Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Theory and computations for the Dirichlet process and related models: An overview ☆

Alejandro Jara

Department of Statistics, Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, Chile

ARTICLE INFO

Article history:
Received 3 March 2016
Received in revised form 7 November 2016
Accepted 9 November 2016
Available online xxxx

Keywords:
Random probability distributions
Hierarchical model
Density estimation
Conditional density estimation

ABSTRACT

Data analysis sometimes requires the relaxation of parametric assumptions in order to gain modeling flexibility and robustness against mis-specification of the probability model. In the Bayesian context, this is accomplished by placing a prior distribution on an infinite-dimensional space, referred to as Bayesian nonparametric models. We provide an overview on the most popular Bayesian nonparametric models for probability distributions and for collections of predictor-dependent probability distributions. The intention of is not to be complete or exhaustive, but rather to touch on areas of interest for the practical use of the priors in the context of a hierarchical model. We give an overview covering the main properties of the basic models and the algorithms for fitting them.

1. Introduction

The definition and study of properties of probability models defined over infinite-dimensional spaces have received increasing attention in the statistical literature because they are the basis for the specification of Bayesian nonparametric (BNP) statistical models [1]. BNP models allow the modeler to gain model flexibility and robustness against mis-specification of a parametric statistical model. They are constructed by defining a stochastic process whose trajectories lie in the infinite-dimensional space of interest, such as the space of all probability measures defined on a given measurable space. The law governing such a process is then used as a prior distribution for the infinite-dimensional parameter in a Bayesian framework.

BNP statistics is a relative new area of statistics. The intersection of Bayesian and nonparametric statistics was almost empty until the sixties and seventies where the first advances were made, primarily on the mathematical formulations. It was only in the early nineties with the advent of sampling based methods, in particular Markov chain Monte Carlo (MCMC) methods, that substantial progress has been made in the area. Posterior distributions over infinite-dimensional spaces are highly complex and hence sampling methods play a key role.

The increase in applications of BNP methods in the statistical literature has been motivated largely by the availability of simple and efficient methods for posterior computation in Dirichlet process mixture (DPM) models [2]. The DPM models incorporate Dirichlet process (DP) priors [3,4] for components in Bayesian hierarchical models, resulting in an extremely flexible class of models. Due to the flexibility and ease in implementation, DPM models are now routinely implemented in a wide variety of applications [1]. Furthermore, a rich theoretical literature showing large support of the prior and an

☆ A. Jara was supported by FONDECYT 1141193 grant.
E-mail address: atjara@uc.cl.
URL: http://www.mat.uc.cl/~ajara.

http://dx.doi.org/10.1016/j.ijar.2016.11.008
0888-613X/

adequate asymptotic behavior of the posterior distribution justify the use of DPM models for the inference on probability measures [5–8].

In this paper, basic BNP methods for probability measures are reviewed. The literature is too vast to attempt even a moderate review, and we instead refer the interested reader to [9–11,1], and references therein, for more general overviews. In this paper we focused mainly on computational aspects. Various relevant priors for probability distributions and collections of predictor-dependent probability measures are reviewed in Sections 2 and 3, respectively. The main computational algorithms are discussed in Section 4. A final discussion section concludes the paper.

2. BNP priors for single probability distributions

2.1. Dirichlet process

The DP was introduced by Ferguson [3] as a prior on the space of probability measures on a given measurable space $(\boldsymbol{\Theta}, \mathbb{B}(\boldsymbol{\Theta}))$, where $\boldsymbol{\Theta}$ is a complete and separable space and $\mathbb{B}(\boldsymbol{\Theta})$ is a σ-field on $\boldsymbol{\Theta}$. Ferguson covered many of their basic properties and applied them to a variety of nonparametric estimation problems, thus providing the first time a Bayesian interpretation for some of the commonly used nonparametric procedures.

Definition 1. Let $M > 0$ and G_0 be a probability measure on $(\boldsymbol{\Theta}, \mathbb{B}(\boldsymbol{\Theta}))$. A DP with parameters (M, G_0) is a random probability measure G on $(\boldsymbol{\Theta}, \mathbb{B}(\boldsymbol{\Theta}))$, which assigns probability $G(B)$ to every (measurable) set $B \in \mathbb{B}(\boldsymbol{\Theta})$ such that for each measurable finite partition $\{B_1, \ldots, B_k\}$ of $\boldsymbol{\Theta}$, the joint distribution of the vector $(G(B_1), \ldots, G(B_k))$ is the Dirichlet distribution with parameters

$$(MG_0(B_1), \ldots, MG_0(B_k)).$$

Using the Kolmogorov's consistency theorem, Ferguson showed the existence of the DP. An alternative proof of the existence of the process is given by Blackwell [12]. The process is usually denoted as $\mathrm{DP}(MG_0)$, where M is the precision parameter and G_0 is a center probability measure. The product MG_0 is usually referred to as the base measure of the DP.

The random probability measure G is uniquely defined by its specified finite-dimensional distributions. The DP arises naturally as an infinite-dimensional analogue of the finite-dimensional Dirichlet distribution, which in turn has its roots in the one-dimensional Beta distribution. In this way, most of the basic properties of DP arise as an extension of the properties of the Dirichlet distribution. As a matter of fact, for every $B \in \mathbb{B}(\boldsymbol{\Theta})$, it follows that

(a) $G_0(B) = 0 \Longrightarrow \Pr(G(B) = 0) = 1$, and $G_0(B) > 0 \Longrightarrow \Pr(G(B) > 0) = 1$.
(b) $E(G(B)) = G_0(B)$.
(c) $Var(G(B)) = \frac{G_0(B)(1-G_0(B))}{1+M}$.

These properties easily follow by the observation that each $G(B)$ is a beta random variable with parameters $MG_0(B)$ and $M(1 - G_0(B))$. These results show the effect of the precision parameter in a DP. If M is large, G is highly concentrated around G_0, thus justifying the terminology.

Ferguson also showed the full weak support of the DP. Full support is a minimum requirement and almost a "necessary" property for a Bayesian model to be considered "nonparametric". In this context, the full support means that the prior probability model assigns positive mass to any neighborhood of every probability measure on a given space. Therefore, the definition of support strongly depends on the choice of a "distance" defining the neighborhoods: for the weak support the neighborhoods are defined using a metric inducing the weak topology.

A DP is almost surely (a.s.) discrete. Ferguson [3] proved this important property by using a gamma process representation of it. Blackwell [12] and Blackwell and MacQueen [13] gave alternative proofs. In the latter case, the DP arises as the mixing measure in de Finetti's representation theorem [14], of the following continuous analogue of the Polya urn scheme:

$$\boldsymbol{\theta}_1 \mid G_0 \sim G_0,$$

and, for $i = 2, 3, \ldots,$

$$\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_{i-1}, M, G_0 \sim \frac{1}{M+i-1} \sum_{j=1}^{i-1} \delta_{\boldsymbol{\theta}_j}(\cdot) + \frac{M}{M+i-1} G_0(\cdot),$$

where $\delta_{\boldsymbol{\theta}}$ is the Dirac measure on $(\boldsymbol{\Theta}, \mathbb{B}(\boldsymbol{\Theta}))$ giving mass one to the point $\boldsymbol{\theta}$. This representation is extremely crucial for MCMC sampling from a DP. The representation also shows that ties are expected among $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$. The expected number of distinct $\boldsymbol{\theta}$'s is, as $n \to \infty$, $M \log \frac{n}{M}$ [15], which is asymptotically much smaller than n. Ferguson further demonstrates that if $\boldsymbol{\theta}_1 \mid G \sim G$, then marginally $G(\boldsymbol{\theta}_1 \in B) = G_0(B)$.

The DP enjoys an important conjugacy property summarized next.

Please cite this article in press as: A. Jara, Theory and computations for the Dirichlet process and related models: An overview, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.11.008

Theorem 1. (Ferguson, 1973) *Let* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \mid G \overset{i.i.d.}{\sim} G$ *and* $G \mid M, G_0 \sim \mathrm{DP}(MG_0)$. *Then,*

$$G \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n, M, G_0 \sim \mathrm{DP}\left(MG_0 + \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}\right).$$

An immediate corollary arising from this theorem which is used often in MCMC developments is the following.

Corollary 1. *Let* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n, \boldsymbol{\theta}_{n+1} \mid G \overset{i.i.d.}{\sim} G$ *and* $G \mid M, G_0 \sim \mathrm{DP}(MG_0)$. *Then*

$$\boldsymbol{\theta}_{n+1} \mid M, G_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim \frac{M}{M+n} G_0 + \frac{1}{M+n} \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}.$$

Sethuraman [16] provides an extremely useful alternative representation of the DP.

Theorem 2. (Sethuraman, 1994) *Let* $V_i \mid M \overset{i.i.d.}{\sim} \mathit{Beta}(1, M)$ *and* $\boldsymbol{\theta}_i \mid G_0 \overset{i.i.d.}{\sim} G_0$, $i = 1, 2, \ldots$. *If*

$$G = \sum_{i=1}^{\infty} W_i \delta_{\boldsymbol{\theta}_i},$$

where $W_1 = V_1$ *and, for* $i = 2, \ldots$, $W_i = V_i \prod_{j=1}^{i-1}(1 - V_j)$, *then* G *is a* $\mathrm{DP}(MG_0)$

Note that this theorem immediately implies that G is discrete with probability one. It may be noted that the masses W_i's are obtained by successive "stick-breaking" with $V_1, V_2, \ldots$ as the corresponding stick-breaking proportions, and allocated to randomly chosen atoms $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ generated from G_0. A consequence of this representation is that if $G \sim \mathrm{DP}(MG_0)$, $\boldsymbol{\theta} \sim G_0$ and $W \sim \mathrm{Beta}(1, M)$, and all of them are independent, then $W\delta_{\boldsymbol{\theta}}(\cdot) + (1 - W)G(\cdot)$ is again a $\mathrm{DP}(MG_0)$. This property leads to important distributional equations for functionals of the DP and could also be used to simulate a Markov chain on the space of the probability distribution with $\mathrm{DP}(MG_0)$ as its stationary distribution [17].

Finally, the DP has a very important conditioning property.

Theorem 3. *If* $A \in \mathbb{B}(\boldsymbol{\Theta})$ *with* $G_0(A) > 0$, *then the random measure* $G\,|_A$, *the restriction of* G *to* A *defined by* $G\,|_A = G(B \mid A) = G(A \cap B)/G(A)$, *is a* DP *with parameters* M *and* $G_0\,|_A$, *and is independent of* $G(A)$.

The argument can be extended to more than one set. Thus the DP locally splits into numerous independent DPs.

2.2. Processes derived from a Dirichlet process

2.2.1. Mixture of Dirichlet processes

DP requires the specification of the base measure $\gamma = MG_0$. Assigning a single base measure to a DP may be difficult or limiting. Antoniak [18] mitigates this problem by considering mixtures of Dirichlet processes (MDP). MDP also arise in censored data problems and hierarchical models in which one level is modeled as a DP.

Definition 2. Let $(\boldsymbol{\Theta}, \mathbb{B}(\boldsymbol{\Theta}))$ be a measurable space and $(U, \mathcal{B}, P)$ a probability space. Let the transition measure $\gamma : U \times \mathcal{X} \mapsto [0, \infty)$ be such that:

(a) $\forall\ u \in U$, $\gamma(u, \bullet)$ is a finite, non-null measure on $(\boldsymbol{\Theta}, \mathbb{B}(\boldsymbol{\Theta}))$, and
(b) $\forall\ A \in \mathbb{B}(\boldsymbol{\Theta})$, $\gamma(\bullet, A)$ is measurable on $(U, \mathcal{B})$.

G is a mixture of Dirichlet processes on $(\boldsymbol{\Theta}, \mathbb{B}(\boldsymbol{\Theta}))$ if for all measurable partitions $\{B_j\}_{j=1}^{k}$ of $\boldsymbol{\Theta}$ we have

$$\mathcal{P}\left(G(B_1) \le y_1, \ldots, G(B_k) \le y_k\right) = \int_U D\left(y_1, \ldots, y_k \mid \gamma(u, B_1), \ldots, \gamma(u, B_k)\right) P(du),$$

where $D(y_1, \ldots, y_k \mid a_1, \ldots, a_k)$ is the CDF of a Dirichlet distribution with parameter $(a_1, \ldots, a_k)$.

We use the notation $G \sim \int \mathrm{DP}(\gamma(u))P(du)$. The main idea behind the MDP is to allow the parameter of the DP γ to be random as well, thus creating a hierarchical model: $u \sim P$, and $G \mid u \sim \mathrm{DP}(\gamma(u))$. In practice, one may propose a parametric family as the base measure and put a hyper-prior on the parameters of that family. The resulting procedure has an intuitive

appeal in that if one is a weak believer in a parametric family, then instead of using a parametric analysis, one may use the corresponding MDP to "robustify" the parametric procedure.

The following lemma due to Antoniak [18] is often used for MCMC sampling schemes.

