Journal of Machine Learning Research 18 (2017) 1-32 Submitted 8/10; Revised 12/14; Published 4/17

Generalized Pólya Urn for Time-Varying Pitman-Yor Processes

François Caron* CARON@STATS.OX.AC.UK
Department of Statistics
University of Oxford
Oxford, UK

Willie Neiswanger* WILLIE@CS.CMU.EDU
Machine Learning Department
Carnegie Mellon University
Pittsburgh, USA

Frank Wood FWOOD@ROBOTS.OX.AC.UK
Department of Engineering Science
University of Oxford
Oxford, UK

Arnaud Doucet DOUCET@STATS.OX.AC.UK
Department of Statistics
University of Oxford
Oxford, UK

Manuel Davy MANUEL.DAVY@VEKIA.FR
VEKIA
165 Avenue de Bretagne
59044 Lille, France

* *Joint first authorship.*

Editor: David Dunson

Abstract

This article introduces a class of first-order stationary time-varying Pitman-Yor processes. Subsuming our construction of time-varying Dirichlet processes presented in (Caron et al., 2007), these models can be used for time-dynamic density estimation and clustering. Our intuitive and simple construction relies on a generalized Pólya urn scheme. Significantly, this construction yields marginal distributions at each time point that can be explicitly characterized and easily controlled. Inference is performed using Markov chain Monte Carlo and sequential Monte Carlo methods. We demonstrate our models and algorithms on epidemiological and video tracking data.

Keywords: Bayesian nonparametrics, clustering, mixture models, sequential Monte Carlo, particle Markov chain Monte Carlo, dynamic models

1. Introduction

This paper introduces a class of dependent Pitman-Yor processes[1] that can be used for time-varying density estimation and classification in a variety of settings. The construction of this class of dependent Pitman-Yor processes is in terms of a generalised Pólya urn scheme where dependencies between distributions that evolve over time, for instance, are induced by simple urn-like operations on counts and the parameters to which they are associated.

Our Pólya urn for time-varying Pitman-Yor processes is expressive per dependent slice, as each is represented by a Pitman-Yor process (infinite) mixture distribution of which the component densities may as usual take any form. The dependency-inducing mechanism is also flexible and easy to control, a claim supported by an applied literature (see Gasthaus et al. (2008); Ji and West (2009); Ozkan et al. (2009); Bartlett et al. (2010); Neiswanger et al. (2014); Jaoua et al. (2014) among others) that has grown around the subsumed version of this work (Caron et al., 2007). The generalised Pólya urn dependent Dirichlet process can be recovered by a specific settings of the parameters in this generalized model.

Most of the emphasis of this paper is on defining the model and describing its statistical characteristics. While additional model complexity does not automatically beget useful expressivity, in the case of this model and others like it (see Lin et al. (2010) for example), we assert that it does. For example, in applications like fully unsupervised visual object detection and tracking there is an increasing need for top-down models that introduce coherence through bias towards physical plausibility. While dependent density estimation techniques abound in the literature, there are few that possess the right combination of interpretability and flexibility to fill the role of top-down priors for such complex applications. We illustrate experimentally that this model succeeds at filling this role.

The remainder of the paper is organized as follows: In Section 2 we review Pitman Yor processes and stationarity. In Section 3, we present our main contribution, first-order stationary Pitman-Yor mixture models. Section 4 describes sequential Monte Carlo (SMC) and Markov chain Monte Carlo (MCMC) algorithms for inference. We demonstrate the models and algorithms in Section 5 on various applications. Finally we discuss some potential extensions of this class of models in Section 6 and Appendix A.

2. Background

Pitman-Yor processes include a wide class of distributions on random measures such as the popular Dirichlet process (Ferguson, 1973) and the finite symmetric Dirichlet-multinomial prior (Green and Richardson, 2001); see for example (Pitman and Yor, 1997). In particular, Dirichlet process mixtures can be interpreted as a generalization of finite mixture models to infinite mixture models and have become very popular over the past few years in statistics and related areas to perform clustering and density estimation (Escobar and West, 1995; Müller and Quintana, 2004; Teh and Jordan, 2010). More general Pitman-Yor processes enjoy greater flexibility and have been shown to provide a better fit to text or image data due to their ability to capture power-law properties of such data (Teh, 2006).

However, there are many situations where we cannot assume that the distribution of the observations is fixed and instead this latter evolves over time. For example, in a clustering

1. A preliminary version of this work has been presented as a conference paper (Caron et al., 2007)

application, the number of clusters and the locations of these clusters may vary over time. This situation also occurs in spatial statistics where the spatial distribution of localized events such as diseases or earthquakes changes over time (Hall et al., 2006). To address such problems, this article introduces a novel class of time-varying first-order stationary Pitman-Yor process mixtures; that is processes which have marginals following the same Pitman-Yor process mixture.

We first briefly recall here standard results about Pitman-Yor process mixtures. Let $t = 1, 2, ...$ denote a discrete-time index. Note that this index is not needed in this section but will become essential in what is to come. For any generic sequence $\{x_m\}$, we define $x_{k:l} = (x_k, x_{k+1}, \ldots, x_l)$. For ease of presentation, we assume that we receive a fixed number n of observations at each time t denoted $\mathbf{z}_t = \mathbf{z}_{1:n,t}$ which are independent and identically distributed samples from

$$F_t(\cdot) = \int_{\mathcal{Y}} f(\cdot|\mathbf{y}) d\mathbb{G}_t(\mathbf{y}) \tag{1}$$

where $f(\cdot|\mathbf{y})$ is the mixed probability density function and $\mathbb{G}_t$ is the mixing distribution distributed according to a Pitman-Yor process

$$\mathbb{G}_t \sim \mathcal{PY}(\alpha, \theta, \mathbb{H}). \tag{2}$$

Here $\mathbb{H}$ is a base probability measure and the real parameters α and θ satisfy either

$$0 \leq \alpha < 1 \text{ and } \theta > -\alpha \tag{3}$$

$$\text{or } \alpha < 0 \text{ and } \theta = -m\alpha \text{ for } m \in \mathbb{N}. \tag{4}$$

The case $\alpha = 0$ and $\theta > 0$ corresponds to the Dirichlet process denoted $DP(\theta, \mathbb{H})$. The random measure $\mathbb{G}_t$ satisfies the following *stick-breaking* representation (Sethuraman, 1994; Pitman, 1996)

$$\mathbb{G}_t = \sum_{j=1}^{\infty} V_{j,t} \delta_{U_{j,t}} \tag{5}$$

with $V_{j,t} = \beta_{j,t} \prod_{i=1}^{j-1}(1 - \beta_{i,t})$, $\beta_{j,t} \overset{\text{ind}}{\sim} \mathcal{B}(1 - \alpha, \theta + j\alpha)$, $U_{j,t} \overset{\text{iid}}{\sim} \mathbb{H}$ where $\mathcal{B}$ denotes the Beta distribution. From (1), we have equivalently, for $k = 1, \ldots, n$, the following hierarchical model

$$\mathbf{y}_{k,t}|\mathbb{G}_t \overset{\text{iid}}{\sim} \mathbb{G}_t, \tag{6}$$

$$\mathbf{z}_{k,t}|\mathbf{y}_{k,t} \overset{\text{iid}}{\sim} f(\cdot|\mathbf{y}_{k,t}). \tag{7}$$

We can also reformulate the Pitman-Yor process mixture by integrating out the mixing measure $\mathbb{G}_t$ and introducing allocation variables $\mathbf{c}_t = c_{1:n,t}$. For any $j \in \mathcal{J}(\mathbf{c}_t)$, where $\mathcal{J}(\mathbf{c}_t)$ is the set of unique values in $\mathbf{c}_t$, we have

$$U_{j,t} \overset{\text{iid}}{\sim} \mathbb{H}, \tag{8}$$

$$\mathbf{z}_{k,t}|U_{c_{k,t}} \sim f(\cdot|U_{c_{k,t},t}).$$

For convenience, we label here the clusters by their order of appearance. We set $c_{1,1} = 1$, $K_1 = 1$ and $\mathbf{m}_1^1 = m_{1:K_1,1}^1$ a vector of size K_1. Then, at time $t = 1$, for $k = 2, \ldots, n$ the

following Pólya urn model describes the predictive distribution of a new allocation variable (Blackwell and MacQueen, 1973; Pitman, 1996)

w.p. $\frac{m^1_{i,1}-\alpha}{k-1+\theta}$, $i \in \{1,\dots,K_1\}$ set $c^1_{k,1}=i$, $m^1_{i,1}=m^1_{i,1}+1$,
w.p. $\frac{K_1\alpha+\theta}{k-1+\theta}$, set $K_1=K_1+1$, $c^1_{k,1}=K_1$, $m^1_{K_1,1}=1$,

where the abbreviation 'w.p.' stands for 'with probability'.

The sequence associated with $c_{1,t},\dots,c_{n,t}$ is exchangeable and induces a random partition of n, that is an unordered collection of $k \leq n$ positive integers with sum n or, equivalently, a random allocation of n unlabeled points into some random number of unlabeled clusters (materialized by a color for example); each cluster containing at least one point. A common way to represent a partition of n is by the number of terms of various sizes; that is the vector of counts $(a_1,\dots,a_n)$ where $\sum_{j=1}^n a_j = k$ and $\sum_{j=1}^n j a_j = n$. Here a_1 is the number of terms of size 1, a_2 is the number of terms of size 2, etc. Following (Antoniak, 1974), we say that $\mathbf{c}_t \in C(a_{1:n})$ if there are a_1 distinct values of $\mathbf{c}_t$ that occur only once, a_2 that occur twice, etc. It can be shown that $\Pr(\mathbf{c}_t \in C(a_{1:n})) = P_n(a_{1:n})$ is given by the two-parameter Ewens sampling formula (Pitman, 1995)

$$P_n(a_1,\dots,a_n) = \frac{n!\prod_{i=1}^{\sum_j a_j}(\theta+(i-1)\alpha)}{\prod_{i=1}^n(\theta+i-1)}\prod_{i=1}^n\left(\frac{\prod_{j=1}^{i-1}(j-\alpha)}{i!}\right)^{a_i}\frac{1}{a_i!}. \tag{9}$$

In this paper, we introduce a class of statistical models with dependencies between the distributions $\{F_t\}$ and mixing distributions $\{\mathbb{G}_t\}$ while preserving (1) and (2) at any time t. This allows us to explicitly characterize the marginal model at each time step (more generally, at each value of the covariate) and have fine control over prior parameterization of the marginal model. Methods that preserve stationarity in this way have found use in other Bayesian nonparametric models (Teh et al., 2011; Griffin and Steel, 2011), and allow us to achieve good empirical performance in epidemiological and video applications in Section 5. We briefly review constructions in the literature for dependent processes below.

2.1 Literature review

Several authors have considered previously the problem of defining dependent nonparametric models, and in particular dependent Dirichlet processes for time series and spatial models, see e.g. Foti and Williamson (2013) for a recent review.

In an early contribution, Cifarelli and Regazzini (1978) introduced dependencies between distributions by defining a parametric model on the base distribution $\mathbb{H}_s$ dependent on a covariate s and $\mathbb{G}_s \sim \mathrm{DP}(\theta,\mathbb{H}_s)$. This approach is different from ours as we follow here the setting introduced by MacEachern et al. (1999) and introduce dependencies directly on two successive mixing distributions while $\mathbb{H}$ is fixed.

The great majority of recent papers use the stick-breaking representation (5) to introduce dependencies. Under this representation, a realization of a Dirichlet process is represented by two (infinite dimensional) vectors of weights $V_{1:\infty,s}$ and cluster locations $U_{1:\infty,s}$. Dependency with respect to a covariate s is introduced on $V_{1:\infty,s}$ in (Griffin and Steel, 2006; Rodriguez and Dunson, 2011; Arbel et al., 2014), on $U_{1:\infty,s}$ in (MacEachern, 2000; Iorio et al., 2004; Gelfand et al., 2005; Caron et al., 2008) and on both cluster locations and weights in (Griffin and Steel, 2009).

An alternative approach consists of considering the mixing distribution to be a convex combination of independent random probability measures sampled from a Dirichlet process. The dependency is then introduced through some weighting coefficients; e.g. (Dunson et al., 2006; Dunson and Park, 2007; Müller et al., 2004). More recently, various classes of dependent Dirichlet processes were proposed by Griffin (2011), Rao and Teh (2009), Huggins and Wood (2014) and Lin et al. (2010) which rely on the representation of the DP as a normalized random measure.

