Bayesian Analysis (2021) **16**, Number 2, pp. 371–395

Predictive inference with Fleming–Viot-driven dependent Dirichlet processes

Filippo Ascolani[*], Antonio Lijoi[†,§] and Matteo Ruggiero[‡]

Abstract. We consider predictive inference using a class of temporally dependent Dirichlet processes driven by Fleming–Viot diffusions, which have a natural bearing in Bayesian nonparametrics and lend the resulting family of random probability measures to analytical posterior analysis. Formulating the implied statistical model as a hidden Markov model, we fully describe the predictive distribution induced by these Fleming–Viot-driven dependent Dirichlet processes, for a sequence of observations collected at a certain time given another set of draws collected at several previous times. This is identified as a mixture of Pólya urns, whereby the observations can be values from the baseline distribution or copies of previous draws collected at the same time as in the usual Pólya urn, or can be sampled from a random subset of the data collected at previous times. We characterize the time-dependent weights of the mixture which select such subsets and discuss the asymptotic regimes. We describe the induced partition by means of a Chinese restaurant process metaphor with a *conveyor belt*, whereby new customers who do not sit at an occupied table open a new table by picking a dish either from the baseline distribution or from a time-varying offer available on the conveyor belt. We lay out explicit algorithms for exact and approximate posterior sampling of both observations and partitions, and illustrate our results on predictive problems with synthetic and real data.

MSC2020 subject classifications: Primary 62F15; secondary 62G25, 62M20.

Keywords: Chinese restaurant, conveyor belt, random partition, hidden Markov model, generalized Pólya urn, predictive distribution.

1 Introduction and summary of results

Bayesian nonparametric methodology has undergone a tremendous development in the last decades, often standing out among competitors for flexibility, interpretability and computational convenience. See, for example, Hjort et al. (2010); Müller et al. (2015); Ghosal and van der Vaart (2017). The cornerstone of Bayesian nonparametrics is the sampling model based on the Dirichlet process (Ferguson, 1973), whereby

$$Y_i \mid X = x \overset{\text{iid}}{\sim} x, \qquad X \sim \Pi_\alpha. \tag{1.1}$$

[*]Department of Decision Sciences, Bocconi University, via Röntgen 1, 20136 Milano, Italy, filippo.ascolani@phd.unibocconi.it

[†]Department of Decision Sciences, Bocconi University, via Röntgen 1, 20136 Milano, Italy, antonio.lijoi@unibocconi.it

[‡]Collegio Carlo Alberto and ESOMAS Department, University of Torino, C.so Unione Sovietica 218/bis, 10134, Torino, Italy, matteo.ruggiero@unito.it

[§]Also affiliated to Collegio Carlo Alberto, Torino and BIDSA, Bocconi University, Milano, Italy.

 https://doi.org/10.1214/20-BA1206

Here, given a sampling space $\mathcal{Y}$, we use X to denote a random probability measure (RPM) on $\mathcal{Y}$, and observations Y_i are assumed to be independent with distribution x when $X = x$. We also denote by Π_α the distribution X induces on the space $\mathcal{P}(\mathcal{Y})$ of probability measures on $\mathcal{Y}$, with $\alpha = \theta P_0$, $\theta > 0$ and P_0 a probability measure on $\mathcal{Y}$. Notable properties of the Dirichlet process are its large weak support and conjugacy, whereby the conditional RPM X, given observations $Y_1, \ldots, Y_n$ from (1.1), is still a Dirichlet process with updated parameter $\alpha + \sum_{i=1}^n \delta_{Y_i}$.

The great appeal offered by the relative simplicity of the Dirichlet process boosted a number of extensions, among which some of the most successful are mixtures of Dirichlet processes (Antoniak, 1974), Dirichlet process mixtures (Lo, 1984), Pólya trees (Mauldin et al., 1992; Lavine, 1992), Pitman–Yor processes (Perman et al., 1992; Pitman and Yor, 1997), Gibbs-type random measures (Gnedin and Pitman, 2005; De Blasi et al., 2015), normalised random measures with independent increments (Regazzini et al., 2003; Lijoi et al., 2005, 2007), to mention a few. The common thread linking all the above developments is the assumption of exchangeability of the data, equivalent to the conditional independence and identity in distribution in (1.1) by virtue of de Finetti's Theorem. This can be restrictive when modelling data that are known to be generated from partially inhomogeneous sources, as for example in time series modelling or when the data are collected in subpopulations. Such framework can be accommodated by *partial exchangeability*, a weaker type of dependence whereby observations in two or more groups of data are exchangeable within each group but not overall. If groups are identified by a covariate value $z \in \mathcal{Z}$, then observations are exchangeable only if their covariates have the same value.

One of the most active lines of research in Bayesian nonparametrics in recent years aims at extending the basic paradigm (1.1) to this more general framework. Besides pioneering contributions, recent progresses have stemmed from MacEachern (1999), who called a collection of RPMs $\{X_z, z \in \mathcal{Z}\}$ indexed by a finite-dimensional measurement $z \in \mathcal{Z}$ a *dependent Dirichlet process* (DDP) if each marginal measure X_z is a Dirichlet process with parameter that depends on z.

Here we focus on DDPs with temporal dependence, and replace z with $t \in [0, \infty)$ representing time. Previous contributions in this framework include Dunson (2006); Caron et al. (2007); Rodriguez and ter Horst (2008); Griffin and Steel (2010); Caron and Teh (2012); Mena and Ruggiero (2016); Caron et al. (2017); Gutierrez et al. (2016); Canale and Ruggiero (2016); Kon Kam King et al. (2020). Many proposals in this area start from the celebrated stick-breaking representation of the Dirichlet process (Sethuraman, 1994), whereby X in (1.1) is such that

$$X \stackrel{d}{=} \sum_{i \geq 0} V_i \prod_{j=1}^{i-1} (1 - V_j)\, \delta_{Y_i}, \qquad V_i \stackrel{\text{iid}}{\sim} \text{Beta}(1, \theta), \qquad Y_i \stackrel{\text{iid}}{\sim} P_0, \tag{1.2}$$

and the temporal dependence is induced by letting each V_i and/or Y_i depend on time in a way that preserves the marginal distributions. This approach has many advantages, among which: simplicity and versatility, since inducing dynamics on V_i or Y_i allows for a variety of solutions; flexibility, since under mild conditions the resulting processes have

large support (cf. Barrientos et al. 2012); ease of implementation, since strategies for posterior computation based on MCMC sampling are readily available. However, the stick-breaking structure makes the analytical derivation of further posterior information, like for example characterizing the predictive distribution of the observations, often a daunting task. This typically holds for other approaches to temporal Bayesian nonparametric modelling as well. Determining explicitly such quantities would not only give a deeper insight into the model posterior properties, which otherwise remain obscure to a large extent, but also provide a further tool for direct application or as a building block in more involved dependent models, whose computational efficiency would benefit from an explicit computation.

In this paper, we provide analytical results related to the posterior predictive distribution of the observations induced by a class of temporal DDP models driven by Fleming–Viot processes. The latter are a class of diffusion processes whose marginal values are Dirichlet processes. The continuous time dependence is a distinctive feature of our proposal, compared to the bulk of literature in the area. In particular, here we complement previous work done in Papaspiliopoulos et al. (2016), which focussed on identifying the laws of the dependent RPMs involved, by investigating the distributional properties of future observations, characterized as a mixture of Pólya urn schemes, and those of the induced partitions.

More specifically, in Section 2 we detail the statistical model we adopt, which directly extends (1.1) by assuming a hidden Markov model structure whereby observations are conditionally *iid* given the marginal value of a Fleming–Viot-driven DDP. We recall some key properties of this model, and include a new result on the weak support of the induced prior. In Section 3 we present our main results. Conditioning on samples, with possibly different sizes, collected at p times $0 = t_0 < \cdots < t_{p-1} = T$, we characterize the predictive distribution of a further sequence drawn at time $T + t$. This task can be seen as a dynamic counterpart to obtaining the predictive distribution of $Y_{k+1}|Y_1, \ldots, Y_k$ for any $k \geq 1$ in (1.1), when the RPM X is integrated out, which yields the Pólya urn

$$\mathbb{P}(Y_{k+1} \in A|Y_1, \ldots, Y_k) = \frac{\theta}{\theta + k} P_0(A) + \frac{k}{\theta + k} P_k(A), \tag{1.3}$$

for any Borel set A of $\mathcal{Y}$, where P_k denotes the empirical distribution of $(Y_1, \ldots, Y_k)$. In the hidden Markov model framework, we identify the predictive distribution of observations from the DDP at time $T + t$ to be a time-dependent mixture of Pólya urn schemes. This can be thought of as being generated by a latent variable which selects a random subset of the data collected at previous times, whereby every component of the mixture is a classical posterior Pólya urn conditioned to a different subset of the past data. We characterize the mixture weights, where the temporal dependence arises, and derive an explicit expression for the correlation between observations at different time points. Furthermore, we discuss two asymptotic regimes of the predictive distribution – as the time index diverges, which recovers (1.3), and as the current sample size diverges, which links the sequence at time $T + t$ with its de Finetti measure – and lay out explicit algorithms for exact and approximate sampling from the predictive. Next, we discuss the induced partition at time $T + t$ and derive an algorithm for sampling from its distribution. The partition sampling process is interpreted as a *Chinese restaurant*

with conveyor belt, whereby arriving customers who do not sit at an already occupied table, open a new table by choosing a dish either from the baseline distribution P_0 or from a temporally dependent selection of dishes that run through the restaurant on a conveyor belt, which in turn depends on past dishes popularity. We defer all proofs to the Supplementary Material (Ascolani et al. 2020). Finally, Section 4 illustrates the use of our results for predictive inference through synthetic data and through a dataset on the Karnofsky score related to a Hodgkins lymphoma study.

