixing measure, as in (6.2), and we proceeded by assuming BNP priors on the mixing measure.

There is another feature of hierarchical models like (6.2) with a discrete BNP prior $p(H)$ that naturally leads to a mixture model. Consider the posterior predictive distribution $F_{n+1}(y_{n+1}) = p(y_{n+1} \mid \mathbf{y})$. With an argument similar to (6.10) for $i = n+1$, but without conditioning on y_{n+1} and with instead an additional convolution with $p(y_{n+1} \mid z_{n+1} = k, \mathbf{y})$, we find

$$F_{n+1}(y_{n+1}) \equiv p(y_{n+1} \mid \mathbf{y}) \propto \sum_{k=1}^{G_+} n_k p(y_{n+1} \mid \mathbf{y}_k^\star) + \alpha h_0(y_{n+1}). \tag{6.17}$$

Let $F(y) = \int p(y \mid \theta)\, dH(\theta)$, as earlier, and let $\overline{F} = \mathrm{E}\,(F \mid \mathbf{y})$ denote the posterior expectation. Then $F_{n+1} = \overline{F}$. That is, the posterior predictive distribution (6.17) coincides with the posterior expectation of the random probability measure. This is easily seen by considering $\mathrm{P}(y_{n+1} \le c \mid \mathbf{y}) = \mathrm{E}(\mathrm{P}(y_{n+1} \le c \mid F, \mathbf{y}) \mid \mathbf{y}) = \mathrm{E}(F(c) \mid \mathbf{y})$. Here, we overload notation to let $F(c)$ indicate the cdf under the probabilty measure F.

In the outlined construction the nature of F_{n+1} as a mixture model arises from the implied random partition $p(\mathcal{C}_n)$ under i.i.d. sampling $\theta_i \sim H$ from the discrete random probability measure $H = \sum \eta_g \delta_{\mu_g}$. In that case the mixture model F_{n+1} is just another manifestation of the assumed mixture model $F(y) = \sum \eta_g p(y \mid \theta_g)$, and does not introduce fundamentally new structure. In fact, any exchangeable random partition $p(\mathcal{C}_n)$ can be argued to arise from such a construction. See, for example, Lee et al. (2013b) for a review. The attraction of exchangeable random partitions is coherence and mathematical tractability.

However, if the inference goal is a posterior predictive distribution in the form of a mixture model, as in (6.17), then the same form can be achieved with any underlying random partition model $p(\mathcal{C}_n)$, including possibly non-exchangeable random partitions.

6.4.1 Locally weighted mixtures

An attractive general framework for random partitions are the product partition models (PPMs; Hartigan, 1990) which take the form

$$p(\mathcal{C}_n = \{C_1, \ldots, C_{G_+}\}) \propto \prod_{j=1}^{G_+} c(C_j)$$

for some functions $c(C_j)$, which are known as the *cohesion* functions. The cohesion functions are restricted to be non-negative functions of C_j, but in principle any such function is valid. If $c(C)$ is only a function of the size of C, then the resulting model for $\mathcal{C}_n$ is invariant under permutations of the indices. If additionally $p(\mathcal{C}_n) = \sum_{z_{n+1}} p(\mathcal{C}_{n+1})$, then we are back to exchangeable partitions. For example, with $c(C) = \alpha \times (|C| - 1)!$ the PPM reduces to the Pólya urn (6.6). More general, it can be shown that the family of all PPM models with cohesion function $c(C) = c(|C|)$ that depend on C only indirectly through the cardinality and that define exchangeable random partitions coincides with the family of so-called Gibbs-type priors. See, for example, De Blasi et al. (2014) for a discussion. A subset of the Gibbs-type priors, in turn, are the NGG models that we discussed before (De Blasi et al., 2014). Another notable Gibbs-type prior is the Pitman–Yor model, of which incidentally the DP is again a special case (De Blasi et al., 2014).