Although these previous approaches have merits, we believe that it is possible to build more intuitive models in the time domain based on Pólya urn-type schemes. In (Zhu et al., 2005; Ahmed and Xing, 2008; Blei and Frazier, 2011), time-varying Pólya urn models were proposed but these models do not marginally preserve a Dirichlet process. The only model we know of which satisfies this property is presented in (Walker and Muliere, 2003). The authors define a joint distribution $p(\mathbb{G}_1, \mathbb{G}_2)$ such that $\mathbb{G}_1$ and $\mathbb{G}_2$ are marginally $\mathrm{DP}(\theta, \mathbb{H})$ by introducing m artificial auxiliary variables $\mathbf{w}_i \overset{\text{iid}}{\sim} \mathbb{G}_1$ and then $\mathbb{G}_2|\mathbf{w}_{1:m} \sim \mathrm{DP}(\theta + m, \frac{\theta\mathbb{H}+\sum_{i=1}^{m}\delta_{\mathbf{w}_i}}{\theta+m})$. An extension to time series is discussed in (Srebro and Roweis, 2005). An important drawback of this approach is that it requires introducing a very large number m of auxiliary variables to model strongly dependent distributions. When inference is performed, these auxiliary variables need to be inferred from the data and the resulting algorithm can be computationally intensive.

2.2 Contributions and Organization

The models developed here are based on a Pólya urn representation of the Pitman-Yor process but do not require introducing a large number of auxiliary variables to model strongly dependent distributions.

To obtain a first-order stationary Pitman-Yor process mixture using such an approach, we need to ensure that any time t

(A) the sequence $\mathbf{c}_t$ induces a random partition distributed according to the two-parameter Ewens sampling formula (9),

(B) for $j \in \mathbf{c}_t$, the $U_{j,t}$s are identically and independently distributed from $\mathbb{H}$.

The main contribution of this paper consists of defining models satisfying (**A**) using a generalized Pólya urn prediction rule based on the consistence properties under specific deletion procedures of the Ewens sampling formula (Kingman, 1978; Gnedin and Pitman, 2005). Ensuring (**B**) can be performed using standard methods from the time series literature; e.g. (Joe, 1997; Pitt and Walker, 2005).

Our models allow us to modify both the cluster locations and their weights. Furthermore, they rely on simple and intuitive birth and death procedures. By using a Pólya urn approach, the models are defined on the space of partitions; i.e. the labelling of the class to which each data belongs is irrelevant. From a computational point of view, it is usually easier to design efficient MCMC and SMC algorithms for inference based on this finite-dimensional representation than the infinite-dimensional stick-breaking representation, where slice sampling (Walker, 2007; Kalli et al., 2011) or retrospective sampling (Papaspiliopoulos and Roberts, 2008) techniques can be used.

3. Stationary Pitman-Yor Process Mixtures

We present here some models ensuring property (**A**) is satisfied. We then briefly discuss in Section 3.2 how to ensure (**B**).

3.1 Stationary Pitman-Yor Processes

The main idea behind our models consists at each time step t of

• deleting randomly a subset of the allocations variables sampled from time 1 to $t-1$ which had survived the previous $t-1$ deletion steps,

• sampling n new allocation variables corresponding to the n observations $\mathbf{z}_t$.

For any $t \geq 2$, we have generated the allocation variables $\mathbf{c}_{1:t-1}$ corresponding to $\mathbf{z}_{1:t-1}$ from time 1 to $t-1$. We denote by $\mathbf{c}_{1:t-1}^{t-1}$ the subset of $\mathbf{c}_{1:t-1}$ corresponding to variables having survived the deletion steps from time 1 to $t-1$, and we denote by $\mathbf{c}_{1:t-1}^{t}$ the subset corresponding to those having survived from time 1 to t. Let K_{t-1} be the number of clusters created from time 1 to $t-1$. We denote by $\mathbf{m}_{t-1}^{t-1}$ the vector of size K_{t-1} containing the size of the clusters associated to $\mathbf{c}_{1:t-1}^{t-1}$, and we denote by $\mathbf{m}_{t-1}^{t}$ the vector containing the size of clusters associated to $\mathbf{c}_{1:t-1}^{t}$. Hence, these vectors have zero entries corresponding to 'dead' clusters. The introduction of $\mathbf{m}_{t-1}^{t-1}$ and $\mathbf{m}_{t-1}^{t}$ simplifies the presentation of the procedure but note that, from a practical point of view, there is obviously no need to store these vectors of increasing dimension. It is only necessary to store the size of the non-empty clusters and their associated labels.

At time 1, we just generate $\mathbf{c}_1$ according to a standard Pólya urn described in the introduction. At time $t \geq 2$ we have $\mathbf{c}_{1:t-1}^{t-1} = \left(\mathbf{c}_{1:t-2}^{t-1}, \mathbf{c}_{t-1}\right)$ and we sample $\mathbf{c}_{1:t}^{t} = \left(\mathbf{c}_{1:t-1}^{t}, \mathbf{c}_{t}\right)$ as follows. We first obtain $\mathbf{c}_{1:t-1}^{t}$ by deleting a random number of allocation variables from $\mathbf{c}_{1:t-1}^{t-1}$ according to one of the following rules.

• **Uniform deletion**: delete each allocation variable in $\mathbf{c}_{1:t-1}^{t-1}$ with probability $1-\rho$ where $0 \leq \rho \leq 1$. This is statistically equivalent to sampling a number r from a binomial distribution $\mathcal{B}\text{in}(\sum_k m_{k,t-1}^{t-1}, 1-\rho)$ and then removing r items uniformly from $\mathbf{c}_{1:t-1}^{t-1}$ to obtain $\mathbf{c}_{1:t-1}^{t}$. For $\rho = 0$, the partitions $\mathbf{c}_t$ and $\mathbf{c}_{t+1}$ are independent whereas for $\rho = 1$ we have a static Pitman-Yor process.

• **Deterministic deletion**: delete the allocation variables $\mathbf{c}_{t-r}$ from $\mathbf{c}_{1:t-1}^{t-1}$, where $r \in \mathbb{N}$ and $r < t$.

• **Cluster deletion** (for α, θ verifying condition (3) with $\theta \geq 0$): compute the following discrete probability distribution over the set of non-empty clusters

$$\pi_{k,t} = \frac{(\sum_\ell m_{\ell,t-1}^{t-1} - m_{k,t-1}^{t-1})\gamma + m_{k,t-1}^{t-1}(1-\gamma)}{(\sum_\ell m_{\ell,t-1}^{t-1})(1-\gamma+(K_{t-1}-1)\gamma)}$$

where $\sum_k \pi_{k,t} = 1$, $\gamma = \frac{\alpha}{\alpha+\theta}$ then sample an index from this distribution and delete the corresponding cluster to obtain $\mathbf{c}_{1:t-1}^{t}$. The cluster deletion allows us to model large potential 'jumps' in the distributions of the observations. Dependent on the value of γ, we

delete clusters with a probability independent of their size, proportional to their size or proportional to the size of the partition minus their size (Gnedin and Pitman, 2005). The introduction of the Pitman-Yor process gives us this extra modeling flexibility here. In particular, the following three cases are of special interest:

Size-biased deletion For $\alpha = 0$ and $\theta > 0$, and hence $\gamma = 0$, each cluster of size r is selected with probability proportional to r. This result is known as Kingman's characterization of the Ewens family of $(0, \theta)$ partition structures (Kingman, 1978).

Unbiased (uniform) deletion For $\gamma = \frac{\alpha}{\alpha+\theta} = \frac{1}{2}$, given that there are l clusters, each cluster is chosen with probability $\frac{1}{l}$ i.e., each cluster is deleted independently of its size. For $0 \leq \alpha \leq 1$, the (α, α) partition structures are the only partition structures invariant under uniform deletion.

Cosize-biased deletion For $\gamma = 1$ (hence $\theta = 0$), each cluster of size r is selected with probability proportional to the size $n - r$ of the remaining partition.

It is also possible to consider any mixture and composition of these deletion procedures. For example, we can pick w.p. ξ the uniform deletion strategy and w.p. $1 - \xi$ the cluster deletion strategy or perform one uniform deletion followed by one cluster deletion or a deterministic deletion etc. Finally, after these deletion steps, we sample the allocation variables $\mathbf{c}_t$ according to a standard Pólya urn scheme based on the surviving allocation variables $\mathbf{c}^t_{1:t-1}$.

To summarize, the generalized Pólya urn scheme proceeds as in Algorithm 1, where $\mathcal{I}\left(\mathbf{m}^t_t\right)$ and $\left|\mathcal{I}\left(\mathbf{m}^t_t\right)\right|$ denote respectively the indices corresponding to the non-zero entries of $\mathbf{m}^t_t$ and the number of non-zero entries.

Algorithm 1 Generalized Pólya Urn

At time $t = 1$

- Set $c^1_{1,1} = 1$, $m^1_{1,1} = 1$ and $K_1 = 1$.
- For $k = 2, ..., n$

w.p. $\frac{m^1_{i,1}-\alpha}{k-1+\theta}$, $i \in \{1, \ldots, K_1\}$ set $c^1_{k,1} = i$, $m^1_{i,1} = m^1_{i,1} + 1$,
w.p. $\frac{K_1\alpha+\theta}{k-1+\theta}$, set $K_1 = K_1 + 1$, $c^1_{k,1} = K_1$, $m^1_{K_1,1} = 1$.

At time $t \geq 2$

- Kill a subset of $\mathbf{c}^{t-1}_{1:t-1}$ using a mixture/composition of uniform, size-biased and deterministic deletions to obtain $\mathbf{c}^t_{1:t-1}$ (hence $\mathbf{m}^t_{t-1}$) and set $\mathbf{m}^t_t = \mathbf{m}^t_{t-1}$, $K_t = K_{t-1}$.
- For $k = 1, ..., n$

w.p. $\frac{m^t_{i,t}-\alpha}{\sum_i m^t_{i,t}+\theta}$, $i \in \mathcal{I}\left(\mathbf{m}^t_t\right)$ set $c^t_{k,t} = i$, $m^t_{i,t} = m^t_{i,t} + 1$,
w.p. $\frac{\left|\mathcal{I}\left(\mathbf{m}^t_t\right)\right|\alpha+\theta}{\sum_i m^t_{i,t}+\theta}$, set $K_t = K_t + 1$, $c^t_{k,t} = K_t$, $m^t_{K_t,t} = 1$.

The proof that $\mathbf{c}_t$ satisfies $(\mathbf{A})$ is a direct consequence of the remarkable consistence properties under deletion of the Ewens sampling formula which have been first established in (Kingman, 1978) for the one-parameter case and then extended to the two-parameter case in (Gnedin and Pitman, 2005).

Proposition. At any time $t \geq 1$, $\mathbf{c}_t$ induces a random partition distributed according to the Ewens sampling formula.

Proof. We prove by induction a stronger result; that is $\mathbf{c}_{1:t}^t$ induces a random partition following (9). At time 1, this is trivially true as $\mathbf{c}_1^1 = \mathbf{c}_1$ is generated according to a standard Pólya urn. Assume it is true at time $t-1$. For the deterministic and uniform deletion, exchangeability of the partition ensures that $\mathbf{c}_{1:t-1}^t$ also induces a random partition following (9). For the cluster deletion procedure, consistency follows from the results of (Kingman, 1978, pp. 3 and 5) and (Gnedin and Pitman, 2005, p. 4) on regenerative partitions. Finally, as $\mathbf{c}_t$ is sampled according to a standard Pólya urn scheme based on the surviving allocation variables $\mathbf{c}_{1:t-1}^t$ then $\mathbf{c}_{1:t}^t$ indeed induces by construction a random partition following (9). Thanks to exchangeability, it implies that $\mathbf{c}_t$ also induces a random partition distributed according to the Ewens sampling formula.■

So far, so as to simplify presentation, we have considered that the number of allocation variables at each time t is fixed to a value n corresponding to the number of observations received at each time instant. The value of n impacts on the number of alive allocation variables and the correlations between successive vectors $\mathbf{m}_{1:t}^t$. More precisely, the statistical model is not consistent in the Kolmogorov sense. For example, let $\pi_1(c_{1,1}, c_{1,2})$ and $\pi_2(c_{1,1}, c_{2,1}, c_{1,2}, c_{2,2})$ be two models defined for $n = 1$ and $n = 2$, then

$$\left(\sum_{c_{2,1}} \sum_{c_{2,2}} \pi_2(c_{1,1}, c_{2,1}, c_{1,2}, c_{2,2}) \right) \neq \pi_1(c_{1,1}, c_{1,2}).$$

This lack of consistency is shared by other models based on the Pólya urn construction (Zhu et al., 2005; Ahmed and Xing, 2008; Blei and Frazier, 2011). Blei and Frazier (2011) provide a detailed discussion on this issue and describe cases where this property is relevant or not.

It is nonetheless possible to define a slightly modified version of our model that is consistent under marginalisation, at the expense of an additional set of latent variables. This is described in Appendix C.

3.2 Stationary Models for Cluster Locations

To ensure we obtain a first-order stationary Pitman-Yor process mixture model, we also need to satisfy (**B**). This can be easily achieved if for $k \in \mathcal{I}(\mathbf{m}_t^t)$

$$U_{k,t} \sim \begin{cases} p\left(\cdot | U_{k,t-1}\right) & \text{if } k \in \mathcal{I}(\mathbf{m}_{t-1}^t) \\ \mathbb{H} & \text{otherwise} \end{cases}$$

where $\mathbb{H}$ is the invariant distribution of the Markov transition kernel $p\left(\cdot|\cdot\right)$. In the time series literature, many approaches are available to build such transition kernels based on copulas (Joe, 1997) or Gibbs sampling techniques (Pitt and Walker, 2005).