2 Fleming–Viot dependent Dirichlet processes

We consider a class of dependent Dirichlet processes with continuous temporal covariate. Instead of inducing the temporal dependence through the building blocks of the stick-breaking representation (1.2), we let the dynamics of the dependent process be driven by a Fleming–Viot (FV) diffusion. FV processes have been extensively studied in relation to population genetics (see Ethier and Kurtz (1993) for a review), while their role in Bayesian nonparametrics was first pointed out in Walker et al. (2007) (see also Favaro et al., 2009). A loose but intuitive way of thinking a FV diffusion is of being composed by infinitely-many probability masses, associated to different locations in the sampling space $\mathcal{Y}$, each behaving like a diffusion in the interval $[0, 1]$, under the overall constraint that the masses sum up to 1. In addition, locations whose masses touch 0 are removed, while new locations are inserted at a rate which depends on a parameter $\theta > 0$. As a consequence, the random measures X_t and X_s, with $t \neq s$, will share some, though not all, their support points.

The transition function that characterizes a FV process admits the following natural interpretation in Bayesian nonparametrics (cf. Walker et al., 2007). Initiate the process at the RPM $X_0 \sim \Pi_\alpha$, and denote by D_t a time-indexed latent variable taking values in $\mathbb{Z}_+$. Conditional on $D_t = m \in \mathbb{Z}_+$, the value of the process at time t is a posterior DP X_t with law

$$X_t \mid (D_t = m, Y_1, \ldots, Y_m) \sim \Pi_{\alpha + \sum_{i=1}^m \delta_{Y_i}} \qquad Y_i \mid X_0 \overset{\text{iid}}{\sim} X_0. \tag{2.1}$$

Here, the realisation of the latent variable D_t determines how many atoms m are drawn from the initial state X_0, to become atoms of the posterior Dirichlet from which the arrival state is drawn. Such D_t is a pure-death process, which starts at infinity with probability one and jumps from state m to state $m-1$ after an exponentially distributed waiting time with inhomogenous parameter $\lambda_m = m(\theta + m - 1)/2$. The transition probabilities of D_t have been computed by Griffiths (1980); Tavaré (1984), and in particular

$$\mathbb{P}(D_t = m \mid D_0 = \infty) = d_m(t), \tag{2.2}$$

where

$$d_m(t) = \sum_{k=m}^{\infty} e^{-\lambda_k t} (-1)^{k-m} \frac{(\theta + 2k - 1)(\theta + m)_{(k-1)}}{m!(k-m)!},$$

where $\theta_{(k)} = \theta(\theta + 1) \cdots (\theta + k - 1)$ is the ascending factorial or Pochhammer symbol, with $\theta_{(0)} = 1$. Here the fact that $D_0 = \infty$ almost surely should be understood as an

entrance boundary, i.e., the process decreases from infinity at infinite speed so that at each $t > 0$ the value of D_t is finite. The unconditional transition of the FV process is thus obtained by integrating $D_t, Y_1, \ldots, Y_{D_t}$ out of (2.1), leading to

$$P_t(x, \mathrm{d}x') = \sum_{m=0}^{\infty} d_m(t) \int_{\mathcal{Y}^m} \Pi_{\alpha+\sum_{i=1}^m \delta_{y_i}}(\mathrm{d}x') x(\mathrm{d}y_1) \cdots x(\mathrm{d}y_m). \tag{2.3}$$

This was first found by Ethier and Griffiths (1993). It is known that Π_α is the invariant measure of P_t if $X_0 \sim \Pi_\alpha$, in which case all marginal RPMs X_t are Dirichlet processes with the same parameter. In particular, the death process D_t determines the correlation between RPMs at different times. Indeed, a larger t implies a lower m with higher probability, hence a decreasing (on average) number of support points will be shared by the random measures X_0 and X_t when t increases. On the contrary, as $t \to 0$ we have $D_t \to \infty$, which in turn implies infinitely-many atoms shared by X_0 and X_t, until the two RPMs eventually coincide. See Lijoi et al. (2016) for further discussion.

For definiteness, we formalize the following definition.

Definition 1. A Markov process $\{X_t\}_{t\geq 0}$ taking values in the space of atomic probability measures on $\mathcal{Y}$ is a *Fleming–Viot dependent Dirichlet process* with parameter α, denoted $X_t \sim \text{FV-DDP}(\alpha)$, if $X_0 \sim \Pi_\alpha$ and its transition function is (2.3).

Seeing a FV-DDP as a collection of RPMs, one is immediately led to wonder about the support properties of the induced prior. The weak support of a FV-DDP is the smallest closed set in $\mathcal{B}\{\mathcal{P}(\mathcal{Y})^{\mathbb{R}_+}\}$ with probability one, where $\mathcal{P}(\mathcal{Y})$ is the set of probability measures on $\mathcal{Y}$ and $\mathcal{B}\{\mathcal{P}(\mathcal{Y})^{\mathbb{R}_+}\}$ is the Borel σ-field generated by the product topology of weak convergence. Barrientos et al. (2012) investigated these aspects for a large class of DDPs based on the stick-breaking representation of the Dirichlet process. Since no such representation is known for the FV process, our case falls outside that class. The following proposition states that a FV-DDP has full weak support, relative to the support of P_0.

Proposition 1. *Let $\alpha = \theta P_0$ and $\mathcal{Y}$ be the support of P_0. Then the weak support of a* FV-DDP(α) *is given by $\mathcal{P}(\mathcal{Y})^{\mathbb{R}_+}$.*

In order to formalize the statistical setup, we cast the FV-DDP into a hidden Markov model framework. A hidden Markov model is a double sequence $\{(X_{t_n}, Y_{t_n}), n \geq 0\}$ where X_{t_n} is an unobserved Markov chain, called hidden or *latent signal*, and Y_{t_n} are conditionally independent observations given the signal. The signal can be thought of as the discrete-time sampling of a continuous time process, and is assumed to completely specify the distributions of the observations, called *emission distributions*. While the literature on hidden Markov models has mainly focussed on finite-dimensional signals, infinite-dimensional cases have been previously considered in Beal et al. (2002); Van Gael et al. (2008); Stepleton et al. (2009); Yau et al. (2011); Zhang et al. (2014); Papaspiliopoulos et al. (2016).

Here we take X_{t_n} to be a FV-DDP as in Definition 1, evaluated at p times $0 = t_0 < \cdots < t_{p-1} = T$. The sampling model is thus

$$Y^i_{t_n} \mid X_{t_n} = x \overset{\text{iid}}{\sim} x, \quad i = 1, \ldots, n_{t_n}, \qquad X_t \sim \text{FV-DDP}(\alpha), \tag{2.4}$$

where n_{t_n} is the number of observations collected at time t_n. It follows that any two variables $Y^i_{t_n}, Y^j_{t_m}$ are conditionally independent given X_{t_n} and X_{t_m}, with product distribution $X_{t_n} \times X_{t_m}$.

In addition, similarly to mixing a DP with respect to its parameter measure as in Antoniak (1974), one could also consider randomizing the parameter α in (2.4), e.g. by letting $\alpha = \alpha_\gamma$ and $\gamma \sim \pi$ on an appropriate space.

In the following, we will denote for brevity $\mathbf{Y}_n := \mathbf{Y}_{t_n}$ and $\mathbf{Y}_{0:T} := (\mathbf{Y}_0, \dots, \mathbf{Y}_T)$, where $\mathbf{Y}_i$ is the set of n_i observations collected at time t_i. We will sometimes refer to $\mathbf{Y}_{0:T}$ as the *past values*, since the inferential interest will be set at time $T + t$. We will also denote by $(y^*_1, \dots, y^*_K)$ the K distinct values in $\mathbf{Y}_{0:T}$, where $K \leq \sum_{i=0}^{T} n_i$. In this framework, Papaspiliopoulos et al. (2016) showed that the conditional distribution of the RPM X_T, given $\mathbf{Y}_{0:T}$, can be written as

$$\mathcal{L}(X_T|\mathbf{Y}_{0:T}) = \sum_{\mathbf{m}\in\mathbf{M}} w_{\mathbf{m}} \Pi_{\alpha+\sum_{j=1}^{K} m_j \delta_{y^*_j}}, \tag{2.5}$$

where the weights $w_{\mathbf{m}}$ can be computed recursively. In particular, $\mathbf{M}$ is a finite convex set of vector multiplicities $\mathbf{m} = (m_1, \dots, m_K) \in \mathbb{Z}^K_+$ determined by $\mathbf{Y}_{0:T}$, which identify the mixture components in (2.5) with strictly positive weight. We will call $\mathbf{M}$ the set of *currently active indices*. In particular, $\mathbf{M}$ is given by the points that lie between the counts of $(y^*_1, \dots, y^*_K)$ in $\mathbf{Y}_T$, which is the bottom node in a K-dimensional graph in Z^K_+, and the counts of $(y^*_1, \dots, y^*_K)$ in $\mathbf{Y}_{0:T}$, which is the top node. For example, if $T = 1$ suppose we observe $\mathbf{Y}_0 = (y^*_1, y^*_2)$ for some values $y^*_1 \neq y^*_2$ and $\mathbf{Y}_1 = \mathbf{Y}_0$, hence $K = 2$. Then the top node is $(2, 2)$ since in $\mathbf{Y}_{0:1}$ there are 2 of each of (y^*_1, y^*_2) and the bottom node is $(1, 1)$ which is the counts of (y^*_1, y^*_2) in $\mathbf{Y}_1$. Cf. Figure 1. Observations with $K = 3$ distinct values would instead generate a 3-dimensional graph, with the origin $(0, 0, 0)$ linked to 3 level-1 nodes $(1, 0, 0), (0, 1, 0), (0, 0, 1)$, and so on. In general, each upper level node is obtained by adding 1 to one of the lower node coordinates.