However, abandoning the restriction to exchangeable partitions allows for other interesting variations of PPM models. Consider, for example, the problem of clustering patients in a clinical trial. Usually important baseline covariates x_i are available for each patient, and it might be desireable to favor clusters of patients who are more homogeneous with respect to these baseline covariates. Let $\mathbf{x}_k^\star = (x_i;\ i \in C_k)$ denote the baseline covariates arranged by cluster, and similarly for $\mathbf{y}_k^\star$. Müller et al. (2011) introduce the PPMx model by replacing the cohesion function $c(C_k)$ in a PPM with $c(C_k)g(\mathbf{x}_k^\star)$. Here $g(\mathbf{x}_k^\star)$ is termed a similarity function. It is any function that penalizes for lack of heterogeneity of $\mathbf{x}_k^\star$. For example, for categorical covariates $g(\mathbf{x}_k^\star) = 1/m_k$ could be related to the number m_k of distinct values x_i, $i \in C_k$. The implied posterior predictive distribution is a locally weighted mixture. Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{z} = (z_1, \ldots, z_n)$, and write $y = y_{n+1}$, $x = x_{n+1}$ and $z = z_{n+1}$ for the variables of a future patient. Then

$$\eta_k(x) \equiv \mathrm{P}(z = k \mid x, \theta, \mathbf{x}, \mathbf{z}) \propto \frac{g(\mathbf{x}_k^\star \cup x)c(C_k \cup \{n+1\})}{g(\mathbf{x}_k^\star)c(C_k)}$$

for $k = 1, \ldots, G_+, G_+ + 1$, with the understanding that the denominator evaluates as 1 for $k = G_+ + 1$. And the posterior predictive distribution for a future observation $y = y_{n+1}$ with covariates $x = x_{n+1}$, still conditional on θ and $\mathbf{z}$, becomes

$$p(y \mid x, \mathbf{x}, \mathbf{z}) = \sum_{k=1}^{G_+} \eta_k(x) p(y \mid x, \mathbf{x}_k^\star, \mathbf{y}_k^\star) + \eta_{G_+ + 1}(x) h_0(y \mid x), \tag{6.18}$$

where $p(y \mid x, \mathbf{x}_k^\star, \mathbf{y}_k^\star) = \int p(y \mid x, \theta_k^\star)\, dp(\theta_k^\star \mid \mathbf{x}_k^\star, \mathbf{y}_k^\star)$ and $h_0(y \mid x) = \int p(y \mid x, \theta)\, dH_0(\theta)$. The form of the predictive distribution is a mixture over the clusters of a random partition, similar to F_{n+1} in (6.17), but now as a locally weighted mixture with weights $\eta_k(x)$ that are indexed by the covariate x. Marginalizing with respect to θ and $\mathbf{z}$, the posterior predictive distribution $p(y_{n+1} \mid x_{n+1}, \mathbf{y})$ adds additional posterior averaging with respect to $p(\theta, \mathbf{z} \mid \mathbf{y})$, that is, $F_{n+1}(y \mid x, \mathbf{y}) = \mathrm{E}(p(y \mid x, \theta, \mathbf{x}, \mathbf{z}) \mid \mathbf{y})$. The form of (6.18) as a mixture with covariate dependent weights defines a mixture of experts model as discussed in Chapter 12.

6.4.2 Conditional regression

An alternative, natural way to introduce similar covariate dependent mixtures, again as a posterior predictive distribution $p(y_{n+1} \mid \mathbf{y}, \mathbf{x}, x_{n+1})$ in a BNP model, is the following construction. We proceed as if the pairs (x_i, y_i) were independent random samples from a joint distribution $(x_i, y_i) \sim F$, complete the model with a DPM on F, and report inference on F. This is introduced in Müller et al. (1996) and Park & Dunson (2010). The construction is easiest when both x_i and y_i are continuous. Consider a DPM of normal kernels, mixing with respect to location and scale. Write the DPM as a hierarchical model as in (6.8), using the kernel $p(x, y \mid \theta = (\mu, \Sigma)) = \mathcal{N}(\mu, \Sigma)$:

$$(x_i, y_i \mid \theta_i) \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \Sigma_i), \quad \theta_i \sim H \text{ and } H \sim \mathcal{DP}(\alpha, H_0).$$

