Collegio Carlo Alberto

Geometric Weight Priors and their Applications in Bayesian Nonparametrics

Ramsés H. Mena

No. 225
November 2011

Carlo Alberto Notebooks

www.carloalberto.org/working_papers

Geometric Weight Priors and their Applications in Bayesian Nonparametrics

Ramsés H. Mena[1]

November 2011

Abstract

Bayesian nonparametric techniques rely on suitable construction of random probability measures. The canonical example is without doubt the Dirichlet process, however in some situations different models are more suitable or preferred. Whereas most available alternatives to the Dirichlet process tend to generalized it in order to overcome certain prediction or fitting drawbacks, some of these issues might be rather solved with simpler models. Here we will review one of these simpler nonparametric priors, which can be seen as generated through a set of ordered weights within a species sampling model representation. We discuss various aspects of these random distributions as well as some of their applications in nonparametric mixtures, covariate or time dependent settings.

Keywords: Dirichlet process; Geometric distribution; Gibbs sampler; Measure-valued process; Nonparametric mixture model; Nonparametric regression; Random partitions.

1. Introduction. The chapter by [Walker (2011)] stresses the importance of having a dependent sample so that learning, about the distribution that generates the observations, can take place. It was also noted that if one assumes that the nature of the phenomenon under study generates independent and identically distributed (i.i.d.) observations, hence the dependence in the sample needed for Bayesian learning reduces to exchangeability. Whether one shares the notion of a correct model or prefers to think of the Bayesian approach to inference as implied by assuming exchangeability among the observations, the role of such a dependence property is apparent.

From de Finetti's representation theorem we see that a set of random variables $(X_i)_{i\geq 1}$ defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in a complete separable metric space $\mathbb{X}$, endowed with the Borel σ-field $\mathscr{X}$, is exchangeable if and only if for any $n \geq 1$

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} \mathsf{P}(A_i)\, \mathsf{Q}(\mathsf{dP}), \qquad A_i \in \mathcal{X} \tag{1}$$

where $\mathcal{P}_{\mathbb{X}}$ denotes the set of probability measures on $(\mathbb{X}, \mathcal{X})$. An interpretation immediately follows from the above representation, namely the unknown, say P, that allows to desegregate the join law of $(X_1, X_2, \ldots)$ into conditional independent and identical measures, is random and uniquely driven by Q. Therefore it is evident its relation to Bayesian statistics and thus the importance of specifying such a distribution Q for a random probability measure (r.p.m.), P.

Currently there are various approaches to define r.p.m.s, namely via extensions of finite dimensional distributions (Ferguson, 1973; Lijoi *et al.*, 2005); suitable normalization of

[1]Ramsés H. Mena is Associate Professor of Statistics at the Universidad Nacional Autónoma de México and an invited Research Fellow at Collegio Carlo Alberto. Email: ramses@sigma.iimas.unam.mx

stochastic processes (Ferguson, 1973; Regazzini *et al.*, 2003); representations through predictive distributions and species sampling models (Pitman, 1996); stick-breaking constructions (Sethuraman and Tiwari, 1982; Sethuraman, 1994; Ishwaran and James, 2001), etc. Up to date reviews of these approaches and their applications can be found in Walker *et al.* (1999) and Hjort *et al.* (2010). Each of these constructions provide different strengths either from an analytical perspective, from an numerical perspective or when used in particular applications or for extensions to non-exchangeable contexts. Without doubt the canonical example is Ferguson (1973) Dirichlet process, whose different constructions have in part served as gateway for the above extensions, aside from still being the main r.p.m. used in applications. Remember that for a partition $(B_1, \ldots, B_k)$ of $\mathbb{X}$ and a finite measure $\alpha > 0$ on $(\mathbb{X}, \mathcal{X})$, Ferguson (1973) defined the Dirichlet process $(\mathcal{D}_\alpha)$ as the stochastic process having finite dimensional distributions $(\mathsf{P}(B_1), \ldots, \mathsf{P}(B_k)) \sim \mathsf{Dir}(\alpha(B_1), \ldots, \alpha(B_k))$, where $\mathsf{Dir}(a_1, \ldots, a_k)$ denotes the Dirichlet distribution over the $(k-1)$-dimensional simplex.

Although tractable generalizations, such as the two parameter Dirichlet process (Pitman, 1995), have sometimes proven to be more adequate, the dominance of the Dirichlet process still prevails in the applied community. This is perhaps due to its mathematical and computational tractability which in part are due to its Blackwell and MacQueen (1973) Pólya urn representation which establishes that the $\mathcal{D}_{\theta \mathsf{P}_0}$ can be seen as the limit of predictive distributions

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n) = \frac{\theta}{\theta + n} \mathsf{P}_0(\cdot) + \frac{n}{\theta + n} \mathsf{P}_n(\cdot), \qquad n \geq 1 \tag{2}$$

with $X_1 \sim \mathsf{P}_0$, $\mathsf{P}_0(\cdot) := \alpha(\cdot)/\theta$, $\theta := \alpha(\mathbb{X})$ and $\mathsf{P}_n(\cdot) := n^{-1} \sum_{i=1}^n \delta_{X_i}(\cdot)$. Generalizations of the above urn representation of predictive distribution typically involve complicated weights which in turn are cumbersome to incorporate in MCMC algorithms (see for instance Lijoi *et al.*, 2005, 2007b). Indeed, most constructive approaches of r.p.m.s seek to generalize the Dirichlet process, resulting in richer r.p.m.s but at the same time posing additional complications when applying or studying such objects.

Here we will undertake a somehow different rationale, namely instead of searching for another construction to generalize the Dirichlet process we look for a somehow simpler r.p.m. that in turn simplifies some of its applications to non exchangeable situations via the construction of dependent processes. Before bringing out the main idea let us first review some well-know facts about the Dirichlet process and some other nonparametric priors.

1.1. Disentangling the total mass parameter of the Dirichlet process. A general approach to construct discrete r.p.m.s can be derived from the simple definition of a purely discrete measure on $(\mathbb{X}, \mathcal{X})$, that is

$$\mathsf{P}(B) = \sum_{i=1}^{\infty} \mathsf{w}_i \, \delta_{\mathsf{Z}_i}(B), \quad B \in \mathcal{X} \tag{3}$$

but with the difference that in this case the weights w_i and locations Z_i are random, that is $\sum_i \mathsf{w}_i = 1$ a.s., and independent of $\mathsf{Z}_i \overset{\text{iid}}{\sim} \mathsf{P}_0$, with P_0 a non-atomic measure on $(\mathbb{X}, \mathcal{X})$. Such general class of models are termed proper species sampling models (see Pitman, 1996). It easily follows that $\mathbb{E}[\mathsf{P}] = \mathsf{P}_0$, thus calling such a term the prior guess at the shape of P, which clearly turns into an important component when applying and studying (3).

In particular Sethuraman (1994) (see also Sethuraman and Tiwari, 1982) proved that if

the weights are given in the following stick-breaking form

$$\mathsf{w}_1 = \mathsf{v}_1 \quad \text{and} \quad \mathsf{w}_i = \mathsf{v}_i \prod_{j=1}^{i-1}(1-\mathsf{v}_j) \tag{4}$$

with $\mathsf{v}_i \overset{\text{iid}}{\sim} \text{Be}(1,\theta)$, $\theta > 0$, hence the Dirichlet process is recovered, *i.e.* $\mathsf{P} \sim \mathcal{D}_{\theta \mathsf{P}_0}$. The two parameter Poisson-Dirichlet proceses ($\mathcal{PD}(\sigma,\theta)$) follows from picking $\mathsf{v}_i \overset{\text{ind}}{\sim} \text{Be}(1-\sigma, \theta+i\sigma)$ with $\theta > -\sigma$ and $\sigma \in [0,1)$ (cf. Ishwaran and James, 2001).

An appealing feature of the Dirichlet process, and other r.p.m.s constructed as in (3), is its almost surely discreteness (see Blackwell, 1973), which also follows intuitively from the Pólya urn form of (2) where a positive probability is assigned to ties, *i.e.* $\mathbb{P}[X_i = X_j] > 0$ for any $i \neq j$. This implies that $\mathsf{X}^{(n)} := (X_1, \ldots, X_n)$ contains $\mathsf{K}_n \leq n$ distinct observations $(X_1^*, \ldots, X_{\mathsf{K}_n}^*)$ with corresponding frequencies $\mathbf{N}_{\mathsf{K}_n} = (\mathsf{N}_1, \ldots, \mathsf{N}_{\mathsf{K}_n})$ such that $\sum_{j=1}^{\mathsf{K}_n} \mathsf{N}_j = n$. Hence selecting an exchangeable sample of size n driven by $\mathcal{D}_{\theta\,\mathsf{P}_0}$ induces a partition into groups $\{\mathcal{G}_1, \ldots, \mathcal{G}_{\mathsf{K}_n}\}$, $\mathcal{G}_j := \{i : X_i = X_j^*\}$, $j = 1, \ldots, \mathsf{K}_n$ each with probability

$$\mathbb{P}[\{\mathsf{K}_n = k\} \cap \{\mathsf{N}_1 = n_1, \ldots, \mathsf{N}_{\mathsf{K}_n} = n_k\}] = \frac{\theta^k}{(\theta)_n} \prod_{i=1}^{k}(n_i - 1)! \tag{5}$$

where $n_j = \#\mathcal{G}_j$ and $(\theta)_n := \theta(\theta+1)\cdots(\theta+n-1)$. Notice that K_n and $\mathbf{N}_{\mathsf{K}_n}$ are sufficient for the grouping probabilities, namely they depend only on the clustering structure among the X_i's and not their actual values, *e.g.* if $n = 4$ and $k = 2$, the groups $\{\{X_1, X_2, X_3\}, \{X_4\}\}$ and $\{\{X_1, X_2, X_4\}, \{X_3\}\}$ receive (a priori) the same grouping probability as both are encoded by the same integer composition of 4, $(n_1, n_2) = (3, 1)$. In fact, the support of the above grouping probabilities is in bijection with the set of partitions of a set with n elements, $[n] = \{1, \ldots, n\}$, here denoted by $\mathcal{P}_{[n]}$, and whose cardinality is identified with the nth Bell number, B_n. In particular, such a number can be obtained through the recursion $\mathsf{B}_{n+1} = \sum_{k=0}^{n} \binom{n}{k} \mathsf{B}_k$, with $\mathsf{B}_1 = \mathsf{B}_0 = 1$. Due to the clear symmetry in the right hand side of (5) these probabilities, hereafter denoted by $\Pi_k^{(n)}(n_1, \ldots, n_k)$, can be typically encoded to be in bijection with smaller combinatorial structures, such as the space of compositions, or the space to integer partitions. These kind of distributions are known in the literature as exchangeable partition probability functions (EPPFs) and are widely used to study random partitions in areas such as population genetics (Ewens, 1972), combinatorics (Hansen, 1994), Economics (Aoki, 2004) and excursion theory (Pitman, 1996) among many others (see also Ewens and Tavarè, 1997). Within Bayesian nonparametric inference they can be applied to depict the clustering structure among observations in hierarchical mixture models (see Lo, 1984) or to give robust solutions to species sampling problems (see Lijoi *et al.*, 2007a).

Summing expression (5) over all integer compositions corresponding to set partitions for a fixed k one obtains the prior probability on the number of distinct values

$$\mathbb{P}[\mathsf{K}_n = k] = \frac{\theta^k}{(\theta)_n} |s(n,k)|, \qquad k = 1, \ldots, n \tag{6}$$

where $s(n,k)$ identifies the Stirling number of the first kind. Figure 1 shows the above probabilities for $n = 50$ and various choices of θ. Clearly the parameter θ is quite informative on the a priori number of groups.

