Nonparametric Bayesian models through probit stick-breaking processes

Abel Rodriguez

University of California, Santa Cruz, USA

David B. Dunson

Duke University, Durham, USA

Summary. We describe a novel class of Bayesian nonparametric priors based on stick-breaking constructions where the weights of the process are constructed as probit transformations of normal random variables. We show that these priors are extremely flexible, allowing us to generate a great variety of models while preserving computational simplicity. Particular emphasis is placed on the construction of rich temporal and spatial processes, which are applied to two problems in finance and ecology.

Keywords: Nonparametric Bayes; Random Probability Measure; Stick-breaking Prior; Mixture Model; Data Augmentation; Spatial Data; Time Series

1. Introduction

Bayesian nonparametric (BNP) mixture models have become extremely popular in the last few years, with applications in fields as diverse as finance (Kacperczyk et al., 2003; Rodriguez & ter Horst, 2009), econometrics (Chib & Hamilton, 2002; Hirano, 2002), image analysis (Han et al., 2008; Orbanz & Buhlmann, 2008), genetics (Medvedovic & Sivaganesan, 2002; Dunson et al., 2008), medicine (Kottas et al., 2002) and auditing (Laws & O'Hagan, 2002). In the simple case where we are interested in estimating a single distribution from an independent and identically distributed sample $y_1, \ldots, y_n$, nonparametric mixtures assume that observations arise from a convolution

$$y_j \sim \int k(\cdot|\phi)G(\mathrm{d}\phi)$$

where $k(\cdot|\phi)$ is a given parametric kernel indexed by ϕ, and G is a mixing distribution, which is assigned a flexible prior. For example, assuming that G follows a Dirichlet process (DP) prior (Ferguson, 1973; Blackwell & MacQueen, 1973; Ferguson, 1974; Sethuraman, 1994) leads to the well known Dirichlet process mixture (DPM) models (Lo, 1984; Escobar, 1994; Escobar & West, 1995).

Many recent developments in BNP mixture models have concentrated on models for *collections* of distributions defined on an appropriate space S (e.g., $S \subset \mathbb{R}^2$ for spatial processes and $S \subset \mathbb{N}$ for temporal processes observed in discrete time). Unlike traditional parametric models, where only a limited number of the features of the distribution are allowed to change with covariates, these models provide additional flexibility by allowing the mixing distribution $G_{\mathbf{s}}$ to change with $\mathbf{s} \in S$ while inducing dependence among the members of the collection. A number of different approaches have been developed with this goal in mind, including those based on mixtures of independent processes (Müller et al., 2004; Dunson, 2006; Griffin & Steel, 2006a; Dunson et al., 2007), and approaches that induce dependence in the weights and/or atoms of different $G_{\mathbf{s}}$ (MacEachern, 1999, 2000; DeIorio et al., 2004; Gelfand et al., 2005; Teh et al., 2006; Griffin & Steel, 2006b; Duan et al., 2007; Rodriguez & Ter Horst, 2008; Rodriguez et al., 2008). In this paper we pursue this last direction to construct dependent nonparametric priors.

The design of BNP models requires a delicate balance between the need for simple and computationally efficient algorithms and the need for priors with large enough support. Introducing dependence only in the atoms of $G_{\mathbf{s}}$, which has been an extremely popular approach, typically leads to relatively simple computational algorithms but does not afford sufficient flexibility. For example, MacEachern (2000) shows that constant-weight models cannot accommodate collections of independent distributions. On the other hand, inducing complex dependence structure in the weights (e.g., periodicities) can be a hard task and typically leads to complex and inefficient computational algorithms, limiting the applicability of the models.

This paper proposes a novel approach to construct rich and flexible families of nonparametric priors that allow for simple computational algorithms. Our approach, which we call a probit stick-breaking process (PSBP), uses a stick-breaking construction similar to the one underlying the Dirichlet process (Sethuraman, 1994; Ongaro & Cattaneo, 2004), but replaces the characteristic beta distribution in the definition of the sticks by probit transformations of normal random variables. Therefore, the resulting construction for the weights of the process is reminiscent of the continuation ratio probit model popular in survival analysis (Agresti, 1990; Albert & Chib, 2001). Although we emphasize the construction of temporal and spatial models, this strategy is extremely flexible and allows us to create all sorts of nonparametric models, such as nonparametric random effects and ANOVA models, as well as nonparametric regression models. Indeed, a similar approach has been used to create nonparametric factor models and nonparametric variable selection procedures in Rodriguez et al. (2009) and Chung & Dunson (2009)

The remaining of the paper is organized as follows: After a brief review of the literature on stick-breaking priors, Section 2 describes the probit stick-breaking processes for a single probability measure and studies its theoretical properties. The PSBP is extended in Section 3 to model collections of distributions. This section also presents some specific examples, including temporal and spatial models for distributions. Section 4 discusses efficient computational implementation for PSBP models using collapsed samplers. In Section 5, we present two illustrations where the probit stick-breaking process is used to construct flexible nonparametric spatial and temporal models in the context of applications to finance and environmental sciences. Finally, Section 6 presents our conclusions and future research directions.

2. Stick-breaking priors for discrete distributions

Let $(\mathcal{X}, \mathcal{B})$ be a complete and separable metric space (typically $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{B}$ are the Borel sets on $\mathbb{R}^n$), and let $G \in \mathcal{G}$ be its associated probability measure. G follows a stick-breaking prior with centering measure G_0 and shape measure $H_1, H_2, \ldots$ if and only if it admits a representation of the form:

$$G(\cdot) = \sum_{l=1}^{L} w_l \delta_{\boldsymbol{\theta}_l}(\cdot) \tag{1}$$

where the *atoms* $\{\boldsymbol{\theta}_l\}_{l=1}^{L}$ are independent and identically distributed from G_0 and the stick-breaking weights are defined $w_l = u_l \prod_{r<l}(1 - u_r)$, where the *stick-breaking ratios* are independently distributed $u_l \sim H_l$ for $l < L$ and $u_L = 1$. The number of atoms L can be finite (either known or unknown) or infinite. For example, taking $L = \infty$, and having $u_l \sim \mathsf{Beta}(1 - a, b + la)$ for $0 \leq a < 1$ and $b > -a$ yields the two-parameter Poisson-Dirichlet Process, also known as the Pitman-Yor Process (Ishwaran & James, 2001), with the choice $a = 0$ and $b = \eta$ resulting in the Dirichlet Process (DP) (Ferguson, 1973, 1974; Sethuraman, 1994).

The use of beta random variables to define stick-breaking priors is customary because it endows the process with some interesting and useful properties. For example, Pitman-Yor processes can be characterized by a generalized Pólya urn (Pitman, 1995, 1996; Ishwaran & James, 2001), which has some computational advantages. Although having a predictive rule for the process is an appealing property, we argue in this paper that other distributions on the stick-breaking weights can be used to create very flexible nonparametric priors while preserving computational simplicity.

In the sequel, instead of using beta random variables, we will concentrate on stick-breaking weights constructed as

$$u_l = \Phi(\alpha_l) \qquad\qquad \alpha_l \sim \mathsf{N}(\mu, \sigma^2)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function for the standard normal distribution. Setting $\mu = 0$ and $\sigma = 1$ trivially leads to $u_l \sim \mathsf{Uni}[0, 1]$ (and therefore to a DP with precision parameter $\eta = 1$) while different mean parameters produce distributions for u_l that are right skewed (if $\mu < 0$) or left skewed (if $\mu > 0$). For a finite L, the construction of the weights ensures that $\sum_{l=1}^{\infty} w_l = 1$. When $L = \infty$, it is easy to check that

$$\sum_{l=1}^{\infty} \mathsf{E}(\log(1 - u_l)) = -\infty$$

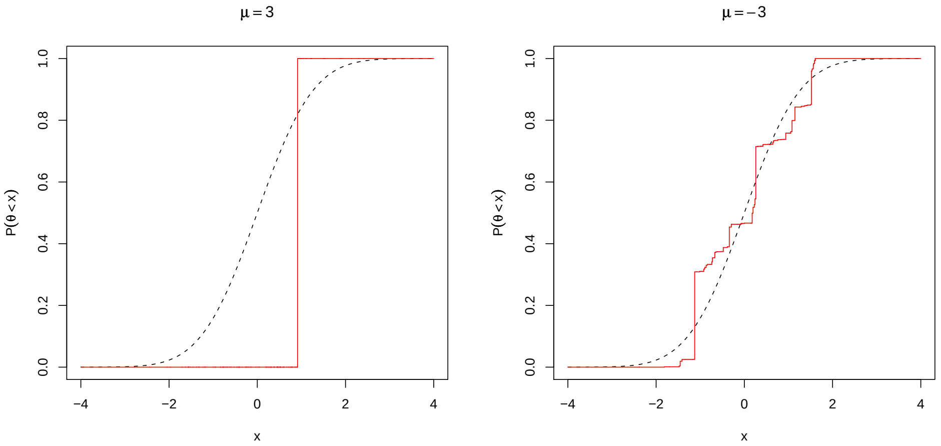

Fig. 1. Random realizations of a probit stick-breaking process with a standard Gaussian centering measure (dashed line) with $\sigma = 1$. The two plots demonstrate the effect of the precision parameter μ on the realizations.

and therefore $\sum_{l=1}^{\infty} w_l = 1$ almost surely (see Appendix A). Similarly, for any measurable set $B \in \mathcal{B}$ the first and second moments are given by

$$\mathsf{E}(G(B)) = G_0(B)$$

$$\mathsf{Var}(G(B)) = G_0(B)(1 - G_0(B))\beta_2 \left\{ \frac{1 - (1 - 2\beta_1 + \beta_2)^L}{2\beta_1 - \beta_2} \right\}$$

where $\beta_1 = \mathsf{E}(u_l) = \Phi(\mu/\sqrt{1+\sigma^2})$ and $\beta_2 = \mathsf{E}(u_l^2) = \mathsf{Pr}(T_1 > 0, T_2 > 0)$, with (T_1, T_2) following a joint bivariate normal distribution such that $\mathsf{E}(T_i) = \mu$, $\mathsf{Var}(T_i) = 1 + \sigma^2$ and $\mathsf{Cov}(T_1, T_2) = \sigma^2$ (see Appendix A). Therefore, we can interpret G_0 as the centering measure and μ and σ^2 as controlling the variance of the sampled distributions around the mean G_0. Indeed, note that, since $1 - 2\beta + \beta_2 < 1$, then $\lim_{L \to \infty} V(G(B)) = G_0(B)(1 - G_0(B)\beta_2/(2\beta_1 - \beta_2)$. Also, note that $\mathsf{Var}(G(B))$ is increasing in μ, and as $\mu \to \infty$ the random distribution G becomes a point mass at a random location $\boldsymbol{\theta}$ almost surely. Figure 1 serves to illustrate the properties just discussed.