Combining the stationary Pitman-Yor and cluster locations models, we can summarize the full model by the following Bayesian network in Figure 1. It can also be summarized using a Chinese restaurant metaphor (see Figure 2).

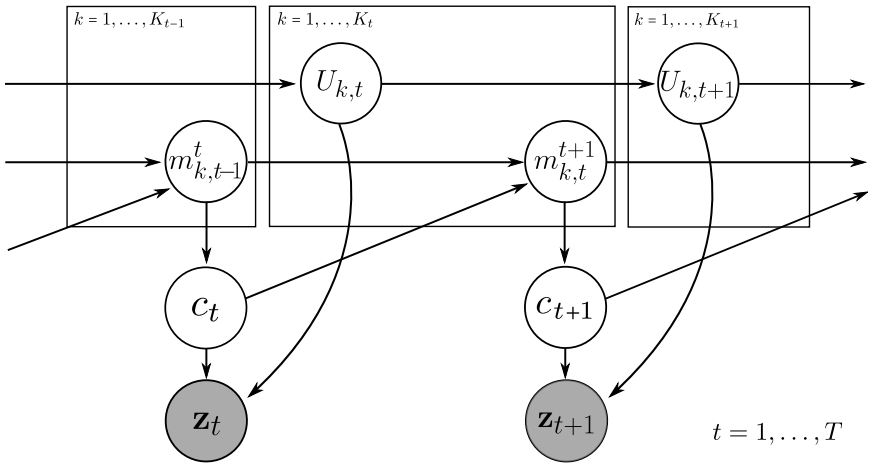


Figure 1: A representation of the time-varying Pitman-Yor process mixture as a directed graphical model, representing conditional independencies between variables. All assignment variables and observations at time t are denoted c_t and $\mathbf{z}_t$, respectively.

3.3 Properties of the Models

Under the uniform deletion model, the number $A_t = \sum_i m^t_{i,t-1}$ of alive allocation variables at time t can be written as

$$A_t = \sum_{j=1}^{t-1} \sum_{k=1}^{n} X_{j,k}$$

where $X_{j,i}$ are independently distributed from a Bernoulli of parameter ρ^j. A_t is therefore distributed from a Poisson binomial (also called Pólya Frequency) distribution (Pitman, 1997). The asymptotic mean and variance of the distribution of A_t are respectively $\frac{n\rho}{1-\rho}$ and $\frac{n\rho}{1-\rho^2}$, and the distribution $\Pr(A_t = k) \propto a_{k,t-1}$ where $a_{k,t}$, $k = 0, \ldots, nt$ satisfies the algebraic identity

$$\prod_{i=1}^{t} \left(x + \frac{1-\rho^i}{\rho^i} \right)^n = \sum_{k=0}^{nt} a_{k,t} x^k. \tag{10}$$

Its stationary distribution is obtained by taking the limit as $t \to \infty$.

Clearly the sequence $\{\mathbf{c}_t\}$ is not Markovian but $\left\{\mathbf{c}^t_{1:t}\right\}$ and $\left\{\mathbf{c}^t_{1:t-1}\right\}$ and the associated vectors $\left\{\mathbf{m}^t_{1:t}\right\}$ and $\left\{\mathbf{m}^t_{1:t-1}\right\}$ are for the uniform and cluster deletions. The transition probabilities for these processes are quite complex. However it can be shown easily for example that, for the uniform deletion model, we have for $k \in \mathcal{I}(\mathbf{m}^t_{t-1})$

$$\begin{aligned}
\mathbb{E}\left[m^{t+1}_{k,t} | \mathbf{m}^t_{t-1} \right] &= \mathbb{E}\left[\mathbb{E}\left[m^{t+1}_{k,t} \middle| m^t_{k,t} \right] | \mathbf{m}^t_{t-1} \right] \\
&= \rho \mathbb{E}\left[m^t_{k,t} | \mathbf{m}^t_{t-1} \right] \\
&= \rho \left(m^t_{k,t-1} + n \frac{m^t_{k,t-1} - \left| \mathcal{I}\left(\mathbf{m}^t_{t-1}\right) \right| \alpha}{\theta + \sum_k m^t_{k,t-1}} \right)
\end{aligned}$$

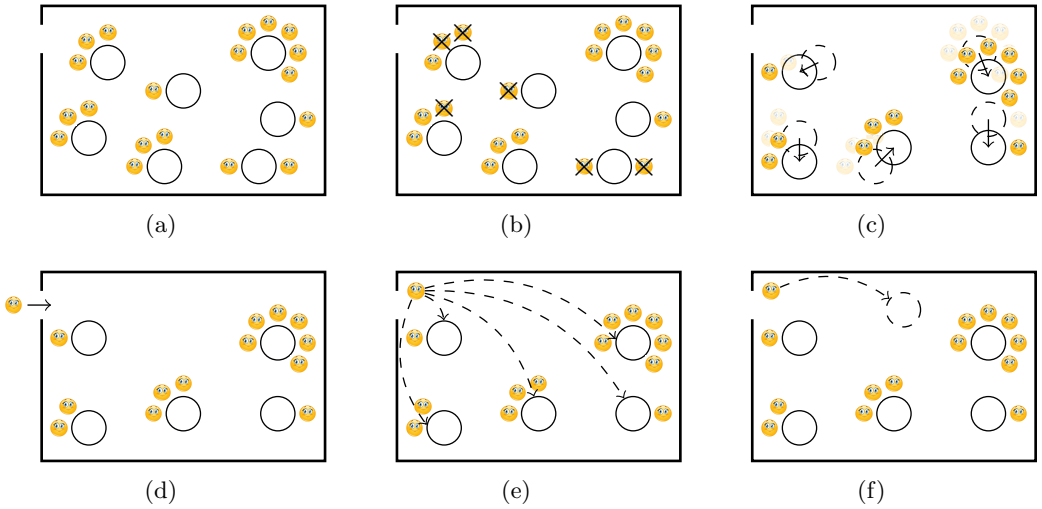


Figure 2: Illustration of the uniform deletion time-varying Pitman-Yor process mixture. Consider a restaurant with a countably infinite number of tables. (a) At time t, there are a certain number of customers in the restaurant, shared between several tables. Each customer remains seated at her/his table (w.p. ρ), or leaves the restaurant (w.p. $(1-\rho)$). (b) Once this choice has been made by each customer, empty tables are removed, and a certain number of customers remain in the restaurant. (c) Each table that is still occupied has its location evolved according to the transition kernel $p(\cdot|\cdot)$. (d) A new customer enters the restaurant and either (e) sits at a table with a probability proportional to the number of people at this table or (f) sits alone at a new table whose location has distribution $\mathbb{H}$. When $n-1$ other new customers enter the restaurant, repeat operations (d)-(f).

and

$$\mathbb{E}[\sum_{k\notin\mathcal{I}(\mathbf{m}_{t-1}^t)} m_{k,t}^{t+1}|\mathbf{m}_{t-1}^t] = \frac{\rho n\left[\left|\mathcal{I}\left(\mathbf{m}_{t-1}^t\right)\right|\alpha+\theta\right]}{\theta+\sum_k m_{k,t-1}^t}.$$

For any deletion model, it can be additionally shown that we have, conditional on $\mathbf{m}_{t-1}^t$ (Pitman, 1996)

$$\mathbb{G}_t = \sum_{j\in\mathcal{I}(\mathbf{m}_{t-1}^t)} P_j U_{j,t} + R_t\mathbb{G}_t^\star$$

where $\mathbb{G}_t^\star \sim \mathcal{PY}(\alpha,\theta+\alpha\left|\mathcal{I}\left(\mathbf{m}_{t-1}^t\right)\right|,\mathbb{H})$ and

$$\left(\{P_j|j\in\mathcal{I}(\mathbf{m}_{t-1}^t)\},R_t\right)|\mathbf{m}_{t-1}^t \sim \mathcal{D}(\{m_{j,t-1}^t-\alpha|j\in\mathcal{I}(\mathbf{m}_{t-1}^t)\},\left|\mathcal{I}(\mathbf{m}_{t-1}^t)\right|\alpha+\theta)$$

where $\mathcal{D}$ is the standard Dirichlet distribution. Moreover, $\mathbb{G}_t^\star$ and $\left(\{P_j|j\in\mathcal{I}(\mathbf{m}_{t-1}^t)\},R_t\right)$ are statistically independent. In particular, in the Dirichlet process case, we have

$$\mathbb{G}_t|\mathbf{c}_{1:t-1}^t,U_{j,t} \sim DP\left(\theta+\left|\mathcal{I}\left(\mathbf{m}_{t-1}^t\right)\right|,\frac{1}{\theta+\left|\mathcal{I}\left(\mathbf{m}_{t-1}^t\right)\right|}(\theta\mathbb{H}+\sum_k m_{k,t}^{t-1}\delta_{U_{k,t}})\right).$$

We display a Monte Carlo estimate of $\mathrm{corr}\big(\int \mathbf{y}\mathbb{G}_t(d\mathbf{y}), \int \mathbf{y}\mathbb{G}_{t+\tau}(d\mathbf{y})\big)$ when $\mathbb{H}(\mathbf{y})$ is a standard normal distribution for different values of ρ, ξ and r resp. for the uniform, size-biased and deterministic deletions, and θ (with $\alpha = 0$) in Fig. 3. The correlations decrease faster as ρ, r and ξ decrease as expected.

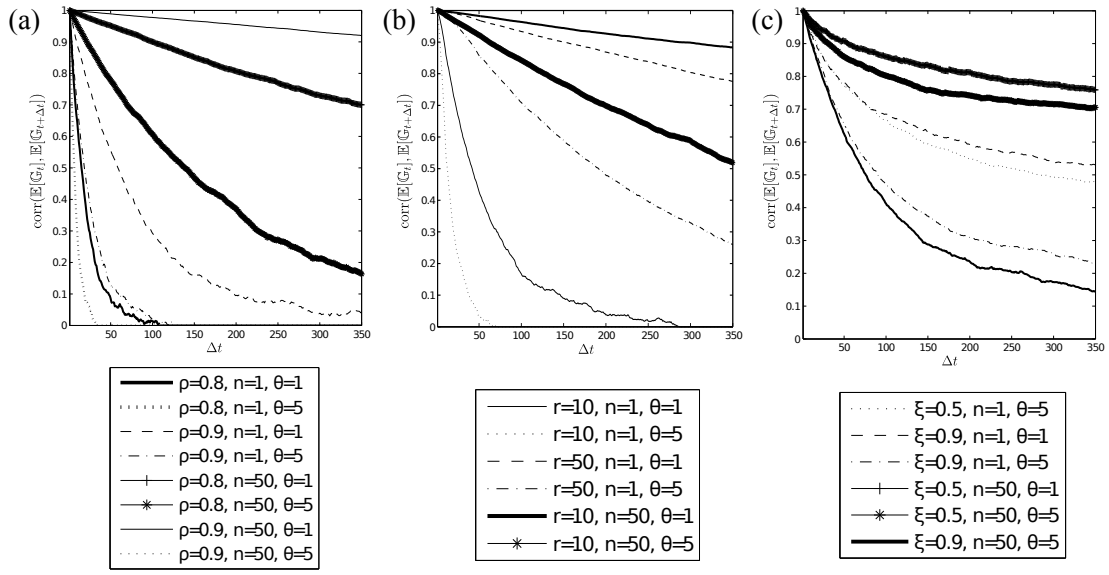


Figure 3: Plots of $\mathrm{corr}\big(\int \mathbf{y}\mathbb{G}_t(d\mathbf{y}), \int \mathbf{y}\mathbb{G}_{t+\Delta t}(d\mathbf{y})\big)$ approximated by Monte Carlo simulations as a function of the time difference Δt for (a) uniform (b) deterministic and (c) cluster deletions, for different values of θ and (a) ρ, (b) r and (c) ξ.

4. Bayesian Inference in Time-Varying PYPM

Bayesian inference is based on the posterior distribution of the cluster assignment variables $\mathbf{c}_t$, the vectors $\mathbf{m}_{t-1}^t$ and $U_{1:K_t}$ given by $p(\mathbf{c}_{1:t}, \mathbf{m}_{1:t-1}^{1:t}, U_{1:K_t}|\mathbf{z}_{1:t})$ at time t. We describe sequential Monte Carlo methods to fit our models.