We note here that the presence of $d_m(t)$ in (2.3) makes the computations with FV processes in principle intractable, yielding in general infinite mixtures difficult to simulate from (cf. Jenkins and Spanò 2017). It is then remarkable that conditioning on past data one is able to obtain conditional distributions for the signal given by finite mixtures as in (2.5).

3 Predictive inference with FV-DDPs

3.1 Predictive distribution

In the above framework, we are primarily interested in predictive inference, which requires obtaining the predictive distribution of $Y^1_{T+t}, \dots, Y^k_{T+t}|\mathbf{Y}_{0:T}$, that is the marginal distribution of a k-sized sample drawn at time $T + t$, given data collected up to time T, when the random measures involved are integrated out. See Figure 2. Note that by virtue of the stationarity of the FV process, if $X_0 \sim \Pi_\alpha$, then $\mathbb{P}(Y_t \in A) = P_0(A)$ for any $t \geq 0$. Note also that if one mixes model (2.4) by randomizing the parameter

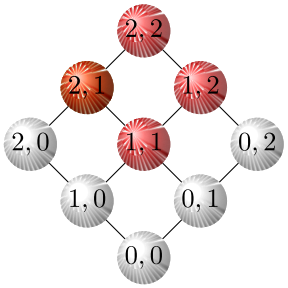

Figure 1: Red indices in the graph identify active mixture components at time T, i.e. the set $\mathbf{M}$ in (2.5), corresponding to points $\mathbf{m} \in \mathbb{Z}_+^K$ with positive weight. In this example $K = 2$, and the graph refers to $\mathbf{M}$ at time $T = 1$ if we observe $\mathbf{Y}_0 = (y_1^*, y_2^*) = \mathbf{Y}_1$.

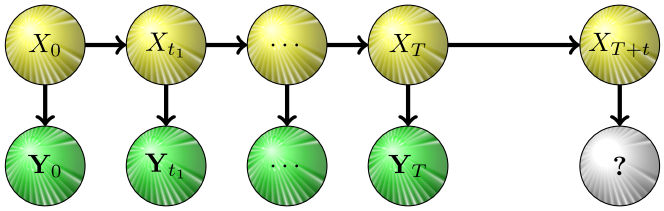

Figure 2: The predictive problem depicted as a graphical model. The upper yellow nodes are nonobserved states of the infinite-dimensional signal, the lower green nodes are conditionally independent observed data whose distribution is determined by the signal, the light gray node is the object of interest.

measure $\alpha = \alpha_\gamma$ as mentioned above, the evaluation of the predictive distributions is of paramount importance for posterior computation. Indeed, one needs the distribution of $\gamma|\mathbf{Y}_{0:T}$, and if for example γ has discrete support on $\mathbb{Z}_+$ with probabilities $\{p_j, j \in \mathbb{Z}_+\}$, then

$$\mathbb{P}(\gamma = j|\mathbf{Y}_{0:T}) \propto p_j \mathbb{P}(\mathbf{Y}_{0:T}|j) \propto p_j \mathbb{P}(\mathbf{Y}_0|j)\mathbb{P}(\mathbf{Y}_1|\mathbf{Y}_0, j) \cdots \mathbb{P}(\mathbf{Y}_T|\mathbf{Y}_{0:T-1}, j).$$

Denote for brevity $Y_{T+t}^{1:k} := (Y_{T+t}^1, \ldots, Y_{T+t}^k)$ the k values drawn at time $T + t$. For $\mathbf{m} \in \mathbb{Z}_+^K$, let $\{\mathbf{n} \in \mathbb{Z}_+^K : \mathbf{n} \leq \mathbf{m}\}$ be the set of nonnegative vectors such that $n_i \leq m_i$ for all i. Define also $|\mathbf{n}| := \sum_{j=1}^K n_i$, and

$$L(\mathbf{M}) := \{\mathbf{n} \in \mathbb{Z}_+^K : \mathbf{n} \leq \mathbf{m}, \mathbf{m} \in \mathbf{M}\} \tag{3.1}$$

to be all the points in $\mathbb{Z}_+^K$ lying below the top node of $\mathbf{M}$. E.g., if $\mathbf{M}$ is given by the red nodes in Figure 1, then $L(\mathbf{M})$ is given by all nodes shown in the figure.

Proposition 2. *Assume (2.4), and let the law of X_T given data $\mathbf{Y}_{0:T}$ be as in (2.5), where the weights $w_\mathbf{m}$ have been computed recursively. Then, for any Borel set A of $\mathcal{Y}$,*

the first observation at time $T+t$ has distribution

$$\mathbb{P}\left(Y_{T+t} \in A | \mathbf{Y}_{0:T}\right) = \sum_{\mathbf{n} \in L(\mathbf{M})} p_t(\mathbf{M}, \mathbf{n}) \left(\frac{\theta}{\theta + |\mathbf{n}|} P_0(A) + \frac{|\mathbf{n}|}{\theta + |\mathbf{n}|} P_{\mathbf{n}}(A) \right) \tag{3.2}$$

and the $(k+1)$st observation at time $T+t$, given the first k, has distribution

$$\begin{aligned} \mathbb{P}\left(Y_{T+t}^{k+1} \in A | \mathbf{Y}_{0:T}, Y_{T+t}^{1:k}\right) &= \sum_{\mathbf{n} \in L(\mathbf{M})} p_t^{(k)}(\mathbf{M}, \mathbf{n}) \\ &\times \left(\frac{\theta}{\theta + |\mathbf{n}| + k} P_0(A) + \frac{|\mathbf{n}|}{\theta + |\mathbf{n}| + k} P_{\mathbf{n}}(A) + \frac{k}{\theta + |\mathbf{n}| + k} P_k(A) \right) \end{aligned} \tag{3.3}$$

where

$$P_{\mathbf{n}} = \frac{1}{|\mathbf{n}|} \sum_{i=1}^{K} n_i \delta_{y_i^*}, \qquad P_k = \frac{1}{k} \sum_{j=1}^{k} \delta_{Y_{T+t}^j} \tag{3.4}$$

and $(y_1^*, \ldots, y_K^*)$ are the distinct values in $\mathbf{Y}_{0:T}$.

Before discussing the details of the above statement, a heuristic read of (3.2) is that the first observation at time $T+t$ is either a draw from the baseline distribution P_0, or a draw from a random subset of the past data points $\mathbf{Y}_{0:T}$, identified by the latent variable $\mathbf{n} \in L(\mathbf{M})$. Given how $L(\mathbf{M})$ is defined, Y_{T+t} can therefore be thought of as being drawn from a mixture of Pólya urns, each conditional on a different subset of the data, ranging from the full dataset to the empty set. Indeed, recall from Section 2 that the top node of $\mathbf{M}$, hence of $L(\mathbf{M})$ in (3.1), is the vector of multiplicities of the distinct values $(y_1^*, \ldots, y_K^*)$ contained in the entire dataset $\mathbf{Y}_{0:T}$. The probability weights associated to each lower node $\mathbf{n} \in L(\mathbf{M})$ are determined by a death process on $L(\mathbf{M})$, that differs from D_t in (2.2). In particular this is a Markov process that jumps from node $\mathbf{m}$ to node $\mathbf{m} - \mathbf{e}_i$ after an exponential amount of time with parameter $m_i(\theta + |\mathbf{m}| - 1)/2$, with $\mathbf{e}_i$ being the canonical vector in the ith direction. The weight associated with node $\mathbf{n} \in L(\mathbf{M})$ is then given by the probability that such death process is in $\mathbf{n}$ after time t, if started from any node in $\mathbf{M}$. For example, if $\mathbf{M}$ is as in Figure 1, than the weight of the node $(0, 2)$ is given by the probability that the death process is in $(0, 2)$ after time t if started from any other node of $\mathbf{M}$. Being a non increasing process, the admissible starting nodes are $(2, 2)$, $(1, 2)$ and $(0, 2)$ itself. Figure 3 highlights the first two of these paths.

