Chapter 5

Dependent Dirichlet Processes and Other Extensions

5.1. Dependent Extensions of the DP

Many applications involve families of probability models $\mathcal{G} = \{G_x : x \in X\}$. For example, G_x could be the distribution of the time to progression for a patient with baseline covariates x. In that case, a prior $p(G_x : x \in X)$ would provide a nonparametric alternative to the popular but restrictive proportional hazards model. More generally, a nonparametric prior $p(\mathcal{G})$ can be used to define a fully non-parametric regression $p(y \mid x) = G_x(y)$. Another typical applications of prior models $p(\mathcal{G})$ is in the construction of mixed effects models, where $p(\mathcal{G})$ is used to define a random effects distribution $G_x(\cdot)$ for patients with covariates x.

Most popular prior models for $\mathcal{G}$ in the recent literature are based on extensions of the Dirichlet process (DP) model discussed in Chapter 3, which are collectively known as dependent Dirichlet process (DDP) models. We first consider the simplest case, with finitely many dependent RPMs $\mathcal{G} = \{G_j,\ j = 1, \ldots, J\}$ that are judged to be exchangeable, i.e., the prior model $p(\mathcal{G})$ should be invariant with respect to any permutation of the indices. This case could arise, for example, as a prior model for unknown random effects distributions G_j in related studies, $j = 1, \ldots, J$. To keep the upcoming discussion specific we will continue to refer to this motivating example. In words, we wish to define a prior probability model $p(\mathcal{G})$ that allows us to borrow strength across the J studies. Patients under study j_1 should inform inference about patients enrolled in another related study $j_2 \neq j_1$. Two extreme modeling choices would be (i) to pool all patients and assume one common random effects distribution, or (ii) to assume J distinct random effects distributions with independent priors. Formally the earlier choice assumes $G_j \equiv G$, $j = 1, \ldots, J$ with a prior $p(G)$. The latter assumes $G_j \sim p(G_j)$, independently, $j = 1, \ldots, J$. We refer to the two choices as extremes since the first choice implies maximum borrowing of strengths, and the other choice implies no borrowing of strength. In most applications, the desired level of borrowing strength is somewhere in-between these two extremes.

Figure 5.1 illustrates the two modeling approaches. Note that in Figure 5.1 we added a hyperparameter η to index the prior model $p(G_j \mid \eta)$ and $p(G \mid \eta)$, which was implicitly assumed fixed. The use of a random hyperparameter η allows for some borrowing of strength even in the case of conditionally independent $p(G_j \mid \eta)$. Learning across studies can happen through learning about the hyperparameter η. This is exactly the construction in Cifarelli and Regazzini (1978), which was used in, among others, Muliere and Petrone (1993) and Mira and Petrone (1996). However, the nature of the learning across studies is determined by the parametric form of η. This is illustrated in Figure 5.2. Assume $G_j \sim \mathsf{DP}(\alpha, G^\star_\eta)$, independently, $j = 1, 2$ and a base measure $G^\star_\eta = \mathsf{N}(m, B)$ with unknown hyperparameter $\eta = (m, B)$.

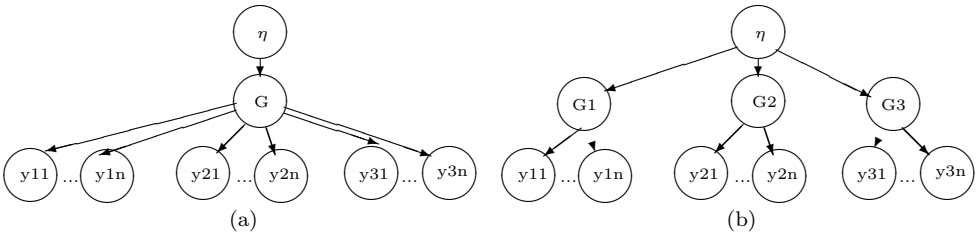


FIG 5.1. *One common RPM G (panel a) versus distinct RPMs G_j, independent across studies (panel b).*

In this case, prediction for a future study G_3 can not possibly learn about the multimodality of G_1 and G_2, beyond general location and orientation.

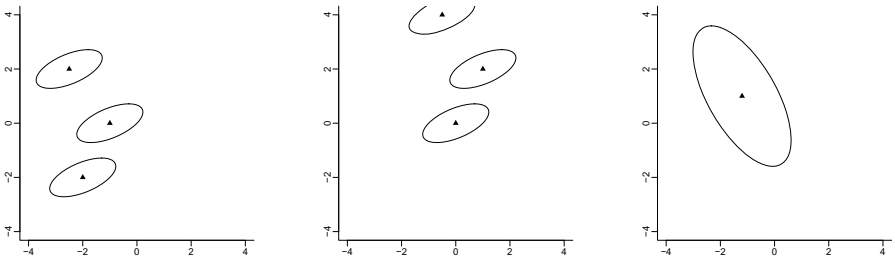
A natural next step in the model elaboration would now be to consider more complex choices for the hyperparameter η. Ideally, when $G^\star = \eta$ is an RPM itself, then we could potentially achieve arbitrary learning across the studies. This is exactly the construction of the hierarchical and nested DPs. See §5.4.2 and §5.4.3. However, these approaches are not suitable to model more general types of data such as spatial and/or temporal data. In §5.2 we introduce a still more general and widely used extensions of the DP that achieves the desired borrowing of strength while preserving the computational advantages of the DP.

5.2. Dependent DP (DDP)

MacEachern (1999) introduced what has meanwhile become by far the most commonly used prior for dependent RPMs, $p(G_x : x \in X)$. The model is known as the dependent DP (DDP). The beauty of the model is the elegance and simplicity of the construction. Recall the stick breaking representation of a DP random measure $G \sim \mathsf{DP}(M, G^\star)$, where

$$G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_h}(\cdot),$$

$m_h \sim G^\star$, independently across h and $w_h = v_h \prod_{g<h}(1-v_h)$ with $v_h \sim \mathsf{Beta}(1, M)$,



(a) $E(G_1 \mid \boldsymbol{y})$ (b) $E(G_2 \mid \boldsymbol{y})$ (c) Prediction for G_3

FIG 5.2. *$G_j \sim \mathsf{DP}(M, G^\star)$ with common $G^\star = \mathsf{N}(m, B)$. Learning across studies is restricted to the parametric form of η.*

i.i.d. The DDP uses the same construction for each G_x,

$$G_x(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{m_{x,h}(\cdot)}. \tag{5.1}$$

Here $\boldsymbol{m}_h = \{m_{x,h} : x \in X\}$ are independent realizations from a stochastic process on X (such as a Gaussian process), and the w_hs are constructed as before. Keeping $m_{x,h}$ independent across h ensures that each G_x marginally follows a DP prior. The simple, yet powerful idea of the DDP construction is to introduce dependence over x, i.e., to link the G_x through dependent locations of the point masses. Implicit in the notation used in (5.1) is the definition of weights w_h that are common across x. This variation of the DDP model is sometimes referred to as "common weight" (or "single p") DDP. However, the proposal in MacEachern (1999) is more general than (5.1), allowing also varying weights $w_{x,h}$. This more general construction is used, for example, in the time series DDP proposed in Nieto-Barajas *et al.* (2008) who define a DDP prior for a time series $\{G_t,\ t = 1, \ldots, T\}$ of random probability measures by introducing dependence of the weights $w_{t,h}$ (see §5.5.2).

Griffin and Steel (2006) define another interesting variation of the basic DDP by keeping both sets of parameters, locations and weights, unchanged across x. Instead they use permutations of how the weights are matched with locations. The permutations change with x. One advantage of such models is the fact that the support of G_x remains constant over x, a feature that can be important for extrapolation beyond the observed data.

5.3. ANOVA DDP

De Iorio *et al.* (2004) and De Iorio *et al.* (2009) define the ANOVA DDP as a variation of the DDP that is particularly useful for multivariate categorical covariates x. For illustration, assume $x = (u, v)$ for two categorical factors u and v. For example $G_{u,v}$ could be the random effects distribution for patients who are treated in related multi-arm clinical studies. Here u could be indexing related studies and v could be the different treatment arms.

The simplest form of dependence for a set $\{m_{x,h}\}$ of random variables indexed by two categorical covariates $x = (u, v)$ is an ANOVA model with main effects for u and v. This is exactly the model used in De Iorio *et al.* (2004). In particular, we assume $m_{x,h} = \mu_h + \alpha_{h,u} + \beta_{h,v}$ for $u \in \{0, \ldots, U\}$ and $v \in \{0, \ldots, V\}$, and assign normal priors on $\mu_h, \alpha_{h,u}$ and $\beta_{h,v}$, with $\alpha_{h,0} = \beta_{h,v} = 0$ for identifiability. Also, let $\boldsymbol{\theta}_h = (\mu_h, \alpha_{h,u}, \beta_{h,v},\ u = 1, \ldots, U,\ v = 1, \ldots, V)'$ denote the column vector of all ANOVA effects. Finally, let $G^\star(\boldsymbol{\theta}_h)$ denote the joint normal prior on the ANOVA effects. We write $\{G_x,\ x \in X\} \sim \mathsf{ANOVA\ DDP}(G^\star, M)$.

Implementation becomes particularly easy when the ANOVA DDP model is used in a DPM model, i.e., the random G_x is convoluted with an additional kernel, for example

$$y_i \mid m_i \sim \mathsf{N}(m_i, s^2), \qquad m_i \mid x_i = x, \mathcal{G} \sim G_x,$$

with $\{G_x,\ x \in X\} \sim \mathsf{ANOVA\ DDP}(G^\star, M)$. In this case, inference can be reduced to a standard DP mixture of normal model. To do so, let $\boldsymbol{d}_i$ denote a design vector that selects the relevant ANOVA factors corresponding to x_i for observation y_i. We can equivalently write

$$y_i \mid \boldsymbol{\theta}_i \sim \mathsf{N}(\boldsymbol{d}_i'\boldsymbol{\theta}_i, s^2), \qquad \boldsymbol{\theta}_i \mid F \sim F, \tag{5.2}$$

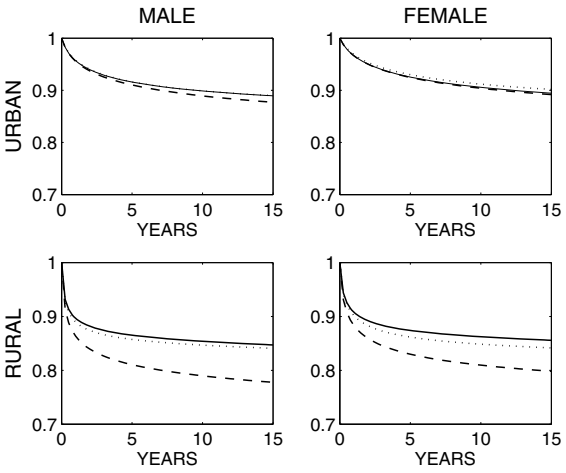


FIG 5.3. *Posterior estimated survival functions $E(S_x \mid \text{data})$, arranged by sex and rural vs. urban birth place. The three curves in each panel correspond to the three birth cohorts.*

with $F \sim \mathsf{DP}(F^\star, M)$, a DP mixture of normal linear models. The mixing measure G in the DP mixture is a probability model for complete vectors of ANOVA factors, $F = \sum_h w_h \delta_{\theta_h}$. Inference can therefore proceed like in a standard DP mixture model. Although the description implicitly assumed univariate outcomes y_i, extending the model to multivariate outcomes is straightforward using corresponding multivariate ANOVA models for $m_{x,h}$.

