Statistical Science
2022, Vol. 37, No. 1, 24–41
https://doi.org/10.1214/20-STS819

The Dependent Dirichlet Process and Related Models

Fernando A. Quintana, Peter Müller, Alejandro Jara and Steven N. MacEachern

Abstract. Standard regression approaches assume that some finite number of the response distribution characteristics, such as location and scale, change as a (parametric or nonparametric) function of predictors. However, it is not always appropriate to assume a location/scale representation, where the error distribution has unchanging shape over the predictor space. In fact, it often happens in applied research that the distribution of responses under study changes with predictors in ways that cannot be reasonably represented by a finite dimensional functional form. This can seriously affect the answers to the scientific questions of interest, and therefore more general approaches are indeed needed. This gives rise to the study of fully nonparametric regression models. We review some of the main Bayesian approaches that have been employed to define probability models where the complete response distribution may vary flexibly with predictors. We focus on developments based on modifications of the Dirichlet process, historically termed dependent Dirichlet processes, and some of the extensions that have been proposed to tackle this general problem using nonparametric approaches.

Key words and phrases: Related random probability distributions, Bayesian nonparametrics, nonparametric regression, quantile regression.

1. INTRODUCTION

We review the popular class of dependent Dirichlet process (DDP) models. These define a widely used fully nonparametric Bayesian regression for a response $\boldsymbol{y} \in \mathscr{Y}$, based on a set of predictors $\boldsymbol{x} \in \mathscr{X} \subseteq \mathbb{R}^p$. Despite a barrage of related literature over the past 25 years, to date there is no good review of such models. This paper fills this gap.

Fernando A. Quintana is Professor, Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile, and Deputy Director, ANID—Millennium Science Initiative Program—Millennium Nucleus Center for the Discovery of Structures in Complex Data, Santiago, Chile (e-mail: quintana@mat.uc.cl). Peter Müller is Professor, Department of Statistics and Data Science, University of Texas at Austin, Austin, Texas, USA (e-mail: pmueller@math.utexas.edu). Alejandro Jara is Associate Professor, Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile, and Director, ANID—Millennium Science Initiative Program—Millennium Nucleus Center for the Discovery of Structures in Complex Data, Santiago, Chile (e-mail: atjara@uc.cl). Steven N. MacEachern is Professor, Department of Statistics, Ohio State University, Columbus, Ohio, USA (e-mail: snm@stat.osu.edu).

Fully nonparametric regression can be seen as an extension of traditional regression models, where, starting from some elements in $\mathscr{X}$ and a corresponding set of responses in $\mathscr{Y}$, the goal is to model the distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$. Standard linear regression models proceed under the assumption of a Gaussian distribution for $\boldsymbol{y} \mid \boldsymbol{x}$ with a mean modeled as a linear combination of $\boldsymbol{x}$. Further extensions of this idea to exponential families gave rise to the popular class of generalized linear models, where a transformation of the mean response is modeled as a linear combination of $\boldsymbol{x}$. Many other similar extensions are available. We focus on a nonparametric version of this idea, which involves going beyond the notion that the effect of predictors is restricted to change some particular functional of the response distribution, such as the mean, a quantile, or the parameters in a generalized linear model.

The fully nonparametric regression problem that we focus on arises when we assume that $\boldsymbol{y}_i \mid F_{\boldsymbol{x}_i} \overset{\text{ind}}{\sim} F_{\boldsymbol{x}_i}$, $i = 1, \ldots, n$. The parameter of interest is the complete set of predictor-dependent random probability measures $\mathscr{F} = \{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\}$, where $F_{\boldsymbol{x}}$ is a probability measure defined on the response sample space $\mathscr{Y}$, whose elements can flexibly change with the values of the predictors $\boldsymbol{x}$, that is, the entire shape of the distribution can change with $\boldsymbol{x}$. From a Bayesian point of view, the fully nonparametric regression model is completed by defining a

prior distribution for $\mathscr{F}$, which is taken to be the probability law of a probability measure-valued stochastic process with index $\boldsymbol{x}$. At the risk of abusing notation, we use from now on the same symbol to refer to the probability measure and its cumulative distribution function (CDF). The distinction should be clear from the context.

Several popular approaches have been developed to formalize Bayesian inference for such nonparametric regression. These include additive random tree models like the BART (Chipman, George and McCulloch, 2010), approaches based on basis expansions such as wavelet regression and more. Also, there is of course extensive literature on non-Bayesian approaches to nonparametric regression. Many of these approaches are based on a model of the form

$$\boldsymbol{y}_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

with $E(\epsilon_i) = 0$, and are concerned with finding a function $f : \mathscr{X} \to \mathscr{Y}$ such that $\|y_i - f(\boldsymbol{x}_i)\|$ is small, for f in some some class, often represented as being spanned by some basis functions. Such methods include the following: Under *local averaging* $f(\boldsymbol{x})$ is estimated from those $\boldsymbol{y}_i$'s such that $\boldsymbol{x}_i$ is "close" to $\boldsymbol{x}$; *local modeling* estimates $f(\boldsymbol{x})$ by locally fitting some function or kernel such as a Gaussian function or a polynomial; *global modeling or least squares estimation* finds f that minimizes $\frac{1}{n}\sum_{i=1} \|\boldsymbol{y}_i - f(\boldsymbol{x}_i)\|^2$ in the class; and *penalized modeling* is based on finding f that minimizes $\frac{1}{n}\sum_{i=1} \|\boldsymbol{y}_i - f(\boldsymbol{x}_i)\|^2 + J_n(f)$ in the class, where $J_n(f)$ is a penalization term, such as $J_n(f) = \lambda_n \int_{\mathscr{X}} |f''(t)|^2\, dt$. See, for example, Györfi et al. (2002), Klemelä (2014), Faraway (2016) and references within. Many of these classical frequentist approaches could be construed to imply nonparametric Bayesian models, but they are not usually cast as prior probability models for a family $\mathscr{F}$ of random probability measures indexed by covariates.

In the Bayesian nonparametric (BNP) literature, the problem of defining priors over related random probability distributions has received increasing attention over the past few years. To date, most of the BNP priors to account for the dependence of a set of probability distributions on predictors are generalizations and extensions of the celebrated Dirichlet process (DP) (Ferguson, 1973, 1974) and Dirichlet process mixture (DPM) models (Lo, 1984). A DPM model defines a random probability measure as

$$f(\boldsymbol{y} \mid G) = \int_{\Theta} \psi(\boldsymbol{y}, \boldsymbol{\theta}) G(d\boldsymbol{\theta}), \quad \boldsymbol{y} \in \mathscr{Y}, \tag{1}$$

where $\psi(\bullet, \boldsymbol{\theta})$ is a continuous density function, for every $\boldsymbol{\theta} \in \Theta$, and G is a discrete random probability measure with a DP prior. If G is DP with parameters (M, G_0), where $M \in \mathbb{R}_0^+$ and G_0 is a probability measure on Θ, written as $G \mid M, G_0 \sim \mathrm{DP}(MG_0)$, then the trajectories of the process can be a.s. represented by the stick-breaking representation (Sethuraman, 1994):

$$G(B) = \sum_{h=1}^{\infty} w_h \delta_{\boldsymbol{\theta}_h}(B), \tag{2}$$

where B is any measurable set, $\delta_{\boldsymbol{\theta}}(\cdot)$ is the Dirac measure at $\boldsymbol{\theta}$, $w_h = V_h \prod_{\ell<h}(1 - V_\ell)$, with $V_h \mid M \overset{\text{iid}}{\sim} \mathrm{Be}(1, M)$, $\boldsymbol{\theta}_h \mid G_0 \overset{\text{iid}}{\sim} G_0$, and the $\{w_h\}$ and $\{\boldsymbol{\theta}_h\}$ collections are independent. Discussion of properties and applications of DPs can be found, for instance, in Müller et al. (2015). Many BNP priors for nonparametric regressions $\mathscr{F} = \{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\}$ are based on extensions of model (1). They incorporate dependence on predictors via the mixing distribution in (1), by replacing G with $G_{\boldsymbol{x}}$, and the prior specification problem is related to the modeling of the collection of predictor-dependent mixing probability measures $\{G_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\}$.

Consider first the simplest case, where a finite number of dependent RPMs $\mathcal{G} = \{G_j, j = 1, \ldots, J\}$ are judged to be exchangeable so that the prior model $p(\mathcal{G})$ should accordingly be invariant with respect to all permutations of the indices. Consider, for example, an application to borrowing strength across J related clinical studies. This can be achieved, for example, through joint modeling of study-specific effects distributions G_j for $j = 1, \ldots, J$. A main aim here is that subjects under study j_1 should inform inference about subjects enrolled in a different but related study $j_2 \neq j_1$. Two extreme modeling choices would be (i) to pool all patients and assume one common effects distribution, or (ii) to assume J distinct distributions with independent priors. Formally, the earlier choice assumes $G_j \equiv G$, $j = 1, \ldots, J$, with a prior $p(G)$, such as $G \sim \mathrm{DP}(M, G_0)$. The latter assumes $G_j \sim p(G_j)$, independently, $j = 1, \ldots, J$. We refer to the two choices as extremes since the first choice implies maximum borrowing of strength, and the other choice implies no borrowing of strength. In most applications, the desired level of borrowing strength is somewhere in-between these two extremes.

Figure 1 illustrates the two modeling approaches. Note that in Figure 1 we added a hyperparameter η to index the prior model $p(G_j \mid \eta)$ and $p(G \mid \eta)$, which was implicitly assumed fixed. The use of a random hyperparameter η allows for some borrowing of strength even in the case of conditionally independent $p(G_j \mid \eta)$. Learning across studies can happen through learning about the hyperparameter η. However, the nature of the learning across studies is determined by the parametric form of η. This is illustrated in Figure 2. Assume $G_j \sim \mathrm{DP}(M, G_\eta^\star)$, independently, $j = 1, 2$, and a base measure $G_\eta^\star = \mathrm{N}(m, B)$ with unknown hyperparameter $\eta = (m, B)$. In this case, prediction for a future study G_3 can not possibly learn about the multimodality of G_1 and G_2, beyond general location and orientation.

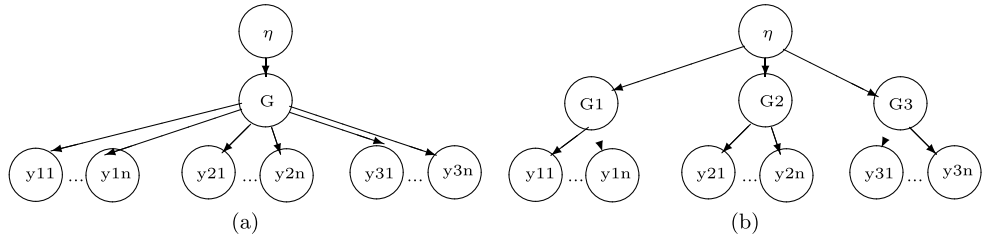
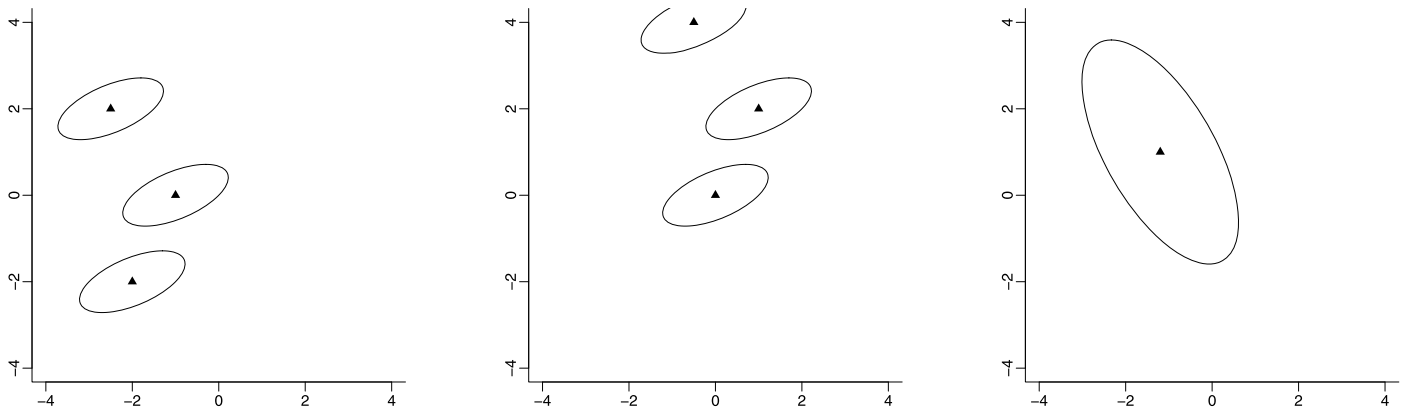


FIG. 1. *One common RPM G (panel a) versus distinct RPMs G_j, independent across studies (panel b). Here η is a fixed hyperparameter.*

The previous simple example illustrates the need to develop classes of models with the ability to relate collections of nonparametric distributions in more complex fashions. When this collection is indexed by a set of predictors $\boldsymbol{x} \in \mathscr{X}$, the nonparametric regression approach mentioned earlier arises, and the definition of a prior on this collection enables one to borrow information across the distributions for responses, $F_{\boldsymbol{x}}$. For modeling, one important property is the notion of distributions changing smoothly with respect to $\boldsymbol{x} \in \mathscr{X}$, just as is the case of generalized linear models in the scale of the transformed mean. The smoothness could be expressed as continuity of $F_{\boldsymbol{x}}$ (with respect to some conveniently chosen topology) or as the notion that $F_{\boldsymbol{x}}$ "approaches" $F_{\boldsymbol{x}_0}$ as $\boldsymbol{x} \to \boldsymbol{x}_0$, for instance, $\mathrm{Corr}\{F_{\boldsymbol{x}}(A), F_{\boldsymbol{x}_0}(A)\} \to 1$ as $\boldsymbol{x} \to \boldsymbol{x}_0$ for any event A. Many of the models to be discussed later satisfy some version of this property.

