General Stick-Breaking Priors

STA 941. Surya Tokdar

1 Pitman-Yor Process

If $Z_1, \ldots, Z_n \overset{\text{IID}}{\sim} P$, $P \sim \text{DP}(a, \pi)$ with π non-atomic, then it can be shown that $K_n =$ the number of distinct elements among $Z_1, \ldots, Z_n$ satisfies:

$$\lim_{n\to\infty} \frac{1}{\log n} \mathbb{E}K_n = a, \quad \frac{1}{\log n}(K_n - a\log n) \to 0 \text{ a.s..}$$

To see the first result, let $D_1 = 1$ and $D_i = I(Z_i \not\in \{Z_1, \ldots, Z_{i-1}), i \ge 1$, so that $K_n = D_1 + \cdots + D_n$. We know that $\mathbb{E}D_i = a/(a + i - 1)$ and hence,

$$\mathbb{E}K_n = \sum_{i=1}^{n} \frac{a}{a+i-1} \asymp a \int_0^{n-1} \frac{1}{a+x} dx \asymp a \log n.$$

The second result follows from SLLN provided $\sum_{i=1}^{\infty} \mathbb{V}\text{ar}(D_i)/(\log i)^2 < \infty$, which holds since $\mathbb{V}\text{ar}(D_i) = a(i-1)/(a+i-1)^2$ and $\sum_{n>1} 1/(n \log^p n) < \infty$ for any $p > 1$.

Such logarithmic growth rates of K_n maybe undesirable in some applications. Furthermore, the sizes of these K_n clusters show an exponential tail behavior that might be undesirable as well. Arrange the clusters from largest to smallest with sizes $m_1 \ge m_2 \ge \cdots \ge m_{K_n}$. For any fixed $k \ge 1$, define $V_k = \lim_{n\to\infty} m_k/n$. Then $\mathbb{E}V_k \asymp \exp(-k/a)$. However, in many natural databases in language processing, image segmentation, etc., cluster size distributions exhibit a power-law tail decay, i.e., $V_k \sim k^{-\gamma}$ for some $\gamma > 0$. See Sudderth and Jordan (2009) and Goldwater et al. (2011) for some illuminating discussions.

A simple extension of DP that offers more flexible clustering rates and cluster size tail behaviors is the Pitman-Yor process (Ishwaran and James, 2001). The stick-breaking definition is almost identical to DP's, i.e.,

$$P(\cdot) = \sum_{l=1}^{\infty} w_l \delta_{\lambda_l}(\cdot)$$

with $\lambda_l \overset{\text{IID}}{\sim} \pi$ and $w_l = \beta_l \prod_{j<l}(1-\beta_j)$ but the break proportions β_l are not IID; instead $\beta_l \overset{\text{IND}}{\sim} Be(1-b, a+bl)$ for some $b \in [0,1)$ and $a > -b$. The extra parameter b is often called the discount parameter. Denote this process by $\text{PY}(a, b, \pi)$. Notice that $\text{PY}(a, 0, \pi) = \text{DP}(a, \pi)$.

Clearly, more general definitions are possible by letting $\beta_l \overset{\text{IND}}{\sim} G_l$ for well chosen sequences of distributions G_l on $(0,1)$. The stick-breaking process is a valid (random) probability measure on $\Lambda = \text{supp}(\pi)$ as long as $\sum_{l=1}^{\infty} \mathbb{E}\log(1-\beta_l) = -\infty$. However, the DP and the PY processes have several well known properties leading to computational

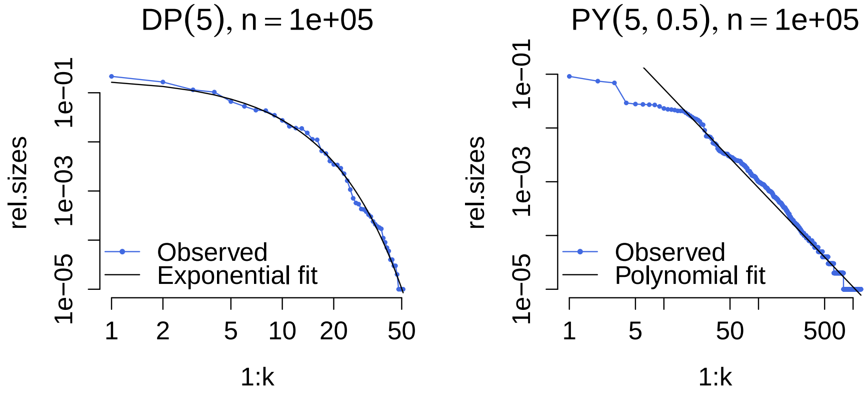

Figure 1: Number of clusters and cluster size distributions. Here $n = 100,000$. One sequence of $Z_1, Z_2, \ldots, Z_n$ is generated from each of $\mathrm{DP}(5, -)$ and $\mathrm{PY}(5, 0.5, -)$. Number of clusters can be inferred from the x-axis range. Both axes are plotted in logarithmic scale.