Example 13 (ANOVA DDP) *De Iorio* et al. *(2009) use an ANOVA DDP prior to implement non-parametric survival regression with multiple covariates. We apply their model to analyze data on childhood mortality in Columbia (Somoza, 1980). The dataset includes observations for 1437 children (using only the oldest child for each mother) and covariates including gender (binary), birth cohort (categorical with 3 levels), and an indicator for the child being born in a rural area (binary). The dataset includes extensive censoring, with 87% of the children alive at the time of observation.*

Let $S_x(t)$ denote the probability of a child with covariates x surviving beyond time t. The ANOVA DDP defines a prior probability model on $\mathcal{G} = \{S_x;\ x \in X\}$. Figure 5.3 shows point estimates of the survival functions for all combinations of gender, rural and birth cohort as $E(S_x \mid \text{data})$. However, posterior inference under the nonparametric ANOVA DDP model delivers more than point estimates. Figure 5.4 shows pointwise central 95% posterior probability intervals for $S_x(t)$.

5.4. Multilevel Modeling of Exchangeable RPMs

5.4.1. Weighted Mixtures of DPs

We consider now the problem of modeling exchangeable collections of RPMs, i.e., $\mathcal{G} = \{G_j : j = 1, \ldots, n\}$, where the prior $p(\mathcal{G})$ is invariant with respect to the order in which the G_js are included in the model. The data are $(y_{i,j})$, where $j = 1, \ldots J$ denotes the study under which the observations were generated and $i = 1, \ldots, I_j$ indexes observations within study j.

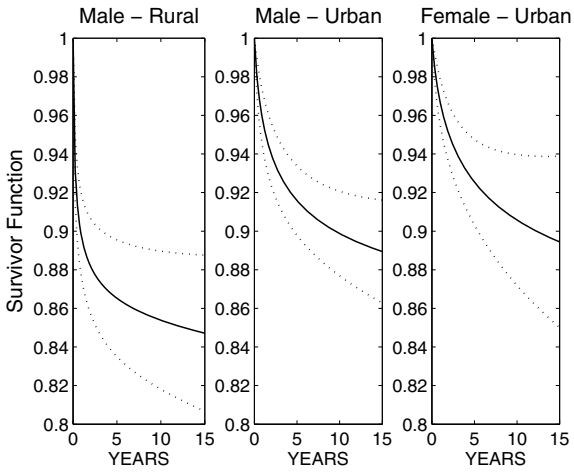


FIG 5.4. *Posterior estimated survival functions (solid lines) with 95% credible interval (dotted lines). All three panels represent children belonging to the third birth cohort. The labels indicate the levels of the other two covariates.*

Müller *et al.* (2004) and more recently Griffin *et al.* (2010) and Kolossiatis *et al.* (2012) use a construction based on the superposition of random measures. By sharing part of the probability mass across different studies the construction creates the desired dependence. Figure 5.5 illustrates the idea.

In Müller *et al.* (2004) each of the RPMs G_j is defined as a combination of a common F_0 and a study-specific F_j. Let

$$G_j \mid \epsilon, F_j, F_0 = \epsilon F_0 + (1-\epsilon)F_j, \qquad F_0 \sim (M, H), \qquad F_j \sim \mathsf{DP}(M, H), \tag{5.3}$$

with $y_{i,j} \sim \int p(y_{i,j} \mid \theta) G_j(\mathrm{d}\theta)$. The model is completed with a prior on ϵ,

$$p(\epsilon) = \pi_0 \delta_0 + \pi_1 \delta_1 + (1 - \pi_0 - \pi_1)\mathsf{Beta}(a, b),$$

where $\mathsf{Beta}(x; a, b)$ denotes a beta distributed random variable x with parameters (a, b). Note that this prior on ϵ includes point masses on 0 and 1, allowing for the two extreme cases of common and conditionally independent G_j across studies.

Example 14 (Dependent RPMs) *Müller* et al. *(2004) use the hierarchical model* (5.3) *as a prior probability model for random effects distributions in two related studies. The data are log white blood cell counts over time for breast cancer patients in*

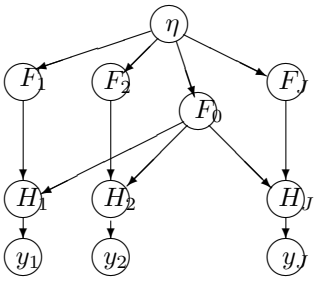


FIG 5.5. *Hierarchical composition of RPMs.*

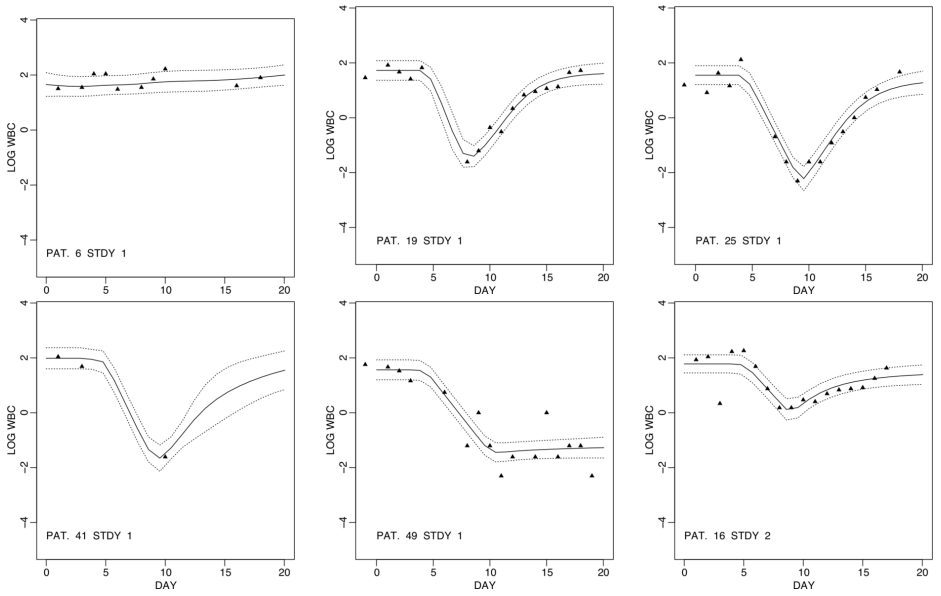


FIG 5.6. *Some typical patients. The data show y_{ijk} for 6 arbitrarily selected patients from studies $j = 1$ and $j = 2$. The triangles are the observed WBC. The solid line shows the posterior fitted mean curve, and the dotted lines show 95% central HPD intervals for the mean curve.*

two related studies. Figure 5.6 shows the data for some selected patients from the two studies.

The model includes a non-linear regression mean curve $f(t; \theta_{ji})$ for blood count data for patient i, in study j, $i = 1, \ldots, n_i$ and $j = 1, 2$. The mean curve is indexed with patient-specific random effects θ_{ij}. The random effects θ_{ij} are 9-dimensional and are assumed to arise from a study-specific random effects distribution G_j. The model is completed with the hierarchical prior in (5.3) for $\{G_1, G_2, G_3\}$, including a future third study $j = 3$. Figure 5.7 shows posterior inference for the unknown distributions F_0, F_1 and G_1 as bivariate scatterplots of random draws from the posterior means $E(F_0 \mid Data)$, $E(F_1 \mid Data)$ and $E(G_1 \mid Data)$. Figure 5.8 shows posterior predictive inference for a patient from a future third study $j = 3$.

Note that, in equation (5.3), the marginal prior for G_j is constructed as a sum of two RPMs that follow DP priors. Hence, the implied marginal prior $p(G_j)$ is not in general a DP itself. For many applications this might not be a concern. If desired, however, it is possible to construct the combination of the two RPMs F_0 and F_j such that the implied marginal prior $p(G_j)$ is again in the same family as $p(F_j)$; such a model is developed in Kolossiatis *et al.* (2012). In particular, assuming a DP prior for $p(F_j)$ it is possible to choose $p(\epsilon)$ such that $p(G_j)$ is a DP prior again.

