[1]

LATENT STICK-BREAKING PROCESSES

ABEL RODRÍGUEZ, DAVID B. DUNSON, AND ALAN E. GELFAND

ABSTRACT. We develop a model for stochastic processes with random marginal distributions. Our model relies on a stick-breaking construction for the marginal distribution of the process, and introduces dependence across locations by using a latent Gaussian copula model as the mechanism for selecting the atoms. The resulting latent stick-breaking process (LaSBP) induces a random partition of the index space, with points closer in space having a higher probability of being in the same cluster. We develop an efficient and straightforward MCMC algorithm for computation and discuss applications in financial econometrics and ecology.

1. INTRODUCTION

Recent literature in Bayesian nonparametrics has repeatedly focused on models for collections of distributions based on extensions of the Dirichlet process and other stick-breaking priors (Ferguson, 1973; Sethuraman, 1994; Ishwaran and James, 2001; Ongaro and Cattaneo, 2004). These extensions

[1]Abel Rodriguez is Assistant Professor, Department of Applied Mathematics and Statistics, University of California, Mailstop SOE2, Santa Cruz, CA 95064, abel@soe.ucsc.edu. David B. Dunson is Professor, Department of Statistical Sciences, Duke University, Box 90251, Durham, NC 27708, dunson@stat.duke.edu. Alan E. Gelfand is Professor, Department of Statistical Sciences, Duke University, Box 90251, Durham, NC 27708, alan@stat.duke.edu.
This work was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Key words and phrases. Nonparametric Bayes; Random probability measure; Random stochastic processes; Option pricing; Point-referenced counts .

The authors would like to thank the Associate Editor and two anonymous referees, whose comments greatly contributed to improve the quality of the manuscript.

are generated by introducing dependence in the weights and/or atoms of the stick breaking construction (MacEachern, 2000; DeIorio et al., 2004; Gelfand et al., 2005; Teh et al., 2006; Griffin and Steel, 2006b; Dunson and Park, 2007; Duan et al., 2007; Rodriguez et al., 2007), or by considering mixtures of independent samples from one or more Dirichlet processes (Müller et al., 2004; Griffin and Steel, 2006a; Dunson et al., 2007). In contrast, this paper focuses on models for prior distributions on stochastic processes on an index space $D \subset \mathbb{R}^d$ with rich but constant marginal distributions. We no longer focus on building different distributions for each possible value of the index space, but instead construct a stochastic process where observations at different locations are dependent but have a common (albeit unknown) marginal distribution. We call the resulting construction, a latent stick-breaking process (LaSBP). As the name indicates, we rely on a stick-breaking construction to represent the unknown marginal distribution, while introducing an underlying (latent) Gaussian process on D to drive the selection of the atoms at each location $\mathbf{s} \in D$, as suggested in MacEachern (2007).

As a motivation, consider the problem of estimating species abundance from point counts at pre-specified locations. It is clear that abundance changes spatially with the specific characteristics of micro-habitats, and modeling these changes is key to understanding population dynamics. However, biologists are interested in not only the spatial distribution of a species but also its local and global abundance. For example, when studying abundance of European Starlings (a non-native bird species) in North Carolina, one may be interested in the marginal distribution of the number of individual birds observed at randomly selected locations. In this context, the underlying Gaussian process introduced above can be interpreted as a latent, location-specific habitat quality score, which changes spatially making higher rates more likely in certain regions. Similarly, we could consider the price of a financial asset to be a random variable whose volatility changes in time according to an underlying score reflecting market sentiments, with the common marginal distribution providing a summary of the

return levels that can be expected from the asset. Current nonparametric Bayes models for related problems focus on priors for dependent collections of conditional distributions, which do not allow for the inference of a common marginal distribution.

Due to the discrete nature of its marginal distribution, LaSBP models induce a random segmentation of the index space D. This segmentation is spatially coherent, in the sense that, the closer two locations are in space, the larger the probability of them being assigned to the same cluster. Typically, the segmentation has an interesting interpretation in specific applications. For example, the partitions can be identified with micro-habitats in species distribution applications. Nonetheless, smooth interpolations and predictions can be recovered through model averaging. The literature in signal and image segmentation is enormous, some interesting reviews include Pal and Pal (1993), Pham et al. (2000), and Suri et al. (2005). Mixture models are the most relevant for our discussion. In contrast to our method, many of them do not incorporate information on spatial location in the prior for allocating observations to mixture components. One interesting exception is the model developed in Figueiredo (2005) and Figueiredo et al. (2007) for image segmentation, which uses a latent conditionally autoregressive (CAR) structure to define spatially dependent weights, in a similar spirit to Rodriguez et al. (2007).

Since the selection process in the LaSBP inherits the characteristics of the underlying Gaussian process, our models can easily accommodate a rich variety of correlation structures, including nonstationary processes. This is especially interesting because (1) typical change-point models (like those used in finance) assume a stationary Markovian structure in order to simplify computation, and (2) nonstationary dependence structures are hard to specify in other nonparametric models like the order-dependent Dirichlet process (Griffin and Steel, 2006b). Also, the model does not require replicates at each location, giving it an advantage over Gelfand et al. (2005) for the purpose of our applications.

In addition, the use of order constraints in the definition of the LaSBP allows us to generate skewed mixing distributions. For scale mixtures (as in Section 7.1), this feature yields models that favor heavier tails for the marginal distribution of the observations than those obtained by dependent Dirichlet process (DDP) mixtures. For location mixtures (as in Section 7.2), the LaSBP favors distributions that have a heavy right or left tail. This is appealing in many applications, but can be avoided through suitable adjustment of the distributions of the stick-breaking weights if desired.

The LaSBP can also be formulated as a Gaussian copula process with a common and unknown marginal distribution, which is assigned a stick-breaking prior. The literature on copula modeling is extensive; Nelsen (1999), among others, provides an introduction. Within a quasi-Bayesian setting, Hoff (2007) has recently developed a multivariate nonparametric model for mixed discrete and continuous data that uses Gaussian copulas and marginal set likelihoods. Although similar in spirit to the LaSBP, this model avoids the specification of the marginal distributions by using a pseudo-likelihood, and is formulated as a procedure to obtain a finite dimensional distribution and not a process. Also, Kottas et al. (2005) formulate a model for ordinal data using a nonparametric specification for the copula. Again, the goals of this model are different from ours, as they focus on contingency tables and build a mixture model for the copula describing the correlation across entries in the table.

The LaSBP is related to the hybrid Dirichlet process (hDP) models in Petrone et al. (2007), who also propose the use of Gaussian copula processes to build dependence across locations. However, while we focus on problems where prediction and interpolation are the main goal, the hDP is designed for local surface clustering. Although the hDP is, in principle, a more flexible model (as it allows for more general atoms than the LSBP), it can produce anomalous behaviors if the prediction surface is expected to be smooth (see Section 3). The construction of the LaSBP ensures that the resulting surfaces will be smooth by introducing an order constraint on the atoms of the stick-breaking construction. For some applications (e.g, the stochastic volatility model in Section 7.1), the infinite

hidden Markov model (iHMM) (Beal et al., 2001; Fox et al., 2008; van Gael et al., 2008), which is closely connected to the hierarchical Dirichlet process (Teh et al., 2006), is a viable alternative to the LaSBP. However, the iHMM assumes by construction that the index space D for the stochastic process is discrete and unidimensional, which prevents its use in other interesting applications such as the one presented in Section 7.2. In addition, the iHMM relies on a Markovian assumption, which only allows a very restrictive local dependence structure in allocation to mixture components, while the LaSBP is much more flexible in using a latent Gaussian process to drive the allocation.

This article is organized as follows: Section 2 reviews the literature on stick-breaking processes and sets the stage for the LaSBP model. Sections 3 and 4 define the latent stick-breaking process and discuss some of its properties, with Section 5 describing an efficient MCMC algorithm for posterior computation. Section 6 discusses extensions of the model to multivariate problems. Section 7 presents two applications in financial econometrics and ecology. Finally, in section 8 we present a short discussion and some future directions.

2. STICK-BREAKING PRIORS

Consider the probability spaces $(\mathbf{\Theta}, \mathcal{B}, P)$ and $(\mathbf{P}, \mathcal{C}, Q)$ such that $P \in \mathbf{P}$. Typically, $\mathbf{\Theta} \subset \mathbb{R}^q$, $\mathcal{B}$ corresponds to the Borel σ-algebra of subsets of $\mathbb{R}^d$ and $\mathbf{P}$ is the space of probability measures over $(\mathbf{\Theta}, \mathcal{B})$. We say that P follows from a stick-breaking distribution on $\mathbf{P}$ if

$$P(\cdot) = \sum_{l=1}^{L} w_l \delta_{\theta_l^*}(\cdot)$$

with $\theta_l^* \overset{iid}{\sim} H$ for some baseline probability measure H on $(\mathbf{\Theta}, \mathcal{B})$ and $w_l = v_l \prod_{k=1}^{l-1}(1 - v_k)$ where $v_l \sim \mathsf{beta}(a_l, b_l)$ for $l < L$ and $v_l = 1$ for $l = L$. The number of atoms L can be finite (either known or unknown) or infinite. The class of stick-breaking priors includes some well-known processes. For

example, taking $L = \infty$, $a_l = 1 - a$ and $b_l = b + la$ for $0 \leq a < 1$ and $b > -a$ yields the two-parameter Poisson-Dirichlet Process (Pitman, 1996), with the choice $a = 0$ and $b = \alpha$ resulting in the Dirichlet Process with baseline measure H and precision parameter α (Sethuraman, 1994).

For $L < \infty$, setting $v_L = 1$ ensures that $\sum_{l=1}^{L} w_l = 1$ with probability one. If $L = \infty$, Ishwaran and James (2001) show that the weights add up to 1 (and therefore the process yields a well-defined probability distribution) almost surely if and only if $\sum_{l=1}^{\infty} \mathsf{E}(\log(1 - v_k)) = -\infty$. For any measurable set $B \in \mathcal{B}$, $P(B)$ is a random variable whose moments are given by

$$\mathsf{E}(P(B)) = H(B)$$

$$\mathsf{V}(P(B)) = H(B)(1 - H(B)) \sum_{l=1}^{L} \left[\frac{a_l(1 + a_l)(a_l + b_l)}{(a_l + b_l)^2(a_l + b_l + 1)} \prod_{k=1}^{l-1} \frac{b_k(1 + b_k)(a_k + b_k)}{(a_k + b_k)^2(a_k + b_k + 1)} \right]$$

Therefore, H describes the expected shape of the distributions, while the sequences $\{a_l\}_{l=1}^{L}$ and $\{b_l\}_{l=1}^{L}$ control the variance of the sampled distributions around H.