tractability that are hard to replicate with other choices. A key property is the so-called Polýa urn scheme representation of any $Z_1, \ldots, Z_n \overset{\text{IID}}{\sim} P$, $P \sim PY(a, b, \pi)$ given as follows (Pitman, 1995, Proposition 9): $Z_1 \sim \pi$ and for $i \geq 1$,

$$Z_{i+1} | (Z_1, \ldots, Z_i) \sim \frac{a + bK_i}{a + i} \pi + \sum_{c=1}^{K_i} \frac{\mathcal{N}_c^i - b}{a + i} \delta_{Z_c^*} \tag{1}$$

where K_i is the number of unique values amongst $\{Z_1, \ldots, Z_i\}$ with Z_c^*, $c = 1, \ldots, K_i$, denoting these unique values and $\mathcal{N}_c^i = \#\{1 \leq j \leq i : Z_j = Z_c^*\}$.

From (1), it follows that

$$\mathbb{E}K_{i+1} = \mathbb{E}K_i + \frac{a + b\mathbb{E}K_i}{a + i}$$

which implies,

$$\mathbb{E}K_n = \frac{a}{b} \left\{ \prod_{j=1}^{n} \frac{a + b + j - 1}{a + j - 1} - 1 \right\}$$

by induction on n. Stirling's approximation then implies

$$\mathbb{E}K_n \asymp \frac{\Gamma(a + 1)}{b\Gamma(a + b)} n^b.$$

See Pitman (2002, §3.3) for more details.

Therefore, the number of clusters under a PY process prior grows much more rapidly than the $\log n$ rate offered by a DP. Moreover, the cluster size distribution also shows a power law under PY. As before, define V_k as the limiting relative size of the k-th largest cluster. Pitman and Yor (1997, Proposition 17) show that

$$\mathbb{E}V_k \asymp D_{a,b} k^{-1/\alpha}$$

for some constant $D_{a,b}$ (given by a complicated but computable expression involving gamma functions). Figure 1 shows a comparison of both cluster size and relative cluster size distributions of a $\mathrm{DP}(5, -)$ and a $\mathrm{PY}(5, 0.5, -)$ [no need to specify π since partition structure does not depend on it].

Clearly, model fitting with a PY-mixture model where observations $Y_1, \ldots, Y_n$ are taken as $Y_j \overset{\text{IND}}{\sim} g(\cdot|\lambda_j)$ with $\lambda_j \overset{\text{IID}}{\sim} P$, $P \sim \mathrm{PY}(a, b, \pi)$, may proceed exactly as in the case of a DP-mixture model. In every iteration of the MCMC, one first runs one cycle of updates to regenerate the clustering pattern (either draw a new label or assign to one of the existing clusters) followed by another cycle of updates of the cluster specific parameters.

2 Priors on covariate dependent distributions

For mortality rate analysis, we previously focused on the longitudinal nature of the data and motivated a linear mixed effects model because a linear regression fit of crude rate on extreme weather variables, time and the interactions of time with hospital density, median income and population density, threw up residuals that for many counties were either all positive or all negative. Figure 2 shows a plot of 2012 prediction errors of mortality rates from the same model fit (recall training data stopped at 2011). Clearly there is a spatial pattern, indicating that a more accurate model should allow the random effects distribution to vary spatially.

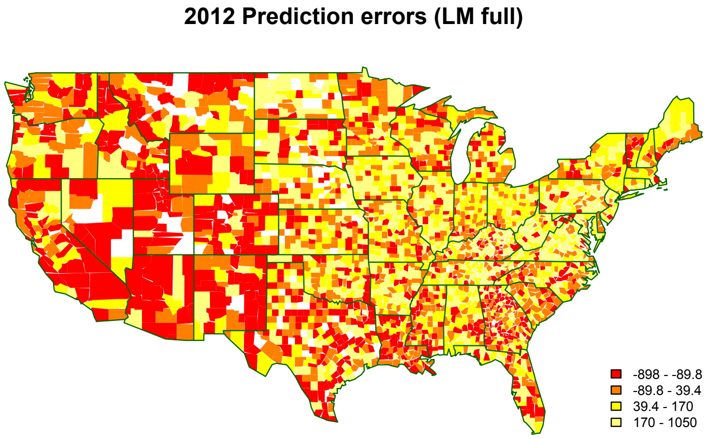

Figure 2: 2012 prediction errors from a linear model analysis of mortality rates.