The construction is easiest described by using a representation of the DP prior as a normalized gamma process. Let $\mu \sim \mathsf{GaP}(M\, G^\star)$ denote a gamma process, i.e., $\mu(B) \sim \mathsf{Gamma}(M\, G^\star(B), 1)$ for any measurable set B. Without loss of generality assume $J = 2$. Kolossiatis *et al.* (2012) use independent gamma processes $\mu_j \sim \mathsf{GaP}(M\, G^\star)$, $j = 0, 1, \ldots, 2$. Then

$$F_j(B) = \frac{\mu_j(B)}{\mu_j(X)}, \qquad j = 0, 1, 2$$

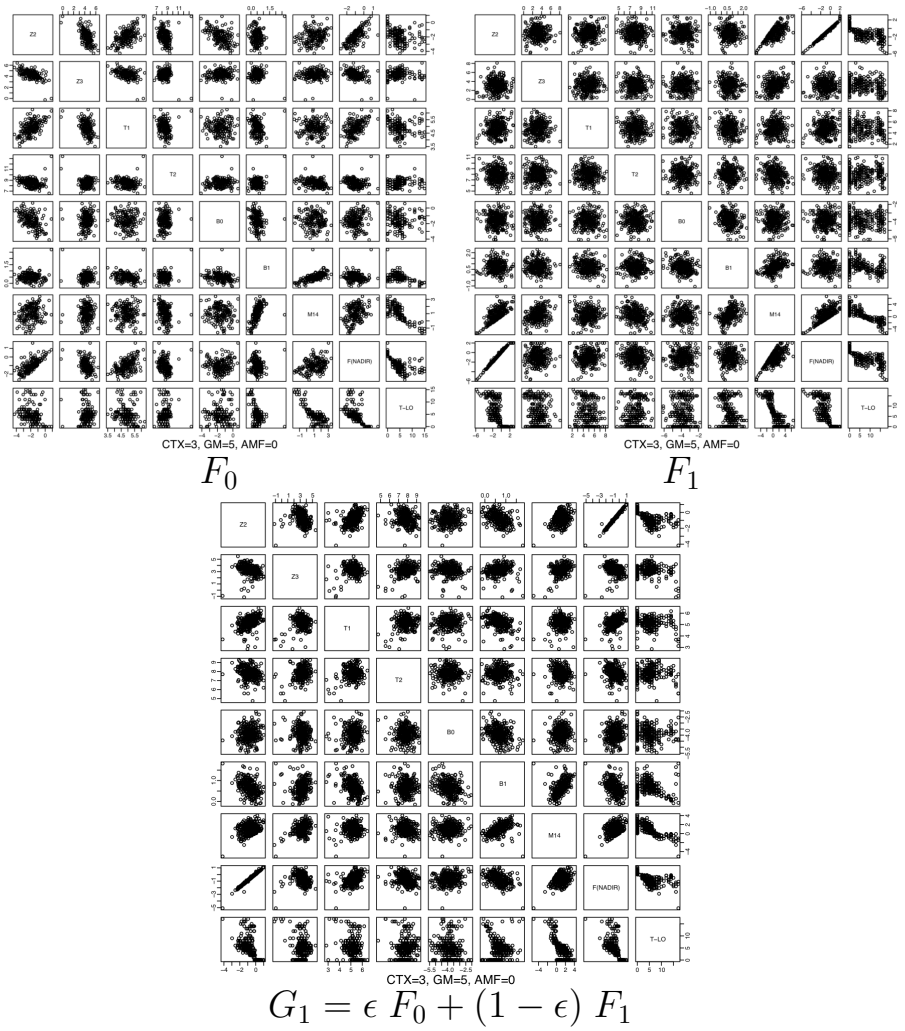
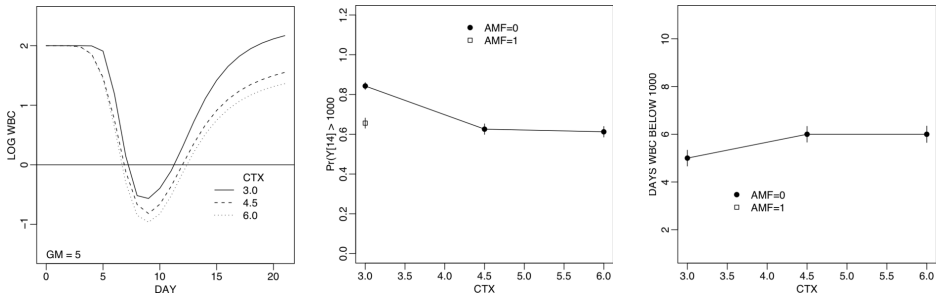


FIG 5.7. *Posterior estimated distributions* $E(F_0 \mid Data)$, $E(F_1 \mid Data)$ *and* $E(G_1 \mid Data)$.



(a) Estimated WBC profile

(b) $Pr\{WBC > 1000\}$ on day 14

(c) $E(T_{14} \mid y)$ T_{14} = days WBC below 1000.

FIG 5.8. *Posterior predictive inference for a hypothetical future patient in a future study* $j = 3$.

is a DP random probability. If we define the prior for ϵ as

$$\epsilon = \mu_0(X)/\left(\mu_0(X) + \mu_1(X)\right),$$

then

$$G_j(B) = \epsilon F_0(B) + (1-\epsilon)F_1(B) = \frac{\mu_0(B) + \mu_1(B)}{\mu_0(X) + \mu_1(X)}$$

follows marginally a DP prior again since the sum of the two gamma processes μ_1 and μ_2 is a gamma process again, $\mu_0 + \mu_1 \sim \mathsf{GaP}\left((M_0 + M_1)\, G^\star\right)$.

5.4.2. Hierarchical DP

Consider conditionally independent DP priors for each $G_j \in \mathcal{G} = \{G_j;\ j = 1, \ldots, J\}$, i.e., $G_j \sim \mathsf{DP}(M, G_0)$, independently. Recall the discussion in §5.1. The simplest way to borrow strength across G_js is through the base measure G_0. However, if G_0 is modeled parametrically, then the specific form of that parametric family determines and limits how information can be shared across the G_js. For example, if $G_0 = \mathsf{N}(\phi, 0)$ is indexed with a location parameter ϕ then borrowing strength across the G_j can only be through that location parameter. To avoid this limitation we could instead use a nonparametric prior for G_0, e.g., $G_0 \sim \mathsf{DP}(B, H)$. This is exactly the construction of the hierarchical Dirichlet process (HDP) of Teh *et al.* (2006). An early version of the same model appears in Escobar and Tomlinson (1999).

As a DP random measure, G_0 is discrete $G_0(\cdot) = \sum_{h=1}^{\infty} w_l \delta_{\widetilde{\theta}_h}(\cdot)$. Any θ drawn from G_0 is necessarily equal to one of the $\widetilde{\theta}_h$. In other words, the atoms of G_j agree with those of G_0. Hence, G_j can be written as

$$G_j(\cdot) = \sum_{h=1}^{\infty} \varpi_{j,h} \delta_{\widetilde{\theta}_h}(\cdot),$$

where $p\{(\varpi_{j1}, \varpi_{j2}, \ldots, \varpi_{jJ}) \mid (w_1, w_2, \ldots, w_J)\} = \mathsf{Dir}(w_1, w_2, \ldots, w_J)$ for any finite J. In other words, all the G_js use the same set of atoms but assign different (albeit related) weights to them (see Figure 5.9).

One implication of sharing the same atoms $\widetilde{\theta}_h$ across all G_js is that HDP mixture (HDPM) models allow co-clustering across different groups. Similar to (3.9) the HDPM can be written as a hierarchical model

$$y_{i,j} \mid \theta_{i,j} \sim p(y_{i,j} \mid \theta_{i,j}), \quad \theta_{i,j} \mid G_j \sim G_j, \quad G_j \mid G_0 \sim \mathsf{DP}(M, G_0), \quad G_0 \sim \mathsf{DP}(B, H),$$

where $j = 1, \ldots, J$ and $i = 1, \ldots, I_j$. Let $\{\theta_r^{\star\star};\ r = 1, \ldots, k\}$ denote the unique values among all θ_{ij}, $j = 1, \ldots, J$, $i = 1, \ldots, I_j$. Using ties to define clusters $S_r = \{(ij) :\ \theta_{ij} = \theta_r^{\star\star}\}$ we get a random partition model where $y_{i,j}$ and $y_{i',j'}$ can be assigned to the same cluster, even if they belong to different studies $j \neq j'$.

An appealing feature of the HDP is that it inherits the simple form of the predictive probability distribution from the DP prior. Conditional on G_0, the predictive probability distribution for $\theta_{ij} \sim G_j$ is unchanged from (3.3). Let k_i^j denote the number of distinct values $\{\theta_{hj}^\star, h = 1, \ldots, k_i^j\}$ among the draws $\{\theta_{1,j}, \ldots, \theta_{i-1,j}\}$ and $n_{i-1,h}^j = \sum_{\ell=1}^{i-1} I(\theta_{\ell j} = \theta_{hj}^\star)$. Then

$$\theta_{i,j} \mid \theta_{i-1,j}, \ldots, \theta_{1,j} \sim \sum_{h=1}^{k_i^j} \frac{n_{i-1,h}^j}{M + i - 1} \delta_{\theta_{h,j}^\star} + \frac{M}{M + i - 1} G_0, \qquad 1 \leq i \leq I_j,$$

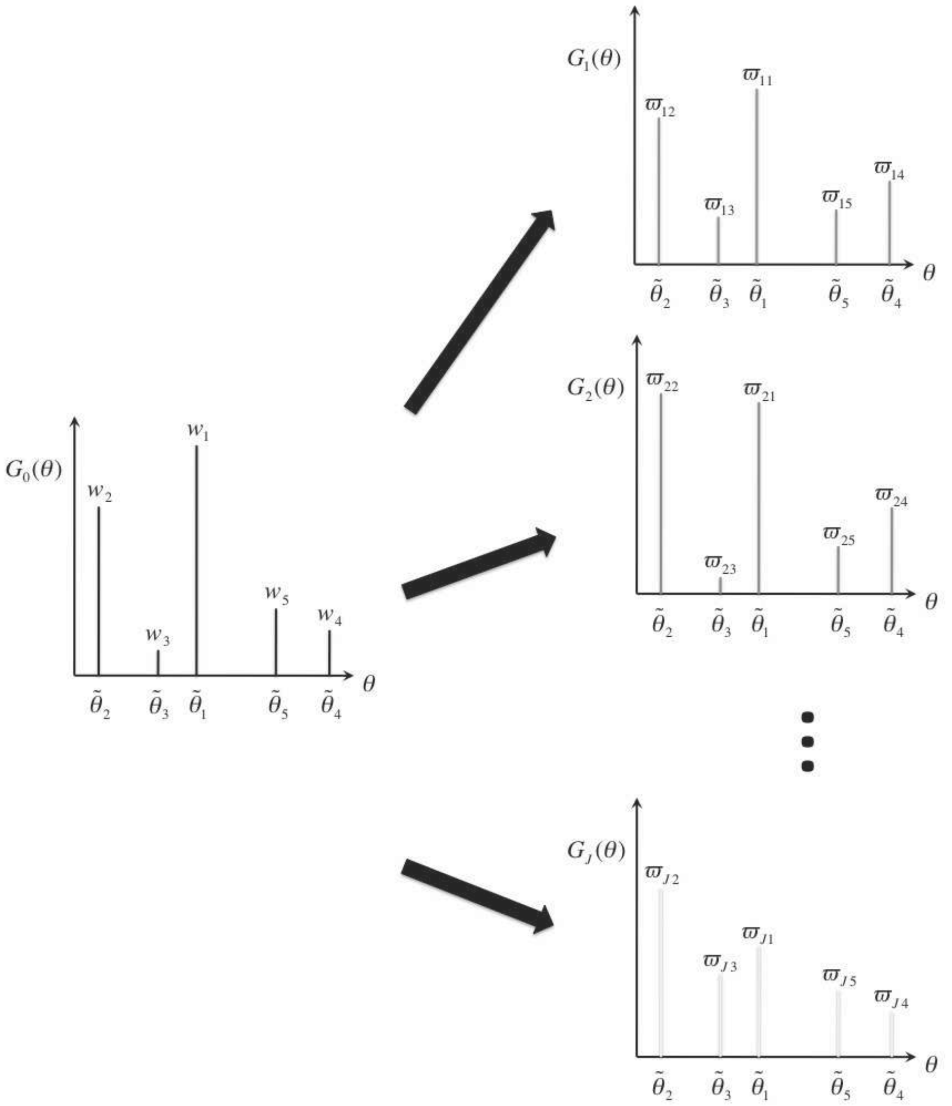


FIG 5.9. *Stylized representation of the hierarchical Dirichlet process. For each distribution, the location of the vertical lines on the horizontal axis corresponds to the value of the atoms* ($\tilde{\theta}_h$)*, while the height corresponds to the weight associated with it. The group-specific distributions* $G_1, \ldots, G_J$ *are conditionally independent draws from a Dirichlet process with baseline measure* G_0*, so the atoms are drawn from it. But, since the baseline measure* G_0 *is also drawn from a DP, it is discrete, and the atoms of the* G_j*s have to be identical to those originally drawn to construct* G_0*.*