As before, let $\theta_k^\star = (\mu_k^\star, \Sigma_k^\star)$, $k = 1, \ldots, G_+$, denote the unique values of θ_i, $i = 1, \ldots, n$, with multiplicities n_k. Let $f(y \mid x, \theta_k^\star)$ denote the conditional normal density in y given x that is implied by the multivariate normal $\mathcal{N}(\mu_k^\star, \Sigma_k^\star)$, and let $\eta(x \mid \theta_k^\star)$ denote the marginal normal density in x under $\mathcal{N}(\mu_k^\star, \Sigma_k^\star)$. Similarly, let $f_0(y \mid x)$ and $\eta_0(x)$ denote the implied conditional and marginal when $\theta^\star$ is generated from H_0, that is, $f_0(y \mid x) = \int f(y \mid x, \theta)\, dH_0(\theta)$ and $\eta_0(x) = \int \eta(x \mid \theta)\, dH_0(\theta)$. The posterior predictive distribution for

a future observation $y = y_{n+1}$ with covariates $x = x_{n+1}$ takes the form

$$p(y \mid x, \theta^\star) \propto \alpha\, \eta_0(x) f_0(y \mid x) + \sum_{k=1}^{G_+} n_k\, \eta(x \mid \theta_k^\star)\, f(y \mid x, \theta_k^\star),$$

similar to (6.18). The construction introduced an additional – one could argue, inappropriate – factor in the likelihood by including x_i in the hypothetical multivariate response (x_i, y_i). The approach only works easily when x_i and y_i are both continuous. In the case of mixed data types it is more natural to separate the two, as in the PPMx model.

Dependent Dirichlet process mixtures

Finally, recall the construction of DDP model. This provides yet another approach to construct covariate dependent mixture models, now on the basis of a BNP prior $p(H_x,\ x \in X)$ on the mixing measure in a mixture, now indexed with the covariate x. This construction is not much used.

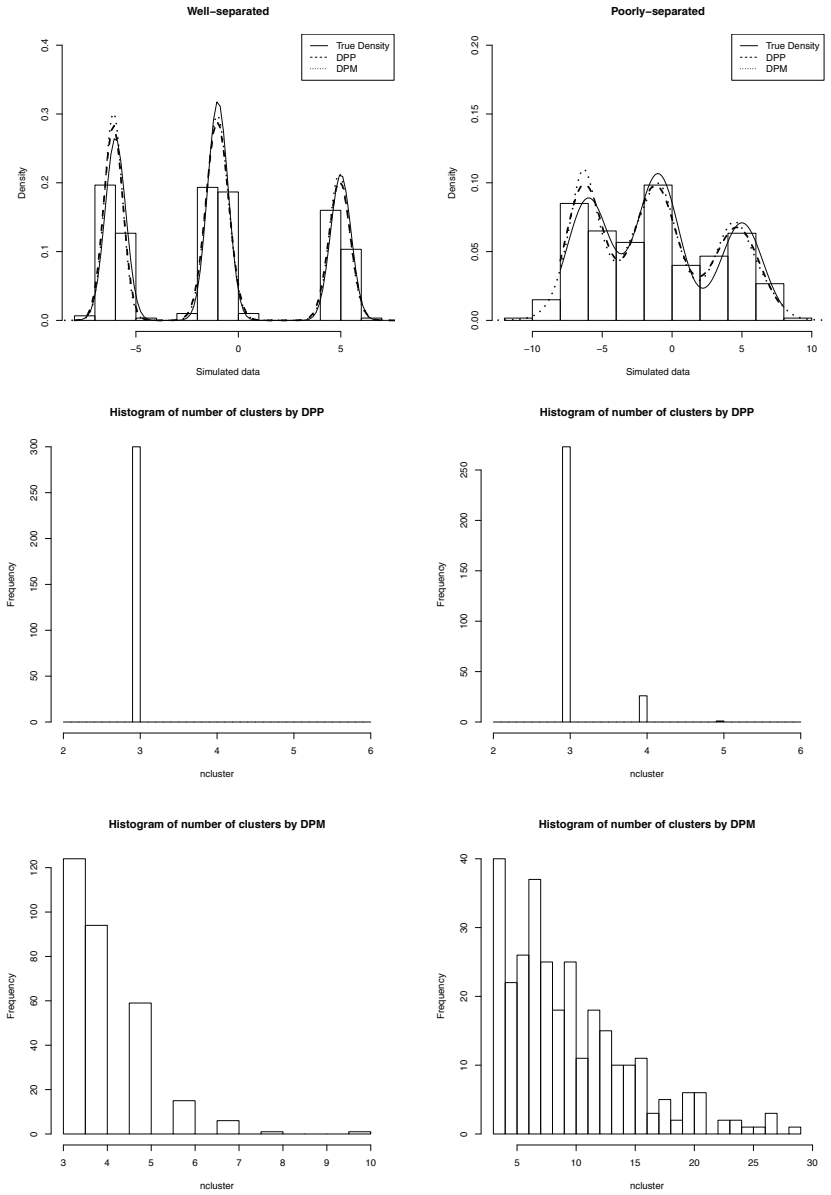
6.5 Repulsive Mixtures (Determinantal Point Process)