To sample approximately from the sequence of distributions $p(\mathbf{c}_{1:t}, \mathbf{m}_{1:t-1}^{1:t}, U_{1:K_t}|\mathbf{z}_{1:t})$ as t increases, we propose an SMC method also known as particle filter. In this approach, the posterior distribution is approximated by a large collection of random samples—termed particles—which are propagated through time using Sequential Importance Sampling with resampling steps; see (Doucet et al., 2001) for a review of the literature. MacEachern et al. (1999) developed a sequential importance sampling method for Beta-Binomial Dirichlet process mixtures and Fearnhead (2004) proposed a SMC method for general conjugate Dirichlet process mixtures. In these papers, even in cases where the mixed distribution $\mathbb{G}$ can be integrated out analytically, the fact that $\mathbb{G}$ is a static parameter leads to an accumulation of the Monte Carlo errors over time (Kantas et al., 2009; Poyiadjis et al.,

2011). This prevents using such techniques for large datasets with long time sequences. We deal here with time-varying models whose forgetting properties limit drastically this accumulation of errors.

We present in Algorithm 2 a general SMC algorithm. It relies on importance distributions denoted generically by $q(\cdot)$ and is initialized with weights $w_0^{(i)} = N^{-1}$ and cluster sizes $m_0^{0\ (i)} = 0$ for $i = 1, ..., N$. The simplest case of this algorithm consists of selecting the prior as an importance distribution. Note that we use $\overline{A}$ to denote the complementary of the set $A \subset \mathbb{N}$ in $\mathbb{N}$.

Algorithm 2 Sequential Monte Carlo for Time-Varying Pitman-Yor mixture processes

At time $t \geq 1$

- For each particle $i = 1, .., N$
  - Sample $\widetilde{\mathbf{m}}_{t-1}^{t\ (i)} | \mathbf{m}_{t-1}^{t-1\ (i)} \sim \Pr(\mathbf{m}_{t-1}^{t} | \mathbf{m}_{t-1}^{t-1\ (i)})$
  - Sample $\widetilde{\mathbf{c}}_t^{(i)} \sim q(\mathbf{c}_t | \widetilde{\mathbf{m}}_{t-1}^{t\ (i)}, U^{(i)}_{\mathcal{I}(\widetilde{\mathbf{m}}_t^{(i)}),t-1}, z_t)$
  - For $k \in \mathcal{J}(\widetilde{\mathbf{c}}_t^{(i)}) \cap \overline{\mathcal{I}(\widetilde{\mathbf{m}}_{t-1}^{t\ (i)})}$, sample $\widetilde{U}_{k,t}^{(i)} \sim q(U_{k,t} | \{\mathbf{z}_{j,t} | \widetilde{c}_{j,t}^{(i)} = k\})$
  - For $k \in \mathcal{I}(\widetilde{\mathbf{m}}_{t-1}^{t\ (i)})$, sample $\widetilde{U}_{k,t}^{(i)} \sim \begin{cases} q(U_{k,t} | U_{k,t-1}^{(i)}, \{\mathbf{z}_{j,t} | \widetilde{c}_{j,t}^{(i)} = k\}) & \text{if } k \in \mathcal{J}(\widetilde{\mathbf{c}}_t^{(i)}) \\ p(U_{k,t} | U_{k,t-1}^{(i)}) & \text{otherwise} \end{cases}$
- For $i = 1, .., N$, update the weights

$$\begin{aligned} \widetilde{w}_t^{(i)} \quad \propto w_{t-1}^{(i)} & \frac{p(\mathbf{z}_t | \widetilde{U}_t^{(i)}, \widetilde{\mathbf{c}}_t^{(i)}) \Pr(\widetilde{\mathbf{c}}_t^{(i)} | \widetilde{\mathbf{m}}_{t-1}^{t\ (i)})}{q(\widetilde{\mathbf{c}}_t^{(i)} | \widetilde{\mathbf{m}}_{t-1}^{t\ (i)}, U^{(i)}_{\mathcal{I}(\widetilde{\mathbf{m}}_t^{(i)}),t-1}, \mathbf{z}_t)} \times \prod_{k \in \mathcal{I}(\widetilde{\mathbf{m}}_{t-1}^{t\ (i)})} \frac{p(\widetilde{U}_{k,t}^{(i)} | U_{k,t-1}^{(i)})}{q(\widetilde{U}_{k,t}^{(i)} | U_{k,t-1}^{(i)}, \{\mathbf{z}_{j,t} | \widetilde{c}_{j,t}^{(i)} = k\})} \\ & \times \prod_{k \in \mathcal{J}(\widetilde{\mathbf{c}}_t^{(i)}) \cap \overline{\mathcal{I}(\widetilde{\mathbf{m}}_{t-1}^{t\ (i)})}} \frac{\mathbb{H}(\widetilde{U}_{k,t}^{(i)})}{q(\widetilde{U}_{k,t}^{(i)} | \{\mathbf{z}_{j,t} | \widetilde{c}_{j,t}^{(i)} = k\})} \end{aligned} \tag{11}$$

  with $\sum_{i=1}^N \widetilde{w}_t^{(i)} = 1$.
- Resampling. Compute $N_{\text{eff}} = \left[\sum \left(\widetilde{w}_t^{(i)} \right)^2 \right]^{-1}$. If $N_{\text{eff}} \leq N_T$, multiply the particles with large weights and remove the particles with small weights, resulting in a new set of particles denoted $\cdot_t^{(i)}$ with weights $w_t^{(i)} = 1/N$. Otherwise, rename the particles and weights by removing the $\widetilde{\cdot}$.

In this algorithm, N_T is a threshold triggering the resampling step; typically we set $N_T = N/2$. The posterior $p(\mathbf{c}_{1:t}, \mathbf{m}_{1:t-1}^{1:t}, U_{1:K_t} | \mathbf{z}_{1:t})$ is approximated using the set of weighted samples $\left\{ w_t^{(i)}, \left(\mathbf{c}_{1:t}^{(i)}, \mathbf{m}_{1:t-1}^{1:t\ (i)}, U_{1:K_t}^{(i)} \right) \right\}$. In cases where $\mathbb{H}$ is a conjugate prior for f, we can integrate out the cluster locations and sample only the allocation variables.

We also develop a particle MCMC (PMCMC) inference algorithm (Andrieu et al., 2010) for this model. PMCMC is a method that uses SMC as an intermediate sampling step to move efficiently through high dimensional state spaces. It is an alternative to other common forms of MCMC, such as single-site Gibbs sampling, which we have found suffer from worse empirical performance due to quasi-ergodicity (i.e. poor mixing as the Markov chain becomes stuck in posterior modes). We implement the iterated-conditional SMC form of PMCMC (Andrieu et al., 2009, 2010), which involves iterating through the SMC algorithm described above while conditioning on a sampled particle at each iteration. While

this algorithm does not operate in an online fashion, as SMC does, we show in Section 5 that it yields improved performance in tasks such as video object tracking (shown in Section 5).

An alternative approach consists of modelling the observations as follows (we discuss here the deterministic deletion model but the uniform deletion can be defined similarly). In this model, the predictive distribution of $\mathbf{z}_t$ conditional on $c_t = k$ only depends on the observations assigned to the same cluster in the time interval $\{t-r+1, ..., t-1\}$ and

$$\begin{aligned} p(\mathbf{z}_t|c_t = k, \mathbf{z}_{1:t-1}, c_{1:t-1}) &= p(\mathbf{z}_t|\{\mathbf{z}_j : c_j = k \text{ for } j = t-r+1, \dots, t-1\}) \\ &= \frac{\int_{\mathcal{Y}} f(\mathbf{z}_t|\mathbf{y}) \prod_{\{j\in\{t-r+1,\dots t-1\}:c_j=k\}} f(\mathbf{z}_j|\mathbf{y}) d\mathbb{H}(\mathbf{y})}{\int_{\mathcal{Y}} \prod_{\{j\in\{t-r+1,\dots t-1\}:c_j=k\}} f(\mathbf{z}_j|\mathbf{y}) d\mathbb{H}(\mathbf{y})}. \end{aligned} \tag{12}$$

This distribution can be computed in closed-form in the conjugate case. We propose the simple deterministic Algorithm 3 to approximate the posterior in the context of this model. The posterior $p(\mathbf{c}_{1:t}|\mathbf{z}_{1:t})$ is approximated using the set of weighted samples $\left\{w_t^{(i)}, \mathbf{c}_{1:t}^{(i)}\right\}$. This algorithm uses a deterministic selection step. Alternatively, we could have used the stochastic resampling procedure of Fearnhead (2004).

Algorithm 3 $N-$best algorithm for Deterministic Deletion

At time $t \geq 1$

- Set $w_{0,t}^{(i)} = w_{n,t-1}^{(i)}$
- For $k = 1, \dots, n$
  - For each particle $i = 1, .., N$
    - For $j \in \mathcal{J}(\{c_{1:k-1,t}^{(i)}, c_{1:t-1}^{(i)}\}) \cup \{0\} \cup \{c_{new}\}$, let $\widetilde{c}_{k,t}^{(i,j)} = j$ and compute the weight

$$\widetilde{w}_{k,t}^{(i,j)} \propto w_{k-1,t}^{(i)} p\left(\mathbf{z}_{k,t}|\, \mathbf{z}_{1:k-1,t}, \mathbf{z}_{1:t-1}, c_{1:k-1,t}^{(i)}, c_{1:t-1}^{(i)}, \widetilde{c}_{k,t}^{(i,j)}\right) p\left(\widetilde{c}_{k,t}^{(i,j)} \middle|\, c_{1:k-1,t}^{(i)}, c_{1:t-1}^{(i)}\right) \tag{13}$$

  - Keep the N particles $\left(c_{1:k-1,t}^{(i)}, c_{1:t-1}^{(i)}, \widetilde{c}_{k,t}^{(i,j)}\right)$ with highest weights $\widetilde{w}_{k,t}^{(i,j)}$, rename them $\left(c_{1:k,t}^{(i)}, c_{1:t-1}^{(i)}\right)$ and denote $w_{k,t}^{(i)}$ the associated weights.

5. Applications

In this Section, we demonstrate the models and algorithms on simulated data, modeling of the spread of a disease, and multi-object tracking.

5.1 Sequential Time-Varying Density Estimation

We consider the synthetic problem of estimating sequentially time-varying densities F_t on the real line using observations $\mathbf{z}_t$. We assume the observations $\mathbf{z}_t$ (where $n = 1$) follow a time-varying Dirichlet process with both uniform and size-biased deletion, a Gaussian mixed density and normal-inverse Wishart base distribution, whose pdf is given by

$$p(\mu, \Sigma|\mu_0, \kappa_0, \nu_0, \Lambda_0) \propto |\Sigma|^{-\frac{\nu_0+p+1}{2}} \exp\left[-\frac{1}{2} tr(\Lambda_0\Sigma^{-1} - \frac{\kappa_0}{2}(\mu-\mu_0)^T\Sigma^{-1}(\mu-\mu_0))\right]$$

where $p = 1$ in this case. To keep the presentation simple, we assume here that the hyperparameters of the base distribution are assumed fixed and known $\mu_0 = 0$, $\kappa_0 = 0.1$, $\nu_0 = 2$ and $\Lambda_0 = 1$. Following Pitt et al. (2002), we introduce a time-varying model on the cluster location using r auxiliary variables $\omega_{j,t}$, $j = 1, \ldots, r$

$$\omega_{j,t} \sim \mathcal{N}\left(\mu_t, \frac{\Sigma_t}{\tau}\right) \text{ for all } j$$
$$(\mu_{t+1}, \Sigma_{t+1})|\omega_{1:r,t} \sim \mathcal{N}iW(\mu_0', \kappa_0', \nu_0', \Lambda_0')$$

where $\tau = 0.5$ and $r = 4000$ are fixed parameters, $\kappa_0' = \kappa_0 + r\tau$, $\nu_0' = \nu_0 + r$, $\mu_0' = \frac{r\tau}{\kappa_0 + r\tau}\overline{w} + \frac{\kappa_0}{\kappa_0 + r\tau}\mu_0$ and $\Lambda_0' = \Lambda_0 + \frac{\kappa_0 r\tau}{\kappa_0 + r\tau}(\overline{\omega} - \mu_0)(\overline{\omega} - \mu_0)^T + \tau\sum_{j=1}^n(\omega_{j,t} - \overline{\omega})(\omega_{j,t} - \overline{\omega})^T$ and $\overline{\omega} = \frac{1}{n}\sum_{j=1}^r \omega_{j,t}$. The DP scale parameter is $\theta = 3$. Instead of fixing ρ, we assume it is time-varying with $\rho_t|\rho_{t-1} \sim \mathcal{B}(a_\rho, a_\rho\frac{1-\rho_{t-1}}{\rho_{t-1}})$ where $a_\rho = 1000$, such that $\mathbb{E}[\rho_t|\rho_{t-1}] = \rho_{t-1}$ and $var(\rho_t|\rho_{t-1}) = \frac{\rho_{t-1}^2(1-\rho_{t-1})}{a_\rho + \rho_{t-1}}$. Note that the resulting model is still first-order stationary. We select a mixture of uniform and cluster deletions with $\xi = 0.98$. The observations $\mathbf{z}_t$ are generated for $t = 1, \ldots, 1000$ from a sequence of mixtures of normal distributions, see Figure 4. Abrupt changes occur at times $t = 301$ and $t = 601$ where modes of the true density appear/disappear whereas the mode moves smoothly from 0 to -1.5 between $t = 701$ and $t = 850$. For illustration purposes, we compute the average number of alive allocation variables $N_{t|t}$ as follows

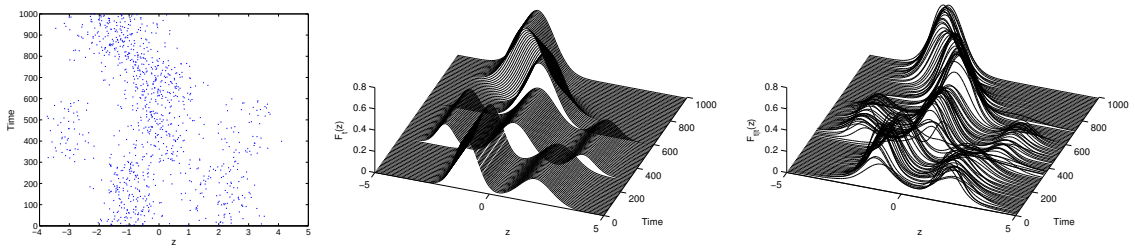
$$N_{t|t} = \mathbb{E}\left[\sum_{k=1}^{K_t} m_{k,t}^t \middle| \mathbf{z}_{1:t}\right]. \tag{14}$$

A SMC algorithm is implemented with 1000 particles (Doucet et al., 2001); this algorithm is a slight generalization of the Algorithm 2 where ρ_t is also sampled. Results are displayed in Figure 4 and the estimates of $N_{t|t}$ and $\rho_{t|t} = \mathbb{E}\left[\rho_t|\,\mathbf{z}_{1:t}\right]$ are plotted in Figure 5.