The structure of the stick-breaking weights is reminiscent of the continuation ratio probit models used in discrete-time survival analysis (Agresti, 1990; Albert & Chib, 2001). In this setting, the stick-breaking weight w_l represents the hazard of an individual "dying" at time l. Unlike the Dirichlet process, the probit stick-breaking prior does not form a conjugate family on the space of probability measures, in the sense that the posterior distribution for G is not a probit stick-breaking distribution. However this is not an obstacle for computation (see Section 4).

Since a distribution G sampled from a PSBP will be discrete almost surely, a sequence of values sampled from G has a positive probability of showing ties. In a PSBP mixture, the pattern of these ties in the parameters induces a partition of the observations into groups. Therefore, when the PSBP is used for clustering purposes it is important to understand the structure of the partitions generated by the model. In particular, we are interested in how the expected number clusters (corresponding to the distinct number of values in a sample from G) grows as the sample size n grows, and on how uniformly are the observations assigned to the clusters. For non-atomic centering measures, these are controlled exclusively by the precision parameters μ and σ. The top panel of Figure 2 shows the expected number of groups against the logarithm of the number of observations for $\sigma = 1$ and different values of μ, and compares the clustering properties of the PSBP against the Dirichlet process. These expected values were approximated using a Monte Carlo method that involves retrospective sampling (Roberts & Papaspiliopoulos, 2008). In order to simplify interpretation, we compare processes with the same value of $\beta_1 = \mathsf{E}(u_l)$ (remember that for the DP, $\mathsf{E}(u_l) = 1/(1+\eta)$, while for the PSBP $\mathsf{E}(u_l) = \Phi(\mu/\sqrt{1+\sigma^2})$). In first place, we note that the rate of growth of the number of clusters with the sample size in the probit stick-breaking process is logarithmic, just as in the Dirichlet process. Also, since the processes are equivalent for $\beta = 0.5$, the curves agree. However, the distribution of the number of clusters under the probit stick-breaking process seems to be more right (left) skewed for $\beta > 0.5$ ($\beta < 0.5$). In any case, the differences are small, even for sample sizes of about 2000 observations. This pattern is similar for values of σ different from 1.

The bottom panel of Figure 2 shows curves relating the expected number of clusters to the expected size of the largest cluster, for a sample of $n = 200$ observations. Each curve corresponds to a different value of σ, while points on the curve are generated by changing μ between -1 and 1. This plot hints at the different roles μ and σ play in controlling the clustering structure of the model: for small σ, the size of the largest cluster declines more rapidly with a decreasing value of μ, and hence an increasing expected number of clusters, than for larger values of σ. Indeed, very large values of σ tend to induce a single very large cluster accompanied by many smaller clusters.

In the case of finite PSBP models where $L < \infty$, it is important to verify that the behavior of model is in some sense consistent as the number of components grows. In particular we would like to check whether, as $L \to \infty$, the finite model converges to the infinite process. This is important both for computational reasons (as finite truncations can provide a simple algorithm for model fitting) and for the robustness of the model (if there is inconsistency, then the model will typically be sensitive to the choice of L). As Ishwaran & James (2001) point out, this property cannot be taken for granted and must be checked.

The following result shows that, for the probit stick-breaking process, truncations are

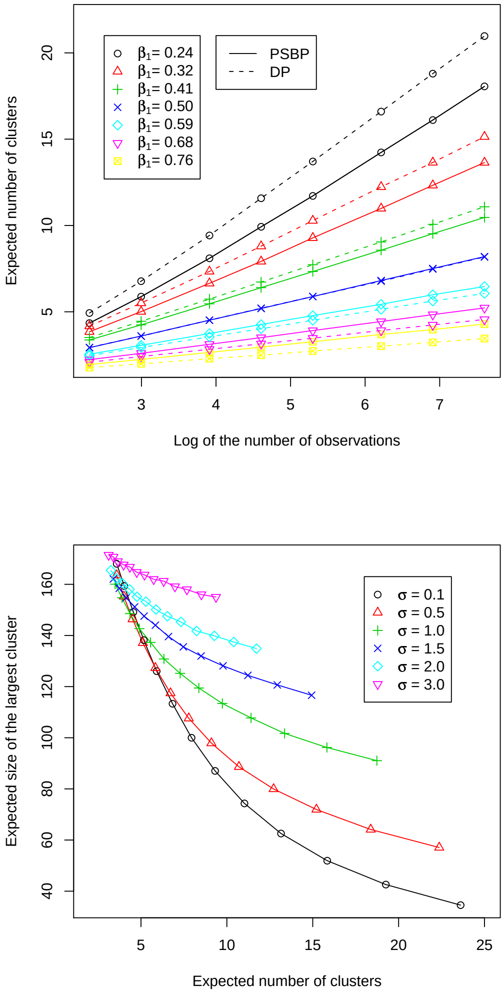

Fig. 2. Clustering structure generated by the probit stick-breaking process. Top panel shows the expected number of clusters under the PSBP with $\sigma = 1$, compared against the Dirichlet process. Note that both models show a logarithmic rate of growth in the number of observations in the sample. However, the probit stick-breaking process grows more slowly than the DP for $\beta_1 > 0.5$ ($\mu < 0$) and faster for $\beta_1 < 0.5$ ($\mu > 0$). Bottom panel shows the expected size of the largest cluster versus the expected number of clusters under different combinations of μ and σ. These plots correspond to samples of $n = 200$ observations.

indeed good approximations. The proof, which can be seen Appendix B, is a straightforward extension of Ishwaran & James (2001) and Ishwaran & James (2003).

THEOREM 1. *Let G^L be a random distribution drawn from a PSBP with L components, baseline measure G_0 and variance parameters μ and σ^2, and let G^∞ denote the case $L = \infty$. In addition, for a sample of size n, $\mathbf{y} = (y_1, \ldots, y_n)$, let*

$$p^L(\mathbf{y}) = \mathsf{E}_{G^L}\left\{ \prod_{i=1}^{n} \int k(y_i|\phi_i) G^L(d\phi_i) \right\}$$

where E_{G^L} denotes the expectation with respect to the law of the random distribution G^L, and $p^\infty(\mathbf{y})$ defined similarly. Then

$$||p^L(\mathbf{y}) - p^\infty(\mathbf{y})|| \leq 4\left(1 - \left\{1 - \left[\Phi\left(-\frac{\mu}{\sqrt{1+\sigma^2}}\right)\right]^{L-1}\right\}^n\right)$$

where $||p^L(\mathbf{y}) - p^\infty(\mathbf{y})||$ denotes the total variation distance between $p^L(\mathbf{y})$ and $p^\infty(\mathbf{y})$.

Note $\lim_{L\to\infty} ||p^L(\mathbf{y}) - p^\infty(\mathbf{y})|| = 0$, and therefore the finite process converges in total variation norm (and therefore in distribution) to the infinite process. As a consequence we have the following Corollary.

COROLLARY 1. *The posterior distribution based on a L-finite PSBP converges in distribution to the one based on the infinite PSBP as $L \to \infty$, as long as the predictives are non-degerate.*

Corollary 1 is especially important for computational purposes. Indeed, it ensures that samples obtained from the posterior distribution of the truncated process can be used to generate arbitrarily accurate inferences on measurable functionals of the infinite process. In practice, the number of atoms does not need to be extremely large. Indeed, note that

$$\left[\Phi\left(-\frac{\mu}{\sqrt{1+\sigma^2}}\right)\right]^{L-1} = \exp\left\{(L-1)\log\left[\Phi\left(-\frac{\mu}{\sqrt{1+\sigma^2}}\right)\right]\right\}$$

Since $\Phi\left(-\frac{\mu}{\sqrt{1+\sigma^2}}\right) < 1$, the rate of decay of the $||p^L(\mathbf{y}) - p^\infty(\mathbf{y})||$ is exponential in L, just like with the Dirichlet process. In Figure 3 we demonstrate the behavior of $||p^L(\mathbf{y}) - p^\infty(\mathbf{y})||$ as a function of μ and L for $n = 1000$ and $\sigma = 1$. Note that for $\beta_1 = 0.26$ (which roughly corresponds to $\eta = 3$ in the Dirichlet process), about 50 atoms are enough for a reasonable approximation, while for $\beta_1 = 17$ (which is roughly equivalent to $\eta = 5$), more than 70 atoms seem to yield no visible additional benefit.