In many applications investigators seek to interpret the components of a mixture model, or equivalently, the clusters in a random partition model, as scientifically meaningful structure. For example, researchers have recognized that breast cancer is highly heterogeneous and different subgroups involve different disease mechanisms. Depending on the nature of the experiment and the data, a mixture model such as (6.1) could be used to recover such latent subgroups. For example, investigators might identify molecular drivers based on gene expression profiles of different tumor subtypes. In such applications, maximizing the diversity of inferred molecular drivers helps identify and interpret these distinct mechanisms.

However, the use of independent priors for component-specific parameters μ_g in (6.1) and, similarly, $\theta_g^\star$ in (6.2) does not provide any prior regularization to favor such diversity. To the contrary, the independent prior gives rise to concerns about overfitting which generates redundant mixture components with similar parameters, leading to unnecessarily complex models and poor interpretability. In particular, this compromises the interpretation of the mixture components as biologically meaningful structure. Rousseau & Mengersen (2011) argue that concerns are partially mitigated with carefully chosen priors which lead to redundant structure being asymptotically removed; see also Chapter 4. Alternatively, Petralia et al. (2012) proposed a class of repulsive priors for mixture components. The proposed repulsive prior is based on a distance metric in which small distances are penalized. However, posterior computations are complex and do not easily scale to higher-dimensional problems.

Determinantal Point Process

Xu et al. (2016) argue for the use of a determinantal point process (DPP) to build an alternative prior that imposes the desired repulsive prior regularization across mixture components. The DPP is a point process that generates random configurations $\{\mu_1, \ldots, \mu_G\}$ in a way that highly favors very diverse atoms μ_g. The cardinality G is part of the random configuration. The use of DPP priors for statistical inference in mixtures was first proposed in aff (????). Xu et al. (2016) generalize the setup by discussing more general inference for



$\sigma = 0.5$ $\sigma = 1.5$

FIGURE 6.5
The top panel plots show the histograms of two simulated data sets with true density (solid), estimated density by DPP prior (dashed) and DPM (dotted). The middle and bottom panels present the histograms of the estimated number of clusters by DPP prior and DPM, respectively.

latent structures and by developing an efficient transdimensional posterior MCMC scheme that allows inference across G.

The DPP defines a point process on $C \subseteq \mathbb{R}^d$. We first define a point process for a finite state space, $S = \{\omega_1, \ldots, \omega_R\}$, for example a grid in $\mathbb{R}^d$ (in that case R would be the

number of grid cells). Let C denote an $(R \times R)$ positive semidefinite matrix, constructed, for example, as $C_{ij} = C(\omega_i, \omega_j)$ with a covariance function $C(\omega_i, \omega_j)$. Let C_A denote the submatrix of rows and columns indicated by $A \subseteq S$. We define a random point configuration $X = \{\mu_1, \ldots, \mu_G\}$. The μ_g will later become the component-specific parameters, when we use the DPP to construct a prior probability model for a mixture. We define

$$\mathrm{P}(X = A) \propto \det(C_A) \tag{6.19}$$

as a probability distribution on the 2^n possible point configurations $X \subset S$. This defines a subclass of DPPs known as L-ensembles. It is easy to see why the DPP defines a repulsive point process if one interprets the determinant as the volume of a parallelotope spanned by the column vectors of C_A. Equal or similar column vectors span less volume than very diverse ones. A good review of DPP models for finite state spaces, including the derivation of the normalizing constant in (6.19), appears in Kulesza & Taskar (2012).