An early reference on predictor-dependent DP models is Cifarelli and Regazzini (1978), 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 (1993), who considered a linear regression model for the centering distribution of the form $G^0_{\boldsymbol{x}} \equiv N(\boldsymbol{x}'\beta, \sigma^2)$, where $\beta \in \mathbb{R}^p$ is a vector of regression coefficients, and $N(\mu, \sigma^2)$ stands for a normal distribution with mean μ and variance σ^2. This is the type of construction illustrated in Figure 2. Similar models were discussed by Mira and Petrone (1996) and Giudici, Mezzetti and Muliere (2003). Linking the related nonparametric models through a regression on the baseline parameters of nonparametric models, however, limits the nature of the trajectories and the type of dependent processes that can be thus generated. Indeed, realizations of the resulting process $\mathcal{G} = \{G_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\}$ are not continuous as a function of the predictors. The very limited type of association structure motivated the development of alternative extensions of the DP model to a prior for $\mathcal{G}$. In this paper, we provide an overview of the main constructions of such predictor-dependent extensions of DP priors and their main properties. The discussion centers on different ways of constructing the nonparametric component of models. A few of the many successful types of applications that have been proposed are mentioned. In reviewing the various models to be presented, we discuss some of the main corresponding works without attempting to provide a complete catalog of references. We include a brief discussion of other popular constructions of dependent DP random measures, without the explicit notion of a conditioning covariate x.

While we focus on DP-based constructions, we note that several interesting alternatives to develop predictor-driven random probability measures have been considered



(a) $E(G_1 \mid \boldsymbol{y})$ (b) $E(G_2 \mid \boldsymbol{y})$ (c) Prediction for G_3

FIG. 2. *$G_j \sim \mathrm{DP}(M, G^\star)$ with common $G^\star = \mathrm{N}(m, B)$. Learning across studies is restricted to the parametric form of η. The obvious common structure of G_1 and G_2 as defining three well separated clusters can not be learned by the model, which is restricted to learning through the common hyperparameters η.*

in the recent literature. Tokdar, Zhu and Ghosh (2010) develop a logistic Gaussian process that allows for smoothly varying dependence on conditioning variables. Still using Gaussian process priors, but starting from a rather different construction, Jara and Hanson (2011) proposed another alternative, putting the Gaussian process prior on the (logit transformation) of the branch probabilities in a Pólya tree prior (Lavine, 1992). Another covariate-dependent extension of the Pólya tree model was introduced in Trippa, Müller and Johnson (2011) who define a dependent multivariate process for the branch probabilities based on a simple gamma process construction.

Finally, although issues pertaining to the implementation of posterior simulation are relevant for practical application of these methods, our discussion does not focus on computational aspects. In Section 2, we describe MacEachern's dependent Dirichlet process (DDP) and its basic properties. In Section 3, we discuss the main variations and alternative constructions to MacEachern's DDP. In Section 4, we discuss approaches to handle endogenous predictors. In Section 5, we discuss the implied partition structure of DDP models. In Section 6 we illustrate the main approaches. A final discussion in Section 7 concludes the article, including some thoughts on future research directions.

2. DEPENDENT DIRICHLET PROCESS (DDP)

We start our discussion with the general definition of DDP and then give details for popular special cases.

2.1 General Definition

MacEachern (1999, 2000) introduced the DDP model as a flexible class of predictor-dependent random probability distributions. The key idea behind the DDP construction is to define a set of random measures that are marginally (i.e., for every possible predictor value $\boldsymbol{x} \in \mathscr{X}$) DP-distributed random measures. In this framework, dependence is introduced through a modification of the stick–breaking representation of each element in the set,

$$G_{\boldsymbol{x}}(\bullet) = \sum_{h=1}^{\infty} \underbrace{\left\{ V_h(\boldsymbol{x}) \prod_{\ell<h} [1 - V_\ell(\boldsymbol{x})] \right\}}_{w_h(\boldsymbol{x})} \delta_{\boldsymbol{\theta}_h(\boldsymbol{x})}(\bullet), \tag{3}$$

where $V_h(\boldsymbol{x})$, $h \in \mathbb{N}$, are $[0,1]$-valued independent stochastic processes with index set $\mathscr{X}$ and $\mathrm{Be}(1, M_{\boldsymbol{x}})$ marginal distributions, and $\boldsymbol{\theta}_h(\boldsymbol{x})$, $h \in \mathbb{N}$, are independent stochastic processes with index set $\mathscr{X}$ and $G_{\boldsymbol{x}}^0$ marginal distributions. The processes associated to the weights and atoms are independent. From an intuitive viewpoint, the constructed DDP can be thought of as taking an ordinary DP and modifying some of its components (i.e., weights and atoms) according to the type of desired indexing or functional dependence of predictors $\boldsymbol{x} \in \mathscr{X}$. Conditions on the $V_h(\boldsymbol{x})$ and $\boldsymbol{\theta}_h(\boldsymbol{x})$ processes can be established to ensure smoothness of the resulting random measures $G_{\boldsymbol{x}}(\bullet)$ when $\boldsymbol{x}$ ranges over $\mathscr{X}$.

Canonical DDP construction. MacEachern (1999, 2000) defined and provided a canonical construction of the DDP by using transformations of two independent sets of stochastic processes, $Z_{\mathscr{X}}^{V_h} = \{Z_h^V(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{X}\}$, and $Z_{\mathscr{X}}^{\boldsymbol{\theta}_h} = \{Z_h^{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{X}\}$, for $h \geq 1$, the former used for defining $\{V_h(\boldsymbol{x})\}$, and the latter for defining $\{\boldsymbol{\theta}_h(\boldsymbol{x})\}$. To induce the desired marginal distributions for $\{V_h(\boldsymbol{x})\}$ and $\{\boldsymbol{\theta}_h(\boldsymbol{x})\}$, MacEachern resorted to the well-known inverse transformation method (see, e.g., Devroye, 1986). For instance, let $Z(x)$ denote a zero-mean Gaussian process on $\mathscr{X} = \mathbb{R}$ having constant variance σ^2. Let $\Phi(\cdot)$ and $B(\cdot)$ denote the cumulative distribution functions of the $N(0,1)$ and $\mathrm{Be}(1, M)$ distributions, respectively. Then $V(x) = B^{-1}(\Phi(\sigma^{-1}Z(x)))$ is a stochastic process on $\mathscr{X}$ that satisfies $V(x) \sim \mathrm{Be}(1, M)$ for all $x \in \mathscr{X}$. The same type of transformation can be applied to construct suitable atom processes $\{\theta_h(x), h \geq 1\}$ such that $\theta_h(x) \sim G_0$ for all $x \in \mathscr{X}$ and $h \geq 1$.

Practical application of this general model requires specification of its various components, which has traditionally motivated the adoption of some specific forms. The most commonly used DDPs assume that covariate dependence is introduced either in the atoms or weights, leaving the other as a collection of random variables exhibiting no covariate indexing, so that the basic DP definition is partially modified but the distributional properties retained. We review these forms in the next section.

Support and an alternative definition. One particularity of MacEachern's DDP definition is that given the sets of stochastic processes, $Z_{\mathscr{X}}^{V_h} = \{Z_h^V(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{X}\}$ and $Z_{\mathscr{X}}^{\boldsymbol{\theta}_h} = \{Z_h^{\boldsymbol{\theta}}(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{X}\}$, and all other parameters involved in the transformations described above, the collection of dependent probability distributions given in (3) are not random: they are just deterministic functions of these quantities. To facilitate the study of theoretical properties of the DDP, Barrientos, Jara and Quintana (2012) gave an alternative definition. This alternative definition exploits the connection between copulas and stochastic processes. Since under certain regularity conditions a stochastic process is completely characterized by its finite-dimensional distributions, it is possible –and useful– to define stochastic processes with given marginal distributions via copulas. The basic idea is to specify the collection of finite dimensional distributions of a process through a collection of copulas and marginal distributions.

Copulas are functions that are useful for describing and understanding the dependence structure between random variables. If H is a d-variate CDF with marginal CDFs given by $F_1, \ldots, F_d$, then by Sklar's theorem (Sklar, 1959), there exists a copula function $C : [0,1]^d \longrightarrow [0,1]$ such that $H(t_1, \ldots, t_d) = C(F_1(t_1), \ldots, F_d(t_d))$, for all

$t_1, \ldots, t_d \in \mathbb{R}$, and this representation is unique if the marginal distributions are absolutely continuous. Thus by the probability integral transform, a copula function is a d-variate CDF on $[0,1]^d$ with uniform marginals on $[0,1]$, which fully captures the dependence among the associated random variables, irrespective of the marginal distributions.

Let $\mathcal{C}^V_{\mathscr{X}} = \{C^V_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d} : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \mathscr{X}, d > 1\}$ and $\mathcal{C}^\theta_{\mathscr{X}} = \{C^\theta_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d} : \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d \in \mathscr{X}, d > 1\}$ be two sets of copulas satisfying Kolmogorov's consistency conditions. In Barrientos, Jara and Quintana's (2012) definition, $V_h(\boldsymbol{x})$, $h \in \mathbb{N}$, are $[0,1]$-valued independent stochastic processes with index set $\mathscr{X}$, with common finite dimensional distributions determined by the set of copulas $\mathcal{C}^V_{\mathscr{X}}$, and $\text{Be}(1, M_{\boldsymbol{x}})$ marginal distributions. Similarly, $\boldsymbol{\theta}_h(\boldsymbol{x})$, $h \in \mathbb{N}$, are independent stochastic processes with index set $\mathscr{X}$, with common finite dimensional distributions determined by the set of copulas $\mathcal{C}^\theta_{\mathscr{X}}$, and $G^0_{\boldsymbol{x}}$ marginal distributions. This alternative construction produces a definition of the DDP exactly as in (3), and in particular, the interpretation of the DDP obtained as modifying a basic DP persists. Furthermore, based on this alternative definition, Barrientos, Jara and Quintana (2012) established basic properties of MacEachern's DDP and other dependent-stick breaking processes. Specifically, they provided sufficient conditions for the full weak support of different versions of the process and also to ensure smoothness of trajectories of $G_{\boldsymbol{x}}(\bullet)$ as $\boldsymbol{x}$ ranges over $\mathscr{X}$. In addition, they also characterized the Hellinger and Kullback–Leibler support of mixtures induced by different versions of the DDP and extended the results to the general class of dependent stick-breaking processes.

2.2 The Single-Weights DDP

MacEachern considered the case of common weights across the values of $\boldsymbol{x}$, also referred to as "single-weights" DDP model, defined as

$$G_{\boldsymbol{x}}(\bullet) = \sum_{h=1}^{\infty} \underbrace{\left\{ V_h \prod_{\ell<h} [1 - V_\ell] \right\}}_{w_h} \delta_{\boldsymbol{\theta}_h(\boldsymbol{x})}(\bullet) \tag{4}$$
$$= \sum_{h=1}^{\infty} w_h \delta_{\boldsymbol{\theta}_h(\boldsymbol{x})}(\bullet),$$

where the V_h's are i.i.d. $\text{Be}(1, M)$ random variables, which are common across all levels of $\boldsymbol{x}$. The $\boldsymbol{\theta}_h(\boldsymbol{x})$'s are independent stochastic processes with index set $\mathcal{X}$ and marginal distributions $G^0_{\boldsymbol{x}}$. In the literature, to this day, this is the most popular form of DDP, mainly due to the fact that posterior simulation can be implemented using the same type of sampling algorithms available for the case of the DP.