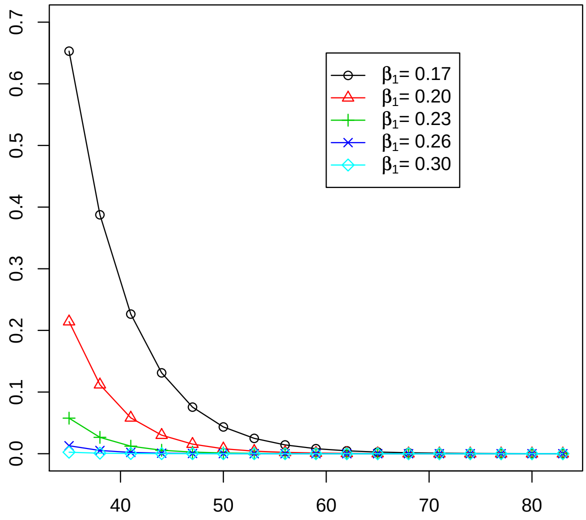

Fig. 3. Bounds on the distance between the infinite stick-breaking process and its corresponding truncation for $\sigma = 1$ and different values of μ. The curves are indexed by $\beta_1 = \Phi(\mu/\sqrt(1+\sigma^2))$.

3. Dependent probit stick-breaking processes for collections of distributions

In order to extend the single-distribution model in Section 2 to a prior on a collection of distributions, we could replace the set of atoms $\{\boldsymbol{\theta}_l\}_{l=1}^L$ and latent random variables $\{\alpha_l\}_{l=1}^L$ with a sequence of independent stochastic processes $\{\boldsymbol{\theta}_l(\mathbf{s}) : \mathbf{s} \in S\}_{l=1}^L$ and $\{\alpha_l(\mathbf{s}) : \mathbf{s} \in S\}_{l=1}^L$ so that,

$$\begin{aligned} y_j(\mathbf{s}) \sim f_{\mathbf{s}} &= \int k(\cdot|\boldsymbol{\phi})G_{\mathbf{s}}(\mathrm{d}\boldsymbol{\phi}) \\ G_{\mathbf{s}}(\cdot) &= \sum_{l=1}^{L} w_l(\mathbf{s})\delta_{\boldsymbol{\theta}_l(\mathbf{s})}(\cdot) \\ w_l(\mathbf{s}) &= \Phi(\alpha_l(\mathbf{s}))\prod_{r<l}(1-\Phi(\alpha_r(\mathbf{s}))). \end{aligned} \qquad (2)$$

$\{\alpha_l(\mathrm{s}) : \mathrm{s} \in S\}_{l=1}^{L}$ has Gaussian margins and $\{\boldsymbol{\theta}_l(\mathrm{s})\}_{l=1}^{L}$ are independent and identically distributed sample paths from a given stochastic process. Models that incorporate dependent weights have a number of theoretical and practical advantages over models that only use dependent atoms like the constant weight models described in DeIorio et al. (2004), Gelfand et al. (2005) and Rodriguez & Ter Horst (2008). On one hand, models with non-constant weights have richer support. Indeed it is well known that constant-weights models cannot generate a set of independent measures (MacEachern, 2000).

This extension of PSBPs to dependent PSBPs parallels the extension of Dirichlet processes to the dependent case by MacEachern (1999, 2000). However, for dependent PSBPs it is much more straightforward to accommodate varying weights without sacrificing computational tractability. In the sequel, we focus our attention on PSBP models with constant atoms where $\boldsymbol{\theta}_l(\mathrm{s}) = \boldsymbol{\theta}_l \sim G_0$ for all $\mathrm{s} \in S$, but adding dependence in the atoms is straightforward. The resulting class $\mathcal{M} = \{G_{\mathbf{s}} : \mathrm{s} \in S\}$ is such that $G_{\mathbf{s}}$ marginally follows a probit stick-breaking process for each $\mathrm{s} \in S$. Therefore for any set $B \in \mathcal{B}$,

$$\begin{aligned}\mathsf{E}(G_{\mathbf{s}}(B)) &= G_0(B)\\ \mathsf{Var}(G_{\mathbf{s}}(B)) &= G_0(B)(1 - G_0(B))\beta_2 \left\{ \frac{1 - (1 - 2\beta_1(\mathrm{s}) + \beta_2(\mathrm{s}))^L}{2\beta_1(\mathrm{s}) - \beta_2(\mathrm{s})} \right\}\end{aligned}$$

One important property of dependent PSBP models with constant atoms is their smoothness. A simple definition of process smoothness for the PSBP can be obtained by considering the distance (in some appropriate topology) between realizations of the process at nearby locations. In particular, for any fixed point $\mathrm{s}_0 \in S$, we can define a stochastic process $\{Z_{\mathbf{s}_0}(\mathrm{s}) : \mathrm{s} \in S\}$ such that

$$Z_{\mathbf{s}_0}(\mathrm{s}) = \int |G_{\mathbf{s}_0}(\mathrm{d}\boldsymbol{\phi}) - G_{\mathbf{s}}(\mathrm{d}\boldsymbol{\phi})| \tag{3}$$

gives the total variation distance between $G_{\mathbf{s}}$ and $G_{\mathbf{s}_0}$ (note that this is indeed a stochastic process since both distributions are random). The stochastic process $Z_{\mathbf{s}_0}(s)$ is said to be almost surely continuous at s if $\lim_{\mathbf{s}' \to \mathbf{s}} Z_{\mathbf{s}_0}(\mathrm{s}') = Z_{\mathbf{s}_0}(\mathrm{s})$. If the process is almost surely continuous for every $\mathrm{s} \in S$, then it is said to have continuous realizations (Banerjee & Gelfand, 2003).

THEOREM 2 (SMOOTHNESS OF DEPENDENT PSBP MODELS). *Let $\{\boldsymbol{\theta}_l\}_{l=1}^{L}$ be an independent and identically distributed sequence from some centering distribution G_0 and $\{\alpha_l(\mathrm{s}) : \mathrm{s} \in S\}_{l=1}^{L}$ be an independent sequence of stochastic processes with continuous realizations and Gaussian marginals, both defining a dependent PSBP with constant atoms. Also, let $Z_{\mathbf{s}_0}(\mathrm{s})$ be as defined in* (3). *For any* $\mathrm{s}_0 \in S$,

(a) $Z_{\mathbf{s}_0}(\mathrm{s})$ *also has continuous realizations almost surely.*
(b) $\lim_{\mathbf{s} \to \mathbf{s}_0} Z_{\mathbf{s}_0}(\mathrm{s}) = 0 = Z_{\mathbf{s}_0}(\mathrm{s}_0)$ *almost surely.*

The proof of Theorem (2) can be seen in Appendix C. Conditions for the almost sure continuity of the random processes $\{\alpha_l(\mathrm{s}) : \mathrm{s} \in S\}_{l=1}^{L}$ are given in Kent (1989). It is worth emphasizing that continuity in these results does not refer to the draws from the random distribution $G_{\mathbf{s}}$ (which is almost surely discrete), but to the similarity between realizations at nearby locations.

The covariance structure generated by the model is another important property to understand. For any Borel set B we have that, a priori,

$$\mathsf{Cov}(G_{\mathbf{s}}(B), G_{\mathbf{s}'}(B)) = \frac{\beta_2(\mathrm{s},\mathrm{s}')\left\{1 - [1 - \beta_1(\mathrm{s}) - \beta_1(\mathrm{s}') + \beta_2(\mathrm{s},\mathrm{s}')]^L\right\}}{\beta_1(\mathrm{s}) + \beta_1(\mathrm{s}') - \beta_2(\mathrm{s},\mathrm{s}')} \times G_0(B)\{1 - G_0(B)\}$$

where the expression for $\beta(\mathrm{s},\mathrm{s}')$ can be seen in Appendix D. Therefore, the process for the distributions will typically be nonstarionary. In the following subsections, we discuss some examples of dependent PSBPs.

3.1. Dependent PSBPs with latent Gaussian processes

A particularly interesting example of a dependent PSBP arises by letting $\alpha_l(\mathrm{s})$ in equation (2) be a Gaussian process over S with mean μ and covariance function $\sigma^2\gamma(\mathrm{s},\mathrm{s}')$, for $\mathrm{s} \in S \subset \mathbb{R}^q$. More concretely, given observations associated with locations (or predictors) $\mathrm{s}_1, \ldots, \mathrm{s}_n$, the joint distribution for the realizations of the latent processes $\alpha_l(\mathrm{s})$ at these locations is given by

$$\begin{pmatrix} \alpha_l(\mathrm{s}_1) \\ \alpha_l(\mathrm{s}_2) \\ \vdots \\ \alpha_l(\mathrm{s}_n) \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \gamma(\mathrm{s}_1,\mathrm{s}_2) & \ldots & \gamma(\mathrm{s}_1,\mathrm{s}_n) \\ \gamma(\mathrm{s}_2,\mathrm{s}_1) & 1 & \ldots & \gamma(\mathrm{s}_2,\mathrm{s}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(\mathrm{s}_n,\mathrm{s}_1) & \gamma(\mathrm{s}_n,\mathrm{s}_2) & \ldots & 1 \end{pmatrix} \right)$$

Letting $\gamma(\mathrm{s},\mathrm{s}') \to 0$ for all s and s′ leads to independent estimates at each location, while $\gamma(\mathrm{s},\mathrm{s}') \to 1$ leads to a common nonparametric prior for all locations. Models of this type can be used for time series observed in continuous time ($S = \mathbb{R}^+$), or to construct models for spatial data ($S \subset \mathbb{R}^2$). In particular, this construction allows us to easily generate spatial processes for discrete and non-Gaussian distributions. Even more, we can introduce multivariate atoms, leading to a simple procedure to construct non-stationary, non-separable multivariate spatial-temporal processes. By interpreting S as a space of predictors, this construction also allows us to generate flexible nonparametric regression models with heteroscedastic errors, as discussed in Griffin & Steel (2006b).