In Figure 4, we display the filtered density estimate $F_{t|t} = \mathbb{E}\left[F_t|\mathbf{z}_{1:t}\right]$ which manages to track the slow and abrupt changes of the true density. In Figure 5, we see that the model adapts to F_t quickly by also estimating ρ_t: ρ_t suddenly decreases at times $t = 300$ where the modes of the density suddenly change. $N_{t|t}$ follows a similar evolution: whenever F_t does not evolve, we use as many previously collected observations as possible to estimate the density by letting $N_{t|t}$ increases. When F_t changes abruptly, $N_{t|t}$ also decreases abruptly and the model quickly gets rid of the old clusters. Moreover, the cluster deletion procedure allows us to only delete irrelevant clusters. This is illustrated at $t = 600$ where the two minor modes disappear whereas the main mode is preserved.

5.2 Foot and Mouth Disease

Foot and Mouth disease is a highly transmissible viral infection which can spread rapidly among livestock. The epidemic which started in February 2001 and ended in October 2001 saw 2000 cases of death in farms throughout England. Although complex models have been designed in the epidemiology literature relying for example on the spatial distribution of farms (Keeling et al., 2001), we propose to use our model to estimate clusters of cases of Foot and Mouth disease. Let $\mathbf{z}_{k,t}$ be the two-dimensional geographical position of case k



(a) Data (b) True density (c) Estimated density

Figure 4: (a) Data (b) True density and (c) filtered density estimate. Abrupt changes occur at times $t = 301$ and $t = 601$. The mode of the density evolves smoothly between times $t = 700$ and $t = 850$.

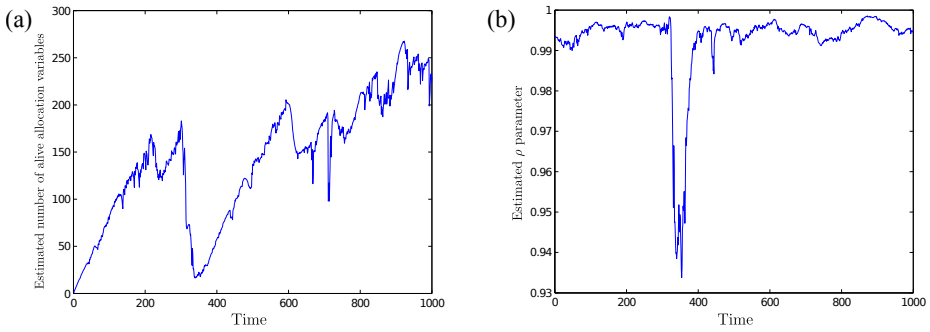
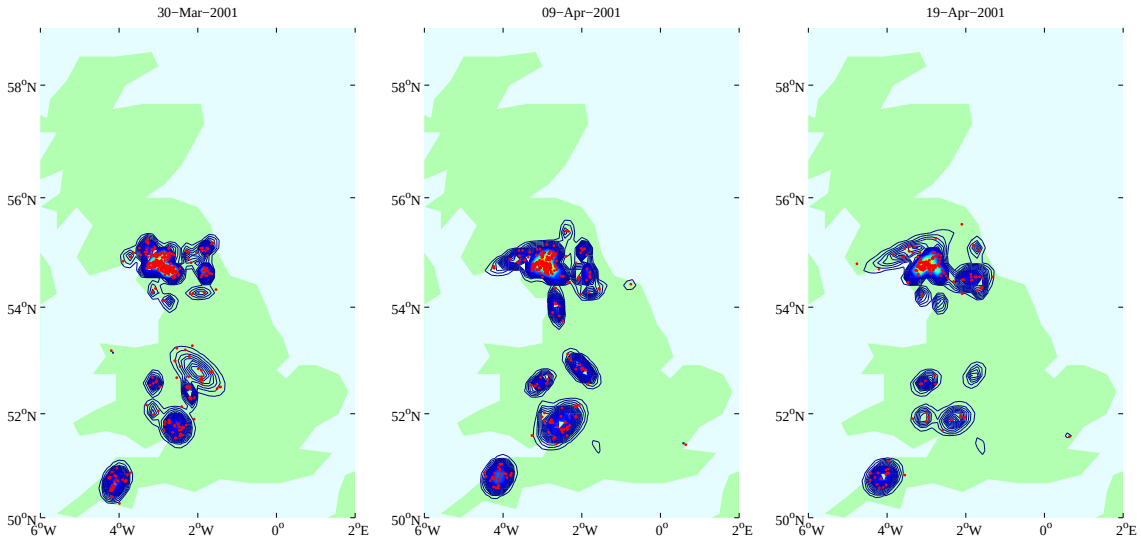


Figure 5: (a) Time evolution of the average number of alive clusters. (b) Time evolution of $\rho_{t|t}$.

at day number t. We assume that $\mathbf{z}_t$ follows a time-varying Dirichlet process mixture with a Gaussian mixed pdf whereas $\mathbb{H}$ is a Normal inverse Wishart distribution. We consider a deterministic deletion model where $r = 7$; seven days being an upper bound for the incubation of the disease. For the cluster locations, we use the model defined by Eq. (12). Inference is performed using the $N-$best algorithm 3 with $N = 1000$ particles. The marginal maximum a posteriori estimates of the cluster locations for some days during the epidemic are represented in Figure 6.

We computed the predictive cumulative density function $\Pr(\mathbf{z}_{k,t} \leq z | \mathbf{z}_{1:k-1,t}, \mathbf{z}_{1:t-1})$ for each disease case. The inverse Gaussian cumulative density function transform of it is plotted in Figure 7 for each of the two dimensions. As can be shown, our simple model does fit the data relatively well, considering that no prior information about farm locations is used here.



(a) (b) (c)

Figure 6: Evolution of the predictive density over time. Red dots represent foot and mouth cases over the last 7 days.

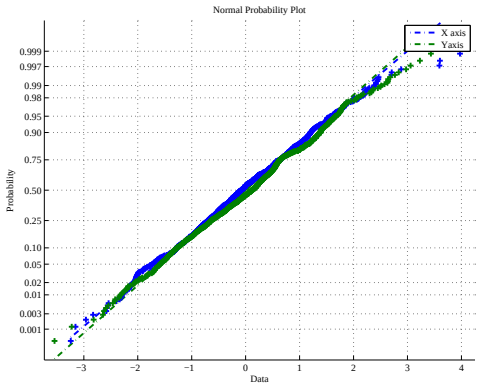


Figure 7: Normal qq-plot of the transformed predictive distribution for foot-and-mouth disease dataset for the two dimensions.

5.3 Object Tracking in Videos

The time-varying Pitman-Yor process can also be used as a prior in models used to find, track, and learn representations of arbitrary objects in a video without a predefined method

for object detection[2]. We present a model that localizes objects via unsupervised tracking while learning a representation of each object, avoiding the need for pre-built detectors. This model uses the Pitman-Yor prior to capture the uncertainty in the number and appearance of objects and requires only spatial and color video data that can be efficiently extracted via frame differencing. Nonparametric mixture models have been used in the past for a variety of computer vision tasks, including tracking (Topkaya et al., 2013; Neiswanger et al., 2014), trajectory clustering (Wang et al., 2011), and video retrieval (Li et al., 2008).

To find and track arbitrary video objects, we model spatial and color features that are extracted as objects travel within a video scene. The model isolates independently moving video objects and learns object models for each that capture their shape and appearance. The learned object models allow for tracking through occlusions and in crowded videos. The unifying framework is a time-varying Pitman-Yor process mixture, where each component is a (time-varying) object model. This setup allows us to estimate the number of objects in a video and track moving objects that may undergo changes in orientation, perspective, and appearance.

We describe the form of the extracted data, define the components of the model, and demonstrate this model on three real video datasets: a video of foraging ants, where we show improved performance over other detection-free methods; a human tracking benchmark video, where we show comparable performance against object-specific methods designed to detect humans; and a T cell tracking task where we demonstrate our method on a video with a large number of objects and show how our unsupervised method can be used to automatically train a supervised object detector.

Data. At each frame t, we assume we are given a set of N_t foreground pixels, extracted via some background subtraction method (such as those detailed in (Yilmaz et al., 2006)). These methods primarily segment foreground objects based on their motion relative to the video background. For example, an efficient method applicable for stationary videos is frame differencing: in each frame t, one finds the pixel values that have changed beyond some threshold, and records their positions $\mathbf{z}^s_{t,n} = (z^{s_1}_{t,n}, z^{s_2}_{t,n})$. In addition to the position of each foreground pixel, we extract color information. The spectrum of RGB color values is discretized into V bins, and the local color distribution around each pixel is described by counts of surrounding pixels (in an $m \times m$ grid) that fall into each color bin, denoted $\mathbf{z}^c_{t,n} = (z^{c_1}_{t,n}, \ldots, z^{c_V}_{t,n})$. Observations are therefore of the form

$$\mathbf{z}_{t,n} = (\mathbf{z}^s_{t,n}, \mathbf{z}^c_{t,n}) = (z^{s_1}_{t,n}, z^{s_2}_{t,n}, z^{c_1}_{t,n}, \ldots, z^{c_V}_{t,n}). \tag{15}$$

Examples of spatial pixel data extracted via frame differencing are shown in Figure 8 (a)-(g).

The time-varying Pitman-Yor process mixture of objects We define an object model, $\mathrm{F}(U_{k,t})$, which is a distribution over pixel data, where $U_{k,t}$ represents the parameters of the k^{th} object at time t. We wish to keep our object model general enough to be applied to arbitrary video objects, but specific enough to learn a representation that can aid in tracking. In this paper, we model each object with

$$\mathbf{z}_{t,n} \sim \mathrm{F}(U_{k,t}) = \mathcal{N}(\mathbf{z}^s_{t,n}|\mu_{k,t}, \Sigma_{k,t}) Mult(\mathbf{z}^c_{t,n}|\delta_{k,t}) \tag{16}$$

2. A preliminary version of this work has been presented as a conference paper (Neiswanger et al., 2014). Matlab code is available at `https://github.com/willieneis/DirichletProcessTracking`.