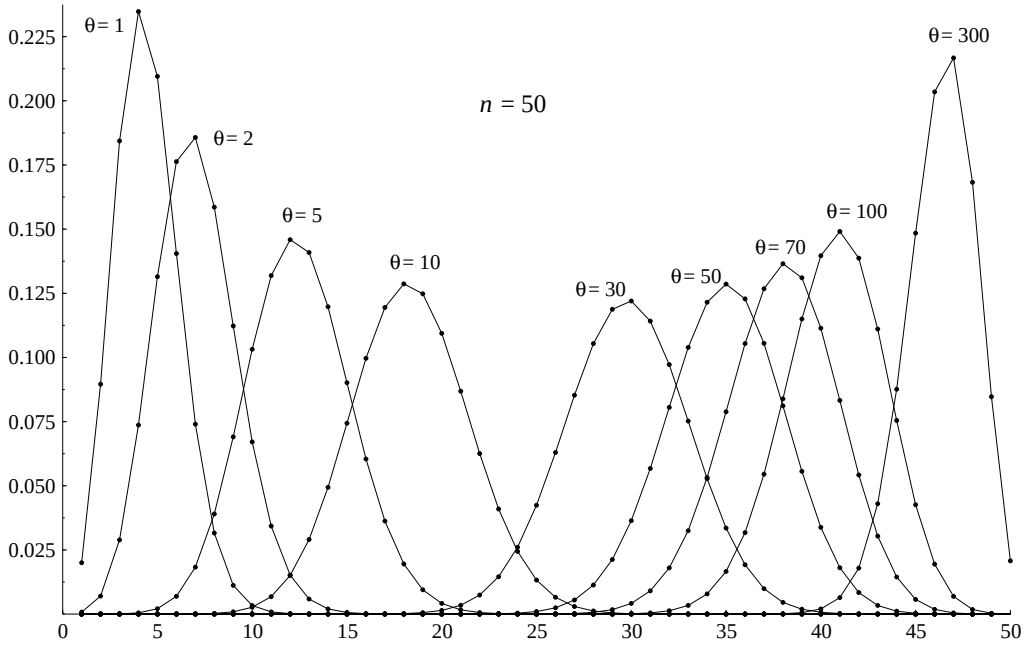

Figure 1: Prior probability on the number of different species corresponding to a sample of size 50 driven by $\mathcal{D}_{\theta \mathsf{P}_0}$ as θ varies.

Remark 1. The total mass parameter θ corresponding to a Dirichlet process is highly informative on the a priori number of groups.

One of the most popular applications of the Dirichlet process, and also other discrete r.p.m's, is when used as building block in hierarchical mixture models. Indeed following Lo (1984), a random density could be constructed as

$$\begin{aligned} Y_i \mid X_i &\overset{\text{ind}}{\sim} f(Y_i \mid X_i) \\ X_i \mid \mathsf{P} &\overset{\text{iid}}{\sim} \mathsf{P} \\ \mathsf{P} &\sim \mathsf{Q} \end{aligned} \tag{7}$$

where $f(\cdot \mid X)$ is typically a Lebesgue probability density and therefore used to model data of a continuous nature. Notice that if $f(\cdot \mid X) = \delta_X(\cdot)$ then we recover the no-mixture case. This means that we can model a set of $\mathbb{Y}$-valued observables $(Y_i)_{i \geq 1}$, also defined on $(\Omega, \mathscr{F}, \mathbb{P})$, though the random density

$$\mathsf{f}(y) = \int_{\mathbb{X}} f(y \mid x) \, \mathsf{P}(\mathrm{d}x)$$

namely the discrete r.p.m., P, incorporates its clustering structure at the level or latent variables X_i's. In particular, when $Q = \mathcal{D}_{\theta \mathsf{P}_0}$, Escobar (1988, 1994) and Escobar and West (1995) used this model, together with its representation (2), to propose one of the first Bayesian nonparametric MCMC implementations that in turn aid to promote the use of Bayesian nonparametric techniques and the Dirichlet process (see also Lenk, 1988; Berry and Christensen, 1979, for other earlier numerical approaches in Bayesian nonparametrics) .

Using representation (3) of a discrete r.p.m. the above model can be written as

$$\mathsf{f}(y) = \int_{\mathbb{X}} f(y \mid x)\, \mathsf{P}(\mathrm{d}x) = \sum_{i=1}^{\infty} \mathsf{w}_i\, f(y \mid \mathsf{Z}_i) \tag{8}$$

with $\mathsf{Z}_i \overset{\text{iid}}{\sim} \mathsf{P}_0$, namely an infinite mixture model. This clearly establishes a connection with the theory of finite mixture models typically used in model-based clustering (see Banfield and Raftery, 1993) and density estimation problems. In fact, if we are interested in the clustering among the observables Y_i's, it seems natural to think it as being induced by a clustering at the latent level of the X_i's. In particular, if we denote by $\mathsf{p}_n \in \mathcal{P}_{[n]}$ a particular partition of $\mathsf{X}^{(n)}$, and thus of $\mathsf{Y}^{(n)}$, where each Y_i is iid from the random density (8), hence we can compute (cf. Lijoi and Prünster, 2010) the posterior probability

$$\mathbb{P}(\mathsf{p}_n \mid \mathsf{Y}^{(n)}) \propto \mathbb{P}(\mathsf{p}_n)\mathbb{P}(\mathsf{Y}^{(n)} \mid \mathsf{p}_n) \tag{9}$$

where

$$\mathbb{P}(\mathsf{p}_n) = \Pi_k^{(n)}(n_1, \ldots, n_k)$$

is the corresponding EPPF to P, *e.g.* expression (5) when $\mathsf{P} \sim \mathcal{D}_{\theta \mathsf{P}_0}$, and

$$\mathbb{P}(\mathsf{Y}^{(n)} \mid \mathsf{p}_n) = \prod_{j=1}^{k} \int_{\mathbb{X}} \prod_{i \in \mathcal{G}_j} f(y_i \mid x_j) \mathsf{P}_0(\mathrm{d}x_j), \tag{10}$$

sometimes termed the clustering likelihood of $\mathsf{Y}^{(n)}$ for a given partition $\{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$. Although computing (9) can be simplified for specific choices of f and P_0, its direct evaluation throughout all its support, namely the subset $\mathcal{P}_{[n]}^k \subset \mathcal{P}_{[n]}$ of partitions of size k, is computationally unfeasible when n is large as a Stirling number of the second type, $\mathsf{S}_{n,k}$, of evaluations would be needed. Unlike the EPPF, the posterior distribution (9) is not longer exchangeable due to the effect of the kernel f in the observations Y_i's when evaluated in a particular group partition. Clearly, the problem becomes larger when we are also interested in the number of groups as we would need to compute

$$\mathbb{P}[\mathsf{K}_n = k] \propto \sum_{\mathsf{p}_n \in \mathcal{P}_{[n]}^k} \Pi_k^{(n)}(n_1, \ldots, n_k) \prod_{j=1}^{k} \int_{\mathbb{X}} \prod_{i \in \mathcal{G}_j} f(y_i \mid x_j) \mathsf{P}_0(\mathrm{d}x_j) \tag{11}$$

for $k = 1, \ldots, n$. Therefore resorting to MCMC or other kind of numeric methods is imperative for real applications. At the outset, building $\mathcal{P}_{[n]}$-valued Markov chains does not appear to be an easy task, luckily this is implicitly done with most MCMC methods based on Pólya urn schemes (cf. Escobar and West, 1995) as track of the k and frequencies $(n_1, \ldots, n_k)$ can easily be recorded iteration-wise. See also Lau and Green (2007) and the references therein for other approaches.

An important observation here is that when inferring about the clustering structure of $\mathsf{Y}^{(n)}$ using quantities such as (9) and (11) the interpretation is different to that typically obtained in analyses based on finite mixture models (cf. Richardson and Green, 1997), as under this latter approach the number of clusters is explicitly identified by the number of components in the mixture. Having said this, we could still say that the former approach, based on the posterior probabilities on partitions, is also "model based" as it also depends on the choice of f, P_0 and particular model for r.p.m.

Example 1. For computationally feasibility let us assume we observe the following small data set $y = (-1.522, -1.292, -0.856, -0.104, 2.388, 3.080, 3.313, 3.415, 3.922, 4.194)$, namely $n = 10$ with a $\mathsf{B}_{10} = 115975$ possible groups. Assuming the hierarchical model (7), with $\mathsf{Q} = \mathcal{D}_{\theta \mathsf{P}_0}$,

$$f(\cdot \mid \mu, \xi) = \mathsf{N}(\cdot \mid \mu, \xi^{-1}) \quad \text{and} \quad \mathsf{P}_0(\mathrm{d}\mu, \mathrm{d}\xi) = \mathsf{N}(\mu \mid 0, (\tau\,\xi)^{-1})\mathsf{Ga}(\xi \mid \alpha, \beta)\mathrm{d}\mu\,\mathrm{d}\xi, \tag{12}$$

for $\tau, \alpha, \beta > 0$, the intra partition likelihood (10) becomes

$$\prod_{j=1}^{k} \left\{ \frac{\tau}{n_j + \tau} \right\}^{\frac{1}{2}} \frac{\beta^{\alpha}\,\Gamma(\alpha + \frac{n_j}{2})}{(2\pi)^{\frac{n_j}{2}}\,\Gamma(\alpha)\left[\beta + \frac{S_j}{2}\right]^{\alpha + \frac{n_j}{2}}}$$

where $S_j = \sum_{i \in \mathcal{G}_j} y_i^2 - n_j\,\bar{y}_j^2/(n_j + \tau)$ and $\bar{y}_j = n_j^{-1} \sum_{i \in \mathcal{G}_j} y_i$. Hence by direct evaluation across the 115975 possible partitions we can compute the posterior probabilities (9) and (11). Figure 2 shows an histogram of the data from where it is evident that two groups is the most intuitive choice. In particular, if we set $\tau = 0.1$ and $\alpha = \beta = 1$, we obtain the values in Table 1, from where it is clear that the choice of total mass parameter θ influences the posterior clustering probabilities. Indeed, when $\theta = 0.5$ or $\theta = 1$, the posterior mode of (9) sits on the most intuitive partition, say $\mathsf{p} = \{\{y_1, y_2, y_3, y_4\}, \{y_5, y_6, y_7, y_8, y_9, y_{10}\}\}$, with probabilities 0.522 and 0.332 respectively whereas when $\theta = 5$ the posterior mode is on $\mathsf{p}^* := \{\{y_1\}, \{y_2\}, \{y_3\}, \{y_4\}, \{y_5, y_6, y_7, y_8, y_9, y_{10}\}\}$ with probability 0.0204. To some extent, these observations could have been predicted since the prior expectation of K_n is given by

$$\mathbb{E}[\mathsf{K}_n] = \sum_{i=1}^{n} \frac{\theta}{\theta + i - 1}$$

which for $\theta = 0.5, \theta = 1$ and $\theta = 5$ one it yields $2.13, 2.93$ and 5.84 respectively, thus slightly coinciding with the posterior results of Table 1. Clearly the effect of the prior could be diminished as the simple size n increases. □

Remark 2. The total mass parameter θ corresponding to a Dirichlet process strongly influences the posterior inference on the number of groups in a Bayesian non parametric mixture model.

The remark that the choice of θ, in Dirichlet processes applications, strongly influences the prior and posterior results is not new, indeed in order to attain consistent results, *e.g.* within MCMC implementations based on Pólya urn schemes, one typically needs to assign a prior to θ and then incorporate it within the learning process of the sampler as noticed by Escobar and West (1995). Although such an effect of an informative prior can be attenuated with other choices of r.p.m.s such as the two parameter Poisson-Dirichlet or normalized generalized gamma processes (cf. Lijoi *et al.*, 2007b), a similar effect still remains present for their corresponding parameters.

In fact such a total mass parameter takes different interpretations in the literature, *e.g.* the scale parameter of the $\mathcal{D}_{\theta \mathsf{P}_0}$, namely it regulates the concentration mass around the prior mean at the shape P_0, it can be thought as the prior sample size (cf. Sethuraman

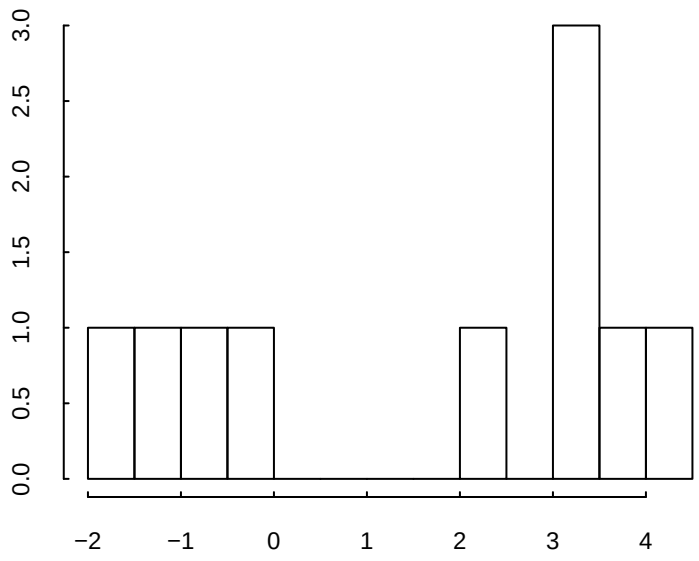

Figure 2: Histogram for the small data set.

k	$\theta = 0.5$	$\theta = 1$	$\theta = 5$
1	0.019469	0.00619	0.000071
2	**0.591630**	0.37634	0.021504
3	0.312288	**0.39729**	0.113509
4	0.067986	0.17298	0.247113
5	0.008033	0.04088	**0.291972**
6	0.000568	0.00578	0.206592
7	0.000025	0.00051	0.090763
8	$6.74\,\mathsf{E}{-7}$	0.00003	0.024486
9	$1.03\,\mathsf{E}{-8}$	$8.38\,\mathsf{E}{-7}$	0.003740
10+	$6.85\,\mathsf{E}{-11}$	$1.12\,\mathsf{E}{-8}$	0.000249

Table 1: Exact posterior probabilities (11) for the number of groups K_n corresponding to the small data set and different choices of θ. In bold face the modal probabilities.

and Tiwari, 1982), it can also be used to match some moments of linear functionals of the Dirichlet process (cf. Walker and Mallick, 1997). Furthermore it arises in other areas such as populations genetics, where it takes the role of the mutation rate in a Wright-Fisher-type population model (cf. Ewens, 1972).