A priori, the covariance of the process under (2) is

$$\mathsf{Cov}(y(\mathrm{s}), y(\mathrm{s}')) = \frac{\beta_2(\mathrm{s},\mathrm{s}')\left\{1 - [1 - \beta_1(\mathrm{s}) - \beta_1(\mathrm{s}') + \beta_2(\mathrm{s},\mathrm{s}')]^L\right\}}{\beta_1(\mathrm{s}) + \beta_1(\mathrm{s}') - \beta_2(\mathrm{s},\mathrm{s}')} \mathsf{E}_{G_0}(\mathsf{Var}(\mathbf{y}|\theta)),$$

which means that a priori the process is nonstationary, even if the underlying processes driving the atoms are stationary. Indeed, this shows one of the limitations induced by using models with constant atoms, as in this case it is impossible to center the nonparametric process on a stationary process (except for the trivial, random noise model).

A posteriori, the covariance of the process can be computed by conditioning on the values of the stick-breaking weights and atoms:

$$\begin{aligned}\mathsf{Cov}(y(\mathbf{s}), y(\mathbf{s}') | \{w_l(\mathbf{s})\}_{l=1}^{L}, \{\boldsymbol{\theta}_l\}_{l=1}^{L}) &= \sum_{l=1}^{L}\sum_{r=1}^{L} w_l(\mathbf{s}) w_r(\mathbf{s}') \mathsf{E}(\mathbf{y}|\boldsymbol{\theta}_l)\mathsf{E}(\mathbf{y}|\boldsymbol{\theta}_k) \\ &\quad - \left(\sum_{l=1}^{L} w_l(\mathbf{s})\mathsf{E}(\mathbf{y}|\boldsymbol{\theta}_l)\right)\left(\sum_{l=1}^{L} w_l(\mathbf{s}')\mathsf{E}(\mathbf{y}|\boldsymbol{\theta}_l)\right)\end{aligned}$$

Other functionals of interest can be calculated in a similar fashion. To predict the distributions at a new location $\mathbf{s}_{n+1}$, we can interpolate the latent field $\alpha(\mathbf{s})$ to obtain $\alpha(\mathbf{s}_{n+1})$ and compute $\{w_l(\mathbf{s}_{n+1})\}_{l=1}^{L}$.

3.2. Dependent PSBPs with latent Markov random fields

Consider now a model for distributions that evolve in discrete time, as in Griffin & Steel (2006b); Caron et al. (2007, 2008) and Griffin (2008). For $t = 1, \ldots, T$ let

$$\begin{aligned} y_t &\sim \int k(\cdot|\boldsymbol{\phi}) G_t(\mathrm{d}\boldsymbol{\phi}) & G_t(\cdot) &= \sum_{l=1}^{L} w_{lt}\delta_{\boldsymbol{\theta}_l}(\cdot) \\ w_{lt} &= \Phi(\alpha_{lt})\prod_{r<l}(1-\Phi(\alpha_{rt})) & \alpha_{lt} &= \mathbf{A}_t'\boldsymbol{\eta}_{lt} \end{aligned}$$

We can induce dependence in the weights through a general autoregressive process of the form

$$\boldsymbol{\eta}_{lt}|\boldsymbol{\eta}_{l,t-1} \sim \mathsf{N}(\mathbf{B}_t\boldsymbol{\eta}_{l,t-1}, \mathbf{W}_t)$$

By appropriately choosing the structural parameters $\mathbf{A}_t$, $\mathbf{B}_t$ and $\mathbf{W}_t$ a number of different evolution patterns can be accommodated. For example, letting $\mathbf{A}_t$ be a $p \times 1$ vector and $\mathbf{B}_t$ be a $p \times p$ matrix such that

$$\mathbf{A}_t = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad \mathbf{B}_t = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

we induce a form-free periodic model for densities (see West & Harrison (1997), Chapter 7), where G_{t+p} is centered around G_t, in the sense that $\mathsf{E}(\alpha_{l,t+p}|\alpha_{l,t+p-1},\ldots,\alpha_{lt}) = \alpha_{lt}$. In particular, setting $\mathbf{W}_t = \mathbf{0}$ leads to a common G_t for all time points. Other patterns like trends, periodicities and autoregressive processes can be similarly modeled, providing additional flexibility over other nonparametric models. Also, this approach can be generalized to construct spatial (or spatial-temporal) models for aerial data by considering a two (or three) dimensional Gaussian random Markov field, in the spirit of Figueiredo (2005) and Figueiredo et al. (2007).

3.3. Random effect models for distributions

Finally, consider a situation where multiple observations are obtained for each one of I populations, and our goal is to borrow information nonparametrically across them while assuming exchangeability both between and within populations. Specifically, for $j = 1,\ldots,n_i$ and $i = 1,\ldots,I$ assume that data y_{ij} corresponds to the j-th observation from the i-th population. In a parametric setting, a natural model for this situation is a random effects model. For the nonparametric case, assume that for some parametric kernel $k(\cdot|\boldsymbol{\phi})$,

$$y_{ij} \sim \int k(\cdot|\boldsymbol{\phi}) G_i(\mathrm{d}\boldsymbol{\phi}) \qquad\qquad G_i(\cdot) = \sum_{l=1}^{\infty} w_{il}\delta_{\theta_l}(\cdot)$$

where

$$w_{il} = \Phi(\alpha_{il}) \prod_{r<l} (1 - \Phi(\alpha_{ir})) \qquad \alpha_{il} \sim \mathsf{N}(\alpha_l^0, 1) \qquad \alpha_l^0 \sim \mathsf{N}(\mu, 1)$$

The common prior for $\{\alpha_{il}\}_{i=1}^{I}$ allows us to borrow information across populations by shrinking the stick-breaking ratios corresponding to the l-th mixture component towards a common value. This formulation is reminiscent of the hierarchical Dirichlet process (HDP) (Teh et al., 2006); indeed this random effect model for distributions can be used as an alternative to the HDP. However, unlike the HDP, a generalization that includes covariates is straightforward by letting

$$\alpha_{il} = \mathbf{x}_{ij}'\boldsymbol{\eta}_{il} \qquad \boldsymbol{\eta}_{il} \sim \mathsf{N}(\boldsymbol{\eta}_l^0, 1) \qquad \boldsymbol{\eta}_l^0 \sim \mathsf{N}(\mu, 1)$$

where $\mathbf{x}_{ij}$ is the vector of covariates specific to observation y_{ij}.

4. Posterior sampling

In this section we demonstrate that a collapsed Markov chain Monte Carlo (MCMC) sampler (Robert & Casella, 1999; Ishwaran & James, 2001) can be constructed to fit the dependent PSBPs mixtures described in 2. Our algorithms borrow on ideas previously used to fit Bayesian continuation-ratio probit models in survival analysis (Albert & Chib, 2001).

First, we concentrate on the case $L < \infty$. For each observation $y_j(\mathbf{s}_i)$, corresponding to replicate j under condition/location i, $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$, introduce an indicator variables $\xi_j(\mathbf{s}_i)$ such that $\xi_j(\mathbf{s}_i) = l$ if and only if observation $y_j(\mathbf{s}_i)$ is sampled from mixture component l. The use of these latent variables is standard in mixture models; conditional on the indicators the full conditional distribution of the component-specific parameters for a model with constant atoms is given by

$$p(\boldsymbol{\theta}_l | \cdots) \propto G_0(\mathrm{d}\boldsymbol{\theta}_l) \prod_{\{(i,j) | \xi_j(\mathbf{s}_i) = l\}} k(y_j(\mathbf{s}_i) | \boldsymbol{\theta}_l).$$

If the centering measure G_0 is conjugate to the kernel $k(\cdot|\boldsymbol{\theta})$, sampling from this distribution is straightforward. In non-conjugate settings, sampling can still be carried out using a Metropolis-Hastings step. Similarly, if the atoms are not constant then the observations on each component correspond to draws from a single stochastic process and sampling of its parameters can be carried out using standard simulation algorithms.