Lemma 1. (Antoniak, 1974) *Assume that* $G \sim \int \mathrm{DP}(MG_{\boldsymbol{\eta}})P(dM, d\boldsymbol{\eta})$ *and* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \mid G \overset{i.i.d.}{\sim} G$, *where* $G_{\boldsymbol{\eta}}$ *is nonatomic. Then,*

$$P(dM, d\boldsymbol{\eta} \mid \boldsymbol{\theta}_{1:n}) \propto \left\{ \prod_{j=1}^{d(\boldsymbol{\theta}_{1:n})} g_{\boldsymbol{\eta}}(\boldsymbol{\theta}_j^*) \right\} \frac{M^{d(\boldsymbol{\theta}_{1:n})}\Gamma(M)}{\Gamma(M+n)} P(dM, d\boldsymbol{\eta}),$$

where $\boldsymbol{\theta}_{1:n} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, $g_{\boldsymbol{\eta}}$ *is the density of* $G_{\boldsymbol{\eta}}$ *w.r.t. Lebesgue measure,* $\Gamma(\cdot)$ *is the gamma function, and* $d(\boldsymbol{\theta}_{1:n})$ *denotes the number of distinct values of* $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$, *denoted by the set* $\{\boldsymbol{\theta}_j^*\}_{j=1}^{d(\boldsymbol{\theta}_{1:n})}$.

2.2.2. Dirichlet process mixture models

As described in Section 1, in DPM models the discrete nature of a DP is exploited to define a mixture model with infinitely many components of some simple parametric form $f(\cdot \mid \boldsymbol{\theta})$. This approach is commonly employed to induce a BNP prior for densities. Interestingly, if the first stage of in a hierarchical model is given by a parametric distribution and the second stage is modeled with a DP prior, then the posterior distribution of the process is a MDP.

Theorem 4. (Antoniak, 1974) *If* $X_i \mid \boldsymbol{\theta}_i \overset{ind.}{\sim} f(\cdot \mid \boldsymbol{\theta}_i)$, $i = 1, \ldots, n$, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \mid G \overset{i.i.d.}{\sim} G$, *and* $G \mid M, G_0 \sim \mathrm{DP}(MG_0)$. *Then*

$$G \mid \boldsymbol{X}_{1:n} \sim \int \mathrm{DP}(MG_0 + \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}) P(d\boldsymbol{\theta}_{1:n} \mid \boldsymbol{X}_{1:n}),$$

where $P(d\boldsymbol{\theta}_{1:n} \mid \boldsymbol{X}_{1:n})$ *is the posterior distribution arising from* $X_i \mid \boldsymbol{\theta}_i \overset{ind.}{\sim} f(\cdot \mid \boldsymbol{\theta}_i)$ *and* $\boldsymbol{\theta}_i \mid G_0 \overset{i.i.d.}{\sim} G_0$, $i = 1, \ldots, n$.

Because of this result, DPM were originally called mixtures of DPs. Because models are not usually named for the properties of their posterior distribution, this terminology has been avoided in the more recent literature.

The choice of the appropriate parametric distribution depends on the underlying sample space. If the underlying density function is defined on the entire real line, a location-scale kernel is appropriate. On the unit interval, beta distributions form a flexible family and provide interesting connections with approximations based on Bernstein polynomials [19,20]. On the positive half line, mixtures of gamma, Weibull or lognormal may be used. On a m-dimensional simplex space

$$\Delta_m = \left\{ (y_1, \ldots, y_m) \in [0,1]^m : \sum_{i=1}^{m} y_i \leq 1 \right\},$$

Dirichlet distributions form a rich family [21].

2.2.3. Other models derived from a Dirichlet process

Invariant DPs (IDP) were considered by Dalal [22]. Suppose that a prior on the space of the probability measures symmetric around zero needs to be specified. With this aim, an alternative is to consider $G \mid M, G_0 \sim \mathrm{DP}(MG_0)$ and let $\overline{G}(B) = (G(B) + G(-B))/2$, where $-B = \{x : -x \in B\}$. Another way of randomly generating symmetric probabilities is to consider a Dirichlet process G on $[0, \infty)$ and unfold it to $\overline{G}$ on $\mathbb{R}$ by defining $\overline{G}(B) = \overline{G}(-B) = \frac{1}{2}G(B)$. Those techniques are particularly helpful for constructing priors on the error distribution in regression models. The model parameters are not identified without some restriction on G, and symmetry around zero may be considered a reasonable condition on G ensuring identification of the model parameters.

Conditional DPs (CDP) arise when it is necessary to constrain the support of the DP, for example, to the set of all probability distribution with a given quantile. If $\{B_1, \ldots, B_k\}$ is a finite partition then the conditional distribution of G given $\{G(B_j) = w_j, j = 1, \ldots, k\}$, where G is a $\mathrm{DP}(MG_0)$ and $w_j \geq 0$, $\sum_{j=1}^{k} w_j = 1$, is called a conditional DP. By the conditioning property of the DP mentioned before, it follows that the above process may be written as $G = \sum_{j=1}^{k} w_j G_j$, where each G_j is a DP on B_j. Consequently G is a countable mixture of DP with "orthogonal" supports. A particular case of CDP is the Doss' median-0 DP [23,24]. Another variant of the CDP can be found in [25], including applications to binary regression models.

2.3. Generalizations of the Dirichlet process

2.3.1. Tail-free and Polya trees

The concept of tail-free (TF) process was introduced by Freedman [26] and Fabius [27], and chronologically precedes that of the DP. Assume that $\mathbb{R}$ is the sample space. Let $E = \{0, 1\}$ and E^m be the m-fold Cartesian product $E \times \ldots \times E$, where $E^0 = \emptyset$. Further, set $E^* = \bigcup_{m=0}^{\infty} E^m$. A TF process is defined by allocations of random probabilities to sets in a nested

sequence of partitions of the sample space. Let $\pi_0 = \{\mathbb{R}\}$, $\pi_1 = \{B_0, B_1\}$, $\pi_2 = \{B_{00}, B_{01}, B_{10}, B_{11}\}, \ldots$, be a sequence of partitions of $\mathbb{R}$, such that for every $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in E^*$, with $\varepsilon_i \in E$, $B_\varepsilon = B_{\varepsilon 0} \cup B_{\varepsilon 1}$ and $B_{\varepsilon 0} \cap B_{\varepsilon 1} = \emptyset$. Assume that $w_1 < w_2$ for every $w_1 \in B_{\epsilon 0}$ and for every $w_2 \in B_{\epsilon 1}$, and that B_ε is a left open right closed interval unless ε is a string of one's only. Further assume that $\bigcup_{m=0}^{\infty} \pi_m$ is a generator for the Borel σ-field on $\mathbb{R}$. Note that this is ensured if the collection of right end points of B_ε is dense in $\mathbb{R}$. A probability G may then be described by specifying all the conditional probabilities $\{Y_\varepsilon = G(B_{\varepsilon 0} \mid B_\varepsilon) : \varepsilon \in E^*\}$. A prior of G may thus be defined by specifying the joint distributions of all Y_ε's. The specification may be written in a tree form. The different hierarchy in the tree signifies prior specification of different levels. A prior for G is said to be tail-free with respect to the sequences of partitions $\{\pi_m\}_0^\infty$ if the collections $\{Y_\emptyset\}, \{Y_0, Y_1\}, \{Y_{00}, Y_{01}, Y_{10}, Y_{11}\}, \ldots$, are mutually independent. Note that, variables within the same hierarchy need not be independent; only the variables at different levels are required to be so.

The concept of TF admits the DP as an important special case. The DP is TF with respect to any sequence of partitions. Indeed, the DP is the only prior that has this distinguished property. A Polya tree (PT) is also a special case of a TF process, where besides across row independence, the random conditional probabilities are also independent within row and have beta distributions. The concept of PT was originally considered by Ferguson [4], and later studied thoroughly by Mauldin et al. [28], and Lavine [29,30]. A practical discussion on the use of the model can be found in Christensen et al. [31]. The prior can be seen as the de Finetti measure in a generalized Polya urn scheme [28]. The connection with Polya urn schemes justifies the name of PT [32] and allows elegant proofs of many of their properties.

Definition 3. (Lavine, 1992) Let $\{\pi_m\}$ be a sequence of binary partitions as before and $\{\alpha_\varepsilon : \varepsilon \in E^*\}$ be a collection of nonnegative numbers. A random probability measure G on Ω is said to be a PT with parameters $(\{\pi_m\}, \{\alpha_\varepsilon : \varepsilon \in E^*\})$, if there exist a collection $\mathcal{Y} = \{Y_\varepsilon : \varepsilon \in E^*\}$ of random variables such that the following hold:

(i) The collection $\mathcal{Y}$ consists of mutually independent random variables,
(ii) For each $\varepsilon \in E^*$, Y_ε has a beta distribution with parameters $\alpha_{\varepsilon 0}$ and $\alpha_{\varepsilon 1}$,
(iii) The random probability measure G is related to $\mathcal{Y}$ through the relations

$$G\left(B_{\varepsilon_1 \ldots \varepsilon_m}\right) = \left(\prod_{j=1, \varepsilon_j = 0}^{m} Y_{\varepsilon_1 \ldots \varepsilon_{j-1}}\right) \left(\prod_{j=1, \varepsilon_j = 1}^{m} Y_{\varepsilon_1 \ldots \varepsilon_{j-1}}\right), \; m = 1, 2, \ldots,$$

where the factors are $Y_\emptyset$ or $1 - Y_\emptyset$ if $j = 1$.

Note that from the TF property and the beta distribution of the Y_ε's it is not difficult to find expressions for the moments of probabilities of sets in the partition under a PT. For example, the mean and the variance are given by

$$E\left(G(B_{\varepsilon_1 \cdots \varepsilon_m})\right) = \prod_{j=1}^{m} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_j}}{\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 1}},$$

and

$$Var\left(G(B_{\varepsilon_1 \cdots \varepsilon_m})\right) = \prod_{j=1}^{m} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_j}(\alpha_{\varepsilon_1 \cdots \varepsilon_j} + 1)}{(\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 1})(\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 1} + 1)} - \prod_{j=1}^{m} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_j}^2}{(\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1} 1})^2},$$

respectively. PT also enjoy a conjugacy result summarized in the following theorem.

Theorem 5. *Let* $X_1, \ldots, X_n \mid G \overset{iid}{\sim} G$, *and*

$$G \mid \{\pi_m\}, \{\alpha_\varepsilon : \varepsilon \in E^*\} \sim \mathrm{PT}\left(\{\pi_m\}, \{\alpha_\varepsilon : \varepsilon \in E^*\}\right).$$

Then,

$$G \mid \boldsymbol{X}_{1:n}, \{\pi_m\}, \{\alpha_\varepsilon : \varepsilon \in E^*\} \sim \mathrm{PT}\left(\{\pi_m\}, \{\alpha_\varepsilon^* : \varepsilon \in E^*\}\right),$$

where $\{\alpha_\varepsilon^* = \alpha_\varepsilon + n_\varepsilon : \varepsilon \in E^*\}$, *and* n_ε *is the number of* $X_1, \ldots, X_n$ *in* B_ε.

The class of PT contains all DP, characterized by the relation that $\alpha_{\varepsilon 0} + \alpha_{\varepsilon 1} = \alpha_\varepsilon$ for all ε [4]. The advantage of a DP over a more general PT is that DP are the only TF processes in which the choice of $\{\pi_m\}$ does not affect inference. PT have an advantage in that conditions can be set on $\{\alpha_\varepsilon : \varepsilon \in E^*\}$ such that G is absolutely continuous with probability one [4,33–35].

Please cite this article in press as: A. Jara, Theory and computations for the Dirichlet process and related models: An overview, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.11.008

Theorem 1.121 and Lemma 1.124 (pp. 66–68) in Schervish [36] provide results for general TF processes, and a simple tool to evaluate the continuity of a PT.

Lavine [29] provides the marginal predictive density with respect to a dominating measure λ for an observation $X \mid G \sim G$, denoted by g_λ, where $G \mid \{\pi_m\}, \{\alpha_\varepsilon : \varepsilon \in E^*\} \sim \mathrm{PT}(\{\pi_m\}, \{\alpha_\varepsilon : \varepsilon \in E^*\})$.

Theorem 6. (Lavine, 1992) *For $x \in \Omega$, let $\varepsilon_1, \varepsilon_2, \ldots$ be the infinite sequence of 0's and 1's such that $x \in B_{\varepsilon_1 \cdots \varepsilon_m}$ for all $m = 1, 2, \ldots$. Then,*

$$g_\lambda(x \mid \{\pi_m\}, \{\alpha_\varepsilon : \varepsilon \in E^*\}) = \lim_{m\to\infty} \frac{\mathcal{P}\left(X \in B_{\varepsilon_1 \cdots \varepsilon_m}\right)}{\lambda\left(B_{\varepsilon_1 \cdots \varepsilon_m}\right)},$$
$$= \lim_{m\to\infty} \frac{\prod_{j=1}^{m} \frac{\alpha_{\varepsilon_1 \cdots \varepsilon_j}}{\alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}0} + \alpha_{\varepsilon_1 \cdots \varepsilon_{j-1}1}}}{\lambda\left(B_{\varepsilon_1 \cdots \varepsilon_m}\right)},$$

where the first equality holds for λ-almost all x.

These results show that when λ is the Lebesgue measure, a PT prior with appropriate choice of the parameters may be used for density estimation and the posterior mean can be computed analytically.