This brings us to the general modeling context where one is interested in specifying a prior distribution on a collection of probability measures P_x on some space Λ indexed by a Euclidean variable $x \in \mathcal{X}$. In the mortality analysis, each P_x sits on $\Lambda = \mathbb{R}^q$, where q is the dimension of the random effect and $\mathcal{X}$ is a subset of $\mathbb{R}^2$, giving perhaps the central latitude-longitude information for each county. Below we discuss how DP and DP type prior specifications could be extended to allow covariate information. But before we go there, we need to take a look at probability measures on the space of functions (or curves) $\Lambda^{\mathcal{X}} = \{\lambda(\cdot) : \mathcal{X} \to \Lambda\}$.

2.1 Random elements in functions spaces, the Gaussian process

We are familiar with stochastic processes being defined as a collection of random variables indexed over, usually, a nice Euclidean subspace. For BNP modeling, it is more useful to view stochastic processes as elements in a well behaved function space, such as a Banach space of a Hilbert space.

For example, by a Gaussian process $\xi = (\xi(x) : x \in \mathcal{X})$ we usually mean a stochastic process for which there are functions $m : \mathcal{X} \to \mathbb{R}$ and $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, such that for any $k \in \mathbb{N}$ and any $\{x_1, \ldots, x_k\} \subset \mathcal{X}$, the random vector $(\xi(x_1), \ldots, \xi(x_k))$ has a k dimensional Gaussian distribution with mean $(m(x_1), \ldots, m(x_k))$ and covariance matrix with (i, j)-th element $C(x_i, x_j)$, $1 \le i, j \le k$. The covariance function C needs to be non-negative definite for the covariance matrix to be valid, and this imposes some restrictions on construction of Gaussian processes. Let $\mathrm{GP}(m, C)$ denote such a process.

Without loss of generality assume $m \equiv 0$, because if $\xi \sim \mathrm{GP}(0, C)$ then for any function $m : \mathcal{X} \to \mathbb{R}$, $m + \xi \sim \mathrm{GP}(m, C)$. When C satisfies

$$C(s, s) + C(t, t) - 2C(s, t) \le K\|t - s\|^\gamma, \forall s, t \in \mathcal{X}, \tag{2}$$

for some $K, \gamma > 0$, there exists a stochastic process $\xi \sim \mathrm{GP}(0, C)$ with continuous sample paths, i.e., with probability one the map $x \mapsto \xi(x)$ is continuous. In such cases, it is more useful to think of ξ as a random element of the Banach space $C(\mathcal{X})$ – the linear space of all real continuous functions on $\mathcal{X}$ equipped with the supremum norm [C, C^d, C_b etc. are accepted notation to denote spaces of continuous, d-times differentiable, continuous with bound b, etc. These are not to be confused with the C I have used for the covariance function.]

In fact, an alternative way to define a Gaussian process whose sample paths belong to a separable Banach space[1]. A random element ξ of a separable Banach space $(B, \|\cdot\|_B)$ is called *Gaussian* if the distribution of the scalar variable $b^*\xi$ is Gaussian for every $b^* \in B^*$, the dual space of B. We call ξ zero-mean if $b^*\xi$ has mean zero for every $b^* \in B^*$. We won't pursue this definition any further, but keep this in mind to always

[1] *Separable* means to have a countable dense subset. $C(\mathcal{X})$ is separable when $\mathcal{X}$ is a compact Euclidean subset, a result that follows from Weierstrass approximation theorem which asserts any continuous function is a limit of polynomials with rational coefficients. $C(\mathcal{X})$ is not separable when $\mathcal{X}$ is unbounded. This forces us to restrict to compact $\mathcal{X}$.