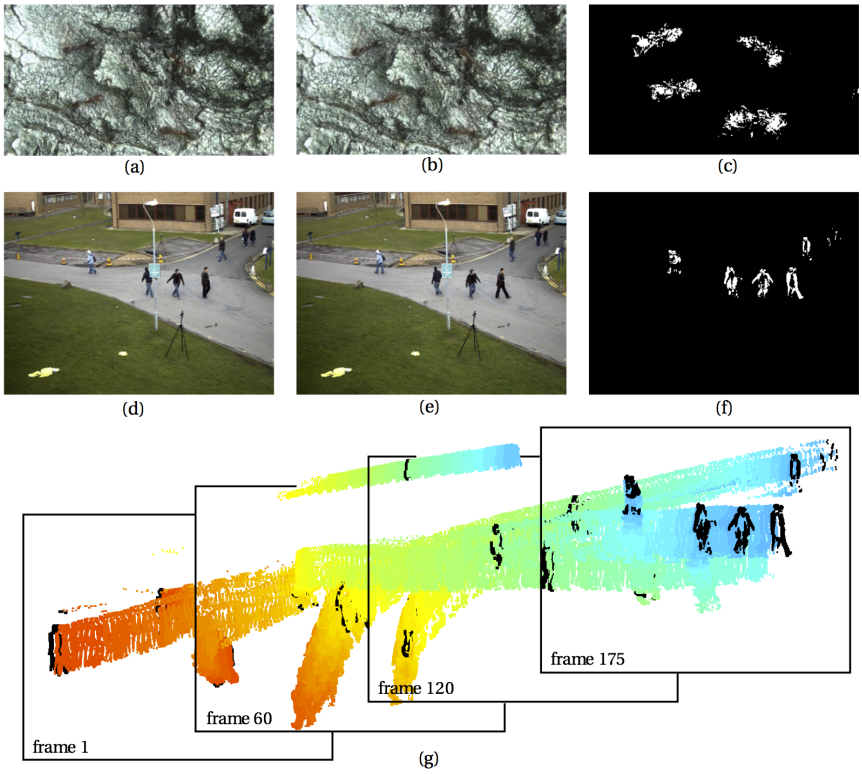


Figure 8: (a - f) Two pairs of consecutive frames and the spatial observations $\mathbf{z}^s_{t,n}$ extracted by taking the pixel-wise frame difference between each pair. (g) The results of frame differencing over a sequence of images (from the PETS2010 dataset).

where object parameters $U_{k,t} = \{\mu_{k,t}, \Sigma_{k,t}, \delta_{k,t}\}$, and $\sum_{j=1}^{V} \delta^j_{k,t} = 1$. The object model captures the objects' locus and extent with the multivariate Gaussian and color distribution with the multinomial. We demonstrate in the following experiments that this representation can capture the physical characteristics of a wide range of objects while allowing objects with different shapes, orientations, and appearances to remain isolated during tracking. We define $\mathbb{H}$, which represents the prior distribution over object parameters, to be

$$\mathbb{H}(\mu_{k,t}, \Sigma_{k,t}, \delta_{k,t}|\mu_0, \kappa_0, \nu_0, \Lambda_0, q_0) = \mathcal{N}iW(\mu_{k,t}, \Sigma_{k,t}|\mu_0, \kappa_0, \nu_0, \Lambda_0)\mathcal{D}(\delta_{k,t}|q_0). \tag{17}$$

To satisfy stationarity, we introduce a set of M auxiliary variables $\mathbf{a}^k_t = (a^k_{t,1}, \ldots, a^k_{t,M})$ for cluster k at time t (Pitt and Walker, 2005). Due to the form of the object parameter priors (i.e the base distribution of the Dirichlet process) and the object model, we can easily apply this auxiliary variable method. With this addition, object parameters do not directly depend on their values at a previous time, but are instead dependent through an intermediate sequence of variables. This allows the cluster parameters at each time step to be marginally distributed according to the base distribution $\mathbb{H}$ while maintaining simple time varying behavior. We can therefore sample from the transition kernel using

$U_{k,t} \sim T(U_{k,t-1}) = T_2 \circ T_1(U_{k,t-1})$, where

$$a^k_{t,1:M} \sim T_1(U_{k,t-1}) = \mathcal{N}(\mu_{k,t-1}, \frac{\Sigma_{k,t-1}}{\tau}) Mult(\delta_{k,t-1}) \tag{18}$$

$$\mu_{k,t}, \Sigma_{k,t}, \delta_{k,t} \sim T_2(a^k_{t,1:M}) = \mathcal{N}iW(\mu_M, \kappa_M, \nu_M, \Lambda_M)\mathcal{D}(q_M) \tag{19}$$

where we choose $\tau = 1$ for all experiments, and formulae for $\mu_M, \kappa_M, \nu_M, \Lambda_M$ and q_M are given in Appendix B.

We apply uniform deletion of allocation variables in this application, to maintain generality of the method over a variety of object types in videos. Additionally, we chose hyperparameter values by simulating from the prior and inspecting the simulated parameters for similarity with existing data. We found that the algorithm performance was not very sensitive to the hyperparameter values settings, and remained fairly consistent over a wide range of settings. In the following experiments we perform inference using both the SMC and PMCMC inference algorithms with $N = 100$ particles, and compare performance of both algorithms.

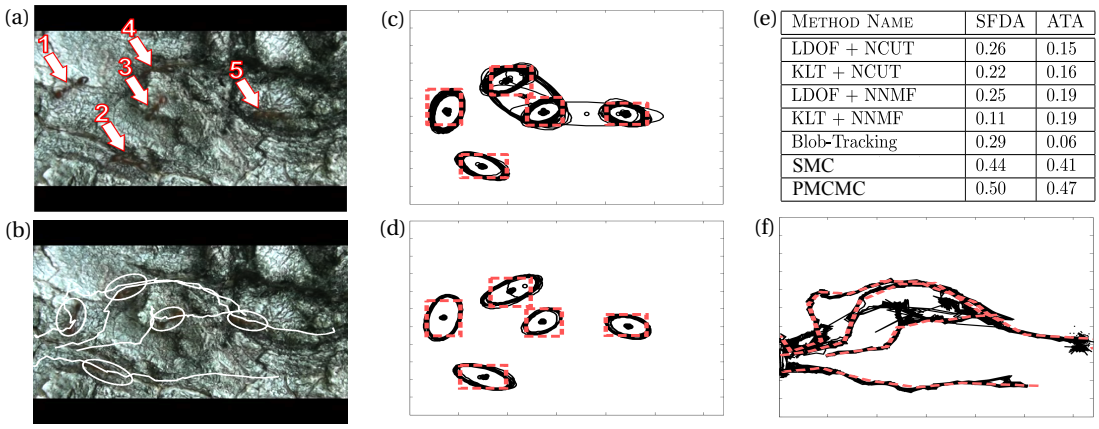
Detection-free comparison methods. Detection-free tracking strategies aim to find and track objects without any prior information about the objects' characteristics nor any manual initialization. One type of existing strategy uses optical flow or feature tracking algorithms to produce short tracklets, which are then clustered into full object tracks. We use implementations of Large Displacement Optical Flow (LDOF) (Brox and Malik, 2011) and the Kanade-Lucas-Tomasi (KLT) feature tracker (Tomasi and Kanade, 1991) to produce tracklets[3]. Full trajectories are then formed using the popular normalized-cut (NCUT) method (Shi and Malik, 2000) to cluster the tracklets or with a variant that uses non-negative matrix factorization (NNMF) to cluster motion using tracklet velocity information (Cheriyadat and Radke, 2009)[4]. We also compare against a detection-free blob-tracking method, where extracted foreground pixels are segmented into components in each frame (Stauffer and Grimson, 2000) and then associated with the nearest neighbor criterion (Yilmaz et al., 2006).

Performance metrics. For quantitative comparison, we report two commonly used performance metrics for object detection and tracking, known as the sequence frame detection accuracy (SFDA) and average tracking accuracy (ATA) (Kasturi et al., 2008). These metrics compare detection and tracking results against human-authored ground-truth, where SFDA$\in [0, 1]$ corresponds to detection performance and ATA$\in [0, 1]$ corresponds to tracking performance, the higher the better. We authored the ground-truth for all videos with the Video Performance Evaluation Resource (ViPER) tool (Doermann and Mihalcik, 2000).

Insect tracking. The video contains six ants with a similar texture and color distribution as the background. The ants are hard to discern, and it is unclear how a predefined detection criteria might be constructed. Further, the ants move erratically and the spatial data extracted via frame differencing does not yield a clear segmentation of the

3. The LDOF implementation can be found at `http://www.seas.upenn.edu/~katef/LDOF.html` and the KLT implementation at `http://www.ces.clemson.edu/~stb/klt/`.
4. The NCUT implementation can be found at `http://www.cis.upenn.edu/~jshi/software/` and the NNMF implementation at `http://www.ornl.gov/~czx/research.html`.

objects in individual frames. A still image from the video, with ant locations shown, is given in Figure 3(a).



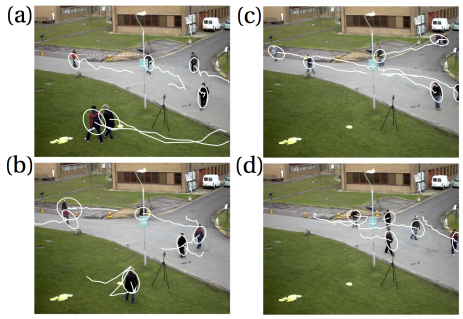
Method Name	SFDA	ATA
LDOF + NCUT	0.26	0.15
KLT + NCUT	0.22	0.16
LDOF + NNMF	0.25	0.19
KLT + NNMF	0.11	0.19
Blob-Tracking	0.29	0.06
SMC	0.44	0.41
PMCMC	0.50	0.47

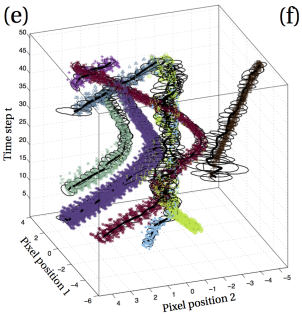
Figure 9: The ants in (a) are difficult to discern (positions labeled). We plot 100 samples from the inferred posterior over object parameters (using SMC (c) and PMCMC (d)) with ground-truth bounding boxes overlaid (dashed). PMCMC proves to give more accurate object parameter samples. We also plot samples over object tracks (sequences of mean parameters) using PMCMC in (f) , and its MAP sample in (b). We show the SFDA and ATA scores for all comparison methods in (e).

We compare the SMC and PMCMC inference algorithms, and find that PMCMC yields more accurate posterior samples (3(d)) than SMC (3(c)). Note that we run PMCMC as described in Section 4 for 100 iterations, where the first pass is equivalent to the SMC algorithm. Ground-truth bounding boxes (dashed) are overlaid on the posterior samples. The MAP PMCMC sample is shown in 3(b) and posterior samples of the object tracks are shown in 3(f), along with overlaid ground-truth tracks (dashed). SFDA and ATA performance metrics for all comparison methods are shown in 3(e). Our model yields higher metric values than all other detection-free comparison methods, with PMCMC inference scoring higher than SMC. The comparison methods seemed to suffer from two primary problems: very few tracklets could follow object positions for an extended sequence of frames, and clustering tracklets into full tracks sharply decreased in accuracy when the objects came into close contact with one another.

Comparisons with detector-based methods. Next, we aim to show that our general-purpose algorithm can compete against state of the art object-specific algorithms, even when it has no prior information about the objects. We use a benchmark human-tracking video from the International Workshop on Performance Evaluation of Tracking and Surveillance (PETS) 2009-2013 conferences (Ellis and Ferryman, 2010), due to its prominence in a number of studies (listed in Figure 10(f)). It consists of a monocular, stationary camera, 794 frame video sequence containing a number of walking humans. Due to the large number of frames and objects in this video, we perform inference with the SMC algorithm only.

Our model is compared against ten object-specific detector-based methods from the PETS conferences. These methods all either leverage assumptions about the orientation, position, or parts of humans, or explicitly use pre-trained human detectors. For example, out of the three top scoring comparison methods, (Breitenstein et al., 2009) uses a state of the art pedestrian detector, (Yang et al., 2009) performs head and feet detection, and (Conte et al., 2010) uses assumptions about human geometry and orientation to segment humans and remove shadows.





(f)

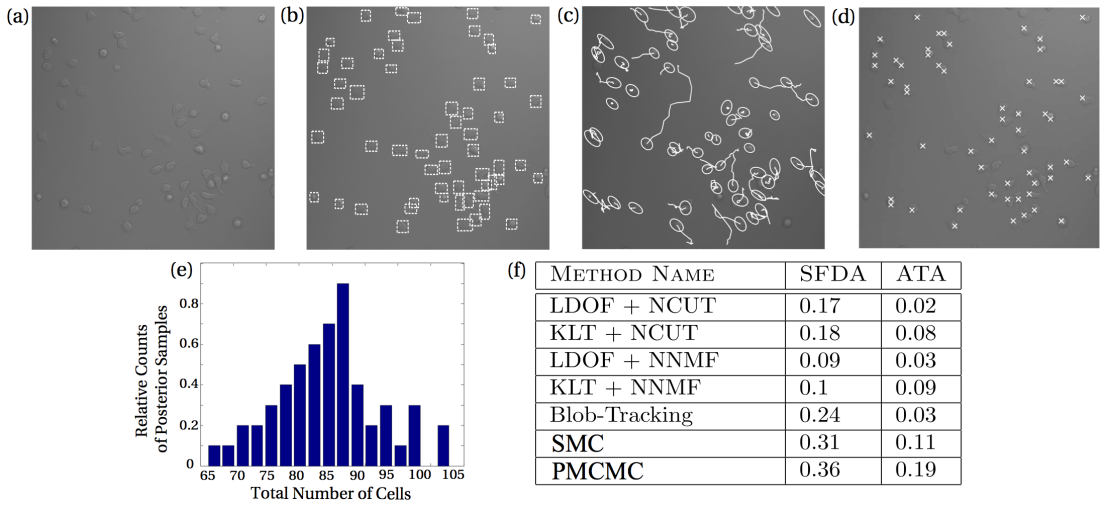
Method Name	SFDA	ATA
Breitenstein	0.57	0.30
Yang	0.55	0.45
Conte	0.53	0.06
SMC	0.51	0.30
Berclaz	0.48	0.15
Alahi 1	0.43	0.04
Alahi 2	0.42	0.05
Bolme 1	0.41	NA
Ge	0.38	0.04
Bolme 2	0.34	NA
Arsic	0.18	0.02

Figure 10: Results on the PETS human tracking benchmark dataset and comparison with object-detector-based methods. The MAP object parameter samples are overlaid on four still video frames (a-d). The MAP object parameter samples are also shown for a sequence of frames (a 50 time-step sequence) along with spatial pixel observations (e) (where the assignment variables $c_{t,n}$ for each pixel are represented by marker type and color). The SFDA and ATA performance metric results for our model and ten human-specific, detection-based tracking algorithms are shown in (f), demonstrating that the our model achieves comparable performance to these human-specific methods. Comparison results were provided by the authors of (Ellis and Ferryman, 2010).