2.2.1 *The ANOVA-DDP and linear DDP models.* One of the earliest versions of DDP models was the ANOVA-DDP of De Iorio et al. (2004). Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be a vector of responses (possibly vector-valued) for each of n subjects, and suppose that $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is a corresponding set of covariates. Assume each $\boldsymbol{x}_i$ is in turn a vector of c categorical covariates, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ic})$. Interpret $\boldsymbol{x}_i$ as factors in an ANOVA model, and let d_i denote corresponding design vectors. Assume then that $\boldsymbol{x}_i$ contains all the desired main effects and interactions, as well as desired identifiability constraints. Note that the covariate space $\mathscr{X}$ in this setup is discrete, and so we have a finite number of RPMs. The idea of the ANOVA-DDP models is to encode the covariate dependence in the form of simple linear regressions for the atom processes $\{\theta_h(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{X}\}$. Specifically, this approach uses $\theta_h(\boldsymbol{x}) = \lambda_h' d_{\boldsymbol{x}}$ for $h \geq 1$ where $\{\lambda_h : h \geq 1\}$ is a sequence of i.i.d. random vectors with distribution G_0 and $d_{\boldsymbol{x}}$ is the design vector that corresponds to a generic combination of observed factorial covariates $\boldsymbol{x}$. The model just described implies that (4) becomes

$$G_{\boldsymbol{x}}(\bullet) = \sum_{h=1}^{\infty} w_h \delta_{\lambda_h' d_{\boldsymbol{x}}}(\bullet),$$

that is, a DP mixture of linear models $\lambda_h' d_{\boldsymbol{x}}$. Each element of the collection $\mathcal{G} = \{G_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\}$ has a DP prior distribution with atoms given by $\{\lambda_h' d_{\boldsymbol{x}} : h \geq 1\}$. The elements of $\mathcal{G}$ are correlated because they share a common set of weights and the atoms are originated as linear combinations computed from a single set of parameters, namely $\{w_h : h \geq 1\}$ and $\{\lambda_h : h \geq 1\}$.

To accommodate a continuous response, De Iorio et al. (2004) extended the above construction through a convolution with a continuous kernel, for example, a normal kernel, leading to

$$y_i \mid G_{\boldsymbol{x}_i} \overset{\text{ind}}{\sim} \int N(y_i \mid \mu, \phi)\, dG_{\boldsymbol{x}_i}(\mu)$$
$$= \int N\big(y_i \mid \lambda' d_{\boldsymbol{x}_i}, \phi\big)\, dG(\lambda).$$

The model can be restated by breaking the mixture with the introduction of latent parameters:

$$y_i \mid \lambda_i, \phi \sim N\big(\lambda_i' d_i, \phi\big), \qquad \lambda_1, \ldots, \lambda_n \mid G \overset{\text{iid}}{\sim} G, \tag{5}$$
$$G \sim \text{DP}(M, G_0).$$

The last expression highlights the nature of the model as just a DP mixture of, in this case, normal linear models. The same simplification is possible whenever the atoms $\{\theta_h(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{X}\}$ are indexed by a finite-dimensional parameter vector, like the linear model $\theta_h(\boldsymbol{x}) = \lambda_h' d_{\boldsymbol{x}}$ in this case. The model in (5) is completed with a suitable prior for the precision parameter ϕ, for example, $\phi \sim Ga(a, b)$ if a scalar, or $\phi \sim \text{Wishart}(\nu, S)$ if a matrix. The above

model can be easily modified to mix over scale parameters as well. An immediate consequence of (5) is that the induced marginal distribution for a single response y with design vector $d_{\boldsymbol{x}}$ then becomes a flexible infinite mixture model:

$$y \sim \sum_{h=1}^{\infty} w_h N(y \mid \lambda_h' d_{\boldsymbol{x}}, \phi). \tag{6}$$

We remark here that the hierarchical structure leading to (5) reflects a common practice in the use and application of the DDP. Since marginally each element of the $\mathcal{G}$ family is almost surely discrete (because it is drawn from a DP), models for discrete outcomes are frequently built on convolving the DPs with a continuous kernel, thus yielding a mixture of continuous distributions, which is itself a continuous distribution. In the ANOVA-DDP model of De Iorio et al. (2004), the normal kernel plays precisely this role.

De la Cruz-Mesía, Quintana and Müller (2007) applied the ANOVA-DDP construction to model random effects for longitudinal hormone profiles of pregnant women, where the dependence was on a normal/abnormal pregnancy indicator. This setting was particularly useful for classification purposes. More recently, Gutiérrez et al. (2019) use the ANOVA-DDP framework to propose a multiple testing procedure for comparing several treatments against a control. A further extension of the ANOVA-DDP construction was given in De Iorio et al. (2009), who considered the modeling of nonproportional hazards for survival analysis. They considered a cancer clinical trial, where interest centered on whether high doses of a treatment are more effective than lower doses. The data included additional discrete and continuous covariates, so the model was under the extended ANCOVA-style framework that adds linear combinations of continuous covariates to the ANOVA factorial design.

This same idea can be extended to linear combinations of any given set of covariates, giving rise to the *linear* DDP (LDDP). Specifically, such models involve a linear combination of a set of covariates, as in, for example, general linear models, and so the infinite mixture on the right-hand side of (6) becomes $\sum_{h=1}^{\infty} w_h N(y \mid \lambda_h' \boldsymbol{x}, \phi)$, where $\boldsymbol{x}$ is now the generic value of the (typically vector-valued) covariate. As earlier, the weights $\{w_h\}$ follow a DP-style stick-breaking specification. An analogous expression for a more general kernel function k can be immediately derived. The same type of construction was explored in Jara et al. (2010) in the context of doubly censored outcomes. Their model involves an interval-valued response, corresponding to the observed onset and event times (cavities in the teeth of children from Flanders, Belgium, in their example). Associated with each such response is a latent bivariate vector of true onset and event times, and these are modeled (in the logarithmic scale) using a linear DDP defined in terms of covariates that include deciduous second molars health status and the age at which children started brushing.

2.2.2 *Spatial DDP.* Gelfand, Kottas and MacEachern (2005) define what can be interpreted as a spatial case of a common weight DDP (4) for G_s, with $s \in D \subset \mathbb{R}^d$ being spatial locations and $\theta_h(s)$ generated by a baseline GP, as in the common-weight DDP. However, the focus is not on G_s as in (4), but instead on $\boldsymbol{\theta}_D \sim \sum w_h \delta_{\boldsymbol{\theta}_{h,D}}$, where $\boldsymbol{\theta}_{h,D} = \{\theta_h(s), s \in D\}$. Let $\boldsymbol{s} = (s_1, \ldots, s_n)$ denote a set of n locations at which observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ are made. They consider repeated observations $\boldsymbol{y}_t$, $t = 1, \ldots, T$, with occasion-specific covariates $\boldsymbol{x}_t$. Writing a mixture with respect to a DP random measure as a hierarchical model, they assume

$$\boldsymbol{y}_t \mid \boldsymbol{\theta}_t, \boldsymbol{\beta}, \tau^2 \overset{\text{ind}}{\sim} N(\boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{\theta}_t, \tau^2 \boldsymbol{I}), \qquad \boldsymbol{\theta}_t \mid G^{\eta} \overset{\text{iid}}{\sim} G^{\eta},$$
$$G^{\eta} \sim \text{DP}(M, G_0^{\eta}),$$

where $G_0^{\eta} \equiv N(\boldsymbol{0}, \sigma^2 \boldsymbol{H}(\eta))$ and $\boldsymbol{H}(\eta)$ is a suitable covariance function depending on hyperparameters η.

Dunson and Herring (2006) considered a model for a collection of random functions based on a finite set of latent trajectories described by Gaussian processes. The observations are thus seen as arising from the convolution of a smooth latent trajectory and a noisy Gaussian process. Their motivation came from the study of the relationship between disinfection by-products in the water in early pregnancy and later outcomes. Specifically, denoting by g_i the stochastic process, that is, $\{g_i(t) : t > 0\}$, associated with subject $1 \le i \le n$, Dunson and Herring (2006) assume that

$$g_i = \gamma_i + \epsilon_i, \qquad \gamma_i \overset{\text{iid}}{\sim} G, \qquad \epsilon_i \overset{\text{iid}}{\sim} \text{GP}(\boldsymbol{H}(\eta)),$$

where γ_i is the latent trajectory, and $\text{GP}(\boldsymbol{H}(\eta))$ denotes a Gaussian process with covariance function $\boldsymbol{H}(\eta)$. Their approach specifies the RPM G as $G(\cdot) = \sum_{h=1}^{k} p_h \delta_{\Theta_h}(\cdot)$ with $\Theta_h \sim \text{GP}(\boldsymbol{H}(\eta_{\kappa_h}))$, that is, a finite mixture of atoms given by Gaussian processes with suitable covariance functions. By choosing $\kappa_h = \kappa$ for all h and $(p_1, \ldots, p_k) \sim \text{Dir}(M/k, \ldots, M/k)$, the resulting RPM G approaches $G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\Theta_h}(\cdot)$ as $k \to \infty$ with DP-style weights (see, e.g., Green and Richardson, 2001).

2.2.3 *Dynamic DDP.* The DDP framework has also been used to model dynamic phenomena, by means of a sequence of random distributions that evolve in time. Caron et al. (2008) considered a dynamic linear model formulation to solve this problem, where the state and observation noise distributions where modeled as DP mixtures using two independent DPs so that the mean of the underlying processes is allowed to change in time.

Rodriguez and ter Horst (2008) considered a related model, based on a DDP formulation, where now the atoms

in the infinite mixture are allowed to change in time. Letting y_{it} denote the ith observation at time $1 \le t \le T$, they proposed the model

$$y_{it} \mid G_t \sim \int N(\boldsymbol{F}_{it}'\boldsymbol{\theta}_t, \sigma^2)\, dG_t(\boldsymbol{\theta}_t, \sigma^2),$$

$$G_t(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{(\boldsymbol{\theta}_{ht}^*, \sigma_h^{*2})}(\cdot),$$

$$\boldsymbol{\theta}_{ht}^* \sim N(\boldsymbol{H}_t \boldsymbol{\theta}_{h,t-1}^*, \sigma_h^{*2} \boldsymbol{W}_t),$$

completed with conjugate priors for σ_h^{*2} and $\boldsymbol{\theta}_{h,0}^*$. Matrices $\boldsymbol{F}_{it}$, $\boldsymbol{H}_t$ and $\boldsymbol{W}_t$ are assumed known and can be used to represent many patterns such as trends, periodicity, etc. The resulting model for $\mathcal{G} = \{G_t : 1 \le t \le T\}$ is thus a DDP, where the components of the atoms controlling the distribution means evolve in time in an autoregressive fashion.

Di Lucca et al. (2013) considered a model for a sequence of random variables $\{y_t : t \ge 1\}$ featuring a general autoregressive formulation by means of $y_t \mid (y_{t-1}, \ldots, y_{t-p}) = \boldsymbol{y} \sim G_{\boldsymbol{y}}$ and the problem of defining a prior for $\mathcal{G} = \{G_{\boldsymbol{y}} : \boldsymbol{y} \in \mathscr{Y}\}$. They discussed a general prior DDP model of the form $G_{\boldsymbol{y}}(\cdot) = \sum_{h=1}^{\infty} w_h(\boldsymbol{y}) \delta_{\boldsymbol{y}}(\cdot)$. Lau and So (2008) considered similar types of model, where each atom can be expressed as an infinite mixture of autoregressions of order p. Di Lucca et al. (2013) focused on the particular single-weights case and an order $p = 1$ process where the atom processes are expressed as simple linear autoregression: $\theta_h(\boldsymbol{y}) = \beta_h + \alpha_h y$. The full model in this case can be expressed as

$$\begin{aligned} y_t \mid y_{t-1} = y, \alpha_t, \beta_t, \sigma^2 &\sim N(\beta_t + \alpha_t y, \sigma^2), \\ (\beta_t, \alpha_t) \mid G &\overset{\text{iid}}{\sim} G, \\ G &\sim \mathrm{DP}(M, G_0). \end{aligned} \tag{7}$$

However, they also considered the case when atoms are defined as $\theta_h(y) = b + a_h y + \mathrm{OU}(\rho, \tau^2)$, where $\mathrm{OU}(\rho, \tau^2)$ denotes the Ornstein–Uhlenbeck process, a particular Gaussian process with covariance function of the form $\mathrm{Cov}[\theta(s), \theta(t)] = \tau^2 \rho^{|s-t|}$. Di Lucca et al. (2013) extended this approach for sequences of binary outcomes defined in terms of an autoregressive process Z_t with a flexible DDP prior distribution, where dependence is on the previous p binary responses.

An interesting variation of a dynamic DDP construction is proposed by Ascolani, Lijoi and Ruggiero (2021) who define a family $\mathcal{G} = \{G_t, t \ge 0\}$ of dependent random probability measures indexed by time. Their construction is motivated by a Fleming–Viot process. The random probability measures G_t share some, but not all atoms. The set D_t of atoms in the original G_0 which are shared in G_t is defined as a pure death process over time. Importantly, each G_t marginally remains a DP random measure. They refer to the model as the Fleming–Viot-DDP. In Prünster and Ruggiero (2013) this construction is applied to model market shares over time. Mena and Ruggiero (2016) construct another common-atoms DDP over time by setting up a Wrights–Fisher diffusion on the fractions $v_{t,\ell}$ in the stick-breaking construction of the marginal DP prior for G_t.