Remark 3. In practical implementations based on the Dirichlet process, the total mass parameter θ needs to be incorporated in the learning process.

A relevant remark is at hand: some numerical techniques to estimate Bayesian nonparametric mixture models, based on Dirichlet processes and on other more general stick-breaking r.p.m.s, are based on truncations of the infinite series representation (3), namely

$$\mathsf{P}_T(\cdot) = \sum_{i=1}^{T} \mathsf{w}_i \, \delta_{\mathsf{Z}_i}(\cdot) \tag{13}$$

for $T < \infty$, and thus rely on a L_1 error bound to determine the value of the truncation point T (see Theorem 2 in Ishwaran and James, 2001). The problem is typically that stick-breaking weights are not necessarily ordered and therefore such a bound is dependent on the parameters corresponding to the weights. Notice that when weights are ordered we could simply take maximum possible weight to set a unique error bound. This consideration, together with our previous remarks about θ, imply that we would also need to incorporate T in the MCMC analysis, which is certainly neither commonly done nor an easy task. For instance, if we denote by $\mathcal{L}(\mathsf{Y})$ the marginal density of Y under the mixture model (7) with $\mathsf{Q} = \mathcal{D}_{\theta \mathsf{P}_0}$ and $\mathcal{L}_T(\mathsf{Y})$ the one corresponding to the above truncation, hence Ishwaran and James (2001) show that $||\mathcal{L}(\mathsf{Y}) - \mathcal{L}_T(\mathsf{Y})|| \approx 4n \exp\{-(T-1)/\theta\}$. This means that if we have $n = 50$ data and a big number of distinct values k, *i.e.* big θ, then we would need a relatively large truncation point. Of course we could potentially use a deliberately huge value for T, however, this is not always possible. Mostly when $\mathcal{D}_{\theta \mathsf{P}_0}$ is instrumental for other more complex models such as those found in some dependent Dirichlet processes applications.

Remark 4. Unordered random weights in species sampling models might result in truncations methods highly dependent on their parameters.

So we have underlined some issues of the Dirichlet process that in part are mainly result of the randomness of its weights, and so clearly reflected on its only parameter θ. The objective of this work is then to present a somehow easier r.p.m. that to some extent overcome such issues. In Section 2 we present the geometric weights prior as one of such simpler models, some estimation aspects and an intuitive derivation through a more general class of r.p.m.s are also presented. Being a simpler object it also makes it appealing to extensions to non-exchangeable contexts, in Section 3 some applications to construct dependent nonparametric processes are presented. In particular its application to build models for Bayesian nonparametric regression and to construct probability measure-valued diffusion processes is explored. Some examples illustrating the proposed models are also presented. Some concluding remarks are deferred to Section 4.

2. Geometric weights priors. A simple inspection to representation (3) of a discrete random probability measure tell us that there are two main sources of randomness, namely the weights and the locations. From the observations in the above discussion we learned that we need to put a prior on the parameters of the weights, particularly on the total mass parameter corresponding to the Dirichlet process. In part, this can be thought as originated for

an excess of randomness on the corresponding weights. In fact, as noticed in the chapter by [Walker (2011)], the lack of order in the Dirichlet process weights, induces a non identifiability problem. In other words, the fact that the stick-breaking weights are not deterministically ordered provokes that mass in a particular location $B \in \mathcal{X}$, *i.e.* $\mathsf{P}(B)$, can be attained by many different combinations of weights, w_i's and locations Z_i's, this issue is clearly inherited by mixtures based in such a prior. Hence one wonders whether one can define a r.p.m. with simpler weights, that from one side makes better use of the availability of infinite locations to attain a particular mass and at the same time solve the identifiability issue. As we will see below the answer to this question is positive and it turns out to be very simple, namely by removing a level of randomness in the stick-breaking weights.

Instead of considering weights as in (4), consider

$$\omega_i := \mathbb{E}\left[\mathsf{w}_i\right] = \mathbb{E}\left[\mathsf{v}_i \prod_{j=1}^{i-1}(1-\mathsf{v}_j)\right] = \mu_i(\psi)\prod_{j=1}^{i-1}(1-\mu_j(\psi))$$

with $\mu_i(\psi) := \mathbb{E}[\mathsf{v}_i]$ and where ψ here generically denotes the parameter (or parameters) corresponding to the distribution of the v_i's. Next, assume that the v_i's have been chosen such that $\omega_i > \omega_{i+1}$, that is $\mu_{i+1}(\psi) < \mu_i(\psi)(1-\mu_i(\psi))^{-1}$ for all i. Notice also that when randomizing ψ the arguments in the right hand side product of the above expression are not longer independent.

For the Dirichlet process we have $\mathsf{v}_i \overset{\text{iid}}{\sim} \text{Be}(1,\theta)$, thus setting $\lambda := \mu_i(\theta) = (1+\theta)^{-1}$, we have

$$\omega_i = \lambda(1-\lambda)^{i-1} \tag{14}$$

namely geometric weights. Accordingly with the idea or randomizing θ, we could also assign a prior distribution to λ, say $\lambda \sim \text{Be}(a,b)$. In other words, we can consider the following r.p.m.

Definition 1. Let $\mathcal{P}_{\mathbb{X}}$ the set of probability measures on $(\mathbb{X},\mathcal{X})$. We called a r.p.m. $\mathsf{P} \in \mathcal{P}_{\mathbb{X}}$ a geometric weights prior, $\mathcal{GWP}(a,b)$ if

$$\mathsf{P}(\mathsf{B}) = \lambda\sum_{i=1}^{\infty}(1-\lambda)^{i-1}\,\delta_{\mathsf{Z}_i}(B), \quad B \in \mathcal{X} \tag{15}$$

with $\mathsf{Z}_i \overset{\text{iid}}{\sim} \mathsf{P}_0$ and $\lambda \sim \text{Be}(a,b)$, $a,b>0$.

Notice that $\mathsf{P} \sim \mathcal{GWP}(a,b)$ is an almost surely discrete random probability measure and that, as for the Dirichlet process, and other species sampling models, the probability measure P_0 can be thought of as the prior guess at the shape of P, since $\mathbb{E}[\mathsf{P}] = \mathsf{P}_0$.

In some manner, the above definition simplifies that of $\mathcal{D}_{\theta\mathsf{P}_0}$ with random θ, as one level of randomness in the weights has been removed by replacing them with their expected values while keeping the required prior on θ. Indeed, it is a somehow simpler object than the Dirichlet process, as it is constructed through random weights that depend on only one Beta random variable, instead of on an infinite number of them as in the Dirichlet process. In other words the stick is always broken with the same Beta random variable in order to construct the weights.

In fact, at first sight the random probability measure provided by (15) can be misinterpreted as a special case of the Dirichlet process, since the weights of the former can be

obtained as in (4) by letting $(\mathsf{v}_1, \mathsf{v}_2, \ldots)$ be all equal to the same realization of a Beta random variable, so that w_i, $i = 1, 2, \ldots$, are mixed geometric with Beta(a, b) as the mixing distribution. However this proves not to be the case. First because in the Dirichlet process case the parameter a of the Beta distribution is constrained to be one. And more significantly because the Dirichlet process, $\mathsf{P} \sim \mathcal{D}_{\theta \mathsf{P}_0}$ is characterized by the distributional equation

$$\mathsf{P} \stackrel{d}{=} \mathsf{v}_1 \delta_{\mathsf{Z}_1} + (1 - \mathsf{v}_1)\mathsf{P} \tag{16}$$

with $\mathsf{v}_1 \sim \text{Be}(1, \theta)$ and $Z_1 \sim \mathsf{P}_0$, stochastically independent of P (cf. Sethuraman, 1994). The same procedure applied to $\mathsf{P}^* \sim \mathcal{GWP}(a, b)$ yields

$$\mathsf{P}^* \stackrel{d}{=} \lambda\, \delta_{\mathsf{Z}_1} + (1 - \lambda)\mathsf{P}^* \tag{17}$$

The crucial difference between these two cases is that in (16), P is independent of $(\mathsf{v}_1, \mathsf{Z}_1)$, while in (17), P^* is independent of Z_1 but not of λ. Hence we are dealing with a different random probability measure.

A key issue that arises when considering whether to use a random probability measure of type (15) is the impact of the strong constraint on inference in terms of flexibility, namely one wonders whether it has full support. This could be a more relevant concern when one compares the Dirichlet process, or other more complex r.p.m.s, with the rather simplistic r.p.m. of Definition 1. The following proposition, whose proof easily follows from results in Ongaro and Cattaneo (2004), justifies the use of the proposed model for inferential purposes.

Proposition 1. Let $\mathscr{P}$ be the probability distribution induced on $\mathcal{P}_{\mathbb{X}}$ by random probability measures $\mathsf{P} \sim \mathcal{GWP}(a, b)$. Then the support of $\mathscr{P}$ in the topology of weak convergence on $\mathcal{P}_{\mathbb{X}}$ is given by all probability measures $\mathsf{G} \in \mathcal{P}_{\mathbb{X}}$ such that the support of G is included in that of P_0.

Therefore the geometric weights prior constitutes a valid alternative to the Dirichlet process in Bayesian nonparametric applications. In particular, it all points out towards a good prior candidate for mixture models, *i.e.* it has simpler and ordered weights, full support, same prior guess at the shape driven by our choice of P_0. However, before going into illustrations we provide with an alternative derivation of $\mathsf{P} \sim \mathcal{GWP}(a, b)$ that allows us to have a better intuition of why such a simpler object works just as fine.

2.1. An intuitive derivation. Walker (2007) proposed an alternative method for posterior inference based on the nonparamteric mixture model (7) that overcomes the infinite summation in (8). By augmenting (8) through a latent uniform variable one can slice the corresponding infinite summation by considering the joint density

$$\mathsf{f}(y, u) = \sum_{i=1}^{\infty} \mathbb{I}(u < \mathsf{w}_i)\, f(y \mid \mathsf{Z}_i) \tag{18}$$

with conditional density

$$\mathsf{f}(y \mid u) = \frac{1}{|\mathsf{A}_u|} \sum_{i \in \mathsf{A}_u} f(y \mid \mathsf{Z}_i), \tag{19}$$

where

$$\mathsf{A}_u := \{j : u < \mathsf{w}_j\} \tag{20}$$

and $|\mathsf{A}_u| := \sum_{i=1}^{\infty} \mathbb{I}(u < \mathsf{w}_i)$. Notice that given u the random set of component-indexes in the mixture, A_u, is finite, *i.e.* $|\mathsf{A}_u| < \infty$. Based on this idea Walker (2007) proposed a Gibbs sampler for posterior analysis, based on the nonparametric mixture model (7), that avoids truncations such as (13). See also Kalli *et al.* (2011) for more efficient implementations.