In practice, it would be difficult to elicit the family $\{\alpha_\varepsilon : \varepsilon \in E^*\}$ and the partition $\{\pi_m\}$. Lavine [29] provides a canonical construction of the PT prior on $\Omega \subset \mathbb{R}$. Let G_0 be a probability measure on Ω. Then, [29] proposed to center the PT around G_0, with CDF $G_0(t)$, by taking each level m of the partition $\{\pi_m\}$ to coincide with quantiles $G_0(k/2^m)$, $k = 0, 1, \ldots 2^m$, and further taking $\alpha_{\varepsilon 0} = \alpha_{\varepsilon 1}$ for all $\varepsilon \in E^*$. A given $\varepsilon = \varepsilon_1 \cdots \varepsilon_m \in E^m$ can be thought of as a base-2 representation of a decimal number N. If $B_{\varepsilon_1 \cdots \varepsilon_m}$ is defined to be the interval $\left(G_0^{-1}\{N/2^m\}, G_0^{-1}\{(N+1)/2^m\}\right)$, then $E\left(G(B_{\varepsilon_1 \cdots \varepsilon_m})\right) = \prod_{j=1}^{m} E(Y_{\varepsilon_1 \cdots \varepsilon_j}) = G_0(B_{\varepsilon_1 \cdots \varepsilon_m})$. Thus G_0 have a similar role that the center measure of a DP.

Once the PT is centered around a probability measure, G_0, the family $\{\alpha_\varepsilon : \varepsilon \in E^*\}$ further determines how much G can "deviate" from G_0, much like the role of the precision parameter M in a DP, and thus how quickly a random sample "takes over" the centering distribution G_0. Besides, the PT will have infinitely many more parameters which may be used to describe the prior belief. To avoid specifying too many parameters, a default method is adopted, where one chooses α_ε as a function depending on the length of the string ε, ρ. Recall that the choice of α_ε directly controls the type of trajectories generated by the process. Lavine [29] considered $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = \rho(m) = m^2$ is a "sensible canonical choice". Walker and Mallick [37] considered $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = \rho(m) = cm^2$, where $c > 0$. Any function $\rho(m)$ such that $\sum_{m=1}^{\infty} \rho(m)^{-1} < \infty$ guarantees G to be absolutely continuous a priori. By considering different values of c, Hanson and Johnson [38] found the family $\alpha_{\varepsilon_1 \cdots \varepsilon_m} = \rho(m) = cm^2$ to be sufficiently rich to capture interesting features of the distributions under consideration, and c has the same effect than the precision parameter M of a DP.

Although the prior mean distribution function may have a smooth Lebesgue density, the randomly sampled densities from a simple PT are very rough, being nowhere differentiable. In fact, the posterior predictive densities of future observations computed from PT have noticeable jumps at the boundaries of partitioned sets and that a choice of centering distribution G_0 that is particularly unlike the sample distribution of the data will make convergence of the posterior very slow. Note also that a practical implementation requires some meaningful elicitation of the centering distribution. Often a class of target measures can be identified, like the normal family, but it is hard to choose a single member of the family.

To overcome these difficulties it is natural to consider a centering measure which contains unspecified parameters, G_θ, and a further prior is given to these hyperparameters. The resulting hierarchical prior is therefore a mixture of PTs (MPT). The additional parameter will average out jumps to yield smooth densities [38]. Note, however, that the TF property is lost. Of particular interest is the situation where θ is a scale parameter and G is forced to have median zero [38]. By using similar arguments to the ones considered by Hanson and Johnson [38] it is possible to prove that when θ is a location parameter the posterior expected density is continuous everywhere. Another possible way of creating MPT is by keeping the partition fixed and varying the α_ε's parameters. As the partitions do not vary, the resulting density is discontinuous everywhere just like a usual PT.

Computation with PT might be hindered by the need to update the infinite number of parameters which describe the tree. The finite PT (FPT), which is also called a "partially specified Polya tree" [30], addresses this concern. The finite PT is constructed to be identical to the PT up to a finite pre-specified level J. However, the PT parameters in the set $\{\alpha_\varepsilon : \varepsilon \in E^*\}$ are updated only to level J in the FPT.

Lavine [30] discusses two scenarios for which it might be reasonable to update only to a pre-specified level J. The first scenario is when the parameters in $\{\alpha_\varepsilon : \varepsilon \in E^*\}$ are constructed to increase rapidly enough as the level of the tree increases. The posterior updating of the distributions of Y_ε beyond level J does not affect the prior strongly. This is explained by the behavior of the parameters of the prior beta distribution and the number of available data points as the level of the partition increases. On the one hand, the bigger the level of the partition, the bigger the parameters of the prior beta distribution. Thus, the prior variance goes to zero as the level of the partition increases and the beta prior distribution concentrates the mass around its mean (0.5). On the other hand, there are less data points in each set as the level of the partition increases and the posterior distribution of Y_ε is pretty much concentrated on 0.5 as the level of the partition increases. Therefore, beyond level J, the Y_ε's do not have much impact on the final random probability distribution.

Please cite this article in press as: A. Jara, Theory and computations for the Dirichlet process and related models: An overview, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.11.008

The second scenario in which FPT are appealing arises from concerns of prior elicitation. It might be possible to elicit prior information about parameters near the top of the PT and information about aspects of the distribution such as shape and modality, but it could be unreasonable to expect to elicit meaningful prior distributions for each and every parameter of the PT prior.

Lavine [30] detailed how such a level can be chosen by placing bounds on the posterior predictive distribution at a point. Hanson and Johnson [38] suggested the rule of thumb $J \approx \log_2(n)$ and [39] suggested $J \approx \log_2(n/N)$, where N is a "typical" number of observations falling into each set at level J when there is reasonable comfort in the centering family.

Multivariate version of PT priors have been also developed [40,39,41]. Jara et al. [41] proposed a scalable multivariate MPT that allows the user to separates the choice of the partition from the choice of the centering parameters, by considering a multivariate normal centering distribution and random partitions obtained by mixing over of the decomposition of the covariance matrix. An advantage of this approach is that the partition dependence associated with standard PT is avoided and, at the same time, the MPT inherits the mean-centered property around a multivariate normal distribution. This model has been also employed for proposing an automatic independent sampler Metropolis–Hastings algorithm [42].

2.3.2. Priors obtained from random series representations

Sethuraman [16] infinite series representation creates several possibilities of generalizing the DP and considering alternatives to DPM models, by changing the distribution of the weights, the support points, or the number of terms. Natural candidates follow from truncating the infinite series representation, $\sum_{i=1}^{\infty} W_i \delta_{\boldsymbol{\theta}_i}$, by using some appropriately chosen value of N and considering $\sum_{i=1}^{N} W_i \delta_{\boldsymbol{\theta}_i}$. An example of this procedure is the ϵ-DP proposed by Muliere and Tardella [43], where N is chosen such that the total variation distance between the DP and the truncation is bounded by a given ϵ. Another variation is the Dirichlet-multinomial process introduced by Muliere and Secchi [44]. Here the random probability measure is, for some finite N, given by

$$G(\cdot) = \sum_{i=1}^{N} W_i \delta_{\boldsymbol{\theta}_i}(\cdot),$$

where

$$(W_1, \ldots, W_N) \mid M, N \sim \text{Dirichlet}(M/N, \ldots, M/N),$$

and

$$\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N \mid G_0 \overset{i.i.d.}{\sim} G_0.$$

More generally, Pitman [45] described a class of models

$$G(\cdot) = \sum_{i=1}^{\infty} W_i \delta_{\boldsymbol{\theta}_i}(\cdot) + \left(1 - \sum_{i=1}^{\infty} W_i \delta_{\boldsymbol{\theta}_i}(\cdot)\right) G_0,$$

where, for a continuous distribution G_0, we have that $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots \overset{i.i.d.}{\sim} G_0$, assumed independent of the non-negative random variables W_i. The weights W_i are constrained by $\sum_{i=1}^{\infty} W_i \leq 1$ a.s. The model is known as Species Sampling Model (SSM), with the interpretation of W_i as the relative frequency of the ith species in a list of species present in a certain population, and $\boldsymbol{\theta}_i$ as the tag assigned to that species. If $\sum_{i=1}^{\infty} W_i = 1$ the SSM is called proper and the corresponding probability measure is discrete. The class of SSM includes as special cases the DP and some normalized random measures [46], among many others. In fact, the stick-breaking priors studied by Ishwaran and James [47] are also a special case of SSM, adopting the form $\sum_{i=1}^{N} W_i \delta_{\boldsymbol{\theta}_i}$, where $1 \leq N \leq \infty$. The weights are defined as $W_i = \prod_{j=1}^{i-1}(1 - V_j)V_i$ with $V_i \sim \text{Beta}(a_i, b_i)$, independently, for given sequences $(a_1, a_2, \ldots)$ and $(b_1, b_2, \ldots)$. Stick-breaking priors are quite general, including not only the Dirichlet-multinomial process and the DP as special cases, but also a two-parameter DP extension, known as the Poisson–Dirichlet (PD) process [48], and the beta two-parameter process [49].

The PD processes belong to the class of species sampling models and admits the DP prior as an important special case. The PD process can also be defined as $G(B) = \sum_{i=1}^{\infty} W_i \delta_{\boldsymbol{\theta}_i}(B)$ where the random weights W_i are independent for the $\boldsymbol{\theta}_i$'s and the $\boldsymbol{\theta}_i$ are i.i.d. from a distribution G_0. The weights have the following stick-breaking representation $W_i = V_i \prod_{j<i}(1 - V_j)$, where $V_i \overset{ind.}{\sim} \text{Beta}(1-\gamma, M + j\gamma)$, where either $\gamma = -\kappa < 0$ and $M = \varsigma\kappa$, for some $\kappa > 0$ and $\varsigma = 2, 3, \ldots$, or $0 \leq \gamma < 1$ and $M > -\gamma$. In applications of the PD model the parameter space is usually restricted to $\mathcal{A} = \{(\gamma, M) \in \mathbb{R}^2 : 0 \leq \gamma < 1, M > -a\}$, because it is large enough to include two important special cases. When $\gamma = 0$, the DP(MG_0) follows. When $0 < \gamma < 1$ and $M = 0$, the PD$(\gamma, 0)$ process corresponds to the normalized γ-stable subordinator introduced by Kingman [50]. The DP and stable law are key processes because they represent the canonical measures of the PD process [48].

3. BNP priors for collections of probability distributions

Suppose that we observe regression data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i) : i = 1, \ldots, n\}$, where $\boldsymbol{x}_i \in \mathcal{X} \subseteq \mathbb{R}^p$ is a p-dimensional vector of predictors and $\boldsymbol{y}_i$ is an outcome vector. Rather than assuming an unknown functional form for the mean function or another

functional, as is usually done in nonparametric regression, under the framework of fully nonparametric regression the problem is cast as inference for a family of predictor-dependent distributions

$$\left\{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^p\right\},$$

where $\boldsymbol{y}_i \mid \boldsymbol{x}_i \overset{ind.}{\sim} F_{\boldsymbol{x}_i}$. Therefore, from a Bayesian point of view, the specification of a fully nonparametric regression model requires the definition of a probability model for the set of predictor-dependent probability measures $\{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$, allowing the complete shape of the elements of $\{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathcal{X}\}$ to change flexibly with the values of $\boldsymbol{x}$.

3.1. Dependent Dirichlet process and its extensions

An early reference on predictor-dependent DP models is Cifarelli and Regazzini [51], who defined a model for related probability measures, by introducing a regression model in the centering measure of a collection of independent DP random measures. This approach is used, for example, by Muliere and Petrone [52], who considered a linear regression model for the centering distribution $G_{\boldsymbol{x}}^0 \equiv N(\boldsymbol{x}'\boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients, and $N(\mu, \sigma^2)$ stands for a normal distribution with mean μ and variance σ^2, respectively. Similar models were discussed in [53] and [54]. Linking nonparametric models through the centering distribution, however, limits the nature of the dependence of the process. A more flexible construction, the dependent Dirichlet process (DDP), was proposed by MacEachern [55,56]. The key idea behind the DDP is the construction of a set of random measures that are DPs marginally, i.e. that for every predictor value the random probability measure G_{bx} is a DP. In this framework, dependence is introduced through a modification of the stick-breaking representation of each element in the set.

MacEachern [55,56] generalized the stick-breaking representation of a DP by considering

$$G_{\boldsymbol{x}}(B) = \sum_{j=1}^{\infty} W_j(\boldsymbol{x}) \delta_{\boldsymbol{\theta}_j(\boldsymbol{x})}(B),$$

where the support points $\boldsymbol{\theta}_j(\boldsymbol{x})$, $j = 1, \ldots,$ are independent stochastic processes with index set $\mathcal{X}$ and $G_{\boldsymbol{x}}^0$ marginal distributions, and the weights take the form $W_j(\boldsymbol{x}) = V_j(\boldsymbol{x}) \prod_{k<j}[1 - V_k(\boldsymbol{x})]$, where $\{V_j(\boldsymbol{x}) : j \geq 1\}$ are independent stochastic processes with index set $\mathcal{X}$ and Beta$(1, M_{\boldsymbol{x}})$ marginal distributions. MacEachern [56] also studied a version of the process with predictor-independent weights, $G_{\boldsymbol{x}}(B) = \sum_{j=1}^{\infty} W_j \delta_{\boldsymbol{\theta}_j(\boldsymbol{x})}(B)$.

