ISSN 2279-9362

Collegio Carlo Alberto

Dynamic density estimation with diffusive Dirichlet mixtures

Ramsés H. Mena
Matteo Ruggiero

No. 369
October 2014

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Dynamic density estimation with diffusive Dirichlet mixtures

Ramsés H. Mena

Universidad Nacional Autónoma de México

Matteo Ruggiero*

University of Torino and Collegio Carlo Alberto

October 9, 2014

We introduce a new class of nonparametric prior distributions on the space of continuously varying densities, induced by Dirichlet process mixtures which diffuse in time. These select time-indexed random functions without jumps, whose sections are continuous or discrete distributions depending on the choice of kernel. The construction exploits the widely used stick-breaking representation of the Dirichlet process and induces the time dependence by replacing the stick-breaking components with one-dimensional Wright–Fisher diffusions. These features combine appealing properties of the model, inherited from the Wright–Fisher diffusions and the Dirichlet mixture structure, with great flexibility and tractability for posterior computation. The construction can be easily extended to multi-parameter GEM marginal states, which include for example the Pitman–Yor process. A full inferential strategy is detailed and illustrated on simulated and real data.

Keywords: density estimation, Dirichlet process, hidden Markov model, nonparametric regression, Pitman–Yor process, Wright–Fisher diffusion.

1 Introduction

Bayesian nonparametric inference has undergone a tremendous development in the last two decades. This has been stimulated not only by significant theoretical advances, but also by the availability of new and efficient computational methods that have made inference, based on analytically intractable posterior distributions, feasible. A recent extensive survey of the state-of-the-art of the discipline can be found in Hjort, Holmes, Müller and Walker (2010).

In this paper we tackle the problem of estimating continuously varying distributions, given data points observed at different time intervals, possibly not equally spaced. More specifically, we consider the following setting, stated as an assumption for ease of later reference.

*matteo.ruggiero@unito.it; http://sites.carloalberto.org/ruggiero

Assumption 1. *The data generating process is assumed to be a random function* $g : \mathbb{X} \times [0,T] \to \mathbb{R}_+$, $0 < T < \infty$, *where* $\mathbb{X}$ *is a locally compact Polish space,* g *is continuous in the sense that* $\sup_{|s-t|<\delta} \sup_{x\in\mathbb{X}} |g(x,s) - g(x,t)| \to 0$ *as* $\delta \to 0$, *and sections* $g(\cdot,t)$ *are densities absolutely continuous with respect to a common dominating measure.*

Here we assume that given $g(\cdot,t)$, the data are such that at time t_i

$$y_{t_{i,1}}, \ldots, y_{t_{i,k}} \overset{iid}{\sim} g(\cdot, t_i), \tag{1}$$

independently of the past observations $y_{t_j,h}$, for all $j < i$ and all h. Hence observations $y_{t_i,j}$ are exchangeable for fixed i and varying j, but only partially exchangeable in general. Our goal is to define an appropriate prior for g and carry out inference on its entire shape. We will deal with both single and multiple data points available at every time t_i, obtained by letting $k = 1$ and $k > 1$ in (1), and refer to these two settings as *single data* and *multiple data*, respectively. These are of separate interest as in some frameworks only single data points are available (e.g., financial indices); it is then natural to wonder about the model performance in such settings, with structural lack of information, while evaluating the precision gain that can be obtained when more information is available.

The commonly recognised cornerstone for density estimation in a Bayesian nonparametric framework is the Dirichlet process mixture model, introduced by Lo (1984) and recalled in Section 2 below. Here the idea is to extend some advantages of using a Dirichlet process mixture to the case when g is considered to be the realisation of a measure-valued process. Hence we aim at inducing a prior on the space of functions as in Assumption 1 by constructing a measure-valued process, with continuous sample paths and marginal states given by a Dirichlet process mixture, which is suitable for nonparametric inference in continuous time. Different types of temporal dependence induced on the observations are of interest. Here we focus on Markovianity for the mixing measure. Although potential applications of assuming a Markovian mixing measure concern a variety of fields such as econometrics, finance and medicine among others, this approach has been object of only a limited amount of research, so far, in the literature on Bayesian nonparametric inference. On the other hand, in a context of temporal dependence, the Markov property for the mixing measure considerably simplifies the finite dimensional distributions structure, and, most importantly, it need not impose a Markovian dependence on the actual observations.

Thus our setting can be regarded in two ways. One is that of a nonparametric regression, where the mixing measure is indexed by the covariate t, and the observations are partially exchangeable. A second interpretation, given our above requests, is that of a hidden Markov

model, whereby the unobserved signal is an infinite-dimensional, or measure-valued, Markov process. Conditionally on the knowledge of the signal $g(\cdot, t)$, the observations are independent of each other, and on the past observations, with emission distribution given by the signal state, as in (1).

The paper is organized as follows. Section 2 recalls the essential preliminary notions, like that of Dirichlet process and Dirichlet process mixture, and reviews to some extent the literature on the so-called dependent processes in Bayesian nonparametrics. These are in fact measure-valued processes, indexed by time or more generally by covariates, specifically designed for inferential purposes in Bayesian nonparametric frameworks with non-exchangeable data. Section 3 moves from the stick-breaking representation of the Dirichlet process to construct a class of diffusive Dirichlet processes. The idea consists in replacing the Beta-distributed components of the stick-breaking weights with one-dimensional Wright–Fisher diffusions. This yields a time-dependent process with purely-atomic continuous paths and marginal states given by Dirichlet processes. In order to have a statistical model suitable for estimating functions as in Assumption 1, we then define a diffusive Dirichlet mixture by considering an appropriate hierarchy whose top level is given by a diffusive Dirichlet process.

A challenging aspect of statistical models which involve diffusions regards the computational side, even when this is performed via simulation techniques with the aid of a computer. In a nutshell, this is due to the often encountered intractability of the transition density, in the fortunate cases when this is known explicitly. Here we devise a strategy for posterior computation based on Gibbs sampling and slice sampling, with the latter used both on the instant-wise infinite-dimensional mixing measure and on the transition density of the time-dependent components. After outlining the algorithm for posterior computation in Section 4, in Section 5 we illustrate the use of diffusive Dirichlet mixtures on two sets of simulated data and on real financial data. Section 6 collects some concluding remarks and briefly highlights possible extensions, concerned with the model parametrisation and with the aim of relaxing some model constraints. All proofs and the algorithm details are deferred to the Appendix.

2 Dependent processes in Bayesian nonparametrics

The Dirichlet process, introduced by Ferguson (1973) and widely accepted as the main breakthrough in the history of Bayesian nonparametric statistics, is a discrete random probability measure defined as follows. Let $\mathbb{X}$ be a Polish space endowed with the Borel sigma algebra $\mathcal{B}(\mathbb{X})$, $\mathscr{P}(\mathbb{X})$ be the space of Borel probability measures on $\mathbb{X}$, and let α be a nonatomic, finite and non-null measure on $\mathbb{X}$. A $\mathscr{P}(\mathbb{X})$-valued random variable Q is said to be a Dirichlet pro-

cess with parameter α, denoted $Q \sim \mathscr{D}_\alpha$, if for any measurable partition $A_1, \ldots, A_k$ of $\mathbb{X}$, the vector $(Q(A_1), \ldots, Q(A_k))$ has the Dirichlet distribution with parameters $(\alpha(A_1), \ldots, \alpha(A_k))$. A second construction of the Dirichlet process, still due to Ferguson (1973), exploits the idea of normalising the jumps of a gamma subordinator by their sum, locating the jumps at independent and identically distributed points sampled from $\alpha/\alpha(\mathbb{X})$. This strategy has been followed for constructing several other nonparametric priors, among which the normalised inverse-Gaussian process (Lijoi, Mena and Prünster, 2005) and the normalised generalized gamma process (Lijoi, Mena and Prünster, 2007). A later construction of the Dirichlet process, formalized by Sethuraman (1994) and particularly useful for our purposes, is usually referred to as the *stick-breaking representation*. This states that the law of a Dirichlet process coincides with the law of the discrete random probability measure

$$S = \sum_{i=1}^{\infty} w_i \delta_{x_i}, \tag{2}$$

obtained by letting

$$w_1 = v_1, \qquad w_i = v_i \prod_{j=1}^{i-1} (1 - v_j), \qquad v_i \overset{iid}{\sim} \text{Beta}(1, \theta), \tag{3}$$

and $x_i \sim^{iid} \alpha/\theta$, where $\theta = \alpha(\mathbb{X})$ and the v_i's and x_i's are mutually independent. The stick-breaking construction has received a wide appreciation from the Bayesian community. In particular this is due to the fact that it greatly facilitates the implementation of posterior simulation, using Markov chain Monte Carlo techniques that exploit the slice sampler (Damien, Wakefield and Walker, 1999; Walker, 2007) or the retrospective sampler (Papaspiliopoulos and Roberts, 2008). Such strategies for avoiding infinite computations, without deterministically truncating the random measure, can be used for example with Dirichlet process mixtures (Lo, 1984). The latter are a very popular class of models which amplifies the use of Dirichlet processes to a wider spectrum of statistical applications, most notably density estimation and data clustering, by modelling observations according to the random density

$$f_S(y) = \int K(y|\, x) S(\mathrm{d}x) = \sum_{i=1}^{\infty} w_i K(y|\, x_i) \tag{4}$$

where $K(\cdot|y)$ is a kernel density and S is as in (2). Thanks to the large support properties of the Dirichlet prior, the above model results in a very flexible class of distributions: for instance, any density on the real line can be recovered as an appropriate Dirichlet mixture of normal densities. For this and other examples see Lo (1984), Section 3 and Ghosh and Ramamoorthi (2003), Section 5.