The transition probabilities of this death process are

$$p_{\mathbf{m},\mathbf{n}}(t) = p_{|\mathbf{m}|,|\mathbf{n}|}(t) \mathrm{HG}(\mathbf{m} - \mathbf{n}; \mathbf{m}, |\mathbf{m} - \mathbf{n}|), \qquad \mathbf{0} \leq \mathbf{n} \leq \mathbf{m}, \tag{3.5}$$

where $\mathrm{HG}(\mathbf{i}; \mathbf{m}, |\mathbf{i}|)$ is the multivariate hypergeometric probability function evaluated at $\mathbf{i}$, namely

$$\mathrm{HG}(\mathbf{i}; \mathbf{m}, |\mathbf{i}|) = \frac{\binom{\mathbf{m}_1}{\mathbf{i}_1} \cdots \binom{\mathbf{m}_l}{\mathbf{i}_l}}{\binom{|\mathbf{m}|}{|\mathbf{i}|}}, \quad l = \dim(\mathbf{m})$$

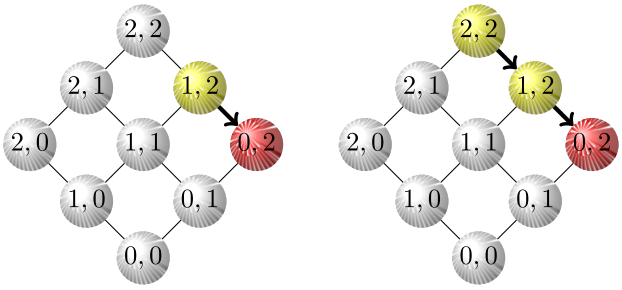

Figure 3: The weight associated to an index $\mathbf{n} \in L(\mathbf{M})$ at time $T+t$ is determined by the probability that the death process reaches $\mathbf{n}$ from any active index $\mathbf{m} \in \mathbf{M}$ at time T. For $\mathbf{M}$ as in Figure 1, the weight of the mixture component with index $\mathbf{n} = (0, 2)$, i.e., no atoms y_1^* and 2 atoms y_2^*, is the sum of the probabilities of reaching node $(0, 2)$ via the path starting from $(1, 2)$ (left) and from $(2, 2)$ (right).

with $\dim(\mathbf{m})$ denoting the dimension of vector $\mathbf{m}$, while $p_{|\mathbf{m}|,|\mathbf{n}|}(t)$ is the probability of descending from level $|\mathbf{m}|$ to $|\mathbf{n}|$ (see Lemma 1 in the Supplementary Material). Hence, in general, the probability of reaching node $\mathbf{n} \in L(\mathbf{M})$ from any node in $\mathbf{M}$ is

$$p_t(\mathbf{M}, \mathbf{n}) = \sum_{\mathbf{m} \in \mathbf{M}, \mathbf{m} \geq \mathbf{n}} w_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t). \tag{3.6}$$

In conclusion, with probability $p_t(\mathbf{M}, \mathbf{n})$ the first draw at time $T+t$ will be either from P_0, with probability $\theta/(\theta + |\mathbf{n}|)$, or a uniform sample from the subset of data identified by the multiplicity vector $\mathbf{n}$.

After each draw, the weights associated to each node need to be updated according to the likelihood that the observation was generated by the associated mixture component, similarly to what is done for mixtures of Dirichlet processes. Specifically,

$$p_t^{(k+1)}(\mathbf{M}, \mathbf{n}) \propto p_t^{(k)}(\mathbf{M}, \mathbf{n}) p(y_{T+t}^{k+1} \mid y_{T+t}^{1:k}, \mathbf{n}), \tag{3.7}$$

where

$$p(y_{T+t}^{k+1} \mid y_{T+t}^{1:k}, \mathbf{n}) := \frac{\theta p_0(y_{T+t}^{k+1}) + \sum_{i=1}^{K} n_i \delta_{y_i^*}(\{y_{T+t}^{k+1}\}) + \sum_{j=1}^{k} \delta_{y_{T+t}^j}(\{y_{T+t}^{k+1}\})}{\theta + |\mathbf{n}| + k} \tag{3.8}$$

is the predictive distribution of the $(k+1)$st observation given the previous k and conditional on $\mathbf{n}$, for P_0 discrete with density p_0. An analogous formula holds when P_0 is diffuse, that takes into account the different atoms in each component of the mixture.

Concerning the general case for the $(k+1)$st observation at time $T+t$, trivial manipulations of (3.3) provide different interpretative angles. Rearranging the term in brackets one obtains

$$\frac{\theta_{\mathbf{n}}}{\theta_{\mathbf{n}} + k} P_{0,\mathbf{n}} + \frac{k}{\theta_{\mathbf{n}} + k} P_k, \tag{3.9}$$

which bears a clear structural resemblance to (1.3). Here

$$\theta_{\mathbf{n}} = \theta + |\mathbf{n}|, \qquad P_{0,\mathbf{n}} := \frac{\theta}{\theta + |\mathbf{n}|} P_0 + \frac{|\mathbf{n}|}{\theta + |\mathbf{n}|} P_{\mathbf{n}}$$

play the role of concentration parameter and baseline probability measure (i.e., the initial urn configuration), respectively. Thus (3.3) can be seen as a mixture of Pólya urns where the base measure has a randomised discrete component $P_{\mathbf{n}}$. Unlike in (1.3), observations not drawn from empirical measure P_k of the current sample can therefore be drawn either from P_0 or from the empirical measure $P_{\mathbf{n}}$, where past observations are assigned multiplicities $\mathbf{n}$ with probability $p_t^{(k)}(\mathbf{M}, \mathbf{n})$.

An alternative interpretation is obtained by expanding the sum in (3.3) to obtain a single generalised Pólya urn, written in compact form as

$$\mathbb{P}\big(Y_{T+t}^{k+1} \in \cdot \,|\, \mathbf{Y}_{0:T}, Y_{T+t}^{1:k}\big) = A_k P_0(\cdot) + \sum_{i=1}^{K} C_{i,k} \delta_{y_i^*}(\cdot) + B_k P_k(\cdot), \tag{3.10}$$

where A is a Borel set of $\mathcal{Y}$. In this case, the first observation is either from P_0 or a copy of a past value $\mathbf{Y}_{0:T}$, namely

$$Y_{T+1}^1 \sim \begin{cases} P_0 & \text{w.p. } A_0 \\ \delta_{y_i^*} & \text{w.p. } C_{i,0}, \end{cases}$$

while the $(k+1)$st can also be a copy of one of the first k current observations $Y_{T+t}^{1:k}$, namely

$$Y_{T+1}^{k+1} \sim \begin{cases} P_0 & \text{w.p. } A_k \\ \delta_{y_i^*} & \text{w.p. } C_{i,k} \\ P_k & \text{w.p. } B_k. \end{cases}$$

The pool of values to be copied is therefore given by past values $\mathbf{Y}_{0:T}$ and current, already sampled observations $Y_{T+t}^{1:k}$.

As a byproduct of Proposition 2, we can evaluate the correlation between observations at different time points.

Proposition 3. *For $t, s > 0$, let Y_t, Y_{t+s} be from (2.4). Then*

$$\mathrm{Corr}(Y_t, Y_{t+s}) = \frac{e^{-\frac{\theta}{2}s}}{\theta + 1}.$$

Unsurprisingly, the correlation decays to 0 as the lag s goes to infinity. Moreover,

$$\mathrm{Corr}(Y_t, Y_{t+s}) \to \frac{1}{\theta + 1}, \quad \text{as } s \to 0$$

which is the correlation of two observations from a DP as in (1.1).

3.2 Sampling from the predictive distribution

In order to make Proposition 2 useful in practice, we provide an explicit algorithm to sample from the predictive distribution (3.3), which can be useful *per se* or for approximating posterior quantities of interest. Exploiting (3.9) and the fact that (3.3) can be seen as a mixture of Pólya urns, we can see $\mathbf{n} \in \mathbb{Z}_+^K$ as a latent variable whereby, given $\mathbf{n}$, sampling proceeds very similarly to a usual Pólya urn.

Recalling that $|\mathbf{n}| = \sum_{j=1}^K n_i$, a simple algorithm for the $(k+1)$st observation would therefore be:

- sample $\mathbf{n} \in L(\mathbf{M})$ w.p. $p_t^{(k)}(\mathbf{M}, \mathbf{n})$;
- sample from $P_0, P_{\mathbf{n}}$ or P_k with probabilities proportional to $\theta, |\mathbf{n}|, k$ respectively;
- update weights $p_t^{(k)}(\mathbf{M}, \mathbf{n})$ to $p_t^{(k+1)}(\mathbf{M}, \mathbf{n})$ for each $\mathbf{n} \in L(\mathbf{M})$.

A detailed pseudo-code description is provided in Algorithm 1.

Algorithm 1 Exact sampling from (3.3).

Input: - active nodes at time T: $\mathbf{M}$
- precision parameter: θ
- last mixture weights $p_t^{(k)}(\mathbf{M}, \mathbf{n})$, $\mathbf{n} \in L(M)$
- past unique observations: $y_1^*, \dots, y_K^*$
- current observations: $y_{T+t}^1, \dots, y_{T+t}^k$

Sample $\mathbf{n}$ w.p. $p_t^{(k)}(\mathbf{M}, \mathbf{n})$, $\mathbf{n} \in L(\mathbf{M})$

Sample Y from $P_0, P_{\mathbf{n}}$ or P_k w.p. $\frac{\theta}{\theta+|\mathbf{n}|+k}, \frac{|\mathbf{n}|}{\theta+|\mathbf{n}|+k}, \frac{k}{\theta+|\mathbf{n}|+k}$ respectively

Set $y_{T+t}^{k+1} = Y$

Update parameters:

for $\mathbf{n} \in L(\mathbf{M})$ and $p(y_{T+t}^{k+1} \mid y_{T+t}^{1:k})$ as in (3.8) **do**

$p_t^{(k+1)}(\mathbf{M}, \mathbf{n}) = p_t^{(k)}(\mathbf{M}, \mathbf{n}) p(y_{T+t}^{k+1} \mid y_{T+t}^{1:k})$

Normalize $p_t^{(k+1)}(\mathbf{M}, \mathbf{n})$

A possible downside of the above sampling strategy is that when the set $L(\mathbf{M})$ is large, updating all weights may be computationally demanding. Indeed, the size of the set $L(\mathbf{M})$ is $|L(\mathbf{M})| = \prod_{j=1}^K (1 + m_j)$, where m_j is the multiplicity of y_j^* in the data, which can grow considerably with the number of observations (cf. also Proposition 2.5 in Papaspiliopoulos and Ruggiero 2014). It is however to be noted that, due to the properties of the death process that ultimately governs the time-dependent mixture weights, typically only a small portion of these will be significantly different from zero. Figure 4 illustrates this point by showing the nodes in $\{0, \dots, 50\}$ with weight larger than 0.05 at different times, if at time 0 there is a unit mass at the node 50, when $\theta = 1$. A deeper investigation of these aspects in a similar, but parametric, framework, can be found in Kon Kam King et al. (2020).