Barrientos et al. [57] provided an alternative definition of MacEachern's DDP, as well as extensions to more general stick-breaking constructions. The alternative definition exploits the connection between Copulas and stochastic processes and makes explicit the parameters of the process. They provided a characterization of the weak support of the two versions of MacEachern's DDP and of a version of the DDP where only the weights depend on predictors. They also provide sufficient conditions for the full Hellinger support of mixture models induced by DDP priors, and characterize their Kulback–Leibler support.

De Iorio et al. [58,59] proposed a particular version of the DDP, where the component of the atoms defining the location in a DDP mixture of normals model follows a linear regression model $\boldsymbol{\theta}_l(\boldsymbol{x}) = (\boldsymbol{x}'\boldsymbol{\beta}_l, \sigma_l^2)$. An advantage of this model for related random probability measures, referred to as the linear DDP (LDDP), is that it can be represented as DPM of linear (in the coefficients) regression models, when the model is convolved with a normal kernel, given by

$$Y_i \mid G \overset{ind.}{\sim} \int N\left(Y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2\right) dG(\boldsymbol{\beta}, \sigma^2),$$

and

$$G \mid M, G_0 \sim \mathrm{DP}\left(MG_0\right).$$

Versions of this single weights DDP have been applied, for instance, to survival [59,60], spatial modeling [61], functional data [62], longitudinal data analysis [63] and ROC curve modeling [64].

3.2. Other models for related probability measures

Other extensions of the DP for dealing with related probability distributions include the DPM mixture of normals model for the joint distribution of the response and predictors [65], the hierarchical mixture of DPM [66], the hierarchical DP [67], the order-based DDP model [68], the nested DP [69], the predictor-dependent weighted mixture of DP [70], the kernel-stick breaking process [71], the matrix-stick breaking process [72], the local DP [73], the logit-stick breaking processes [74], the probit-stick breaking processes [75,76], the cluster-X model [77], the product partition model with regression on covariates [78], the dependent skew DP model [79], and the dependent Bernstein polynomials model [80], among many others. Dependent neutral to the right processes and correlated two-parameter Poisson–Dirichlet processes have been proposed by [81] and [82], respectively, by considering suitable Lévy copulas. The general class of dependent normalized completely random measures has been discussed, for instance, by [83].

Please cite this article in press as: A. Jara, Theory and computations for the Dirichlet process and related models: An overview, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.11.008

Based on a different formulation of the problem, Tokdar et al. [84] and Jara and Hanson [85] proposed alternatives to convolutions of dependent stick-breaking approaches, which yield conditional probability measures with density w.r.t. Lebesgue measure, without the need of convolutions. Motivated by problems with continuous predictors, Tokdar et al. [84] developed Bayesian density regression model based on logistic Gaussian processes and subspace projection. Motivated by problems with continuous and discrete predictors, Jara and Hanson [85], proposed mixtures of dependent tail-free processes, where the dependence is introduced by replacing the TF conditional probabilities Y_ε by logistic Gaussian processes $Y_\varepsilon(\boldsymbol{x})$.

4. Computational aspects

A critical advantage of BNP over parametric Bayesian analyses is the ability to incorporate uncertainty at the level of infinite-dimensional parameters. However, this flexibility increases the computational complexity of the analysis. Much of the rapid development of BNP models in the last decades has been a direct result of advances in simulation-based computational methods, particularly MCMC [1]. The introduction of MCMC methods in the area began with the work of Escobar [86] on DP, later published in Escobar [87] and Escobar and West [88]. The following sections provide some discussion on the computational aspects for posterior sampling for two important prior probability models.

4.1. Fitting a DPM model

MCMC algorithms [86,87], sequential imputations [89,90], predictive recursions [91,92] and variational methods [93,94] have been used for fitting models including DP priors. In this section we focus the attention on MCMC methods for three reasons: (i) they allow the construction of general purpose algorithms, (ii) they have been used successfully in the posterior sampling under DP priors, and (iii) they allow for full posterior inferences on the parameters of interest.

To illustrate the main MCMC algorithms we will consider the following abstract DPM model

$$X_i \mid G \overset{i.i.d}{\sim} \int f(\cdot \mid \boldsymbol{\theta}) G(d\boldsymbol{\theta}), \tag{1}$$

and

$$G \mid M, \boldsymbol{\eta} \sim \mathrm{DP}(MG_{\boldsymbol{\eta}}), \tag{2}$$

and

$$(M, \boldsymbol{\eta}) \sim p(M)p(\boldsymbol{\eta}). \tag{3}$$

4.1.1. Marginal approaches

The use of MCMC methods for the model given by expressions (1)–(3) is not feasible since this would require the imputation of an infinite-dimensional parameter G. Escobar [86] avoided this problem by marginalizing analytically the DP from the model, thus providing the first mechanism for fitting a DPM model. Any class of algorithms where the DP is integrated out from the hierarchical model are referred to as marginal approaches, and are based on the following representation of the DPM model given by expressions (1)–(3), which is obtained by introducing unit-specific parameters,

$$\begin{aligned} &X_i \mid \boldsymbol{\theta}_i \overset{ind.}{\sim} f(\cdot \mid \boldsymbol{\theta}_i), \\ &\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \mid G \overset{i.i.d.}{\sim} G, \\ &G \mid M, \boldsymbol{\eta} \sim \mathrm{DP}(MG_{\boldsymbol{\eta}}), \end{aligned}$$

and

$$(M, \boldsymbol{\eta}) \sim p(M)p(\boldsymbol{\eta}).$$

4.1.1.1. Escobar's Gibbs sampler Under the previous formulation of the DPM model, Escobar [86] proposed a Gibbs sampler approach to explore the posterior distribution of the finite-dimensional parameters $P(d\boldsymbol{\theta}_{1:n}, dM, d\boldsymbol{\eta} \mid \boldsymbol{X}_{1:n})$. The Gibbs sampling approach is based on the full conditional distributions

$$\begin{aligned} &(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}^{(i)}, M, \boldsymbol{\eta}, \boldsymbol{X}_{1:n}), \quad i = 1, \ldots, n, \\ &(M \mid \boldsymbol{\theta}_{1:n}, \boldsymbol{\eta}, \boldsymbol{X}_{1:n}), \\ &(\boldsymbol{\eta} \mid \boldsymbol{\theta}_{1:n}, M, \boldsymbol{X}_{1:n}), \end{aligned}$$

where $\boldsymbol{\theta}^{(i)} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_n)$. Note that, by Corollary 1 above,

$$P(d\boldsymbol{\theta}_i \mid \boldsymbol{\theta}^{(i)}, M, \boldsymbol{\eta}) \propto MG_{\boldsymbol{\eta}}(d\boldsymbol{\theta}_i) + \sum_{j \neq i} \delta_{\boldsymbol{\theta}_j}(d\boldsymbol{\theta}_i).$$

Now, by noticing that $X_i \perp\!\!\!\perp \boldsymbol{\theta}_{1:n} \mid \boldsymbol{\theta}_i$, Bayes rule's yields

$$P(d\boldsymbol{\theta}_i \mid \boldsymbol{\theta}^{(i)}, M, \boldsymbol{\eta}, \boldsymbol{X}_{1:n}) \propto f(X_i \mid \boldsymbol{\theta}_i) \left\{ MG_{\boldsymbol{\eta}}(d\boldsymbol{\theta}_i) + \sum_{j \neq i} \delta_{\boldsymbol{\theta}_j}(d\boldsymbol{\theta}_i) \right\}.$$

The normalization constant of this probability measure is given by

$$\begin{aligned} \int f(X_i \mid \boldsymbol{\theta}_i) \left\{ G_{\boldsymbol{\eta}}(d\boldsymbol{\theta}_i) + \sum_{j \neq i} \delta_{\boldsymbol{\theta}_j}(d\boldsymbol{\theta}_i) \right\} &= M \int f(X_i \mid \boldsymbol{\theta}_i) G_{\boldsymbol{\eta}}(d\boldsymbol{\theta}_i) + \sum_{j \neq i} f(X_i \mid \boldsymbol{\theta}_i), \\ &\equiv Ma(X_i, \boldsymbol{\eta}) + \sum_{j \neq i} b(X_i, \boldsymbol{\theta}_j). \end{aligned}$$

Therefore, it follows that

$$\boldsymbol{\theta}_i \mid \boldsymbol{\theta}^{(i)}, M, \boldsymbol{\eta}, \boldsymbol{X}_{1:n} \sim q_0 P(d\boldsymbol{\theta}_i \mid M, \boldsymbol{\eta}, X_i) + \sum_{j \neq i} q_j \delta_{\boldsymbol{\theta}_j}(d\boldsymbol{\theta}_i),$$

where $q_0 \propto Ma(X_i, \boldsymbol{\eta})$ and $q_j \propto b(X_i, \boldsymbol{\theta}_j)$.

An advantage of Escobar's Gibbs sampling is that the computational effort is, in principle, independent of the dimensionality of $\boldsymbol{\theta}$. However, it may suffer from some potential drawbacks. Firstly, note that practical implementation requires the evaluation of $a(X_i, \boldsymbol{\eta})$ for different values of X_i and $\boldsymbol{\eta}$. Clearly, the computations are straightforward if the integrations required to compute $a(X_i, \boldsymbol{\eta})$ may be cheaply performed. When $G_{\boldsymbol{\eta}}$ is a conjugate prior, then the marginal distribution is known analytically. When non-conjugate priors are used, as is appropriate in many contexts, the Gibbs sampler implementation requires that an often difficult numerical integration be performed.

On the other hand, note that a realization of the Polya urn scheme partitions the vector $\boldsymbol{\theta}$ into a batch of clusters. In the case of a continuous $G_{\boldsymbol{\eta}}$ the $\boldsymbol{\theta}_i$ belonging to different clusters assume different values with probability one. Note that this view of the DP also leads to the representation of $\boldsymbol{\theta}_{1:n}$ as $(\boldsymbol{\theta}^*, \boldsymbol{s})$, where the vector $\boldsymbol{s}$ contains the clustering of the $\boldsymbol{\theta}_j$'s and $\boldsymbol{\theta}^*$ contains the $k \leq n$ unique locations of the clusters. The relationship between $\boldsymbol{\theta}_{1:n}$ and $(\boldsymbol{\theta}^*, \boldsymbol{s})$ is given by $\boldsymbol{\theta}_i = \boldsymbol{\theta}^*_{s_i}$. A possible drawback of Escobar's Gibbs sampler algorithm is that the locations of the clusters or groups of parameters could essentially become fixed moving only rarely.

4.1.1.2. Other marginal approaches for conjugate DPM models Bush and MacEachern [95] proposed an approach for improving the mixing of Escobar's algorithm. In their proposal a second stage is added to the original Gibbs sampler, wherein the cluster locations are moved. In their approach, updating the latent mixture parameters $\boldsymbol{\theta}_{1:n}$ proceeds with the equivalent parametrization $(\boldsymbol{\theta}^*, \boldsymbol{s})$. Each Gibbs sampling scan consists of picking a new value for each s_i from its conditional distribution given X_i, $\boldsymbol{\theta}^*$ and $\boldsymbol{s}^{(i)} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$, and then picking a new value for each coordinate in $\boldsymbol{\theta}^*$ from its conditional distribution given the X_i's for which $s_i = r$, $i = 1, \ldots, n$. The required conditional probabilities for s_i can easily be computed and are given by

$$P\left(s_i = j \mid \boldsymbol{s}^{(i)}, \boldsymbol{\theta}^*, M, \boldsymbol{\eta}, X_i\right) \propto \begin{cases} \frac{n_j^{(i)}}{n-1+M} f(X_i \mid \boldsymbol{\theta}^*_j), & j \in \{1, \ldots, k^{(i)}\}, \\ \frac{M}{n-1+M} \int f(X_i \mid \boldsymbol{\theta}) G_{\boldsymbol{\eta}}(d\boldsymbol{\theta}), & j = k^{(i)} + 1, \end{cases}$$

$i = 1, \ldots, n$, where $n_r \sum_{j=1}^n I(s_j)_{\{s_j = r\}}$ is the size of the r-th cluster, with $I(\cdot)_A$ being the indicator function of the set A, $n_r^{(i)} = \sum_{j=1, j \neq i}^n I(s_j)_{\{s_j = r\}}$ is the size of the r-th cluster, after the i-th subject is removed, and $k^{(i)} \leq n$ is the dimension of the vector $\boldsymbol{\theta}^*$, after the i-th subject is removed. Note that $\boldsymbol{\theta}^*$ and its dimension remain unchanged when the i-th subject is not alone in a given cluster. On the other hand, $\boldsymbol{\theta}^*_{s_i}$ is removed from $\boldsymbol{\theta}^*$ when the cluster in which the i-th subject has been previously assigned is a singleton. In the latter case, it is also assumed the that cluster indicators are relabeled correspondingly to match the changes in $\boldsymbol{\theta}^*$ and, thus, every coordinate of $\boldsymbol{s}^{(i)}$ belongs to the set $\{1, \ldots, k^{(i)}\}$. Finally, note that under this Gibbs sampling, when the updating of s_i chooses a value not equal to any other $\boldsymbol{s}^{(i)}$, the corresponding value for $\boldsymbol{\theta}^*_{s_i}$ is chosen from the posterior distribution of $\boldsymbol{\theta}$, based on the prior $G_{\boldsymbol{\eta}}$ and the single observation X_i. In such a case, $\boldsymbol{\theta}^*$ is also updated by incorporating the new coordinate $\boldsymbol{\theta}^*_{s_i}$.