Conditional on the component specific parameters and the realized values of the weights $\{w_l(\mathbf{s}_1)\}_{l=1}^{L}, \ldots, \{w_l(\mathbf{s}_n)\}_{l=1}^{L}$ at the observed locations, the full conditional distribution for the indicators is multinomial with probabilities given by

$$\mathsf{Pr}(\xi_j(\mathbf{s}_i) = l | \cdots) \propto w_l(\mathbf{s}_i) k(y_j(\mathbf{s}_i) | \boldsymbol{\theta}_l).$$

In order to sample the value of the latent processes $\{\alpha_l(\mathbf{s}_1)\}_{l=1}^{L}, \ldots, \{\alpha_l(\mathbf{s}_n)\}_{l=1}^{L}$ and the corresponding weights $\{w_l(\mathbf{s}_1)\}_{l=1}^{L}, \ldots, \{w_l(\mathbf{s}_n)\}_{l=1}^{L}$, for each $i = 1, \ldots, n$ and $l = 1, \ldots, L-1$ we introduce a collection of conditionally independent latent variables $z_{jl}(\mathbf{s}_i) \sim \mathsf{N}(\alpha_l(\mathbf{s}_i), 1)$. If we define $\xi_j(\mathbf{s}_i) = l$ if and only if $z_{jl}(\mathbf{s}_i) > 0$ and $z_{jr}(\mathbf{s}_i) < 0$ for $r < l$, we have

$$\begin{aligned} \mathsf{Pr}(\xi_j(\mathbf{s}_i) = l) &= \mathsf{Pr}(z_{jl}(\mathbf{s}_i) > 0, z_{jk}(\mathbf{s}_i) < 0 \text{ for } r < l) \\ &= \Phi(\alpha_l(\mathbf{s}_i)) \prod_{r<l} (1 - \Phi(\alpha_r(\mathbf{s}_i))) = w_l(\mathbf{s}_i). \end{aligned} \tag{4}$$

independently for each i. This data augmentation scheme simplifies computation as it allows us to implement another Gibbs sampling scheme. Indeed, conditionally on the value of the latent process and the indicator variables, we can impute the augmented variables by sampling from its full conditional distribution,

$$z_{jl}(\mathbf{s}_i) | \cdots \sim \begin{cases} \mathsf{N}(\alpha_l(\mathbf{s}_i), 1) \mathbf{1}_{\mathbb{R}^-} & l < \xi_j(\mathbf{s}_i) \\ \mathsf{N}(\alpha_l(\mathbf{s}_i), 1) \mathbf{1}_{\mathbb{R}^+} & l = \xi_j(\mathbf{s}_i) \end{cases},$$

where $\mathsf{N}(\mu, \tau^2) \mathbf{1}_{\Omega}$ denotes the normal distribution with mean μ and variance τ^2 truncated to the set Ω.

In turn, conditional on the augmented variables, the latent processes can be sampled by taking advantage of the normal priors for $\alpha_l(\mathbf{s})$. The details of this step are specific to the problem being considered; for example, for the spatial model described in Section 3.1 observed without replicates, we have

$$(\alpha_l(\mathbf{s}_1), \ldots, \alpha_l(\mathbf{s}_n))' \sim \mathsf{N}\left(\left[\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma^2}\mathbf{I} \right]^{-1} \left[\mu\boldsymbol{\Sigma}^{-1}\mathbf{1} + \frac{1}{\sigma^2}\mathbf{z}_l \right], \left[\boldsymbol{\Sigma}^{-1} + \frac{1}{\sigma^2}\mathbf{I} \right]^{-1} \right)$$

where $\mathbf{z}_l = (z_l(\mathbf{s}_1), \ldots, z_l(\mathbf{s}_n))'$ and $\mathbf{1}$ is a column vector of ones. Similarly, for the models in Section 3.2, the corresponding augmented models on the latent variables $z_{it} \sim \mathsf{N}(\alpha_{it}, 1)$ results in a series of dynamic linear models (West & Harrison, 1997) and a Forward-Backward algorithm (Carter & Kohn, 1994; Fruhwirth-Schnatter, 1994) can be used to efficiently sample the latent process.

Once the latent processes have been updated, the weights can be computed using (4). If unknown parameters remain in the specification of the latent process (spatial correlations, evolution variances, the mean μ of the latent process), these can typically be sampled conditionally of the value of the imputed latent processes.

In the case $L = \infty$, we can easily extend this algorithm to generate a slice sampler, as discussed in Walker (2007) and Papaspiliopoulos (2009). Alternatively, the results at the end of Section 2 suggest that a finite PSBP with a large number of components (≈ 50, depending on the value of μ) can be used instead (Ishwaran & James, 2001; Ishwaran & Zarepour, 2002).

5. Illustrations

This section presents two applications of the PSBP model. We first discuss an application of the discrete time PSBP introduced in Section 3.2, and then we move to an application of the spatial nonparametric models described in Section 3.1. All computations were carried out using the algorithm for a finite PSBP with $L = 50$ components. In both cases, inferences are based on 100,000 samples of the MCMC obtained after a burn-in period of 10,000 iterations. No evidence of lack of convergence was found from the visual inspection of trace plots or the application of the Gelman-Rubin convergence test (Gelman & Rubin, 1992).

5.1. Multivariate stochastic volatility

In this section we use the PSBP model to construct a multivariate stochastic volatility model that allows for the joint distribution of returns across multiple assets to evolve in time. Therefore, the model accommodates not only time varying volatilities for the assets, but also time varying means and correlations, providing added flexibility to traditional stochastic volatility models. The data set under consideration consists of the weekly returns of the S&P500 (US stock market) and FTSE100 (UK stock market) indexes covering the ten-year period between

February 02, 1999 and April 02, 2009, for a total of 522 observations. Figure 4 shows the evolution of these returns in time (on which the current financial crises can be clearly seen in the increased volatility of returns), while Figure 5 shows a scatterplot of the returns generated by both indexes (which reveals that there is a strong correlation among the assets).

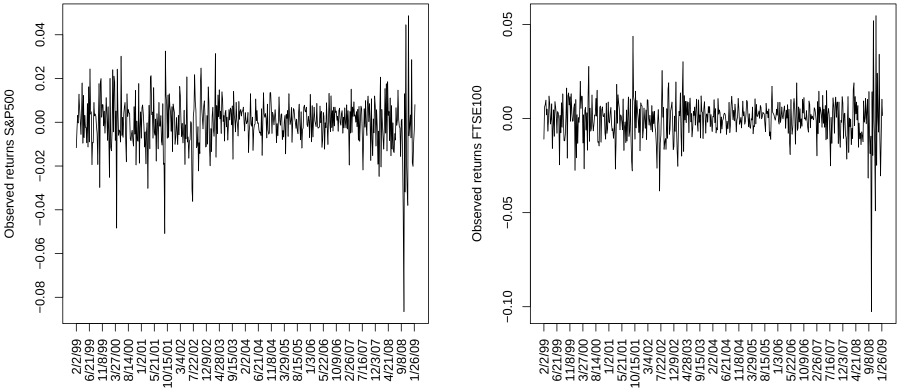

Fig. 4. Time series plots for the weekly returns on the S&P500 and FTSE indexes between February 02, 1999 and February 02, 2009. Different volatility levels are readily visible in the plots

We model $\mathbf{r}_t = (r_t^1, r_t^2)'$, the joint log-return of the S&P500 and the FTSE100 at time t, as following a normal distribution with time varying mean vector $\boldsymbol{\mu}_t$ and covariance matrix $\boldsymbol{\Sigma}_t$. In turn, we let $(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \sim G_t$, where G_t evolves in time according to a discrete time PSBP (see Section 3.2) with a normal-inverse Wishart centering measure. Based on historical information, the parameters of the centering measure were chosen so that the mean of the returns is 0, the annualized volatility is centered around 12%, and the expected correlation is 0.75. For the latent structure we assume $A_t = 1$, $B_t = 1$ and $W_t = U$, leading to a specification reminiscent of a random walk mixing distribution, where G_{t+1} is a priori "centered" around G_t. We used a Gamma hyperprior for $1/U$ with two degrees of freedom and mean 1, and a standard normal prior for η_0.

Figure 6 shows the estimated volatilities for both assets under the PSBP model (solid lines). These can be contrasted against estimates obtained from the Bayesian stochastic volatility model described in Jacquier et al. (1994) and Kim et al. (1998) (dashed lines). Both sets of estimates have very similar features, however, the series estimated using the PSBP are smoother, presenting less short term fluctuations. This is most likely due to the heavy tails

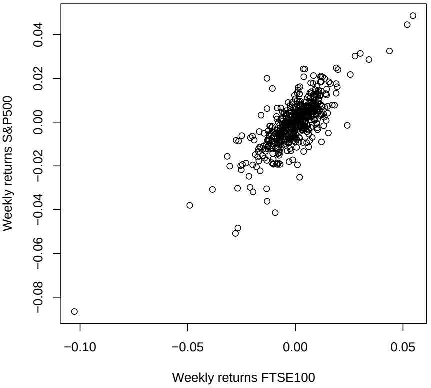

Fig. 5. Scatter plot for weekly returns on the S&P500 and FTSE indexes between February 02, 1999 and February 02, 2009. The plot clearly suggests that the returns of both indexes are strongly correlated.

in the conditional distributions induced by the use of a mixture likelihood, which allow us to explain outliers without the need to increase the volatility.

To help emphasize the additional flexibility afforded by the model, we present in Figures 7 and 8 the estimated correlation across assets and the estimated expected (annualized) returns. Many multivariate stochastic volatility models available in the literature assume that the correlation across assets is constant, and most of them also assume a constant mean for the returns. The results from the PSBP suggest that these assumptions might not be supported by the data. First, note the negative association between correlation and expected returns: the correlation among the two assets tends to decrease when the returns are high, and to increase when returns decrease. This leverage effect has been blamed for the failure of traditional pricing models in the aftermath of the current financial crises. Additionally, note that the historical mean of returns for both the S&P500 and FTSE100 indexes turns out to be negative over this period. This result is highly influenced by the dismal returns realized during the last 18 months. However, it is hardly believable that the expected returns on assets is a negative constant. Instead, a time varying pattern such as the one depicted in Figure 8 is more reasonable, as it nicely correlates with the business cycle.