In Figure 4(a-d), the MAP sample from the posterior distribution over the object parameters is overlayed on the extracted data over a sequence of frames. The first 50 frames from the video are shown in 4(e), where the assignment of each data point is represented by color and marker type. We show the SFDA and ATA values for all methods in 4(f), and can see that our model yields comparable results, receiving the fourth highest SFDA score and tying for the second highest ATA score.

Tracking populations of T cells. Automated tracking tools for cells are useful for cell biologists and immunologists studying cell behavior (Manolopoulou et al., 2012). We present results on a video containing T cells that are hard to detect using conventional methods due to their low contrast appearance against a background (Figure 5(a)). Furthermore, there are a large number of cells (roughly 60 per frame, 92 total). In this experiment, we aim to demonstrate the ability of our model to perform a tough detection task while scaling up to a large number of objects. Ground-truth bounding boxes for the cells at a single frame are shown in 5(b) and PMCMC inference results (where the MAP sample is

plotted) are shown in in 5(c). A histogram illustrating the inferred posterior over the total number of cells is shown in 5(e). It peaks around 87, near the true value of 92 cells.



Method Name	SFDA	ATA
LDOF + NCUT	0.17	0.02
KLT + NCUT	0.18	0.08
LDOF + NNMF	0.09	0.03
KLT + NNMF	0.1	0.09
Blob-Tracking	0.24	0.03
SMC	0.31	0.11
PMCMC	0.36	0.19

Figure 11: T cells are numerous, and hard to detect due to low contrast images (a). For a single frame, ground-truth bounding boxes are overlaid in (b), and inferred detection and tracking results are overlaid in (c). A histogram showing the posterior distribution over the total number of cells is shown in (e). The SFDA and ATA for the detection-free comparison methods are shown in (f). Inferred cell positions (unsupervised) were used to automatically train an SVM for supervised cell detection; SVM detected cell positions for a single frame are shown in (d).

Manually hand-labeling cell positions to train a detector is feasible but time consuming; we show how unsupervised detection results from our model can be used to automatically train a supervised cell detector (a linear SVM), which can then be applied (via a sliding window across each frame) as a secondary, speedy method of detection (Figure 5(d)). This type of strategy in conjunction with our model could allow for an ad-hoc way of constructing detectors for arbitrary objects on the fly, which could be taken and used in other vision applications, without needing an explicit predefined algorithm for object detection.

6. Discussion

In this article, we have presented a class of first-order stationary Pitman-Yor processes for time-varying density estimation and clustering. These models are based on a simple generalized Pólya urn sampling scheme whose validity follows from the consistence properties under specific deletion rules of the two-parameter Ewens sampling formula. We have proposed SMC and PMCMC methods to fit these models and have demonstrated them on several applications. The model proposed in the present paper has also been successfully

applied to dynamic spike sorting (Gasthaus et al., 2008) and to cell fluorescent microscopic imaging tracking (Ji and West, 2009).

There are numerous potential extensions to this work. We have focused on specifying directly the marginal distribution of the allocation variables, the underlying infinite-dimensional process $\{\mathbb{G}_t\}$ being integrated out. This has allowed us to develop intuitive models and simple algorithms to fit them. However, it would be interesting to explore whether it is possible to obtain an explicit representation for $\{\mathbb{G}_t\}$ and whether this can be related to the class of models described in (Griffin, 2011).

The uniform and deterministic deletion steps can be applied to any hierarchical model, as long as the predictive distribution is known. We could therefore develop time-varying versions of exchangeable models such as the Bernoulli trips introduced by Walker et al. (1999). The same methodology could also be extended to dynamic social networks. In this context, a classical model is Friend I, which assumes a finite Pólya Urn scheme as the reinforcement procedure for friends interactions (Skyrms and Pemantle, 2000). In order to take into account fading memory, a model called 'discounted from the past' has been proposed, where the weights of the Pólya Urn are discounted at each time step– an approach similar to that of Zhu et al. (2005). However this model has poor asymptotic properties, as each individual always ends up being friend with a single individual (Skyrms and Pemantle, 2000). We conjecture that uniform deletion rules would have nicer properties.

Finally in some applications, it would be interesting to develop models allowing clusters to merge and split over time. We believe that it is possible to build first-order stationary models including such merge-split mechanisms exploiting once more the remarkable properties of the Poisson-Dirichlet distributions under splitting and merging transformations; see e.g. (Pitman, 2002).

Acknowledgments

The authors thank the Action Editor and referees for their helpful comments and Dr Miles Thomas, DEFRA Food and Environment Research Agency, for providing the Foot-and-Mouth data. F. Caron acknowledges the support of the European Commission under the Marie Curie Intra-European Fellowship Programme[5]. A. Doucet's research is partly funded by EPSRC (EP/K000276/1 and EP/K009850/1). Part of this work has been supported by the BNPSI ANR project ANR-13-BS-03-0006-01.

Appendix A. Extensions

A.1 Species Sampling Priors

The consistence properties under size-biased deletion are remarkable properties specific to the two-parameter Ewens sampling formula (Gnedin and Pitman, 2005). However, the uniform and deterministic deletion steps can be extended to any model with a given prediction rule which ensures exchangeability over the data. In particular, the class of species sampling priors - which includes the two-parameter Ewens sampling formula - is of particular interest as it enjoys these two properties (Pitman, 1996; Lee, 2009). A sequence $\{\mathbf{y}_n\}$ is called a

5. The contents reflect only the authors views and not the views of the European Commission.

species sampling sequence if the following prediction rule holds

$$\begin{aligned} \mathbf{y}_1 &\sim \mathbb{G}_0 \\ \mathbf{y}_{n+1}|\mathbf{y}_1,\ldots,\mathbf{y}_n &\sim \sum_{j=1}^{k} \frac{p(n_1,\ldots,n_j+1,\ldots,n_k)}{p(n_1,\ldots,n_j,\ldots,n_k)}\delta_{U_j} + \frac{p(n_1,\ldots,n_j,\ldots,n_k,1)}{p(n_1,\ldots,n_j,\ldots,n_k)}\mathbb{G}_0 \end{aligned} \tag{20}$$

where $\mathbb{G}_0$ is a base distribution, U_j and n_j, $j = 1,\ldots k$, are respectively the set of distinct values within $\mathbf{y}_{1:n}$ and their relative occurrences. Here $p : \cup_{k=1}^{\infty}\mathbb{N}^k \to [0,1]$ is a *symmetric* function of k-tuples of non-negative integers with sum n called the *Exchangeable Partition Probability Function*. This function must satisfy

$$p(1) = 1 \text{ and } p(n_1,\ldots,n_k) = \sum_{j=1}^{k} p(n_1,\ldots,n_j+1,\ldots,n_k) + p(n_1,\ldots,,n_k,1).$$

To obtain first-order stationary species sampling process, we can apply the uniform and deterministic deletion steps followed by the predictive step (20).

A.2 Time-varying Hierarchical and Coloured Dirichlet Processes

Various hierarchical extensions of the Dirichlet process have been proposed such as the popular hierarchical Dirichlet process (Teh et al., 2006) and the coloured Dirichlet process (Green, 2010). We show here how time-varying stationary versions of these models can easily be designed.

The hierarchical Dirichlet process consists of embedded Dirichlet processes where we have at the top of the hierarchy

$$\mathbb{G} \sim \mathrm{DP}(\alpha, \mathbb{G}_0)$$

then for each group $k = 1,\ldots,d$

$$\mathbb{G}_k|\mathbb{G} \sim \mathrm{DP}(a, \mathbb{G})$$

and finally for each item $i = 1,\ldots,n_k$ within group k

$$\begin{aligned} \mathbf{y}_{i,k}|\mathbb{G}_k &\sim \mathbb{G}_k \\ \mathbf{z}_{i,k}|\mathbf{y}_{i,k} &\sim f(\cdot|\mathbf{y}_{i,k}). \end{aligned}$$

A popular application of this model is topic clustering. In this case, a group k represents a document and data $\mathbf{z}_{i,k}$ represent words within documents. The 'mother' mixing distribution $\mathbb{G}$ represents the overall mixture over topics, and each 'child' mixing distribution $\mathbb{G}_k$ represents the mixture associated to a document k, sampled from a Dirichlet process of base measure $\mathbb{G}$. This model can be straightforwardly extended to take into account a time-varying base measure $\mathbb{G}_t$. Similarly to (Teh et al., 2006), a 'Chinese restaurant franchise' formulation can be expressed based on the set of alive allocation variables at time t.

The coloured Dirichlet process consists of a finite mixture of Dirichlet process mixtures where

$$\pi_{1:K} \sim \mathcal{D}(a_1,\ldots,a_K).$$

Here $\mathcal{D}$ denotes the Dirichlet distribution. For each group $k = 1, \ldots, K$, we have

$$\mathbb{G}_k \sim \mathrm{DP}(\alpha_k, \mathbb{G}_{0,k}).$$

Finally for each item $i = 1, \ldots, n$, we have

$$\begin{aligned} c_i|\pi_{1:K} &\sim \pi_{1:K}, \\ \mathbf{y}_i|c_i &\sim \mathbb{G}_{c_i}, \\ \mathbf{z}_i|\mathbf{y}_i &\sim f(\cdot|\mathbf{y}_i). \end{aligned}$$

The main application of such models is for clustering data in presence of some background noise. Again, a time-varying version of this model, where the weights and cluster locations evolve over time, can straightforwardly be developed by deleting allocation variables as described in Section 3 and sampling them conditionally on the set of alive variables as described in (Green, 2010).

Appendix B. Auxiliary Variable Formulae

The parameters $\mu_M, \kappa_M, \nu_M, \Lambda_M$ and q_M in Eq. (19) can be written as

$$\kappa_M = \kappa_0 + M \tag{21}$$

$$\nu_M = \nu_0 + M \tag{22}$$

$$\mu_M = \frac{\kappa_0}{\kappa_0 + M}\mu_0 + \frac{M}{\kappa_0 + M}\overline{\mathbf{a}}^s \tag{23}$$

$$\Lambda_M = \Lambda_0 + S_{\mathbf{a}^s} \tag{24}$$

$$q_M = q_0 + \sum_{i=1}^{M} \mathbf{a}_i^c \tag{25}$$

where $\mathbf{a}^s$ and $\mathbf{a}^c$ respectively denote the spatial and color components of the auxiliary variables, and $\overline{\mathbf{a}}$ and $S_{\mathbf{a}}$ respectively denote the sample mean and sample covariance of the auxiliary variables.

Appendix C. Kolmogorov-Consistent Model

As mentioned in Section 3, the model is not Kolmogorov-consistent. Consider a sequence of distributions $\pi_n(\mathbf{c}_{1:n,1:t})$, $n = 1, 2, \ldots$, where there are n allocation variables at each time t, then, except in the trivial cases $\rho = 0$ or $\rho = 1$

$$\sum_{\mathbf{c}_{n,1:t}} \pi_n(\mathbf{c}_{1:n,1:t}) \neq \pi_{n-1}(\mathbf{c}_{1:n-1,1:t}). \tag{26}$$

It is possible to construct a slightly different model that satisfies the above equality. Consider that there are p latent variables $\widetilde{c}_{k,t}$ which evolve according to the generalized Pólya urn defined in Algorithm 1. We denote by $\widetilde{\mathbf{c}}_{1:t-1}^{t-1}$ the subset of $\widetilde{\mathbf{c}}_{1:t-1}$ corresponding to variables having survived the deletion steps from time 1 to $t-1$, and we denote by $\widetilde{\mathbf{c}}_{1:t-1}^{t}$ the subset corresponding to those having survived from time 1 to t. Let $\widetilde{K}_{t-1}$ be the number of

clusters created from time 1 to $t-1$. We denote by $\widetilde{\mathbf{m}}_{t-1}^{t-1}$ the vector of size $\widetilde{K}_{t-1}$ containing the size of the clusters associated to $\widetilde{\mathbf{c}}_{1:t-1}^{t-1}$, and we denote by $\widetilde{\mathbf{m}}_{t-1}^{t}$ the vector containing the size of clusters associated to $\widetilde{\mathbf{c}}_{1:t-1}^{t}$.