The second step added by Bush and MacEachern [95], the cluster locations are re-sampled from the conditional posterior distribution. The latter distribution for $\boldsymbol{\theta}^*_j$ is obtained by combining the baseline prior $G_{\boldsymbol{\eta}}$ with the likelihood $\prod_{l:s_l=j} f(X_l \mid \boldsymbol{\theta}^*_j)$. That is, it is the posterior based on a random sample of observations which are drawn from the same $\boldsymbol{\theta}^*_j$. Note that, as was the case for Escobar's original Gibbs sampling method, this approach is feasible if it is possible to compute $a(X_i, \boldsymbol{\eta})$ needed for the normalizing constant b.

MacEachern [96] also proposed to analytically integrate over the $\boldsymbol{\theta}^*_j$'s, eliminating them from the algorithm. In this case, the state of the Markov chain then consists only of the s_j's which are updated in a Gibbs sampling scheme with the following conditional probabilities,

$$P\left(s_i = j \mid \boldsymbol{s}^{(i)}, \boldsymbol{\theta}^*, M, \boldsymbol{\eta}, X_i\right) \propto \begin{cases} \frac{n_j^{(i)}}{n-1+M} \int f(X_i \mid \boldsymbol{\theta}) H_j^{(i)}(d\boldsymbol{\theta}), & j \in \{1, \ldots, k^{(i)}\}, \\ \frac{M}{n-1+M} \int f(X_i \mid \boldsymbol{\theta}) G_{\boldsymbol{\eta}}(d\boldsymbol{\theta}) & j = k^{(i)} + 1, \end{cases}$$

where $H_j^{(i)}$ is the posterior distribution of $\boldsymbol{\theta}$ based on the prior $G_{\boldsymbol{\eta}}$ and all observations X_l for which $l \neq i$ and $s_l = j$. Of course, practical implementation is feasible if it is possible to integrate analytically over the $\boldsymbol{\theta}_j^*$, as will generally be the case when $G_{\boldsymbol{\eta}}$ is the conjugate prior.

Finally, Jain and Neal [97] and Dahl [98] proposed merge-split samplers for conjugate DPM models. Being merge-split algorithms [99,100], both methods try to update groups of indices in one update. Dahl proposal borrows ideas from the sequential importance sampling. The sampler, referred to as SAMS, proposes splits by sequentially allocating observations to one of two split components using allocation probabilities that are conditional on previously allocated data. My own experience with these merge-split algorithms is not satisfactory for moderate to large datasets. This is explained by the use of "uniformly" generated partitions as candidate movements, which generates huge number of unlikely accepted candidates. Alternative algorithms trying to fixing this problem has been proposed by Fong et al. [101].

4.1.1.3. Marginal approaches for non-conjugate DPM models The algorithms described so far cannot easily be applied to models where $G_{\boldsymbol{\eta}}$ is not a conjugate prior for $f(\cdot \mid \boldsymbol{\theta})$, as the integral $\int f(X_i \mid \boldsymbol{\theta}) G_{\boldsymbol{\eta}}(d\boldsymbol{\theta})$ will then usually not be analytically tractable. Sampling form the posterior distribution, H_j, may also be hard when the prior is not conjugate.

MacEachern and Müller [102] presented a framework that allow auxiliary values for $\boldsymbol{\theta}$ drawn from $G_{\boldsymbol{\eta}}$ to be used to define a valid Markov chain sampler in this context. Denoting by k the number of distinct elements in $\boldsymbol{\theta}_{1:n}$, they proposed an alternative parametrization by augmenting $\boldsymbol{\theta}^*$ to

$$\{\underbrace{\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^*}_{\boldsymbol{\theta}_F^*}, \underbrace{\boldsymbol{\theta}_{k+1}^*, \ldots, \boldsymbol{\theta}_n^*}_{\boldsymbol{\theta}_E^*}\}.$$

The augmentation relies upon the constraint that there will be no gaps in the values of the s_i, i.e., $n_j > 0$ for $j = 1, \ldots, k$, and $n_j = 0$ for $j = k+1, \ldots, n$. This justifies the name, "no gaps", of the algorithm. They interpreted $\boldsymbol{\theta}_E^*$ as potential but not yet used cluster locations, the empty clusters, and $\boldsymbol{\theta}_F^*$ as full clusters. In this augmented model the Gibbs sampler is simplified, such that the evaluation of integrals is replaced by simple likelihood evaluations,

$$P\left(s_i = j \mid \boldsymbol{s}^{(i)}, \boldsymbol{\theta}^*, M, \boldsymbol{\eta}, X_i\right) \propto \begin{cases} n_j^{(i)} f(X_i \mid \boldsymbol{\theta}_j^*), & j \in \{1, \ldots, k^{(i)}\}, \\ \frac{M}{k^{(i)}+1} f(X_i \mid \boldsymbol{\theta}_{k^{(i)}+1}^*), & j = k^{(i)} + 1. \end{cases}$$

Again, in the re-sampling step, the conditional posterior distribution of θ_j^* is obtained by combining the baseline prior $G_{\boldsymbol{\eta}}$ with the likelihood $\prod_{l:s_l=j} f(X_l \mid \boldsymbol{\theta}_j^*)$. Note that this algorithm can be applied to any model for which we can sample from $G_{\boldsymbol{\eta}}$ and compute $f(X_l \mid \boldsymbol{\theta})$ regardless of whether $G_{\boldsymbol{\eta}}$ is the conjugate prior for $f(\cdot \mid \boldsymbol{\theta})$. As noted by Neal [103], however, there is a puzzling inefficiency in the algorithm's mechanism for setting s_i to a value different from all other s_j, that is, for assigning an observation to a newly created mixture component. The probability of such a change is reduced from what one might expect by a factor of $k^{(i)} + 1$, with a corresponding reduction in the probability of the opposite change. Neal [103] described a similar algorithm without this inefficiency and proposed three Metropolis–Hastings algorithms with partial Gibbs sampling to update the configurations.

The Gibbs sampler with auxiliary variables of Neal [103] (algorithm 8) is similar to the approach of MacEachern and Müller [102], but differs in that the auxiliary parameters are regarded as existing only temporarily. The main idea behind his algorithm is that if $s_i \neq s_l$ for all $l \neq i$, then it must be associated with one of the m auxiliary parameters. Under this algorithm, the conditional probabilities for the cluster indicators are given by

$$P\left(s_i = j \mid \boldsymbol{s}^{(i)}, \boldsymbol{\theta}^*, M, \boldsymbol{\eta}, X_i\right) \propto \begin{cases} n_j^{(i)} f(X_i \mid \boldsymbol{\theta}_j^*), & j \in \{1, \ldots, k^{(i)}\}, \\ \frac{M}{m} f(X_i \mid \boldsymbol{\theta}_j^*), & j \in \{k^{(i)} + 1, \ldots, k^{(i)} + m\}. \end{cases}$$

Note that when $m = 1$, algorithm 8 of Neal [103] closely resembles the "no gaps" algorithm of MacEachern and Müller [102]. The difference is that the probability of changing s_i from a component shared with other observations to a new singleton component is approximately $k^{(i)} + 1$ times greater with algorithm 8 and the same is true for the reverse change. When M is small this seems to be a clear benefit, since the probabilities for other changes are affected only slightly. Using a simulated dataset, Neal [103] also showed that the auxiliary Gibbs sampler (with a properly chosen tuning parameter) has the best computational efficiency of one-at-a-time non-conjugate samplers for DPM models.

MacEachern and Müller [102] have also developed an algorithm based on a "complete" scheme for mapping from the $\boldsymbol{\theta}^*$ to $\boldsymbol{\theta}_i$. It requires maintaining n values for $\boldsymbol{\theta}^*$, which may be inefficient when $k < n$. Finally, Dahl [98] also proposed a version of his SAMS sampler for non-conjugate DPM.

Please cite this article in press as: A. Jara, Theory and computations for the Dirichlet process and related models: An overview, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.11.008

4.1.2. Posterior sampling approaches for non-standard DPM models

Doss [104], Florens and Rolin [105], and Hanson and Johnson [106] discussed posterior algorithms in non-standard hierarchical models. These models arise when the likelihood, viewed as a function of $X_1, \ldots, X_n$, is a simple indicator function

$$\prod_{i=1}^{n} I_{A_i^{\theta}}(X_i),$$

where A_i^{θ} is a subset of the sample space. This situation gives rise to algorithmic possibilities which are unavailable or very difficult to implement under standard hierarchical models. Examples of this type of models arise in survival analysis and categorical data analysis. Under these models it is not necessary to sample the complete process G, but only the needed parts of G. This is possible to be performed by using the properties of DP. Specifically, note that by Theorem 3 above, a representation of the process G has the following form:

$$G = \sum_{j=1}^{N} G_j G^j.$$

where j indexes the sets that define a finite partition of the sample space $\{B_1, \ldots, B_N\}$ (defined by the data), $G_j = G(B_j)$ and $G^j(\cdot) = G(\cdot|B_j) = G|_{B_j}$, with the G_j's being Dirichlet distributed random variables and the G^j's being independent DPs. Therefore, $G \mid \boldsymbol{X}_{1:n}, \theta, M, \boldsymbol{\eta}$ can be sampled by first sampling $\{G_j\} \mid \boldsymbol{X}_{1:n}, \theta, M, \boldsymbol{\eta}$ by using Ferguson's definition of DP and then by sampling each $G^j \mid \{G_j\}, \boldsymbol{X}_{1:n}, \theta, M, \boldsymbol{\eta}$ using the stick-breaking representation of the DP. Note that the implementation of this strategy requires to sample the latent data $X_1, \ldots, X_n$ form G subject to the constraint that $X_i \in A_i^{\theta}$. Because of the a.s. discreteness of DP, this can be performed using the standard inverse cumulative distribution function method for discrete variables. This idea has been extended by Papaspilopoulos and Roberts [107] for standard-hierarchical DPM models.

4.1.3. Augmented posterior sampling approaches

The MCMC approaches described in Section 4.1.1 avoid the use of approximations of the DPM model by integrating out completely the infinite-dimensional parameter G. Alternative classes of MCMC approaches, that also do not rely on parametric approximations of the model, augment the posterior distribution with pieces of the infinite-dimensional process, such that the computation is exact (up to a Monte Carlo error) for the finite-dimensional posterior associated with the remaining parameters of the model. These alternative classes include the slice sampling algorithm, originally proposed by Walker [108], and the retrospective sampling algorithm proposed by Papaspiliopoulos and Roberts [107].

To describe these algorithms, first notice that the mixture model induced by a DP given by expressions (1)–(3), can be equivalently written as

$$X_i \mid \boldsymbol{\theta}^*_{1,\infty}, \boldsymbol{V}_{1:\infty} \overset{i.i.d.}{\sim} \sum_{j=1}^{\infty} W_j f(\cdot \mid \boldsymbol{\theta}^*_j),$$

$$V_j \mid M \overset{i.i.d.}{\sim} \text{Beta}(1, M),$$

$$\boldsymbol{\theta}^*_j \mid \boldsymbol{\eta} \overset{i.i.d.}{\sim} G_{\boldsymbol{\eta}},$$

$$(M, \boldsymbol{\eta}) \sim p(M)p(\boldsymbol{\eta}),$$

where $\boldsymbol{V}_{1:\infty} = (V_1, \ldots, V_\infty)$ and $\boldsymbol{\theta}^*_{1:\infty} = (\boldsymbol{\theta}^*_1, \ldots, \boldsymbol{\theta}^*_\infty)$. The key idea behind these algorithms is to find, at each step of the Gibbs sampler, a finite number of elements in $\boldsymbol{V}_{1:\infty}$ and $\boldsymbol{\theta}^*_{1:\infty}$ that need to be sampled to produce a valid Markov chain with correct stationary distribution for the remaining parameters.