where the unique values $\theta^{\star}_{hj}$ in turn are draws from G_0. We have a second instance of the Pólya urn (3.3) for a sequence of draws $\theta^{\star}_{hj} \sim G_0$. A minor notational complication arises by the double index $_{hj}$. Index the $\theta^{\star}_{hj}$ in sequence by running the first index h faster than the second index j, i.e., first we list all unique values among $\{\theta_{i1},\ i = 1, \ldots, I_1\}$, then all additional unique values among $\{\theta_{i2},\ i = 1, \ldots, I_2\}$ that have not yet been recorded, etc. Recall that $\theta^{\star\star}_r$, $r = 1, \ldots, k^{\star}$, are the unique

values among the $\theta^\star_{hj}$. Let $m_{h,j,r}$ be the number of elements among $\{\theta^\star_{11}, \ldots, \theta^\star_{h-1,j}\}$ equal to $\theta^{\star\star}_r$, and let $R_{h,j} = h + \sum_{j'<j} k_{j'}$. We get the predictive probability function

$$\theta^\star_{h,j} \mid \theta^\star_{h-1,j}, \ldots, \theta^\star_{1,1} \sim \sum_{r=1}^{k^\star_{h-1,j}} \frac{m_{h-1,j,r}}{B + R_{h-1,j} - 1} \delta_{\theta^{\star\star}_r} + \frac{B}{B + R_{h-1,j} - 1} H,$$

where $k^\star_{h,j}$ is the number of unique values among $\theta^\star_{11}, \ldots, \theta^\star_{h-1,j}, \theta^\star_{hj}$. The two Pólya urns can be combined to define a collapsed Gibbs sampler similar to §3.3.1, for details, see Teh *et al.* (2006).

Example 15 (Modeling documents using bag of words models) *One of the most illuminating applications of the hierarchical Dirichlet process mixture model is in modeling a collection of documents (corpora) using bag-of-words models.*

Bag-of-words models ignore the order in which observations appear in the text. Individual words are treated as categorical data with a document-specific distribution. Let $y_{i,j} \in \{1, \ldots, D\}$ be a categorical variable such that $y_{i,j} = d$ indicates that the i-th word in the j-th document is the d-th word in a dictionary of size D. In the simplest bag-of-words model, words from document j are assumed i.i.d. with $p(y_{i,j} = d \mid \theta_j) = \theta_{j,d}$. Information is shared across documents through a random-effects distribution on the θ_js, so that $\theta_j \mid G \sim G$. For example, for a nonparametric specification we could set $G \sim \mathsf{DP}(M, G_0)$ with G_0 being a (finite) Dirichlet distribution. This type of models can be considered "single topic" models because all words within a document come from the same probability distribution over words.

A natural extension of this idea is to treat each document as being composed of multiple topics, with topics being shared across documents. For example, a document on the effect of singing on child health might deal with the topics "music" (which places high probability on words such as "song," "melody" and "piano") and "medicine" (which emphasizes words such as "health," "symptom" and "treatment"), while another document about the entertainment options available in San Francisco this weekend might involve again the topic "music," along with the topic "outdoor activities" (focusing on words such as "hike," "ocean" and "sun"). In this extended model, topics corresponds to a different probability distribution over the dictionary, and words are still independently drawn from one of the multiple topics in the document. Such a model can be implemented using a HPDM with sampling model

$$p(y_{i,j} = d \mid \theta_{i,j}) = \theta_{i,j,d}$$

and HPDM prior

$$\theta_{i,j} \mid G_j \sim G_j, \quad G_j \mid G_0 \sim \mathsf{DP}(M, G_0), \quad G_0 \sim \mathsf{DP}(B, \mathsf{Dir}(\eta)).$$

Here $G_0 = \sum w_h \widetilde{\theta}_h$ is a distribution of multinomial probability vectors (over the dictionary) $\widetilde{\theta}_h$. Each $\widetilde{\theta}_h$ corresponds to a topic. Each $G_j = \sum \varpi_{jh} \widetilde{\theta}_h$ is a mixture of topic-specific multinomial probabilities and is the distribution over words for document j. The weights ϖ_{jh} are the relative weights of the topics. To observe a word in document j we first select a topic by drawing $\theta_{ij} \sim G_j$, and then an actual word with $p(y_{ij} = d \mid \theta_{ij}) = \theta_{ijd}$.

5.4.3. Nested DP

In the HPD, information is shared across the distributions $G_1, \ldots, G_J$ by sharing the atoms of the stick-breaking construction. This allows us to cluster draws

across groups, but tells us nothing about how the distributions themselves should be grouped together. There are no ties, $p(G_j = G_\ell) = 0$ for $j \neq \ell$. The nested Dirichlet process (NDP), first introduced in Rodríguez *et al.* (2008), is an alternative construction for the collection $G_1, \ldots, G_J$ which does allow for clustering of the groups as well as clustering of the observations themselves.

Like the HDP, the NDP is a hierarchical model involving two Dirichlet processes,

$$G_j \mid Q \sim Q, \qquad\qquad Q \sim \mathsf{DP}(M, \mathsf{DP}(B, G_0)).$$

Hence, in the NDP the baseline measure for the first Dirichlet process is given by the second Dirichlet process *rather than by a random distribution drawn from it.* The random probability measure Q is a distribution on distributions. Alternatively, we could write NDP in terms of a stick-breaking construction,

$$G_j \mid Q \sim Q, \qquad\qquad Q = \sum_{k=1}^{\infty} w_k \delta_{\widetilde{G}_k},$$

where $w_k = z_k \prod_{h<k}(1 - z_h)$, $z_k \sim \mathsf{Beta}(1, M)$, and $\widetilde{G}_k \sim \mathsf{DP}(B, G_0)$. Hence the first level of the hierarchy generates a distribution on RPMs with point masses corresponding to the random distributions $\widetilde{G}_1, \widetilde{G}_2, \ldots$. These random distributions are then specified nonparametrically through draws from a common Dirichlet process, so that $\widetilde{G}_k = \sum_{l=1}^{\infty} \varpi_{k,l} \delta_{\widetilde{\theta}_{k,l}}$, with $\varpi_{k,l} = v_{k,l} \prod_{h<l}(1 - v_{k,l})$, $v_{k,l} \sim \mathsf{Beta}(1, B)$, and $\widetilde{\theta}_{k,l} \sim G_0$ independently for every k and l.

Writing the NPD in terms of its stick-breaking construction highlights the nature of the NDP as two-level clustering. First, the model clusters similar distributions together, by sampling from $G_j \sim Q = \sum w_k \delta_{\widetilde{G}_k}$. Then the model clusters observations only across distributions that have already been clustered together.

An alternative characterization for the NDP is as a model for random partitions of a set of random distributions. To see this consider the partition $\rho = \{S_1, \ldots, S_K\}$ of $\{G_1, \ldots, G_J\}$, where $S_k = \{j : G_j = G_k^\star\}$, so that the $G_k^\star$s denote a set of distinct random random measures. Then, the definition for Q implies that

$$p(\rho) = \frac{M^K (M-1)! \prod_{k=1}^{K} (n_k - 1)!}{(M + J - 1) K!},$$

where n_k denotes the umber of distributions in the set S_k. This is (3.5), with an additional $K!$ in the denominator to reflect the lack of ordering of the clusters in ρ. From there we can write

$$\begin{aligned} p(G_1, \ldots, G_J) &= p(G_1^\star, \ldots, G_K^\star \mid \rho) p(\rho) \\ &= \left\{ \prod_{k=1}^{K} \mathsf{DP}(G_k^\star \mid B, G_0) \right\} \left\{ \frac{M^K (M-1)! \prod_{k=1}^{K} (n_k - 1)!}{(M + J - 1) K!} \right\}. \end{aligned} \tag{5.4}$$

Finally, since sampling from G_j induces clusters we get two-level clustering.

Computation for NDP mixtures can be easily carried out by replacing the DP with almost sure truncations as the ones discussed in §3.4 (see Rodríguez *et al.*, 2008 for details). Alternatively, we can derive MCMC algorithms that avoid truncating the process by extending the representation in (5.4). To do so, condition on $\rho = \{S_1, \ldots, S_K\}$ and define K sets of partitions $\sigma_1, \ldots, \sigma_K$ with $\sigma_k = \{R_{k,1}, \ldots, R_{k,L_k}\}$, such that the k-th set is associated with the observations drawn from the distribution assigned to S_k. In other words, $R_{k,l} = \{\theta_{i,j} : \theta_{i,j} = \theta_{k,l}^\star\}$ where $\theta_{k,1}^\star, \ldots, \theta_{k,L_k}^\star$

denotes a set of L_k unique draws from $G_k^\star$. Since each $G_k^\star$ is independently drawn from a DP, the implied prior on each σ_k is

$$p(\sigma_k) = \frac{B^{L_k}(B-1)!\prod_{l=1}^{L_k}(m_{k,l}-1)!}{(B+\bar{I}_k-1)L_k!},$$

where $\bar{I}_k$ is the number of observations generated from distributions in group R_k, and $L_k < \bar{I}_k$ is the number of clusters associated with them. Hence, the joint distribution on $G_1, \ldots, G_K$ can be written in terms of ρ, (σ_k) and $(\theta^\star_{k,l})$,

$$p(\rho, (\sigma_k), (\theta^\star_{k,l}) \mid \mathbf{y}) = \left\{\prod_{j=1}^{J}\prod_{i=1}^{I_j} p(y_{i,j} \mid \theta_{i,j})\right\}\left\{\prod_{k=1}^{K}\prod_{l=1}^{L_k} p(\theta^\star_{k,l})\right\}\left\{\prod_{k=1}^{K} p(\sigma_k)\right\} p(\rho).$$

MCMC algorithms can be generated by devising proposal distributions that modify ρ, (σ_k) and/or $(\theta^\star_{k,l})$. One such proposal distribution is discussed in Müller and Nieto-Barajas (2008).