Many developments which derive from this modelling approach have been proposed. A first possibility is to replace the Dirichlet process in (4) by letting S be some other discrete non parametric prior. See Lijoi and Prünster (2010) and references therein. A different direction, which currently represents a major research frontier of the area, is to extend S in (4) in order to accommodate forms of dependence more general than exchangeability. Besides pioneering contributions stimulated by Cifarelli and Regazzini (1978), this line of research was initiated by MacEachern (1999; 2000), who proposed a class of *dependent processes*, that is a collection of random probability measures

$$\left\{ S_u = \sum_{i=1}^{\infty} w_i(u)\delta_{x_i(u)},\ u \in \mathbb{U} \right\} \tag{5}$$

where the weights w_i and/or the atoms x_i depend on some covariate $u \in \mathbb{U}$. The current literature on the topic includes, among other contributions: De Iorio, Müller, Rosner and MacEachern (2004), who proposed a model with an ANOVA-type dependence structure; Gelfand, Kottas and MacEachern (2005), who apply the dependent Dirichlet process to spatial modelling by using a Gaussian process for the atoms; Griffin and Steel (2006), who let the dependence on the random masses be directed by a Poisson process; Dunson and Park (2008), who construct an uncountable collection of dependent measures based on a stick-breaking procedure with kernel-based weights; Rodriguez and Dunson (2011), who replace the Beta random variables in (3) with a transformation of Gaussian processes via probit links. See also Dunson, Pillai and Park (2007), who define a Dirichlet mixture of regression models; Dunson, Xue and Carin (2008), who propose a matrix-valued stick-breaking prior; Duan, Guindani and Gelfand (2007) and Petrone, Guindani and Gelfand (2009) for other developments of dependent priors for functional data; Fuentes-García, Mena and Walker (2009) for a dependent prior for density regression; Trippa, Müller and Johnson (2011), who define a dependent process with Beta marginals.

Of particular interest for the purposes of this paper are the developments of dependent processes where the space $\mathbb{U}$ indexes time. In this regard we mention, among others, Dunson (2006), who models the dependent process as an autoregression with Dirichlet distributed innovations, whereas in Griffin and Steel (2010) the innovation is reduced to a single atom sampled from the centering measure; Caron et al. (2008), who model the noise in a dynamic linear model with a Dirichlet process mixture; Caron, Davy and Doucet (2007), who develop a time-varying Dirichlet mixture with reweighing and movement of atoms; Rodriguez and Ter Horst (2008), who induce the dependence in time only via the atoms, by making them into an heteroskedastic random walk.

Here we aim at constructing a measure-valued diffusion whose realisations are functions as in Assumption 1. In this respect, the infinite-dimensional diffusions, found in the literature, which are related to Bayesian nonparametric priors are not suitable for efficient inference. Just to mention a few cases, this holds for example for the infinitely-many-alleles model (Ethier and Kurtz, 1981; 1986), related to the Dirichlet process; for its two-parameter version (Petrov, 2009), related to the two-parameter Poisson–Dirichlet distribution; and for normalised inverse-Gaussian diffusions (Ruggiero, Walker and Favaro, 2013), related to normalised inverse-Gaussian random measures. The reasons are mainly due to the fact that an inferential strategy based on these dependent processes, given our current knowledge of their properties, would oblige to update single atoms of a Pólya urn scheme (see, e.g., Ruggiero and Walker, 2009; Favaro, Ruggiero and Walker, 2009), or otherwise face serious computational issues. To the best of our knowledge, the only model which satisfies the given requirements, among which the Feller property, and allows efficient inference is given in Mena, Ruggiero and Walker (2011). However, this is based on decreasingly ordered weights, so, despite showing good performance, this feature can nonetheless be considered restrictive for certain applications. The model developed in the next section removes this constraint by letting only the weights' means be ordered, as happens for the Dirichlet process.

3 Diffusive Dirichlet process mixtures

In this section we elaborate on a construction provided in Feng and Wang (2007), in order to develop a class of measure-valued processes, suitable to be used in a statistical model, with the sought-after features outlined in the Introduction. Consider the special case of (5) given by the collection of random probability measures

$$P_t = \sum_{i=1}^{\infty} w_i(t)\delta_{x_i}, \quad t \geq 0, \quad x_i \overset{iid}{\sim} G, \tag{6}$$

where $\sum_{i\geq 1} w_i(t) = 1$ for all $t \geq 0$ and G is a non atomic probability measure on $\mathbb{X}$. Here the atoms are random but do not vary with time, and the dependence is induced only via the weights. This setting suffices for guaranteeing enough modelling flexibility (see Section 6 for more comments on this point), whereas the opposite scheme, obtained by inducing the dependence only via the atoms, may be unsatisfactory. See for example Rodriguez and Dunson (2011) for a discussion. Recall now the stick-breaking structure of the Dirichlet process weights (3). A natural extension for having time dependence with continuity in t is to let w_i diffuse in time, in a way that retains the marginal distributions. A simple way

to achieve such result is to let each component v_i diffuse in $[0,1]$, with fixed marginals, and perform the same construction as in (3). This can be obtained by letting the v_i's be one-dimensional Wright–Fisher diffusions, characterised as the unique solution in $[0,1]$ of the stochastic differential equation

$$\mathrm{d}v(t) = \frac{1}{2}[a(1-v(t)) - bv(t)]\mathrm{d}t + \sqrt{v(t)(1-v(t))}\mathrm{d}B(t), \tag{7}$$

where $a, b \geq 0$ and $B(t)$ denotes a standard Brownian motion. See for example Karlin and Taylor (1981), Section 15.2. For our purposes it will suffice to highlight the following properties of Wright–Fisher diffusions $v(t)$ with parameters (a,b) as in (7), henceforth denoted $v(\cdot) \sim \mathrm{WF}(a,b)$:

- when $v(t)$ approaches 0 (resp. 1), the diffusion coefficient converges to 0 and the drift approaches $a/2$ (resp. $-b/2$), thus keeping the diffusion inside $[0,1]$;
- when $a, b \geq 1$, the points 0 and 1 are both *entrance boundaries*, implying (essentially) that they are never touched for $t > 0$;
- when $a, b > 0$, $v(t)$ has invariant distribution given by a $\mathrm{Beta}(a,b)$;
- when $a, b > 0$, $v(t)$ is strongly ergodic, that is, irrespective of the initial distribution, the law of $v(t)$ will converge to the invariant measure as t diverges.

A typical behavior of $v(t)$, together with the occupancy frequencies plotted against the invariant distribution, is shown in Figure 1, for values $(a,b) = (1,4)$. Note how these parameters make the trajectory occupy the half interval containing the mean value of $\mathrm{Beta}(1,4)$ for most of the time.

The idea is then to let every v_i in (3) vary, independently of the other components, according to a Wright–Fisher diffusion with parameters $(1,\theta)$, that is $v_i(\cdot) \sim^{iid} \mathrm{WF}(1,\theta)$, with $v_i(\cdot) = \{v_i(t), t \geq 0\}$, and to construct diffusive stick-breaking weights

$$w_1(t) = v_1(t), \qquad w_i(t) = v_i(t)\prod_{j<i}(1-v_j(t)), \qquad v_i(\cdot) \overset{iid}{\sim} \mathrm{WF}(1,\theta). \tag{8}$$

The process $w(\cdot) = \{w(t), t \geq 0\}$ defined above for the vector of weights $w(t) = (w_1(t), w_2(t), \ldots)$ has been characterised by Feng and Wang (2007), who investigate its sample path properties. It is clear that the constraints $a = 1$ and $b = \theta$ are by no means essential but only chosen to preserve the connection with the Dirichlet process (see Proposition 3.3 below). Section 6 will briefly discuss possible extensions. With this formulation, (6) defines a family of dependent random probability measures which retain at every time point the stick-breaking structure

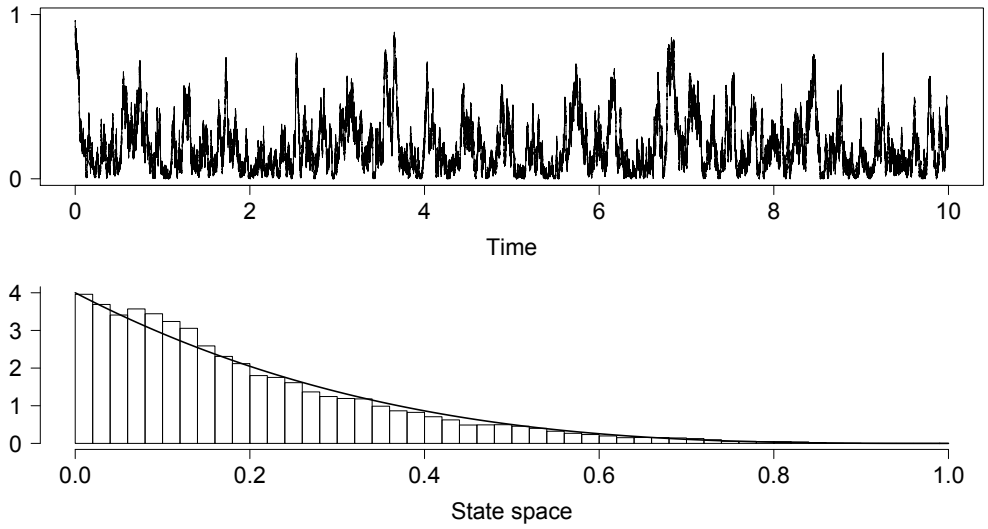

Figure 1: Above: an approximated sample-path of a Wright–Fisher diffusion, with $a = 1$ and $b = 4$. Below: ergodic frequencies of the above sample-path against invariant distribution Beta$(1, 4)$.

featured by (3). For ease of reference we summarise the construction in the following definition.

Definition 3.1. *A family of dependent random probability measures $P = \{P_t,\, t \geq 0\}$ with representation (6)-(7)-(8) is said to be a* diffusive Dirichlet process.

Besides studying $w(\cdot)$, Feng and Wang (2007) also consider a construction more general than (6), where the atoms $(x_1, x_2, \dots,)$ are let to be a Markov process on $\mathbb{X}^\infty$. However their model is too general for our purposes and its properties are hard to establish without further assumptions on the model components. Hence we need to formalise the path properties of P, as these cannot be deduced directly from either of the models considered in their paper.