In order to increase the flexibility of the model, most applications use convolutions of the form $\int K(\cdot|\boldsymbol{\theta})P(d\boldsymbol{\theta})$, where P follows a stick-breaking prior and K is a kernel, to model unknown probability distributions. Under fairly general conditions on the kernel K, mixtures of this type yield models that are dense (in the L^1 sense) in the space of absolutely continuous distributions (Lo, 1984).

One way to extend the stick-breaking priors to *collections* of distributions on a space D is to replace the random variables involved in the definition of the random distribution with stochastic processes on D. This is the strategy behind the dependent Dirichlet process (MacEachern, 2000; DeIorio et al., 2004; Gelfand et al., 2005; Griffin and Steel, 2006b). Inherently, all of these approaches assume that, conditional on the stick-breaking parameters, the probability of assigning observations to atoms is

known. For example, for the SDP (Gelfand et al., 2005),

$$\mathsf{Pr}(\theta(\mathbf{s}) = \theta_i^*(\mathbf{s}), \theta(\mathbf{s}') = \theta_j^*(\mathbf{s}') | \{\theta_l^*(\mathbf{s})\}_{l=1}^L, \{w_l(\mathbf{s})\}_{l=1}^L) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

for all $\mathbf{s}, \mathbf{s}' \in D$ and all $i, j \in \{1, \ldots, L\}$. Therefore, conditional on the atoms and weights, the selection mechanism is constant for all locations $\mathbf{s} \in S$. In the sequel, we explore an alternative approach, which constructs a process where the selection probabilities are not conditionally independent across locations while the atoms and weights are.

3. LATENT STICK-BREAKING PROCESS

Our goal is to construct a stochastic process $\{\theta(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ on $(\mathbf{\Theta}, \mathcal{B})$ such that $\theta(\mathbf{s}) \sim P$ for all $\mathbf{s} \in D$ and some unknown P, but $\mathsf{Cov}(\theta(\mathbf{s}), \theta(\mathbf{s}')) \neq 0$. To do this, consider the general approach: define a stochastic process $\{U(\mathbf{s}) : \mathbf{s} \in D\}$ with uniform marginals and discrete distribution D, then define $\theta(\mathbf{s}) = P^-(U(\mathbf{s}))$, where P is a random probability distribution and P^- is its generalized inverse. Then clearly $\theta(\mathbf{s})$ is distributed according to P and $\theta(\mathbf{s})$ and $\theta(\mathbf{s}')$ are correlated. The question of model development becomes how to choose the process $\{U(\mathbf{s})\}_{\mathbf{s} \in D}$ and a prior for P. In the sequel, we do this through the triplet

$$\left\{\{z(\mathbf{s}) : \mathbf{s} \in D\}, \{v_l\}_{l=1}^L, \{\theta_l^*\}_{l=1}^L\right\} \tag{1}$$

The *latent* Gaussian process $z(\mathbf{s})$ is such that $z(\mathbf{s}) \sim \mathsf{N}(0, 1)$ for all $\mathbf{s} \in D$ and $\mathsf{Cor}(z(\mathbf{s}), z(\mathbf{s}')) = \gamma(\mathbf{s}, \mathbf{s}')$. The sequences of *stick breaking ratios* $\{v_l\}_{l=1}^L$ is such that $v_l \sim \mathsf{beta}(a_l, b_l)$ for $l < L$ and $v_L = 1$. The sequence of *atoms* $\{\theta_l^*\}_{l=1}^L$ is constructed by imposing an order constraint on a sample from a baseline measure H on $(\mathbf{\Theta}, \mathcal{B})$ using the following mechanism: sample $\theta_1^* \sim H$ and for $l > 1$ draw $\theta_l^* \sim H_l$, where H_l is defined as the restriction of H to the set $S_l = \{\theta : \theta > \theta_{l-1}^*\}$, i.e.,

$H_l(B) = H(B \cap S_l)/H(S_l)$ for any measurable set $B \in \mathcal{B}$. Now, for any finite set of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$, and any n let the joint distribution $G(\theta(\mathbf{s}_1), \ldots, \theta(\mathbf{s}_n))$ be such that

$$G\left(\theta(\mathbf{s}_1) = \theta^*_{l_1}, \ldots, \theta(\mathbf{s}_n) = \theta^*_{l_n} \middle| \{\theta^*_l\}_{l=1}^L, \{v_l\}_{l=1}^L\right) = \mathsf{Pr}\Big(z(\mathbf{s}_1) \in [\Phi^{-1}(\pi_{l_1-1}), \Phi^{-1}(\pi_{l_1})), \ldots, z(\mathbf{s}_n) \in [\Phi^{-1}(\pi_{l_n-1}), \Phi^{-1}(\pi_{l_n}))\Big) \tag{2}$$

where $\Phi(\cdot)$ denotes the cumulative standard normal distribution function and $\pi_l = 1 - \prod_{k \leq l}(1 - v_k)$ is the proportion of the unit stick assigned to the first l atoms, with $\pi_0 = 0$. We call this hierarchical prior a Latent Stick Breaking Process (LaSBP), denoted $\mathsf{LaSBP}_L(\{a_l\}_{l=1}^L, \{b_l\}_{l=1}^L, H, \gamma)$. Referring back to our example in ecology, we can think of $z(\mathbf{s})$ as the quality of the habitat at location $\mathbf{s}$, $\{\theta^*_l\}_{l=1}^L$ as the set of possible abundance rates that can potentially be observed, and $\{v_l\}_{l=1}^L$ as the marginal prior probability of observing the different rates.

Theorem 1. *For every $n > 1$, the $\mathsf{LaSBP}_L(\{a_l\}_{l=1}^L, \{b_l\}_{l=1}^L, H, \gamma)$ satisfies Kolmogorov's conditions,*

(1) *Symmetry:*

$$G(\theta(\mathbf{s}_1) \in A_1, \ldots, \theta(\mathbf{s}_n) \in A_n) = G(\theta(\mathbf{s}_{\nu_1}) \in A_{\nu_1}, \ldots, \theta(\mathbf{s}_{\nu_n}) \in A_{\nu_n})$$

for any permutation $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ of the set $\{1, \ldots, n\}$

(2) *Consistency :*

$$G(\theta(\mathbf{s}_1) \in A_1, \ldots, \theta(\mathbf{s}_n) \in A_n, \theta(\mathbf{s}_{n+1}) \in (-\infty, \infty)) = G(\theta(\mathbf{s}_1) \in A, \ldots, \theta(\mathbf{s}_n) \in A_n)$$

where $A_1, \ldots, A_n$ are measurable subsets of $\boldsymbol{\Theta}$.

The proof is in Appendix A. Predictions on the value $\theta(\mathbf{s}')$ for an unobserved location $\mathbf{s}'$ depend on the predictions for $z(\mathbf{s}')$, in particular

$$\begin{aligned}
&\mathsf{Pr}(\theta(\mathbf{s}') = \theta_l^* | \theta(\mathbf{s}_1), \ldots, \theta(\mathbf{s}_n), \{z(\mathbf{s}_i)\}_{i=1}^n, \{\theta_l^*\}_{l=1}^L, \{v_l\}_{l=1}^L) \\
&\qquad = \mathsf{Pr}\left(z(\mathbf{s}') \in [\Phi^{-1}(\pi_{l-1}), \Phi^{-1}(\pi_l)) | z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)\right) \\
&\qquad = \int_{\Phi^{-1}(\pi_{l-1})}^{\Phi^{-1}(\pi_l))} \phi\left(t | \boldsymbol{\Psi}'_{\mathbf{s}',\mathbf{s}} \boldsymbol{\Psi}_{\mathbf{s},\mathbf{s}}^{-1} \mathbf{z}; 1 - \boldsymbol{\Psi}'_{\mathbf{s}',\mathbf{s}} \boldsymbol{\Psi}_{\mathbf{s},\mathbf{s}}^{-1} \boldsymbol{\Psi}_{\mathbf{s}',\mathbf{s}}\right) dt
\end{aligned}$$

where $\phi(\cdot|a; b)$ denotes the density of the univariate Gaussian distribution with mean a and variance b, $\boldsymbol{\Psi}_{\mathbf{s}',\mathbf{s}}$ is a column vector with entries $[\boldsymbol{\Psi}_{\mathbf{s}',\mathbf{s}}]_i = \gamma(\mathbf{s}_i, \mathbf{s}')$, $\boldsymbol{\Psi}_{\mathbf{s},\mathbf{s}}$ is a $n \times n$ matrix with entries $[\boldsymbol{\Psi}_{\mathbf{s},\mathbf{s}}]_{ij} = \gamma(\mathbf{s}_i, \mathbf{s}_j)$ and $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n))'$. Note that the sequences $\{v_l\}_{l=1}^L$ and $\{\theta_l^*\}_{l=1}^L$ define a random distribution

$$P(\cdot) = \sum_{l=1}^{L} w_l \delta_{\theta_l^*}(\cdot)$$

where the probability weights are defined as $w_l = v_l \prod_{k<l}(1 - v_k)$ and satisfy $\sum_l w_l = 1$ almost surely. The random distribution P will be the unknown marginal distribution of the process, and the latent process $z(\mathbf{s})$ plays no role in its definition. Note that, taking $L = \infty$, $a_l = 1$, $b_l = b$ and removing the order constraint on $\{\theta_l^*\}_{l=1}^L$, the random P follows a Dirichlet process with baseline measure bH. In contrast, the hDP$_k$ uses a so-called symmetric Dirichlet process prior, i.e., the weights $\{w_l\}_{l=1}^L$ follow a Dirichlet distribution, $\mathsf{Dir}(\alpha/k, \ldots, \alpha/k)$.

In order to gain some intuition into this construction, we show in Figure 1 three random realizations of the $\theta(\mathbf{s})$ surface on $D = [0, 1]^2$ associated with three different marginal distributions P sharing the same underlying latent process $z(\mathbf{s})$, which was generated using an Exponential correlation function. For this simulation, we took $L = 100$ and $a_l = 1$, $b_l = \alpha$, for three different values of α. The latent process was sampled at 3600 locations over an evenly spaced 60×60 grid, and we used a standard Gaussian baseline measure H to generate the atoms of the process. The resulting surfaces are tiled, segmenting the space according to the level of the underlying Gaussian process. The parameter

α influences the roughness of the segmentation (essentially, the number of different levels that the random surface takes), the baseline measure H controls the actual level of the tiles, and the correlation function γ influences the relative size of the different tiles.

Figure 1 also provides insight into the order constraint we introduced in the definition of the process. The order constraint allows us to link the behavior of the atoms of the stick-breaking construction to the behavior of the latent process. In particular, it enforces the internal consistency condition, $z(\mathbf{s}) > z(\mathbf{s}') \Rightarrow \theta(\mathbf{s}) \geq \theta(\mathbf{s}')$. In the species abundance application, this means that higher habitat quality scores should imply a larger number of sightings. In a similar spirit, a more nervous market should be subject to larger price changes. Without the order constraint forcing adjacent atoms to have similar values, predictions from the model might yield unreasonably small or large values, while label switching can introduce multimodality in the predictive distribution at new locations.