One possible concern with our hierarchical specification is whether there is enough information in the observations to identify the precision parameter η_0 or the parameters controlling the latent process. This does not seem to be an issue in our case, as the poste-

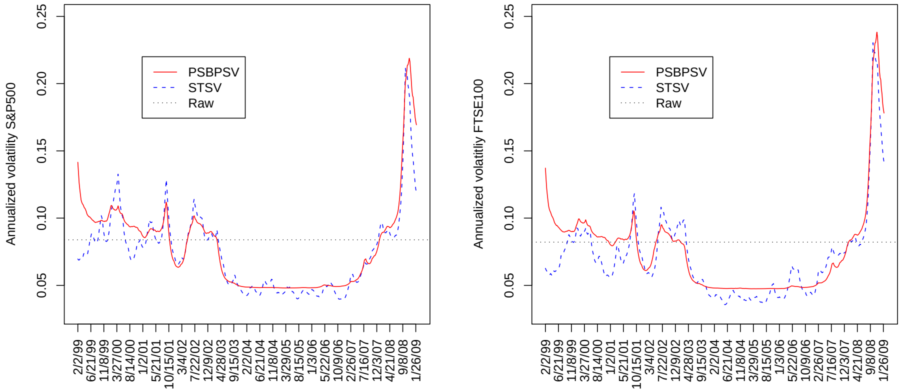

Fig. 6. Estimated volatilities for the S&P500 and the FTSE100 indexes under the PSBP model. We compare our estimates with those obtained from the stochastic volatility model of Jacquier et al. (1994) and Kim et al. (1998) (labeled STSV). The horizontal line corresponds to a standard deviation of returns over the 10-year period. The spikes in volatility at the end of the series correspond to the current financial crises and are clearly apparent in both indexes.

rior distributions for the parameters U and η_0 show substantial learning from the data. The posterior distribution for U has a posterior mean of 0.1189 (symmetric 95% credible band, $(0.0744, 0.1848)$), while the mass parameter η_0 has a posterior mean of 0.0322 (95% credible interval, $(-0.0842, 0.1446)$), both substantially different from their prior distributions.

5.2. *Spatial processes for count data*

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`. Here, we focus on modeling the abundance of *Zenaida macroura*, commonly known as the Mourning Dove, in North Carolina during the 2006-2007 winter season. We use information from 27 parties (see Figure 9); since the diameter of the circles is very small compared to the

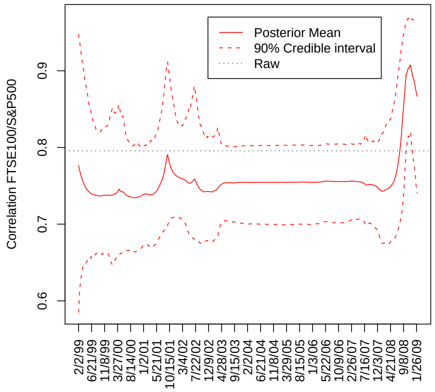

Fig. 7. Estimated correlation between the S&P500 and the FTSE100 indexes under the PSBP model. The horizontal line corresponds to a simple correlation of returns over the 10-year period. In line with recent discussions in the literature, the model demonstrates that correlation among assets tends to increase in times of financial distress.

size of the region under study, we treat the data as point referenced to the center of the circle.

Specifically, we let $y(\mathbf{s})$ stand for the number of birds observed at location $\mathbf{s}$ (expressed in latitude and longitude) and assume that $y(\mathbf{s}) \sim \mathsf{Poi}(h(\mathbf{s})\lambda(\mathbf{s}))$, where $h(\mathbf{s})$ represents the number of man-hours invested at location $\mathbf{s}$. Next, we assume that $\lambda(\mathbf{s}) \sim G_{\mathbf{s}}$, where $G_{\mathbf{s}}$ follows a spatial PSBP driven by underlying Gaussian processes with mean α, exponential covariance function and common variance and correlation parameters σ^2 and ρ (see Section 3.1). Based on historical data from previous CBC censuses we assume an exponential centering measure G_0 with mean 0.15 sighting/man-hour. Priors for the parameters of the latent Gaussian processes σ^2 and ρ are also taken to be exponential with unit mean, while the prior for α is set to a standard normal distribution.

Figure 9 shows the mean of the predictive distribution for the number of sightings per man-hour over a 90×30 grid overlaid on the North Carolina map. The map shows a higher expected number of sighting in the northern area of the coastal plain region, with a lower expected number of sighting in the Piedmont plateau. This is a very reasonable result as it is well known that the Mourning Dove favors open and semi-open habitats, such as farms, prairie, grassland, and lightly wooded areas, while avoiding swamps and thick forest (Kaufman, 1996, page 293).

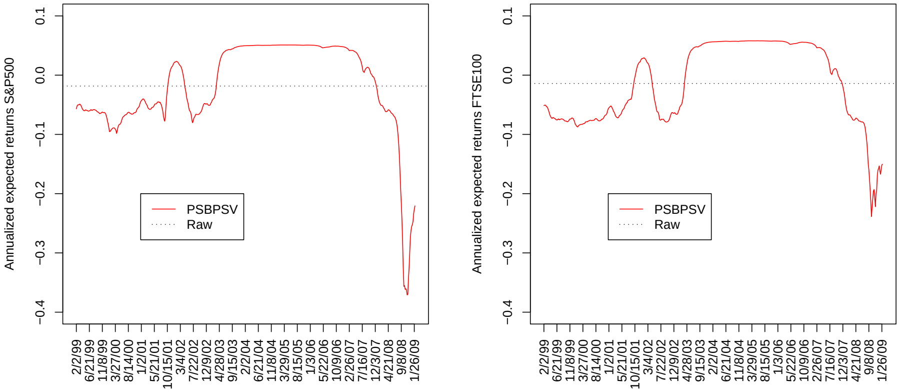

Fig. 8. Annualized expected returns for the S&P500 and the FTSE100 indexes under the PSBP model. The horizontal line corresponds to a simple average of returns over the 10-year period.

In Figure 10 we present density estimates for 4 locations in North Carolina; three of them correspond to places where parties were active (Greenville, Wayne County and Greensboro), while the fourth correspond to a location in the Blue Ridge mountain near Waynesville where no data was observed (for an out-of-sample prediction exercise). The resulting distributions tend to have heavy tails and might present multimodality, as in the case of Greenville. This is in contrast to the predictive distributions that would be obtained from a Poisson generalized linear model with a logarithmic link and spatial random effects that follow a Gaussian process, which would be unimodal.

As before, there is substantial learning in the structural parameters of the model. The posterior mean of the correlation parameter ρ is 0.179 (95% credible band $(0.033, 0.717)$), for σ^2 it is 1.203 (95% credible band $(0.574, 2.032)$) and for μ it is -0.274 (95% credible band $(-0.451, 0.133)$).

6. Discussion

One of the main advantages of the PSBP formulation is their generality and flexibility; the probit formulation allow us to extend all the traditional Bayesian models based on hierarchical linear models to generate equivalent models for distributions with little additional cost. In addition to any possible combination of the models discussed in this paper, we can easily

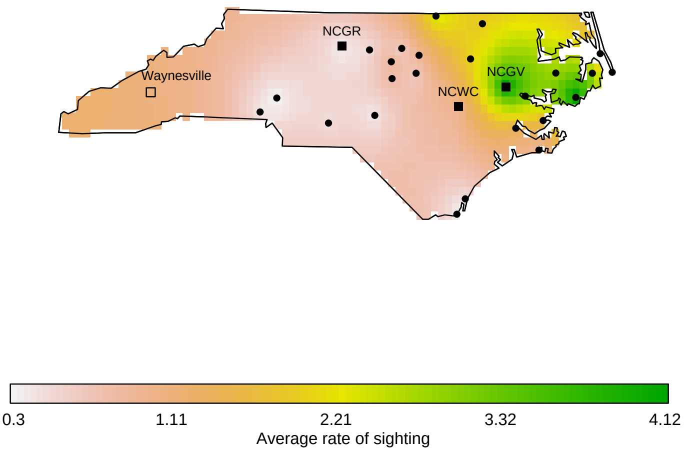

Fig. 9. Estimated expected rate of sightings (per man-hour) for the Mourning Dove. Filled dots correspond to the 27 locations where observations were collected. Squared dots represent locations where density estimation is carried out, filled squares represent locations for in-sample-predictions, while the empty square corresponds to a point of out-of-sample prediction.

create ANOVA, mixed effects and clustering procedures, among others.

Our discussion in this paper has focused on models where the mean of the latent stochastic processes driving the weights is common to all components. As we discussed in Section 2, this implies that the number of components grows roughly at a logarithmic rate as the number of observations increase. Simulation experiments demonstrate that faster growth rates can be obtained by having the mean of α_l decrease linearly (or more generally, polynomially) with l. This is reminiscent of the behavior of Poisson-Dirichlet processes, where growth rates follow a power law for discount parameters greater than 0.