Example 16 (Assesing quality of care in US hospitals) *The Department of Health and Human Services makes available to the public a series of (self-reported) quality of care measures for U.S. hospitals at http://www.hospitalcompare.hhs.gov/. To illustrate the characteristics of the NPM, we consider a model for one of these measures (the percentage of patients who received the appropriate initial antibiotic). The information on hospitals is nested within states, so a NDP mixture is a natural alternative to identify underperforming/overperforming states, as well as underperforming/overperforming hospitals within each group of states. The model clusters states j into sets of states with matching distribution of hospitals. All hospitals within each such set of states are then clustered into sets of hospitals with matching distribution of quality of care measures.*

More specifically, let $y_{i,j}$ be the (suitable transformed) quality of care measurement in hospital $i = 1, \ldots, I_j$ of state $j = 1, \ldots, J$. Then the model becomes

$$y_{i,j} \mid \theta_{i,j} \sim \mathsf{N}(\mu_{i,j}, \tau^2_{i,j}), \quad (\mu_{i,j}, \tau^2_{i,j}) \mid G_j \sim G_j, \quad G_j \mid Q \sim Q, \quad Q = \sum_{k=1}^{\infty} w_k \delta_{\widetilde{G}_k},$$

where $\widetilde{G}_k \sim \mathsf{DP}\left\{B, \mathsf{N}(\mu \mid \mu_0, \sigma/^2\kappa_0)\mathsf{IGamma}(\sigma^2 \mid \nu_0, \tau_0^2)\right\}$.

Figure 5.10 presents the resulting density estimates for four states representative of the clusters generated by the NDP. Note that the estimates demonstrate slightly different levels of skewness in addition to different means.

Example 17 (Document clustering in multi-topic models) *To help clarify the differences between the NDP and the HDP, consider an alternative extension of the bag-of-words model discussed in Example 15, where a nested DP is used to model G_j, the document-specific topic distribution,*

$$y_{i,j} \mid \theta_{i,j} \sim \mathsf{Multinom}(\theta_{i,j}), \quad \theta_{i,j,} \mid G_j \sim G_j, \quad G_j \mid Q \sim Q, \quad Q = \sum_{k=1}^{\infty} w_k \delta_{\widetilde{G}_k},$$

and $\widetilde{G}_k \sim \mathsf{DP}(B, \mathsf{Dir}(\eta))$. The structure on Q implies that documents with matching distributions G_j will be clustered together, something that does not happen under the HDP model. On the other hand, since the $\widetilde{G}_k$s are drawn independently, topics are shared only among documents assigned to a common cluster, but not across clusters of documents.

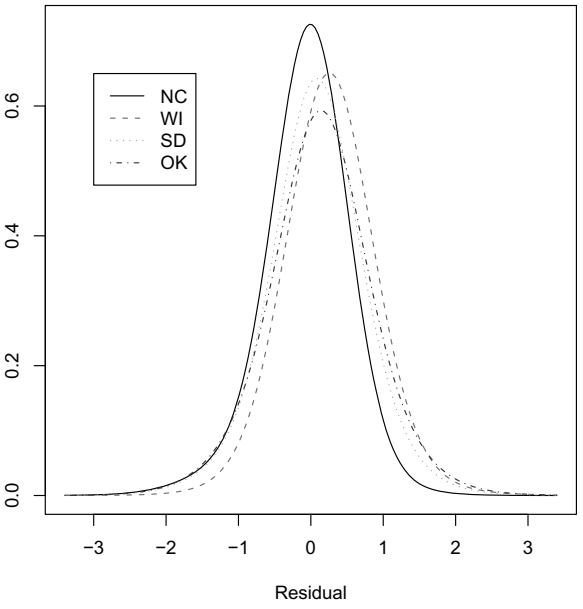


FIG 5.10. *Mean predictive density for four states under the NDP model: North Carolina (NC), Wisconsin (WI), South Dakota (SD) and Oklahoma (OK).*

5.5. DP Models for Time Course Data

5.5.1. Dynamic DP

The "single-p" DDP model can be used to construct collections of random distributions that evolve in discrete time. Consider a setting where at each time $t = 1, \ldots, T$ we collect observations $y_{t,1}, \ldots, y_{t,n_t}$ from a model that is written as a convolution of a normal linear model with a mixing measure G_t on the linear regression parameters:

$$y_{t,i} \mid \theta_{t,i} \sim \mathsf{N}(y_{t,i} \mid x_{t,i}\theta_{t,i}, \sigma^2), \qquad \theta_{t,i} \mid G_t \sim G_t,$$

where $x_{t,i}$ is a row vector. To create a flexible model for the mixing distribution we let

$$G_t(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\widetilde{\theta}_{t,h}},$$

and define the atoms sequentially by setting

$$\widetilde{\theta}_{0,h} \sim \mathsf{N}(\mathbf{m}_0, \mathbf{C}_0), \qquad \widetilde{\theta}_{t,h} \mid \widetilde{\theta}_{t-1,h} \sim \mathsf{N}(\mathbf{B}_t\widetilde{\theta}_{t-1,h}, \mathbf{W}_t). \tag{5.5}$$

If the weights are defined by the stick breaking prior, as in (5.1), then the model becomes a DDP with point masses $\boldsymbol{m}_h$ replaced by the stochastic process $\widetilde{\theta}_h = (\widetilde{\theta}_{th} : t = 1, 2, \ldots)$ defined by the Markov model (5.5).

This type of dynamic DDP was introduced in Rodríguez and Ter Horst (2008). It can be interpreted as a type II multiprocess model, in the sense of West and Harrison (1997). Since the weights (w_h) are independent of t, the model can be rewritten as a regular DP mixture model where $\mathbf{\Theta}_i = (\theta'_{0,i}, \theta'_{1,i}, \ldots, \theta'_{T,i})$ and

$$y_{t,i} \mid \theta_{t,i} \sim \mathsf{N}(y_{t,i} \mid x_{t,i}\theta_{t,i}, \sigma^2), \qquad \mathbf{\Theta}_t \mid \widetilde{G} \sim \widetilde{G}, \qquad \widetilde{G} \sim \mathsf{DP}(M, \widetilde{G}_0),$$

where the baseline measure $\widetilde{G}_0$ is the multivariate normal distribution induced by (5.5). Using this representation in terms of a single DPM allows us to create a slight generalization where we also mix over the observational variance σ^2. In that case,

$$y_{t,i} \mid \theta_{t,i} \sim \mathsf{N}(y_{t,i} \mid x_{t,i}\theta_{t,i}, \sigma_i^2), \qquad (\mathbf{\Theta}_t, \sigma_i^2) \mid \widetilde{G} \sim \widetilde{G}, \qquad \widetilde{G} \sim \mathsf{DP}(M, \widetilde{G}_0),$$

where $\widetilde{G}_0$ is defined as

$$\widetilde{\theta}_0 \mid \sigma^2 \sim \mathsf{N}(\mathbf{m}_0, \sigma^2\mathbf{C}_0), \quad \widetilde{\theta}_t \mid \widetilde{\theta}_{t-1}, \sigma^2 \sim \mathsf{N}(\mathbf{B}_t\widetilde{\theta}_{t-1}, \sigma^2\mathbf{W}_t), \quad \sigma^2 \sim \mathsf{IGamma}(\nu_0, V_0).$$

Another generalization, in which σ^2 is constant across components but allowed to evolve with time, is presented in Rodríguez and ter Horst (2010).

As with other DDP models, the representation in terms of a simple DPM also allows us to employ all the computational tools described in §3.3 to make inferences on this model. An important consideration, no matter which algorithm is used, is that efficient sampling for the atoms can be accomplished using Forward-Backward algorithms (Carter and Kohn, 1994; Frühwirth-Schnatter, 1994).

An appealing feature of this type of dynamic DDP models is their flexibility. By appropriately choosing the structural parameters x_t, $\mathbf{B}_t$ and $\mathbf{W}_t$ a number of different evolution patterns can be accommodated, including trends, periodicities and dynamic regressions.

Example 18 (Autoregressive models for distributions) *Consider a mixture of order-p autoregressive processes where $x_{it} = (1, 0, 0, \ldots, 0)$ is a vector of length p and*

$$\mathbf{B}_t = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

is a $p \times p$ matrix. The model is completed by setting a prior on the vector of autoregressive coefficients $(\phi_1, \ldots, \phi_p)$. To simplify computation, this can be chosen as a multivariate Gaussian distribution.

Example 19 (Modeling the evolution of claim distributions) *Rodríguez and Ter Horst (2008) use the dynamic DDP to model the value of travel reimbursement claims in a major international development bank between January 2005 and May 2007. They use a simple random walk model where $y_{t,i} \sim \mathsf{N}(y_{t,i} \mid \theta_{t,i}, \sigma_i^2)$ and $\widetilde{\theta}_t \mid \widetilde{\theta}_{t-1}, \sigma^2 \sim \mathsf{N}(\widetilde{\theta}_{t-1}, \sigma^2 U)$. Figure 5.11 shows the smoothed and one-step-ahead density estimates generated by the model for five months (January to May, 2007).*

5.5.2. Time Series DDP

Nieto-Barajas *et al.* (2008) introduce another variation of DDP models that is suitable as a prior for a time series of random probability measures.Their construction