As an illustration of this phenomenon, consider a very simple setup where $L = 3$ and the process is observed at three locations (labeled s_1, s_2 and s_3 in Figure 2). Let the stick-breaking weights be such that $\Phi^{-1}(\pi_1) = -1$ and $\Phi^{-1}(\pi_2) = 1$ and the underlying $z(\mathbf{s})$ take the values $z(\mathbf{s}_1) = -2$, $z(\mathbf{s}_2) = 2$ and $z(\mathbf{s}_3) = 0$. This implies that the component membership indicators are $\xi(\mathbf{s}_1) = 1$, $\xi(\mathbf{s}_2) = 3$ and $\xi(\mathbf{s}_3) = 2$. Also, let the atoms take the value $\theta_1^* = -5$, $\theta_2^* = 5$ and $\theta_3^* = 0$ (this is a valid combination of values if the order constraint is removed).

Consider now the value of the underlying process and the corresponding predictions at location $\mathbf{s}_4$. Since the Gaussian process is continuous, predictions for $z(\mathbf{s}_4)$ will be close to 0, making $\xi(\mathbf{s}_4) = 2$ (and a prediction close to 5) highly likely, even though $\mathbf{s}_3$ is far away from $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_4$. This is undesirable if we are interested in smooth interpolations, as in that case we would expect $\theta(\mathbf{s}_4) < 0$. This feature can be particularly acute with small samples, as there may be little information to distinguish between multiple latent surfaces consistent with different label permutations.

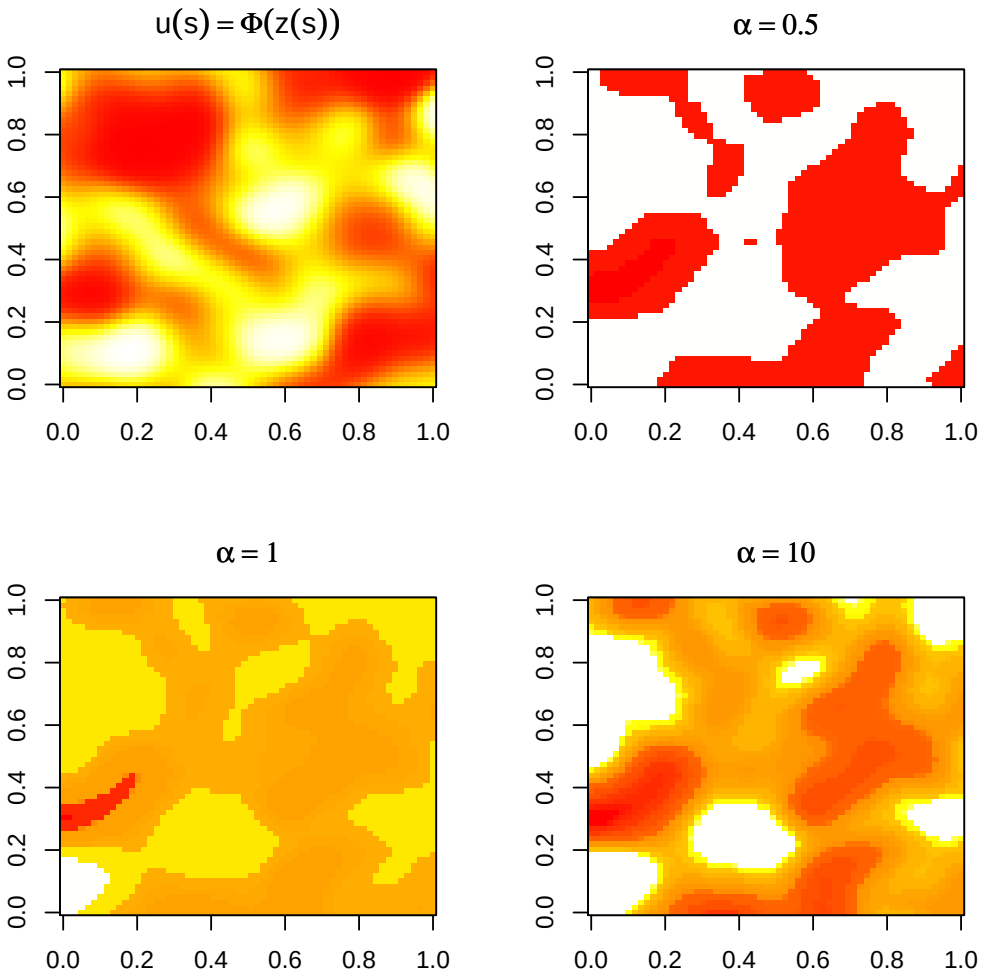


FIGURE 1. Realizations from a LaSBP process on $[0,1]^2$ with a standard Gaussian baseline measure. We illustrate the effect of different concentrations while keeping the underlying Gaussian process (shown in the upper left panel) constant to simplify interpretation. In the color scale, white corresponds to large values, yellow to medium values and red to low values of the process.

In most applications, the LaSBP will not be used to model the observable $Y(\mathbf{s})$ directly, but will be used to construct LaSBP mixtures. Briefly, we will assume that $Y(s)|\theta(s) \sim f(\cdot|\theta(\mathbf{s}))$, where $\theta(\mathbf{s})$

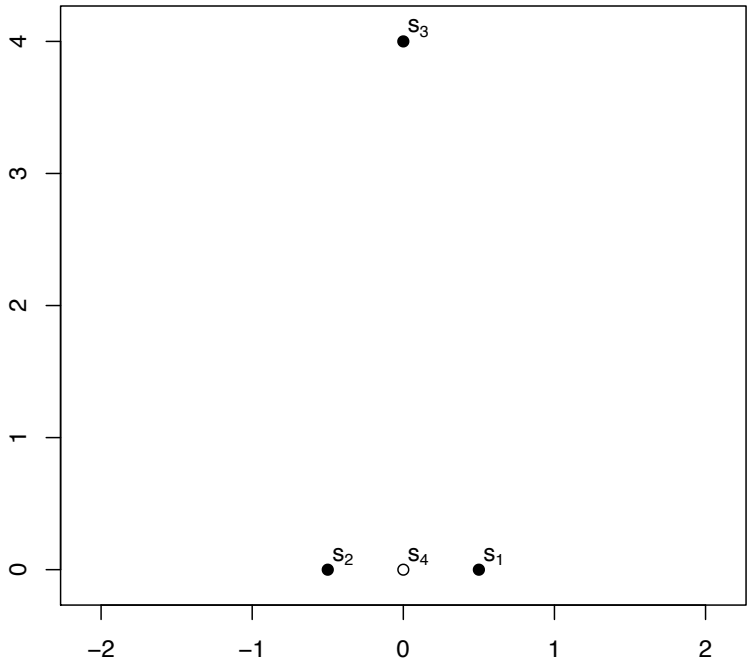


FIGURE 2. Locations in our toy example. Black points correspond to locations where data is observed, the white point is a location where prediction is desired.

is a scalar, and $f(\cdot|\theta(\mathbf{s}))$ is a density function, e.g. $\mathsf{N}(\mu, \theta(\mathbf{s}))$ (as in Section 7.1) or $\mathsf{Poi}(\theta(s))$ (as in Section 7.2), and let $\theta(\mathbf{s}) \sim \mathsf{LaSBP}_L(\{a_l\}_{l=1}^L, \{b_l\}_{l=1}^L, H, \gamma)$. This implies that $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ are correlated since $\theta(\mathbf{s})$ and $\theta(\mathbf{s}')$ are correlated, and that marginally, $Y(s)|P \sim \sum_{l=1}^L w_l f(\cdot|\theta_l^*)$, which defines the same marginal for any $\mathbf{s}$ (although possibly multimodal).

4. PROPERTIES

In what follows, let $\xi(\mathbf{s}) = l$ iff $z(\mathbf{s}) \in [\Phi^{-1}(\pi_{l-1}), \Phi^{-1}(\pi_l))$ be the labeling process indicating the membership of locations to components of the stick-breaking process and $\beta_l(\mathbf{s}) = \mathsf{Pr}(\xi(\mathbf{s}) =$

l) represent the marginal probability that a realization of the process at location $\mathbf{s}$ is assigned to component l. Similarly, let $\beta_{lk}(\mathbf{s},\mathbf{s}') = \mathsf{Pr}(\xi(\mathbf{s}) = l, \xi(\mathbf{s}') = k|\{v_l\}_{l=1}^L)$ be the joint probability of a realization of the process at locations $\mathbf{s}$ and $\mathbf{s}'$ taking values θ_l^* and θ_k^* respectively. By definition, $\beta_k(\mathbf{s}) = \mathsf{Pr}(\xi(\mathbf{s}) = k|\{v_l\}_{l=1}^L) = w_k$ and

$$\beta_{lk}(\mathbf{s},\mathbf{s}') = \mathbf{\Phi}_{\gamma(\mathbf{s},\mathbf{s}')}(\Phi^{-1}(\pi_l), \Phi^{-1}(\pi_k)) - \mathbf{\Phi}_{\gamma(\mathbf{s},\mathbf{s}')}(\Phi^{-1}(\pi_{l-1}), \Phi^{-1}(\pi_k)) - \mathbf{\Phi}_{\gamma(\mathbf{s},\mathbf{s}')}(\Phi^{-1}(\pi_l), \Phi^{-1}(\pi_{k-1})) + \mathbf{\Phi}_{\gamma(\mathbf{s},\mathbf{s}')}(\Phi^{-1}(\pi_{l-1}), \Phi^{-1}(\pi_{k-1}))$$

where $\mathbf{\Phi}_r(\cdot,\cdot)$ denotes the cumulative distribution of the standard bivariate normal with correlation r. Note that these probabilities are random *a priori* since $w_1, \ldots, w_L$ are random.

First, consider the expectation of the LaSBP process $\theta(\mathbf{s})$. If ψ is a measurable function on $(\mathbf{\Theta}, \mathcal{B})$,

$$\mathsf{E}(\psi(\theta(\mathbf{s}))) = \sum_{l=1}^{L} \mathsf{E}(w_l)\mathsf{E}(\psi(\theta_l^*)) = \sum_{l=1}^{L} \frac{a_l}{a_l + b_l} \left[\prod_{k<l} \frac{b_k}{a_k + b_k}\right] \mathsf{E}(\psi(\theta_l^*)) > \mathsf{E}_H(\psi(\theta_1^*))$$

Due to the order constraint, some care must be exercised when interpreting the baseline measure in our model. Although H still plays a role controlling the average level of the process, it is not true in general that the level is independent of the choice of the sequences a_l and b_l, as with typical stick-breaking processes. Also, the average level of the process will be strictly larger than the mean of the baseline measure H.