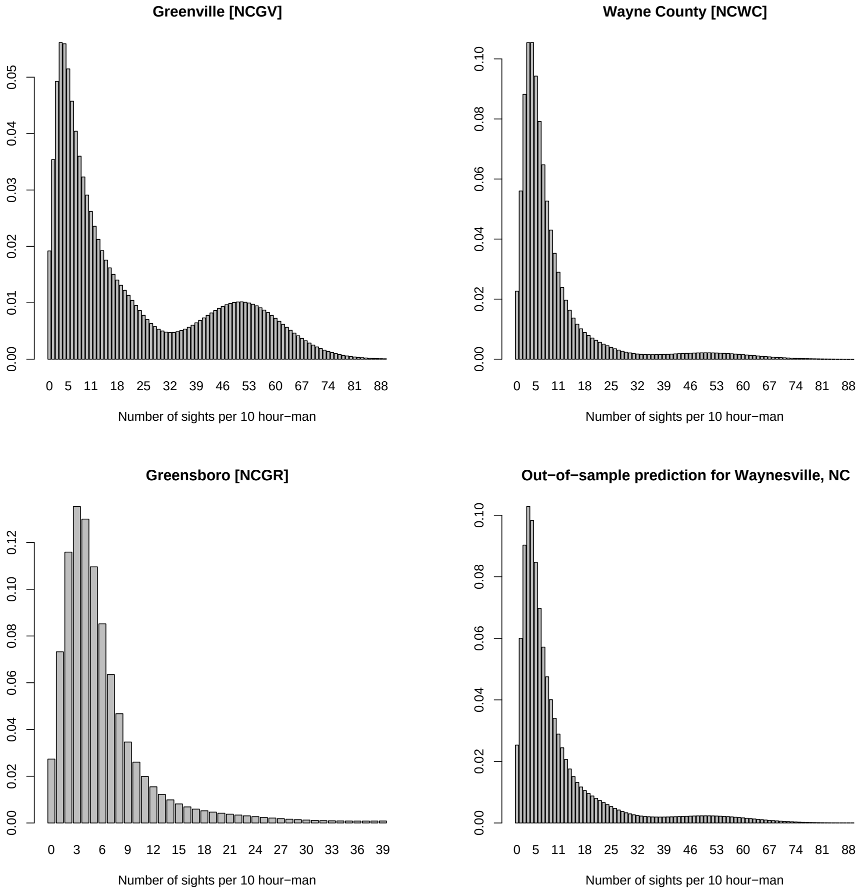

Fig. 10. Density estimates for four NC locations. The top two panels and the lower left panel correspond to three locations where observations were collected (in-sample predictions), while the bottom right panel corresponds to an out-of-sample prediction for a location in the Blue Ridge mountains next to Waynesville, NC.

A. Properties of the weights of the PSBP

Let $u_l = \Phi(z_l)$ where $z_l \sim \mathsf{N}(\mu, \sigma^2)$. The expectation of $\beta_1 = \mathsf{E}(u_l)$ can be easily computed using a change of variables,

$$\begin{aligned}\beta_1 = \mathsf{E}(u_l) &= \int_{-\infty}^{\infty} \Phi(z_l) \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \frac{(z_l - \mu)^2}{\sigma^2} \right\} dz_l \\ &= \int_S \frac{1}{2\pi\sigma} \exp\left\{ -\frac{1}{2} \left[x^2 + \frac{(z_l - \mu)^2}{\sigma^2} \right] \right\} dx dz_l\end{aligned}$$

where $S = \{(x, z_l) : -\infty < z_l < \infty,\ \infty < x < z_l\}$. Applying the change of variables $t_1 = z_l - x$ and $t_2 = z_l$ we get

$$\begin{aligned}\mathsf{E}(u_l) &= \int_0^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma} \exp\left\{ -\frac{1}{2} \left[(t_2 - t_1)^2 + \frac{(t_2 - \mu)^2}{\sigma^2} \right] \right\} dt_2 dt_1 \\ &= \int_0^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1+\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(t_1 - \mu)^2}{1+\sigma^2} \right\} dt_1 = 1 - \Phi\left(-\frac{\mu}{\sqrt{1+\sigma^2}} \right) \\ &= \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}} \right) = \mathsf{Pr}(T_1 > 0)\end{aligned}$$

where $T_1 \sim \mathsf{N}(\mu, 1 + \sigma^2)$. Now, using Jensen's inequality,

$$\begin{aligned}\mathsf{E}\left[\log(1 - u_l)\right] &\leq \log\left[1 - \mathsf{E}(u_l)\right] \\ &= \log(1 - \Phi(\mu/\sqrt{1+\sigma^2})) < 0\end{aligned}$$

Therefore, $\sum_{l=1}^{\infty} \mathsf{E}\left[\log(1 - u_l)\right] = -\infty$ and, by theorem 2 in Ishwaran & James (2001), $\sum_{l=1}^{\infty} w_l = 1$ almost surely. A similar calculation can be used to compute the second central moment of u_l,

$$\begin{aligned}\beta_2 = \mathsf{E}(u_l^2) &= \int_{-\infty}^{\infty} \left[\Phi(z_l)\right]^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \frac{(z_l - \mu)^2}{\sigma^2} \right\} dz_l \\ &= \int_S \left(\frac{1}{2\pi} \right)^{3/2} \frac{1}{\sigma} \exp\left\{ -\frac{1}{2} \left[x^2 + y^2 + \frac{(z_l - \mu)^2}{\sigma^2} \right] \right\} dx dy dz_l\end{aligned}$$

where now $S = \{(x, y, z_l) : -\infty < z_l < \infty,\ \infty < x < z_l,\ \infty < y < z_l\}$. Using the change of variables $t_1 = z_l - x$ and $t_2 = z_l - y$ and $t_3 = z_l$ we get

$$\begin{aligned}\mathsf{E}(u_l^2) = &\int_0^{\infty} \int_0^{\infty} \int_{-\infty}^{\infty} \left(\frac{1}{2\pi} \right)^{3/2} \frac{1}{\sigma} \times \\ &\qquad \exp\left\{ -\frac{1}{2} \left[(t_3 - t_1)^2 + (t_3 - t_2)^2 + \frac{(t_3 - \mu)^2}{\sigma^2} \right] \right\} dt_3 dt_1 dt_2 \\ = &\ \mathsf{Pr}(T_1 > 0, T_2 > 0)\end{aligned}$$

where $\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim \mathsf{N}\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix} ; \begin{pmatrix} 1+\sigma^2 & \sigma^2 \\ \sigma^2 & 1+\sigma^2 \end{pmatrix} \right)$.

The argument can be directly extended to higher order moments. In general, the p-th moment can be obtained from the cumulative distribution function of a p-variate normal distribution,

$$\beta_p = \mathsf{E}(u_l^p) = \mathsf{Pr}(T_1 > 0, \ldots, T_p > 0)$$

where $\mathbf{T} = (T_1, \ldots, T_p)'$ follows a multivariate normal distribution with $\mathsf{E}(T_i) = \mu$, $\mathsf{Var}(T_i) = 1 + \sigma^2$ and $\mathsf{Cov}(T_i, T_j) = \sigma^2$.

B. Truncations of PSBP models

Note that Theorem 2 in Ishwaran & James (2001) applies directly. Therefore,

$$||p^L(\mathbf{y}) - p^{\infty}(\mathbf{y})|| \leq 4 \left[1 - \mathsf{E}\left\{ \left(\sum_{s=1}^{L-1} w_s \right)^n \right\} \right].$$

Now, by Jensen's inequality and the results in Appendix A

$$\begin{aligned} \left[1 - \mathsf{E}\left\{ \left(\sum_{s=1}^{L-1} w_s \right)^n \right\} \right] &\leq \left[1 - \left(\sum_{s=1}^{L-1} \mathsf{E}(w_s) \right)^n \right] \\ &= \left(1 - \left\{ 1 - \left[\Phi\left(-\frac{\mu}{\sqrt{1+\sigma^2}} \right) \right]^{L-1} \right\}^n \right) \end{aligned}$$

C. Proof of Theorem 2

Consider first the case when L is finite. Note that, if the collection $\{\alpha_l(\mathbf{s})\}_{l=1}^{L}$ has continuous realizations, so does $\{w_l(\mathbf{s})\}_{l=1}^{L}$ because it is a continuous transformation of the latent processes. Now,

$$\begin{aligned} |Z_{\mathbf{s}_0}(\mathbf{s}) - Z_{\mathbf{s}_0}(\mathbf{s}')| &= \left| \int |G_{\mathbf{s}_0}(\mathrm{d}\boldsymbol{\phi}) - G_{\mathbf{s}}(\mathrm{d}\boldsymbol{\phi})| - \int |G_{\mathbf{s}_0}(\mathrm{d}\boldsymbol{\phi}) - G_{\mathbf{s}'}(\mathrm{d}\boldsymbol{\phi})| \right| \\ &\leq \int |G_{\mathbf{s}}(\mathrm{d}\boldsymbol{\phi}) - G_{\mathbf{s}'}(\mathrm{d}\boldsymbol{\phi})| \\ &= 2 \sup_{B \in \mathcal{B}} |G_{\mathbf{s}}(B) - G_{\mathbf{s}'}(B)| \\ &\leq 2 \sup_{B \in \mathcal{B}} \sum_{l=1}^{L} |w_l(\mathbf{s}) - w_l(\mathbf{s}')| \, \delta_{\boldsymbol{\theta}_l}(B) \end{aligned}$$

Due to the almost sure continuity of $\{w_l(\mathrm{s})\}_{l=1}^{L}$, for any $\epsilon > 0$ there exists a Δ such that $|\mathrm{s} - \mathrm{s}'| < \Delta$ implies $|w_l(\mathrm{s}) - w_l(\mathrm{s}')| < \epsilon/(2L)$, which in turn implies that

$$\begin{aligned}|Z_{\mathbf{s}_0}(\mathrm{s}) - Z_{\mathbf{s}_0}(\mathrm{s}')| &\leq 2 \sup_{B \in \mathcal{B}} \sum_{l=1}^{L} |w_l(\mathrm{s}) - w_l(\mathrm{s}')| \, \delta_{\boldsymbol{\theta}_l}(B) \\ &\leq 2 \sum_{l=1}^{L} |w_l(\mathrm{s}) - w_l(\mathrm{s}')| < \epsilon\end{aligned}$$