Here we are interested in the conditional distribution (19), as it can also be used to construct other kind of random densities, namely

$$\mathsf{f}(y \mid \mathsf{A}) = \frac{1}{|\mathsf{A}|} \sum_{i \in \mathsf{A}} f(y \mid \mathsf{Z}_i), \tag{21}$$

where A is a random finite subset of $\mathbb{N}_+$ and, as before, $\mathsf{Z}_i \overset{\text{iid}}{\sim} \mathsf{P}_0$. This somehow resembles the approach undertaken in finite mixture models, but with some differences. First notice that if we assume model (21) for each observation, Y_j, then there will be a random set A_j for each of them, whereas in the finite mixture model approach the number of components, N, suffices for all observations. In fact this is one of the reasons for having complex weight and location specifications in the finite mixture approach, *i.e.* to build a richer model able to allocate the required mass in a particular location. On the other hand, for Bayesian nonparametric mixtures such as those based on the Dirichlet process and other discrete r.p.m.s, there are an infinite number of locations, Z_i's, at our disposal to allocate a particular mass, therefore the complexity in the weights modulating a r.p.m. can be relaxed. In particular, notice that the random set A corresponding to the Dirichlet process might have gaps, *i.e.* we can have realizations of the sort of $\{2, 5, 10, 345, 1004\}$, namely it is not a consecutive sequence of integers from 1 to N as it is typically for finite mixture models. In fact, having an infinite number of locations at our disposal, we do not see a clear need to have index sets with gaps.

Hence, an idea is to consider the random set $\mathsf{A} := \{1, \ldots, \mathsf{N}\}$ so (21) reduces to

$$\mathsf{f}(y \mid \mathsf{N}) = \frac{1}{\mathsf{N}} \sum_{i=1}^{\mathsf{N}} f(y \mid \mathsf{Z}_i)$$

where N is random and modeled through a distribution supported on $\mathbb{N}_+$, namely $q_N(\cdot \mid \lambda)$, where for now λ denotes a generic parameter of the chosen distribution for N.

If we write out the model by marginalizing over N then we have

$$\mathsf{f}(y) = \sum_{l=1}^{\infty} \frac{1}{l} \sum_{i=1}^{l} f(y \mid \mathsf{Z}_i)\, q_{\mathsf{N}}(l \mid \lambda)$$

which can also be written as

$$\mathsf{f}(y) = \sum_{i=1}^{\infty} \omega_i\, f(y \mid \mathsf{Z}_i) \tag{22}$$

where

$$\omega_i = \sum_{l=i}^{\infty} \frac{q_{\mathsf{N}}(l \mid \lambda)}{l} \tag{23}$$

If we further randomize λ and assign a prior for it, *e.g.* $\lambda \sim \pi$, then (22) becomes a mixture based on a species sampling model (or simply a species sampling model if we take $f(\cdot \mid \mathsf{Z}_i) =$

$\delta_{Z_i}(\cdot)$). Clearly the weights (23) sum up to one (a.s. if λ is random) and are in decreasing order as one sees that $\omega_{i+1} = \omega_i - q_{\mathsf{N}}(i \mid \lambda)/i$.

Therefore (22) and (23) provide with an alternative to the stick-breaking way of constructing species sampling models. By changing the choice of $q_{\mathsf{N}}(\cdot \mid \lambda)$ and the prior for λ, we obtain different r.p.m.s.

Of particular interest here is when we assume $\mathsf{N} \sim \text{Neg-Bin}(2, \lambda)$, that is

$$q_{\mathsf{N}}(l \mid \lambda) = l\lambda^2(1-\lambda)^{l-1}\,\mathbb{I}(l \in \mathbb{N}_+)$$

which evaluating (23) allows us to recover the geometric weights, that is

$$\omega_i = \lambda(1-\lambda)^{i-1}$$

Hence, the r.p.m. of Definition 1 can be recovered as a r.p.m.s of the sort

$$\mathsf{P}(B) = \mathbb{E}_{q_{\mathsf{N}}}\left[\frac{1}{\mathsf{N}}\sum_{i=1}^{\mathsf{N}}\delta_{\mathsf{Z}_i}(B)\right] \tag{24}$$

where $\mathsf{Z}_i \overset{\text{iid}}{\sim} \mathsf{P}_0$, $q_{\mathsf{N}} = \text{Neg-Bin}(2, \lambda)$ and $\lambda \sim \text{Be}(a, b)$.

2.2. Posterior inference. Based on the construction of the previous subsection Fuentes *et al.* (2010) proposed an algorithm for posterior inference under the following setting. Suppose we have a sample $\mathsf{Y}^{(n)} := (Y_1, \ldots, Y_n)$ modeled by the nonparametric mixture model (7) where P follows a r.p.m. as in (24). Introducing a latent variable d_i that, given N_i, indicates from which component Y_i came from we can rewrite the nonparametric mixture model as

$$\begin{aligned} Y_i \mid \mathsf{Z}^{(n)}, d_i, \mathsf{N}_i &\overset{\text{ind}}{\sim} f(Y_i \mid \mathsf{Z}_{d_i}) \\ d_i \mid \mathsf{N}_i &\overset{\text{ind}}{\sim} \mathsf{U}\{1, \ldots, \mathsf{N}_i\} \\ \mathsf{N}_i &\overset{\text{iid}}{\sim} q_{\mathsf{N}}(\cdot \mid \lambda) \\ \lambda &\sim \pi \end{aligned} \tag{25}$$

with $\mathsf{Z}_i \overset{\text{iid}}{\sim} \mathsf{P}_0$. Following Fuentes *et al.* (2010) a Gibbs sampler algorithm, for the case $q_{\mathsf{N}} = \text{Neg-Bin}(2, \lambda)$ and $\lambda \sim \text{Be}(a, b)$, *i.e.* for a mixture model (7) when $\mathsf{P} \sim \mathcal{GMP}(a, b)$, reduces to sample from the following full conditional distributions

$$f(\mathsf{Z}_j \mid \cdots) \;\propto\; \mathsf{P}_0(\mathsf{Z}_j)\prod_{d_i=j} f(y_i \mid \mathsf{Z}_j), \qquad \text{for } j = 1, \ldots, M \tag{26}$$

$$\mathbb{P}(d_i = l \mid \cdots) \;\propto\; f(Y_i \mid X_l)\,\mathbb{I}(l \in \{1, \ldots, \mathsf{N}_i\})$$

$$\mathbb{P}(\mathsf{N}_i = j \mid \cdots) \;=\; \lambda(1-\lambda)^{j-1}\mathbb{I}(j \geq d_i)$$

$$f(\lambda \mid \cdots) \;=\; \text{Be}\left(\lambda \mid a + 2n, b + \sum_{i=1}^{n}\mathsf{N}_i - n\right) \tag{27}$$

for $i = 1, \ldots, n$ and where $M = \max\{\mathsf{N}_1, \ldots, \mathsf{N}_n\}$. Clearly, the Gibbs sampler simplifies when f and P_0 form a conjugate pair.

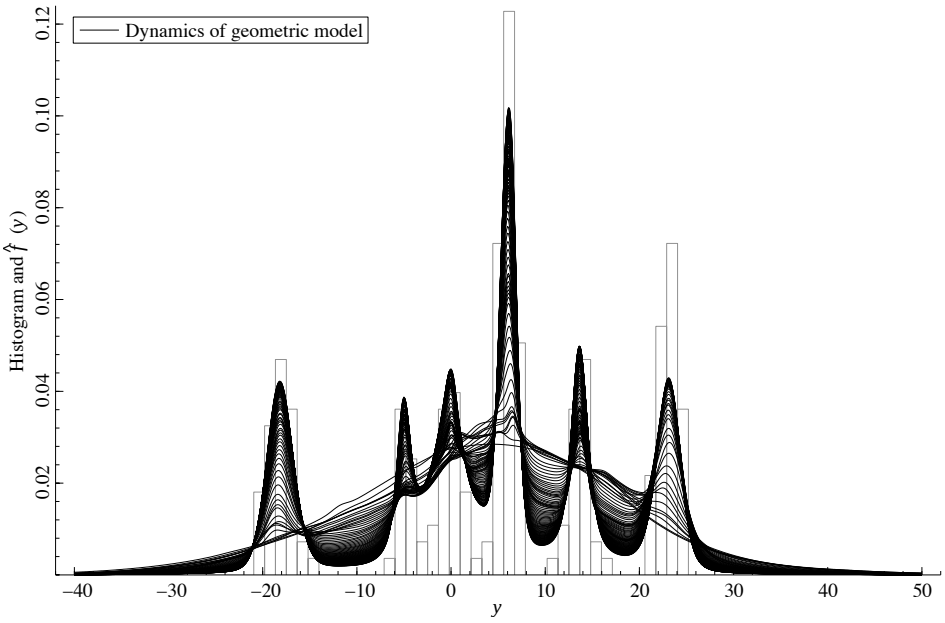

Figure 3: Dynamics of the density estimator, based on the geometric weights prior, through the first 100 iterations of the Gibbs sampler algorithm for the mean-scale mixtures data set. The hyper–parameters are given by $(\tau, \alpha, \beta, a, b) = (100, 0.5, 0.5, 0.5, 0.5)$ and initial values $N_i = 10$ and $d_i \in \{1, \ldots, 10\}$ for all $i = 1, \ldots, 240$.

Example 2. In order to illustrate the performance of the above mixture of $\mathcal{GWP}(a, b)$, let us consider 240 data points coming from a mean-variance mixture of six normal distributions with weights $(0.17, 0.08, 0.125, 0.2, 0.125, 0.21)$ and mean-variance parameters given by $(-18, 2), (-5, 1), (0, 1), (6, 1), (14, 1)$ and $(23, 125)$. We assume the same kernel and prior guess at the shape specifications as in (12).

Figures 3 and 4 show the dynamics of the density estimator for the first 100 iterations based on mixtures of $\mathcal{GWP}(a, b)$ and on mixtures of $\mathcal{D}_{\theta \mathsf{P}_0}$ respectively. From Figure 3 we note that the availability of an unlimited number of Z_j's to represent a particular cluster location always results in a improvement in subsequent iterations whereas for the Dirichlet process case, Figure 4, the algorithm requires several iterations to detect a good candidate for the Z_j representing a particular location. This feature is better appreciated in the mode welling around -18, which can be thought as being far from the overall mean of the data. It can also be observed at the tails of the density estimators in Figure 3, where for the initial iterations a bigger mass, than that shown for the Dirichlet process case, is allocated. □

It is probably worth mentioning that the above drawback of nonparametric mixtures based on the Dirichlet process, and also on other discrete r.p.m.s, has received considerable attention in the Bayesian nonparametric literature resulting in algorithms that aim to accelerate the identification of good candidates for the Z_j's identifying particular cluster locations. (see for instance MacEachern and Müller, 1998). However, despite these efforts, this issue is not fully resolved.

Figure 5 shows the estimates for both, the Dirichlet process the geometric weights prior

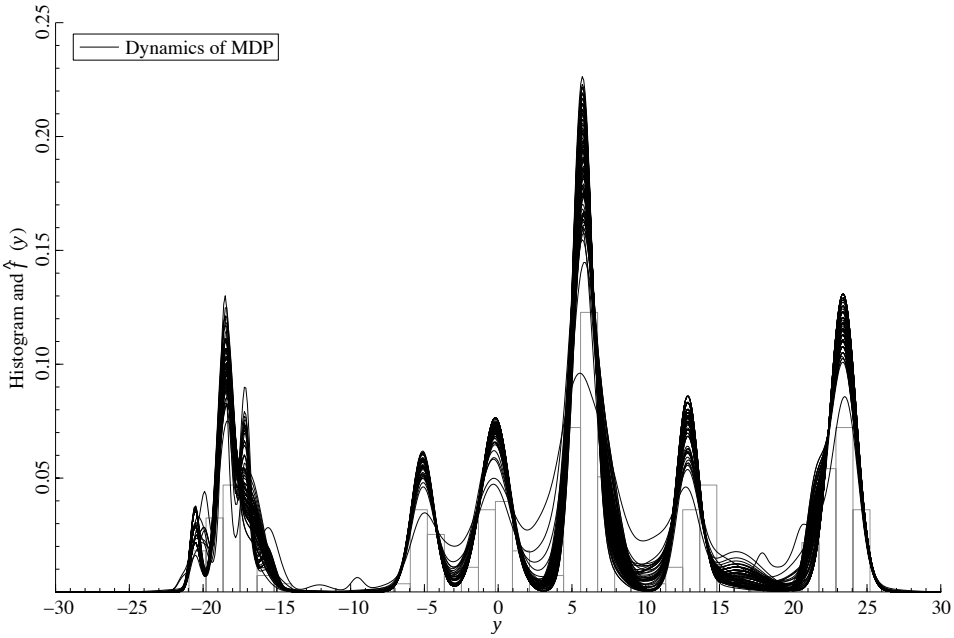

Figure 4: Dynamics of the density estimator for the mixture of Dirichlet process, through the first 100 iterations of the Gibbs sampler algorithm for the mean-scale mixtures data set.

cases, at a convergent stage. This figure also compares the true model that generated the observations, as we can see both approaches can be thought as being relatively satisfactory, however the approach based on $\mathcal{GWP}(a, b)$ appears to be closer to the true model.