2.3 The Single-Atoms DDP

A parallel construction to the common weights DDP in the previous section considers a set of common atoms across all values of $\boldsymbol{x}$. This is the so called "single-atoms" DDP model, for which (3) takes the form

$$G_{\boldsymbol{x}}(\bullet) = \sum_{h=1}^{\infty} \underbrace{\Big\{ V_h(\boldsymbol{x}) \prod_{\ell<h} [1 - V_\ell(\boldsymbol{x})] \Big\}}_{w_h(\boldsymbol{x})} \delta_{\boldsymbol{\theta}_h}(\bullet), \tag{8}$$

where $V_h(\boldsymbol{x})$, $h \in \mathbb{N}$, are $[0,1]$-valued independent stochastic processes with index set $\mathscr{X}$ and marginal distributions $\mathrm{Be}(1, M_{\boldsymbol{x}})$. The locations $\boldsymbol{\theta}_h$, $h \in \mathbb{N}$, are independent with marginal distributions G^0, and the $\{V_h(\boldsymbol{x})\}$ and $\{\boldsymbol{\theta}_h\}$ collections are mutually independent.

Under the single-atoms model, all the covariate-dependence is expressed through the weights of the stick-breaking representation. One advantage of doing so is that, unlike the single-weights case, the implied prior probability model on partitions changes with the values of $x \in \mathscr{X}$. This is important when the implied partition is of interest. Another important feature is that problems related to extrapolation of $\boldsymbol{\theta}_h(\boldsymbol{x})$ are avoided, which could otherwise arise for inference for a new value of $\boldsymbol{x}$ beyond the range of the observed data. This is the case because under the single-atoms DDP all atoms are linked with observed data, in contrast to the single-weights DDP which includes atoms for new covariate values that are not linked with any observed data.

Duan, Guindani and Gelfand (2007) describe a model motivated by the analysis of spatially varying responses. Let $\{y(s) : s \in D\}$ be a stochastic process indexed by locations in a set $D \subset \mathbb{R}^d$, and let $s_1, \ldots, s_n$ the locations at which observations are collected. Their general construction involves a RPM G over the space of surfaces of D having finite-dimensionals adopting the following form: for any $s_1, \ldots, s_n \in D$ and $A_1, \ldots, A_n$ Borel-measurable sets in $\mathbb{R}$,

$$\begin{aligned} &P\big(y(s_1) \in A_1, \ldots, y(s_n) \in A_n\big) \\ &\quad = \sum_{i_1=1}^{\infty} \cdots \sum_{i_n=1}^{\infty} p_{i(s_1),\ldots,i(s_n)} \delta_{\theta_{i(s_1)}}(s_1) \cdots \delta_{\theta_{i(s_n)}}(s_n), \end{aligned}$$

where the θ_j's are i.i.d. from G_0 and the weights $\{p_{i(s_1),\ldots,i(s_n)}\}$ determine the site-specific joint selection probabilities. Conditions can be given so that the above specification follows a DP at any given location.

Always in the spatial context, specifically of modeling for hurricane surface wind fields, Reich and Fuentes (2007) propose a general framework that includes the single-atoms DDP as a special case. Their model is specially designed for spatial dependence as well, so that the covariates are geographical coordinates. Letting $\boldsymbol{s}$ denote such coordinates, their construction involves weights computed as $w_1(\boldsymbol{s}) = V_1(\boldsymbol{s})$ and $w_h(\boldsymbol{s}) = V_h(\boldsymbol{s}) \prod_{\ell=1}^{h-1}(1 - V_\ell(\boldsymbol{s}))$ for $h > 1$, where $V_h(\boldsymbol{s}) = \omega_h(\boldsymbol{s})V_h$, and $V_h \overset{\text{iid}}{\sim} \text{Beta}(a, b)$. The function $\omega_h(\boldsymbol{s})$ is centered at knot $\boldsymbol{\psi}_h = (\psi_{h1}, \psi_{h2})$, and the spread is controlled by parameters $\boldsymbol{e}_h = (e_{h1}, e_{h2})$. Reich and Fuentes (2007) discuss several possible choices for the ω_h functions and related parameters.

Griffin and Steel (2006) define another interesting variation of the basic DDP by keeping both sets of parameters, locations and the fractions (V_h), unchanged across $\boldsymbol{x}$. They use instead permutations of how the weights are matched with locations. The permutations change with $\boldsymbol{x}$. One advantage of such models is the fact that the support of $G_{\boldsymbol{x}}$ remains constant over $\boldsymbol{x}$, a feature that can be important for extrapolation beyond the observed data. A modification of this idea was explored by Griffin and Steel (2010) to generate what they called the DP regression smoother. The construction is centered over a class of regression models, and dependence is on the weights. More recently, similar ideas are used by Griffin and Steel (2011) to construct a family of prior distributions for a sequence of time dependent general RPMs that include the DDP setting as a special case. Another simple sequence of time-dependent DDPs was proposed by Gutiérrez, Mena and Ruggiero (2016), with a Markov chain structure for the sequence of time-varying sticks, and with application to the analysis of air quality data.

3. VARIATIONS OF MACEACHERN'S DDP

In this section, we discuss a variety of models extending the original definition (3). Many of these extensions are based on constructing independent weights and atoms processes indexed by covariates, but that do not necessarily produce a DP-distributed random measure. From an intuitive viewpoint, these classes of models can be seen as taking the basic DP construction and altering some of their basic components in terms of predictors $\boldsymbol{x} \in \mathscr{X}$ to a form that may differ from the initial distributional properties. While this typically modifies the marginal DP property, the extra flexibility allows one to tailor the properties of the model to fit specific applications.

3.1 Weighted Mixture of DPs (WMDP)

Dunson, Pillai and Park (2007) proposed a data-based prior using the observed predictors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. For every $\boldsymbol{x} \in \mathscr{X} \subset \mathbb{R}^p$, they considered the following construction

$$G_{\boldsymbol{x}}(\bullet) = \sum_{j=1}^{n} \left(\frac{\gamma_j K(\boldsymbol{x}, \boldsymbol{x}_j)}{\sum_{\ell=1}^{n} \gamma_\ell K(\boldsymbol{x}, \boldsymbol{x}_\ell)} \right) G_j(\bullet),$$

with

$$\gamma_j \mid \kappa \overset{\text{iid}}{\sim} \Gamma(\kappa, n\kappa), \qquad G_j \mid M, G_0 \overset{\text{iid}}{\sim} \text{DP}(M, G_0),$$

where $K : \mathscr{X} \times \mathscr{X} \longrightarrow \mathbb{R}^+$ is a bounded kernel function. The choice of K impacts the degree of borrowing of information from the neighbors in estimating the distribution at any particular predictor value $\boldsymbol{x}$. Some choices are discussed in the original technical report. In the paper, they considered

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\{\psi \|\boldsymbol{x} - \boldsymbol{x}'\|^2\}, \quad \psi \mid \mu_\psi, \sigma_\psi^2 \sim \text{LN}(\mu_\psi, \sigma_\psi^2), \tag{9}$$

where $\text{LN}(a, b)$ denotes the log-normal distribution with parameters $a \in \mathscr{R}$ and $b > 0$. With this choice, the resulting model for a given $\boldsymbol{x}$ borrows more heavily from those G_j's for which the corresponding $\boldsymbol{x}_j$ is close to $\boldsymbol{x}$. One primary application of this particular construction is in the context of *density regression* that is, in measuring how a probability distribution on the space of responses $\mathscr{Y}$ changes according to predictors $\boldsymbol{x} \in \mathscr{X}$.

3.2 Kernel Stick-Breaking

The kernel stick-breaking process (KSBP) was introduced by Dunson and Park (2008). For all $\boldsymbol{x} \in \mathscr{X} \subset \mathbb{R}^p$, the KSBP is defined as follows:

$$G_{\boldsymbol{x}}(\bullet) = \sum_{h=1}^{\infty} \left\{ W(\boldsymbol{x}; \boldsymbol{\Gamma}_h, V_h) \prod_{\ell<h} (1 - W(\boldsymbol{x}; \Gamma_\ell, V_\ell)) \right\} \times G_h(\bullet), \tag{10}$$

where $W(\boldsymbol{x}; \Gamma_h, V_h) = V_h K(\boldsymbol{x}, \boldsymbol{\Gamma}_h)$, with $K : \mathscr{X} \times \mathscr{X} \longrightarrow [0, 1]$, for example, as given in (9), $V_h \mid a_h, b_h \overset{\text{ind}}{\sim} \text{Be}(a_h, b_h)$, $\boldsymbol{\Gamma}_h \mid H \overset{\text{iid}}{\sim} H$ (random kernel locations), and $G_h \mid \mathcal{G} \overset{\text{iid}}{\sim} \mathcal{G}$ (random probability measures). The KSBP thus begins with an infinite sequence of basis random distributions $\{G_h\}$ and then constructs covariate-dependent random measures by mixing according to distance from the random locations Γ_h, with stick-breaking probabilities that are defined as a kernel multiplied by Beta-distributed weights. It is also possible to simplify the definition of KSBP, adopting the particular form

$$G_{\boldsymbol{x}}(\bullet) = \sum_{h=1}^{\infty} \left\{ W(\boldsymbol{x}; \boldsymbol{\Gamma}_h, V_h) \prod_{\ell<h} ((1 - W(\boldsymbol{x}; \Gamma_\ell, V_\ell)) \right\} \times \delta_{\boldsymbol{\theta}_h}(\bullet),$$

where $W(\boldsymbol{x}; \Gamma_h, V_h) = V_h K(\boldsymbol{x}, \boldsymbol{\Gamma}_h)$, with $K : \mathscr{X} \times \mathscr{X} \longrightarrow [0, 1]$, $V_h \mid M \overset{\text{iid}}{\sim} \text{Be}(1, M)$, $\boldsymbol{\Gamma}_h \mid H \overset{\text{iid}}{\sim} H$ (random kernel locations), and $\boldsymbol{\theta}_h \mid G_0 \overset{\text{iid}}{\sim} G_0$. This amounts

to replacing the random measure $G_h(\bullet)$ defined in (10) by just a single atom $\boldsymbol{\theta}_h$. Compared to the former, this latter version of KSBP greatly reduces model complexity while still retaining some flexibility.

3.3 Probit and Logit Stick-Breaking

Chung and Dunson (2009) introduced a modification of the stick-breaking representation for DPs where the Beta random variables are replaced by normally distributed random variables transformed using the standard normal CDF. They refer to the resulting measure as the probit-stick breaking (PSB) process. The PSB is defined by

$$G(\bullet) = \sum_{h=1}^{\infty} \left\{ \Phi(\eta_h) \prod_{\ell<h} (1 - \Phi(\eta_\ell)) \right\} \delta_{\boldsymbol{\theta}_h}(\bullet), \tag{11}$$

where $\eta_h \mid \mu \overset{\text{iid}}{\sim} N(\mu, 1)$ and $\boldsymbol{\theta}_h \mid G_0 \overset{\text{iid}}{\sim} G_0$. If $\mu = 0$, (11) reduces to a regular DP with $M = 1$, that is, uniformly distributed sticks. Chung and Dunson (2009) also consider a covariate-dependent version of the PSB to model sets of related probability distributions. This is done by replacing the η_h variables with suitable stochastic processes or regression functions. For instance, if $\{\eta_h(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{X}\}$ denote independent Gaussian processes with unit variance, a dependent PSB can be defined as

$$G_{\boldsymbol{x}}(\bullet) = \sum_{h=1}^{\infty} \left\{ \Phi(\eta_h(\boldsymbol{x})) \prod_{\ell<h} [1 - \Phi(\eta_\ell(\boldsymbol{x}))] \right\} \delta_{\boldsymbol{\theta}_h}(\bullet). \tag{12}$$

A similar modification can be obtained by taking $\eta_h(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\gamma}_h$. More generally, let $\eta_h(\boldsymbol{x}) = \alpha_h + f_h(\boldsymbol{x})$ with $\alpha_h \sim N(\mu, 1)$ and $f_h : \mathbb{R}^p \to \mathbb{R}$ an unknown regression function, characterized by finitely many parameters $\boldsymbol{\phi}_h$, with $\boldsymbol{\phi}_h \sim \boldsymbol{H}$. Denote this model as PSBP$(\mu, \boldsymbol{H}, G_0)$. One main focus of the proposal in Chung and Dunson (2009) was variable selection. To that end, they assume the model

$$y \mid \boldsymbol{x} \sim f(y \mid \boldsymbol{x}) = \int N(y \mid \boldsymbol{x}'\boldsymbol{\beta}, \tau^{-1}) \, dP_{\mathscr{X}}(\boldsymbol{\beta}, \tau),$$

$$P_{\mathscr{X}} = \{P_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\} \sim \text{PSBP}(\mu, \boldsymbol{H}, G_0),$$

where the variable selection structure is here introduced in $\boldsymbol{H}$ and in G_0, and by considering inclusion/exclusion indicators at the level of the atoms in (12). See further discussion on PSBP in Rodríguez and Dunson (2011). A related construction, termed the logit-stick breaking process was proposed in Ren et al. (2011), which essentially replaces the probit by a logit link in (11). Applications of logit-stick breaking processes to density regression can be found in Rigon and Durante (2021).