To this end, denote by $C_{\mathscr{P}(\mathbb{X})}([0,\infty))$ the space of continuous functions from $[0,\infty)$ to $\mathscr{P}(\mathbb{X})$. Here P denotes the $C_{\mathscr{P}(\mathbb{X})}([0,\infty))$-valued random element $\{P_t, t \geq 0\}$, and P_t its coordinate projection at t, so that $P_t \in \mathscr{P}(\mathbb{X})$. We endow $\mathscr{P}(\mathbb{X})$ with the topology induced by the total variation norm, so that elements of $C_{\mathscr{P}(\mathbb{X})}([0,\infty))$ have modulus of continuity

$$\omega(P,\delta) = \sup_{|s-t|<\delta} \sup_{A\in\mathcal{B}(\mathbb{X})} |P_t(A) - P_s(A)|,$$

and $P \in C_{\mathscr{P}(\mathbb{X})}([0,\infty))$ if and only if $\omega(P,\delta) \to 0$ as $\delta \to 0$, that is if $P_{t+s} \to P_t$ in total variation distance as $s \to 0$. See Billingsley (1968), Chapter 2. The fact that the diffusive Dirichlet Process P in Definition 3.1 has continuous sample paths in total variation, should be intuitive from the construction, since the only time-varying quantities are diffusion processes mapped through a continuous function. The following Proposition formalises this fact.

Proposition 3.2. *Let P be as in Definition 3.1. Then P is a Feller process with realisations almost surely in $C_{\mathscr{P}(\mathbb{X})}([0,\infty))$.*

The Feller property for P guarantees certain desirable path properties which, among other things, yield the well-definedness of the process and its Markovianity. We refer the reader to Ethier and Kurtz (1986), Chapter 4, for more details on Feller operators. However, the continuity of sample paths is not implied by the Feller property and is proven separately. In particular, it is such continuity that will allow, after embedding P in an appropriate statistical model, to select almost surely continuous functions as in Assumption 1.

Since the stationary distribution of the Wright–Fisher diffusion, used in (8), is the same distribution used in (3) for the stick-breaking components, it is also intuitive that the process of Definition 3.1 is stationary with respect to the law of a Dirichlet process, as stated by the next proposition.

Proposition 3.3. *Let P be as in Definition 3.1 and, for a finite non-null measure α on $\mathbb{X}$, let $\mathscr{D}_\alpha$ denote the law of a Dirichlet process. Then P is reversible and stationary with respect to $\mathscr{D}_\alpha$. In particular, if $P_0 \sim \mathscr{D}_\alpha$, then $P_t \sim \mathscr{D}_\alpha$ for every $t > 0$.*

Hence the marginal states of the diffusive Dirichlet process are Dirichlet distributed. Here it is important to note that modelling the above measure-valued process as stationary will not constrain the data to come from a stationary process. This will be more transparent when we will consider the hierarchical statistical model for the data. On the contrary, this aspect will turn into an advantage since it will allow to propagate in time the support properties of the Dirichlet prior. Indeed it is well known that the Dirichlet prior has full weak support. That is, if $\mathbb{X}$ is the support of the parameter measure α, then the support of $\mathscr{D}_\alpha$ in the weak topology is

$$\operatorname{supp}(\mathscr{D}_\alpha) = \{Q \in \mathscr{P}(\mathbb{X}) : \ \operatorname{supp}(Q) \subset \mathbb{X}\}. \tag{9}$$

See Ghosh and Ramamoorthi (2003), Section 3.2.3. Proposition 3.3 then implies that at stationarity P_t, marginally, has support (9). See also Barrientos, Jara and Quintana (2012) for

sufficient conditions for having full weak support in the context of dependent stick-breaking processes.

Another byproduct of Proposition 3.3 is the immediate derivation of the marginal moments of P. In particular, let $P_0 \sim \mathscr{D}_\alpha$, with $\alpha = \theta G$, where $\theta > 0$ and $G \in \mathscr{P}(\mathbb{X})$. Then, for all $t \geq 0$ and $A \in \mathcal{B}(\mathbb{X})$,

$$\mathbb{E}[P_t(A)] = G(A), \qquad \mathrm{Var}[P_t(A)] = \frac{G(A)(1-G(A))}{\theta+1}, \tag{10}$$

where $P_t(A)$ denotes (6) evaluated at the set A. In addition, the following proposition provides an explicit expression for the autocorrelation function of the process.

Proposition 3.4. *Let P be as in Definition 3.1. Then, for any $A \in \mathcal{B}(\mathbb{X})$ and any $t, s > 0$,*

$$\mathrm{Corr}(P_t(A), P_{t+s}(A)) = \frac{(1+\theta)[(2+\theta)+\theta e^{-\lambda s}]}{(2+\theta)(1+2\theta)-\theta e^{-\lambda s}} \tag{11}$$

where $\lambda = (1+\theta)/2$.

As expected, the correlation does not depend on the set A, since the time dependence enters only via the weights and not the locations. Furthermore, it is easily seen that the correlation decays exponentially to $(1+\theta)/(1+2\theta)$ as $s \to \infty$. Although this can perhaps be considered as an undesirable property, the existence of a lower bound for the correlation is a common feature of all dependent processes whose atoms are fixed (see Rodriguez and Dunson, 2011). In Section 6 we will provide more comments on this point and outline a possible extension which aims at relaxing some of the model constraints.

The above dependent process can be used to formulate a dependent mixture model by considering the time-varying density

$$f_{P_t}(y) = \int K(y|x) P_t(\mathrm{d}x), \tag{12}$$

where $K(\cdot|x)$ is a kernel density with parameter x. An equivalent formulation is provided in the form of the hierarchical model

$$\begin{aligned} y|x &\sim K(y|x) \\ x|t, P_t &\sim P_t \\ P &\sim \text{diff-DP}, \end{aligned} \tag{13}$$

where $P \sim$ diff-DP denotes that P is a diffusive Dirichlet process. Since P is the only time-varying component in the statistical model, the dependent mixture inherits the diffusive

behavior from P, and Proposition 3.3 implies that marginally f_{P_t} is a Dirichlet process mixture. Thus the dependent mixture induces a prior distribution on the space $C_{\mathscr{P}(\mathbb{X})}([0,\infty))$ of continuous functions from $[0,\infty)$ to $\mathscr{P}(\mathbb{X})$, which almost surely selects functions $g : \mathbb{X} \times [0,T] \to \mathbb{R}_+$ that satisfy Assumption 1. The choice of kernel K determines the discrete or continuous nature of the sections $g(\cdot,t) = f_{P_t}(\cdot)$ and their support.

4 Posterior computation

We overview the strategy for simulating from the posterior distribution of (12) and of other quantities of interest, such as for example the mean functional process $\eta_t = \mathbb{E}_{f_{P_t}}(y)$, which depicts the average stochastic process driving the observations. Here we highlight the main points of interest, while the fully detailed procedure can be found in Appendix B. For notational simplicity, and in view of the real data example below, we assume a single data setting. However, the strategy allows a straightforward extension to the case of multiple observations at every time point.

Specifically, we assume data $y^{(n)} = (y_{t_1}, \ldots, y_{t_n})$ are observed at times $0 \le t_1 < \ldots < t_n$, where time intervals are not necessarily equally spaced. The target of inference is the data generating time-varying distribution $g(y,t)$, such that $y_{t_i} \sim g(\cdot, t_i)$. We model such g by means of a diffusive Dirichlet mixture, with $g(\cdot,t) = f_{P_t}(\cdot)$. To this end, let $v(\cdot) = ((v_1(t), v_2(t), \ldots), t \ge 0)$ denote a collection of independent Wright–Fisher diffusions defined as in (7), with $v_j(\cdot) \sim^{iid} \mathrm{WF}(1,\theta)$. Let also $x = (x_1, x_2, \ldots)$ be the random locations sampled from a nonatomic probability measure G. Hence the data generating process is modelled as

$$f_{P_t}(y \mid v(t), x) = \sum_{j \ge 1} w_j(t)\, K(y \mid x_j). \tag{14}$$

The model induces dependence among the observations, which are exchangeable for fixed t but partially exchangeable in general.

The infinite dimensionality of the random measure at t is dealt with via slice sampling (Damien, Wakefield and Walker, 1999; Walker, 2007). Specifically, we extend the slice algorithm in Kalli, Griffin and Walker (2011) to augment the above random density by

$$f_{P_t}(y, u, s \mid v(t), x) = \frac{\mathbb{I}(u < \psi_s)}{\psi_s}\, w_s(t)\, K(y \mid x_s) \tag{15}$$

where $s \mapsto \psi_s$ is a decreasing function with known inverse ψ^*, e.g. $\psi_s = e^{-\eta s}$, for $0 \le \eta \le 1$. The latent variable s indexes which of the kernels $K(\cdot \mid x_s)$ better captures the mass at y, and given s, $u \sim U(0, \psi_s)$. For the purpose of estimation it is enough to condition on

$v^{(n)} = \{v(t_i)\}_{i=1}^n$, rather than on the whole path $v(\cdot)$. In this section and in the appendix we will use the notation $\mathcal{L}(z)$ to indicate generically the law of z. The conditional augmented likelihood is given by

$$\mathcal{L}\Big(y^{(n)}, u^{(n)}, s^{(n)} \mid v^{(n)}, x\Big) = \prod_{i=1}^{n} \frac{\mathbb{I}(u_i < \psi_{s_i})}{\psi_{s_i}} \Big[v_{s_i}(t_i) \prod_{k<s_i} (1 - v_k(t_i))\Big] K(y_{t_i} \mid x_{s_i}) \tag{16}$$

where $u^{(n)} = (u_1, \ldots, u_n)$ and $s^{(n)} = (s_1, \ldots, s_n)$, with $u_i = u_{t_i}$ and $s_i = s_{t_i}$. Due to the random truncation induced by the slice sampling method, one is only able to learn about the first m Wright–Fisher processes and locations, where

$$m = \max(\lfloor \psi^*(u_1) \rfloor, \lfloor \psi^*(u_2) \rfloor, \ldots, \lfloor \psi^*(u_n) \rfloor), \tag{17}$$

and $\lfloor A \rfloor$ denotes the integer part of A. Hence, denoting

$$v_{1:m}^{(n)} = (v_1^{(n)}, \ldots, v_m^{(n)}), \qquad x_{1:m} = (x_1, \ldots, x_m), \tag{18}$$

we see that

$$\mathcal{L}\left(v_{1:m}^{(n)}, x_{1:m} \mid y^{(n)}, u^{(n)}, s^{(n)}\right) \propto \mathcal{L}\left(y^{(n)}, u^{(n)}, s^{(n)} \mid v_{1:m}^{(n)}, x_{1:m}\right) \mathcal{L}(v_{1:m}^{(n)}) \, \mathcal{L}(x_{1:m}),$$

while the posterior distribution remains the unchanged prior for $l > m$. Here the m processes and locations are mutually independent, so that

$$\mathcal{L}(v_{1:m}^{(n)}) = \prod_{j=1}^{m} \mathcal{L}(v_j^{(n)}), \qquad \mathcal{L}(x_{1:m}) = \prod_{j=1}^{m} \mathcal{L}(x_j).$$