Conditional on $(\widetilde{\mathbf{c}}_{1:t-1}^{t}, \widetilde{c}_t)$ (or equivalently $(\widetilde{\mathbf{m}}_{t-1}^{t}, , \widetilde{c}_t)$), the allocation variables $c_{k,t}$ of the n observations are simply drawn from a standard Pólya urn. The whole model is represented in Figure 12. As the Markov process is on the latent variables and not on the allocation variables $c_{k,t}$ associated to observations, this model is Kolmogorov-consistent:

$$\sum_{\mathbf{c}_{n,1:t}} \pi_n(\mathbf{c}_{1:n,1:t}) = \pi_{n-1}(\mathbf{c}_{1:n-1,1:t}). \tag{27}$$

Note that we have in this case an additional tuning parameter p. Larger values of p induce higher correlations.

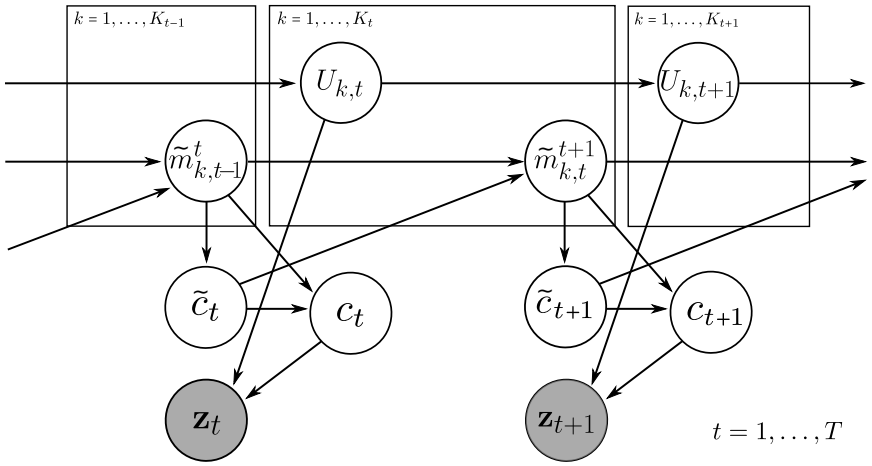


Figure 12: A representation of the time-varying Pitman-Yor process mixture as a directed graphical model, representing conditional independencies between variables. All assignment variables and observations at time t are denoted c_t and $\mathbf{z}_t$, respectively.

References

A. Ahmed and E.P. Xing. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process. In *Proceedings of the Eighth SIAM International Conference on Data Mining*, 2008.

A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst. Sparsity-driven people localization algorithm: Evaluation in crowded scenes environments. In *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, pages 1–8. IEEE, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo for efficient numerical simulation. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 45–60. Springer, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society B*, 72:269–342, 2010.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

J. Arbel, K. Mengersen, and J. Rousseau. Bayesian nonparametric dependent model for the study of diversity for species data. Technical report, arXiv preprint arXiv:1402.3093, 2014.

D. Arsic, A. Lyutskanov, G. Rigoll, and B. Kwolek. Multi camera person tracking applying a graph-cuts based foreground segmentation in a homography framework. In *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, pages 1–8. IEEE, 2009.

N. Bartlett, D. Pfau, and F. Wood. Forgetting counts: Constant memory inference for a dependent hierarchical Pitman-Yor process. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 63–70, 2010.

J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, pages 1–8. IEEE, 2009.

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

D. Blei and P. Frazier. Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, 12:2383–2410, 2011.

D.S. Bolme, Y.M. Lui, BA Draper, and JR Beveridge. Simple real-time human detection using a single correlation filter. In *Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, pages 1–8. IEEE, 2009.

M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Markovian tracking-by-detection from a single, uncalibrated camera. In *Int. IEEE CVPR Workshop on Performance Evaluation of Tracking and Surveillance (PETS'09)*, 2009.

T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, 2011.

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

F. Caron, M. Davy, A. Doucet, E. Duflos, and P. Vanheeghe. Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Transactions on Signal Processing*, 56(1):71–84, 2008.

A. M. Cheriyadat and R. J. Radke. Non-negative matrix factorization of partial track data for motion segmentation. In *IEEE 12th International Conference on Computer Vision*, pages 865–872. IEEE, 2009.

D.M. Cifarelli and E. Regazzini. Nonparametric statistical problems under partial exchangeability. *Annali del'Instituto di Matematica Finianziara dell'Universit di Torino*, 12:1–36, 1978.

D. Conte, P. Foggia, G. Percannella, and M. Vento. Performance evaluation of a people tracking system on pets2009 database. In *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 119–126. IEEE, 2010.

D. Doermann and D. Mihalcik. Tools and techniques for video performance evaluation. In *15th International Conference on Pattern Recognition*, volume 4, pages 167–170. IEEE, 2000.

A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in practice*. Springer-Verlag, 2001.

D. B. Dunson and J.-H. Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2007.

D. B. Dunson, N. Pillai, and J.-H. Park. Bayesian density regression. *Journal of the Royal Statistical Society B*, 69(2):163–183, 2006.

A. Ellis and J. Ferryman. PETS2010 and PETS2009 evaluation of results using individual ground truthed single views. In *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 135–142. IEEE, 2010.

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

P. Fearnhead. Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, 14:11–21, 2004.

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

N. Foti and S. Williamson. A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to appear, 2013.

J. Gasthaus, F. Wood, D. Görür, and Y.W. Teh. Dependent Dirichlet process spike sorting. In *NIPS*, 2008.

W. Ge and R.T. Collins. Evaluation of sampling-based pedestrian detection for crowd counting. In *2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, pages 1–7. IEEE, 2009.

A.E. Gelfand, A. Kottas, and S.N. MacEachern. Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association*, 100(471):1021–1035, 2005. doi: 10.1198/016214504000002078.

A. Gnedin and J. Pitman. Regenerative partition structures. *Electronic Journal of Combinatorics*, 11:1–21, 2005.

P.J. Green. Colouring and breaking sticks: random distributions and heterogeneous clustering. In *Probability and Mathematical Genetics: Papers in Honour of Sir John Kingman*. Cambridge University Press, 2010.

P.J. Green and S. Richardson. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, 28(2):355–375, 2001.

J. E. Griffin. The Ornstein–Uhlenbeck Dirichlet process and other time-varying processes for Bayesian nonparametric inference. *Journal of Statistical Planning and Inference*, 141(11):3648–3664, 2011.

J.E. Griffin and M.F.J. Steel. Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101(473):179–194, 2006.

J.E. Griffin and M.F.J. Steel. Time-dependent stick-breaking processes. Technical report, University of Kent, 2009.

J.E. Griffin and M.F.J. Steel. Stick-breaking autoregressive processes. *Journal of Econometrics*, 162 (2):383–396, 2011.

P. Hall, H. Müller, and P. Wu. Time-dynamic density and mode estimation with applications to fast mode tracking. *Journal of Computational and Graphical Statistics*, 15:82–100, 2006.

J. H. Huggins and F. Wood. Infinite structured hidden semi-Markov models. *arXiv preprint arXiv:1407.0044*, 2014.

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

N. Jaoua, F. Septier, E. Duflos, and P. Vanheeghe. State and impulsive time-varying measurement noise density estimation in nonlinear dynamic systems using Dirichlet process mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1–5, Florence, Italy, 2014.

C. Ji and M. West. Dynamic spatial mixture modeling and its application in Bayesian tracking for cell fluorescent microscopic imaging. In *Joint Statistical Meeting*, 2009.

H. Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, 1997.

M. Kalli, J.E. Griffin, and S.G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, 2011.

N. Kantas, A. Doucet, S. Singh, and J. M. Maciejowski. An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. In *15th IFAC Symposium on System Identification (SYSID), Saint-Malo, France.*, volume 102, page 117, 2009.

R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang. Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 319–336, 2008.

M.J. Keeling, M.E.J. Woolhouse, D.J. Shaw, L. Matthews, M. Chase-Topping, D.T. Haydon, S.J. Cornell, J. Kappey, J. Wilesmith, and B.T. Grenfell. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science*, 294:813–817, 2001.

J.F.C. Kingman. Random partitions in population genetics. *Proceedings of the Royal Society of London*, 361:1–20, 1978.

J. Lee. Species sampling model and its application to Bayesian statistics. Technical report, Seoul National University, 2009.

X. Li, W. Hu, Z. Zhang, X. Zhang, and G. Luo. Trajectory-based video retrieval using Dirichlet process mixture models. In *BMVC*, pages 1–10, 2008.

D. Lin, E. Grimson, and J. W. Fisher III. Construction of dependent Dirichlet processes based on Poisson processes. In *NIPS*. Neural Information Processing Systems Foundation (NIPS), 2010.

S.N. MacEachern. Dependent Dirichlet processes. Technical report, Dept. of Statistics, Ohio State university, 2000.

S.N. MacEachern, M. Clyde, and J.S. Liu. Sequential importance sampling for nonparametric Bayes models: the next generation. *Canadian Journal of Statistics*, 27(2):251–267, 1999.

I. Manolopoulou, M. P. Matheu, M. D. Cahalan, M. West, and T. B. Kepler. Bayesian spatio-dynamic modeling in cell motility studies: Learning nonlinear taxic fields guiding the immune response. *Journal of the American Statistical Association*, 107(499):855–865, 2012.

P. Müller and F. Quintana. Nonparametric Bayesian data analysis. *Statistical Science*, pages 95–110, 2004.

P. Müller, F. Quintana, and G. Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society B*, 66:735–749, 2004.

W. Neiswanger, F. Wood, and E. Xing. The dependent Dirichlet process mixture of objects for detection-free tracking and object modeling. In *Artificial Intelligence and Statistics*, 2014.

E. Ozkan, I. Y. Ozbek, and M. Demirekler. Dynamic speech spectrum representation and tracking variable number of vocal tract resonance frequencies with time-varying Dirichlet process mixture models. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1518–1532, 2009.

O. Papaspiliopoulos and G.O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186, 2008.

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158, 1995.

J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory; Papers in Honor of David Blackwell*. Institute of Mathematical Statistics, California, 1996.

J. Pitman. Probabilistic bounds on the coefficients of polynomials with only real zeros. *Journal of Combinatorial Theory*, 77:279–303, 1997.

J. Pitman. Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. *Combinatorics, Probability and Computing*, 11:501–514, 2002.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25:855–900, 1997.

M.K. Pitt and S.G. Walker. Constructing stationary time series models using auxiliary variables with applications. *Journal of the American Statistical Association*, 100(470):554–564, 2005.

M.K. Pitt, C. Chatfield, and S.G. Walker. Constructing first order stationary autoregressive models via latent processes. *Scandinavian Journal of Statistics*, 29:657–663, 2002.

G. Poyiadjis, A. Doucet, and S. Singh. Particle approximations of the score and observed information matrix in state space models with application to parameter estimation. *Biometrika*, 98(1):65–80, 2011.

V. Rao and Y.W. Teh. Spatial normalized gamma processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

A. Rodriguez and D. B. Dunson. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1), 2011.

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

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

B. Skyrms and R. Pemantle. A dynamic model of social network formation. *Proceedings of the National Academy of Sciences*, 97:9340–9346, 2000.

N. Srebro and S. Roweis. Time-varying topic models using dependent Dirichlet processes. Technical report, Department of Computer Science, University of Toronto, 2005.

C. Stauffer and E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

Y. W. Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006.

Y. W. Teh, C. Blundell, and L. Elliott. Modelling genetic variations using fragmentation-coagulation processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 819–827, 2011.

Y.W. Teh and M.I. Jordan. Hierarchical Bayesian nonparametric models with applications. *Bayesian Nonparametrics*, pages 158–207, 2010.

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

C. Tomasi and T. Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ., 1991.

I. S. Topkaya, H. Erdogan, and F. Porikli. Detecting and tracking unknown number of objects with Dirichlet process mixture models and Markov random fields. In *Advances in Visual Computing*, pages 178–188. Springer, 2013.

S.G. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36(1):45–54, January 2007.

S.G. Walker and P. Muliere. A bivariate Dirichlet process. *Statistics and probability letters*, 64:1–7, 2003.

S.G. Walker, P. Damien, P.W. Laud, and A.F.M. Smith. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society B*, 61(3): 485–527, 1999.

X. Wang, K. T. Ma, G. Ng, and E. L. Grimson. Trajectory analysis and semantic region modeling using nonparametric hierarchical bayesian models. *International Journal of Computer Vision*, 95 (3):287–312, 2011.

J. Yang, PA Vela, Z. Shi, and J. Teizer. Probabilistic multiple people tracking through complex situations. In *11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009.

A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4):13, 2006.

X. Zhu, Z. Ghahramani, and J. Lafferty. Time-sensitive Dirichlet process mixture models. Technical report, Carnegie Mellon University, Pittsburgh, PA, 2005.