3.4 Hierarchical Mixture of DP

Consider again the case $\mathscr{X} = \{1, \ldots, J\}$, as in the example presented in Section 1, and let $\mathcal{G} = \{G_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\} = \{G_1, \ldots, G_J\}$. Motivated by the need to borrow strength across related studies (a situation also arising in applications of meta-analysis), Müller, Quintana and Rosner (2004) proposed a hierarchical DP model. In this construction, the probability distribution for group j is a weighted mixture of independent random measures. Specifically, the probability model for a group is defined as a mixture of a common distribution H_0, shared by all groups, and an idiosyncratic component H_j, which is specific to each group,

$$G_j(\bullet) = \epsilon H_0(\bullet) + (1 - \epsilon) H_j(\bullet), \tag{13}$$

where $\epsilon \in [0, 1]$ controls the level of dependence in the set $\mathcal{G}$, and $H_0, H_1, \ldots, H_J$ are assumed to be independent DPs. The two extreme cases depicted in Figure 1 correspond to $\epsilon = 1$ for panel (a), that is, a single common measure, and $\epsilon = 0$ for panel (b), that is, independent model and no borrowing of strength. Model (13) represents then a tradeoff between these two extreme options, allowing one to borrow strength through the common part, while retaining flexibility for the study-specific part of the model. More recently, Wang and Rosner (2019) used this construction to propose a propensity score-based mixture model to combine subject-level information from randomized and registry studies, their goal being inference on a causal treatment effect.

Extending (13) to the case of continuous predictors can be easily accomplished by combining a study index, j, continuous predictors $\boldsymbol{z}$, and setting up

$$G_{j,\boldsymbol{z}}(\bullet) = \epsilon H_{0,\boldsymbol{z}}(\bullet) + (1 - \epsilon) H_{j,\boldsymbol{z}}(\bullet),$$

where $H_{0,\boldsymbol{z}}, H_{1,\boldsymbol{z}}, \ldots, H_{J,\boldsymbol{z}}$ are now independent MacEachern's DDPs based on the continuous predictors $\boldsymbol{z}$, incorporating dependence on predictors as in the LDDP or ANCOVA-DDP of Section 2.2.1, according to the available covariates types. The construction is easily modified to allow for study-specific variation in the weight assigned to the idiosyncratic component H_j by replacing ϵ with ϵ_j.

A clever variation of this construction is introduced in Kolossiatis, Griffin and Steel (2013) who chose the weight ϵ to ensure that G_j remains marginally a DP again. A more general version of the same construction appears in Camerlenghi et al. (2019a).

3.5 Hierarchical DP of Teh et al. (2006)

In the context of $\mathscr{X} = \{1, \ldots, J\}$, Teh et al. (2006) proposed a model that induces an ANOVA type of dependence. In their construction, referred to as the hierarchical DP (HDP), the random probability measure for the jth group G_j, $j = 1, \ldots, J$, is a DP conditional on a common measure G, which in turn is also a DP,

$$\begin{aligned} G_j \mid M_j, G &\overset{\text{ind}}{\sim} \text{DP}(M_j, G), \quad j = 1, \ldots, J, \\ G \mid M, G_0 &\sim \text{DP}(M, G_0). \end{aligned} \tag{14}$$

A main motivation behind the particular form adopted in (14) was to provide a model that allows for sharing clusters among related subpopulations. Teh et al. (2006) consider the analysis of text, where a primary goal was to share clusters among various documents within a cluster, and also to share clusters among various corpora. The HDP facilitates the construction of clusters at various levels, due to its hierarchical formulation. In fact, this clustering structure can be described in terms of a Chinese restaurant franchise, where at each of a collection of restaurants customers sit at tables organized by dishes, and dishes can be ordered from a global menu available to all restaurants. This construction, if restricted to a single restaurant, reduces to the usual Chinese restaurant process (Aldous, 1985) that is colloquially used to describe the DP.

3.6 The Nested DP

Also in the context of $\mathscr{X} = \{1, \ldots, J\}$, Rodríguez, Dunson and Gelfand (2008) proposed an alternative model, referred to as the nested DP. In their construction the law of the random probability measure for the jth group G_j, $j = 1, \ldots, J$, is an infinite mixture of trajectories of DPs,

$$\begin{aligned} G_j &\overset{\text{ind}}{\sim} \sum_{h=1}^{\infty} \pi_h \delta_{G_h^*}(\bullet), \quad j = 1, \ldots, J, \\ G_h^* \mid M_2, H &\overset{\text{iid}}{\sim} \text{DP}(M_2, H), \end{aligned} \tag{15}$$

where $\pi_h = V_h \prod_{\ell<h}(1 - V_\ell)$, with $V_h \mid M_1 \overset{\text{iid}}{\sim} \text{Be}(1, M_1)$, for $h = 1, 2, \ldots$. The main motivation behind (15) was to construct a clustering of individuals across the different groups, for example, patients within different medical centers. The NDP model aims to simultaneously cluster patients within centers, borrowing information across centers for which similar clusters are detected, and to cluster different centers. This is then a type of multilevel clustering.

By way of comparison, it can be noted that in the HDP of Teh et al. (2006), the random measures in $\mathcal{G} = \{G_1, \ldots, G_J\}$ share the same atoms but assign them different weights, while in the NDP two distributions G_{j_1} and G_{j_2} either share both atoms and weights (i.e., they are identical) or share nothing at all. Thus, the NDP allows for clusters at the level of the responses and also at the level of distributions, while the HDP allows for clusters only at the level of observations.

One of the limitations of the NDP is that for any two random measures G_{j_1}, G_{j_2} it supports only the two extreme cases of either all atoms and weights shared, that is, $G_{j_1} = G_{j_2}$, or no atoms shared, but does not allow any intermediate configuration with some atoms being shared. As a consequence, whenever there are ties of atoms between G_{j_1} and G_{j_2}, the nested structure forces the two random distributions to be identical. For a discussion of this problem, see Camerlenghi et al. (2019a) who introduce the latent nested process as a more general hierarchical prior for random probability measures that avoids this restriction. More recently, Beraha, Guglielmi and Quintana (2020) propose the semi-hierarchical DP as an alternative solution to the limitations inherent to latent nested processes, with the added benefit of computationally efficient implementations to the comparison and clustering of potentially many subpopulations.

Like any discrete random probability measure, the NDP can be used to define random partitions. Model (15) could be written in short as $G_j \sim \text{DP}\{M_1, \text{DP}(M_2, H)\}$. The outer DP, with total mass M_1 gives rise to a partition of $\mathscr{X}$. Consider now samples $y_{ji} \sim G_j$, $i = 1, \ldots, n_j$. The inner DP gives rise to random partitions of $\mathscr{Y}_j = \{1, \ldots, n_i\}$, that is, the NDP defines a nested partition of $\mathscr{X}$ and $\mathscr{Y}_j$, with the prior for the random partitions for $\mathscr{Y}_j$ and $\mathscr{Y}_{j'}$ being equal in distribution when $G_j = G_{j'}$. Curiously, exactly the same random nested partition on $\mathscr{X}$ and $\mathscr{Y}_j$ is implied by the enriched DP (EDP) defined in Wade, Mongelluzzo and Petrone (2011). The EDP defines a random probability measure for pairs (x_i, y_i) as $P_X(x_i)P_{Y|X}(y_i \mid x_i)$, which, as discrete random probability measures, gives rise to the same random nested partition.

3.7 The Product of Independent DPs

Alternatively, Gelfand and Kottas (2001) proposed an approach based on the product of independent random measures. In this construction the distribution for the jth group G_j, $j = 1, \ldots, J$, is given by

$$G_j(\bullet) \equiv H_j(\bullet) \prod_{\ell<j} H_\ell(\bullet), \quad j = 1, \ldots, J,$$

where

$$H_j \mid M_j, H_{0j} \overset{\text{ind}}{\sim} \text{DP}(M_j, H_{0j}), \quad j = 1, \ldots, J.$$

The motivation for this construction arises from the need to define models that induce stochastic ordering for the random group specific distributions G_j. The ordering holds with probability 1 in the prior and so is also satisfied a posteriori.

3.8 Other Constructions

Chung and Dunson (2009) proposed a similar construction, referred to as the local DP, where the stick-breaking weights selected to define the probability weights depend on a set of random locations and their distances to a given predicted value. In this construction, the support points also depend on predictors.

Fuentes-García, Mena and Walker (2009) considered a dependent variation of geometric-weights stick-breaking processes (Mena, Ruggiero and Walker, 2011). In this

construction, the stick-breaking weights are replaced by their expected value, thus reducing the number of parameters.

Petrone, Guindani and Gelfand (2009) proposed a *hybrid* variation of the Dirichlet process that can be also extended to more general discrete random probability measures. Their construction was motivated by functional data analysis. In their context, different curves are expressed as a mixture of a smaller set of canonical curves, where the level of borrowing strength (local clustering) can vary over different portions of the curves.

Another type of construction stems from the fact that the Dirichlet process is also a special case of a normalized random measure with independent increments (NRMI), as described in Regazzini, Lijoi and Prünster (2003). This means that if F has a DP distribution, then it can be expressed in the form

$$F(\bullet) = \frac{\mu(\bullet)}{\mu(\Omega)},$$

where Ω is the space where the DP is defined, and μ is a completely random measure on $(\Omega, \mathcal{B}(\Omega))$, that is, for any collection of disjoint sets $A_1, A_2, \ldots$ in $\mathcal{B}(\Omega)$, the Borel σ-field in Ω, the random variables $\mu(A_1), \mu(A_2), \ldots$ are independent, and $\mu(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} \mu(A_j)$ holds true a.s. See, for example, James, Lijoi and Prünster (2009). As shown in Ferguson (1973), the Dirichlet process arises as the normalized version of a Gamma process. Barrios et al. (2013), Favaro and Teh (2013) and Argiento, Guglielmi and Pievatolo (2010) discuss modeling with mixtures of NRMIs, and in particular discuss practical implementation of posterior simulation for such models. See additional MCMC implementation details in Favaro and Teh (2013). Building on related ideas, Epifani and Lijoi (2010) and Leisen and Lijoi (2011), proposed dependent neutral to the right processes and correlated two-parameter Poisson–Dirichlet processes, respectively, by considering suitable Lévy copulas. The general class of dependent and/or correlated normalized completely random measures has been discussed, for instance, by Griffin, Kolossiatis and Steel (2013) and by Lijoi, Nipoti and Prünster (2014). Griffin and Leisen (2017) discuss various aspects of Lévy copulas and their connections to several popular models. In particular, they use them to define compound random measures. DDPs defined by introducing dependence in NRMIs have also been explored in the literature. Lin, Grimson and Fisher (2010) used this idea to propose a Markov chain of Dirichlet processes, and other extensions to normalized random measured are described in Chen, Ding and Buntine (2012) and in Chen et al. (2013). Finally, Camerlenghi et al. (2019b) study properties of some general hierarchical processes obtained via normalization, including the HDPs discussed earlier in Section 3.5.

4. THE INDUCED CONDITIONAL DENSITY APPROACH

The approaches described so far yield valid inferences when the set of predictors $\boldsymbol{x}$ are fixed by design or are random but exogenous. Notice that the exogeneity assumption permits us to focus on the problem of conditional density estimation, regardless of the data generating mechanism of the predictors, that is, if they are randomly generated or fixed by design (see, e.g., Barndorff-Nielsen, 1973, 1978). Under the presence of endogenous predictors, both the response and the predictors should be modeled jointly.