Example 1. Consider $\theta(\mathbf{s}) \sim LSBP_L(a, b, H)$, where H is an exponential baseline measure with mean λ. This coincides with the illustrations we present in Section 7. In this case, $\theta_l^*|\theta_{l-1}^*$ follows a shifted exponential distribution and $\mathsf{E}(\theta_l^*|\theta_{l-1}^*) = \lambda + \theta_{l-1}^*$. This implies that $\mathsf{E}(\theta_l^*) = l\lambda$ and

$$\begin{aligned}\mathsf{E}(\theta(\mathbf{s})) &= \sum_{l=1}^{L} \lambda l \frac{a}{a+b}\left(\frac{b}{a+b}\right)^{l-1} \\ &= \lambda\frac{a+b}{a}\left[1 - \left(\frac{b}{a+b}\right)^{L+1} - (L+1)\left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)^{L}\right]\end{aligned}$$

Note that

$$\lim_{L\to\infty} \mathsf{E}[\theta(\mathbf{s})] = \lambda\frac{a+b}{a}$$

Therefore the process is well defined (in the sense that it has a finite first moment) even with an infinite number of atoms. A similar argument yields

$$\lim_{L\to\infty} \mathsf{E}[(\theta(\mathbf{s}))^2] = 2\lambda^2\left(\frac{a+b}{a}\right)^2 \quad \text{and} \quad \lim_{L\to\infty} \mathsf{V}[\theta(\mathbf{s})] = \lambda^2\left(\frac{a+b}{a}\right)^2$$

Example 1 illustrates that in many cases we can interpret finite LaSBPs as truncations of infinite processes. In this situation, the theory of truncations developed in Ishwaran and James (2001), Ishwaran and James (2003) or Rodriguez et al. (2007) can be directly modified for the LaSBP. The following theorem gives sufficient conditions for the limiting process to have finite expectation.

Theorem 2. *Let $\theta(\mathbf{s}) \sim LSBP_L(a, b, H)$, where the baseline measure H is such that, for large l,*

$$\mathsf{E}(\theta_l^*|\theta_{l-1}^*) \leq \mu + \lambda\theta_{l-1}^*$$

for some real constants μ and λ such that $0 < \lambda \leq 1$. Then

$$\lim_{L\to\infty} \mathsf{E}(\theta(\mathbf{s})) < \infty$$

Corollary 1. *If $\mathsf{E}(\theta_l^*) < c_1 l^p + c_2\lambda^l$ for some real constants c_1, c_2 and λ with $0 < \lambda \leq 1$ and integer constant p, then $\lim_{L\to\infty} \mathsf{E}(\theta(\mathbf{s})) < \infty$.*

Corollary 2. *if H has a compact support then $\lim_{L\to\infty} \mathsf{E}(\theta(\mathbf{s})) < \infty$.*

The details of the proof are shown in appendix B, and it can be easily extended to provide conditions for finite higher order moments. As the following two examples show, the condition in theorem 2 is generally satisfied when H corresponds to a distribution with light tails.

Example 2. Let $\theta(\mathbf{s}) \sim LSBP_L(a, b, H)$ where H induces a Gaussian distribution with mean μ and variance σ^2. In this case

$$\mathsf{E}(\theta_l^*|\theta_{l-1}^*) = \mu + \sigma \frac{\phi\left(\frac{\theta_{l-1}^*-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\theta_{l-1}^*-\mu}{\sigma}\right)}$$

where ϕ denotes the density of the standard normal distribution. Using some well known properties (Abramowitz and Stegun (1965), page 298),

$$\sigma \frac{\phi\left(\frac{\theta_{l-1}^*-\mu}{\sigma}\right)}{1-\Phi\left(\frac{\theta_{l-1}^*-\mu}{\sigma}\right)} \leq \frac{1}{\sqrt{2\pi}}\left[(\theta_{l-1}^*-\mu) + \sqrt{(\theta_{l-1}^*-\mu)^2+2\sigma^2}\right] \approx \frac{2}{\sqrt{2\pi}}(\theta_{l-1}^*-\mu)$$

for large θ_{l-1}^*. Since $2/\sqrt{2\pi} < 1$, Theorem 1 applies and $\lim_{L\to\infty} \mathsf{E}(\theta(\mathbf{s})) < \infty$.

Example 3. Consider now a baseline measure inducing a Pareto distribution with density

$$p(\theta) = \frac{\alpha}{\beta}\left(\frac{\beta}{\theta}\right)^{\alpha+1} \quad \theta \geq \beta$$

The conditional density reduces to

$$p(\theta_l^*|\theta_{l-1}^*) = \frac{\alpha}{\theta_{l-1}^*}\left(\frac{\theta_{l-1}^*}{\theta_l^*}\right)^{\alpha+1} \quad \theta_l^* \geq \theta_{l-1}^*$$

and $\mathsf{E}(\theta_l^*|\theta_{l-1}^*) = \frac{\alpha}{\alpha-1}\theta_{l-1}^*$. Therefore $\lim_{L\to\infty} \mathsf{E}(\theta(\mathbf{s})) < \infty$ iff $b/(a+b) < (\alpha-1)/\alpha$.

The covariance of the LaSBP $\theta(\mathbf{s})$ process is given by

$$\mathsf{Cov}(\theta(\mathbf{s}), \theta(\mathbf{s}')) = \sum_{l=1}^{L}\sum_{k=1}^{L}\left[\mathsf{E}(\beta_{lk}(\mathbf{s},\mathbf{s}'))\mathsf{E}(\theta_l^*\theta_k^*) - \mathsf{E}(\beta_l(\mathbf{s}))\mathsf{E}(\beta_k(\mathbf{s}'))\mathsf{E}(\theta_l^*)\mathsf{E}(\theta_k^*)\right]. \tag{3}$$

Although we do not have a closed form expression for the covariance function, it is clear from (3) that the process on $\theta(\mathbf{s})$ is isotropic, stationary or non-stationary if $z(\mathbf{s})$ has these properties.

Realizations from the LaSBP $\theta(\mathbf{s})$ are *discontinuous* almost surely. This is clear from Figure 1, and is a consequence of the stick-breaking construction. However, note that

$$\beta_{lk}(\mathbf{s}, \mathbf{s}') \to \beta_l(\mathbf{s})\beta_k(\mathbf{s}') \quad \text{as} \quad \gamma(\mathbf{s}, \mathbf{s}') \to 0$$

$$\beta_{lk}(\mathbf{s}, \mathbf{s}') \to \begin{cases} \beta_l(\mathbf{s}) & l = k \\ 0 & \text{otherwise} \end{cases} \quad \text{as} \quad \gamma(\mathbf{s}, \mathbf{s}') \to 1$$

In particular, this implies that $\mathsf{Pr}(\theta(\mathbf{s}) = \theta(\mathbf{s}')) \to 1$ as $\mathbf{s}' \to \mathbf{s}$ almost surely. Therefore the selection mechanism (which assigns atoms to locations) is continuous even if the resulting realizations are not. This result also shows that, as the range of the latent Gaussian process goes to zero, samples at different locations become iid observations from P. On the other hand, as the range grows, the LaSBP becomes a constant random surface on D and the corresponding LaSBP mixture reduces to a parametric model with a prior H on the unknown parameter.

It is important to note that $\theta(\mathbf{s})$ and $\theta(\mathbf{s}')$ are never independent. Indeed, even if $\gamma(\mathbf{s}, \mathbf{s}') = 0$, implying that $\mathsf{Pr}(\xi(\mathbf{s}) = l, \xi(\mathbf{s}') = k | \{v_r\}_{r=1}^L) = \mathsf{Pr}(\xi(\mathbf{s}) = l | \{v_r\}_{r=1}^L)\mathsf{Pr}(\xi(\mathbf{s}') = k | \{v_r\}_{r=1}^L) = \beta_l(\mathbf{s})\beta_k(\mathbf{s}')$, it is clear from (3) that, in general, $\mathsf{Cov}(\theta(\mathbf{s}), \theta(\mathbf{s}')) \neq 0$. It is also interesting to formally consider the effect of the stick-breaking distribution on the surface. Assume for simplicity that $a_l = a$ and $b_l = b$ for all l, which we will do in most practical applications. Note that as $\frac{a}{a+b} \to 1$ we have $w_1 \to 1$ and the model degenerates again into a flat random surface. Therefore, if the true surface is indeed close to flat, identifiability issues could arise between the spatial range and the parameters of the stick-breaking distributions.

Finally, we also note that, due to the order constraint on the atoms, the LaSBP with constant a and b tends to generate skewed marginal distributions as the process will also encourage larger weights for smaller values of θ. This provides our model with particular flavor that differentiates it from the regular Dirichlet process and the Dirichlet-multinomial model underlying the hDP in Petrone et al.

(2007). More precisely, the weights arising from a stick-breaking construction are stochastically ordered. Under the order constraint for the θ_l^*'s, this implies that smaller θ_l^*s receive stochastically larger weights. For scale LaSBP mixtures (as in Sections 7.1 and 7.2), this feature yields models that favor heavier tails for the marginal distribution of the observations than those obtained by dependent Dirichlet process (DDP) mixtures. For location mixtures, the LaSBP favors distributions that have a heavy right or left tail which, in our experience, can be appealing property in many applications. Indeed, there is often prior information available that one tail of the distribution is heavier and is more subject to secondary modes, and our specification allows a natural incorporation of such prior information. In the applications we have encountered, this is almost always the case. To provide a few additional examples: 1) in studies of premature delivery, the natural response is the gestational age at delivery. The right tail tends to be very stable across populations and is very light, while the left tail (corresponding to premature deliveries) is quite heavy & varies substantially across ethnic groups and with other factors; 2) in studies of the level of a pollutant, the left tail (corresponding to unusually low pollution levels) tends to be very light relative to the right tail (corresponding to high pollution levels); and (3) similar behavior is observed typically for continuous adverse health responses, such as frequency of DNA strand breaks in cells. In applications where this property not appealing, it can be mitigated through non-constant sequences $\{a_l\}_{l=1}^L$ and $\{b_l\}_{l=1}^L$ controlling the stick breaking weights.