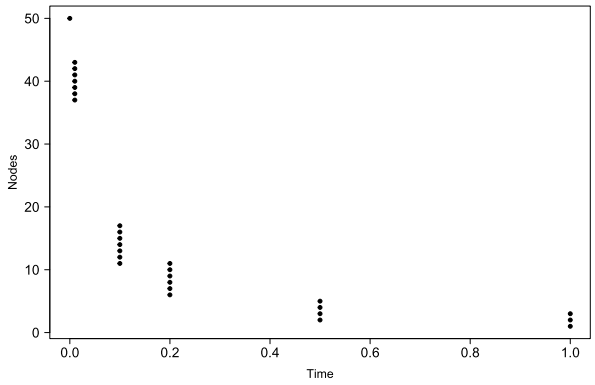

Figure 4: Nodes in $\{0, \ldots, 50\}$ (black dots) with probability of being reached by the death process bigger than .05 after lags .01, .1, .2, .5 and 1 (horizontal axis). Starting with mass 1 at the point 50, only a handful of nodes have significant mass after these lags.

Hence an approximate version of the above algorithm can be particularly useful to exploit this aspect. We can therefore target a set $\tilde{\mathbf{M}} \subset L(\mathbf{M})$ such that $|\tilde{\mathbf{M}}| \ll |L(\mathbf{M})|$ and $\sum_{\mathbf{n}\in\tilde{\mathbf{M}}} p_t(\mathbf{M}, \mathbf{n}) \approx 1$ by inserting a Monte Carlo step in the algorithm and simulate the death process with a large number of particles. The empirical frequencies of the particles landing nodes will then provide an estimate of the weights $p_t(\mathbf{M}, \mathbf{n})$ in (3.2), and one can retain only those above a certain threshold. Furthermore, the simulation of the multidimensional death process can be factorised into simulating a one-dimensional death process, which simply tracks the number of steps down the graph, and hypergeometric sampling for choosing the landing node within the reached level. A simple algorithm for simulating the death process is as follows: for $i = 1, \ldots, N$,

- draw $\mathbf{m}$ with probability $w_{\mathbf{m}}$ and set $m = |\mathbf{m}|$;
- run a one-dimensional death process from m, and let n be the landing point after time t;
- draw $\mathbf{n}^{(i)} \sim \mathrm{HG}(n, \mathbf{m}/|\mathbf{m}|)$;

and return $\{\mathbf{n}^{(i)}, i = 1, \ldots, N\}$. Note, in turn, that the simulation of the death process trajectories does not require to evaluate its transition probabilities (3.5), which are prone to numerical instability, and can instead be straightforwardly set up in terms of successive exponential draws by repeating the following cycle: for $i \geq 1$,

- draw $Z_i \sim \mathrm{Exp}(m(\theta + m - 1)/2)$
- if $\sum_{j\leq i} Z_j < t$ set $m = m - 1$ else return $n = m - i + 1$ and exit cycle.

Algorithm 2 outlines the pseudocode for sampling approximately from (3.3) according to this strategy.

Algorithm 2 Approximate sampling from (3.3).

Input: - active nodes at time T: $\mathbf{M}$
- time to propagate: t
- precision parameter: θ
- mixture weights at time T: $w_{\mathbf{m}}$
- past unique observations: $y_1^*, \ldots, y_K^*$
- number of Monte Carlo iterates: N
- weights threshold: $\varepsilon \geq 0$

$\tilde{\mathbf{M}} = \emptyset$; $w = \emptyset$
for $i \in 1 : N$ **do**
  Sample $\mathbf{m}$ w.p. $w_{\mathbf{m}}$, $\mathbf{m} \in \mathbf{M}$
  $n = |\mathbf{m}|$; $s = t$
  for $j \geq 1$ **do**
    Sample Z from $\text{Exp}(n(\theta + n - 1)/2)$ and set $s = s - Z$
    if $s > 0$ and $n > 0$ **then**
      Set $n = n - 1$
    else
      Return n and exit cycle.
  Sample $\mathbf{n} \sim \text{HG}(n, \mathbf{m}/|\mathbf{m}|)$
  if $\mathbf{n} \notin \tilde{M}$ **then**
    Add $\mathbf{n}$ to $\tilde{\mathbf{M}}$ and add 1 to w
  else
    Add 1 to the corresponding element of w
Normalize w.
Retain weights $w > \varepsilon$ and normalize again.
Apply algorithm 1 with $\mathbf{M} = \tilde{\mathbf{M}}$ and $p_t(\mathbf{M}, \mathbf{n}) = w$

3.3 Partition structure and Chinese restaurants with conveyor belt

A sample from (3.3) will clearly feature ties among the observations, since there are two discrete sources for the data, namely $P_{\mathbf{n}}$ and P_k. A fundamental task concerning sampling models with ties is to characterize the distributional properties of the induced random partition. We say that a random sample $(Y_1, \ldots, Y_n)$ induces a partition with frequencies $(n_1, \ldots, n_K)$ if $\sum_{i=1}^K n_i = n$ and grouping the observed values gives multiplicities $(n_1, \ldots, n_K)$. The distribution of a random partition generated by an exchangeable sequence is encoded in the so-called exchangeable partition probability function, which for the Dirichlet process was found in Antoniak (1974) to be

$$p(n_1, \ldots, n_k) = \frac{\theta^k}{\theta_{(n)}} \prod_{i=1}^{k} (n_i - 1)!, \tag{3.11}$$

with $\theta_{(n)}$ as in Section 2. The sampling scheme on the space of partitions associated to the Dirichlet process is generally depicted through a Chinese restaurant process (Pitman, 2006): the first customer sits at a table and orders a dish from the menu P_0,

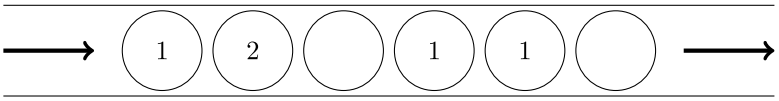

Figure 5: Schematic depiction of a conveyor belt running through the Chinese restaurant. The conveyor makes available to the customers only a time-varying selection from a pool of dishes. The Figure depicts the current selection, given by three dishes of type 1, one of type 2 and two empty slots from which previously available dishes have been removed.

while successive customers either sit at an existing table j, with probability proportional to its current occupancy n_j, and receive the same dish as the other occupants, or sit at an unoccupied table, with probability proportional to θ, and order from P_0.

To account for random partitions induced by a FV-DDP, one can think of a *conveyor belt* typical of some Chinese restaurants, which delivers a non constant selection of dishes that customers can choose to pick up. See Figure 5. In the context of (3.3), each new customer on day $T+t$ faces a different configuration $\mathbf{n}$ of dishes available on the conveyor belt, determined by the weights $p_t^{(k)}(\mathbf{M}, \mathbf{n})$. This depends on the following factors: (i) which dishes were most popular on day T, the greater the popularity, the higher their multiplicity in the nodes of $\mathbf{M}$, hence the greater their average multiplicity on the conveyor on day $T+t$ as determined by $\mathbf{n}$; (ii) the removal of dishes that showed symptoms of food spoilage before the first customer arrives, as determined by the temporal component; (iii) previous customers' choices, as the kitchen readjusts the conveyor at each new customer by reinforcing the most popular dishes, as determined by the update (3.7).

Schematically, the Chinese restaurant process with conveyor belt proceeds as follows. The first customer at time $T+t$ arrives at the restaurant, finds the configuration $\mathbf{n}$ on the conveyor belt, then picks a dish

- from the conveyor belt, with probability $|\mathbf{n}|/(\theta+|\mathbf{n}|)$
- from the menu P_0, with probability $\theta/(\theta+|\mathbf{n}|)$

and sits at the first table. The kitchen then readjusts the offer on the conveyor belt based on the first customer's choice, through (3.7). The $(k+1)$st customer arrives at the restaurant, finds a configuration $\mathbf{n}'$ on the conveyor belt, then

- with probability $m_j/(\theta+|\mathbf{n}'|+k)$ sits at table j and receives the same dish as the other occupants, m_j being the current table occupancy
- otherwise picks a dish
  - from the conveyor belt, with probability $|\mathbf{n}'|/(\theta+|\mathbf{n}'|+k)$
  - from the menu P_0, with probability $\theta/(\theta+|\mathbf{n}'|+k)$

  and sits at a new table.

Note that node $\mathbf{0}$ has always positive probability, in which case the conveyor belt is

empty and (3.3) reduces to (1.3). Hence a customer facing the configuration $\mathbf{n} = \mathbf{0}$ is entering a usual Chinese restaurant.