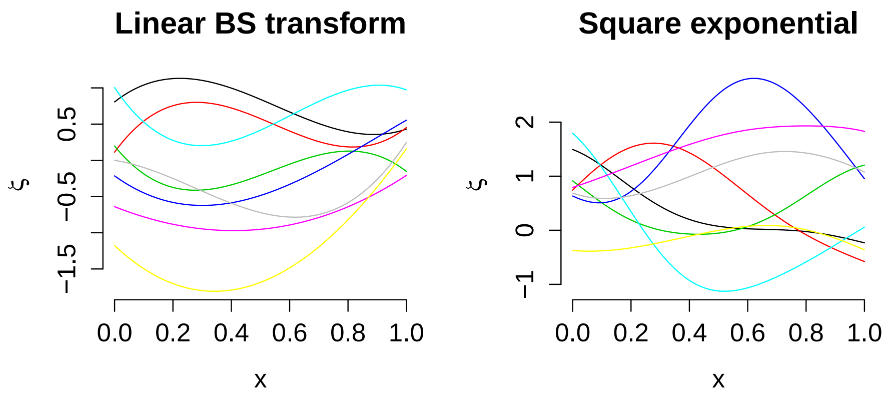

Figure 3: Draws of $\xi(x)$ from two different Gaussian processes.

think of a GP as a probability distribution on the space of functions/curves/surfaces and so on, i.e., view $\xi = (\xi : \xi(x), x \in \mathcal{X})$ as a whole instead of collection of random variables.

Many specifications of the covariance function C are known that ensure continuous and differentiable sample paths of ξ. Two specific examples to keep in mind are:

- Linear covariance: $C(s,t) = (1, h^T(s))S(1, h^T(t))^T$, where h is a given transformation. This arises for Gaussian processes defined as: $\xi(x) = (1, h^T(x))\gamma$ with $\gamma \sim N(0, \Sigma)$. It's useful to define $S = \rho^2 R$, with R being a correlation matrix.

- Square-exponential covariance: A more flexible specification of smooth Gaussian processes could be obtained with $C(s,t) = \rho^2 \exp\{-\psi^2 \|s-t\|^2\}$, which is strictly positive definite [will see a proof later].

Figure 3 shows draws of ξ from these two different processes, with $\rho = 1$, $R = \mathrm{I}$, $\psi = 2$, $\mathcal{X} = (0,1)$ and $h(x)$ denoting B-spline transforms with 3 degrees of freedom.

2.2 Dependent Dirichlet process priors

The stick-breaking representation of the Dirichlet process makes it conceptually easy to extend it to a process whose realizations are a collection of probability distributions $\{P_x : x \in \mathcal{X}\}$ all defined on a common space Λ. In particular, one may take

$$P_x(\cdot) = \sum_{l=1}^{\infty} w_l(x)\delta_{\lambda_l(x)}(\cdot) \tag{3}$$

where λ_l's are independent Λ-valued stochastic processes on $\mathcal{X}$ and $w_l(x) = \beta_l(x)\prod_{j<l}(1-\beta_j(x))$ with β_l's being independent $(0,1)$ valued stochastic processes on $\mathcal{X}$. If the λ_l processes are IID and also the β_l processes are both IID and satisfy $\beta_l(x) \sim Be(1,a)$ at

every $x \in \mathcal{X}$, then for every fixed x, the random probability measure $P_x \sim \mathrm{DP}(a, \pi_x)$, where π_x is the distribution of $\lambda_l(x)$. Such extensions are generally referred to as "dependent Dirichlet processes" (DDP), originating in MacEachern (1999, 2000).

2.2.1 Constant-weight DDP

Any practicable specification of (3) requires a fair amount of structure on the stochastic process valued atoms and/or the stochastic process valued weights. The first wave of DDP models mostly used a "constant weight" approach with $\beta_l(x) \equiv \beta_l$, and all variations across $\mathcal{X}$ were encoded through the atoms $\lambda_l(x)$, e.g., by drawing λ_ls from a Gaussian process distribution. Two prominent examples are De Iorio et al. (2004) and Gelfand et al. (2005) who use, respectively, a linear GP and an exponential GP (i.e., without the square on $\|s - t\|$ above).