5. INFERENCE FOR MIXTURES OF LASBP

For inference purposes, we develop an MCMC algorithm. We assume $a_l = a$, $b_l = b$, $L < \infty$ and use a blocked sampler (Ishwaran and James, 2001). If $L = \infty$, a retrospective sampler (Roberts and Papaspiliopoulos, 2008) can be easily obtained as an extension. In order to facilitate computation, the latent processes, $z(\mathbf{s})$ and $\xi(\mathbf{s})$ are sampled explicitly. Assuming that n data points $y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)$ have been collected, one for each location $\mathbf{s}_1, \ldots, \mathbf{s}_n$, and using the familiar bracket notation, the joint

distribution is given in this case by

$$\prod_{i=1}^{n}[y(\mathbf{s}_i)|\theta^*,\xi(\mathbf{s}_i)]\times\prod_{i=1}^{n}[\xi(\mathbf{s}_i)|z(\mathbf{s}_i),\mathbf{v}]\times[z(\mathbf{s}_1),\ldots,z(\mathbf{s}_n)|\lambda]\times\prod_{l=1}^{\infty}[v_i|\alpha]\times\prod_{l=1}^{\infty}[\theta_l]\times[\alpha]\times[\lambda]$$

After setting up initial values for all parameters in the model, the algorithm proceeds by sequentially updating the parameters according to the following steps:

(1) Jointly update the latent processes $\xi(\mathbf{s})$ and $z(\mathbf{s})$, one location at a time, by first sampling $\xi(\mathbf{s}_i)$ from a discrete distribution such that

$$\Pr(\xi(s_i)=l)\propto p(y(\mathbf{s}_i)|\theta_l)\times \Pr(z(\mathbf{s}_i)\in[\Phi^{-1}(\pi_{l-1}),\Phi^{-1}(\pi_l))|z(\mathbf{s}_1),\ldots,z(\mathbf{s}_{i-1}),z(\mathbf{s}_{i+1}),\ldots,z(\mathbf{s}_n))$$

where the prior probability of component l can be obtained by a univariate integration, and then sampling $z(\mathbf{s}_i)$ from the restricted univariate normal distribution defined by

$$[z(\mathbf{s}_i)|z(\mathbf{s}_1),\ldots,z(\mathbf{s}_{i-1}),z(\mathbf{s}_{i+1}),\ldots,z(\mathbf{s}_n)]\mathbf{1}_{\Omega_l}$$

where $\Omega_l=\{z(\mathbf{s}_i):z(\mathbf{s}_i)\in[\Phi^{-1}(\pi_{\xi(\mathbf{s}_i)-1}),\Phi^{-1}(\pi_{\xi(\mathbf{s}_i)}))\}$. Note that, if $\mathbf{s}_i=\mathbf{s}_j$ for some j, then the prior probability for observation i being assigned to component $\xi(\mathbf{s}_j)$ is one and therefore $\xi(\mathbf{s}_i)=\xi(\mathbf{s}_j)$ and $z(\mathbf{s}_i)=z(\mathbf{s}_j)$, as expected.

(2) Sample the stick-breaking ratios one at a time from their full conditional

$$[v_l|\mathbf{v}_{(-l)}]\propto v_l^{a-1}(1-v_l)^{b-1}\mathbf{1}_{A_l}$$

where $A_l=\left\{v_l:q_l^l\leq v_l<q_l^u\right\}$ and

$$q_l^l = \max_{\{i:\xi(\mathbf{s}_i)\geq l\}} \left\{ 1 - \frac{1-\Phi(z(\mathbf{s}_i))}{\prod_{k\leq\xi(\mathbf{s}_i),k\neq l}(1-v_k)} \right\}$$
$$q_l^u = \min_{\{i:\xi(\mathbf{s}_i)\geq l+1\}} \left\{ 1 - \frac{1-\Phi(z(\mathbf{s}_i))}{\prod_{k\leq\xi(\mathbf{s}_i)-1,k\neq l}(1-v_k)} \right\}$$

Note that v_l depends on $\mathbf{v}_{(-l)}$ only through the constraint on the support. Also, v_l depends on $z(\mathbf{s})$ only through $z(\mathbf{s}_i) : \xi(\mathbf{s}_i) \geq l$. Therefore, for $l > l^* = \max_i\{\xi(\mathbf{s}_i)\}$, v_l is conditionally independent from $\mathbf{v}_{(-l)}$.

(3) Sample the atoms $\{\theta_l\}$. The set of full conditionals is given by

$$p(\theta_1|\cdots) \propto h(\theta_1) \prod_{\{i:\xi(\mathbf{s}_i)=1\}} p(\mathbf{y}(\mathbf{s}_i)|\theta_1)$$

and

$$p(\theta_l|\cdots) \propto h(\theta_l|\theta_{l-1}) \prod_{\{i:\xi(\mathbf{s}_i)=l\}} p(\mathbf{y}(\mathbf{s}_i)|\theta_l) \qquad l > 1$$

where h_l is the density associated with the measure H_l.

(4) The prior parameters on the stick-breaking ratios a and b can be jointly sampled using a random-walk Metropolis-Hastings algorithm.

(5) The parameters of the underlying Gaussian process can be sampled conditional on the current realization $z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)$ using a random-walk Metropolis-Hastings algorithm, as is customary in Gaussian process models.

(6) Since the conditional moves in step 1 can raise some concerns about mixing rates, we suggest to additionally sample the whole Gaussian process jointly. This can be easily done by noting that the joint full conditional is given by

$$[z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n)|\cdots]\cdots \sim \mathsf{N}(\mathbf{0}, \mathbf{\Psi})\,\mathbf{1}_\Omega$$

where $[\mathbf{\Psi}]_{ij} = \gamma(\mathbf{s}_i, \mathbf{s}_j)$, $\Omega = \cap_{i=1}^n \Omega_i$ and $\Omega_i = \{z(\mathbf{s}_i) : \Phi^{-1}(\pi_{\xi(\mathbf{s}_i)-1}) \leq z(\mathbf{s}_i) < \Phi^{-1}(\pi_{\xi(\mathbf{s}_i)})\}$ as before. Due to the form of the restrictions, this is a relatively straightforward step.

The supplemental material discusses some details related to efficient computational implementation for this model.

6. MULTIVARIATE LATENT STICK-BREAKING PROCESSES

The order constraint in the definition of the LaSBP makes it hard to directly extend our formulation to multivariate atoms. One alternative we explore here is to use a latent multivariate Gaussian process to induce dependence in the selection mechanism of multiple stick-breaking representations. As an example, consider a two dimensional process $\boldsymbol{\theta}(\mathbf{s}) = (\theta_1(\mathbf{s}), \theta_2(\mathbf{s}))'$ with marginal distributions P_1 and P_2 with stick-breaking representations $P_i(\cdot) = \sum_{l=1}^L w_{il}\delta_{\theta^*_{il}}(\cdot)$ with $\theta^*_{il} \leq \theta^*_{i,l+1}$ for $i = 1, 2$ as defined in Section 3 for the univariate LaSBP, with components independent for $i = 1, 2$. Also, let $(z_1(\mathbf{s}), z_2(\mathbf{s}))' = \mathbf{z}(\mathbf{s}) = \mathbf{A}\mathbf{x}(\mathbf{s})$, where $x_i(\mathbf{s})$ is a standardized Gaussian process with correlation function $\gamma_i(\mathbf{s}, \mathbf{s}')$ and $\mathbf{A}$ is a matrix such that

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ \eta & \sqrt{1-\eta^2} \end{pmatrix} \Rightarrow \mathbf{A}'\mathbf{A} = \begin{pmatrix} 1 & \eta \\ \eta & 1 \end{pmatrix}.$$

Letting $\mathbf{a}_j$ be the j-th column of $\mathbf{A}$, the corresponding cross-covariance matrix for $\mathbf{z}(\mathbf{s})$ is given by $\Sigma_{\mathbf{z}(\mathbf{s}),\mathbf{z}(\mathbf{s}')} = \gamma_1(\mathbf{s}, \mathbf{s}')\mathbf{a}_1\mathbf{a}_1' + \gamma_1(\mathbf{s}, \mathbf{s}')\mathbf{a}_2\mathbf{a}_2'$. In particular, this implies that $\mathsf{V}(z_i(\mathbf{s})) = 1$ for $i = 1, 2$ and $\mathsf{Cov}(z_1(\mathbf{s}), z_2(\mathbf{s})) = \mathsf{Cor}(z_1(\mathbf{s}), z_2(\mathbf{s})) = \eta$. As in Section 3, we can define

$$G(\theta_1(\mathbf{s}_1) = \theta^*_{1l_1}, \ldots, \theta_1(\mathbf{s}_n) = \theta^*_{1l_n}, \theta_2(\mathbf{s}_{n+1}) = \theta^*_{2l_{n+1}}, \ldots, \theta_2(\mathbf{s}_m) = \theta^*_{2l_m} | \{\theta_{1l}, \theta_{2l}\}_{l=1}^L, \{v_{1l}, v_{2l}\}_{l=1}^L) =$$

$$\mathsf{Pr}\left(z_1(\mathbf{s}_1) \in [\Phi^{-1}(\pi_{1,l_1-1}), \Phi^{-1}(\pi_{1,l_1})), \ldots, z_1(\mathbf{s}_n) \in [\Phi^{-1}(\pi_{1,l_n-1}), \Phi^{-1}(\pi_{1,l_n})),\right.$$

$$\left. z_2(\mathbf{s}_{n+1}) \in [\Phi^{-1}(\pi_{2,l_{n+1}-1}), \Phi^{-1}(\pi_{2,l_{n+1}})), \ldots, z_2(\mathbf{s}_m) \in [\Phi^{-1}(\pi_{2,l_m-1}), \Phi^{-1}(\pi_{2,l_m}))\right)$$

where process $\theta_1(\mathbf{s})$ is observed in locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$ and process $\theta_2(\mathbf{s})$ is observed in locations $\mathbf{s}_{n+1}, \ldots, \mathbf{s}_m$. Although these two sets of location will typically agree, the process is well defined even if that is not the case.

Choosing $\eta = 0$ implies that $\theta_1(\mathbf{s})$ is independent from $\theta_2(\mathbf{s})$. More generally, we can calculate the cross-covariance function for this process by noting that

$$\mathsf{Cov}(\theta_i(\mathbf{s}), \theta_j(\mathbf{s}')) = \sum_{l=1}^{L}\sum_{k=1}^{L} \mathsf{E}[\beta_{lk}^{ij}(\mathbf{s}, \mathbf{s}')]\mathsf{E}(\theta_{il}^*)\mathsf{E}(\theta_{jk}^*) - \mathsf{E}[\beta_l^i(\mathbf{s})]\mathsf{E}[\beta_k^j(\mathbf{s})]\mathsf{E}(\theta_{il}^*)\mathsf{E}(\theta_{jk}^*) \tag{4}$$

where, as in Section 4, $\beta_{lk}^{ij}(\mathbf{s}, \mathbf{s}') = \mathsf{Pr}(\Phi(z_i(\mathbf{s})) \in [\pi_{i,l-1}, \pi_{i,l}), \Phi(z_j(\mathbf{s}')) \in [\pi_{j,k-1}, \pi_{j,k}) | \{v_{1l}, v_{2l}\}_{l=1}^{L})$ and $\beta_l^i(\mathbf{s}) = \mathsf{Pr}(\Phi(z_i(\mathbf{s})) \in [\pi_{i,l-1}, \pi_{i,l}) | \{v_{il}\}_{l=1}^{L}) = w_{il}$. As expected, (4) reduces to (3) when $i = j$.