For a continuous state space $C \subseteq \mathbb{R}^d$, we define an L-ensemble by a density $f(X)$ with respect to the unit-rate Poisson process as

$$f(X) \propto \det(C_X), \tag{6.20}$$

for $X = \{\mu_1, \ldots, \mu_G\}$. As before, C_X is a $(G \times G)$ matrix with (i, j)th entry defined by a continuous covariance function $C(\mu_i, \mu_j)$. We write $X \sim \mathrm{DPP}(C)$, or $X \sim \mathrm{DPP}(C, \phi, \tau)$ when ϕ and τ are included as unknown hyperparameters in the definition of the covariance function $C(\mu_i, \mu_j)$.

Determinantal point process mixtures

Xu et al. (2016) propose the DPP mixture model as the sampling model (6.1) with prior

$$(\mu_1, \ldots, \mu_G) \sim \mathrm{DPP}(C), \quad \eta \mid G, \delta \sim \mathcal{D}(\delta, \ldots, \delta).$$

The number of terms G in the mixture is part of the random point configuration. Posterior inference is implemented as transdimensional MCMC. Essentially, the density (6.20) with respect to the unit-rate Poisson process can be used in a reversible jump MCMC as if it were a density with respect to the Lebesgue measure. In the Metropolis–Hastings acceptance probability the density $f(X)$ is replaced by $f(X)/G!$ (Xu et al., 2016).

Figure 6.5 summarizes a small simulation study using the proposed DPP mixture model, with a normal kernel $y_i \mid \mu_g \sim \mathcal{N}(\mu_g, \sigma^2)$ and an additional gamma prior on the common precision parameter, $1/\sigma^2 \sim \mathcal{G}(a_0, b_0)$. We used a squared exponential covariance function. The implemented model also includes hyperparameters for the covariance function. See Xu et al. (2016) for details about inference for the covariance function hyperparameters. The top panel shows the simulation truth, density estimates under the DPP model and under an equivalent DPM model for two simulated data sets. The middle and bottom panels show the posterior distribution of the number G of clusters under the DPP prior and DPM, respectively. Both the DPP and DPM models lead to very similar density estimates. However, inference under the DPP reports far fewer clusters, simply because of its repulsive feature.

6.6 Concluding Remarks

Starting from an interpretation of a generic mixture model as an expectation with respect to a random mixing measure, we discussed BNP approaches to inference in mixture models.

In some cases the BNP nature of a mixture model is only one of perspective. For example, inference under a finite mixture model can be meaningfully seen as inference under a BNP model when the focus is on inference for the implied (finite) discrete mixing measure or when the finite mixing measure is constructed as an approximation of an infinite discrete BNP prior. This is the case, for example, for the popular finite DP prior.

We reviewed BNP priors for the mixing measure in mixtures of a parametric kernel with respect to a nonparametric model. Taking this perspective, we excluded in particular a discussion of mixtures of BNP models, that is, mixtures of nonparametric models with respect to parametric (hyper)priors. A common example of such models are mixtures of Pólya tree models which are used to mitigate the lack of smoothness of density estimates under a Pólya tree prior (Hanson & Johnson, 2002).

Another related large class of BNP models generalizes the random partition models that we briefly described in this chapter. A partition is a family of non-overlapping subsets of $[n]$. More general feature allocation models generate instead possibly overlapping subsets, also without the requirement that the union be $[n]$. One of the most widely used feature allocation models is the Indian buffet process; see, for example, Broderick et al. (2013) for a good review of random partitions and feature allocation models.

Finally, another interesting class of BNP models in mixtures that we did not discuss in this review are hierarchical extensions of the DPM. The hierarchical DP of Teh et al. (2006) defines multiple DPM models, with a hierarchical hyperprior that allows the mixtures to share atoms across submodels. Similar models are discussed in Rodríguez et al. (2008) and Lee et al. (2013a), with different notions of hierarchical extensions and ways of linking the submodels; see also Chapter 17 for further BNP models and their application in finance.

Bibliography

(????).

Argiento, R., Guglielmi, A. & Pievatolo, A. (2010). Bayesian density estimation and model selection using nonparametric hierarchical mixtures. *Computational Statistics & Data Analysis* **54**, 816–832.

Barrios, E., Nieto-Barajas, L. E. & Prünster, I. (2013). A study of normalized random measures mixture models. *Statistical Science* **28**, 313–334.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics* **1**, 353–355.