The constant weight specification lends valuable computational tractability by preserving the Polýa Urn scheme property of a single DP. Essentially, the collection $\{P_x : x \in \mathcal{X}\}$ with P_x as in (3) with $\beta_l(\cdot) \equiv \beta_l$ and λ_l drawn IID from some probability distribution Π on $\Lambda^{\mathcal{X}}$ can be identified with a single DP

$$\tilde{P}(\cdot) = \sum_l w_l \delta_{\lambda_l}(\cdot)$$

on $\Lambda^{\mathcal{X}}$ with precision a and base measure Π. Any $Z_i \overset{\text{IND}}{\sim} P_{x_i}$, $i = 1, \ldots, n$ could be equivalently represented by the hierarchical representation:

$$\begin{aligned} Z_i &= \zeta_i(x_i), i = 1, \ldots, n \\ \zeta_1, \ldots, \zeta_n &\overset{\text{IID}}{\sim} \tilde{P} \\ \tilde{P} &\sim \mathrm{DP}(a, \Pi). \end{aligned}$$

Figure 4 shows a draw from a DDP-mixture model: $f(y|x) = \int N(y|\mu, \sigma^2) dP_x(\mu)$, $x \in [0, 1]$, where the underlying $\tilde{P} \sim \mathrm{DP}(a, \Pi)$ with Π denoting the zero-mean Gaussian process distribution with square-exponential covariance kernel with parameters $\rho = 1$ and $\psi = 2$. The weights and the curve-valued atoms are shown on the top row. The resulting P_xs, for $x \in \{0, 1/3, 2/3, 1\}$ are shown in the other panels, along with the normal mixture $f(y|x)$, with $\sigma = 0.5$.

The fact that a single DP random measure $\tilde{P}$ induces a constant-weight DDP $\{P_x : x \in \mathcal{X}\}$ means that an urn scheme for $(Z_1, \ldots, Z_n)$ could be devised by augmenting them with a vector of *cluster labels* $c_1, \ldots, c_n$, which must satisfy $c_i = c_j$ iff $\zeta_i = \zeta_j$, but otherwise arbitrary. These cluster labels are NOT transient, i.e., they cannot be generated on the fly as needed from the information contained in Z_is, but must be maintained and updated throughout within a bigger Markov chain sampler for the $Z_1, \ldots, Z_n$. This is because we may not be able to infer the ties in the functions ζ_is simply from the ties in Z_is. The conditional distribution of (c_i, Z_i) given (c_{-i}, Z_{-i})

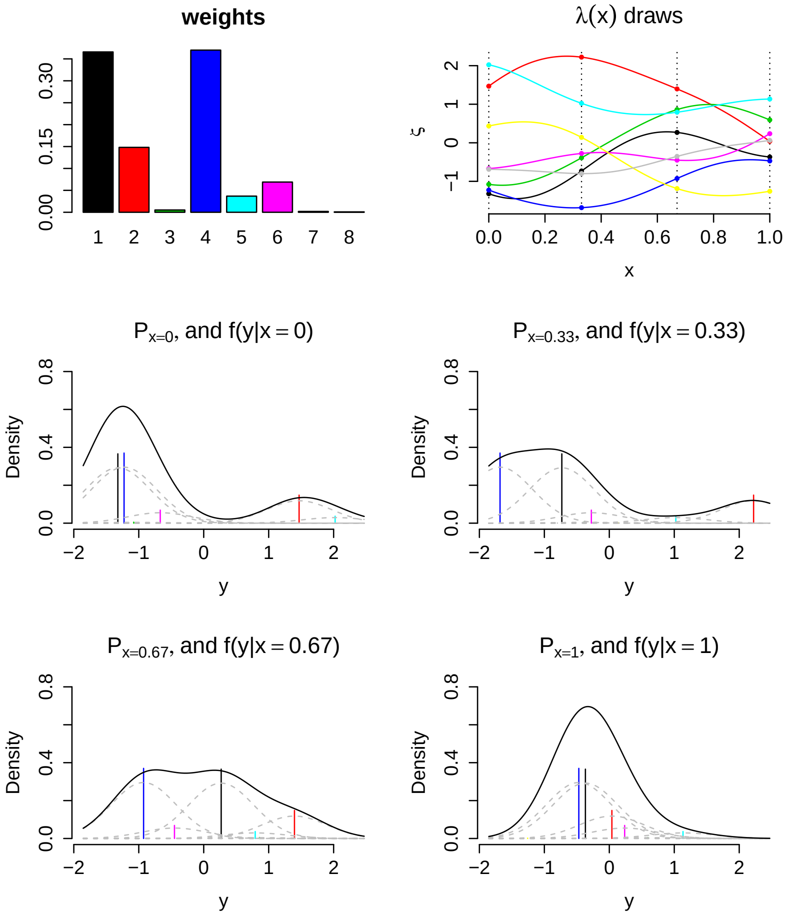

Figure 4: Constant-weight DDP.