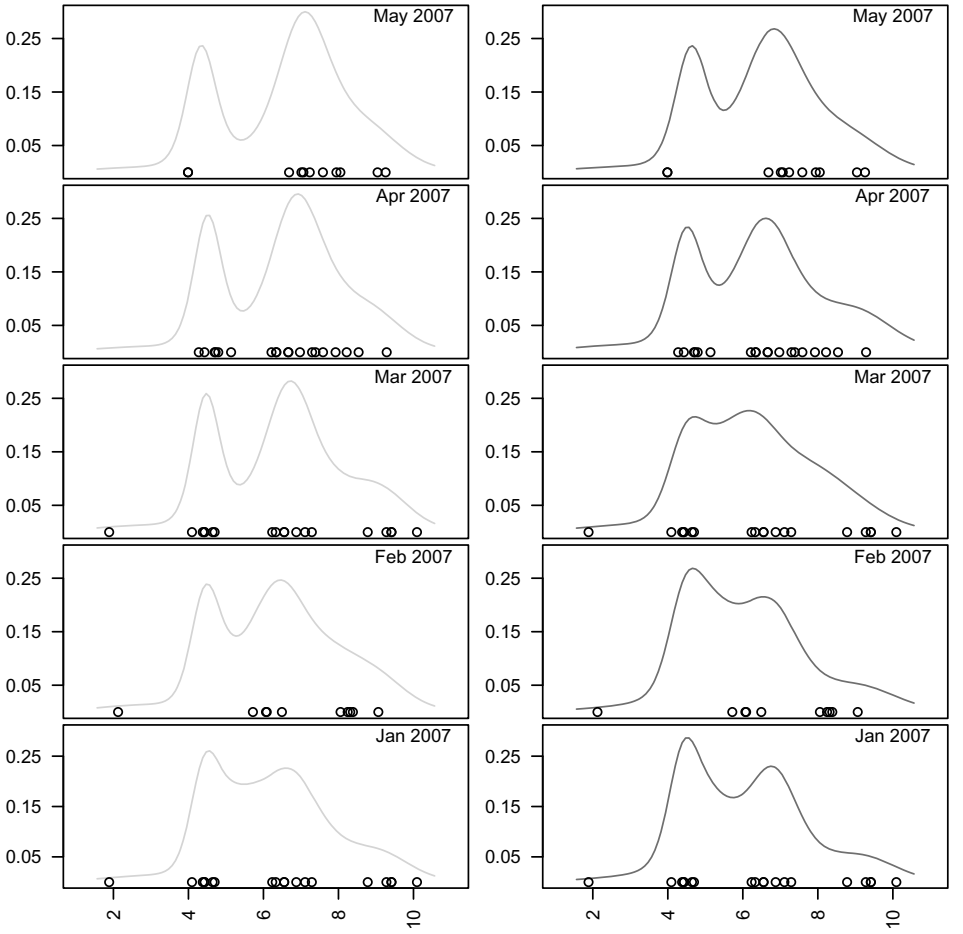


FIG 5.11. *Dynamic density estimates* $p(y_t \mid D_T)$ *and one-step ahead predictive distributions,* $p(y_t \mid D_{t-1})$ *for claims in 2007. Dots correspond to actual observations.*

is very straightforward. Recall the representation of a DP random measure by the stick breaking construction

$$G_t = \sum_{h=1}^{\infty} w_{th}\delta_{\widetilde{\theta}_h}, \tag{5.6}$$

where w_{th} are weights specific to G_t and $\widetilde{\theta}_h$ are point masses. Recall that the weights are defined by iterative stick breaking as $w_{th} = v_{th}\prod_{g<h}(1 - v_{tg})$ with beta distributed fractions v_{th}. The locations of the point masses are assumed to be common across all t. The use of the representation (5.6) already reveals that the proposed construction will be a variation of a common location DDP, i.e., all RPMs G_t have the same atoms $\widetilde{\theta}_h$ and only differ by the weights w_{ht}.

Nieto-Barajas *et al.* (2008) achieve the desired serial dependence by introducing a sequence of latent binomial variables $z_{th} \sim \mathsf{Bin}(k, v_{th})$ and replacing the prior for v_{th} in the stickbreaking construction of the DP prior by

$$v_{th} \mid z_{t-1,h} \sim \mathsf{Beta}\left(1 + z_{t-1,h}, M + \{k - z_{th}\}\right),$$

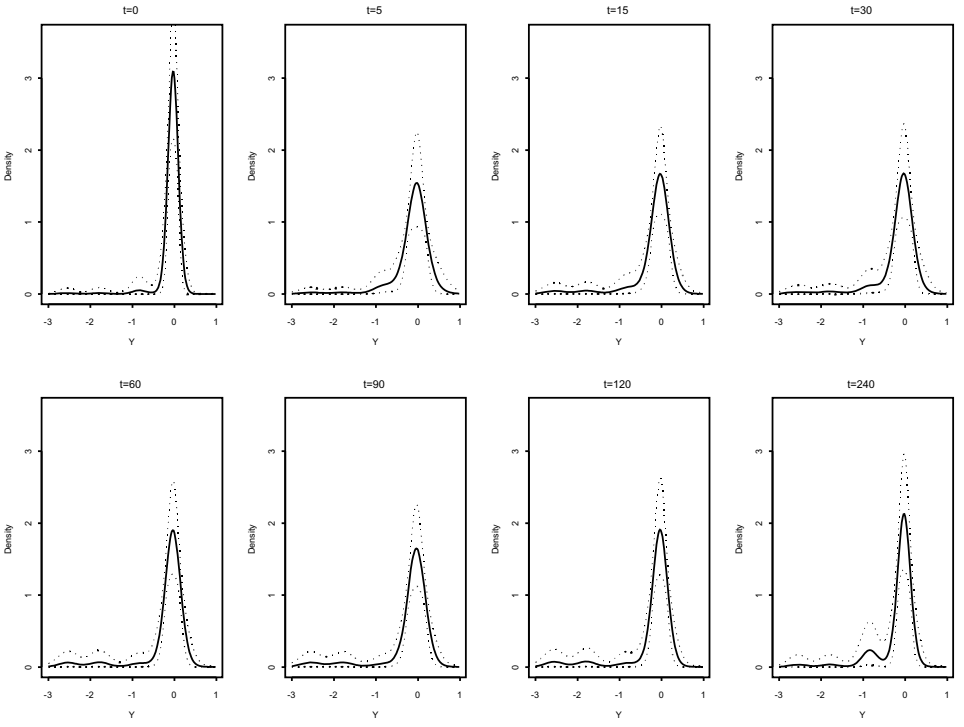


FIG 5.12. *Estimated distributions for* $t \in \{0, 5, 15, \ldots, 240\}$ *under a tsDDP model. The dotted curves show pointwise central 95% credible intervals.*

$t = 2, \ldots, T$. The marginal distribution of v_{th} remains unchanged $v_{th} \sim \mathsf{Beta}(1, M)$, and thus $G_t \sim \mathsf{DP}$, remains unchanged. The choice of k controls the level of dependence, with larger k implying higher dependence. Nieto-Barajas *et al.* (2008) use the model to analyze protein activation over time after an initial intervention. Figure 5.12 shows the estimated distributions G_t for an application of the tsDDP model to inference for protein activations over time after an intervention. Most of the proteins are not impacted by the intervention, only some are. This is reflected in a stable peak around 0 over time, and varying weight in the left tail, corresponding to inhibition of some proteins.

5.6. Spatial DDP

A version of the "single p" DDP model that is suitable for point-referenced spatial data is developed in Gelfand *et al.* (2005), and later extended in Duan *et al.* (2007). In the original definition of the spatial DDP, realizations from a Gaussian process are used as atoms in the stick breaking construction,

$$G_S = \sum_{h=1}^{\infty} w_h \delta_{\widetilde{\theta}_{h,S}}, \qquad \widetilde{\theta}_{h,S} = \left\{\widetilde{\theta}(s) : s \in S\right\} \sim \mathsf{GP}\left\{\mu(s), \gamma(s, s')\right\},$$

with $w_h = v_h \prod_{k<h}\{1-v_k\}$ and $v_h \sim \mathsf{Beta}(1, M)$. In this definition, $\mathsf{GP}\left\{\mu(s), \gamma(s, s')\right\}$ denotes a Gaussian process prior with mean function $\mu(s)$ and covariance function $\gamma(s, s')$. See §1.3.1.

The random distributions G_s can be used, for example, to model the distribution associated with a spatial random-effects. Assume that T independent realizations $y_1, \ldots, y_T$ with $y_t = (y_t(s_1), \ldots, y_t(s_m))'$ are available at locations $s_1, \ldots, s_m$. The model in Gelfand *et al.* (2005) implies that

$$y_t \mid \theta_t \sim \mathsf{N}(\theta_t, \sigma^2 I),$$

where $\theta_t = (\theta_t(s_1), \ldots, \theta_t(s_m))'$, $\theta_t \mid \widetilde{G} \sim \widetilde{G}$, and $\widetilde{G}(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\widetilde{\theta}_h}(\cdot)$. The GP prior implies

$$\widetilde{\theta}_h = \begin{pmatrix} \widetilde{\theta}_h(s_1) \\ \widetilde{\theta}_h(s_2) \\ \vdots \\ \widetilde{\theta}_h(s_m) \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \mu(s_1) \\ \mu(s_2) \\ \vdots \\ \mu(s_m) \end{pmatrix}, \begin{pmatrix} \gamma(s_1,s_1) & \gamma(s_1,s_2) & \cdots & \gamma(s_1,s_m) \\ \gamma(s_2,s_1) & \gamma(s_2,s_2) & \cdots & \gamma(s_2,s_m) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(s_m,s_1) & \gamma(s_m,s_2) & \cdots & \gamma(s_m,s_m) \end{pmatrix} \right).$$

In other words, for any finite sample, the spatial DDP reduces to a multivariate DP mixture with a multivariate Gaussian baseline measure whose mean and covariance matrix are structured according to $\mu(s)$ and $\gamma(s, s')$.

An appealing feature of the spatial DPM is that, although it can be centered around a stationary model a priori, it produces a non-stationary model a posteriori. Indeed, note that

$$\mathsf{E}\{y(s) \mid G_S\} = \sum_{h=1}^{\infty} w_h \widetilde{\theta}_h(s),$$

and

$$\begin{aligned} \mathsf{Cov}\{y(s), y(s') \mid G_S\} &= \left\{ \sum_{h=1}^{\infty} \sum_{k=1}^{\infty} w_h w_k \widetilde{\theta}_h(s) \widetilde{\theta}_k(s') \right\} \\ &\quad - \left\{ \sum_{h=1}^{\infty} w_h \widetilde{\theta}_h(s) \right\} \left\{ \sum_{h=1}^{\infty} w_h \widetilde{\theta}_h(s') \right\}, \end{aligned}$$

while, a priori,

$$\mathsf{E}\{y(s)\} = \mu(s), \qquad \mathsf{Cov}\{y(s), y(s')\} = \frac{1}{M+1} \gamma(s, s').$$

An interesting alternative to the spatial DDP is the hybrid DP of Petrone *et al.* (2009). They achieve a more parsimonious representation than the spatial DDP by considering a mixture of unique processes, but with local mixture weights. Each realization can pick up different unique elements at different locations. A related model is discussed in Rodríguez *et al.* (2010).