3. Dependent processes based on geometric weights priors. As we underlined in the introduction the type of dependence driving Bayesian statistics is exchanegability. However there are many phenomena where models with a more structured dependence are needed and would also benefit from a nonparametric approach, *e.g.* regression analysis, time series analysis, stochastic processes, etc. This has motivated that the Bayesian nonparametric community devotes considerable attention to what currently is termed dependent nonparametric processes. The idea is simple, to provide a family of random probability measures linked by a suitable dependence structure, *e.g.* by means of a set of covariates or a time parameter, and use it for drawing inferences on random phenomena in appropriate frameworks, usually with the aid of simulation techniques.

When this dependence structure can be indexed via a continuous time parameter, then we are effectively dealing with probability-measure-valued processes devised for nonparametric inference purposes. Apart from the theoretical developments offered by the probabilistic study of measure-valued processes, whose literature is certainly broad and well established (see for example Ethier and Kurtz, 1986), on the statistical side this is a relatively young area and to date the most productive ideas have involved the Dirichlet process.

Initiated in part with the papers by MacEachern (1999) (see also Feigin and Tweedie, 1989) who introduced the notion of dependent Dirichlet process, the current literature on the topic includes, among others, De Iorio *et al.* (2004) who proposed a model with an ANOVA-type dependence structure, Gelfand *et al.* (2005) who apply the dependent Dirichlet process

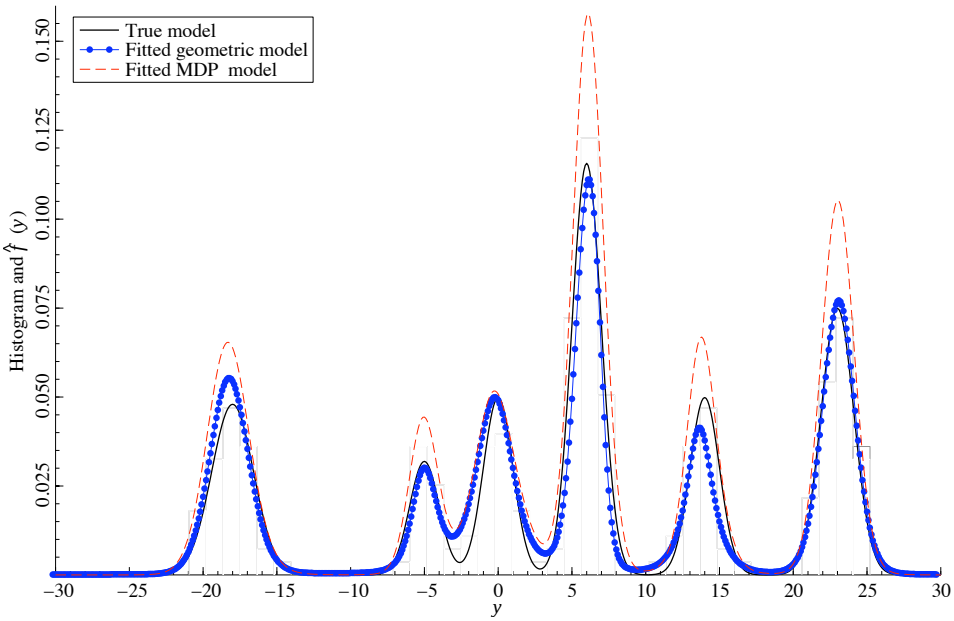

Figure 5: Density estimates for the 6 modes simulated data set based on both $\mathcal{GWP}(a, b)$ and $\mathcal{D}_{\theta \mathsf{P}_0}$. The estimates are based on 10,000 after a burn in period of 2000 iterations. The hyper–parameters are given by $(\mu, \tau, \alpha, \beta, a, b) = (0, 100, 0.5, 0.5, 0.5, 0.5)$ and $N_i = 10$ and $d_i \in \{1, \ldots, 10\}$ for all $i = 1, \ldots, 240$.

to spatial modeling by using a Gaussian process for the atoms, Griffin and Steel (2006) who let the dependence on the random masses be directed by a Poisson process, Caron *et al.* (2006) who model nonparametrically the noise in a dynamic linear model, Dunson and Park (2008) who construct an uncountable collection of dependent random probability measures based on a stick-breaking procedure with kernel-based weights. See also Dunson *et al.* (2007), Rodriguez and Ter Horst (2008), Petrone *et al.* (2009), Rodriguez and Dunson (2011), Griffin and Steel (2010) for other contributions in these directions.

Hence, the main idea is to construct a $\mathcal{P}_{\mathbb{X}}$-valued stochastic process $\{\mathsf{P}_z\}_{z \in \mathcal{Z}}$, which following representation (3) of a random probability measure, extends to

$$\mathsf{P}_z(B) = \sum_{i=1}^{\infty} \mathsf{w}_i(z)\, \delta_{\mathsf{Z}_i(z)}(B), \quad B \in \mathcal{X} \tag{28}$$

where $\{\mathsf{w}_i(z)\}_{i=1}^{\infty}$ and $\{\mathsf{Z}_i(z)\}_{i=1}^{\infty}$ are infinite collections of stochastic processes, indexed by $z \in \mathcal{Z}$. Hence the dependence structure set in the weights and locations drives the dependence relations at the r.p.m. level. Clearly while specifying a dependent nonparametric process we could have both weights and locations dependent or only one of them.

A natural idea when constructing dependent processes is to keep their marginal behavior to be a known r.p.m., *e.g.* a Dirichlet process, a two parameter Dirichlet process, a geometric weights prior, etc. This is easily done by setting strictly stationary processes with the desired marginal distribution. In what follows we will revise a couple of these examples based on geometric weights priors.

3.1. Covariate dependence for regression problems. A well-known problem in general regression analysis is to set a link function between observations and covariates. A Bayesian nonparametric approach to deal with this problem is to propose dependent random densities such as

$$\mathsf{f}_z(y) = \int f(y \mid x)\, \mathsf{P}_z(\mathsf{d}x).$$

where $\{\mathsf{P}_z\}_{z\in\mathcal{Z}}$ follows a covariate dependent nonparametric process.

Fuentes *et al.* (2009) followed this idea and proposed a dependent nonparametric process with marginals $\mathcal{GWP}$ to model $\{\mathsf{P}_z\}_{z\in\mathcal{Z}}$, that is

$$\mathsf{P}_z(\cdot) = \sum_{i=1}^{\infty} \lambda(z)(1-\lambda(z))^{i-1}\, \delta_{\mathsf{Z}_i}(\cdot),$$

where $\mathsf{Z}_i \overset{\text{iid}}{\sim} \mathsf{P}_0$ and

$$\lambda(z) = \frac{e^{\xi(z)}}{1+e^{\xi(z)}} \tag{29}$$

with $\xi := \{\xi(z)\}_{z\in\mathcal{Z}}$ a Gaussian process with continuous mean $\boldsymbol{\mu}$ and continuous covariance function $\boldsymbol{\sigma}$. Namely we retain the iid locations and induce the dependence only through the simple structure of the geometric weights prior.

Some z-dependence properties can be directly studied from the geometric structure of the weights, for instance it is easy to see (see Fuentes *et al.*, 2009) that for $B \in \mathcal{X}$

$$\mathsf{corr}(\mathsf{P}_z(B), \mathsf{P}_{z'}(B)) = \frac{\rho(z,z')}{\sqrt{\rho(z,z)}\sqrt{\rho(z',z')}} \tag{30}$$

where

$$\rho(z,z') = \mathbb{E}\left\{ \frac{\lambda(z)\lambda(z')}{1-[1-\lambda(z)][1-\lambda(z')]}. \right\}$$

When $\lambda(z) \sim \mathcal{LGP}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, *i.e.* it follows the logistic Gaussian process (29), the above expression reduces to

$$\rho(z,z') = \mathbb{E}\left\{ \frac{e^{\xi(z)+\xi(z')}}{e^{\xi(z)}+e^{\xi(z')}+e^{\xi(z)+\xi(z')}} \right\}.$$

Hence, returning to the dependent mixture model we can write

$$\mathsf{f}_z(y) = \lambda(z) \sum_{i=1}^{\infty} (1-\lambda(z))^{i-1}\; f(y \mid \mathsf{Z}_i), \tag{31}$$

The logistic transformation (29) ensures that $0 < \lambda(z) < 1$ as required for the geometric weights prior.

Following the alternative derivation of geometric weights priors of Section 2.1 we can again make use of the latent variable d_i to build up a Gibbs sampler algorithm, which is essentially based on the same full conditionals as in (40) but replacing (27) with the full conditional for $\xi^{(n)} = (\xi_1, \ldots, \xi_n)$, $\xi_i := \xi(z_i)$, which can be updated component-wise via

$$\mathbb{P}(\xi_i \mid \xi_{-i}) \propto \frac{1}{(1+e^{\xi_i})^{\mathsf{N}_i+1}}\, \mathrm{N}\left(\xi_i; \mu_i - \frac{1}{c_{ii}} \sum_{j\neq i} (z_j - \mu_j) c_{ij} + \frac{1}{c_{ii}}, \frac{1}{c_{ii}} \right)$$

where $\mu_i = \boldsymbol{\mu}(z_i)$ and c_{ij} is the ij-term of the precision matrix $\boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\sigma}(z_i, z_j); i, j = 1, \ldots, n\}$. The above density is log-concave and thus can be easily sampled via the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992).

Example 3. Consider a simulated data set with 61 observations coming from

$$Y_i = 0.2\, z_i^3 + \varepsilon_i$$

where $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 0.25)$ and $z = (-3, -2.9, \ldots, 2.9, 3)$. Hence assuming the dependent mixture model (31), with same kernel and prior guess at the shape specifications as in (12), one can implement a MCMC algorithm, based on full conditionals (40) with the above componentwise update of the logistic Gaussian process, to infer about any random functional of the form

$$\boldsymbol{\eta}_z(h) = \int_{\mathbb{Y}} h(y)\, \mathsf{f}_z(y) \mathsf{d}y.$$

This can be done through the Rao-Blackwellized MCMC estimator

$$\tilde{\boldsymbol{\eta}}_{z_i}(h) = \frac{1}{M} \sum_{l=1}^{M} \mathbb{E}_l[h(y) \mid z_i]$$

where M denotes the number of effective iterations in the MCMC. For example, one might be interested in the mean functional $\mathbb{E}_l[y \mid z_i] \approx \mu_{d_i}$, which in practice can be obtained as the updated mean value sampled from (40) in the Gibbs sampler.

Figure 6 shows the MC estimator for the distribution of the mean functional ($h(y) = y$) together with the observed data. For the corresponding Gaussian process we have set $\boldsymbol{\mu}(z) = -|z|$ and $\boldsymbol{\sigma}(z_i, z_j) = e^{-\phi||z_i - z_j||}$. □

Just as in the marginal $\mathcal{GWP}$ case, simpler weights with an infinite number of locations, seems to be enough to allocate the required mass at a particular location. In particular, having the first and largest weight for each covariate point, z, combined with the choice of locations seems enough for regression purposes.

3.2. A nonparametric diffusion process based on geometric weights priors. Although there are currently many examples of dependent nonparametric priors, the current available statistical literature devoted to the study of continuously dependent measures seems little. Indeed, it is of interest to study this case from a statistical perspective as it would allow us to exploit the flexibility of a Bayesian nonparametric approach while enjoying desirable properties such as Markovianity, reversibility, regularity of sample paths for the constructed model, etc.

These features are appealing in various modeling contexts and applications of stochastic processes such as finance or population genetics, where diffusion processes are typically used to model random phenomena evolving in time and structural properties as the aforementioned ones are essential.