In the context of continuous responses and predictors, Müller, Erkanli and West (1996) proposed a DPM of multivariate Gaussian distributions for the complete data $\boldsymbol{d}_i = (y_i, \boldsymbol{x}_i)'$, $i = 1, \ldots, n$, and looked at the induced conditional distributions. Although Müller, Erkanli and West (1996) focused on the mean function only, $m(\boldsymbol{x}) = E(y \mid \boldsymbol{x})$, their method can be easily extended to provide inferences for the conditional density at covariate level $\boldsymbol{x}$. The model is given by

$$\boldsymbol{d}_i \mid G \overset{\text{iid}}{\sim} \int N_k(\boldsymbol{d}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, dG(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

and

$$G \mid M, G_0 \sim \text{DP}(M, G_0),$$

where $k = p + 1$ is the dimension of the complete data vector $\boldsymbol{d}_i$, and the baseline distribution G_0 is the conjugate normal-inverted-Wishart (IW) distribution $G_0 \equiv N_k(\boldsymbol{\mu} \mid \boldsymbol{m}_1, \kappa_0^{-1}\boldsymbol{\Sigma}) \times \text{IW}_k(\boldsymbol{\Sigma} \mid \nu_1, \boldsymbol{\Psi}_1)$. The model is completed with conditionally conjugate priors and hyperpriors on m_1, κ_0 and $\boldsymbol{\Psi}$, and, if desired, a gamma hyperprior on M. The model induces a weight-dependent mixture model for the regression,

$$f_{\boldsymbol{x}}(y) = \sum_{h=1}^{\infty} \omega_h(\boldsymbol{x}) N\big(y \mid \beta_{0h} + \boldsymbol{x}'\boldsymbol{\beta}_h, \sigma_h^2\big), \tag{16}$$

where

$$\omega_h(\boldsymbol{x}) = \frac{w_h N_p(\boldsymbol{x} \mid \boldsymbol{\mu}_{2h}, \boldsymbol{\Sigma}_{22h})}{\sum_{\ell=1}^{\infty} w_\ell N_p(\boldsymbol{x} \mid \boldsymbol{\mu}_{2\ell}, \boldsymbol{\Sigma}_{22\ell})}, \quad h = 1, 2, \ldots,$$

$\beta_{0h} = \mu_{1h} - \boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\boldsymbol{\mu}_{2h}$, $\boldsymbol{\beta}_h = \boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}$, and $\sigma_h^2 = \sigma_{11h}^2 - \boldsymbol{\Sigma}_{12h}\boldsymbol{\Sigma}_{22h}^{-1}\boldsymbol{\Sigma}_{21h}$. Here, the weights w_h follow the usual DP stick-breaking construction, and the remaining elements arise from the standard partition of the vectors of means and (co)variance matrices given by

$$\boldsymbol{\mu}_h = \begin{pmatrix} \mu_{1h} \\ \boldsymbol{\mu}_{2h} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_h = \begin{pmatrix} \sigma_{11h}^2 & \boldsymbol{\Sigma}_{12h} \\ \boldsymbol{\Sigma}_{21h} & \boldsymbol{\Sigma}_{22h} \end{pmatrix},$$

respectively.

The induced conditional density approach of Müller, Erkanli and West (1996) can be easily extended to handle mixed continuous, $\boldsymbol{x}_C$, and discrete predictors, $\boldsymbol{x}_D$,

by considering a DPM model of product of appropriate kernels for discrete k_D and continuous k_D variables,

$$\boldsymbol{d}_i \mid G \overset{\text{iid}}{\sim} \int k_D(\boldsymbol{x}_{iD} \mid \boldsymbol{\theta}_1) k_C(y_i, \boldsymbol{x}_C \mid \boldsymbol{\theta}_2)\, dG(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2), \tag{17}$$

that is, assuming a multiplicative structure in the joint model for $(y, \boldsymbol{x}_D, \boldsymbol{x}_C)$ that mimics conditional independence of $(y, \boldsymbol{x}_C)$ and $\boldsymbol{x}_D$ given suitable parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Similar types of models, but looking only at the induced partition structures, are discussed in Müller and Quintana (2010). In particular, Müller, Quintana and Rosner (2011) proposed a version of (17) that may be viewed as integrating out the random measure G in (17), retaining only the random partition model, while still allowing for covariate dependence in the prior. This approach exploits the connection between the DP and product partition models. See, for example, Quintana and Iglesias (2003).

We introduced the conditional density regression approach assuming endogenous predictors, when the construction of a joint probability model for $(y_i, \boldsymbol{x}_i)$ is natural. However, the same construction can be used to achieve the desired smooth locally weighted mixture of linear regressions even when the $\boldsymbol{x}_i$ are exogenous, or even if they are not random at all. The choice of model depends largely on properties of the model and ease of prior specification, tempered by computational concerns.

5. IMPLIED RANDOM PARTITIONS AND OTHER USES OF THE DDP MODEL

One of the common applications of the DP mixture model (1) is to define a random partition and allow statistical inference on such partitions. Consider an equivalent statement of i.i.d. sampling from (1) as a hierarchical model

$$y_i \mid \theta_i \sim p(y_i \mid \theta_i) \quad \text{and} \quad \theta_i \sim G, \tag{18}$$

$i = 1, \ldots, n$. The discrete nature of the DP random measure G implies positive probabilities of ties among the θ_i with $K \leq N$ unique values $\{\theta_1^\star, \ldots, \theta_K^\star\}$. Defining $S_j = \{i : \theta_i = \theta_j^\star\}$ defines a partition $\{1, \ldots, n\} = \bigcup_j S_j$. A common application of the DP mixture model is to derive inference on such partitions $\rho = \{S_1, \ldots, S_K\}$, and interpret the partitioning subsets as meaningful subpopulations of the experimental units (e.g., patient subpopulations). In anticipation of the upcoming generalization to the DDP, we introduce a slightly different but equivalent definition of the clusters S_j. Recall the representation (2) of a DP random measure, $G = \sum w_h \delta_{\widetilde{\theta}_h}$. Then the non-empty sets $R_h = \{i : \theta_i = \widetilde{\theta}_h\}$ describe the same partition ρ. We switched from indexing clusters by their common unique θ_i values to identifying clusters by the matching atoms in G. Similarly we can set up a model for independent sampling using a DDP prior. Specifically, consider

$$y_i \mid \theta_i \sim p(y_i \mid \theta_i) \quad \text{and} \quad \theta_i \mid x_i = x \sim G_x, \tag{19}$$

$i = 1, \ldots, n$, with a DDP prior on $\mathcal{G} = \{G_x, x \in X\}$. For the moment assume a categorical covariate $x_i \in \{1, \ldots, n_x\}$, and let G_x, $x = 1, \ldots, n_x$ denote the (marginal) random measures, and let $I_x = \{i : x_i = x\}$ denote the subpopulation with covariate x. First, by the earlier argument the model implies a random partition ρ_x of I_x, marginally, for each x. Indexing clusters by the corresponding atom in G_x implicitly defines a joint prior on $\{\rho_x, x \in X\}$, or, alternatively, defines a partition of $\{1, \ldots, n\}$ with clusters S_j that cut across I_x. In particular, the model implies a joint prior on $(\rho_x, \rho_{x'})$ for any $x \neq x'$, and it allows for shared clusters across subpopulations. Different assumptions on various model aspects, such as dispersion in the baseline distribution, or total mass parameter, would have a practical effect on the this joint prior. Curiously, in contrast to the DP mixture model, the DDP model is not commonly used for inference on these implied random partition(s).

Another feature of the DDP model is inference about distributional homogeneity. To be specific, consider again the context of independent sampling in (18) with a categorical covariate $x \in \{1, \ldots, n_x\}$ and let $f_x(y) = \int p(y \mid \theta)\, dG_x(\theta)$ denote the implied marginal distribution of $y_i \mid x_i = x$. In many applications investigators might be interested in the event $f_x = f_{x'}$ for $x \neq x'$. While the DDP prior, short of a pathological special case, implies zero prior probability for exact equality, posterior inference includes meaningful posterior probabilities for $\{d(f_x, f_{x'}) > \epsilon\}$ for any well defined distance of the two distributions. Specifics would depend on particular applications. Related summaries, for example, by displaying posterior means for f_x over x are shown in some papers using DDP priors for density regression. See, for example, Gutiérrez et al. (2019).

6. APPLICATION TO AUTOREGRESSIVE MODELS

We illustrate some of the nonparametric regression models based on DDP models. We implement inference under the ANOVA-DDP or LDDP model (5) and conditional density regression as in (16) to model (auto-)regression on $x_t = y_{t-1}$ in time series data. We specifically employ the LDDP model

$$\begin{aligned} y_t \mid y_{t-1} = y, \beta_{t0}, \beta_{t1}, \sigma_t^2 &\sim N\big(\beta_{t0} + \beta_{t1} y, \sigma_t^2\big), \\ (\beta_{t0}, \beta_{t1}, \sigma_t^2) \mid G &\overset{\text{iid}}{\sim} G, \\ G &\sim \mathrm{DP}\big(M, G_0(\cdot \mid \boldsymbol{\eta})\big), \end{aligned} \tag{20}$$

where $t = 2, \ldots, n$, that is, we mix over the linear coefficients and the variance. The dependence in (20) is conveyed through linear functions of the first lagged response

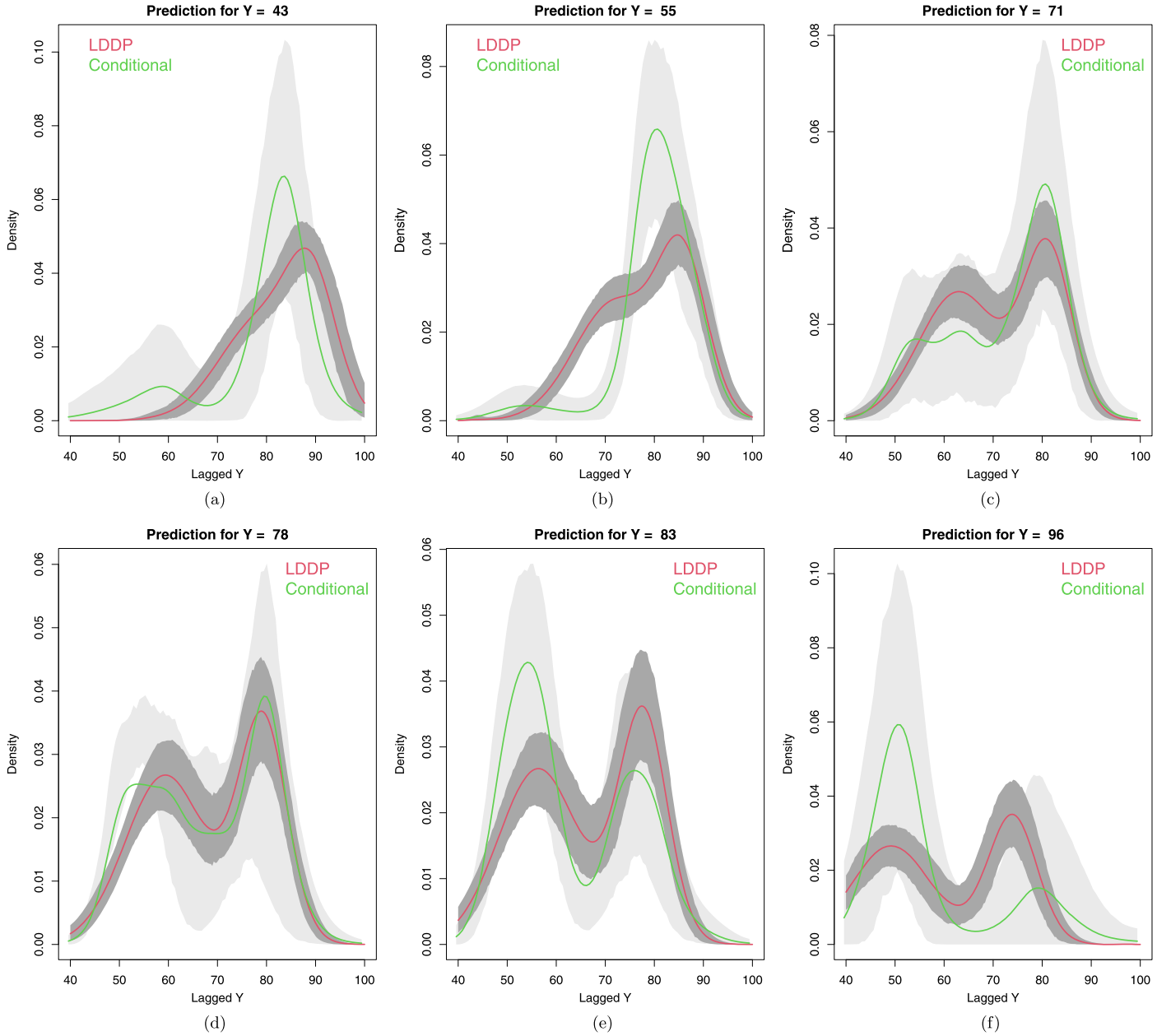


FIG. 3. *Old Faithful Geyser data*: *Posterior estimated G_x, that is, posterior predictive densities (mean and point-wise 95% HPD intervals) for the waiting times at lagged times (a) $y_{t-1} = 43$, (b) $y_{t-1} = 55$, (c) $y_{t-1} = 71$, (d) $y_{t-1} = 78$, (e) $y_{t-1} = 83$ and (f) $y_{t-1} = 96$, including 95% HPD credibility bands. The red curve shows inference under the LDDP model. The green curve shows inference under the conditional density approach.*

in the atoms, keeping common weights. Here, $G_0(\cdot \mid \boldsymbol{\eta})$ is the centering measure with hyperparameters $\boldsymbol{\eta}$. Following Jara et al. (2011), we use $G_0 \equiv N_2(\boldsymbol{\beta} \mid \boldsymbol{\mu}_b, \boldsymbol{S}_b)\Gamma(\sigma^{-2} \mid \tau_1/2, \tau_2/2)$, and complete the prior specification as

$$M \mid a_0, b_0 \sim \Gamma(a_0, b_0),$$
$$\tau_2 \mid \tau_{s_1}, \tau_{s_2} \sim \Gamma(\tau_{s_1}/2, \tau_{s_2}/2),$$
$$\boldsymbol{\mu}_b \mid \boldsymbol{m}_0, \boldsymbol{S}_0 \sim N_p(\boldsymbol{m}_0, \boldsymbol{S}_0),$$
$$\boldsymbol{S}_b \mid \nu, \boldsymbol{\Psi} \sim \mathrm{IW}_p(\nu, \boldsymbol{\Psi}).$$

For this illustration, we consider the Old Faithful geyser data (Härdle, 1991), available as part of the `datasets` library available in `R`, which includes $n = 272$ observations on eruption times (in minutes) and times between eruptions. In the following, we compare inference results under model (20) with inference under density regression, as in (16), again using $x_t = y_{t-1}$, and taking the waiting times between eruptions as the variable y_t of interest. Recall that a conditional density approach is based on a DPM model for $\{(y_t, y_{t-1}) : t = 2, \ldots, n\}$.