It remains to observe that the finite dimensional distributions for each Wright–Fisher process are

$$\mathcal{L}(v_j^{(n)}) = \pi_v(v_j(t_1)) \prod_{i=2}^{n} p_v(v_j(t_i) \mid v_j(t_{i-1})) \tag{19}$$

where $\pi_v = \text{Beta}(1, \theta)$ and p_v denotes the transition density of the Wright–Fisher diffusion. This is known explicitly (Ethier and Griffiths, 1993), but has an infinite series representation. In view of a further implementation of the slice sampler, we resort to the representation of p_v due to Mena and Walker (2009), which reads

$$p_v(v(t) \mid v(0)) = \sum_{m=0}^{\infty} r_t(m) \, D(v(t) \mid m, v(0)), \tag{20}$$

Algorithm 4.1. *Gibbs sampler for diffusive DP mixtures*

1: **input:** data $(y_{t_1}, \ldots, y_{t_n})$ and their recording times $(t_1, \ldots, t_n)$
2: **initial values for:**
   a) Hyper-parameters in G and WF parameters
   b) Upper limit $m^{(0)}$ for the number of locations and WF processes
   c) Membership variables $s^{(0)} = (s_{t_1}, \ldots, s_{t_n})$, taking values in $\{1, \ldots, m^{(0)}\}$
   d) WF processes at $(t_1, \ldots, t_n)$ and corresponding latent variables
   e) Location values
3: for $\mathrm{I} = 1$ to $ITER$
4: Sample slice variables $u_{t_i} \sim U(0, \psi_{s_{t_i}})$ and update the value of m_I
5: if $m^{(\mathrm{I})} > m^{(\mathrm{I}-1)}$ take extra WF processes and their latent variables from priors
6: Update transition density latent variables
7: Update WF processes values
8: Update location values
9: Update WF parameter values
10: Update membership variables $s^{(\mathrm{I})}$ taking values in $\{1, \ldots, m^{(\mathrm{I})}\}$

where $r_t(m)$ is an appropriate deterministic function and $D(v(t) \mid m, v(0))$ a finite mixture of Beta distributions. This leads to an augmentation of p_v similar to that outlined above for f_{P_t}, allowing to avoid the infinite computation.

With the above specification, a Gibbs sampler algorithm can be designed as in Algorithm 4.1. The reader is referred to Appendix B for the algorithm details not included in this section.

5 Illustration

Following the above framework, in this section we illustrate an application of the diffusive Dirichlet process with simulated and real data. More specifically, we consider:

(i) simulated observations sampled at equally spaced intervals from the time-dependent normal density

$$\mathrm{N}(\cos(2t) + t/2, 1/10); \tag{21}$$

(ii) 300 real observations given by daily exchange rates between US dollars and Mexican peso during the period from September 26th, 2008, to December 7th, 2009.

The assumption of data collected at regular intervals here is for computational simplicity only, as the model, through (20), allows for not equally spaced samples.

In order to complete the specification of the diffusive mixture we use a conjugate vanilla choice for the kernel K and centering measure G. Specifically, let

$$K(y \mid x) = \mathrm{N}(y \mid m, v^{-1}),$$
$$G(x) = \mathrm{N}\left(m \mid 0, 1000v^{-1}\right) \mathrm{Gamma}\left(v \mid 10, 1\right),$$

with $x = (m, v)$. The parametrization for G is chosen to achieve a large variance at the location level, to cover all observations with high probability. This is required, as the locations are random but fixed over time, and the weight processes should be able to pick any good candidate location within the data state space at a given time. Running Algorithm 4.1 allows to draw posterior inferences for any functional of the diffusive Dirichlet mixture model. In particular, besides the time-varying density, we are interested in the mean functional

$$\eta_t = \int_{\mathbb{R}} y f_{P_t}(dy). \tag{22}$$

All examples in this section are based on 2000 effective iterations drawn from 10000 iterations thinned each five and after a burn in period of 5000 iterations. We verified practical convergence using the convergence diagnostics of Gelman and Rubin (1992) and of Raftery and Lewis (1992), and neither showed any evidence of convergence problems.

Figure 2 shows the results corresponding to a first dataset simulated from (21), where 100 single observations are collected at equally spaced intervals. The true data generating process is shown as a heat contour, with regions of more intense red being those of higher probability, and presents traits of non-stationarity and seasonality. The black dots are the simulated data, the solid and dashed black lines are the pointwise posterior mode and 95% credible intervals for the mean functional respectively, and the blue dotted lines are the pointwise 95% quantiles of the posterior estimate of the time-varying density. The picture shows that the model correctly captures the non-stationary behaviour with a strong trend and seasonalities. Furthermore, even in this setting with structural lack of instant-wise information, due to the availability of only single data points, the uncertainty of the estimates is reasonably low, as the model instantaneous variance is indirectly learned by the algorithm from the overall data pool. The smoothness level captured by the estimation is also acceptable, given the above considerations on the lack of information and considered that the model is ultimately based on Wright–Fisher components.

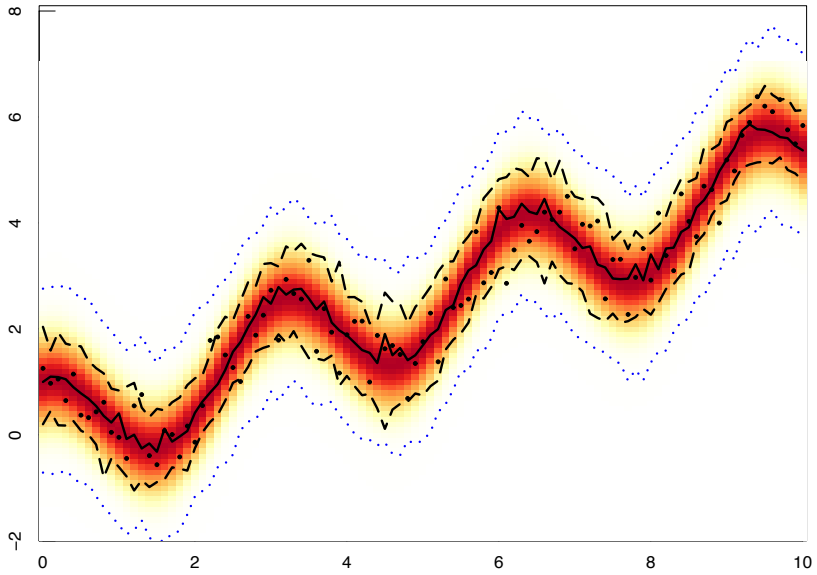

Figure 2: MCMC-based estimation for single data points (black dots) sampled at equally spaced intervals from (21). The picture shows the true model (heat contour), the pointwise posterior mode (solid black line) and 95% credible intervals for the mean functional (dotted black lines), and the pointwise 95% quantiles of the posterior estimate of the time-varying density (blue dotted lines).

In order to investigate the degree of improvement one can gain with multiple data points, we performed the same type of inference as in Figure 2 on a second set of data simulated from (21), where at each of the 100 time points, five data are available. The results, reported in Figure 3, show that the accuracy of the estimation increases satisfactorily, as the true model behaviour is captured with considerably less uncertainty. This is especially true for the mean functional (solid black line), whose credible intervals (dashed black lines) are very narrow. It is also important to note that the smoothness of the true model is also correctly learned by the estimates. This is due to the fact that the rough behaviours of the model subcomponents are confined to levels of the hierarchy where their impact on the final estimation, in presence of enough information, is pooled together and softened by adapting the component-specific volatilities. This however does not prevent the model from capturing quick deviations from

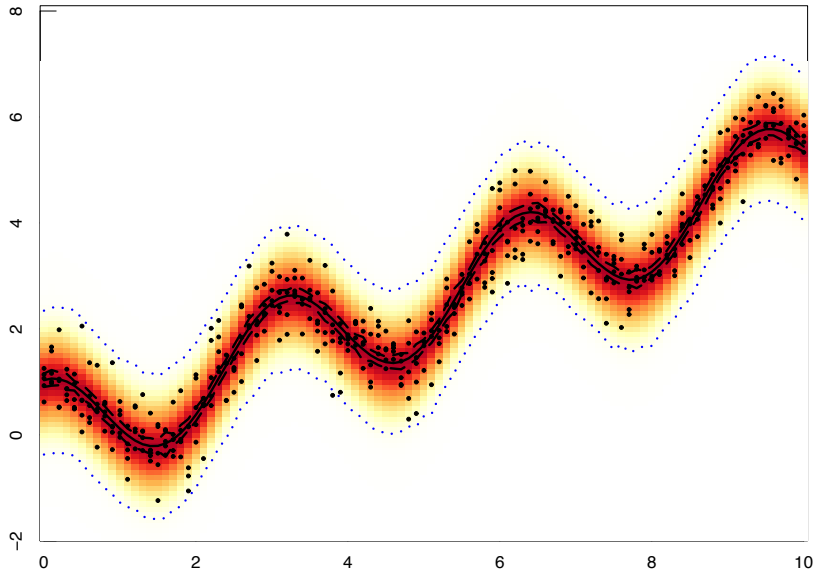

Figure 3: MCMC-based estimation for multiple data points (black dots) sampled at equally spaced intervals from (21). The picture shows the true model (heat contour), the pointwise posterior mode (solid black line) and 95% credible intervals for the mean functional (dotted black lines), and the pointwise 95% quantiles of the posterior estimate of the time-varying density (blue dotted lines).

a smooth trend, as showed by the next example.