5.7. Other Dependent Extensions of the DP

The previous sections have focused mostly on "single p" DDPs. The popularity of this class of models is due to the fact that introducing dependence in the weights of the process, and performing posterior computation in the resulting constructions, is typically difficult. This section discusses dependent generalizations of DP mixtures that induce dependence in the weights of stick-breaking representation by replacing the beta-distributed random variables with more general random variables.

5.7.1. Probit Stick-Breaking Processes

Recall the stick-breaking construction of the DP,

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

where $\widetilde{\theta}_h \sim G_0$, $w_h = v_h \prod_{s<h}(1 - v_s)$ and $v_h \sim \mathsf{Beta}(1, M)$. Instead, consider stick-breaking ratios where $v_h = \Phi(\alpha_h)$ and $\alpha_h \sim \mathsf{N}(\mu, \sigma^2)$. Here $\Phi(x)$ is a standard normal c.d.f. In that case, we say that G follows a probit stick-breaking process (PSBP) with baseline measure G_0 and shape parameters μ and σ, denoted $G \sim \mathsf{PSBP}(\mu, \sigma, G_0)$. In words, the beta prior for the stick breaking model in the DP prior is replaced by a probit model. This simple change greatly simplifies extensions to dependent priors across families of probabilities measures, similar to the DDP.

The probit stick-breaking process has been discussed by Rodríguez *et al.* (2009), Chung and Dunson (2009) and Rodríguez and Dunson (2011), among others. The random distribution G is well defined (in the sense that $\sum_{h=1}^{\infty} w_h = 1$ almost surely), and very flexible. Indeed, from Proposition 3 and Corollary 1 in Ongaro and Cattaneo (2004), the support of the PSBP with respect to the topology of pointwise convergence is the set of absolutely continuous measures with respect to the baseline measure G_0.

The interpretation of the parameters of the PSBP is similar to those in the DP. Indeed, if $\mu = 0$ and $\sigma = 1$ then $v_h \sim \mathsf{Uni}[0, 1]$ and G follows a DP with $M = 1$; also, as $\mu \to \infty$ then $w_1 \to 1$ and G becomes a degenerate distribution at a random location $\widetilde{\theta}_1 \sim G_0$. More generally, for any measurable set B, $\mathsf{E}\{G(B)\} = G_0(B)$ and

$$\mathsf{Var}\{G(B)\} = \frac{\beta_2}{2\beta_1 - \beta_2} G_0(B)\{1 - G_0(B)\},$$

where $\beta_1 = \mathsf{Pr}(T_1 > 0) = \Phi(\mu/\sqrt{1+\sigma^2})$ and $\beta_2 = \mathsf{Pr}(T_1 > 0, T_2 > 0)$, where $(T_1, T_2)'$ follows a bivariate joint distribution with mean $\mathsf{E}(T_i) = \mu$, $\mathsf{Var}(T_i) = 1+\sigma^2$ and $\mathsf{Cov}(T_1, T_2) = \sigma^2$. Hence, G_0 represents the mean of the process, while μ and σ control the variability of G around G_0. Figure 5.13 presents various random samples from PSBPs that illustrate the effect of the parameters on the realizations.

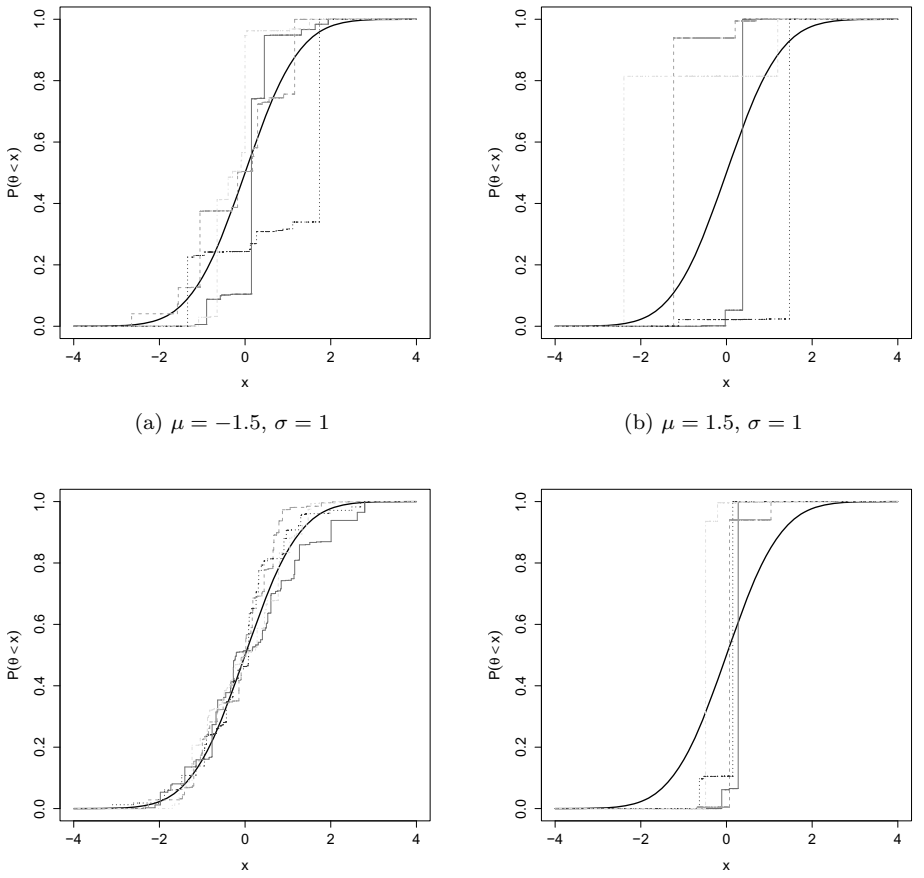
One appealing feature of the PSBP is the computational tractability of PSBP mixture models. In particular, consider a truncated version model

$$y_i \sim p(y_i \mid \theta_i), \qquad \theta_i \mid G \sim G^H,$$

where $G^H(\cdot) = \sum_{h=1}^{H} w_h \delta_{\widetilde{\theta}_h}(\cdot)$ and (w_h) and $(\widetilde{\theta}_h)$ are defined as before but setting $v_H = 1$. In that case, we can introduce random variables $(z_{i,1}), \ldots, (z_{i,H-1})$ such that $z_{i,j} \sim \mathsf{N}(\alpha_h, 1)$ and $\theta_i = \widetilde{\theta}_h$ if and only if $z_{ik} < 0$ for $k < h$ and $z_{ih} \geq 0$. Note that, if we let $s_i = h$ if and only if $\theta_i = \widetilde{\theta}_h$, by integrating out the z_{ih}s we get

$$\mathsf{Pr}(s_i = h) = \mathsf{Pr}(z_{i,1} < 0, \ldots, z_{i,h-1} < 0, z_{i,h} > 0) = \Phi(\alpha_{ih}) \prod_{k<h} \{1 - \Phi(\alpha_{ik})\} = w_h.$$

Conditionally on the auxiliary variables (z_{ih}), the full conditional distribution for α_h is a Gaussian distribution, while conditionally on α_h and the component indicators (s_i), the z_{ih}s are independent and follow (truncated) normal distributions. A similar data augmentation algorithm was originally proposed in the survival analysis literature to fit continuation ratio probit models Albert and Chib (2001).



(a) $\mu = -1.5$, $\sigma = 1$ (b) $\mu = 1.5$, $\sigma = 1$

(c) $\mu = -1.5$, $\sigma = 0.1$ (d) $\mu = 1.5$, $\sigma = 0.1$

FIG 5.13. *Realizations of probit stick-breaking process. The thick line on each Figure corresponds to the same baseline measure G_0 (in this case, a standard normal distribution). The plots demonstrate the effect of the parameters μ and σ (which control how close the realizations are to G_0).*

Alternatively, instead of explicitly truncating the mixing distribution G, designing a slice sampling algorithm similar to the one described §3.3.2 is also possible.

The PSBP can be easily generalized to create dependent probit stick-breaking processes (DPSBP) where dependence is introduced through the weights of the distributions. This is done by replacing the random draws (α_h) by independent realizations $(\alpha_h(x))$ from an appropriate Gaussian process. We illustrate these ideas with two examples.

Example 20 (An alternative to the hierarchical DP) *Consider a situation like the one we described in §5.4.2, where a partially exchangeable set of observations is collected. As before we model*

$$y_{i,j} \mid \theta_{i,j} \sim p(y_{i,j} \mid \theta_{i,j}), \qquad \theta_{i,j} \mid G_j \sim G_j,$$

for $i = 1, \ldots, I_j$ and $j = 1, \ldots, J$. Consider modeling the mixing distributions as

$$G_j(\cdot) = \sum_{h=1}^{\infty} \varpi_{j,h} \delta_{\widetilde{\theta}_h}(\cdot),$$

where $\widetilde{\theta}_h \sim G_0$, $\varpi_{j,h} = \Phi(\alpha_{j,h}) \prod_{k<h}\{1 - \Phi(\alpha_{j,k})\}$, $\alpha_{j,h} \sim \mathsf{N}(\mu_h, \sigma^2)$ and $\mu_h \sim \mathsf{N}(\mu_0, \tau^2)$. This model shares a number of features with the HDP. For example, the collection $G_1, \ldots, G_J$ is exchangeable because the atoms and weights are conditionally independent from each other. Also the distributions are centered around a common mean, in the sense that the transformed weights $(\alpha_{j,1}, \alpha_{j,2}, \ldots)$ are centered around a common value $(\mu_1, \mu_2, \ldots)$, i.e., $\mathsf{E}(\alpha_{j,h}) = \mu_h$ (remember Figure 5.9).

Sampling from this model is straightforward using auxiliary variables. As before, we introduce $z_{i,jh} \sim \mathsf{N}(\alpha_{j,h}, 1)$ and let $s_{i,j} = h$ if and only if $z_{i,j,s} < 0$ for $s < h$ and $z_{i,j,h} \geq 0$. Then, the full conditional distribution for $\alpha_{j,h}$ is normally distributed, i.e.,

$$\alpha_{j,h} \mid \cdots \sim \mathsf{N}\left(\left\{ \frac{1}{\frac{1}{\sigma^2} + n} \right\}^{-1} \left\{ \frac{\mu_h}{\sigma^2} + \sum_{i=1}^{I_j} z_{i,j,h} \right\}, \left\{ \frac{1}{\frac{1}{\sigma^2} + n} \right\}^{-1} \right).$$

On the other hand, the latent variables $z_{i,j,h}$ can be sampled from truncated normal distributions

$$z_{i,j,h} \mid \cdots \sim \begin{cases} \mathsf{N}(\alpha_{j,h}, 1) I(z_{i,j,h} < 0) & h < s_i \\ \mathsf{N}(\alpha_{j,h}, 1) I(z_{i,j,h} \geq 0) & h = s_i \\ \mathsf{N}(\alpha_{j,h}, 1) & h > s_i, \end{cases}$$

where $\mathsf{N}(a, b^2) I(A)$ represents the normal distribution with mean a and variance b^2 truncated to the set A.