In this section we present an approach to construct $\mathcal{P}_{\mathbb{X}}$-valued continuous time stochastic processes $(\mathsf{P}_t)_{t \geq 0}$ (Mena *et al.*, 2011) by simply introducing the time dependence through the weights, while keeping the locations fixed (but random) over time, in the species sampling representation for the $\mathcal{GWP}(a, b)$. That is we will consider

$$\mathsf{P}_t(B) = \lambda_t \sum_{i=1}^{\infty} (1 - \lambda_t)^{i-1}\, \delta_{\mathsf{Z}_i}(B), \qquad B \in \mathcal{X}. \tag{32}$$

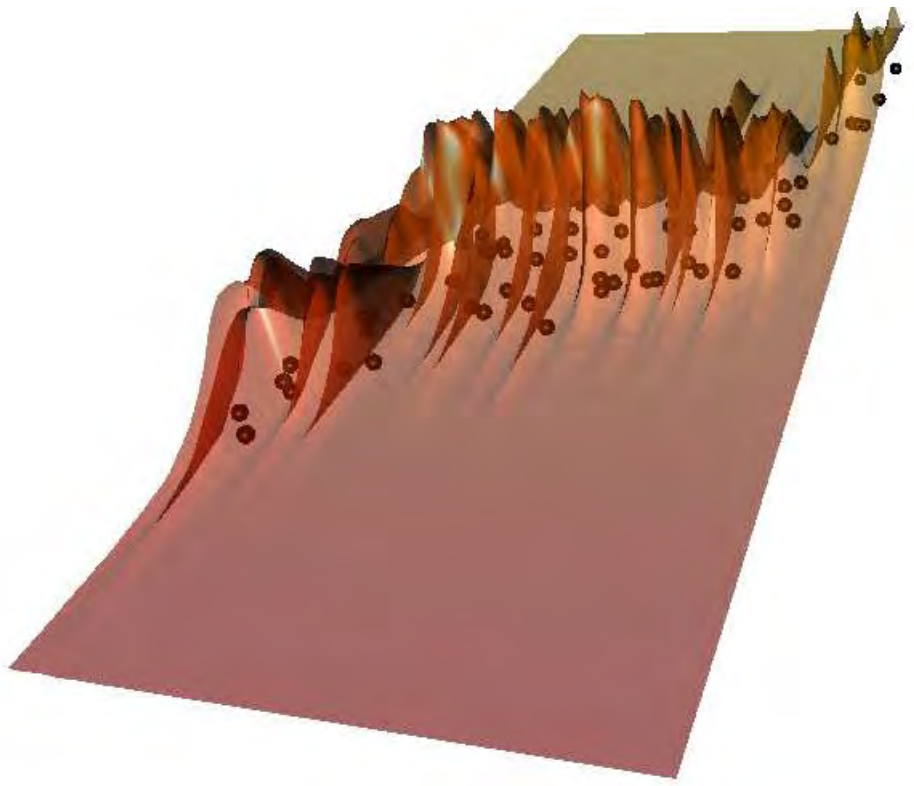

Figure 6: MC estimator for the density of $\eta_z(y)$ for simulated data set. The spheres represent the observed data and the surface the Rao-Blackwellized MC estimator for $\eta_z(y)$. The results are based on 10000 iterations of the Gibbs sampler algorithm.

with $\mathsf{Z}_i \overset{\text{iid}}{\sim} \mathsf{P}_0$ and $(\lambda_t)_{t\geq 0}$ a suitable stochastic process. In order to keep the same marginal r.p.m. $\mathcal{GWP}(a,b)$, we assume $(\lambda_t)_{t\geq 0}$ follows a strictly stationary diffusion process with $\text{Be}(a,b)$ invariant distributions. There are many choices for such a model, here we use a slight generalization of the well-known two-type Wright-Fisher diffusion process which can be described as the solution of the stochastic differential equation (SDE) on $[0,1]$ given by

$$\mathsf{d}\lambda_t = \left[\frac{c}{a+b-1}(a-(a+b)\lambda_t)\right]\mathsf{d}t + \sqrt{\frac{2c}{a+b-1}\lambda_t(1-\lambda_t)}\;\mathsf{d}B_t \tag{33}$$

where $(B_t)_{t\geq 0}$ is a standard Brownian motion. The typical parametrization of the two-type Wright-Fisher with mutation SDE is found when $c=(a+b-1)/2$.

Definition 2. (Mena *et al.*, 2011) A *geometric stick-breaking process* with parameters $a,b,c>0$ is a random process $(\mathsf{P}_t)_{t\geq 0}$ taking values in $\mathcal{P}_{\mathbb{X}}$ defined at each $t\geq 0$ by (32) with $(\lambda_t)_{t\geq 0}$ a two-type Wright-Fisher diffusion and P_0 a nonatomic probability measure on $(\mathbb{X},\mathcal{X})$. We denoted it as $\mathcal{GSB}(a,b,c,\mathsf{P}_0)$.

It can be seen that the Wright-Fisher diffusion is time reversible, strictly stationary with invariant measure $\text{Be}(a,b)$, so an immediate question is whether some of this properties are inherited by the $\mathcal{GSB}(a,b,c,\mathsf{P}_0)$ process.

Before undertaking this problem, first let us note that constructing stationary one dimensional diffusion processes with desired stationary distributions can be done via stochastic differential equations (see Bibby *et al.*, 2005), however this is not entirely useful when, for estimation purposes, one needs to keep track of an analytical form or an exact representation of the corresponding transition density. Here we use an idea to construct continuous time Markov processes (see Mena and Walker, 2009) that allows to have a representation of the transition density of the Wright-Fisher diffusion process. The idea starts with a Gibbs sampler Markov process based on the join density

$$f(y,x) = \text{Po}(y \mid \phi x)\,\text{Ga}(x \mid a,b),$$

form which we can construct a Markov process $(X_t)_{t\geq 0}$ with $\text{Ga}(a,b)$ marginals by the conditional updating $Y_t \mid X_0 \sim \text{Po}(\phi_t X_0)$ and $X_t \mid Y_t \sim \text{Ga}(a+Y_t, b+\phi_t)$. Indeed, such updating leads to the transition density

$$\begin{aligned} p(x_t \mid x_0) &= \sum_{y=0}^{\infty} \text{Ga}(x_t \mid a+y, b+\phi_t)\text{Po}(y \mid \phi_t x_0) \\ &= \frac{e^{-[\phi_t(x_t+x_0)+bx_t]}}{(\phi_t+b)^{-(a+1)/2}\,\phi_t^{(a-1)/2}}\left(\frac{x_t}{x_0}\right)^{\frac{a-1}{2}} \text{I}_{a-1}\left(2\sqrt{x_t\,x_0\phi_t(\phi_t+b)}\right), \end{aligned} \tag{34}$$

with $\phi_t := b(e^{ct}-1)^{-1}$. It can be verified that the above transition corresponds to the solution of a SDE given by

$$\mathsf{d}X_t = c\left(\frac{a}{b}-X_t\right)\mathsf{d}t + \sqrt{\frac{2c}{b}X_t}\,\mathsf{d}B_t$$

known as the Cox-Ingersoll-Ross model for interest rates (see Cox *et al.*, 1985). Therefore having a continuous-time Markov process with $\text{Ga}(a,b)$ invariant distribution suggests a simple transformation of two independent copies of these to obtain a diffusion with $\text{Be}(a,b)$ invariant densities. That is $\lambda_t = X_{1t}/(X_{1t}+X_{2t})$, where $(X_{it})_{t\geq 0}$, $i=1,2$ are independent Markov diffusion processes with $\text{Ga}(a,1)$ and $\text{Ga}(b,1)$ invariant densities respectively and

transition probabilities (34). In fact, it easily follows (see Mena and Walker, 2009) that the transition corresponding to this newly transformed process is given by

$$p(\lambda_t \mid \lambda_0) = \sum_{m=0}^{\infty} p_t(m) D(\lambda_t \mid m, \lambda_0) \tag{35}$$

where

$$p_t(m) = \frac{(a+b)_m \, e^{-m\,c\,t}}{m!}(1 - e^{-c\,t})^{a+b},$$

and

$$D(\lambda_t | m, \lambda_0) = \sum_{k=0}^{m} \mathrm{Be}(\lambda_t | a+k, b+m-k) \, \mathrm{Bin}(k | m, \lambda_0).$$

which again can be seen to correspond to the general class of Beta-binomial diffusion processes given by the solution of the SDE (33). Also, from the Gibbs sampler type construction, it easily follows that such a process is time reversible and has $\mathrm{Be}(a, b)$ stationary distributions.

Having established our approach to construct the diffusion process with $\mathrm{Be}(a, b)$ marginals, we can go back to our questions regarding the dependent nonparametric process of Definition 2. Let $\mathcal{P}_{\mathbb{X}}^{g} \subset \mathcal{P}_{\mathbb{X}}$ be the set of purely atomic probability measures with geometric weights as in Definition 1 and denote $\mathsf{C}_{\mathcal{P}_{\mathbb{X}}^{g}(\mathbb{X})}([0, \infty))$ the space of continuous functions from $[0, \infty)$ to $\mathcal{P}_{\mathbb{X}}^{g}$. Furthermore, for given locations $\mathsf{Z} = \{\mathsf{Z}_i\}_{i=1}^{\infty}$ define the continuous map $\Phi_{\mathsf{Z}}(\lambda) = \sum_{i=1}^{\infty} \lambda(1-\lambda)^{i-1}\delta_{\mathsf{Z}_i}$. Hence, due to the decreasing order of the geometric weights we can set $\Phi^{-1} = g_{\mathsf{Z}}$, where $g_{\mathsf{Z}}(\mathsf{P}) = \mathsf{P}(\{\mathsf{Z}_1\}) = \lambda$ and let $\tilde{B}(a, b) = \mathrm{Be}(a, b) \circ \Phi_{\mathsf{X}}^{-1}$. Therefore we can state the following proposition whose proof can be found in Mena *et al.* (2011)

Proposition 2. Let $(\mathsf{P}_t)_{t\geq 0}$ be a $\mathcal{GSB}(a, b, c, \mathsf{P}_0)$ process on $\mathcal{P}_{\mathbb{X}}^{g}$. Then $(\mathsf{P}_t)_{t\geq 0}$ is a reversible and stationary, with respect to $\tilde{B}(a, b)$, Feller process with sample paths in $\mathsf{C}_{\mathcal{P}_{\mathbb{X}}^{g}(\mathbb{X})}([0, \infty))$.

In other words the stability properties of the Wright-Fisher diffusion process are inherited by the $\mathcal{GSB}(a, b, c, \mathsf{P}_0)$ process at the probability measure-valued level. This is quite intuitive since the locations $(\mathsf{Z}_i)_{i\geq 1}$ are random but fixed across time. An immediate observation would be that letting the Z_i's also vary leads to a more flexible model. While this is certainly true, we have two reasons for keeping the locations fixed. On the probabilistic side this would most likely tear apart the nice properties this process enjoys. But more importantly, on the inference side this is not even needed, in view of Proposition 1 and as the example below will show.

All distributional properties of a diffusion process can be explained through its infinitesimal generator, in particular for the $\mathcal{GSB}(a, b, c, \mathsf{P}_0)$ process the generator can be found in Mena *et al.* (2011), thus providing with a valid alternative to other measure-valued processes found in the literature (cf. Ethier and Kurtz, 1986)

As in the previous section, interest might be on modeling a process taking values on the space of continuous densities. For this purpose the $\mathcal{GSB}(a, b, c, \mathsf{P}_0)$ process can also be incorporated into a dependent nonparametric mixture model given by

$$\mathsf{f}_t(y) = \int f(y \mid x) \, \mathsf{P}_t(\mathsf{d}x) = \lambda_t \sum_{l\geq 1} (1-\lambda_t)^{l-1} \, f(y \mid \mathsf{Z}_l) \tag{36}$$

with $\mathsf{Z}_l \overset{\text{iid}}{\sim} \mathsf{P}_0$. In fact, for a set of observations $\mathsf{Y}^{(n)}$ recordered at times $\{t_i\}_{i=1}^{n}$ and modeled through the above dependent density, Mena *et al.* (2011) proposed a Gibbs sampler algorithm

based on some slice sampler techniques that aid to overcome both the infinite summations, the one in (36) and the one in the representation of the Wright-Fisher transition density (35). The algorithm is a bit more demanding than those used in previous sections and so we briefly discuss it.