The use of dependent Gaussian processes typically provides an interesting interpretation for the model. For example, in the context of the stochastic volatility model described in Section 7.1, the dependence between latent Gaussian processes measures the correlation between market sentiments across different assets. Additionally, sampling for this multivariate extension can proceed along the lines described in section 5, with an additional step to sample the parameters in $\mathbf{A}$ conditionally on all other parameters.

The multivariate extension we describe in this Section is really just a straightforward modification of the univariate case in that the order restrictions are placed on the atoms in each marginal in the same manner as in the univariate case. Hence, the assumptions are not more stringent than in the univariate case, so our previous comments comparing the LaSBP with the hDP in the univariate case still apply. As we discussed in Section 4, it is very often the case that such order restrictions are justified. This is also true in the multivariate case. For example, instead of one pollutant, we may measure multiple pollutants, all of which should have a heavier right tail in their concentration distribution a priori. A posteriori these assumptions are not strictly enforced; if there is sufficient evidence in the data to

refute our a priori (informed) guess of the shape of the distribution, the posterior estimate will be appropriately adapted.

7. ILLUSTRATIONS

7.1. **Modeling stochastic volatility.** In this section we consider an application of the LaSBP to the modeling of stock-market returns. This example is interesting as it allows us to compare the LaSBP with other more standard models in a real life example.

The data set under consideration consists of the weekly returns of the S&P500 index covering the ten-year period between April 21, 1997 and April 9, 2007, for a total of 520 observations. Figure 3 shows the evolution of these returns. The series does not exhibit any long term trend, but different levels of volatility can be clearly seen. In particular, two slightly different regimes are apparent; before May 2003, periods of high-volatility are relatively frequent, while after May 2003, we can appreciate longer low-volatility periods. Indeed, it is well known that financial time series typically exhibit heavy tails and that periods of low/high volatility tend to cluster together.

Similarly to other approaches in the literature (Bollerslev, 1986; Jacquier et al., 1994), we model $r(t)$, the log-return of the S&P500 at time t, as following a normal distribution with constant mean μ but time-varying precision $\phi(t)$. The time-varying precision is then assumed to follow a LaSBP with an exponential baseline measure with mean $1/\phi_0$ and a decreasing (rather than increasing) order constraint . Also, we take $a_l = 1$ and $b_l = \alpha$ for all l. We noted before that this choice places stochastically larger probabilities on larger variances. A prior with this structure is well motivated in this application, because low volatility (high precision) is the norm, while high volatility (low precision) is associated with sporadic shocks to the system. To demonstrate the methodology we use an exponential correlation function with range parameter λ for the latent process, but implementations using more general correlation functions (like Matérn, power exponential or some non-stationary family) is

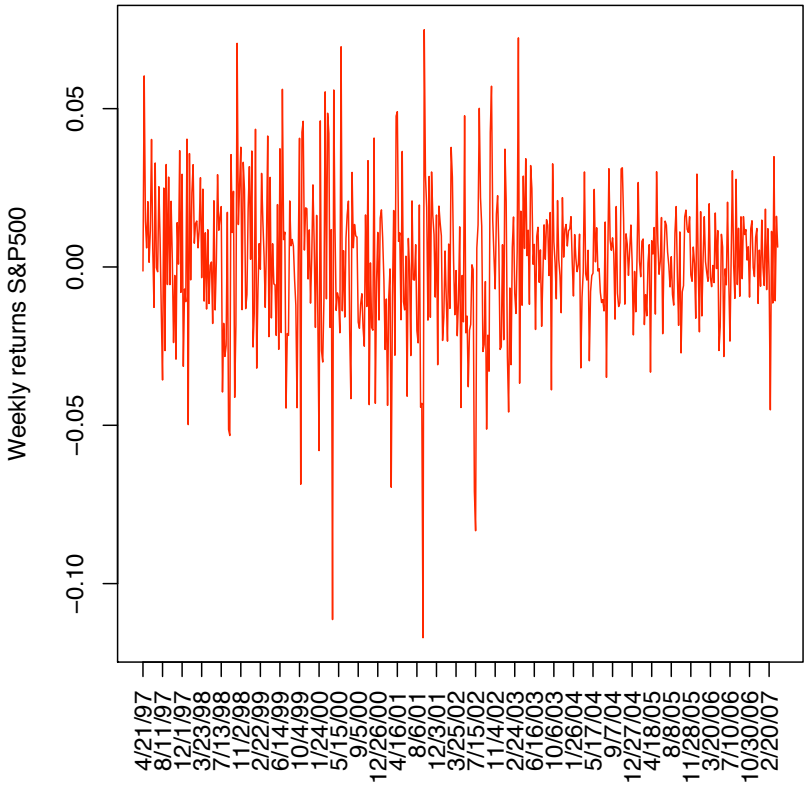


FIGURE 3. Weekly returns on the S&P500 index between April 21, 1997 and April 9, 2007.

just as straightforward. This model results in a marginal distribution that is a scale mixture of Gaussian distributions, yielding a rich model on the class of unimodal distributions that can accommodate heavy tails. Also, the structure of the LaSBP naturally induces volatility clustering, while potentially allowing for richer dependence structures than Markov-switching models.

We complete the model by specifying priors for the different parameters. The mean of the returns is given a normal distribution with mean μ_0 and variance τ^2, the spatial range is given a $\mathsf{Gam}(a_\lambda, b_\lambda)$ distribution, the precision α is assigned a $\mathsf{Gam}(a_\alpha, b_\alpha)$ prior and ϕ_0 is given a $\mathsf{Gam}(\delta, \phi_{00})$ prior. Since the historical annual volatility of the S&P500 is traditionally estimated to be in the range of 12-15%,

we take $\tau^2 = 0.15^2/52$ and $\mathsf{E}(1/\phi_0) = 0.12^2/52$. The prior mean for the average return is taken to be $\mu_0 = 0$, and the degrees of freedom are set to $\delta = 2$. We picked $a_\alpha = b_\alpha = 1$ so that the prior for the precision parameter α is centered around 1 and has variance 1. The prior for the spatial correlation is set as $a_\lambda = 1$ and $b_\lambda = 1/5$, implying that the a-priori range of the latent process driving the correlation is approximately 15 weeks. Forty atoms ($L = 40$) were used in the stick-breaking construction. A small sensitivity analysis was conducted by changing the hyperpriors on σ_0^2, α and λ, as well as L, and results seemed mostly robust to the selection. All results are based on 50,000 iterations obtained after a burn-in period of 5,000 samples. We resampled the latent process every 10 iterations to improve mixing. Visual inspection of the trace plots does not reveal any obvious mixing or convergence problems. Sensitivity analysis was performed by considering values for $\mathsf{E}(1/\phi_0)$ in the range between $0.05^2/52$ and $0.4^2/52$, leading to essentially the same results.

The model identifies between 2 and 11 volatility regimes, with models having between 3 and 6 components receiving the most weight (estimated posterior probabilities 0.27, 0.33, 0.20 and 0.10 respectively). The mean of the precision α is estimated to be 0.700 (median 0.650, symmetric 90% credible interval (0.260 ; 1.312)). The posterior distribution of the correlation parameter λ has a mean of 110.72 (90% credible interval (34.05,241.53)), showing strong evidence of long range dependence in the sample. This feature, which is well known in financial time series, is hard to capture with stochastic volatility models based on low-order autoregressive processes.

The posterior mean weekly return is estimated to be 0.00194 (credible interval (0.00050 ; 0.00336)), a reasonable value given historical evidence. Figure 4 shows the volatility path estimated by the LaSBP model, along with the sample standard deviation and the volatility path obtained from the "standard" stochastic volatility described in Jacquier et al. (1994). Although both models provide similar volatility estimates, there are some interesting differences. For example, the low volatility period between October 2003 and December 2006 is estimated by the LaSBP analysis to be a period

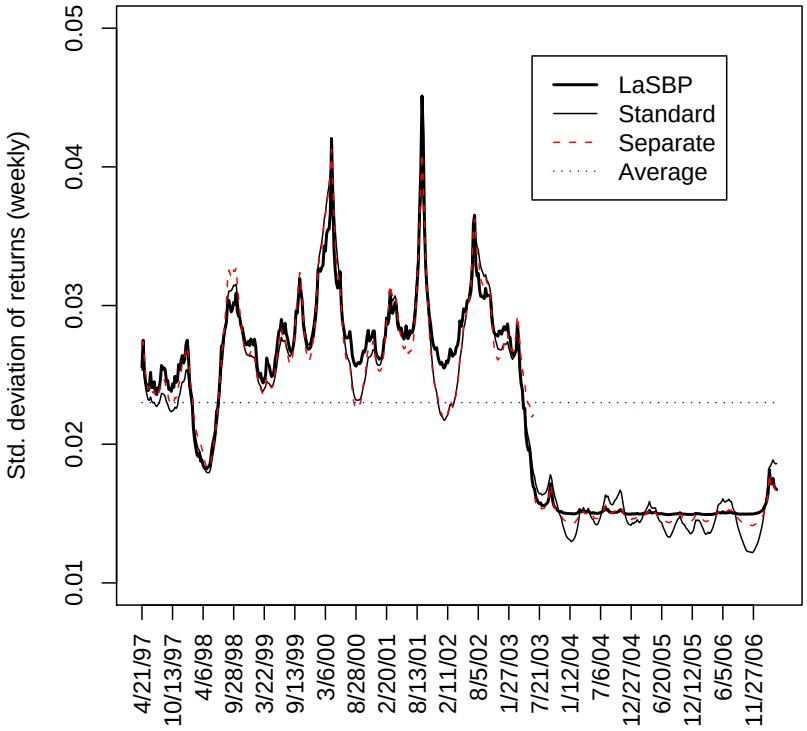


FIGURE 4. Smoothed weekly volatilities for the S&P500 using the LaSBP stochastic volatility model, a standard stochastic volatility model fitted to the whole series, and separate standard stochastic volatility models for the periods before and after 06/16/2003. The flat dotted line corresponds to the empirical standard deviation of the sample.

with essentially constant volatility, while the regular volatility model implies frequent fluctuations. Nonetheless, both models rapidly adapt to the rise in volatility in early 2007. Also the three high volatility peaks in 2000, 2001 and 2002 are more pronounced under our model. In order to better understand the behavior of these models, we also fitted separate standard stochastic volatility models for the periods before and after 06/16/2003. Note that the separate stochastic volatility models fit a