For the case $L = \infty$, write

$$\sum_{l=1}^{\infty} |w_l(\mathrm{s}) - w_l(\mathrm{s}')| = \sum_{l=1}^{L} |w_l(\mathrm{s}) - w_l(\mathrm{s}')| + \sum_{l=L+1}^{\infty} |w_l(\mathrm{s}) - w_l(\mathrm{s}')|$$

and note that

$$\sum_{l=L+1}^{\infty} |w_l(\mathrm{s}) - w_l(\mathrm{s}')| \leq \sum_{l=L+1}^{\infty} w_l(\mathrm{s}) + \sum_{l=L+1}^{\infty} w_l(\mathrm{s}')$$

Now for any $\epsilon > 0$, pick a finite L large enough so that both $\sum_{l=L+1}^{\infty} w_l(\mathrm{s}) < \epsilon/4$ and $\sum_{l=L+1}^{\infty} w_l(\mathrm{s}') < \epsilon/4$. The existence of such L is ensured by the almost sure convergence of the weights (see Appendix A). Since we already showed that there exists a Δ such that $|\mathrm{s} - \mathrm{s}'| < \Delta$ $|w_l(\mathrm{s}) - w_l(\mathrm{s}')| < \epsilon/(4L)$, this implies that

$$|Z_{\mathbf{s}_0}(\mathrm{s}) - Z_{\mathbf{s}_0}(\mathrm{s}')| < \epsilon$$

as desired.

D. Covariance structure in the Dependent PSBPs

Assume that $\alpha_l(\mathrm{s})$ follows a stochastic process with mean function $\mu_l(\mathrm{s})$ and covariance function $\sigma^2\gamma_l(\mathrm{s}, \mathrm{s}')$. Letting $u_l(\mathrm{s}) = \Phi(\alpha_l(\mathrm{s}))$ for all $\mathrm{s} \in S$, we can use a similar change of variables to that used in Appendix A to derive the covariance in the stick-breaking weight between locations s_1 and s_2,

$$\begin{aligned}
\mathsf{E}(u_l(\mathbf{s}_1)u_l(\mathbf{s}_2)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \Phi(z_1)\Phi(z_2)\frac{1}{2\pi}|\boldsymbol{\Sigma}_l|^{-1/2} \\
&\qquad \times \exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_l - \mathbf{z})'\boldsymbol{\Sigma}_l^{-1}(\boldsymbol{\mu}_l - \mathbf{z})\right\} d\mathbf{z} \\
&= \int_0^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}|\exp\left\{-\frac{1}{2}[(x_1 - t_1)^2 + (x_2 - t_2)^2]\right\} \\
&\qquad \times \frac{1}{2\pi}|\boldsymbol{\Sigma}_l|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\mu}_l - \mathbf{x})'\boldsymbol{\Sigma}_l^{-1}(\boldsymbol{\mu}_l - \mathbf{x})\right\} dx_1dx_2dt_1dt_2 \\
&= \mathsf{Pr}(T_1 > 0, T_2 > 0) = \beta_2(\mathbf{s}_1, \mathbf{s}_2)
\end{aligned}$$

where $\mathbf{z} = (z_1, z_2)'$, $\boldsymbol{\mu}_l = (\mu_l(\mathbf{s}_1), \mu_l(\mathbf{s}_2))'$, $\mathbf{x} = (x_1, x_2)'$, $\boldsymbol{\Sigma}_{ij} = \sigma^2\gamma_l(\mathbf{s}_1, \mathbf{s}_2)$ and $\mathbf{T} = (T_1, T_2)' \sim \mathsf{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l + \mathbf{I})$.

References

AGRESTI, A. (1990). *Categorical Data Analysis*. New York: Wiley.

ALBERT, J. H. & CHIB, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics* **57**, 829–836.

BANERJEE, S. & GELFAND, A. E. (2003). On smoothness properties of spatial processes. *Journal of Multivariate Analysis* **84**, 85–100.

BLACKWELL, D. & MACQUEEN, J. B. (1973). Ferguson distribution via Pólya urn schemes. *The Annals of Statistics* **1**, 353–355.

CARON, F., DAVY, M. & DOUCET, A. (2007). Generalized Pòlya urn form time-varying Dirichlet process mixtures. In *Proceedings 23rd Conference on Uncertainty in Artificial Intelligence*.

CARON, F., DAVY, M., DOUCET, A., DUFLOS, E. & VANHEEGHE, P. (2008). Bayesian inference for linear dynamic models with dirichlet process mixtures. *IEEE Transactions on Signal Processing* **56**, 71–84.

CARTER, C. K. & KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553.

CHIB, S. & HAMILTON, B. H. (2002). Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**, 67–89.

CHUNG, Y. & DUNSON, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of American Statistical Association, under revision* .

DEIORIO, M., MÜLLER, P., ROSNER, G. L. & MACEACHERN, S. N. (2004). An anova model for dependent random measures. *Journal of the American Statistical Association* **99**, 205–215.

DUAN, J. A., GUINDANI, M. & GELFAND, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika* **94**, 809–825.

DUNSON, D. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics* **7**, 551–568.

DUNSON, D. B., HERRING, A. H. & MULHERI-ENGEL, S. A. (2008). Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the Royal Statistical Society* **103**, 534–546.

DUNSON, D. B., PILLAI, N. & PARK, J.-H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society, Series B.* **69**, 163–183.

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

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

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

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2** , 615–629.

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

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

FRUHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis* **15**, 183–202.

GELFAND, A. E., KOTTAS, A. & MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100**, 1021–1035.

GELMAN, A. & RUBIN, D. (1992). Inferences from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.

GRIFFIN, J. (2008). The Ornstein-Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference. Technical report, Institute of Mathematics, Statistics and Actuarial Science, University of Kent.

GRIFFIN, J. E. & STEEL, M. F. J. (2006a). Nonparametric inference in time series problems. Technical report, Department of Statistics, University of Warwick.

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

HAN, D. F., LI, W. H. & LI, Z. C. (2008). Semantic image classification using statistical local spatial relations model. *Multimedia Tools and Applications* **39**, 169–188.

HIRANO, K. (2002). Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica* **70**, 781–799.

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

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

ISHWARAN, H. & ZAREPOUR, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica* **12**, 941–963.

JACQUIER, E., POLSON, N. G. & ROSSI, P. E. (1994). Bayesian analysis of stochastic volatility models. *Journal of business and Economic Statistics* **12**, 371–389.

KACPERCZYK, M., DAMIEN, P. & WALKER, S. G. (2003). A new class of Bayesian semiparametric models with applications to option pricing. Technical report, University of Michigan Bussiness School.

KAUFMAN, K. (1996). *Lives of North American Birds*. Houghton Mifflin.

KENT, J. T. (1989). Continuity properties for random fields. *Annals of Probability* **17**, 1432–1440.

KIM, S., SHEPHARD, N. & CHIB, S. (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies* **65**.

KOTTAS, A., BRANCO, M. D. & GELFAND, A. E. (2002). A nonparametric Bayesian modeling approach for cytogenetic dosimetry. *Biometrics* **58**, 593–600.

LAWS, D. J. & O'HAGAN, A. (2002). A hierarchical Bayes model for multilocation auditing. *Journal of the Royal Statistical Society, Series D* **51**, 431–450.

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

MACEACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55.

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

MEDVEDOVIC, M. & SIVAGANESAN, S. (2002). Bayesian infinite mixture model-based clustering of gene expression profiles. *Bioinformatics* **18**, 1194–1206.

MÜLLER, P., QUINTANA, F. & ROSNER, G. (2004). Hierarchical meta-analysis over related non-parametric Bayesian models. *Journal of Royal Statistical Society, Series B* **66**, 735–749.

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

ORBANZ, P. & BUHLMANN, J. M. (2008). Nonparametric bayesian image segmentation. *International Journal of Computer Vision* **77**, 23–45.

PAPASPILIOPOULOS, O. (2009). A note on posterior sampling from Dirichlet process mixture models. Technical report, Department of Economics, Universitat Pompeu Fabra.

PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158.

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

ROBERT, C. P. & CASELLA, G. (1999). *Monte Carlo Statistical Methods.* Springer - Verlag.

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

RODRIGUEZ, A., DUNSON, D. B. & GELFAND, A. E. (2008). The nested Dirichlet process, with Discussion. *Journal of American Statistical Association* **103**, 1131–1154.

RODRIGUEZ, A., DUNSON, D. B. & TAYLOR, J. (2009). Bayesian hierarchically weighted finite mixtures models for samples of distributions. *Biostatistics* **10**, 155–171.

RODRIGUEZ, A. & TER HORST, E. (2009). Measuring expectations in options markets: An application to the s&p500 index. *Quantitative Finance, under revision* .

RODRIGUEZ, A. & TER HORST, E. (2008). Bayesian dynamic density estimation. *Bayesian Analysis* **3**, 339–366.

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

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

WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Commnunications in Statistics, Part B - Simulation and Computation* **36**, 45–54.

WEST, M. & HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models.* Springer - Verlag, New York, second edition edition.