Example 21 (Modeling an uncountable collection of distributions) *Consider an index space $\mathcal{X} \in \mathbb{R}^d$ and an uncountable collection of distributions $G_{\mathcal{X}} = \{G_x : x \in \mathcal{X}\}$. Define*

$$G_x(\cdot) = \sum_{h=1}^{\infty} w_h(x) \delta_{\theta_h}, \qquad w_h(x) = \Phi(\alpha_h(x)) \prod_{k<h} \{1 - \Phi(\alpha_k(x))\},$$

and $\alpha_h(x)$ is a Gaussian process over $\mathcal{X}$ with mean μ and covariance function $\sigma^2 \gamma(x, x')$. Given observations associated with locations $x_1, \ldots, x_n$, the joint distribution for the realizations of the latent processes $\alpha_h(x)$ at these locations is given by

$$\begin{pmatrix} \alpha_h(x_1) \\ \alpha_h(x_2) \\ \vdots \\ \alpha_h(x_n) \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \gamma(x_1, x_2) & \ldots & \gamma(x_1, x_n) \\ \gamma(x_2, x_1) & 1 & \ldots & \gamma(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(x_n, x_1) & \gamma(x_n, x_2) & \ldots & 1 \end{pmatrix} \right).$$

Models of this type can be used for time series observed in continuous time ($\mathcal{X} = \mathbb{R}^+$), or to construct models for spatial data ($\mathcal{X} \subset \mathbb{R}^2$). In particular, this construction allows us to easily generate spatial processes for discrete and non-Gaussian distributions. Even more, we can introduce multivariate atoms, leading to a simple

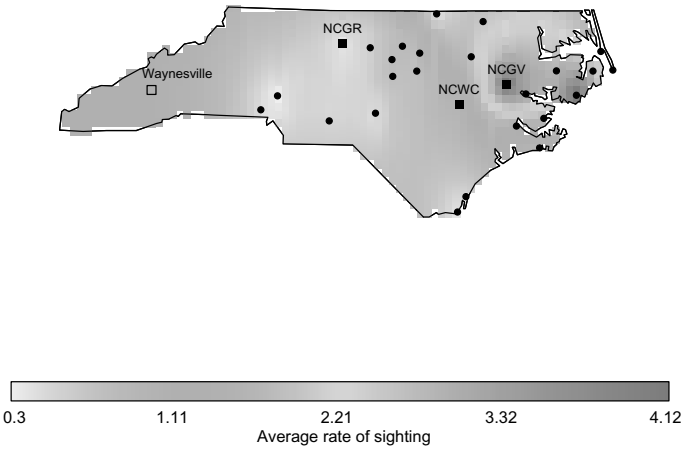


FIG 5.14. *Estimated expected rate of sightings (per man-hour) for the Mourning Dove. Filled dots correspond to the 27 locations where observations were collected. Squared dots represent locations where density estimation is carried out, filled squares represent locations for in-sample-predictions, while the empty square corresponds to a point of out-of-sample prediction.*

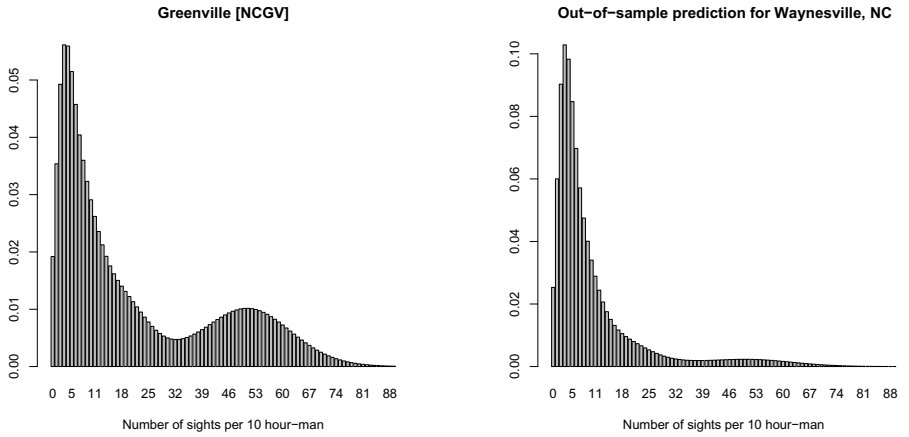


FIG 5.15. *Density estimates for two NC locations. The left panel corresponds an in-sample predictions at Greenville, NC (see also Figure 5.14), while the right panel corresponds to an out-of-sample prediction for a location in the Blue Ridge mountains next to Waynesville, NC.*

procedure to construct non-stationary, non-separable multivariate spatial-temporal processes. By interpreting $\mathcal{X}$ as a space of predictors, this construction also allows us to generate flexible nonparametric regression models with heteroscedastic errors.

Rodríguez and Dunson (2011) use this approach to generate a flexible spatial model for count data, which is used to model bird abundance in North Carolina. Figure 5.14 presents estimates of the expected rate of sightings (per man-hour) for the Mourning Dove, while Figure 5.15 presents predictive distributions for the number of sightings at two different locations.

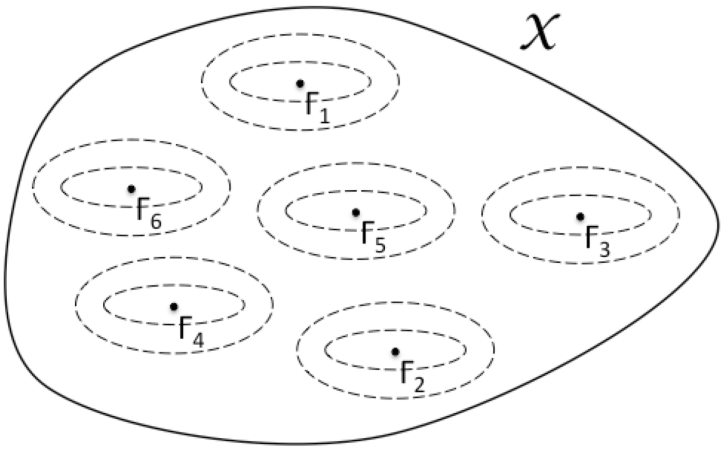


FIG 5.16. *Idealized representation for the construction of the kernel stick-breaking process.*

5.7.2. Kernel Stick-Breaking Processes

The kernel stick-breaking process (KSBP) (Dunson and Park, 2007) is another approach to create a prior over an uncountable collection of distributions $G_{\mathcal{X}} = \{G_x : x \in \mathcal{X} \in \mathbb{R}^d\}$.

In its simplest version, the KSBP is constructed by rebalancing its weights according to the distance between the value of the covariate x and a set of (random) fixed basis locations $\Gamma_1, \Gamma_2, \ldots$. More specifically, given a kernel $K(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to [0,1]$ and a probability measure Q defined on the index space $\mathcal{X}$, draw $\Gamma_h \sim Q$ for $h = 1, 2, \ldots$ and, for every $x \in \mathcal{X}$, create the countable collection $(K(x, \Gamma_h))$ (see Figure 5.16). The distribution G_x is then defined as

$$G_x(\cdot) = \sum_{h=1}^{\infty} w_h(x)\delta_{\widetilde{\theta}_h},$$

where $\widetilde{\theta}_h \sim G_0$, $w_h(x) = u_h(x)\prod_{k<h}\{1 - u_k(x)\}$, $u_h(x) = v_h K(x, \Gamma_h)$ and $v_h \sim \mathsf{Beta}(1, M)$.

Example 22 (KSBP with Gaussian kernels) *Consider a KSBP on $\mathcal{X} = \mathbb{R}^d$ where $K(x, x')$ is a Gaussian kernel, i.e.,*

$$K(x, x') = \exp\left\{-\lambda||x - x'||^2\right\},$$

in which case

$$G_x(\cdot) = \sum_{h=1}^{\infty}\left\{v_h K(x, \Gamma_h)\prod_{k<h}[1 - v_k K(x, \Gamma_k)]\right\}\delta_{\widetilde{\theta}_h}$$

and, for example, $\Gamma_h \sim \mathsf{N}(0, \tau^2)$. Note that, if $\lambda \to 0$, then $K(x, \Gamma_h) = 1$ for every pair (x, Γ_h), and the model reduces to a DP prior. Similarly, if $M \to 0$, G_x becomes a degenerate distribution at a random location $\widetilde{\theta}_1$ for every $x \in \mathcal{X}$. As before, a model of this type can be used for nonparametric regression, and well as for modeling non-stationary, non-separable temporal ($\mathcal{X} = \mathbb{R}^+$), spatial ($\mathcal{X} \subset \mathbb{R}^2$) or spatio-temporal ($\mathcal{X} \subset \mathbb{R}^3$) processes with non-Gaussian marginals.

A slightly more general version of this model can be obtained by replacing the point masses by random distributions drawn from a Dirichlet process and/or by replacing the prior on the v_hs with a more general beta distribution, so that $v_h \sim \mathsf{Beta}(a_h, b_h)$. In any case, the weights of the KSBP satisfy $\sum_{h=1}^{\infty} w_h(x) = 1$ for all $x \in \mathcal{X}$, and each member G_xs is therefore well defined.

Consider now a conditionally independent sequence where $\theta_i \mid \mathcal{G}, x_i \sim G_{x_i}$ and $\mathcal{G}_{\mathcal{X}} = \{G_x : x \in \mathcal{X}\}$ is assigned a KSBP prior. An interesting feature of the KSBP is that the joint distribution for $\theta_1, \ldots, \theta_n \mid x_1, \ldots, x_n$ obtained after integrating the random elements in $\mathcal{G}_{\mathcal{X}}$ can be obtained in closed form. As with the Dirichlet process, this joint distribution is obtained as a product of predictive distributions, each one corresponding to a generalized Pólya urn.

Posterior inference for the KSBP can be accomplish using through a Markov chain Monte Carlo algorithm that combines retrospective sampling and generalized Pólya urn sampling steps. Details can be seen in Dunson and Park (2007).