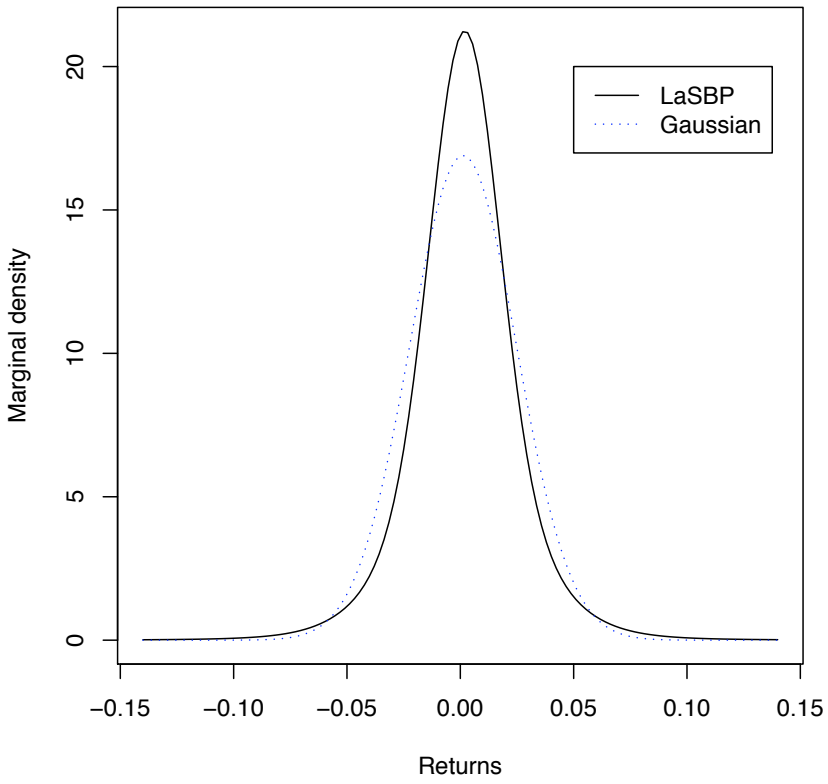


FIGURE 5. Marginal distribution of returns from the S&P500, compared with an empirical fit to the Gaussian distribution.

smoother volatility estimate than the standard model to the second period. This suggests that lack of smoothness in the second period for the standard stochastic volatility model is induced by the lack of flexibility in the model.

Unlike standard volatility models, the LaSBP automatically produces an estimate of the marginal distribution of returns on S&P500, which is shown in Figure 5. As expected from historical evidence, the tails of the estimated marginal distribution are heavier than the tails of a normal distribution with equivalent mean and variance.

7.2. **Understanding species abundance: A model for point-referenced counts.** This Section concentrates on models for spatial count data, in the spirit of Richardson and Green (2001). The Christmas Bird Count (CBC) is an annual census of early-winter bird populations conducted by over 50,000 observers each year between December 14th and January 5th. The primary objective of the Christmas Bird Count is "to monitor the status and distribution of bird populations across the Western Hemisphere". Parties of volunteers follow specified routes through a designated 15-mile diameter circle, counting every bird they see or hear. The parties are organized by compilers who are also responsible for reporting total counts to the organization that sponsors the activity, the Audubon Society. Data and additional details about this survey are available at `http://www.audubon.org/bird/cbc/index.html`.

In this illustration, we concentrate on jointly modeling the abundance of two bird species in North Carolina during the 2006-2007 winter season: *Sturnus vulgaris*, also known as European Starling, and *Zenaida macroura*, also known as Mourning Dove. In this period, forty four circles were active within the state. Since the diameter of the circles is very small compared to the size of the region under study, we treat the data as point referenced to the center of the circle. We elaborate on the model described in Section 6 to construct a bivariate LaSBP mixture of Poisson kernels.

Specifically, letting $y_i(\mathbf{s})$ stand for the number of birds observed at location $\mathbf{s}$ (expressed in latitude and longitude) belonging to species i, we assume $y_i(\mathbf{s}) \sim \mathsf{Poi}(h(\mathbf{s})\lambda_i(\mathbf{s}))$, where $h(\mathbf{s})$ represents the number of man-hours invested at location $\mathbf{s}$. Next, we assume that $(\lambda_1(\mathbf{s}), \lambda_2(\mathbf{s}))$ follows a bivariate LaSBP with exponential baseline measures P_{10} and P_{20} with means λ_{10} and λ_{20}, stick breaking weights $w_{il} = z_{il}\prod_{k<l}(1 - z_{ik})$ with $z_{il} \sim \mathsf{beta}(1, \alpha_i)$, and exponential correlation functions $\gamma_1(\cdot, \cdot)$ and $\gamma_2(\cdot, \cdot)$ with range ρ_1 and ρ_2 respectively.

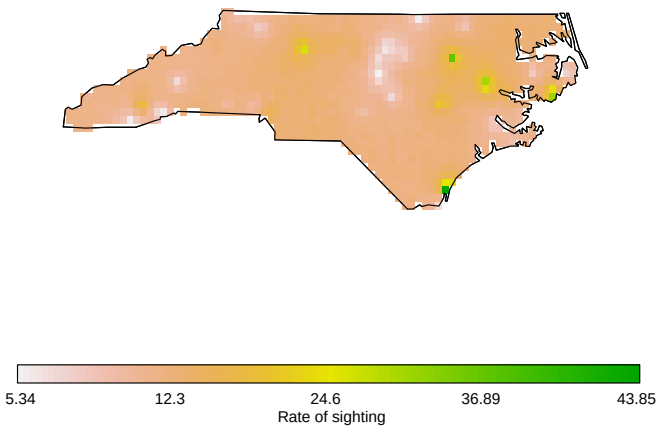
We allow for a hierarchical structure on the mean of the baseline distribution G_0, whose inverse is assumed to follow a gamma distribution, $1/\lambda_{i0} \sim \mathsf{Gam}(2, 2\lambda_{i00})$. Based on historical data, we set

$\lambda_{100} = 10$ and $\lambda_{200} = 4$. The correlation η and range parameters ρ_1 and ρ_2 defining the latent Gaussian processes are assumed unknown, and priors $\eta \sim \mathsf{Uni}[-1, 1]$, $\rho_1 \sim \mathsf{Gam}(1, 1/2)$ and $\rho_2 \sim \mathsf{Gam}(1, 1/2)$ are respectively assigned. Precision parameters are assigned independent priors $\alpha_i \sim \mathsf{Gam}(1, 1)$. Results are based on 100,000 iterations obtained after a burn-in period of 20,000 samples, and the latent process was resampled every 10 iterations. Visual inspection of trace plots reveals no obvious convergence problems, and autocorrelations seem within reasonable levels.

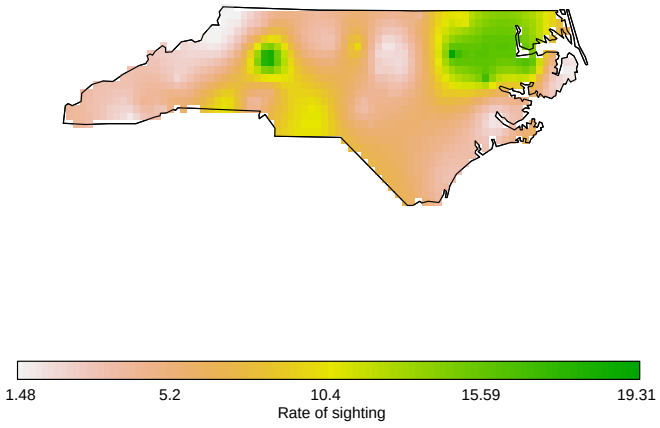
Figure 6 shows the posterior mean surfaces corresponding to the per-man-hour sighting rates for each of the two species. As expected, rates for both species tend to be higher in the coastal plains region; the European Starling and the Mourning Dove tend to be associated with man-altered environments, preferring open country and grasslands, and avoiding woodlands, deserts and swamps. The latent processes appear to be positively correlated, the posterior expectation for η is 0.26 and $(0.12, 0.44)$ is a symmetric, 90% credible interval. However, the difference in the level of spatial dependence between species is striking; while suitable habitats for the Mourning Dove seem to cover large geographical regions, European Starlings seem to concentrate on small microhabitats.

Figure 7 shows estimates of the marginal distribution of the number of sightings in one man-hour for both species under consideration. Both distributions are bimodal and heavy tailed. The smaller mode, which can be interpreted as the baseline rate of sightings, falls around 4 or 6 birds per man per hour for the European Starling and around 2-4 birds for the Mourning Dove. The higher mode, which can be interpreted as corresponding to specially suited habitats, is around 45 sightings per man-hour for the European Sterling and around 23 sightings for the Mourning Dove.

We would like to emphasize that parametric alternatives to our proposed model like a Poisson log-linear model do not allow for a multimodal marginal distribution like the one shown above. On the

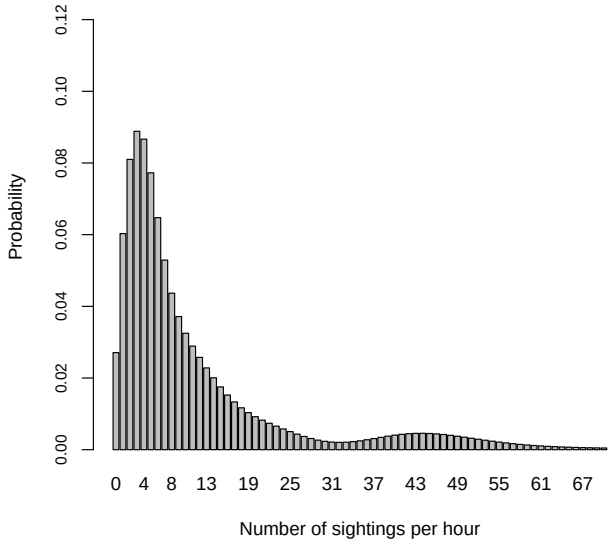


(a)

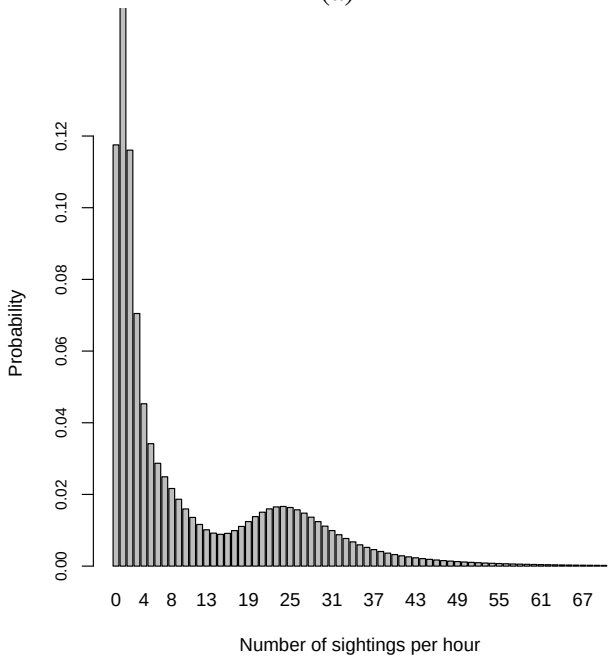


(b)

FIGURE 6. Estimates of the spatial abundance (a) the European Starling and (b) Mourning Dove in North Carolina.