An approach to formally deriving the law of a random partition induced by n observations from X_{T+t} would be to compute

$$\int_{\mathcal{Y}^q} \mathbb{E}\Big[[X_{T+t}(\mathrm{d}y_1)]^{n_1} \cdots [X_{T+t}(\mathrm{d}y_q)]^{n_q}\Big], \qquad q \leq n,$$

which evaluates the probability of all possible configurations of multiplicities $(n_1, \ldots, n_q)$, with $q \leq n$ and $\sum_{h=1}^{q} n_h = n$, irrespective of the values Y_i that generated them. This entails a considerable combinatorial complexity, particularly given by the fact that X_{T+t}, which has a similar representation to (2.5), is given by a mixture of Dirichlet processes whose base measures have partially shared discrete components.

Alternatively, one can derive (3.11) from (1.3), better seen by rewriting P_k in terms of multiplicities of the distinct values, by assuming observations in the same group arrive sequentially, so that the first group has multiplicity n_1 with probability proportional to $\theta(n_1 - 1)!$, the second has multiplicity n_2 with probability proportional to $\theta(n_2 - 1)!$, and so on. Similarly, we can use the results in Proposition 2 to derive the explicit law of a partition induced by a sample from X_{T+t}. The resulting expression, given in Lemma 2 in the Supplementary Material, suffers from the combinatorial complexity due to the possibility of sampling values that start a group both from P_0 and from $P_{\mathbf{n}}$, where $\mathbf{n}$ is itself random. Here instead we provide an algorithm for generating such random partitions, which can be used, for example, to study the posterior distribution of the number of clusters directly, i.e. without resorting to Proposition 2. Mimicking the argument above, we need to

- choose whether to sample a new value from P_0 or from any of the $P_{\mathbf{n}}$'s
- draw the new observation after excluding from $P_{\mathbf{n}}$ the recorded values
- draw the size of the corresponding group.

From (3.10), the probability of drawing a new observation is therefore given by $A_k + \sum_{i \in \mathcal{K}} C_{i,k}$, where $\mathcal{K}$ is the set of past observations still not present in the current sample. The probability of enlarging a group associated to the value y by one is instead

$$\begin{cases} B_k P_k(\{y\}) & \text{if } y \neq y_j^*, \forall j, \\ B_k P_k(\{y\}) + C_{j,k} & \text{if } y = y_j^*. \end{cases}$$

Algorithm 3 outlines the pseudocode for sampling a random partition according to this strategy.

3.4 Asymptotics

We investigate two asymptotic regimes for (3.3). The following Proposition shows that when $t \to \infty$, the FV-DDP predictive distribution converges to the usual Pólya urn (1.3).

Algorithm 3 Sampling random partitions at time $T+t$.

Input: - active nodes at time T: $\mathbf{M}$
- mixture weights at time T: $w_{\mathbf{m}}$
- past unique observations: $y^*_1, \ldots, y^*_K$
- number of observations to draw: n

Initialize $L = 0$, $\mathcal{K} = \{1, \ldots, K\}$ and $\mathcal{D} = \emptyset$
while $L < n$ **do**
  Sample N equal to 0 w.p. A_L and equal to i w.p. $C_{i,L}$, with $i \in \mathcal{K}$.
  if $N = 0$ **then**
    Sample Y from P_0
    Sample l equal to
      - 1 w.p. $A_{L+1} + \sum_{i\in\mathcal{K}} C_{i,L+1}$
      - j w.p. $\left(A_{L+j} + \sum_{i\in\mathcal{K}} C_{i,L+j}\right) \prod_{p=1}^{j-1} B_{L+p} \frac{p}{L+1}$
        with $j = \quad 2, \ldots, n-L$
  else
    Set $Y = y^*_N$ and set $\mathcal{K} = \mathcal{K} \backslash N$
    Sample l equal to
      - 1 w.p. $A_{L+1} + \sum_{i\in\mathcal{K}} C_{i,L+1}$
      - j w.p. $\left(A_{L+j} + \sum_{i\in\mathcal{K}} C_{i,L+j}\right) \prod_{p=1}^{j-1} \left[C_{i,L+p} B_{L+p} \frac{p}{L+1}\right]$
        with $j = \quad 2, \ldots, n-L$
  Set $L = L + l$ and add Y to $\mathcal{D}$.
Return $\mathcal{D}$

Proposition 4. *Under the hypotheses of Proposition 2, we have*

$$\mathcal{L}\left(Y^{k+1}_{T+t} | \mathbf{Y}_{0:T}, Y^{1:k}_{T+t}\right) \xrightarrow{\text{TV}} \frac{\theta}{\theta+k} P_0 + \frac{k}{\theta+k} P_k, \qquad a.s., \text{ as } t \to \infty,$$

with P_k as in (3.4).

Here $\xrightarrow{\text{TV}}$ denotes convergence in total variation distance, and the statement holds almost surely with respect to the probability measure induced by the FV model on the space of measure-valued temporal trajectories. A heuristic interpretation of the above result is that, when the lag between the last and the current data collection point diverges, the information given by past observations $\mathbf{Y}_{0:T}$ becomes obsolete, and sampling from (3.3) approximates sampling from the prior Pólya urn (1.3). This should be intuitive, as very old information, relative to the current inferential goals, should have a negligible effect.

Unsurprisingly, it can be easily proved that an analogous result holds for the distribution of the induced partition, which converges to the EPPF of the Dirichlet process as $t \to \infty$. The proof follows similar lines to that of Proposition 4, and is therefore omitted. In the conveyor belt metaphor, as t increases all dishes on the conveyor belt have been removed due to food spoilage, before the next customer comes in.

The following Proposition shows that when $k \to \infty$ in (3.3), we recover the law of X_{T+t} given $\mathbf{Y}_{0:T}$ as de Finetti measure.

Proposition 5. *Under the hypotheses of Proposition 2, we have*

$$\mathcal{L}\big(Y_{T+t}^{k+1} | \mathbf{Y}_{0:T}, Y_{T+t}^{1:k}\big) \Rightarrow P_{T+t}^*, \quad a.s., \quad as\ k \to \infty,$$

where $P^* \sim \mathcal{L}(X_{T+t} | \mathbf{Y}_{0:T})$.

Here P^* is a random measure with the same distribution as the FV-DDP at time $T+t$ given only the past information $\mathbf{Y}_{0:T}$. Recall for comparison that the same type of limit for (1.3) yields

$$\mathcal{L}(Y_{k+1} | Y_1, \ldots, Y_k) \Rightarrow P^*, \qquad P^* \sim \Pi_\alpha, \quad \text{as } k \to \infty,$$

where Π_α is the de Finetti measure of the sequence and P^* is sometimes called the directing random measure.

4 Illustration

We illustrate predictive inference using FV-DDPs, based on Proposition 2. Besides the usual prior specification of Dirichlet process-based models, which involves the total mass θ and the baseline distribution P_0, here we introduce a parameter $\sigma > 0$ that controls the speed of the DDP. This acts as a time rescaling, whereby the data collection times t_i are rescaled to σt_i. This additional parameter provides extra flexibility for estimation, as it can be used to adapt the prior to the correct time scale of the underlying data generating process.

4.1 Synthetic data

We consider data generated by the model

$$\begin{aligned} Y_t &\sim \frac{1}{2}\mathrm{Po}(\mu_t^{-1}, 0) + \frac{1}{2}\mathrm{Po}(\nu_t^{-1}, 5), \\ \mu_t &= \mu_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{Exp}(1), \\ \nu_t &= \nu_{t-1} + \eta_t, \quad \eta_t \sim \mathrm{Exp}(1), \eta_t \perp\!\!\!\perp \epsilon_t, \end{aligned}$$

where $\mathrm{Po}(\lambda, b)$ denotes a b-translated Poisson distribution with parameter λ, and where $\mu_0^{-1} = \nu_0^{-1} = 5$, for $t = 0, 1, 2, \ldots$. We collect 15 observations at each $t \in \{0, \ldots, 15\}$ and consider one-step-ahead predictions based on the first 5 and 15 data collection times.

We fit the data by using a FV-DDP model as specified in (2.4), with the following prior specification. We consider two choices for P_0, a Negative Binomial with parameters $(2, 0.5)$ and a Binomial with parameters $(99, 0.3)$, which respectively concentrate most of their mass around small values and around the value 30. We consider a uniform prior on θ concentrated on the points $\{.5, 1, 1.5, \ldots, 15\}$. A continuous prior could also

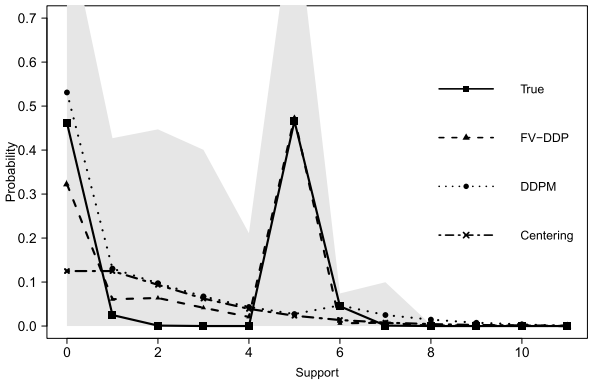

Figure 6: One-step-ahead prediction and 95% pointwise credible intervals, based on 15 data collection times.

be envisaged, at the cost of adding a Metropolis–Hastings step in the posterior simulation, which we avoid here for the sake of computational efficiency. Similarly, for σ we consider a uniform prior on the values $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.5\}$. The estimates are obtained by means of 500 replicates of (3.3) of 1000 observations each, using the approximate method outlined in Algorithm 2 with 10000 Monte Carlo iterates. We also compare the FV-DDP estimate with that obtained using the DDP proposed in Gutierrez et al. (2016). This is constructed from the stick-breaking representation (1.2) by letting

$$V_i(t_n) \sim c\delta_{V'} + (1-c)\delta_{V_i(t_{n-1})}, \qquad c \in (0,1), \qquad V' \sim \text{Beta}(1,\theta),$$

and keeping the locations Y_i fixed. We let the resulting DDP be the mixing measure in a time-dependent mixture of Poisson kernels, which provides additional flexibility to this model with respect to our proposal. Furthermore, we give the competitor model a considerable advantage by training it also with the data points collected at times 6 and 7, which provide information on the prediction targets, and by centering it on the Negative Binomial with parameters $(2, 0.5)$, rather than on the above mentioned mixture, which puts mass closer to where most mass of the true pmf lies.