could be described as

$$P(c_i \notin c_{-i}) = \frac{a}{a+n-1}; \quad P(c_i = c) = \frac{|\mathcal{N}_c^-|}{a+n-1}, c \in c_{-i}. \tag{4}$$

$$Z_i | (Z_{-i}, c_i \notin c_{-i}) \sim \pi_{x_i}(\cdot), \quad Z_i | (Z_{-i}, c_i = c) \sim \pi_{x_i}(\cdot | \lambda(x_j) = Z_j, j \in \mathcal{N}_c^-), c \in c_{-i}, \tag{5}$$

where $\mathcal{N}_c^- = \{j \neq i : c_j = c\}$ and π_x denotes the marginal distribution of $\lambda(x)$ under Π. This leads to possible Gibbs updates of latent parameters in a DDP-mixture model of the form: $y_j \overset{\text{IND}}{\sim} g(\cdot|\theta_j)$, $\theta_j \overset{\text{IND}}{\sim} P_{x_i}$, $(P_x : x \in \mathcal{X}) \sim \text{DDP}$.

2.2.2 Probit stick-breaking process

More recent approaches to DDP have tried to relax the constant weight specification (Dunson and Park, 2008; Chung and Dunson, 2009; Duan et al., 2007). This leads to a fair bit of complication in computing, essentially because urn schemes like (4)-(5) are generally not available in such extensions. Arguably, the most computationally tractable specification of weight shifting DDP comes in the form of a probit stick breaking process (PSBP; Rodriguez and Dunson, 2011). Here one takes

$$\beta_l(x) = \Phi(\xi_l(x))$$

where Φ denotes the standard normal CDF and ξ_ls are independent Gaussian processes. If the Gaussian processes are chosen so that $\mathbb{E}\xi_l \equiv 0$ and $\mathbb{V}\text{ar}\xi_l \equiv 1$, then at every x, the resulting random measure $P_x \sim \text{DP}(1, \pi_x)$. But for other choices, PSBP may not have a DP interpretation. Rodriguez and Dunson (2011) provide careful discussion of the effect of the Gaussian process mean and variance specification on the resulting clustering behaviors, including choices that resemble PY type discount factors. They also advocate, for computational ease, to use constant atoms $\lambda_l(x) \equiv \lambda_l \in \Lambda$. Figure 5 shows a draw from a constant-atom PSBP, with $\lambda_l \overset{\text{IID}}{\sim} N(0,1)$ and $\xi_l \sim \text{GP}(0, C)$ where C is the same square exponential kernel that we used for[2] Figure 4.

PSBP does not have a convenient urn scheme but can be fitted with a Gibbs sampler when the number of stick breaks is truncated at some upper bound N. Rodriguez and Dunson (2011) justify truncation by invoking a general approximation bound on truncation for covariate-free stick-breaking processes, originally due to Ishwaran and James (see e.g., Ishwaran and James, 2001, Theorem 2). Here is a cleaner and more extended take on it.

Let $w_l(\cdot)$, $l = 1, 2, \ldots$ be a sequence of real valued stochastic processes on $\mathcal{X}$ such that for each $x \in \mathcal{X}$, the vector $(w_1(x), w_2(x), \ldots) \in \Delta^\infty$, the infinite dimensional probability simplex. Also let $\lambda_1(\cdot), \lambda_2(\cdot), \ldots$ be a sequence of indepndent, Λ-valued stochastic processes on $\mathcal{X}$, which are also independent of $w_l(\cdot)$s. Let Π denote the joint probability distribution of these stochastic processes. Define the collection of probability measures $\mathcal{P} = \{P_x : x \in \mathcal{X}\}$ on Λ by $P_x(\cdot) = \sum_{l=1}^{\infty} w_l(x)\delta_{\lambda_l(x)}(\cdot)$. Also, for any $N \in \mathbb{N}$, define $\mathcal{P}^N = \{P_x^N : x \in \mathcal{X}\}$ as $P_x^N(\cdot) = \sum_{l=1}^{N} w_l^N(x)\delta_{\lambda_l(x)}(\cdot)$, where $w_l^N(x) = w_l(x)$, $l < N$ and $w_N^N(x) = 1 - \sum_{j<N} w_j^N(x)$. Let $f(y|x) = \int g(y|\lambda)dP_x(\lambda)$ and $f^N(y|x) = \int g(y|\lambda)dP_x^N(\lambda)$. For any $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$, define

$$m(y|x) = \int \prod_{i=1}^{n} f(y_i|x_i)d\Pi,\ m^N(y|x) = \int \prod_{i=1}^{n} f^N(y_i|x_i)d\Pi,\ y = (y_1, \ldots, y_n) \in \mathcal{Y}^n,$$

which are the marginal pdf of $(Y_1, \ldots, Y_n)$ under the full and the N-truncated mixture priors, given covariates $x_1, \ldots, x_n$.