4.1.3.1. The slice sampling algorithm The first complete proposal of the slice sampling algorithm for a DPM model is given by Walker [108]. However, we described here a more efficient version discussed by Kalli et al. [109]. In this approach, uniform latent variables U_i's are introduced such that the joint density is given by

$$f(X_i, U_i \mid \boldsymbol{V}_{1:\infty}, \boldsymbol{\theta}^*_{1:\infty}) = \sum_{j=1}^{\infty} I(U_i, W_i)_{\{U_i < W_j\}} f(X_i \mid \boldsymbol{\theta}^*_j).$$

Given U_i, the number of mixture component is finite,

$$f(X_i \mid U_i, \boldsymbol{V}_{1:\infty}, \boldsymbol{\theta}^*_{1:\infty}) = N_U^{-1} \sum_{j \in A_{U_i}} f(X_i \mid \boldsymbol{\theta}^*_j),$$

where $A_{U_i} = \{k : W_k > U_i\}$ and $N_{U_i} = \sum_{j=1}^{\infty} I(U_i, W_j)_{\{U_i > W_j\}}$. A component indicator variable s_i can be introduced, such that the joint density is given by

$$f(X_i, U_i, s_i \mid \boldsymbol{V}_{1:\infty}, \boldsymbol{\theta}_{1:\infty}) = I(U_i, W_{s_i})_{\{U_i < W_{s_i}\}} f(X_i \mid \boldsymbol{\theta}_j^*).$$

Hence the complete likelihood function for $(\boldsymbol{V}_{1:\infty}, \boldsymbol{\theta}_{\infty}^*)$ is available as a simple product of terms and crucially s_i is finite. Without U_i, s_i can take an infinite number of values.

The full conditional distributions are then given by

$$\begin{aligned}
&\boldsymbol{\theta}_j^* \mid \cdots \propto g_{\boldsymbol{\eta}}(\boldsymbol{\theta}_j^*) \prod_{i:s_i=j} f(X_i \mid \boldsymbol{\theta}_j^*), \quad j = 1, \ldots, \\
&V_j \mid \cdots \sim \text{Beta}\left(1 + \sum_{i=1}^{n} I_{\{s_i=j\}}, \alpha + \sum_{i=1}^{n} I_{\{s_i>j\}}\right), \quad j = 1, \ldots, \\
&U_i \mid \cdots \sim U(0, W_{s_i}), \quad i = 1, \ldots, n, \\
&P(s_i = k \mid \cdots) \propto I(k)_{\{W_k > U_i\}} f(X_i \mid \boldsymbol{\theta}_k^*), \quad i = 1, \ldots, n.
\end{aligned}$$

Obviously, we can not sample all of the $(V_j, \boldsymbol{\theta}_j^*)$. But it is not required to in order to proceed with the chain. We only need to sample up to the integer for which we have found all the appropriate W_k in order to do the fourth step exactly. Since the weights add up to 1, if we find N_i such that

$$\sum_{k=1}^{N_i} W_k > 1 - U_i,$$

then it is not possible for any of the W_k, for $k > N_i$, to be greater than U_i. In the original version Walker sampled the stick-breaking variables conditionally on the uniform latent variables. Please notice that this method also applies for other BNP priors for probability distributions.

4.1.3.2. The retrospective sampling algorithm Doss [104] discussed how to use the inverse cumulative distribution function method to sample random variables from a DP with perfect accuracy. The key idea is that to sample $s_i \sim \sum_{j=1}^{\infty} W_j \delta_j(\cdot)$, we can sample $U_i \sim (0, 1)$, and then set $s_i = j$ if and only if

$$\sum_{l=0}^{j-1} W_l < U_i \leq \sum_{l=1}^{j} W_l,$$

where $W_0 = 0$. Exchanging the order of simulation, between U_i, and the weights and atoms, we can generate a sample from G with perfect accuracy. Papaspiliopoulos and Roberts [107] used this idea and avoid the evaluation of the infinite sum by using a Metropolis–Hastings step. At each scan, denote by N the last component that has data assigned to it. They propose a candidate for s_i from

$$P(s_i = k \mid \cdots) \propto \begin{cases} W_k f(X_i \mid \boldsymbol{\theta}_k^*) & \text{for } k \leq N, \\ W_k M_i & \text{for } k > N. \end{cases}$$

The normalizing constant given by

$$\kappa(N) = \sum_{j=1}^{N} W_j f(X_i \mid \boldsymbol{\theta}_j^*) + \left(1 - \sum_{j=1}^{N} W_j\right) M_i.$$

Papaspiliopoulos and Roberts [107] choose M_i such that posterior probability of allocating i to a new component is greater than the prior:

$$M_i = M_i(N) = \max_{\{k \leq N\}} f(X_i \mid \boldsymbol{\theta}_k^*).$$

The MH acceptance ratio of changing s_i from k to j is given by

$$\begin{cases} 1 & \text{if } j \leq N \text{ and } N = N_{kj}, \\ \min\left\{1, \frac{\kappa(N)}{\kappa(N_{kj})} \frac{M(N_{kj})}{f(X_i \mid \boldsymbol{\theta}_k^*)}\right\} & \text{if } j \leq N \text{ and } N_{kj} \leq N, \\ \min\left\{1, \frac{\kappa(N)}{\kappa(N_{kj})} \frac{f(X_i \mid \boldsymbol{\theta}_j^*)}{M(N_{kj})}\right\} & \text{if } j > N, \end{cases}$$

where N_{kj} denotes the index of the last active component after changing component assignment of s_i form k to j.

4.1.4. Updating hyperparameters

Denoting by k the number of distinct elements in $\boldsymbol{\theta}_{1:n}$, the definition of the DP prior implies that $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_k^* \mid \boldsymbol{\eta} \overset{i.i.d.}{\sim} G_{\boldsymbol{\eta}}$. Therefore, if the prior for $\boldsymbol{\eta}$ and $G_{\boldsymbol{\eta}}$ are a conjugate pair, a Gibbs sampling step can be added to update $\boldsymbol{\eta}$ by a draw from the complete conditional posterior. In general a Metropolis–Hastings type of transition probability might be needed. Updating the total mass parameter M, on the other hand, becomes easy under a gamma prior [88]. From Lemma 1, it follows that

$$P(dM \mid \cdots) \propto \frac{M^k \Gamma(M)}{\Gamma(M+n)} P(dM).$$

If $M \sim \text{Gamma}(a_0, b_0)$, then

$$\begin{aligned} P(dM \mid \cdots) &\propto M^{a_0-1} \exp\{-b_0 M\} M^k \frac{\Gamma(M)}{\Gamma(M+n)}, \\ &= M^{a_0-1} \exp\{-b_0 M\} M^{k-1} (M+n) \frac{\Gamma(M+1)\Gamma(n)}{\Gamma(M+n+1)}, \\ &= M^{a_0+k-2} \exp\{-b_0 M\} (M+n) \int_0^1 \eta^M (1-\eta)^{n-1} d\eta. \end{aligned}$$

The clever algorithm proposed by Escobar and West [88] is based on the introduction of a latent variable γ, such that

$$P(dM, d\gamma \mid \cdots) \propto M^{a_0+k-2} \exp\{-b_0 M\} (M+n) \gamma^M (1-\gamma)^{n-1}.$$

Therefore,

$$\gamma \mid M, k \sim \text{Beta}(M+1, n),$$

and

$$M \mid \gamma, k \sim \pi \text{Gamma}(a_0 + k, n_0 - \log(\gamma)) + (1-\pi)\text{Gamma}(a_0 + k - 1, b_0 - \log(\gamma)),$$

where

$$\pi = \frac{a_0 + k - 1}{n(b_0 - \log(\gamma)) + a_0 + k - 1}.$$

4.1.5. Posterior inferences for functionals of a DP

All of the samplers described in the previous sections for DPM models allow for limited posterior inferences on a given functional $H(G)$ of the infinite-dimensional parameter G, and are typically limited to posterior expectations. Full posterior inference on $H(G)$ requires the imputation of the infinite-dimensional process. Thus, an additional step is typically added to the previous samplers to approximately sample from $P(dH(G) \mid \boldsymbol{X}_{1:n})$, based on samples from the posterior distribution for the unit-specific latent parameters $P(d\boldsymbol{\theta}_{1:n} \mid \boldsymbol{X}_{1:n})$. This is based on the fact that

$$P(dH(G) \mid \boldsymbol{X}_{1:n}) \propto P(dH(G) \mid \boldsymbol{\theta}_{1:n}) \times P(d\boldsymbol{\theta}_{1:n} \mid \boldsymbol{X}_{1:n}),$$

and then the composition sampling method can be applied, along with the conjugacy properties of a DP described in Theorem 1 above. In this approach, the approximate samples of the functional parameter are obtained by truncating the stick-breaking representation of the DP,

$$G_N(\cdot) = \sum_{j=1}^{N} W_j \delta_{\boldsymbol{\theta}_j^*}(\cdot),$$

where the fractions V_j are truncated after N terms, with $V_N = 1$, leaving $W_j = V_j \prod_{k=1}^{j-1}(1 - V_k)$, for $j = 1, \ldots, N-1$, $W_N = \prod_{k=1}^{N-1}(1 - V_k)$, and

$$\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_N^* \mid \boldsymbol{\theta}_{1:n} \overset{i.i.d.}{\sim} \frac{1}{n+M} \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}(\cdot) + \frac{M}{n+M} G_{\boldsymbol{\eta}}(\cdot)$$

and

$$V_1, \ldots, V_{N-1} \mid \boldsymbol{\theta}_{1:n} \overset{i.i.d.}{\sim} \text{Beta}(1, M+n).$$

Gelfand and Kottas [110], suggest to fix N and chose its value by solving the equation

$$E_{\boldsymbol{V}_{1:\infty}|\boldsymbol{X}_{1:\infty}}\left(1-\sum_{j=1}^{N}W_j\right)=1-\left(\frac{n+M}{n+M+1}\right)^N=\epsilon,$$

for an arbitrarily small constant $\epsilon>0$. A random N is considered in the implementations DPM models in DPpackage [111, 112], where the ϵ-DP approximation of Muliere and Tardella [43] is employed.

It is important to stress that under these approaches, only the posterior inference for the infinite-dimensional parameter is prone to an additional approximation error and that the approximation error can be arbitrarily controlled a posteriori. This distinguishes these approaches from the use of parametric approximations of the BNP model [47]. In the later case, the additional approximation is controlled a priori and not a posteriori.

4.1.6. MCMC schemes based on approximations of the DPM model

The use of finite-dimensional approximations for the DPM model, based on truncations of the stick-breaking representation of the DP, allows for straightforward MCMC schemes and can be a good starting point for practitioners that are new to the field and that are willing to bear the costs associated with a parametric approximation. Ishwaran and James [47] originally propose to approximate DPM models. The hierarchical representation of the approximated DPM model is given by

$$\begin{aligned}
&X_i \mid s_i, \boldsymbol{\theta}^*_{1:N} \overset{ind.}{\sim} f(\cdot \mid \boldsymbol{\theta}_{s_i}),\\
&s_i \mid \boldsymbol{V}_{1:N} \overset{i.i.d.}{\sim} \sum_{j=1}^{N} W_j \delta_j(\cdot),\\
&V_j \mid M \overset{i.i.d.}{\sim} \text{Beta}(1, M),\\
&\boldsymbol{\theta}^*_j \mid \boldsymbol{\eta} \overset{i.i.d.}{\sim} G_{\boldsymbol{\eta}},\\
&(M, \boldsymbol{\eta}) \sim p(M)p(\boldsymbol{\eta}).
\end{aligned}$$

The full conditional distributions for the model are given by

$$V_j \mid \cdots \sim \text{Beta}\left(1+n_j, M+\sum_{l=j+1}^{N} n_l\right),$$

$$P(s_i \mid \cdots) \propto \sum_{j=1}^{N} W_j f(X_i \mid \boldsymbol{\theta}_j)\delta_j(\cdot),$$

and

$$\boldsymbol{\theta}^*_j \mid \cdots \propto g_{\boldsymbol{\eta}}(\boldsymbol{\theta}^*_j) \prod_{i:s_i=j} f(X_i \mid \boldsymbol{\theta}^*_j).$$

Under this finite approximation, if an Gamma(a_0, b_0) prior is used for M, then

$$M \mid \cdots \text{Gamma}\left(N+a_0-1, b_0-\sum_{k=1}^{N-1}\log(1-W_k)\right).$$

Ishwaran and Zarepour [49] assess the truncation accuracy by considering the behavior of the moments of the tail probability and Ishwaran and James [113] provided a bound on the truncation error. Let

$$m_\infty(\boldsymbol{X}_{1:n}) = \int \left(\prod_{i=1}^{n}\int f(X_i \mid \boldsymbol{\theta})G(d\boldsymbol{\theta})\right) P(dG).$$

Let

$$m_N(\boldsymbol{X}_{1:n}) = \int \left(\prod_{i=1}^{n}\int f(X_i \mid \boldsymbol{\theta})G(d\boldsymbol{\theta})\right) P(dG_N).$$

Then

$$\begin{aligned}
\int |m_\infty(\boldsymbol{X}_{1:n}) - m_N(\boldsymbol{X})| d\boldsymbol{X}_{1:n} &\leq 4\left[1-E\left\{\left(\sum_{k=1}^{N-1} W_k\right)^n\right\}\right],\\
&\approx 4n\exp\left(-(N-1)/M\right).
\end{aligned}$$