In all cases, we used hyperparameters as in Jara et al. (2011). Results for the analysis are shown in Figure 3. In particular, we show a comparison of posterior inference for G_x for (a) $y_{t-1} = 43$, (b) $y_{t-1} = 55$, (c) $y_{t-1} = 71$,

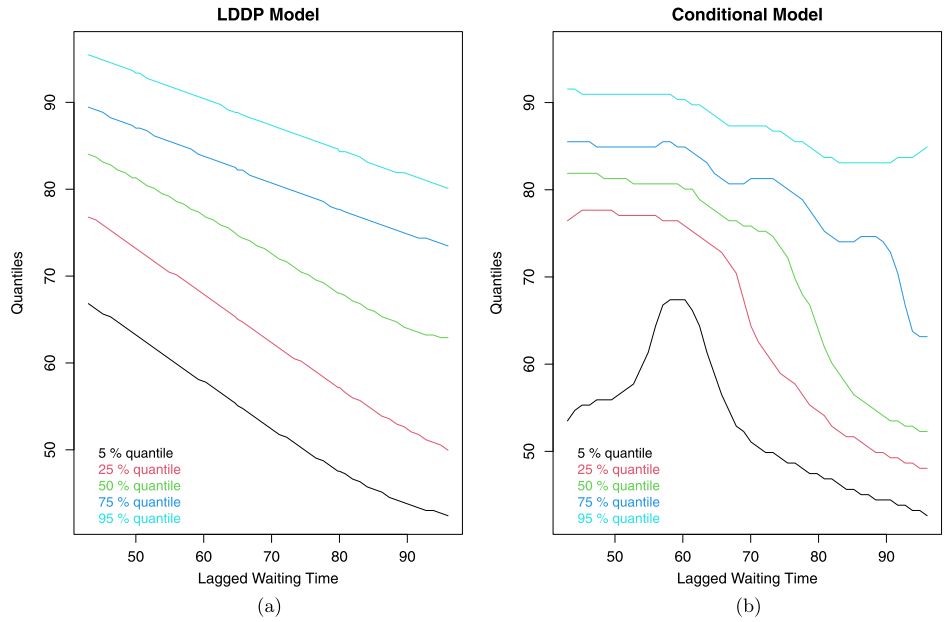


FIG. 4. *Old Faithful Geyser data*: *Estimated quantile curves for the posterior predictive densities under* (*a*) *LDDP model*, *and* (*b*) *Conditional model*. *Each curve corresponds to the indicated quantiles of* G_x *as a function of lagged* $x_t = y_{t-1}$ *values*, *corresponding to* 5% (*black*), 25% (*red*), 50% (*green*), 75% (*blue*) *and* 95% (*light blue*) *quantiles*.

(d) $y_{t-1} = 78$, (e) $y_{t-1} = 83$ and (f) $y_{t-1} = 96$, including 95% HPD credibility bands. The points at which predictions were made correspond to the empirical quintiles of the observed waiting times plus the endpoints of the empirical range. While there are some model-specific differences in the estimated distributions G_x, they both agree on the bimodal nature. Results from both models suggest that the bimodal feature changes as the lagged value for prediction increases, moving from right to left skewness. This finding is more markedly seen in the conditional approach than in the LDDP case. Another point worth noting is that both models produce similar estimates of G_x in a range roughly covering the central quintiles, but as the lagged values veer off from the center of the empirical range the differences among the corresponding estimates become more marked.

To further illustrate how the posterior predictive densities change with lagged waiting times, we show how quantiles change as a function of y_{t-1} values. Figure 4 shows the resulting curves, estimated over a suitable grid of lagged waiting times, for the three quartiles plus the 5% and 95% quantiles. The LDDP model, shown in panel (a), gives rise to patterns that exhibit strong linearly decreasing trends, with a slight increase in the corresponding slopes for higher quantiles. The conditional model is shown in panel (b), and even though there is still a global decreasing trend, the curves are not nearly linear. In fact, they are not even monotonic, suggesting some marked nonlinear features of the data that are not so evidently found under the LDDP model.

7. CONCLUDING REMARKS

DDPs have come a long way since they were originally proposed. By its very definition, a DDP has the potential to incorporate covariate indexing (dependence) either in the atoms or the weights or both. The results in Barrientos, Jara and Quintana (2012) show that under full support of the stochastic processes that are used to convey covariate dependence, the resulting DDP has full support in the space $\mathscr{F} = \{F_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\}$. This holds true for all of the basic DDP constructions: single-atoms, single-weights, and with dependence in both. A natural question is then: which DDP version is the best? There is no final answer to this question, although DDPs with dependence in both atoms and weights are less commonly found, mostly due to the computational complexity related to their implementation. An exception to this is the conditional approach described in Section 4. In broad terms, the single-weights models are typically easier to fit, as the standard algorithms designed to implement posterior simulation in the context of DPs can be applied with relatively minor adjustments. See, for example, the computational aspects in De Iorio et al. (2004). The same applies for the LDDP. On the other hand, the single-atoms models are typically less attractive from a computational viewpoint, mainly due to how covariate dependence is encoded in the definition of the weight processes $\{w_h(\boldsymbol{x}) : \boldsymbol{x} \in \mathscr{X}\}$. However, the single-atoms DDP allows for the prior probability distribution on the partitions to change with $\boldsymbol{x}$, a feature that is not supported by the single-weights DDP. For a formal description of this feature, let $\mathscr{G} = \{G_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\}$ denote the family of random probability measures with DDP

prior, as before. Let $\rho_{\boldsymbol{x}}$ denote the partition of $\{1, \ldots, n\}$ that is implied by a hypothetical sample from $G_{\boldsymbol{x}}$, of size n. Under the single-weights DDP, $p(\rho_{\boldsymbol{x}} \mid \mathcal{G})$ is invariant across $\boldsymbol{x}$; but not so under the single-atoms DDP. This is the case since the prior on the random partition $\rho_{\boldsymbol{x}}$ is determined by the weights in $G_{\boldsymbol{x}}$.

Models for dependent probability distributions do not easily allow for the incorporation of existing prior information about arbitrary functionals. A modeler is unlikely to have prior knowledge about all aspects of a collection of probability measures, but could have real historical prior information about specific functionals (such as the mean or quantile functions). For example, such information could be obtained as the product of applying parametric or (classical) nonparametric approaches to previous data. Furthermore, even in models for single (non-dependent) probability measures, the derivation of the induced distribution for arbitrary functionals is challenging and, thus, usually not exploited. This makes the prior elicitation process difficult. We refer the reader to Lijoi and Prünster (2009) for an exhaustive summary of existing results concerning distributional properties of functional of single and discrete random probability measures.

In the context of a single probability measure, Kessler, Hoff and Dunson (2015) proposed a clever construction of a BNP model with a given distribution on a finite set of functionals. Their approach is based on the conditional distribution of a standard BNP prior, given the functionals of interest. A Metropolis–Hastings MCMC algorithm is proposed to explore the posterior distribution under the marginally specified BNP model, where the standard BNP model is used as a candidate generating model, and that is closely related to the well-known importance-sampling approach for assessing prior sensitivity. Their MCMC algorithm is developed for DP-based models and relies on the marginalization of the random probability measure. Thus, a Monte Carlo approximation of the functionals of interest is employed at any step of the MCMC algorithm to obtain approximated posterior samples of the functionals of interest. The study of extensions of the approach proposed by Kessler, Hoff and Dunson (2015) to the context of sets of predictor-dependent probability measures is a topic of interest for future research.

An interesting topic has been recently brought up by Campbell et al. (2019). They introduced a relaxed version of the notion of exchangeability, *local exchangeability*, which considers bounded changes in total variation norm of the distribution of observations under permutations of data having nearby covariate values. This notion generalizes that of exchangeability and partial exchangeability. The work by Campbell et al. (2019) discusses conditions under which a version of de Finetti's theorem holds in such a way that a DDP is the corresponding de Finetti measure, that is, conditional independence of the observations under a DDP is still true. The study of extensions and applications of these and related results is another topic of interest for future research.

The bulk of work on the DDP and related methods focuses on the family of conditional distributions $\mathcal{G} = \{G_{\boldsymbol{x}} : \boldsymbol{x} \in \mathscr{X}\}$ and models where an observation y is associated with a single value of the covariate $\boldsymbol{x}$. When data are longitudinal, spatial or functional, the observations may be considered to have dependence that cannot be captured by the marginal distributions $G_{\boldsymbol{x}}$. See, for example, Xu, MacEachern and Xu (2015) who separate dependence in financial data series from the marginal distributions. Many open questions remain in this direction.

Finally, the idea of introducing dependence through normalization, for example, as mentioned earlier in Section 3.8 can be further exploited and extended to more general cases, including going beyond the context of DDPs.

ACKNOWLEDGEMENTS

A. Jara's and F. Quintana's research is supported by ANID—Millennium Science Initiative Program—NCN17_059. A. Jara is also supported by Fondecyt grant 1180640, F. Quintana is also supported by Fondecyt grant 1180034. P. Müller acknowledges partial support from grant NSF Grant DMS-1952679 from the National Science Foundation, and under R01 CA132897 from the U.S. National Cancer Institute.

REFERENCES

ALDOUS, D. J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII*—1983. *Lecture Notes in Math.* **1117** 1–198. Springer, Berlin. MR0883646 https://doi.org/10.1007/BFb0099421

ARGIENTO, R., GUGLIELMI, A. and PIEVATOLO, A. (2010). Bayesian density estimation and model selection using nonparametric hierarchical mixtures. *Comput. Statist. Data Anal.* **54** 816–832. MR2580918 https://doi.org/10.1016/j.csda.2009.11.002

ASCOLANI, F., LIJOI, A. and RUGGIERO, M. (2021). Predictive inference with Fleming–Viot-driven dependent Dirichlet processes. *Bayesian Anal.* Advance publication. https://doi.org/10.1214/20-BA1206

BARNDORFF-NIELSEN, O. (1973). On M-ancillarity. *Biometrika* **60** 447–455. MR0345255 https://doi.org/10.1093/biomet/60.3.447

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, Chichester. MR0489333

BARRIENTOS, A. F., JARA, A. and QUINTANA, F. A. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Anal.* **7** 277–309. MR2934952 https://doi.org/10.1214/12-BA709

BARRIOS, E., LIJOI, A., NIETO-BARAJAS, L. E. and PRÜNSTER, I. (2013). Modeling with normalized random measure mixture models. *Statist. Sci.* **28** 313–334. MR3135535 https://doi.org/10.1214/13-STS416

BERAHA, M., GUGLIELMI, A. and QUINTANA, F. A. (2020). The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions. Preprint. Available at arXiv:2005.10287.

CAMERLENGHI, F., DUNSON, D. B., LIJOI, A., PRÜNSTER, I. and RODRÍGUEZ, A. (2019a). Latent nested nonparametric priors (with discussion). *Bayesian Anal*. **14** 1303–1356. MR4044854 https://doi.org/10.1214/19-BA1169

CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019b). Distribution theory for hierarchical processes. *Ann. Statist*. **47** 67–92. MR3909927 https://doi.org/10.1214/17-AOS1678

CAMPBELL, T., SYED, S., YANG, C.-Y., JORDAN, M. I. and BRODERICK, T. (2019). Local exchangeability. Preprint. Available at arXiv:1906.09507.

CARON, F., DAVY, M., DOUCET, A., DUFLOS, E. and VANHEEGHE, P. (2008). Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Trans. Signal Process*. **56** 71–84. MR2439814 https://doi.org/10.1109/TSP.2007.900167

CHEN, C., DING, N. and BUNTINE, W. (2012). Dependent hierarchical normalized random measures for dynamic topic modeling. In *Proceedings of the* 29*th International Conference on Machine Learning* (*ICML*-12) (J. Langford and J. Pineau, eds.). *ICML* '12 895–902. Omnipress, New York.