Theorem 1. *$\|m^N(\cdot|x) - m(\cdot|x)\|_1 \le 2\sum_{i=1}^{n}\sum_{l\ge N}\mathbb{E}w_l(x_i)$.*

Proof. Clearly,

$$\begin{aligned}\|f^N(y_i|x_i) - f(y_i|x_i)\|_1 &= \|\{1 - \sum_{l\le N} w_l(x_i)\}g(y_i|\lambda_N(x_i)) + \sum_{l>N} w_l(x_i)g(y_i|\lambda_l(x_i))\|_1 \\ &\le 2\sum_{l\ge N} w_l(x_i),\end{aligned}$$

[2] In fact, the same draws of the GP were used as atoms in Figure 4 and for the weights in Figure 5.

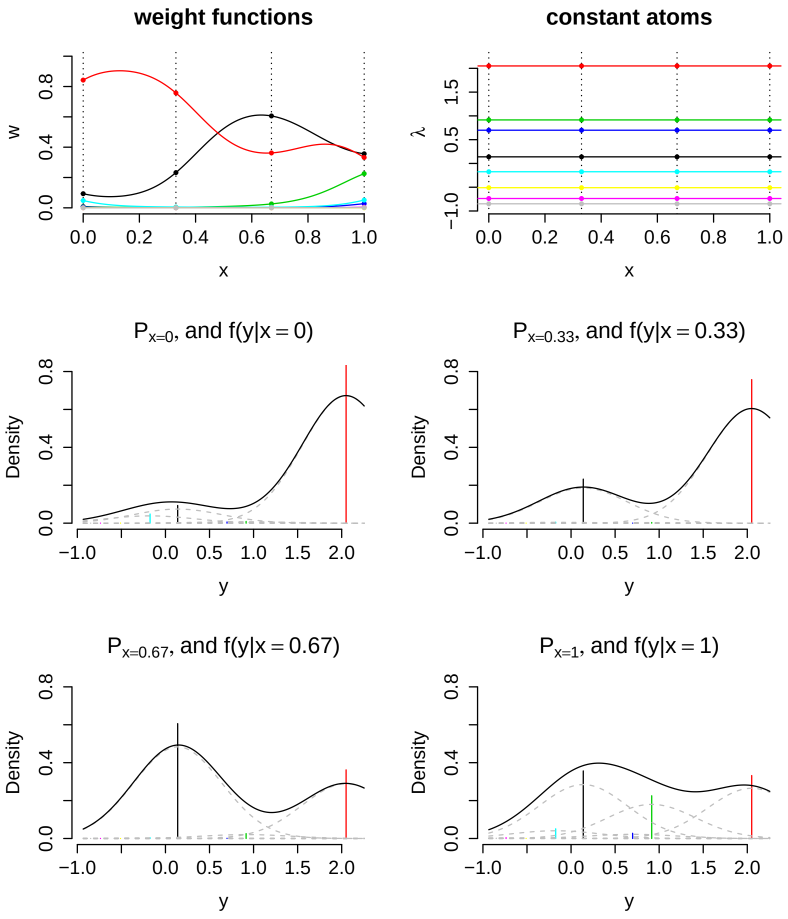

Figure 5: Constant-atom PSBP.

and hence

$$\begin{aligned}\|m^N(\cdot|x) - m(\cdot|x)\|_1 &\le \int \|\prod_i f^N(y_i|x_i) - \prod_i f(y_i|x_i)\|_1 d\Pi \\ &\le \int \sum_i \|f^N(y_i|x_i) - f(y_i|x_i)\|_1 d\Pi \\ &\le 2\sum_i \sum_{l\ge N} \mathbb{E} w_l(x_i)\end{aligned}$$

because $\| \prod_i p_i - \prod_i q_i\|_1 \le \sum_i \|p_i - q_i\|_1$. □

From Theorem 1, for n and x held fixed, the total variation distance between the marginal data distributions under the full prior and the truncated prior vanishes as $N \to \infty$. When $w_l(x)$ is generated by stick-breaking with pieces $\beta_l(x) \sim Be(1, a)$ [marginally, for each x], the upper bound is approximately equal to $n \exp\{-(N-1)/a\}$, and hence one needs $N \approx a\log(n)$ to meet a pre-specified truncation approximation error bound. Unsurprisingly, this is same as the order of $\mathbb{E}K_n$ for DP. Approximation error bounds on the total variation distance between m^N and m translate to approximation error bounds of the resulting posterior distributions, see Ishwaran and James (2002) for more discussion.