First, it is convenient to start by considering the part of the model related to the Wright-Fisher diffusion $(\lambda_t)_{t\geq 0}$. Hence to overcome the infinite summation in (36), we proceed as before and introduce the latent variable d_i that indicates form which component $f(\cdot \mid \mathsf{Z}_i)$ the observation Y_i comes from, namely we have the augmented observations $(t_i, d_i)_{i=1}^n$ and the model can then be written as

$$d_i \mid \lambda_i \sim \text{Geom}(\lambda_i)$$

with $\lambda_i := \lambda_{t_i}$ and corresponding transition density $p(\lambda_i \mid \lambda_{i-1})$ given as in (35), where t has to be replaced by $\boldsymbol{\tau}_i = t_i - t_{i-1}$. Hence, to avoid the infinite summations needed for the transition (35), we introduce a further set of latent variables $(u_i, s_i, k_i)_{i=1}^n$ whereby the augmented transition density is given by

$$p(\lambda_i, u_i, s_i, k_i \mid \lambda_{i-1}) = \mathbb{I}(u_i < g(s_i))\frac{p_i(s_i)}{g(s_i)}\text{Be}(\lambda_i \mid a + k_i, b + s_i - k_i)\text{Bin}(k_i \mid s_i, \lambda_{i-1})$$

where g is a decreasing function with known inverse. Therefore the likelihood with the complete data is given by

$$l(a, b, c) = \text{Be}(\lambda_0 \mid a, b)\prod_{i=1}^{n} p(\lambda_i, u_i, s_i, k_i \mid \lambda_{i-1})\lambda_i(1 - \lambda_i)^{s_i - 1}.$$

Hence if, for instance, we assume independent standard exponential distributions as priors for a, b, c, we see that the full conditionals, *e.g.* $\pi(a \mid b, c, \ldots) \propto l(a, b, c)e^{-a}$ are log-concave and easily sampled through the ARS algorithm. For instance we have

$$\log \pi(c \mid a, b, \cdots) = \sum_{i=1}^{n}\{(a + b)\log(1 - e^{-c\boldsymbol{\tau}_i}) - s_i\, c\, \boldsymbol{\tau}_i\} - c + O$$

where O is a constant which does not depend on c. The full conditionals for a and b follow similarly. The full conditional distribution for k_i is given by

$$\pi(k_i| \ldots) \propto \binom{s_i}{k_i}\frac{\mathbf{1}(k_i \in \{0, 1, \ldots, s_i\})}{\Gamma(a + k_i)\Gamma(b + s_i - k_i)}\left\{\frac{\lambda_i\lambda_{i-1}}{(1 - \lambda_i)(1 - \lambda_{i-1})}\right\}^{k_i}$$

which is clearly easy to sample since k_i can only take a finite number of values. The full conditional for u_i is simply a uniform distribution on $(0, g(s_i))$, where g is chosen at convenience, for example $g(s) = e^{-s}$ or $g(s) = s^{-2}$, so that g^{-1} is known. The benefit of this becomes apparent when we consider the full conditional for s_i. This is indeed given by

$$\pi(s_i| \ldots) \propto \frac{p_i(s_i)}{g(s_i)}\binom{s_i}{k_i}\frac{\Gamma(a + b + s_i)}{\Gamma(b + s_i - k_i)}\{(1 - \lambda_{i-1})(1 - \lambda_i)\}^{s_i}\,\mathbf{1}(k_i \leq s_i \leq g^{-1}(u_i))$$

which by virtue of u_i is restricted to a finite set. The full conditional for λ_i, for $i \neq 0, n$, is given by

$$\pi(\lambda_i| \ldots) = \text{Beta}(1 + a + k_i + k_{i+1}, d_i - 1 + b + s_i + s_{i+1} - k_i - k_{i+1}), \tag{37}$$

whereas

$$\pi(\lambda_0|\ldots) = \text{Beta}(a + k_1, b + s_1 - k_1) \tag{38}$$

and

$$\pi(\lambda_n|\ldots) = \text{Beta}(1 + a + k_n, d_n - 1 + b + s_n - k_n). \tag{39}$$

This deals with the part of the model related to the Wright-Fisher process. For the remaining part of the model, which for a given observation is given by

$$y_i|t_i, \lambda_i, \mathsf{Z}^{(n)} \sim \sum_{l=1}^{\infty} \lambda_i(1-\lambda_i)^{l-1} f(y_i|\mathsf{Z}_l),$$

we procede as before and introduce two latent variables (s_i, v_i) and a deterministic decreasing sequence of numbers (ψ_l) for which $\{l : \psi_l > v\}$ is a known set, such that

$$y_i, v_i, d_i|\lambda_i, \mathsf{Z}^{(n)} \sim \psi_{d_i}^{-1}\mathbf{1}(v_i < \psi_{d_i})\, \lambda_i(1-\lambda_i)^{d_i-1}\, f(y_i|\mathsf{Z}_{d_i}).$$

Namely, we slice the infinite summation again. In order to complete the Gibbs sampler for the model we need to describe how to sample the s_i from their full conditional and also the Z_s's. Now,

$$\pi(d_i|\ldots) \propto \psi_{d_i}^{-1}\lambda_i(1-\lambda_i)^{d_i-1}\, f(y_i|\mathsf{Z}_{d_i})\, \mathbf{1}(d_i \in \{l : \psi_l > v_i\})$$

and clearly the full conditional for v_i is the uniform distribution on $(0, \psi_{d_i})$. Since $\{l : \psi_l > v_i\}$ is a finite set this is easy to sample. Finally, we sample the Z_l's from

$$f(\mathsf{Z}_j \mid \cdots) \quad \propto \quad \mathsf{P}_0(\mathsf{Z}_j) \prod_{d_i=j} f(y_i \mid \mathsf{Z}_j), \qquad \text{for } j = 1, \ldots, M \tag{40}$$

Notice that, as before, we only need to consider a finite number of location updates $(\mathsf{Z}_l)_{l=1}^{M}$ where $M = \max_i M_i$ and $\{1, \ldots, M_i\} = \{l : \psi_l > v_i\}$. We thus have all the full conditional distributions required to implement the Gibbs sampler needed for the estimation of model (36) given a discretely observed trajectory. The following algorithm summarizes the procedure.

Algorithm

1. Select $g(\cdot)$ and $\psi(\cdot)$ functions, *e.g.* $g(x) = \psi(x) = e^{-x}$
2. Set initial values for:
   - Wright-Fisher diffusion parameters (a_0, b_0, c_0)
   - Parameters in the Kernel K and possibly in P_0, *e.g.* θ^0
   - Latent variables needed to overcome infinite summations, $(u_i^0, s_i^0, k_i^0, d_i^0)_{i=1}^n$. For these an initial value for the augmented random probability measure is also needed, e.g. $M^0 = 20$
   - Use these values to initiate $\lambda^0 = (\lambda_i^0)_{i=0}^n$

then for $j = 1, \ldots, I$

3. Update $v^j = (v_i^j)_{i=0}^n$, *i.e.* $v_i^j \sim \mathsf{U}[0, \psi_{s_i^{j-1}}]$, and compute $M^j = \max M_i^j$ with $\{1, \ldots, M_i^j\} = \{l : \psi(l) > v_i^j\}$

4. Update $\lambda^j = (\lambda_i^j)_{i=0}^n$, $\theta^j = (\theta_l^j)_{l=1}^M$ and $(u_i^j, s_i^j, k_i^j, d_i^j)_{i=1}^n$ using the corresponding full conditionals
5. Update (a_j, b_j, c_j), *e.g.* via ARS algorithm

The I iterations can then be used to build a Monte Carlo estimator for f_t of any desired functional of it.

Example 4. In order to illustrate how the modeling scheme described above is able to capture the dynamics of continuous time phenomena we will consider data coming from 251 daily observations (corresponding to a financial year) from the adjusted close quotations of the S&P 500 index during the period 03.03.2008 to 27.02.2009 (the data set can be found at `http://finance.yahoo.com`).

These type of data are typically modeled through parametric diffusion processes, however, one could argue to what extent such restrictive assumptions are justified. For example, in the case of interest rates one could choose among many existing models, such as the Cox-Ingersoll-Ross (CIR) diffusion, the Brennan-Schwartz diffusion or the Duffie-Kan diffusion (see Aït-Sahalia (1996)). Adopting a nonparametric approach based on measure-valued processes provides enough flexibility to avoid such committing assumptions.

As in our previous examples we use the kernel and prior guess at the shape specifications given by (12) and concentrate on the mean functional $\boldsymbol{\eta}_t := \int y\, \mathsf{f}_t(y)\mathsf{d}y$, namely the evolution of the mean, which imitate that of one-dimensional diffusion process. Figure 7 shows the MCMC estimate (heat contours) for the density process, $\hat{f}_t$, and the corresponding mean of the functional $\bar{\boldsymbol{\eta}}_t$ (solid line) for the S&P 500 data set (points). For both data sets the choice of hyper-parameters was $\tau = 1000$ and $(\alpha, \beta) = (10, 1)$. The results are based on 100,000 iterations, after a 20,000 of burn in (thinned each 10), enough to attain a satisfactory convergence of the sampler. Figure 8 shows the Markov chains corresponding to parameters (a, b) and the corresponding posterior densities, the results corresponding to c are proportional to those corresponding to b. A standard convergence analysis was performed, in particular the Gelman and Rubin (1992) visual test and the Raftery and Lewis (1992) diagnosis test were satisfactory. □

It is apparent that the probability measure-valued approach here undertaken is able to capture the dependence induced by these datasets. Furthermore the model adapts well to drastic changes like those observed in the S&P 500 index and typically not captured by (parametric) diffusion process. Note that the strict stationary and reversibility properties of the $\mathcal{GSB}(a, b, c, \mathsf{P}_0)$ process are at the probability measure-valued level and not at the level of observations, therefore certain sudden changes of regime and unstable behaviors might be well captured by this model without compromising the measure-valued stationarity property.

4. Discussion. We have seen that r.p.m.s with simpler weights structure are able to provide good alternatives to more complex nonparametric priors. The key idea is that having simpler weights results in a more efficient use of the infinite collection of locations to assign the required mass to a particular set $B \in \mathcal{X}$. Having simpler weights also result in simpler ways to estimate models and extend them to non-exchangeable contexts as seen in Section 3.

Although for simplicity we mainly concentrated in the geometric weights case, several generalizations are at issue, for instance one could consider the expected weights corresponding to the two-parameter Poisson-Dirichlet process, *i.e.* with corresponding stick-breaking

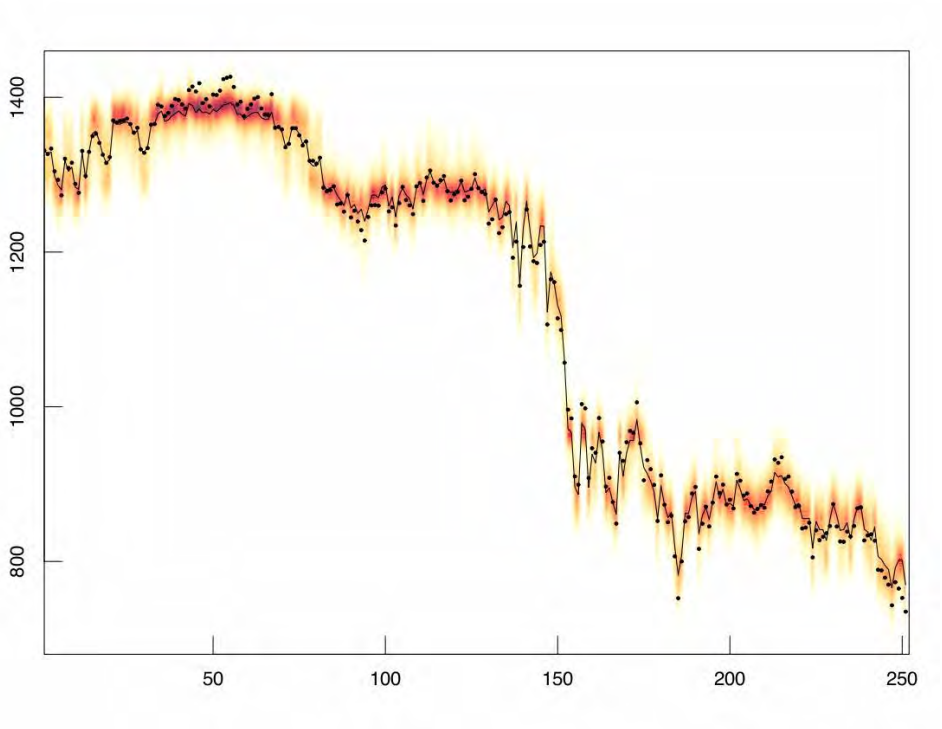

Figure 7: MCMC density estimator for the random density process (36), $\hat{f}_t$, (heat contour), mean of mean functional $\bar{\eta}_t$ (solid) for the S&P 500 data set (dots). The estimates are based on 10000 effective iterations, drawn from 100000 iterations thinned each 10, of the Gibbs sampler algorithm after 20000 iterations of burn in.

weights (4), with $\mathsf{v}_i \overset{\text{ind}}{\sim} \text{Be}(1-\sigma, \theta+i\sigma)$, which implies

$$\omega_i = \frac{1-\sigma}{\theta+1+(i-1)\sigma} \prod_{j=1}^{i-1} \frac{\theta+j\sigma}{\theta+1+(i-1)\sigma}$$

and then $(\sigma, \theta) \sim \pi$. Furthermore, using the alternative derivation in Section 2.1 we could have a different choice of prior for N, *e.g.* $\mathsf{N} \sim \text{Neg-Bin}(r, \lambda)$ which would lead to

$$w_l = \frac{1}{l} \binom{l+r-2}{r-1} \lambda^r (1-\lambda)^{l-1} \, {}_2F_1(1, l+r-1; l+1; \lambda), \tag{41}$$

where ${}_2F_1(a, b; c; \lambda)$ denotes the Gauss hypergeometric function. Both of the above possibilities are clearly more general than the geometric weights while keeping their decreasing feature, however some algorithmic modifications would be needed when randomizing their parameters.