4.2. Fitting a Polya tree model

As the PT prior can be constructed in such a way that it assigns probability one to the set of continuous and even absolutely continuous probability distributions, we consider a model for density estimation in order to illustrate the computational strategies associated to PT priors. Consider the default construction of Lavine [29], where the PT prior is centered around a distribution $G_{\boldsymbol{\eta}}$. Let $\Pi^{\boldsymbol{\eta}} = \{\pi_m^{\boldsymbol{\eta}} : m = 0, 1, \ldots\}$ and $\mathcal{A}^c = \{\alpha_\varepsilon^c : \varepsilon \in E^*\}$. The abstract model is

$$X_1, \ldots, X_n \mid G \overset{iid}{\sim} G,$$
$$G \mid c, \boldsymbol{\eta} \sim PT(\Pi^{\boldsymbol{\eta}}, \mathcal{A}^c),$$

and

$$(c, \boldsymbol{\eta}) \sim p(c)p(\boldsymbol{\eta}).$$

As done by Escobar [87] in the context of DP, it is possible to marginalize the PT and base the inference on the predictive distribution. On marginalization, the infinite-dimensional process G no longer must be partially sampled, and inference is exact up to MCMC error. Theorem 6 is used to find a version of the predictive density w.r.t. Lebesgue measure of $X_{n+1} \mid \boldsymbol{X}_{1:n}, c, \boldsymbol{\eta}$,

$$p(X_{n+1} \mid \boldsymbol{X}_{1:n}, c, \boldsymbol{\eta}) = \lim_{m\to\infty} \left(\prod_{j=1}^{m} \frac{2cj^2 + 2n_{\epsilon(j,\boldsymbol{\eta},X_{n+1})}(\boldsymbol{\eta}, \boldsymbol{X}_{1:n})}{2cj^2 + n_{\epsilon(j-1,\boldsymbol{\eta},X_{n+1})}(\boldsymbol{\eta}, \boldsymbol{X}_{1:n})} \right) g_{\boldsymbol{\eta}}(X_{n+1}),$$

where $\epsilon(j, \boldsymbol{\eta}, \boldsymbol{X}_{1:n})$ denote the binary expansion $\varepsilon_1 \cdots \varepsilon_j \in \{0, 1\}^j$ such that $X \in B^{\boldsymbol{\eta}}_{\varepsilon_1\cdots\varepsilon_j}$, and $n_\epsilon(\boldsymbol{\eta}, \boldsymbol{X}_{1:n})$ is the number of $\boldsymbol{X}_{1:n}$ in the set $B^{\boldsymbol{\eta}}_\epsilon$. Hence the joint posterior distribution of $\boldsymbol{\eta}, c \mid \boldsymbol{X}_{1:n}$ is given by

$$\begin{aligned} p(\boldsymbol{\eta}, c \mid \boldsymbol{X}_{1:n}) &\propto f(\boldsymbol{X}_{1:n} \mid c, \boldsymbol{\eta})p(c)p(\boldsymbol{\eta}) \\ &= \left\{ \prod_{i=1}^{n} f(X_i \mid \boldsymbol{X}_{1:i-1}, c, \boldsymbol{\eta}) \right\} p(c)p(\boldsymbol{\eta}) \\ &= \left\{ \prod_{i=1}^{n} \left[\lim_{m\to\infty} \left(\prod_{j=1}^{m} \frac{2cj^2 + 2n_{\epsilon(j,\boldsymbol{\eta},X_i)}(\boldsymbol{\eta}, \boldsymbol{X}_{1:i-1})}{2cj^2 + n_{\epsilon(j-1,\boldsymbol{\eta},X_i)}(\boldsymbol{\eta}, \boldsymbol{X}_{1:i-1})} \right) g_{\boldsymbol{\eta}}(X_i) \right] \right\} \\ &\quad p(c)p(\boldsymbol{\eta}). \end{aligned}$$

where $\boldsymbol{X}_{1:0} = \emptyset$. This result follows naturally from Theorem 4 of Lavine [30]. Motivated by a median regression model, Hanson and Johnson [38] proposed a Metropolis–Hastings algorithm to obtain posterior inference based on this expression.

An alternative strategy is to consider a FPT by terminating and updating the partition $\Pi^{\boldsymbol{\eta}}$ up to a finite level J. Lavine [30] described the use of FPT to model the errors in regression settings using a Gibbs sampler algorithm for the PT probabilities. Hanson [39] developed algorithms to sample the fully conditional distributions of the FPT parameters $\mathcal{Y}^J$ when the conjugacy property of the PT is lost. Define the vector probabilities of the final partition at level J as $p_{\mathcal{Y}^J} = (p_{\mathcal{Y}^J}(1), p_{\mathcal{Y}^J}(2), \ldots, p_{\mathcal{Y}^J}(2^J))$. Hanson [39] showed that

$$p_{\mathcal{Y}^J}(k) = \prod_{j=1}^{J} Y_{e_j(\lceil k2^{j-J} \rceil)},$$

where $e_j(k) = \varepsilon_1 \cdots \varepsilon_j$ is the j-fold binary representation of $k-1$, and $\lceil x \rceil$ is the ceiling function. Based on this expression, Hanson [39] provided exact formulas for the CDF, density and quantile function of a FPT model. Let $k_{\boldsymbol{\eta}}(j, x)$ be the index $k \in \{1, \ldots, 2^J\}$ such that x falls into set $B^{\boldsymbol{\eta}}_{e_j(k)}$, then the cumulative distribution function, density, and quantile functions are given by

$$G(x \mid \mathcal{Y}^J, \boldsymbol{\eta}) = \left\{ \sum_{k=1}^{k_{\boldsymbol{\eta}}(J,x)-1} p_{\mathcal{Y}^J}(k) \right\} + p_{\mathcal{Y}^J}(k_{\boldsymbol{\eta}}(J, x))[2^J G_{\boldsymbol{\eta}}(x) - k_{\boldsymbol{\eta}}(J, x) + 1],$$

$$g(x \mid \mathcal{Y}^J, \boldsymbol{\eta}) = 2^J p_{\mathcal{Y}^J}(k_{\boldsymbol{\eta}}(J, x)) g_{\boldsymbol{\eta}}(x),$$

and

$$G^{-1}(p \mid \mathcal{Y}^J, \boldsymbol{\eta}) = G^{-1}_{\boldsymbol{\eta}} \left\{ \frac{p - \sum_{k=1}^{N} p_{\mathcal{Y}^J}(k) + N p_{\mathcal{Y}^J}(N)}{2^J p_{\mathcal{Y}^J}(N)} \right\},$$

respectively, where N is such that $\sum_{k=1}^{N-1} p_{\mathcal{Y}^J}(k) < p \leq \sum_{k=1}^{N} p_{\mathcal{Y}^J}(k)$. Those expressions can be used to construct the likelihood and latent variable distribution in different settings. Hanson [39] considers simple Metropolis–Hastings updates of the elements of $\mathcal{Y}^J$, where the candidates $(Y^*_{e_j(k)0}, Y^*_{e_j(k)1})$ are generated from a beta distribution with parameters $(mY_{e_j(k)0}, mY_{e_j(k)1})$, with $m > 0$.

5. Concluding remarks

BNP statistics is a relative new area of statistics which is growing rapidly. A number of themes are in continuous development including theory, methodology and applications. BNP methods are extremely powerful and have a wide range of applicability within several prominent domains of statistics.

In this paper, several BNP approaches have been discussed to allow uncertainty in distributional assumptions and to avoid critical dependence on parametric assumptions in the context of hierarchical models, including a detailed description of the computational methods used for fitting them.

We have not discussed important aspects of BNP methods, including priors for other infinite-dimensional parameters of interest and the evaluation of asymptotic properties. We refer the reader to S. Ghoshal's chapter in [11] for an excellent recent overview on posterior asymptotics for BNP models.

Acknowledgements

A. Jara was supported by FONDECYT 1141193 grant from the Chilean government.

References

[1] P. Müller, F.A. Quintana, A. Jara, T. Hanson, Bayesian Nonparametric Data Analysis, Springer, New York, USA, 2015.
[2] A.Y. Lo, On a class of Bayesian nonparametric estimates I: density estimates, Ann. Stat. 12 (1984) 351–357.
[3] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Stat. 1 (1973) 209–230.
[4] T.S. Ferguson, Prior distribution on the spaces of probability measures, Ann. Stat. 2 (1974) 615–629.
[5] S. Ghosal, J.K. Ghosh, R.V. Ramamoorthi, Posterior consistency of Dirichlet mixtures in density estimation, Ann. Stat. 27 (1999) 143–158.
[6] W. Shen, S.T. Tokdar, S. Ghosal, Adaptive Bayesian multivariate density estimation with Dirichlet mixtures, Biometrika 100 (2003) 623–640.
[7] A. Lijoi, I. Prünster, S. Walker, On consistency of non-parametric normal mixtures for Bayesian density estimation, J. Am. Stat. Assoc. 100 (2005) 1292–1296.
[8] S. Ghosal, A.W. Van der Vaart, Posterior convergence rates of Dirichlet mixtures at smooth densities, Ann. Stat. 35 (2007) 697–723.
[9] D. Dey, P. Müller, D. Sinha, Practical Nonparametric and Semiparametric Bayesian Statistics, Springer, New York, USA, 1998.
[10] T. Hanson, A. Branscum, W. Johnson, Bayesian nonparametric modeling and data analysis: an introduction, in: D.K. Dey, C.R. Rao (Eds.), Bayesian Thinking: Modeling and Computation, in: Handbook of Statistics, vol. 25, Elsevier, Amsterdam, The Netherlands, 2005, pp. 245–278.
[11] N.L. Hjort, C. Holmes, P. Müller, S. Walker, Bayesian Nonparametrics, Cambridge University Press, Cambridge, UK, 2010.
[12] D. Blackwell, Discreteness of Ferguson selection, Ann. Stat. 1 (1973) 356–358.
[13] D. Blackwell, J. MacQueen, Ferguson distributions via Pólya urn schemes, Ann. Stat. 1 (1973) 353–355.
[14] B. de Finetti, Foresight: its logical laws, its subjective sources, in: H.E. Kyburg, H.E. Smokler (Eds.), Studies in Subjective Probability, John Wiley and Sons, New York, USA, 1937, pp. 53–118.
[15] R.M. Korwar, M. Hollander, Contributions to the theory of Dirichlet processes, Ann. Probab. 1 (1973) 705–711.
[16] J. Sethuraman, A constructive definition of Dirichlet prior, Stat. Sin. 2 (1994) 639–650.
[17] P.D. Feigin, R.L. Tweedie, Linear functionals and Markov chains associated with Dirichlet processes, Math. Proc. Camb. Philos. Soc. 105 (1989) 579–585.
[18] C.E. Antoniak, Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems, Ann. Stat. 2 (1974) 1152–1174.
[19] S. Petrone, Bayesian density estimation using Bernstein polynomials, Can. J. Stat. 27 (1999) 105–126.
[20] S. Petrone, Random Bernstein polynomials, Scand. J. Stat. 26 (1999) 373–393.
[21] A.F. Barrientos, A. Jara, F.A. Quintana, Bayesian density estimation for compositional data using random Bernstein polynomials, J. Stat. Plan. Inference 166 (2015) 116–125.
[22] S.R. Dalal, Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions, Stoch. Process. Appl. 9 (1979) 99–107.
[23] H. Doss, Bayesian nonparametric estimation of the median. I. Computation of the estimates, Ann. Stat. 13 (1985) 1432–1444.
[24] H. Doss, Bayesian nonparametric estimation of the median. II. Asymptotic properties of the estimates, Ann. Stat. 13 (1985) 1445–1464.
[25] M.A. Newton, C. Czado, R. Chapell, Bayesian inference for semiparametric binary regression, J. Am. Stat. Assoc. 91 (1996) 142–153.
[26] D. Freedman, On the asymptotic distribution of Bayes' estimates in the discrete case, Ann. Math. Stat. 34 (1963) 1386–1403.
[27] J. Fabius, Asymptotic behavior of Bayes' estimates, Ann. Math. Stat. 35 (1964) 846–856.
[28] R.D. Mauldin, W.D. Sudderth, S.C. Williams, Polya trees and random distributions, Ann. Stat. 20 (1992) 1203–1221.
[29] M. Lavine, Some aspects of Polya tree distributions for statistical modeling, Ann. Stat. 20 (1992) 1222–1235.
[30] M. Lavine, More aspects of Polya tree distributions for statistical modeling, Ann. Stat. 22 (1994) 1161–1176.
[31] R. Christensen, T. Hanson, A. Jara, Parametric nonparametric statistics: an introduction to mixtures of finite Polya trees, Am. Stat. 62 (2008) 296–306.
[32] M. Monticino, How to construct a random probability measure, Int. Stat. Rev. 69 (2001) 153–167.
[33] L.E. Dubins, D.A. Freedman, Random distribution functions, in: Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability, vol. 2, 1967, pp. 183–214.
[34] C.M. Kraft, A class of distribution function processes which have derivatives, J. Appl. Probab. 1 (1964) 385–388.
[35] M. Metivier, Sur la construction de mesures aleatoires presque surement absolument continues par rapport a une mesure donnee, Z. Wahrscheinlichkeitstheor. Verw. Geb. 20 (1971) 332–334.
[36] M.J. Schervish, Theory of Statistics, Springer, New York, USA, 1995.
[37] S.G. Walker, B.K. Mallick, Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing, J. R. Stat. Soc. B 59 (1997) 845–860.
[38] T. Hanson, W.O. Johnson, Modeling regression error with a mixture of Polya trees, J. Am. Stat. Assoc. 97 (2002) 1020–1033.