Broderick, T., Jordan, M. I. & Pitman, J. (2013). Cluster and feature modeling from combinatorial stochastic processes. *Statistical Science* **28**, 289–312.

Bush, C. A. & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Prünster, I. & Ruggiero, M. (2014). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 212–229.

Escobar, M. D. (1988). *Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distributions of the Means.* Unpublished doctoral thesis, Department of Statistics, Yale University.

Escobar, M. D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Favaro, S. & Teh, Y. W. (2013). MCMC for normalized random measure mixture models. *Statistical Science* **28**, 335–359.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Ferguson, T. S. & Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Annals of Mathematical Statistics* **43**, 1634–1643.

Ghoshal, S. & van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference.* Cambridge: Cambridge University Press.

Guindani, M., Sepúlveda, N., Paulino, C. D. & Müller, P. (2014). A Bayesian semi-parametric approach for the differential analysis of sequence counts data. *Applied Statistics* **63**, 385–404.

Hanson, T. & Johnson, W. O. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association* **97**, 1020–1033.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics: Theory and Methods* **19**, 2745–2756.

Hjort, N. L., Holmes, C., Müller, P. & Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge: Cambridge University Press.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

James, L. F., Lijoi, A. & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics* **36**, 76–97.

Jara, A. (2007). Applied Bayesian non- and semi-parametric inference using DPpackage. *Rnews* **7**, 17–26.

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P. & Rosner, G. L. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software* **40**, 1–30.

Karabatsos, G. (2014). Software user's manual for Bayesian regression: Nonparametric and parametric models. Tech. rep., University of Illinois-Chicago.

Kingman, J. F. C. (1993). *Poisson Processes*. Oxford: Oxford University Press.

Klein, J. P. & Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer-Verlag.

Kulesza, A. & Taskar, B. (2012). Determinantal point processes for machine learning. Preprint, arXiv:1207.6083.

Lee, J., Müller, P., Zhu, Y. & Ji, Y. (2013a). A nonparametric Bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association* **108**, 775–788.

Lee, J., Quintana, F., Müller, P. & Trippa, L. (2013b). Defining predictive probability functions for species sampling models. *Statistical Science* **28**, 209–222.

Lijoi, A., Mena, R. H. & Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society, Series B* **69**, 715–740.

Lijoi, A. & Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, N. L. Hjort, C. Holmes, P. Müller & S. G. Walker, eds. Cambridge: Cambridge University Press, pp. 80–136.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates I: Density estimates. *Annals of Statistics* **12**, 351–357.

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, Alexandria, VA.

Müller, P., Erkanli, A. & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.

Müller, P., Quintana, F., Jara, A. & Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Cham: Springer-Verlag.

Müller, P., Quintana, F. & Rosner, G. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics* **20**, 260–278.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Park, J.-H. & Dunson, D. (2010). Bayesian generalized product partition models. *Statistica Sinica* **20**, 1203–1226.

Petralia, F., Rao, V. & Dunson, D. B. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems 25*, F. C. N. Pereira, C. J. C. Burges, L. Bottou & K. Q. Weinberger, eds. Red Hook, NY: Curran Associates, Inc., pp. 1889–1897.

Phadia, E. G. (2013). *Prior Processes and Their Applications*. Heidelberg: Springer-Verlag.

Pitman, J. (2006). *Combinatorial Stochastic Processes*, vol. 1875 of *Lecture Notes in Mathematics*. Berlin: Springer-Verlag. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, with a foreword by Jean Picard.

Regazzini, E., Lijoi, A. & Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Annals of Statistics* **31**, 560–585.

Rodríguez, A., Dunson, D. B. & Gelfand, A. E. (2008). The nested Dirichlet process, with discussion. *Journal of the American Statistical Association* **103**, 1131–1144.

Rossi, P., Allenby, G. & McCulloch, R. (2005). *Bayesian Statistics and Marketing*. New York: John Wiley.

Rossi, P. & McCulloch, R. (2008). *bayesm: Bayesian inference for marketing/microeconometrics*. R package version 2.2-2.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society, Series B* **73**, 689–710.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2006). Sharing clusters among related groups: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566–1581.

Xu, Y., Müller, P. & Telesca, D. (2016). Bayesian inference for latent biologic structure with determinantal point processes (DPP). *Biometrics* **72**, 955–964.