For the illustration with real observations, we concentrate on a challenging set of single data points per observation time, which again provides scarce instantaneous information. Financial data sets as in (ii) are often described with parametric state-space models. This can put serious constraints on the ability of the model to capture the correct marginal distributions and quick deviations from the general pattern. By making the nonparametric assumption that the state-space model follows a diffusive Dirichlet process mixture, we are relaxing such constraints and granting great flexibility to the model. In the interpretation related to hidden Markov models, the unobserved signal here models an evolving distribution, driven by a measure-valued Markov process, and the observations are sampled from the signal states. Note however that this framework does not impose any Markovianity nor stationarity

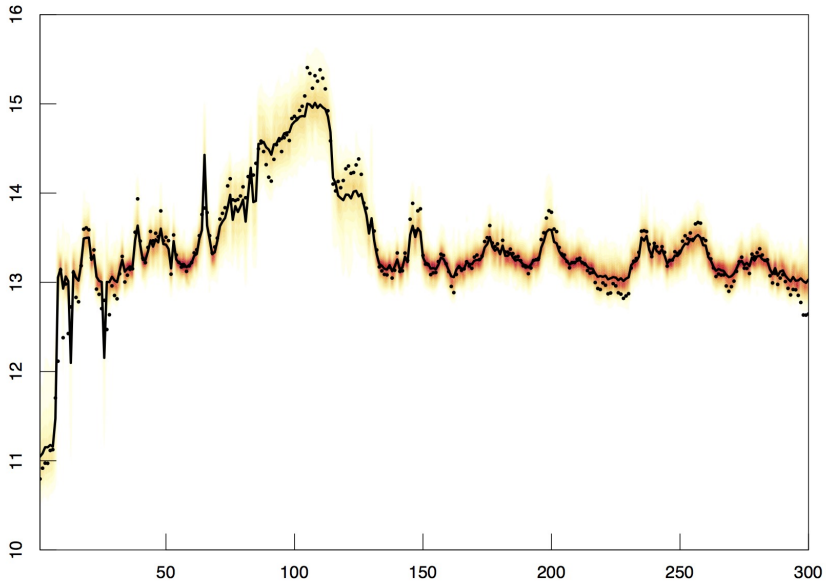

Figure 4: MCMC-based pointwise posterior density estimate (heat contour) and pointwise posterior mode of the mean functional (solid line), based on single data points (black dots) corresponding to dataset (ii).

on the observations.

Figure 4 shows the results on the exchange rate data set, with the horizontal axis representing time and the vertical axis representing the index value. The red heat contour outlines the shape of the pointwise posterior estimate for the time-varying density function f_{P_t}, with regions of more intense red corresponding to higher posterior probability. The black solid line is the pointwise mode of the posterior mean functional (22). The model is able to capture highly volatile behaviours, such as that encountered in the period between 70 and 130. These kind of changes are typically not well recognised by parametric models for time series, which are too rigid to allow for unexpected detours. Figure 5 shows a sub-region of Figure 4, where sudden spikes and different local trends are also shown to be correctly captured, regardless of how abrupt these appear. Another important aspect to be noted is the fact the regions of high estimated probability need not correspond to the regions where data are observed, even having single data points, as in the central sub period around day 225 in Figure 5. This

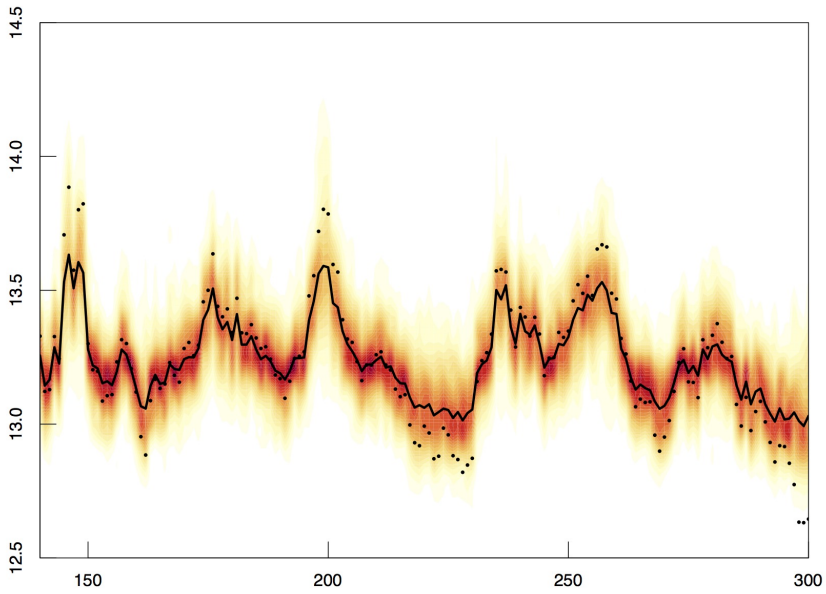

Figure 5: A sub-region of Figure 4.

feature is determined by the model dependence, which allows to borrow information across time, and should not be confused with model rigidity, as the model clearly captures sudden deviations from the trend as that occurring around time 200.

6 Discussion, extensions and future work

We introduced a new class of prior distributions on the space of time-indexed, t-continuous functions $g : \mathbb{X} \times [0, T] \to \mathbb{R}_+$, such that $g(\cdot, t)$ is a density for all $t \in [0, T]$. Such priors are induced by diffusive Dirichlet process mixtures, which extend the Dirichlet process mixture model of Lo (1984) to a framework of Feller measure-valued processes with continuous trajectories. The resulting dependent random density (12) can be used to tackle various statistical problems of interest in many fields such as econometrics, finance and medicine among others, when the random phenomena evolving in continuous time are not satisfactorily modelled parametrically. On the other hand, it can be an alternative to other nonparametric dependent models which are statistically intractable. For example, when the underlying model

structure is linked to population dynamics, it could be desirable to model data by means of Fleming–Viot processes (see Ethier and Kurtz (1993)). However, such a process is computationally intractable in view of inference, and the presented model can then be used to this end, as the main properties of the former such as the stationary measure and path regularity are preserved.

Overall, the introduced model exhibits a good mix of flexibility and structure, yet leaving room for computational efficiency. The full support property of the Dirichlet process and the time dependence of Wright–Fisher diffusions are combined to yield an equilibrated compromise between adaptivity and dependence, and the Gibbs sampler with slice steps is simple to implement. These aspects lead to the ability of jointly detecting smooth and irregular evolutions of the distribution which generates the observations, while borrowing strength between observations when this is needed.

We briefly outline two possible extensions of the model, concerned with certain features that can nonetheless be considered as constraints in certain contexts. The first is a reparametrisation which retains the overall structure, and aims at broadening the set of stationary distributions of the model. The second changes the qualitative features of the model and aims at removing some rigidities. These are concerned with the fact that the correlation is bounded from below, together with the fact that the dependence structure imposes some restrictions on the data generating mechanism.

The construction can be easily extended by relaxing the assumption of identity in distribution of the WF diffusions used in (8). This can be done, for example, by considering the class of GEM diffusions, also developed in Feng and Wang (2007), to allow for more general stationary distributions of the weights, such as two-parameter Poisson–Dirichlet distributions (Pitman, 1995; Pitman and Yor, 1997) or GEM distributions (Johnson, Kotz and Balakrishnan (1997), Chapter 41). A general construction of the latter class of random measures is obtained by taking $v_i \sim^{ind} \text{Beta}(a_i, b_i)$ in (3), while the former corresponds to choosing $a_i = 1 - \sigma$ and $b_i = \theta + i\sigma$ for all $i \geq 1$. By analogy with the construction in Section 3, GEM-type measure-valued diffusions can be defined by letting $v_i(\cdot) \sim^{ind} \text{WF}(a_i, b_i)$ in (8), that is the collection of Wright–Fisher diffusions which induce the time-dependence in the weights of (6) is given by independent, and no longer identically distributed, processes. An appropriate extension of Proposition 3.3 easily follows. One can then replace P in (13) with the resulting measure-valued GEM process to yield additional modelling flexibility. The resulting class of dependent mixture models extends, to a time-dependent framework, the priors considered in Ishwaran and James (2001). Of course there is a trade-off between the amount of flexibility one pursues and the amount of parameters one is willing to deal with in terms of

computational effort. The choice of two-parameter Poisson–Dirichlet distributions probably guarantees extra flexibility at almost no extra cost, as it allows to control more effectively the posterior distribution of the number of clusters (Lijoi, Mena and Prünster, 2007), and only requires one additional step at every Gibbs sampler iteration for updating the posterior distribution of σ.