CHEN, C., RAO, V., BUNTINE, W. and TEH, Y. W. (2013). Dependent normalized random measures. In *Proceedings of the* 30*th International Conference on Machine Learning* (S. Dasgupta and D. McAllester, eds.). *Proceedings of Machine Learning Research* **28** 969–977. PMLR, Atlanta, GA.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat*. **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

CHUNG, Y. and DUNSON, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc*. **104** 1646–1660. MR2750582 https://doi.org/10.1198/jasa.2009.tm08302

CIFARELLI, D. and REGAZZINI, E. (1978). Problemi statistici non parametrici in condizioni di scambialbilita parziale e impiego di medie associative. Technical report. Quaderni Istituto Matematica Finanziaria, Torino.

DE IORIO, M., MÜLLER, P., ROSNER, G. L. and MACEACHERN, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc*. **99** 205–215. MR2054299 https://doi.org/10.1198/016214504000000205

DE IORIO, M., JOHNSON, W. O., MÜLLER, P. and ROSNER, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* **65** 762–771. MR2649849 https://doi.org/10.1111/j.1541-0420.2008.01166.x

DEVROYE, L. (1986). *Nonuniform Random Variate Generation*. Springer, New York. MR0836973 https://doi.org/10.1007/978-1-4613-8643-8

DE LA CRUZ-MESÍA, R., QUINTANA, F. A. and MÜLLER, P. (2007). Semiparametric Bayesian classification with longitudinal markers. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **56** 119–137. MR2359237 https://doi.org/10.1111/j.1467-9876.2007.00569.x

DI LUCCA, M. A., GUGLIELMI, A., MÜLLER, P. and QUINTANA, F. A. (2013). A simple class of Bayesian nonparametric autoregression models. *Bayesian Anal*. **8** 63–87. MR3036254 https://doi.org/10.1214/13-BA803

DUAN, J. A., GUINDANI, M. and GELFAND, A. E. (2007). Generalized spatial Dirichlet process models. *Biometrika* **94** 809–825. MR2416794 https://doi.org/10.1093/biomet/asm071

DUNSON, D. B. and HERRING, A. H. (2006). Semiparametric Bayesian latent trajectory models. Technical report. ISDS Discussion Paper 16, Duke Univ., Durham, NC.

DUNSON, D. B. and PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* **95** 307–323. MR2521586 https://doi.org/10.1093/biomet/asn012

DUNSON, D. B., PILLAI, N. and PARK, J.-H. (2007). Bayesian density regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **69** 163–183. MR2325270 https://doi.org/10.1111/j.1467-9868.2007.00582.x

EPIFANI, I. and LIJOI, A. (2010). Nonparametric priors for vectors of survival functions. *Statist. Sinica* **20** 1455–1484. MR2777332

FARAWAY, J. J. (2016). *Extending the Linear Model with R*: *Generalized Linear, Mixed Effects and Nonparametric Regression Models*. *Chapman & Hall/CRC Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3617626

FAVARO, S. and TEH, Y. W. (2013). MCMC for normalized random measure mixture models. *Statist. Sci*. **28** 335–359. MR3135536 https://doi.org/10.1214/13-STS422

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist*. **1** 209–230. MR0350949

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist*. **2** 615–629. MR0438568

FUENTES-GARCÍA, R., MENA, R. H. and WALKER, S. G. (2009). A nonparametric dependent process for Bayesian regression. *Statist. Probab. Lett*. **79** 1112–1119. MR2510777 https://doi.org/10.1016/j.spl.2009.01.005

GELFAND, A. E. and KOTTAS, A. (2001). Nonparametric Bayesian modeling for stochastic order. *Ann. Inst. Statist. Math*. **53** 865–876. MR1880817 https://doi.org/10.1023/A:1014629724913

GELFAND, A. E., KOTTAS, A. and MACEACHERN, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc*. **100** 1021–1035. MR2201028 https://doi.org/10.1198/016214504000002078

GIUDICI, P., MEZZETTI, M. and MULIERE, P. (2003). Mixtures of products of Dirichlet processes for variable selection in survival analysis. *J. Statist. Plann. Inference* **111** 101–115. MR1955875 https://doi.org/10.1016/S0378-3758(02)00291-4

GREEN, P. J. and RICHARDSON, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Stat*. **28** 355–375. MR1842255 https://doi.org/10.1111/1467-9469.00242

GRIFFIN, J. E., KOLOSSIATIS, M. and STEEL, M. F. J. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **75** 499–529. MR3065477 https://doi.org/10.1111/rssb.12002

GRIFFIN, J. E. and LEISEN, F. (2017). Compound random measures and their use in Bayesian non-parametrics. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **79** 525–545. MR3611758 https://doi.org/10.1111/rssb.12176

GRIFFIN, J. E. and STEEL, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc*. **101** 179–194. MR2268037 https://doi.org/10.1198/016214505000000727

GRIFFIN, J. E. and STEEL, M. F. J. (2010). Bayesian nonparametric modelling with the Dirichlet process regression smoother. *Statist. Sinica* **20** 1507–1527. MR2777334

GRIFFIN, J. E. and STEEL, M. F. J. (2011). Stick-breaking autoregressive processes. *J. Econometrics* **162** 383–396. MR2795625 https://doi.org/10.1016/j.jeconom.2011.03.001

GUTIÉRREZ, L., MENA, R. H. and RUGGIERO, M. (2016). A time dependent Bayesian nonparametric model for air quality analysis. *Comput. Statist. Data Anal*. **95** 161–175. MR3425946 https://doi.org/10.1016/j.csda.2015.10.002

GUTIÉRREZ, L., BARRIENTOS, A. F., GONZÁLEZ, J. and TAYLOR-RODRÍGUEZ, D. (2019). A Bayesian nonparametric multiple testing procedure for comparing several treatments against a control. *Bayesian Anal*. **14** 649–675. MR3959876 https://doi.org/10.1214/18-BA1122

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848

HÄRDLE, W. (1991). *Smoothing Techniques*: *With Implementation in S*. *Springer Series in Statistics*. Springer, New York. MR1140190 https://doi.org/10.1007/978-1-4612-4432-5

JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36** 76–97. MR2508332 https://doi.org/10.1111/j.1467-9469.2008.00609.x

JARA, A. and HANSON, T. E. (2011). A class of mixtures of dependent tail-free processes. *Biometrika* **98** 553–566. MR2836406 https://doi.org/10.1093/biomet/asq082

JARA, A., LESAFFRE, E., DE IORIO, M. and QUINTANA, F. (2010). Bayesian semiparametric inference for multivariate doubly-interval-censored data. *Ann. Appl. Stat.* **4** 2126–2149. MR2829950 https://doi.org/10.1214/10-AOAS368

JARA, A., HANSON, T., QUINTANA, F., MÜLLER, P. and ROSNER, G. L. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *J. Stat. Softw.* **40** 1–30.

KESSLER, D. C., HOFF, P. D. and DUNSON, D. B. (2015). Marginally specified priors for non-parametric Bayesian estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 35–58. MR3299398 https://doi.org/10.1111/rssb.12059

KLEMELÄ, J. (2014). *Multivariate Nonparametric Regression and Visualization*: *With R and Applications to Finance*. *Wiley Series in Computational Statistics*. Wiley, Hoboken, NJ. MR3222314

KOLOSSIATIS, M., GRIFFIN, J. E. and STEEL, M. F. J. (2013). On Bayesian nonparametric modelling of two correlated distributions. *Stat. Comput.* **23** 1–15. MR3018346 https://doi.org/10.1007/s11222-011-9283-7

LAU, J. W. and SO, M. K. P. (2008). Bayesian mixture of autoregressive models. *Comput. Statist. Data Anal.* **53** 38–60. MR2528591 https://doi.org/10.1016/j.csda.2008.06.001

LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235. MR1186248 https://doi.org/10.1214/aos/1176348767

LEISEN, F. and LIJOI, A. (2011). Vectors of two-parameter Poisson–Dirichlet processes. *J. Multivariate Anal.* **102** 482–495. MR2755010 https://doi.org/10.1016/j.jmva.2010.10.008

LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. MR3217444 https://doi.org/10.3150/13-BEJ521

LIJOI, A. and PRÜNSTER, I. (2009). Distributional properties of means of random probability measures. *Stat. Surv.* **3** 47–95. MR2529667 https://doi.org/10.1214/09-SS041

LIN, D., GRIMSON, E. and FISHER, J. W. (2010). Construction of dependent Dirichlet processes based on Poisson processes. In *Advances in Neural Information Processing Systems* 23 (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta, eds.) 1396–1404. Curran Associates, Red Hook.

LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519 https://doi.org/10.1214/aos/1176346412

MACEACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science* Amer. Statist. Assoc., Alexandria, VA.

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

MENA, R. H. and RUGGIERO, M. (2016). Dynamic density estimation with diffusive Dirichlet mixtures. *Bernoulli* **22** 901–926. MR3449803 https://doi.org/10.3150/14-BEJ681

MENA, R. H., RUGGIERO, M. and WALKER, S. G. (2011). Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling. *J. Statist. Plann. Inference* **141** 3217–3230. MR2796026 https://doi.org/10.1016/j.jspi.2011.04.008

MIRA, A. and PETRONE, S. (1996). Bayesian hierarchical nonparametric inference for change-point problems. In *Bayesian Statistics*, 5 (*Alicante*, 1994). *Oxford Sci. Publ.* 693–703. Oxford Univ. Press, New York. MR1425440

MULIERE, P. and PETRONE, S. (1993). A Bayesian predictive approach to sequential search for an optimal dose: Parametric and nonparametric models. *J. Ital. Stat. Soc.* **2** 349–364.

MÜLLER, P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79. MR1399156 https://doi.org/10.1093/biomet/83.1.67

MÜLLER, P. and QUINTANA, F. (2010). Random partition models with regression on covariates. *J. Statist. Plann. Inference* **140** 2801–2808. MR2651966 https://doi.org/10.1016/j.jspi.2010.03.002

MÜLLER, P., QUINTANA, F. and ROSNER, G. (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 735–749. MR2088779 https://doi.org/10.1111/j.1467-9868.2004.05564.x

MÜLLER, P., QUINTANA, F. and ROSNER, G. L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20** 260–278. MR2816548 https://doi.org/10.1198/jcgs.2011.09066

MÜLLER, P., QUINTANA, F. A., JARA, A. and HANSON, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer, New York.

PETRONE, S., GUINDANI, M. and GELFAND, A. E. (2009). Hybrid Dirichlet mixture models for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 755–782. MR2750094 https://doi.org/10.1111/j.1467-9868.2009.00708.x

PRÜNSTER, I. and RUGGIERO, M. (2013). A Bayesian nonparametric approach to modeling market share dynamics. *Bernoulli* **19** 64–92. MR3019486 https://doi.org/10.3150/11-BEJ392

QUINTANA, F. A. and IGLESIAS, P. L. (2003). Bayesian clustering and product partition models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 557–574. MR1983764 https://doi.org/10.1111/1467-9868.00402

REGAZZINI, E., LIJOI, A. and PRÜNSTER, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31** 560–585. MR1983542 https://doi.org/10.1214/aos/1051027881

REICH, B. J. and FUENTES, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Stat.* **1** 249–264. MR2393850 https://doi.org/10.1214/07-AOAS108

REN, L., DU, L., CARIN, L. and DUNSON, D. B. (2011). Logistic stick-breaking process. *J. Mach. Learn. Res.* **12** 203–239. MR2773552

RIGON, T. and DURANTE, D. (2021). Tractable Bayesian density regression via logit stick-breaking priors. *J. Statist. Plann. Inference* **211** 131–142. MR4117446 https://doi.org/10.1016/j.jspi.2020.05.009

RODRÍGUEZ, A. and DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6** 145–177. MR2781811 https://doi.org/10.1214/11-BA605

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. MR2528831 https://doi.org/10.1198/016214508000000553

RODRIGUEZ, A. and TER HORST, E. (2008). Bayesian dynamic density estimation. *Bayesian Anal.* **3** 339–365. MR2407430 https://doi.org/10.1214/08-BA313

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581. MR2279480 https://doi.org/10.1198/016214506000000302

TOKDAR, S. T., ZHU, Y. M. and GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal*. **5** 319–344. MR2719655 https://doi.org/10.1214/10-BA605

TRIPPA, L., MÜLLER, P. and JOHNSON, W. (2011). The multivariate beta process and an extension of the Polya tree model. *Biometrika* **98** 17–34. MR2804207 https://doi.org/10.1093/biomet/asq072

WADE, S., MONGELLUZZO, S. and PETRONE, S. (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Anal*. **6** 359–385. MR2843536 https://doi.org/10.1214/ba/1339616468

WANG, C. and ROSNER, G. L. (2019). A Bayesian nonparametric causal inference model for synthesizing randomized clinical trial and real-world evidence. *Stat. Med*. **38** 2573–2588. MR3962129 https://doi.org/10.1002/sim.8134

XU, Z., MACEACHERN, S. N. and XU, X. (2015). Modeling non-Gaussian time series with nonparametric Bayesian model. *IEEE Trans. Pattern Anal. Mach. Intell*. **37** 372–382.