Under truncation to N components, the PSBP reduces to a finite mixture model, and MCMC computation can be carried by introducing two sets of latent variables. First, as in ordinary finite mixture model, introduce component labels c_i, $i = 1, \ldots, n$ and consider the equivalent representation of the model given by,

$$\begin{aligned}
Y_i &\overset{\text{IND}}{\sim} g(y_i|\lambda_{c_i}(x_i)), \quad i = 1, \ldots, n \\
c_i &\sim \mathit{Mult}(1, (w_1(x_i), \ldots, w_N(x_i)), \quad i = 1, \ldots, n \\
w_l(x) &= \Phi(\xi_l(x)) \prod_{j<l} \{1 - \Phi(\xi_j(x))\}, \quad l = 1, \ldots, N-1 \\
\xi_l &\overset{\text{IID}}{\sim} \text{GP}(m, C), \quad l = 1, \ldots, N-1 \\
\lambda_l &\overset{\text{IID}}{\sim} \Pi.
\end{aligned}$$

The critical step in an MCMC implementation of this model is the update of ξ_ls, for which one can use the standard probit parameter augmentation trick (Albert and Chib, 1993). Note that we only need to track ξ_ls at the observed covariate values. With a slight abuse of notation I will denote $\xi_l(x_i)$ by ξ_{li}. The conditional prior on c_is given ξ_{li}s can be induced by identifying

$$c_i = \min\{1 \le l \le N : Z_{il} > 0\}$$

where $Z_{il} \sim N(\xi_{li}, 1)$, $l = 1, \ldots, N-1$ and $Z_{iN} \equiv 1$. Therefore, a valid update of the ξ_{li}s given the c_is, other parameters and data, can be carried out as,

1. Generate (transiently) Z_{il}, $l = 1, \ldots, c_i$, $i = 1, \ldots, n$ as

$$Z_{il} \sim \begin{cases} N(\xi_{li}, 1)|_{(-\infty,0)} & \text{if } c_i > l \\ N(\xi_{li}, 1)|_{(0,\infty)} & \text{if } c_i = l \end{cases}$$

2. Update $\xi_l = (\xi_{l1}, \ldots, \xi_{ln})$, in parallel across $l = 1, \ldots, N-1$, according to the conjugate model $Z_{jl} \sim N(\xi_{lj}, 1)$, $j \in \{1 \le i \le n : c_i \ge l\}$, $\xi_l \sim N(m, C)$.

References

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association 88*, 669– 679.

Chung, Y. and D. B. Dunson (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association 104*, 1646–1660.

De Iorio, M., P. Müller, G. L. Rosner, and S. N. MacEachern (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association 99*(465), 205–215.

Duan, J. A., M. Guindani, and A. E. Gelfand (2007). Generalized spatial Dirichlet process models. *Biometrika 94*(4), 809–825.

Dunson, D. B. and J. H. Park (2008). Kernel stick-breaking processes. *Biometrika 95*, 307–323.

Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association 100*(471), 1021–1035.

Goldwater, S., T. L. Griffiths, and M. Johnson (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *The Journal of Machine Learning Research 12*, 2335–2382.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

Ishwaran, H. and L. F. James (2002). Approximate dirichlet process computing in finite normal mixtures. *Journal of Computational and Graphical Statistics 11*(3), 508–532.

MacEachern, S. M. (1999). Dependent nonparametric processes. In *Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA, pp. 50–55. American Statistical Association.

MacEachern, S. M. (2000). Dependent Dirichlet processes. Ohio State University Dept. of StatisticsTechnical Report.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields 102*(2), 145–158.

Pitman, J. (2002). Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002. Lecture notes for St. Flour course.

Pitman, J. and M. Yor (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability 25*, 855–900.

Rodriguez, A. and D. B. Dunson (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian Analysis 6*, 145–178.

Sudderth, E. B. and M. I. Jordan (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Advances in Neural Information Processing Systems*, pp. 1585–1592.