Alternatively, one could also look for intersections with other general constructions, such as those via normalizations of known processes, as those would clearly open different ways to construct discrete r.p.m.s with decreasing weights.

We have also seen that from an numerical point of view having a simpler r.p.m. such as the $\mathcal{GWP}$ might lead to more efficient posterior inferences than those obtained via other

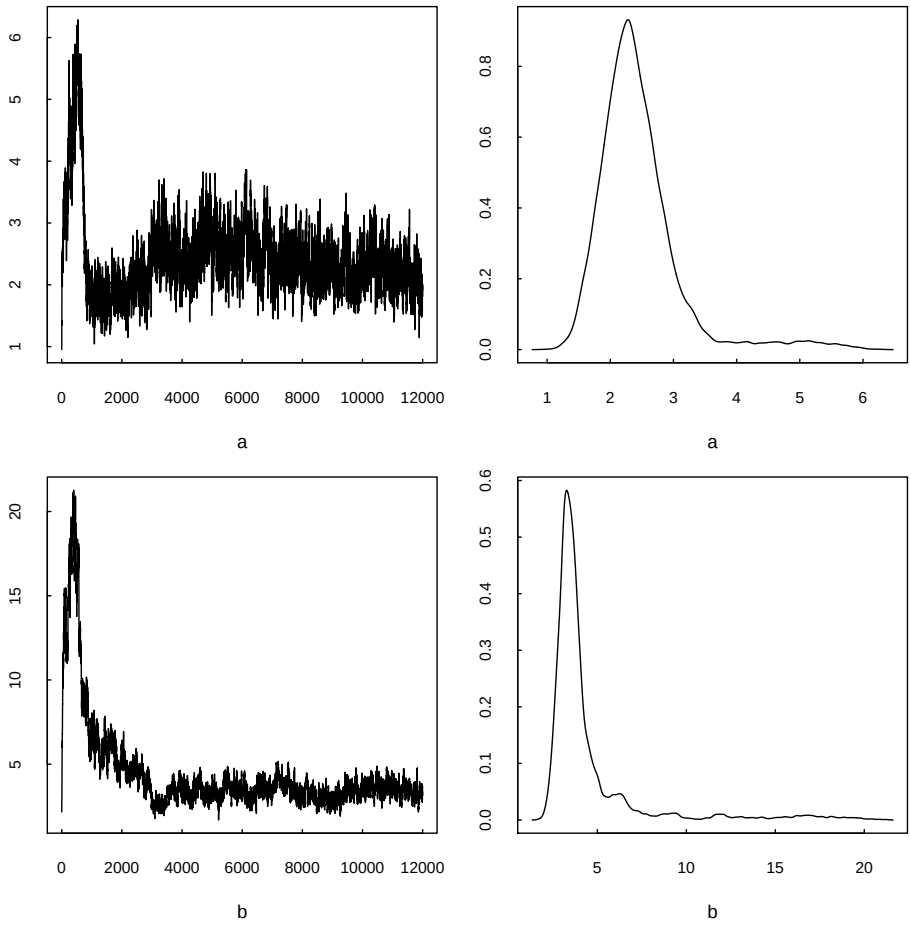

Figure 8: MCMC iterations and posterior densities for parameters (a, b). The estimates are based on 10000 effective iterations, drawn from 100000 iterations thinned each 10 iterations, of the Gibbs sampler algorithm after 20000 iterations of burn in.

nonparametric priors such as the Dirichlet process. However, some important points are at issue: as we mentioned before posterior inferences about the clustering behavior based in quantities such as (9) and (11) are different, and have a different interpretation, than that typically drawn from finite mixture models, in that several "components" from the corresponding infinite mixture model might be used to represent a group. Although this behavior is already seen from Dirichlet process mixtures applications, the observation becomes even stronger when a simple weight structure, such as the one featuring the $\mathcal{GWP}$, is considered, as the need of even more components to attain a particular mass for a given location would be required. Clearly the ultimate word of this kind of inferences, based on the $\mathcal{GWP}$, can be answered through its corresponding EPPF.

Another appealing feature of the simplicity inherent to the $\mathcal{GWP}$ is that it is easily generalized to dependent nonparametric processes. This is in principle conceivable with other more complicated r.p.m. such as the Dirichlet process, however keeping a canonical

construction of the resulting dependent process is not necessarily straightforward. Indeed further inspection of structural properties such as path-continuity, Markovianity, reversibility and other stability properties, is of interest in a number of applications where models need to be encompassed within general theories build upon these properties. The $\mathcal{P}_{\mathbb{X}}$-valued diffusion process presented in Section 3.2 gives an example of such kind of models, bridging the gap between the theory of superprocesses and statistics. In fact, to the best of our knowledge, Section 3.2 presents the first instance of a measure-valued Markovian process applied to model single trajectory phenomena (cf. Example 4). Also notice that the way inference and analyses are done in areas such as finance would be different when using random functionals, since we could speak of objects as the stochastic process driving the mean functional. This clearly constitutes a very appealing feature of probability measure-valued process and posses some challenges within these areas.

Acknowledgements. The present work was conducted while the author was visiting the "de Castro" Statistics Initiative at the Collegio Carlo Alberto. To both institutions goes his gratitude.

References.

Aït-Sahalia, Y., (1996). Nonparametric pricing of interest rate derivative securities. *Econometrica*, **64**, 527–560.

Antoniak, C. E. (1973). Mixture of Dirichlet processes with applications to Bayesian nonparametric problems . *Annals of Statistics*, **2**, 1152–1174.

Aoki, M. (2004) *Modeling Aggregate Behavior and Fluctuations in Economics: Stochastic Views of Interacting Agents* Cambridge University Press.

Banfield, J. D. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.

Berry, D. A. and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Annals of Statistics*, **7**, 558–568.

Bibby, M., Skovgaard, M. and Sørensen, M. (2005). Diffusion-type models with given marginal distribution and autocorrelation function. *Bernoulli*, **11**, 191–220.

Blackwell,D. (1973) Discreteness of Ferguson selections. *Annals of Statistics*, **1**, 356–358.

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

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2006). Bayesian inference for dynamic models with Dirichlet process mixtures. In *International Conference on Information Fusion.*, pp. 1–8. Florence, Italy: INRIA-CCSd-CNRS.

Cox, J. C., Ingersoll, J. E. and Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, **53**, 385–407.

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

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

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

Escobar, M.D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University.

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

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

Ethier, S. N. and Kurtz, T. G. (1986) *Markov processes: Characterization and Convergence.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc. New York

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

Ewens, W. J. and Tavarè, S. (1997). Multivariate Ewens distribution. In *Discrete Multivariate Distributions.*, (Johnson N.S., Kotz S., Balakrishnan N. Eds.), Wiley, New York.

Feigin, P. D. and Tweedie, R. L. (1989). Linear functionals and Markov chains associated with Dirichlet processes. *Mathematical Proceedings of the Cambridge Philosophical Society*, **105**, 579–585.

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

Fuentes-García, R., Mena, R. H. and Walker, S. G. (2009). A nonparametric dependent process for Bayesian regression. *Statisticsa and Probability Letters*, **8**, 112–1119.

Fuentes-García, R., Mena, R. H. and Walker, S. G. (2010). A new Bayesian nonparametric mixture model. *Communications in Statistics – Simulation and Computation*, **39**, 669–682.

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

Gelman, A. and Rubin, D. (1992). Inferences from iterative simulation using multiple sequences. *Statistical Inference*, **7**, 457–472.

Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

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

Griffin, J. E. and Steel, M. F. J. (2010). Stick-breaking autoregressive processes. *Journal of Econometrics*, **162**, 383–396.

Hansen, J.C (1994) Order statistics for decomposable combinatorial structures. *Random Structures and Algorithms*, **5**, 517-533.

Hartigan, J.A. (1975) *Clustering Algorithms.* John Wiley & Sons.

Hjort, N., Holmes, C., Müller, P. and Walker, S.G. (2010) *Bayesian nonparametrics.* Cambridge University Press.

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

Kalli, M., Griffin, J.E. and Walker, S.G. (2011). Slice sampling mixture models. *Statistics and Computing*, **21**, 93-105.

Lau, W.L. and Green, P. L. (2007) Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, **16**, 526-558.

Lenk, P. (1988) The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association*, **83**, 509-516.

Lijoi, A., Mena, R. H. and Prünster, I. (2005) Hierarchical mixture modelling with normalized inverse Gaussian priors. *Journal of the American Statistical Association*, **100**, 1278–1291

Lijoi, A., Mena, R. H. and Prünster, I. (2007a) Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, **94**, 769–786

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

Lijoi, A., Prünster, I. and Walker, S.G. (2008) Bayesian nonparametric estimators from conditional structures. *Annals of Applied Probability*, **18**, 1519–1547

Lijoi, A., Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*, (Hjort, N.L., Holmes, C.C., Müller, P., Walker, S.G. Eds.), Cambridge University Press.

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

Mena, R. H. and Walker, S.G. (2009). On a construction of Markov models in continuous time *METRON - International Journal of Statistics*, **LXVII**, 303–323.

Mena, R. H., Ruggiero, M. and Walker, S.G. (2011). Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling. *Journal of Statistical Planning and Inference* **141**, 3217–3230.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* **23**, 727–741.

MacEachern, S.N. (1998). Computational methods for mixture of Dirichlet process models. In *Practical non-parametric and semiparametric Bayesian statistics* (eds D. Dey, P. Müller and D. Sinha), 23–43. New York: Springer.

MacEachern, S.N. and Müller, P. (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

MacEachern, S.N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the section on Bayesian Statistical Science.*, pp. 50-5. Alexandria, VA: American Statistical Association.

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

Petrone, S. and Guindani, M. and Gelfand, A.E. (1995) Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society. Series B*,**71**, 755-782.

Pitman, J. (1995) Exchangeable and partially exchangeable random partitions. *Probability Theory Reled Fields.*,**102**, 145-158.

Pitman, J. (1996) Some developments of the Blackwell-MacQueen urn scheme. *Statistics, Probability and Game Theory. Papers in honor of David Blackwell.*,**30**, 245-267.

Pitman, J. (2006) *Combinatorial stochastic processes.* Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7-24, 2002. Lecture Notes in Math., Vol. **1875**. Springer, Berlin.

Raftery, A. and Lewis, S. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Inference*, **7**, 493–497.

Regazzini, E., Lijoi, A. and Prünster, I. (2003) Distributional results for means of random measures with independent increments. *Annals of Statistics*, **31**, 560–85.

Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

Rodriguez, A. and Ter Horst, E. (2008). Bayesian dynamic density estimation. *Bayesian Analysis*, **3**, 339-366.

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

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

Sethuraman, J. and Tiwari, R. (1982) Convergence of Dirichlet measures and the interpretation of their parameter. *Proc. Third Purdue Symp. Statist. Decision Theory and Related Topics* , (Gupta, S.S. and Berger, J. Eds.), Academic Press NY.

Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006) Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

Walker, S.G. (2007) Sampling the Dirichlet mixture model with slices. *Communications in Statistics – Simulation and Computation*, **36**, 45–54.

Walker, S.G., Damien, P., Laud, P.W. and Smith, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society. Series B*, **61**, 485–527.

Walker, S.G. and Mallick, B. (1997) A note of the scale parameter of the Dirichlet process. *The Canadian Journal of Statistics*, **25**, 473–479.