Figure 6 shows the results on the one-step-ahead prediction with 15 collection times. The posterior of σ (not shown) concentrates most of the mass on points 0.7 and 0.9, which leads to learning the correct time scale for prediction, resulting in an accurate estimate of the true pmf. The credible intervals are quite wide, and a better precision may be achieved by increasing the number of time points at which the data are recorded.

We compare the previous results with those obtained by choosing σ via out-of-sample validation. This is done here using times 0 to 4 as training and time 5 as test, whereby for each $\sigma \in \{.0001, .001, .01, .1, 0.5, 1, 1.5\}$ we compute the sum of absolute

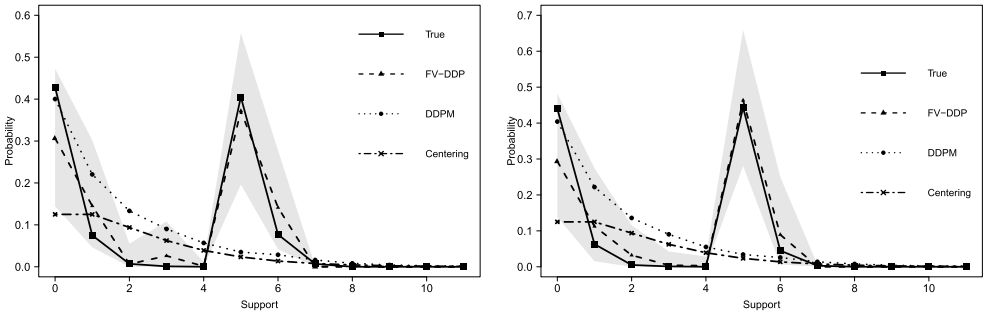

Figure 7: One- (left) and two-step-ahead prediction (right) based on 5 data collection times, with 95% pointwise credible intervals.

errors (SAE) between the FV-DDP posterior predictive mean and the true pmf. These are shown in Table 1, leading to choose $\sigma = .01$.

Table 1: Sum of the absolute error between predicted and true pmf at time 5 for different values of σ.

σ	.0001	.001	.01	.1	.5	1	1.5
SAE	.1410	.1345	**.1064**	.1301	.1261	.1595	.1847

Table 2 shows the posterior weights of relevant values of θ among those with positive prior mass, for the above mentioned choices of P_0 and using the chosen value of σ. The model correctly assigns all posterior probability to the Negative Binomial centering (Binomial not reported in the table), which moves mass towards smaller values as time increases.

Table 2: Relevant posterior weights of θ.

θ	1	1.5	2	3
NegBinom	.5644	.001694	.04702	0.3868

Figure 7 shows the results in this case for the one- and two-step-ahead predictions given only 5 data collection times. The true pmf is correctly predicted by the FV-DDP estimate even in this short horizon scenario, and the associated 95% pointwise credible intervals are significantly sharper if compared to Figure 6, obtained with a longer horizon. The prediction based on the alternative DDP mixture does not infer correctly the target, leading to an associated normalised ℓ_1 distance from the true pmf of 12.72% and 12.84%, compared to 4.95% and 4.90% for the FV-DDP prediction.

4.2 Karnofsky score data

We consider the dataset *hodg* used in Klein and Moeschberger (1997), which contains records on the time to death or relapse and the Karnofsky score for 43 patients with a lymphoma disease. The Karnofsky score (KS) is an index attributed to individual patients, with higher values indicating a better prognosis.

In the framework of model (2.4), we take the times of death or relapse as collection times and let the KS of the survivors at each time be the data. We aim at predicting the future distribution of the KS among the patients who are still in the experiment at that time, which would be an indirect assessment of the effectiveness of the score in describing the patients' prognosis. We also include censored observations (patients leaving the experiment for reasons different from death or relapse), without having them trigger a collection time. The FV-DDP appears as the ideal modeling tool in this framework since it includes a probabilistic mechanism that accounts for the reduced number of observations through different time points.

We train the model up to 42, 108 and 406 days after the start of the experiment, and we make predictions 28, 112 and 144 days ahead, respectively. As regards the prior, we put a uniform distribution on the observed scores (note that new score values cannot appear along the experiment) and we uniformly randomize θ over $\{.5, 1, 1.5, \dots, 15\}$, analogously to Section 4.1. Given the results of the previous subsection for different approaches to selecting σ, here, after transforming the lags in annual, we proceed by selecting σ for each value of θ by maximizing the probability that the death process makes the right number of transitions in the desired laps of time. Some of the selected values for $\sigma_1, \sigma_2, \sigma_3$ for the three different trainings, depending on θ, are shown in Table 3.

Table 3: Choice of σ for some values of θ for the three trainings.

θ	.5	1	1.5	$\cdots$	29	29.5	30
σ_1	0.4947	0.4913	0.4885	$\cdots$	0.3235	0.3266	0.3228
σ_2	0.6059	0.6014	0.5696	$\cdots$	0.3684	0.3130	0.3361
σ_3	0.6149	0.6150	0.5789	$\cdots$	0.3063	0.3018	0.2901

Figure 8 shows the three predictions of the scores distribution. Coherently with the intuition, as the experiment goes by, individuals with higher KS become predominant: from 70 to 230 days the predicted weight associated to a score of 90 increases of more than 10%, and similarly for 100. However the distribution of the scores remains pretty stable, apart from the lowest values, meaning that the highest scoring patients actually had much better prognoses, as showed by the third prediction.

These findings are consistent with the Kaplan-Meyer estimate of the survival function, shown in the bottom right panel, which decreases rapidly between 70 and 230 and flattens after that point, implying that the FV-DDP prediction adapted to the periods of quick change in the underlying distribution and periods of relative steady behaviour.

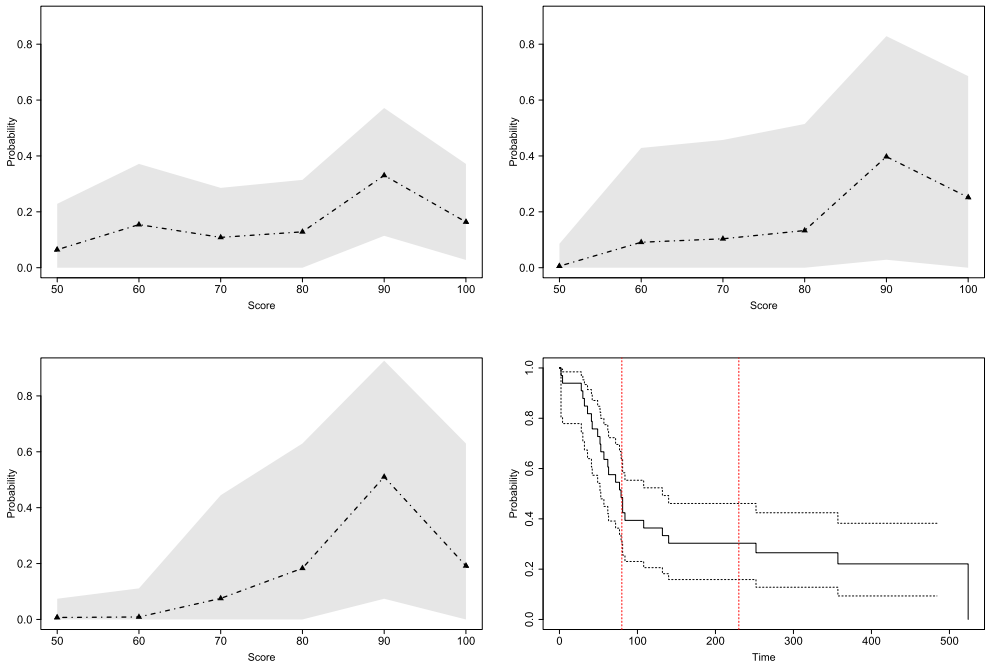

Figure 8: From top left: pmf prediction at 70, 230 and 550 days after the experiment. Bottom right: Kaplan-Meyer estimate of the survival times up to time 550.

5 Discussion

We have derived the predictive distribution for the observations generated by a class of dependent Dirichlet processes driven by a Fleming–Viot diffusion model, which can be characterized as a time-dependent mixture of Pólya urns, and described the induced partition structure together with practical algorithms for exact and approximate sampling of these quantities. An upside of inducing the dynamics through a FV process is that one can implicitly exploit the rich and well understood underlying probabilistic structure in order to obtain manageable closed-form formulae for the quantities of interest. This ultimately relies on the duality with respect to Kigman's coalescent, which was first used for inferential purposes in Papaspiliopoulos and Ruggiero (2014).