A different direction can be considered with the aim of relaxing the dependence structure the model imposes on the data. This is the object of a currently ongoing work by the authors, of which we concisely outline the main ideas. We consider a specific choice for the general dependent process

$$P_t^{(\gamma)} = \sum_{i=1}^{\infty} w_i(t)\delta_{x_i(t)}, \quad t \geq 0,$$

parametrised by $\gamma > 0$, which keeps the model complexity and the implied computational burden relatively low, but yields an autocorrelation which vanishes exponentially fast and whose structure can be inferred from the data. This is obtained, for example, by letting the weights $w(\cdot)$ be as in (8), and by letting the initial atoms $x_i(0) \sim^{iid} G$ be updated one at a time after an interval with $\text{Exp}(\gamma)$ distribution from the last update. The atom to be replaced is chosen, e.g., according to a fixed distribution, and the update is chosen from G. The notation $P^{(\gamma)}$ for the resulting model highlights the role of the intensity γ of the underlying Poisson point process which regulates how often these innovations occur. Such specification extends the model of Section 3 to a model which is still stationary with respect to the law of a Dirichlet process, and has a correlation which decreases to zero exponentially with speed regulated by γ. Informally, $P^{(\gamma)}$ interpolates between the diffusive Dirichlet process, obtained as an appropriate limit of $P^{(\gamma)}$ as $\gamma \to 0$, and a purely discontinuous model $P^{(\infty)}$, whereby in every finite interval infinitely many atoms are updated, so that for every $s > 0$, $P_t^{(\infty)}$ and $P_{t+s}^{(\infty)}$ are uncorrelated. This adds great flexibility to the dependence structure, which can be learned from the data by implementing $P^{(\gamma)}$ in a hierarchical model, similar to (13), augmented with a prior distribution on γ. However, besides the clear advantages related to the correlation which is no longer constrained, the model sample paths are no longer continuous but only continuous in probability, and the Markov property is retained only with respect to the filtration generated by $(w(\cdot), x(\cdot))$ and not with respect to the natural filtration. A further alternative, which retains the path continuity, is the possibility of allowing also the atoms to diffuse. However, this way seems difficult if one is interested in proving minimal theoretical properties for the model.

Ackowledgements

The first author was partially supported by *Consejo Nacional de Ciencia y Tecnología de México* project 131179, during the elaboration of this work. The second author is supported by the European Research Council (ERC) through StG "N-BNP" 306406.

A Proofs

Proof of Proposition 3.2

Define

$$\Delta_\infty = \left\{ z \in [0,1]^\infty : \sum\nolimits_{i\geq 1} z_i = 1 \right\},$$

and let, for any fixed sequence $x = (x_1, x_2, \ldots) \in \mathbb{X}^\infty$, $\mathscr{P}_x$ be the set of purely atomic probability measures with support $x \in \mathbb{X}^\infty$. Denote by $\varphi_x : \Delta_\infty \to \mathscr{P}_x$ the transformation $\varphi_x(\phi(v)) = \sum_{i=1}^\infty w_i \delta_{x_i}$, where $v = (v_1, v_2, \ldots)$ and $\phi : [0,1]^{\mathbb{N}} \to \Delta_\infty$ is defined as

$$\phi_1(v) = v_1, \quad \phi_i(v) = v_i(1-v_1)\ldots(1-v_{i-1}),\ i > 1.$$

Note that the map ϕ is a bijection, with $v_i = w_i/(1 - \sum_{k=1}^{i-1} w_k)$, and that φ_x is continuous as a function of $w = \phi(v)$ in total variation norm, since for every $\varphi_x(w)$ and $\varepsilon > 0$ we can find a neighbourhood

$$U(w,\varepsilon) = \left\{ w' \in \Delta_\infty : \sum\nolimits_{i\geq 1} |w_i - w_i'| < \varepsilon \right\} \tag{A.23}$$

so that $w^* \in U(w,\varepsilon)$ implies $\varphi_x(w^*) \in U_{\mathrm{TV}}(\varphi_x(w),\varepsilon)$, with

$$U_{\mathrm{TV}}(\varphi_x(w),\varepsilon) = \left\{ \varphi_x(w') \in \mathscr{P}_x : \ d_{\mathrm{TV}}(\varphi_x(w),\varphi_x(w')) < \varepsilon \right\}, \tag{A.24}$$

since

$$d_{\mathrm{TV}}(\varphi_x(w),\varphi_x(w')) = \sup_{A\in\mathcal{B}(\mathbb{X})} \left| \sum_{i\geq 1} w_i \delta_{x_i}(A) - \sum_{i\geq 1} w_i' \delta_{x_i}(A) \right| \leq \sum_{i\geq 1} |w_i - w_i'| < \varepsilon.$$

Furthermore, $\varphi_x(w)$ is invertible in w, with continuous inverse

$$\varphi_x^{-1}(P) = (P(\{x_1\}), P(\{x_2\}), \ldots) = w. \tag{A.25}$$

This implies that we can define a Feller semigroup $\{T_x(t)\}_{t\geq 0}$ on $C(\mathscr{P}_x)$ by means of $T_x(t)\psi = [S(t)(\psi\circ\varphi_x)]\circ\varphi_x^{-1}$, where $\{S(t)\}_{t\geq 0}$ is the Feller semigroup on $C(\Delta_\infty)$ corresponding to the

process $w(\cdot)$ and φ_x^{-1} is as in (A.25). Theorem 4.2.7 of Ethier and Kurtz (1986) now implies that for every probability measure ν on $\mathscr{P}_x$, there exists a Markov process P corresponding to $\{T_x(t)\}_{t\geq 0}$ with initial distribution ν and sample paths in $D_{\mathscr{P}_x}([0,\infty))$, the space of right-continuous functions from $[0,\infty)$ to $\mathscr{P}_x$ with left limits, equipped with the Skorohod topology. See Billingsley (1968), Chapter 3, for details. Moreover, being a continuous bijection with continuous inverse, for any fixed x, φ_x is a homeomorphism of Δ_∞ into $\mathscr{P}_x$, from which $\mathscr{P}_x$ is locally compact and separable. Denote now with $p_1(t,P,\mathrm{d}P')$ and $p_2(t,w,\mathrm{d}w')$ the transition functions corresponding to the semigroups $\{T_x(t)\}_{t\geq 0}$ and $\{S(t)\}_{t\geq 0}$ respectively, and define $U(w,\varepsilon)$ as in (A.23) and $U_{\mathrm{TV}}(P,\varepsilon)$ as in (A.24). Then for every $P\in\mathscr{P}_x$ and $\varepsilon>0$, we have

$$t^{-1}p_1(t,P,U_{\mathrm{TV}}(P,\varepsilon)^c)=t^{-1}p_2(t,w,U(w,\varepsilon)^c)\to 0 \qquad \text{as } t\to 0, \tag{A.26}$$

where the identity follows form the fact that the two events are determined by the same subset of elementary events, and the right hand side of (A.26) follows from the continuity of the trajectories of $w(\cdot)$. The result now follows from Ethier and Kurtz (1986), Lemma 4.2.9.

Proof of Proposition 3.3

Since $w(\cdot)$ is reversible (Feng and Wang, 2007), with each component $v_i(\cdot)$ reversible with respect to a Beta$(1,\theta)$ distribution, and the atoms $x_i(t)\equiv x_i$ are trivially reversible and independent of $w(\cdot)$, it follows that P is reversible. The full statement now follows by the fact that $x_i\sim^{iid}G$, and by assuming the initial distribution $v_i(0)\sim\text{Beta}(1,\theta)$ for all $i\geq 1$.

Proof of Proposition 3.4

We have

$$\begin{aligned}\mathbb{E}(P_t(A)P_{t+s}(A)) &= \mathbb{E}\Big(\sum_{i\geq 1} w_i(t)\delta_{x_i}(A)\sum_{j\geq 1} w_j(t+s)\delta_{x_j}(A)\Big)\\ &= \mathbb{E}\Big(\sum_{i\geq 1} w_i(t)w_i(t+s)\delta_{x_i}(A)\\ &\qquad + \sum_{i\geq 1}\sum_{j\neq i\geq 1} w_i(t)w_j(t+s)\delta_{x_i}(A)\delta_{x_j}(A)\Big)\\ &= k_s G(A)+(1-k_s)G^2(A)\end{aligned}$$

where

$$k_s = \mathbb{E}\Big(\sum_{i\geq 1} w_i(t)w_i(t+s)\Big) = \sum_{i\geq 1}\mathbb{E}(w_i(t)w_i(t+s)). \tag{A.27}$$

Here k_s is independent of t by stationarity and $1 - k_s$ is obtained by subtraction, since

$$1 = \sum_{i\geq 1} w_i(t) \sum_{j\geq 1} w_j(t+s) = \sum_{i\geq 1} w_i(t)w_i(t+s) + \sum_{i\geq 1}\sum_{j\neq i\geq 1} w_i(t)w_j(t+s),$$

from which

$$\mathrm{Cov}(P_t(A), P_{t+s}(A)) = k_s G(A)(1 - G(A)), \tag{A.28}$$

and, using (10),

$$\mathrm{Corr}(P_t(A), P_{t+s}(A)) = k_s(1+\theta).$$

Now, from (8) and using independence among the $v_i(\cdot)$'s, we have

$$\begin{aligned}
\mathbb{E}(w_i(t)w_i(t+s)) &= \mathbb{E}\left[\left(v_i(t)\prod_{j<i}(1-v_j(t))\right)\left(v_i(t+s)\prod_{j<i}(1-v_j(t+s))\right)\right] \\
&= \mathbb{E}[v_i(t)v_i(t+s)]\prod_{j<i}\mathbb{E}\Big((1-v_j(t))(1-v_j(t+s))\Big) \\
&= \mathbb{E}[v_i(t)v_i(t+s)]\prod_{j<i}\left(1 - \frac{2}{1+\theta} + \mathbb{E}[v_j(t)v_j(t+s)]\right).
\end{aligned}$$

Since $\mathrm{Corr}[v_i(t), v_i(t+s)] = e^{-\lambda s}$ with $\lambda = (1+\theta)/2$ (cf. Bibby, Skovgaard and Sørensen, 2005), it follows that $E[v_i(t)v_i(t+s)] = (1+\theta)^{-2}[1 + (2+\theta)^{-1}\theta e^{-(1+\theta)s/2}]$, so

$$\begin{aligned}
k_s &= \left(c_1 + c_2 e^{-(1+\theta)s/2}\right)\sum_{i\geq 1}\left(c_1\theta^2 + c_2 e^{-(1+\theta)s/2}\right)^{i-1}, \\
c_1 &= \frac{1}{(1+\theta)^2}, \qquad c_2 = \frac{\theta}{(1+\theta)^2(2+\theta)}
\end{aligned}$$

from which the statement follows by direct computation. Note that by exchanging the limit operation with the sum and the integral (of positive terms) in $\lim_{s\to 0} k_s$, we obtain $\lim_{s\to 0} k_s = (1+\theta)^{-1}$, hence $\lim_{s\to 0}\mathrm{Corr}(P_t(A), P_{t+s}(A)) = 1$.