[39] T. Hanson, Inference for mixtures of finite Polya tree models, J. Am. Stat. Assoc. 101 (2006) 1548–1565.
[40] S.M. Paddock, F. Ruggeri, M. Lavine, M. West, Randomized Polya tree models for nonparametric Bayesian inference, Stat. Sin. 13 (2003) 443–460.
[41] A. Jara, T. Hanson, E. Lesaffre, Robustifying generalized linear mixed models using a new class of mixture of multivariate Polya trees, J. Comput. Graph. Stat. 18 (2009) 838–860.
[42] T. Hanson, J.V.D. Monteiro, A. Jara, The Polya tree sampler: toward efficient and automatic independent Metropolis proposals, J. Comput. Graph. Stat. 20 (1) (2011) 41–62.
[43] P. Muliere, L. Tardella, Approximating distributions of random functionals of Ferguson–Dirichlet priors, Can. J. Stat. 26 (1998) 283–297.
[44] P. Muliere, P. Secchi, A Note on a Proper Bayesian Bootstrap, Tech. rep., Università degli Studi di Pavia, Dipartamento di Economia Politica e Metodi Quantitativ, 1995.
[45] J. Pitman, Some developments of the Blackwell–MacQueen urn scheme, in: T.S. Ferguson, L.S. Shapeley, J.B. MacQueen (Eds.), Statistics, Probability and Game Theory. Papers in Honor of David Blackwell, in: IMS Lecture Notes – Monograph Series, Hayward, California, 1996, pp. 245–268.
[46] E. Regazzini, A. Lijoi, I. Prünster, Distributional results for means of normalized random measures with independent increments, Ann. Stat. 31 (2003) 560–585.
[47] H. Ishwaran, L.F. James, Gibbs sampling methods for stick-breaking priors, J. Am. Stat. Assoc. 96 (2001) 161–173.
[48] J. Pitman, M. Yor, The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator, Ann. Probab. 25 (1997) 855–900.
[49] H. Ishwaran, M. Zarepour, Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models, Biometrika 87 (2000) 371–390.
[50] J.F.C. Kingman, Random discrete distributions, J. R. Stat. Soc. B 37 (1975) 1–22.
[51] D. Cifarelli, E. Regazzini, Problemi statistici non parametrici in condizioni di scambialbilita parziale e impiego di medie associative, Tech. rep., Quaderni Istituto Matematica Finanziaria, Torino, 1978.
[52] P. Muliere, S. Petrone, A Bayesian predictive approach to sequential search for an optimal dose: parametric and nonparametric models, J. Ital. Stat. Soc. 2 (1993) 349–364.
[53] A. Mira, S. Petrone, Bayesian hierarchical nonparametric inference for change-point problems, in: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (Eds.), Bayesian Statistics, vol. 5, Oxford University Press, 1996.
[54] P. Giudici, M. Mezzetti, P. Muliere, Mixtures of Dirichlet process priors for variable selection in survival analysis, J. Stat. Plan. Inference 111 (2003) 101–115.
[55] S.N. MacEachern, Dependent nonparametric processes, in: ASA Proceedings of the Section on Bayesian Statistical Science, American Statistical Association, Alexandria, VA, 1999.
[56] S.N. MacEachern, Dependent Dirichlet Processes, Tech. rep., Department of Statistics, The Ohio State University, 2000.
[57] A.F. Barrientos, A. Jara, F.A. Quintana, On the support of MacEachern's dependent Dirichlet processes and extensions, Bayesian Anal. 7 (2012) 277–310.
[58] M. De Iorio, P. Müller, G.L. Rosner, S.N. MacEachern, An ANOVA model for dependent random measures, J. Am. Stat. Assoc. 99 (2004) 205–215.
[59] M. De Iorio, W.O. Johnson, P. Müller, G.L. Rosner, Bayesian nonparametric non-proportional hazards survival modelling, Biometrics 65 (2009) 762–771.
[60] A. Jara, E. Lesaffre, M. De Iorio, F.A. Quintana, Bayesian semiparametric inference for multivariate doubly-interval-censored data, Ann. Appl. Stat. 4 (2010) 2126–2149.
[61] A.E. Gelfand, A. Kottas, S.N. MacEachern, Bayesian nonparametric spatial modeling with Dirichlet process mixing, J. Am. Stat. Assoc. 100 (2005) 1021–1035.
[62] D.B. Dunson, A.H. Herring, Semiparametric Bayesian Latent Trajectory Models, Tech. rep., ISDS Discussion Paper 16, Duke University, Durham, NC, USA, 2006.
[63] P. Müller, G.L. Rosner, M. De Iorio, S. MacEachern, A nonparametric Bayesian model for inference in related longitudinal studies, J. R. Stat. Soc. C 54 (2005) 611–626.
[64] V. Iñacio, A. Jara, T.E. Hanson, M. de Carvalho, Bayesian nonparametric roc regression modeling, Bayesian Anal. 8 (2013) 623–646.
[65] P. Müller, A. Erkanli, M. West, Bayesian curve fitting using multivariate normal mixtures, Biometrika 83 (1996) 67–79.
[66] P. Müller, F.A. Quintana, G. Rosner, A method for combining inference across related nonparametric Bayesian models, J. R. Stat. Soc. B 66 (2004) 735–749.
[67] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical Dirichlet processes, J. Am. Stat. Assoc. 101 (2006) 1566–1581.
[68] J.E. Griffin, M.F.J. Steel, Order-based dependent Dirichlet processes, J. Am. Stat. Assoc. 101 (2006) 179–194.
[69] A. Rodriguez, D.B. Dunson, A. Gelfand, The nested Dirichlet process, J. Am. Stat. Assoc. 103 (2008) 1131–1154.
[70] D.B. Dunson, N. Pillai, J.H. Park, Bayesian density regression, J. R. Stat. Soc. B 69 (2007) 163–183.
[71] D.B. Dunson, J.H. Park, Kernel stick-breaking processes, Biometrika 95 (2008) 307–323.
[72] D.B. Dunson, Y. Xue, L. Carin, The matrix stick-breaking process: flexible Bayes meta-analysis, J. Am. Stat. Assoc. 103 (2008) 317–327.
[73] Y. Chung, D.B. Dunson, The local Dirichlet process, Ann. Inst. Stat. Math. 63 (2011) 59–80.
[74] L. Ren, L. Du, L. Carin, D.B. Dunson, Logistic stick-breaking process, J. Mach. Learn. Res. 12 (2011) 203–239.
[75] Y. Chung, D.B. Dunson, Nonparametric Bayes conditional distribution modeling with variable selection, J. Am. Stat. Assoc. 104 (2009) 1646–1660.
[76] A. Rodriguez, D.B. Dunson, Nonparametric Bayesian models through probit stick-breaking processes, Bayesian Anal. 6 (2011) 145–178.
[77] P. Müller, F.A. Quintana, Random partition models with regression on covariates, J. Stat. Plan. Inference 140 (2010) 2801–2808.
[78] P. Müller, F.A. Quintana, G.L. Rosner, A product partition model with regression on covariates, J. Comput. Graph. Stat. 20 (2011) 260–278.
[79] F.A. Quintana, Linear regression with a dependent skewed Dirichlet process, Chil. J. Stat. 1 (2010) 35–49.
[80] A.F. Barrientos, A. Jara, F.A. Quintana, Fully nonparametric regression for bounded data using dependent Bernstein polynomials, J. Am. Stat. Assoc. (2016), in press.
[81] I. Epifani, A. Lijoi, Nonparametric priors for vectors of survival functions, Stat. Sin. 20 (2010) 1455–1484.
[82] F. Leisen, A. Lijoi, Vectors of two-parameter Poisson–Dirichlet processes, J. Multivar. Anal. 102 (2011) 482–495.
[83] A. Lijoi, B. Nipoti, I. Prünster, Bayesian inference with dependent normalized completely random measures, Bernoulli 20 (2014) 1260–1291.
[84] S.T. Tokdar, Y.M. Zhu, J.K. Ghosh, Bayesian density regression with logistic Gaussian process and subspace projection, Bayesian Anal. 5 (2010) 1–26.
[85] A. Jara, T. Hanson, A class of mixtures of dependent tail-free processes, Biometrika 98 (2011) 553–566.
[86] M.D. Escobar, Estimating the means of several normal populations by nonparametric estimation of the distributions of the means, Unpublished doctoral thesis, Deparment of Statistics, Yale University, 1988.
[87] M.D. Escobar, Estimating normal means with a Dirichlet process prior, J. Am. Stat. Assoc. 89 (1994) 268–277.
[88] M.D. Escobar, M. West, Bayesian density estimation and inference using mixtures, J. Am. Stat. Assoc. 90 (1995) 577–588.
[89] J.S. Liu, Nonparametric hierarchical Bayes via sequential imputations, Ann. Stat. 24 (1996) 911–930.
[90] S.N. MacEachern, M. Clyde, J.S. Liu, Sequential importance sampling for nonparametric Bayes models: the next generation, Can. J. Stat. 27 (1999) 251–267.
[91] M.A. Newton, F.A. Quintana, Y. Zhang, Nonparametric Bayes methods using predictive updating, in: D. Dey, P. Müller, D. Sinha (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics, Springer, 1998, pp. 45–62.
[92] M.A. Newton, Y. Zhang, A recursive algorithm for nonparametric analysis with missing data, Biometrika 86 (1999) 15–26.

Please cite this article in press as: A. Jara, Theory and computations for the Dirichlet process and related models: An overview, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.11.008

[93] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul, An introduction to variational methods for graphical models, Mach. Learn. 37 (1999) 183–233.
[94] D. Blei, M. Jordan, Variational inference for Dirichlet process mixtures, Bayesian Anal. 1 (2006) 121–144.
[95] C.A. Bush, S.N. MacEachern, A semiparametric Bayesian model for randomised block designs, Biometrika 83 (1996) 275–285.
[96] S.N. MacEachern, Estimating normal means with a conjugate style Dirichlet process prior, Commun. Stat., Simul. Comput. 23 (1994) 727–741.
[97] S. Jain, R.M. Neal, A split-merge Markov Chain Monte Carlo procedure for the Dirichlet process mixture model, J. Comput. Graph. Stat. 13 (2004) 158–182.
[98] D.B. Dahl, Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models, Tech. rep., Technical Report, Texas AM University, USA, 2005.
[99] D.B. Phillips, A.F.M. Smith, Bayesian model comparisons via jump diffusions, in: W.R. Gilks, S. Richardson, D.J. Spiegelhalter (Eds.), Markov Chain Monte Carlo in Practice, Chapman and Hall, New York, USA, 1996, pp. 215–239.
[100] S. Richardson, P.J. Green, On Bayesian analysis of mixtures with an unknown number of components, J. R. Stat. Soc. B 59 (1997) 731–792.
[101] Y. Fong, J. Wakefield, K. Rice, An efficient Markov chain Monte Carlo method for mixture models by neighborhood pruning, J. Comput. Graph. Stat. 21 (2012) 197–216.
[102] S.N. MacEachern, P. Müller, Estimating mixture of Dirichlet process models, J. Comput. Graph. Stat. 7 (2) (1998) 223–338.
[103] R. Neal, Markov chain sampling methods for Dirichlet process mixture models, J. Comput. Graph. Stat. 9 (2000) 249–265.
[104] H. Doss, Bayesian nonparametric estimation for incomplete data via successive substitution sampling, Ann. Stat. 22 (1994) 1763–1786.
[105] J.-P. Florens, J.-M. Rolin, Simulation of Posterior Distributions in Nonparametric Censored Analysis, Papers 98.493, GREMAQ, Toulouse, 1998, available at http://ideas.repec.org/p/fth/gremaq/98.493.html, 1998.
[106] T. Hanson, W.O. Johnson, A Bayesian semiparametric AFT model for interval-censored data, J. Comput. Graph. Stat. 13 (2004) 341–361.
[107] O. Papaspiliopoulos, G.O. Roberts, Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models, Biometrika 95 (2008) 169–186.
[108] S.G. Walker, Sampling the Dirichlet mixture model with slices, Commun. Stat. Simul. Comp. 36 (2007) 45–54.
[109] M. Kalli, J.E. Griffin, S. Walker, Slice sampling mixture models, Stat. Comput. 21 (2011) 93–105.
[110] A.E. Gelfand, A. Kottas, A computational approach for full nonparametric Bayesian inference under Dirichlet Process Mixture models, J. Comput. Graph. Stat. 11 (2002) 289–304.
[111] A. Jara, Applied Bayesian non- and semi-parametric inference using DPpackage, RNews 7 (2007) 17–26.
[112] A. Jara, T. Hanson, F. Quintana, P. Müller, G.L. Rosner, DPpackage: Bayesian semi- and nonparametric modeling in R, J. Stat. Softw. 40 (2011) 1–30.
[113] H. Ishwaran, L.F. James, Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information, J. Comput. Graph. Stat. 11 (2002) 508–532.

Please cite this article in press as: A. Jara, Theory and computations for the Dirichlet process and related models: An overview, Int. J. Approx. Reason. (2016), http://dx.doi.org/10.1016/j.ijar.2016.11.008