(a)



(b)

FIGURE 7. Marginal distribution of sightings for (a) the European Starling and (b) the Mourning Dove in North Carolina.

other hand, nonparametric alternatives like the spatial DP (Gelfand et al., 2005) and the order dependent DP (Griffin and Steel, 2006b) allow for flexible location-specific distributions for the number of sightings, but again do not provide a common estimate like the one in figure 7.

8. Discussion

We have developed and illustrated a novel mechanism to generate stochastic processes with random marginal distributions. This is done by representing the common marginal distribution using a stick-breaking construction and inducing dependence in the selection mechanism of the atoms. Implementation of these models is straightforward using standard Markov chain Monte Carlo algorithms. One important difference between our model and other hybridization schemes is its ability to provide reasonable predictions and interpolations. Indeed, we have shown that some sort of identifiability constraint on the atoms is necessary in order to avoid multimodality in the posterior distribution due to label switching.

LaSBPs are very flexible. Unlike other nonparametric models in the literature, implementation of anisotropic and non-stationary models is straightforward, as is the modeling of discrete or continuous data. In particular, we are currently exploring hierarchical specifications for the latent process that also allow efficient high-dimensional specifications where the correlation between processes varies in space.

Appendix A. Proof of theorem 1

This is a simple consequence of the properties of a Gaussian process, and can be extended for any other consistent selection process (in the Kolmogorov sense). Note that, from equation (2), the finite-dimensional distributions can be represented as mixtures. Therefore, it is enough to prove that the process satisfies Kolmogorov's conditions conditional on $\{v_l\}_{l=1}^L$ and $\{\theta_l^*\}_{l=1}^L$, as mixtures of

consistent processes are consistent. Symmetry is immediate,

$$\begin{aligned}
G\left(\theta(\mathbf{s}_{\nu_1}) = \theta^*_{l_{\nu_1}}, \ldots, \theta(\mathbf{s}_{\nu_n}) = \theta^*_{l_{\nu_n}} \middle| \{\theta^*_l\}_{l=1}^L, \{v_l\}_{l=1}^L\right) \\
&= \mathsf{Pr}\left(z(\mathbf{s}_{\nu_1}) \in [\Phi^{-1}(\pi_{l_{\nu_1}-1}), \Phi^{-1}(\pi_{l_{\nu_1}})), \ldots, z(\mathbf{s}_{\nu_1}) \in [\Phi^{-1}(\pi_{l_{\nu_n}-1}), \Phi^{-1}(\pi_{l_{\nu_n}})\right) \\
&= \mathsf{Pr}\left(z(\mathbf{s}_1) \in [\Phi^{-1}(\pi_{l_1-1}), \Phi^{-1}(\pi_{l_1})), \ldots, z(\mathbf{s}_n) \in [\Phi^{-1}(\pi_{l_n-1}), \Phi^{-1}(\pi_{l_n})\right) \\
&= G\left(\theta(\mathbf{s}_1) = \theta^*_{l_1}, \ldots, \theta(\mathbf{s}_n) = \theta^*_{l_n} \middle| \{\theta^*_l\}_{l=1}^L, \{v_l\}_{l=1}^L\right)
\end{aligned}$$

For consistency

$$\begin{aligned}
\sum_{u \in \{\theta^*_l\}_{l=1}^L} G\left(\theta(\mathbf{s}_1) = \theta^*_1, \ldots, \theta(\mathbf{s}_n) = \theta^*_n, \theta(\mathbf{s}_{n+1}) = u \middle| \{\theta^*_l\}_{l=1}^L, \{v_l\}_{l=1}^L\right) \\
&= \sum_{l_{n+1}=1}^{L} \mathsf{Pr}\Big(z(\mathbf{s}_1) \in [\Phi^{-1}(\pi_{l_1-1}), \Phi^{-1}(\pi_{l_1})), \ldots, z(\mathbf{s}_{n+1}) \in [\Phi^{-1}(\pi_{l_{n+1}-1}), \Phi^{-1}(\pi_{l_{n+1}})\Big) \\
&= \mathsf{Pr}\left(z(\mathbf{s}_1) \in [\Phi^{-1}(\pi_{l_1-1}), \Phi^{-1}(\pi_{l_1})), \ldots, z(\mathbf{s}_{n+1}) \in (-\infty, \infty)\right) \\
&= \mathsf{Pr}\left(z(\mathbf{s}_1) \in [\Phi^{-1}(\pi_{l_1-1}), \Phi^{-1}(\pi_{l_1})), \ldots, z(\mathbf{s}_n) \in [\Phi^{-1}(\pi_{l_n-1}), \Phi^{-1}(\pi_{l_n}))\right) \\
&= G\left(\theta(\mathbf{s}_1) = \theta^*_1, \ldots, \theta(\mathbf{s}_n) = \theta^*_n \middle| \{\theta^*_l\}_{l=1}^L, \{v_l\}_{l=1}^L\right)
\end{aligned}$$

APPENDIX B. PROOF OF THEOREM 2

If $\mathsf{E}(\theta_l|\theta_{l-1}) \leq \mu + \lambda\theta_{l-1}$ then solving the recursion

$$\mathsf{E}(\theta_l) \leq \mu \sum_{k=0}^{l-1} \lambda^k + \lambda^l \mathsf{E}(\theta_1) = \frac{\mu}{1-\lambda} + \lambda^l(\mathsf{E}(\theta_1) - \mu)$$

and

$$\begin{aligned}
\mathsf{E}(\theta(\mathbf{s})) = \sum_{l=1}^{\infty} \mathsf{E}(\theta_l) \frac{a}{a+b}\left(\frac{b}{a+b}\right)^{l-1} &\leq \sum_{l=1}^{\infty} \left[\frac{\mu}{1-\lambda} + \lambda^l(\mathsf{E}(\theta_1) - \mu)\right] \frac{a}{a+b}\left(\frac{b}{a+b}\right)^{l-1} \\
&= \frac{\mu}{1-\lambda} + \frac{a\lambda(E(\theta) - \mu)}{a + b(1-\lambda)}
\end{aligned}$$

only if $\lambda b/(a+b) < 1$. This result can be easily extended to the case where $\mathsf{E}(\theta_l) \leq c_1 l^p + c_2 \lambda^l$ by noting that the expected value of the first term corresponds to the value of the polylogarithm function of order p (Abramowitz and Stegun, 1965), which is finite.

REFERENCES

Abramowitz, M. and I. A. Stegun (Eds.) (1965). *Handbook of Mathematical Functions* (First Edition ed.). Dover Publication.

Beal, M. J., Z. Ghahramani, and C. E. Rasmussen (2001). The infinite hidden markov model. In *Proceedings of Fourteenth Annual Conference on Neural Information Processing Systems*.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* **31**, 307–327.

DeIorio, 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**, 205–215.

Duan, J. A., M. Guindani, and A. E. Gelfand (2007). Generalized spatial Dirichlet process models. *Biometrika , in press*.

Dunson, D. B. and J.-H. Park (2007). Kernel stick-breaking processes. *Biometrika In press*.

Dunson, D. B., N. Pillai, and J.-H. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B.* **69**, 163–183.

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

Figueiredo, M. A., D. S. Cheng, and V. Murino (2007). Clustering under prior knowledge with application to image segmentation. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19*, pp. 401–408. Cambridge, MA: MIT Press.

Figueiredo, M. A. T. (2005). Bayesian image segmentation using Gaussian field priors. In A. Rangarajan, B. Vemuri, and A. L. Yuille (Eds.), *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pp. 74–89. Springer - Verlag, Berlin Heidelberg.

Fox, E., E. Sudderth, M. I. Jordan, and A. Willsky (2008). An hdp-hmm for systems with state persistence. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*.

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**, 1021–1035.

Griffin, J. E. and M. F. J. Steel (2006a). Nonparametric inference in time series problems. Technical report, Department of Statistics, University of Warwick.

Griffin, J. E. and M. F. J. Steel (2006b). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* **101**, 179–194.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics* **1**, 265–283.

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

Ishwaran, H. and L. F. James (2003). Some further developments for stick-breaking priors: finite and infinite clustering and classification. *Sankya* **65**, 577–592.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994). Bayesian analysis of stochastic volatility models. *Journal of business and Economic Statistics* **12**, 371–389.

Kottas, A., P. Müller, and F. Quintana (2005). Nonparametric Bayesian modeling for multivariate ordinal data. *Journal of Computational and Graphical Statistics* **14**, 610–625.

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

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Ohio State University, Department of Statistics.

MacEachern, S. N. (2007). Discussion of "Bayesian nonparametric modelling for spatial data using Dirichlet processes" by gelfand, a. e., guindani, m. and petrone, s. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 8*, pp. 196–198. Oxford University Press.

Müller, P., F. Quintana, and G. Rosner (2004). Hierarchical meta-analysis over related non-parametric Bayesian models. *Journal of Royal Statistical Society, Series B* **66**, 735–749.

Nelsen, R. B. (1999). *An introduction to copulas*. Springer.

Ongaro, A. and C. Cattaneo (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics and Probability Letters* **67**, 33–45.

Pal, N. R. and S. K. Pal (1993). A review on image sementation techniques. *Pattern Recognition* **26**, 1277–1294.

Petrone, S., M. Guindani, and A. E. Gelfand (2007). Hybrid dirichlet processes for functional data. *Journal of the Royal Statistical Society, Series B Submitted*.

Pham, D. L., C. Xu, and J. L. Prince (2000). A survey of current methods in medial image segmentation. *Review of Biomedical Engineering* **2**, 315–337.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In T. S. Ferguson, L. S. Shapeley, and J. B. MacQueen (Eds.), *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, pp. 245–268. Hayward, CA:IMS.

Richardson, S. and P. J. Green (2001). Hidden markov models and disease mapping. *Journal of American Statistical Association* **97**, 1055–1070.

Roberts, G. and O. Papaspiliopoulos (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–186.

Rodriguez, A., D. B. Dunson, and A. E. Gelfand (2007). The nested Dirichlet process, with Discussion. *Journal of American Statistical Association Accepted*.

Rodriguez, A., D. B. Dunson, and J. Taylor (2007). Bayesian hierarchically weighted finite mixtures models for samples of distributions. *Biostatistics Under revision*.

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

Suri, J. S., D. Wilson, and S. Laxminarayan (Eds.) (2005). *Handbook of Biomedical Image Analysis. Volume 2: Segmentation Models. Part B*. Springer.

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

van Gael, J., Y. Saatci, Y.-W. Teh, and Z. Ghahramani (2008). Beam sampling for the infinite hidden markov model. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*.