B Algorithm details

We illustrate the complementary details of the summary of the simulation-based procedure, provided in Section 4, for estimating the time-varying density which generates the data. We will refer to quantities there introduced whenever this does not compromise the readability. Here we provide the general algorithm based on a GEM diffusive mixture as discussed in

Section 6. For ease of the reader, the simplifications implied by choosing a Dirichlet diffusive mixture are made explicit in the last part of the present section.

To recall briefly the relevant notation, let $y^{(n)} = (y_{t_1}, \ldots, y_{t_n})$ be the data points observed at times $(t_1, \ldots, t_n)$, $v(\cdot) = ((v_1(t), v_2(t), \ldots), t \geq 0)$ with $v_j(\cdot) \sim^{ind} \mathrm{WF}(a_j, b_j)$ as in (7), $x = (x_1, x_2, \ldots)$ with $x_j \sim^{iid} G$, for $G \in \mathscr{P}(\mathbb{X})$ nonatomic. So for example π_v in (19) becomes a Beta(a_j, b_j). Note that the (a_j, b_j) parameters must be chosen such that $\sum_{j\geq 1} w_j(0) = 1$, with $w_j(t)$ as in (8) (the Wright–Fisher dynamics would then imply the same holds for all $t \geq 0$). See, for example, Ishwaran and James (2001) for sufficient conditions. Given the discussion in Section 4, it remains to make explicit how to update the random measures locations and weights, the slice and membership variables, and how to use the slice sampling on the Wright–Fisher transition density. We treat these issues separately.

Updating the locations

Since the locations are not time dependent, these are updated as in Kalli, Griffin and Walker (2011). That is

$$\mathcal{L}(x_j \mid \ldots) \propto G(x_j) \prod_{\{i:s_i=j\}} K(y_{t_i} \mid x_j) \tag{A.29}$$

for $j = 1, \ldots, m$, so that only a finite number of locations need to be sampled.

Updating the weights

We need the full conditional distributions for each of the $m \times n$ Wright–Fisher values $v_j(t_i)$, where $j = 1, \ldots, m$, m is as in (17) and $i = 1, \ldots, n$. Hence, for each $j = 1, \ldots, m$ we have

(A.30)

$$\mathcal{L}(v_j(t_1) \mid \ldots) \propto p_v(v_j(t_2) \mid v_j(t_1))\, \pi_v(v_j(t_1))\, v_j(t_1)^{\mathbb{I}(s_1=j)}\, (1 - v_j(t_1))^{\mathbb{I}(s_1>j)},$$

$$\mathcal{L}(v_j(t_i) \mid \ldots) \propto p_v(v_j(t_{i+1}) \mid v_j(t_i))\, p_v(v_j(t_i) \mid v_j(t_{i-1}))\, v_j(t_i)^{\mathbb{I}(s_i=j)}\, (1 - v_j(t_i))^{\mathbb{I}(s_i>j)}, \quad i \neq 1, n,$$

$$\mathcal{L}(v_j(t_n) \mid \ldots) \propto p_v(v_j(t_n) \mid v_j(t_{n-1}))\, v_j(t_n)^{\mathbb{I}(s_n=j)}\, (1 - v_j(t_n))^{\mathbb{I}(s_n>j)}.$$

Note that dropping the dependence on time in the weights processes $w_j(t)$ would yield

$$\begin{aligned} \mathcal{L}(v_j \mid \ldots) &\propto \pi_v(v_j) \textstyle\prod_{\{i:s_i=j\}} \left[v_j \prod_{k<j} (1 - v_k) \right] \\ &\propto \pi_v(v_j)\, v_j^{\sum_{i=1}^n \mathbb{I}(s_i=j)}\, (1 - v_j)^{\sum_{i=1}^n \mathbb{I}(s_i>j)}. \end{aligned}$$

Letting $\pi_{v_j} = \text{Beta}(a_j, b_j)$, the previous simplifies to

$$\mathcal{L}(v_j \mid \ldots) = \text{Beta}(a_j + \sum_{i=1}^{n} \mathbb{I}(s_i = j), b_j + \sum_{i=1}^{n} \mathbb{I}(s_i > j)),$$

which is the usual posterior update in the framework of stick-breaking random probability measures based on Beta distributed stick-breaking component. See Kalli, Griffin and Walker (2011).

Updating the slice and membership variables

For each $i = 1, \ldots, n$ we have

$$\mathcal{L}(u_{t_i} \mid \cdots) = U(0, \psi_{s_{t_i}})$$

and

$$\mathcal{L}(s_{t_i} \mid \ldots) \propto \frac{w_{s_{t_i}}(t_i)}{\psi_{s_{t_i}}} K(y_{t_i} \mid x_{s_{t_i}}(t_i)) \, \mathbb{I}(s_{t_i} \in \{k : \psi_{s_{t_i}} > u_{t_i}\}).$$

Note that since $\{k : \psi_{s_{t_i}} > u_{t_i}\}$ is a finite set, the above distribution involves a finite sampling, namely from $s_{t_i} = 1, \ldots, \lfloor \psi^*(u_{t_i}) \rfloor$, where ψ^* denotes the inverse of ψ.

Slicing the Wright–Fisher transition density

The analytic form for the transition density of the Wright–Fisher diffusion model does not take a simple closed form. One could attempt to use the corresponding spectral representation, which involves an infinite series of orthogonal Jacobi polynomials. However such infinite number of arguments with alternating sign prevents a robust evaluation of the transition density, especially for the extreme points of the state space. Instead, here we opt to use a slightly more general representation given in Mena and Walker (2009). That is, an equivalent formulation of the transition density of the Wright–Fisher diffusion is

$$p_v(v(t) \mid v(0)) = \sum_{m=0}^{\infty} r_t(m) \, D(v(t) \mid m, v(0)) \tag{A.31}$$

where

$$r_t(m) = \frac{(a+b)_m e^{-mct}}{m!} (1 - e^{-ct})^{a+b} \tag{A.32}$$

and

$$D(v(t) \mid m, v(0)) = \sum_{k=0}^{m} \text{Beta}(v(t) | a + k, b + m - k) \text{Bin}(k | m, v(0)). \tag{A.33}$$

Here $\text{Beta}(\cdot \mid a,b)$ is the Beta density with parameters a,b and $\text{Bin}(\cdot \mid m,q)$ is the Binomial probability function with m trials and success probability q. A reparametrization of (7) given by letting $c=(a+b-1)/2$ leads to writing

$$\mathrm{d}v(t)=\left(\frac{c(a-(a+b)v(t))}{a+b-1}\right)\mathrm{d}t+\left(\frac{2c}{a+b-1}v(t)(1-v(t))\right)^{1/2}\mathrm{d}B(t).$$

The above representation is valid for $a+b>1$, in which case 0 and 1 are entrance boundaries. Such condition rules out the inconvenient case of weights $w_j(t)$ in (8) become 0 or 1. A byproduct of the above re-parametrization is that it makes explicit the rate of decay in the autocorrelation function of the Wright–Fisher diffusion, which for the Dirichlet process case (8) reduces to $\text{Corr}[v_j(t),v_j(t+s)]=e^{-(1+\theta)s/2}$. See for example Bibby, Skovgaard and Sørensen (2005).

The representation (A.31) is appealing not only due to the fact that it involves elementary functions, but also and foremost since, unlike in the spectral decomposition, the summands are all positive. It follows that truncations, or rather random truncations such as those invoked by the slice method, are feasible. Indeed, with techniques similar to those used for (15) we can augment the transition density in order to avoid the infinite computation. Introduce then (o_j,k_j,d_j) such that

$$\begin{aligned}&p_t^v(v_j(t),o_j,k_j,d_j\mid v_j(0))\\&=\mathbb{I}(o_j<g(d_j))\frac{r_{j,t}(d_j)}{g(d_j)}\text{Beta}(v_j(t)|a_j+k_j,b_j+d_j-k_j)\text{Bin}(k_j|d_j,v_j(0)),\end{aligned}$$

where as before $d\mapsto g(d)$ is a decreasing function with known inverse g^* and $r_{j,t}(\cdot)$ denotes (A.32) computed on parameters (a_j,b_j,c_j). Augmenting for n observations by means on $(o_j^i,k_j^i,d_j^i)_{i=2}^n$ leads to the likelihood for m processes

$$\begin{aligned}&\mathcal{L}(v_{1:m}^{(n)},o_{1:m}^{(n)},k_{1:m}^{(n)},d_{1:m}^{(n)}\mid a_j,b_j,c_j)\\&\quad=\prod_{j=1}^m\pi_v(v_j(t_1))\prod_{i=2}^n\mathbb{I}(o_j^i<g(d_j^i))\frac{r_{j,\tau_i}(d_j^i)}{g(d_j^i)}\\&\qquad\times\text{Beta}(v_j(t_i)|a_j+k_j^i,b_j+d_j^i-k_j^i)\text{Bin}(k_j^i|d_j^i,v_j(t_{i-1})),\end{aligned}$$

where $\tau_i=t_i-t_{i-1}$, and the the subscript "$1:m$" is interpreted as in (18). Therefore, given a prior $\pi(a_j,b_j,c_j)$, the posterior distribution for (a_j,b_j,c_j) is

$$\mathcal{L}(a_j,b_j,c_j\mid\ldots)\propto\mathcal{L}(v_{1:m}^{(n)},o_{1:m}^{(n)},k_{1:m}^{(n)},d_{1:m}^{(n)}\mid a_j,b_j,c_j)\,\pi(a_j,b_j,c_j). \tag{A.34}$$

Furthermore, the full conditionals for the latent variables $(o_j^i, k_j^i, d_j^i)_{i=2}^n$ for each $j = 1, \dots, m$ are given by $\mathcal{L}(o_j^i \mid \cdots) = U(o_j^i \mid 0, g(d_j^i))$,

$$\mathcal{L}(k_j^i \mid \cdots) \propto \binom{d_j^i}{k_j^i} \frac{\mathbb{I}(k_j^i \in \{0, \dots, d_j^i\})}{\Gamma(a_j + k_j^i)\Gamma(b_j + d_j^i - k_j^i)} \left\{ \frac{v_j(t_i)v_j(t_{i-1})}{(1 - v_j(t_i))(1 - v_j(t_{i-1}))} \right\}^{k_j^i}$$

and

$$\mathcal{L}(d_j^i \mid \cdots) \propto \frac{\Gamma(a_j + b_j + d_j^i)^2 \left[(1 - v_j(t_i))(1 - v_j(t_{i-1}))\right]^{d_j^i}}{e^{d_j^i c_j \tau_i} \Gamma(b_j + d_j^i - k_j^i)\Gamma(d_j^i - k_j^i + 1) g(d_j^i)} \mathbb{I}(k_j^i \le d_j^i \le g^*(o_j^i))$$

The supports of $\mathcal{L}(k_j^i \mid \cdots)$ and $\mathcal{L}(d_j^i \mid \cdots)$ are discrete and bounded, so sampling from such distributions is straightforward, for example via the inverse cumulative distribution function method.