The approach we have described yields dependent RPMs with almost surely discrete realisations. While such a feature perfectly fits the specific illustrations we have discussed, it is not suited to draw inferences with continuous data. An immediate and natural extension of the proposed model, which accommodates continuous outcomes would be to consider dependent mixtures of continuous kernels, whereby the observation y from the RPM at time t becomes a latent variable acting as parameter in a parametric kernel $f(z|y)$. This approach would be in line with the extensive Bayesian literature on semi-parametric mixture models, which has largely used the DP or its var-

ious extensions as mixing measure. It remains however a non trivial exercise to derive in this framework the corresponding formulae for prediction, which we will leave for future investigation.

Supplementary Material

Predictive inference with Fleming–Viot-driven dependent Dirichlet processes (DOI: 10.1214/20-BA1206SUPP; .pdf). Contains all proofs of the results provided above, together with the explicit expression for the transition probabilities of the death process and of the distribution of the partition induced by (3.3).

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174. MR0365969. doi: https://doi.org/10.1214/aos/1176342871. 372, 376, 383

Ascolani, F., Lijoi, A. and Ruggiero, M. (2020). Predictive inference with Fleming–Viot-driven dependent Dirichlet processes – Supplementary Material. *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1206SUPP. 374

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. MR2934952. doi: https://doi.org/10.1214/12-BA709. 373, 375

Beal, M. J., Ghahramani, Z. and Rasmussen, C.E. (2002). The infinite hidden Markov model. *Advances in Neural Information Processing Systems* **14**, 577–585. 375

Canale, A. and Ruggiero, M. (2016). Bayesian nonparametric forecasting of monotonic functional time series. *Electron. J. Stat.* **10**, 3265–3286. MR3572849. doi: https://doi.org/10.1214/16-EJS1190. 372

Caron, F., Davy, M. and Doucet, A. (2007) Generalized Pólya urn for time-varying Dirichlet process mixtures. *Proc. 23rd Conf. on Uncertainty in Artificial Intelligence*, Vancouver. 372

Caron, F., Neiswanger, W., Wood, F., Doucet, A. and Davy, M. (2017). Generalized Pólya urn for time-varying Pitman–Yor processes. *J. Mach. Learn. Res.* **18**, 1–32. MR3634894. 372

Caron, F. and Teh, Y. W.(2012). Bayesian nonparametric models for ranked data. *Neural Information Processing Systems* (NIPS 2012), Lake Tahoe, USA, 2012. 372

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 212–229. doi: https://doi.org/10.1109/TPAMI.2013.217. 372

Dunson, D. B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics* 7, 551–568. MR2490545. doi: https://doi.org/10.1093/biostatistics/kxj025. 372

Ethier, S. N. and Griffiths, R. C. (1993). The transition function of a Fleming–Viot process. *Ann. Probab.* **21**, 1571–1590. MR1235429. doi: https://doi.org/10.1214/aop/1176989131. 375

Ethier, S. N. and Kurtz, T. G. (1993). Fleming–Viot processes in population genetics. *SIAM J. Control Optim.* **31**, 345–386. MR1205982. doi: https://doi.org/10.1137/0331019. 374

Favaro, S., Ruggiero, M. and Walker, S. G. (2009). On a Gibbs sampler based random process in Bayesian nonparametrics. *Electron. J. Stat.* **3**, 1556–1566. MR2578838. doi: https://doi.org/10.1214/09-EJS563. 374

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230. MR0350949. doi: https://doi.org/10.1214/aos/1176342360. 371

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press. MR3587782. doi: https://doi.org/10.1017/9781139029834. 371

Gnedin, A. and Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **325**, 83–102. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 372

Griffin, J. E. and Steel, M. F. J. (2010). Stick-breaking autoregressive processes. *J. Econometrics* **162**, 383–396. MR2795625. doi: https://doi.org/10.1016/j.jeconom.2011.03.001. 372

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theor. Popul. Biol.* **17**, 37–50. MR0568666. doi: https://doi.org/10.1016/0040-5809(80)90013-1. 374

Gutierrez, 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. doi: https://doi.org/10.1016/j.csda.2015.10.002. 372, 388

Hjort, N. L., Holmes, C. C., Müller, P. and Walker, S.G., eds. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge Univ. Press. MR2722988. doi: https://doi.org/10.1017/CBO9780511802478.002. 371

Jenkins, P. A. and Spanò, D. (2017). Exact simulation of the Wright–Fisher diffusion. *Ann. Appl. Probab.* **3**, 1478–1509. MR3678477. doi: https://doi.org/10.1214/16-AAP1236. 376

Kon Kam King, G., Canale, A. and Ruggiero, M. (2020). Bayesian functional forecasting with locally-autoregressive dependent processes. *Bayesian Anal.* **14**, 1121–1141. MR4044848. doi: https://doi.org/10.1214/18-BA1140. 372

Kon Kam King, G., Papaspiliopoulos, O. and Ruggiero, M. (2020). Exact inference for a class of hidden Markov models on general state spaces. *Preprint*. 381

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis Techniques for Censored and Truncated Data*. Springer-Verlag New York 390

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222–1235. MR1186248. doi: https://doi.org/10.1214/aos/1176348767. 372

Lijoi, A., Mena, R. H. and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-gaussian priors. *J. Amer. Statist. Assoc.* **472**, 1278–1291. MR2236441. doi: https://doi.org/10.1198/016214505000000132. 372

Lijoi, A., Mena, R. H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Statist. Soc. Ser. B* **69**, 715–740. MR2370077. doi: https://doi.org/10.1111/j.1467-9868.2007.00609.x. 372

Lijoi, A., Ruggiero, M. and Spanò, D. (2016). On the transition function of some time-dependent Dirichlet and gamma processes. In *JSM Proceedings, Section on Nonparametric Statistics*. Alexandria, VA: American Statistical Association. 375

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Statist.* **12**, 351–357. MR0733519. doi: https://doi.org/10.1214/aos/1176346412. 372

MacEachern, S. N. (1999). Dependent nonparametric Processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statist. Assoc., Alexandria, VA. 372

Mauldin, R. D., Sudderth, W. D. and Williams, S. C. (1992). Polya trees and random distributions. *Ann. Statist.* **20**, 1203–1221. MR1186247. doi: https://doi.org/10.1214/aos/1176348766. 372

Mena, R. H. and Ruggiero, M. (2016). Dynamic density estimation with diffusive Dirichlet mixtures. *Bernoulli* **22**, 901–926. MR3449803. doi: https://doi.org/10.3150/14-BEJ681. 372

Müller, P., Quintana, F. A., Jara, A. and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer. MR3309338. doi: https://doi.org/10.1007/978-3-319-18968-0. 371

Papaspiliopoulos, O. and Ruggiero, M. (2014). Optimal filtering and the dual process. *Bernoulli* **20**, 1999–2019. MR3263096. doi: https://doi.org/10.3150/13-BEJ548. 381, 391

Papaspiliopoulos, O., Ruggiero, M. and Spanò, D. (2016). Conjugacy properties of time-evolving Dirichlet and gamma random measures. *Electron. J. Stat.* **10**, 3452–3489. MR3572856. doi: https://doi.org/10.1214/16-EJS1194. 373, 375, 376

Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92**, 21–39. MR1156448. doi: https://doi.org/10.1007/BF01205234. 372

Pitman, J. (2006). *Combinatorial Stochastic Processes*. Lecture Notes in Math. **1875**. Springer, Berlin. MR2245368. 383

Pitman, J. and Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 372

Rodriguez, A. and ter Horst, E. (2008). Bayesian dynamic density estimation. *Bayes. Anal.* **3**, 339–366. MR2407430. doi: https://doi.org/10.1214/08-BA313. 372

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 372

Sethuraman, J. (1994). A constructive definition of the Dirichlet process prior. *Statist. Sinica* **2**, 639–650. MR1309433. 372

Stepleton, T., Ghahramani, Z., Gordon, G., and Lee, T.-S. (2009). The block diagonal infinite hidden Markov model. *Journal of Machine Learning Research* **5**, 544–551. 375

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetic models. *Theoret. Population Biol.* **26**, 119–164. MR0770050. doi: https://doi.org/10.1016/0040-5809(84)90027-3. 374

Van Gael, V., Saatci, Y., Teh, Y. W. and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning*. 375

Walker, S. G., Hatjispyros S. J. and Nicoleris, T. (2007). A Fleming–Viot process and Bayesian nonparametrics. *Ann. Appl. Probab.* **17**, 67–80. MR2292580. doi: https://doi.org/10.1214/105051606000000600. 374

Yau, C., Papaspiliopoulos, O., Roberts, G. O. and Holmes, C. (2011). Bayesian non-parametric hidden Markov models with applications in genomics. *J. Roy. Statist. Soc. Ser. B* **73**, 37–57. MR2797735. doi: https://doi.org/10.1111/j.1467-9868.2010.00756.x. 375

Zhang, A., Zhu, J. and Zhang, B. (2014). Max-margin infinite hidden Markov models. In *Proceedings of the 31st International Conference on Machine Learning*. 375

Acknowledgments

The authors are grateful to an Associate Editor and two anonymous referees for carefully reading the manuscript and for providing helpful and constructive comments. The second and third authors are partially supported by the Italian Ministry of Education, University and Research (MIUR) through PRIN 2015SNS29B. The third author is also supported by MIUR through "Dipartimenti di Eccellenza" grant 2018-2022. Helpful discussions with Amil Ayoub are gratefully acknowledged by the first author.