Gibbs sampler for diffusive DP mixtures

When instead of a diffusive GEM mixture one chooses the special case of a diffusive Dirichlet process mixture, this corresponds to letting $a_j = 1$, $b_j = \theta$ and $c_j = c$ for all $j = 1, 2, \dots$ in the above derivation, or equivalently to take independent and identically distributed Wright–Fisher processes as in (8). With these specifications the full conditionals (A.30) for the weights processes simplify to

$$\begin{aligned}
\mathcal{L}(v_j(t_1) \mid \cdots) &= \mathrm{Beta}(v_j(t_1) \mid 1 + k_j^2 + \mathbb{I}(s_1 = j); \theta + d_j^2 - k_j^2 + \mathbb{I}(s_1 > j)), \\
\mathcal{L}(v_j(t_i) \mid \cdots) &= \mathrm{Beta}\Big(v_j(t_i) \mid 1 + k_j^i + k_j^{i+1} + \mathbb{I}(s_i = j); \\
&\qquad\qquad \theta + d_j^i + d_j^{i+1} - k_j^i - k_j^{i+1} + \mathbb{I}(s_i > j)\Big), \\
&\qquad i = 2, \dots, n-1, \\
\mathcal{L}(v_j(t_n) \mid \cdots) &= \mathrm{Beta}(v_j(t_n) \mid 1 + k_j^n + \mathbb{I}(s_n = j); \theta + d_j^n - k_j^n + \mathbb{I}(s_n > j)).
\end{aligned}$$

Assuming independent priors for θ and c, and taking the logarithm, the full conditionals (A.34) become

$$\begin{aligned}
\log \mathcal{L}(\theta \mid \cdots) \propto & \log(\pi(\theta)) - m(n-2)\log\Gamma(1+\theta) - m\log\Gamma(\theta) + m\theta \sum_{i=2}^{n} \log(1 - e^{-c\tau_i}) \\
& + \theta \sum_{j=1}^{m}\sum_{i=1}^{n} \log(1 - v_j(t_i)) + 2\sum_{j=1}^{m}\sum_{i=2}^{n} \log\Gamma(1 + \theta + d_j^i) \\
& - \sum_{j=1}^{m}\sum_{i=2}^{n} \log\Gamma(\theta + d_j^i + k_j^i)
\end{aligned}$$

and

$$\log \mathcal{L}(c \mid \cdots) \propto \log(\pi(c)) + m(1+\theta) \sum_{i=2}^{n} \log(1 - e^{-c\tau_i}) - c \sum_{j=1}^{m} \sum_{i=2}^{n} d_j^i \tau_i.$$

These can be sampled using the Adaptive Rejection Metropolis Sampling (ARMS) algorithm.

References

Barrientos, A.F., Jara, A. and Quintana, F.A. (2012). On the Support of MacEachern's Dependent Dirichlet Processes and Extensions. *Bayes. Anal.* **7**, 277–310.

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

Billingsley, P. (1968). *Convergence of probability measures*. Wiley, New York.

Caron, F., Davy. M. and Doucet, A. (2007) Generalized Polya urn for time-varying Dirichlet process mixtures. *Proc. 23rd Conf. on Uncertainty in Artificial Intelligence*, Vancouver.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2008). Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Trans. Sig. Proc.* **56**, 71–84.

Cifarelli, D.M. and Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability: the use of associative means. (Original title: "Problemi statistici non parametrici in condizioni di scambiabilità parziale: impiego di medie associative"). *Quaderni dell'Istituto di Matematica Finanziaria, Univ. of Torino*, **3**(12).

Damien, P., Wakefield, J.C. and Walker, S.G. (1999). Gibbs sampling for Bayesian nonconjugate and hierarchical models using auxiliary variables. *J. Roy. Statist. Soc. Ser. B* **61**, 331–344.

De Iorio, M., Müller, P., Rosner, G. and MacEachern, S.N.(2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99**, 205–215.

Duan, J.A., Guindani, M. and Gelfand, A.E. (2007). Generalized spatial Dirichlet process models. *Biometrika* **94**, 809–825.

Dunson, D.B. (2006). Bayesian dynamic modelling of latent trait distributions. *Biostatistics* 7, 551–568.

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

Dunson, D.B., Pillai, N. and Park, J-H. (2007). Bayesian density regression. *J. Roy. Statist. Soc. Ser. B* **69**, 163–183.

Dunson, D.B., Xue, Y. and Carin, L. (2008). The matrix stick-breaking process: flexible Bayes meta analysis. *J. Amer. Statist. Assoc.* **103**, 317–327.

Ethier, S.N. and Griffiths, R.C. (1993). The transition function of a Fleming–Viot process. *Ann. Probab.* **21**, 1571–1590.

Ethier, S.N. and Kurtz, T.G. (1981). The infinitely-many-neutral-alleles diffusion model. *Adv. Appl. Probab.* **13**, 429–452.

Ethier, S.N. and Kurtz, T.G. (1986). *Markov processes: characterization and convergence*. Wiley, New York.

Ethier, S.N. and Kurtz, T.G. (1993). Fleming–Viot processes in population genetics. *SIAM J. Control Optim.* **31**, 345–386.

Favaro, S., Ruggiero, M. and Walker, S.G. (2009). On a Gibbs sampler based random process in Bayesian nonparametrics. *Electron. J. Statist.* **3**, 1556–1566.

Feng, S. and Wang, F.Y. (2007). A class of infinite-dimensional diffusion processes with connection to population genetics. *J. Appl. Probab.* **44**, 938–949.

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

Fuentes-García, R. , Mena, R. H. and Walker, S.G. (2009). A nonparametric dependent process for Bayesian regression. *Statist. Probab. Lett.* **79**, 1112–1119.

Gelfand, A., Kottas, A. and MacEachern, S. (2005). Bayesian nonparametric spatial modelling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100**, 1021–1035.

Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian nonparametrics*. Springer-Verlag, New York.

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

Griffin, J.E. and Steel, M.F.J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 179–194.

Griffin, J.E. and Steel, M.F.J. (2010). Stick-Breaking Autoregressive Processes. *J. Econometrics* **162**, 383–396.

Hjort, N.L., Holmes, C.C., Mller, P. and Walker, S.G., eds. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge Univ. Press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Ass.* **96**, 161–173.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1997). *Discrete multivariate distributions*. John Wiley & Sons, New York.

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

Karlin, S. and Taylor, H.M. (1981). *A second course in stochastic processes*. Academic Press, New York.

Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modelling with normalised inverse-Gaussian priors. *J. Amer. Statist. Assoc.* **472**, 1278–1291.

Lijoi, A., Mena, R.H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 715–740.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N. L. Hjort, C. C. Holmes, P. Mller and S. G. Walker, eds.) 80–136. Cambridge Univ. Press, Cambridge.

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

MacEachern, S.N. (1999). Dependent Nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statist. Assoc., Alexandria, VA.

MacEachern, S.N. (2000). Dependent Dirichlet processes. *Tech. Rep.*, Ohio State University.

Mena, R.H., Ruggiero, M. and Walker, S.G. (2011). Geometric stick-breaking processes for continuous-time Bayesian nonparametric modelling. *J. Statist. Plann. Inf.* **141**, 3217–3230.

Mena, R.H. and Walker, S.G. (2009). On a construction of Markov models in continuous time. *Metron* **67**, 303–323.

Papaspiliopoulos, O. and Roberts, G.O. (2008). Retrospective MCMC for dirichlet process hierarchical models. *Biometrika* **95**, 169–186.

Petrone, S., Guindani, M. and Gelfand, A.E. (2009). Hybrid Dirichlet mixture models for functional data. *J. Roy. Statist. Soc. Ser. B* **71**, 755–782.

Petrov, L. (2009). Two-parameter family of diffusion processes in the Kingman simplex. *Funct. Anal. Appl.* **43**, 279–296.

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

Pitman, J. and Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.

Raftery, A. and Lewis, S. (1992). One long run with diagnostics: Implementation strate-

gies for Markov chain Monte Carlo. *Statist. Inference* **7**, 493–497.

Rodriguez, A. and Dunson, D.B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayes. Anal.* **6**, 145–178.

Rodriguez, A. and Ter Horst, E. (2008). Bayesian dynamic density estimation. *Bayes. Anal.* **3**, 339–366.

Ruggiero, M. and Walker, S.G. (2009). Countable representation for infinite-dimensional diffusions derived from the two-parameter Poisson–Dirichlet process. *Electron. Comm. Probab.* **14**, 501–517.

Ruggiero, M., Walker, S.G. and Favaro, S. (2013). Alpha-diversity processes and normalised inverse-Gaussian diffusions. *Ann. Appl. Probab.* **23**, 386–425.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statist. Sinica* **2**, 639–650.

Trippa, L., Müller, P. and Johnson, W. (2011). The multivariate beta process and an extension of the Polya tree model. *Biometrika* **98**, 17–34.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Sim. Comput.* **36**, 45–54.