

# A Bayesian hierarchical model for related densities using Pólya trees

Jonathan Christensen

*Duke University, Durham, NC, USA.*

E-mail: jonathan.christensen@duke.edu

Li Ma

*Duke University, Durham, NC, USA.*

E-mail: li.ma@duke.edu

**Summary.** Bayesian hierarchical models are used to share information between related samples and obtain more accurate estimates of sample-level parameters, common structure, and variation between samples. When the parameter of interest is the distribution or density of a continuous variable, a hierarchical model for continuous distributions is required. A number of such models have been described in the literature using extensions of the Dirichlet process and related processes, typically as a distribution on the parameters of a mixing kernel. We propose a new hierarchical model based on the Pólya tree, which allows direct modeling of densities and enjoys some computational advantages over the Dirichlet process. The Pólya tree also allows more flexible modeling of the variation between samples, providing more informed shrinkage and permitting posterior inference on the dispersion function, which quantifies the variation among sample densities. We also show how the model can be extended to cluster samples in situations where the observed samples are believed to have been drawn from several latent populations.

## 1. Introduction

Many statistical applications deal with learning and comparing the distributions of two or more related samples. We may be interested in learning how samples are similar or testing whether they are distinguishably different from each other. Because distributions are complex infinite-dimensional objects, classical approaches work with low-dimensional parameterizations of the distribution. Analysis of variance, for example, reduces distributions to a mean and variance, which are sufficient under the assumption of normality. A wide range of other parametric models within both the Bayesian and frequentist inferential frameworks use other parameterizations of the distribution to reduce the dimensionality of the

problem. Classical nonparametric approaches use features of the samples such as medians (Westenberg, 1948), rank-based scores (Wilcoxon, 1945), or summaries of the empirical distribution functions, as in the Kolmogorov-Smirnov (Kolmogorov, 1933) and Cramér-von Mises tests (Anderson, 1962).

A number of Bayesian nonparametric approaches embrace the infinite-dimensional nature of the problem using extensions of Dirichlet processes citeferguson1973. Among the most well-known, The Hierarchical Dirichlet process (Beal et al., 2002; Teh et al., 2006) builds a hierarchical model using the Dirichlet process, allowing it to share information between samples. Tomlinson and Escobar (1999) provide an early description of a similar model. The Nested Dirichlet process of Rodríguez et al. (2008) takes a different approach, using a Dirichlet process as the base measure of a second Dirichlet process. This induces clustering in the samples, with samples in a cluster being modeled with a single density. Müller et al. (2004) model each sample density as a mixture of two components: one Dirichlet process mixture of Gaussians representing common structure between samples, and a second Dirichlet process mixture of Gaussians representing the idiosyncratic structure of the given sample. A variety of other dependent Dirichlet Processes (MacEachern, 1999) have been described in the literature. Teh and Jordan (2010) give an overview of hierarchical models based on the Dirichlet process.

While the Dirichlet process has been the basis of most of the work in this area, work has also been done on hierarchical extensions of other priors. For example, Teh (2006) defines a Hierarchical Pitman-Yor process, taking advantage of the more flexible clustering structure of the Pitman-Yor process over the Dirichlet process. Camerlenghi et al. (2017) consider hierarchical models based on Normalized Random Measures (Barrios et al., 2013), which includes the Dirichlet process as a special case.

An alternative approach for modeling densities within the Bayesian nonparametric framework is to use a model derived from Pólya trees (Freedman, 1963). Pólya trees are a class of tail-free prior in which an infinite recursive binary partition is placed on the sample space, and probability mass assigned to the elements of the partition by a corresponding infinite sequence of Beta-distributed random variables. A special case gives the familiar Dirichlet process (Ferguson, 1973), but the Pólya tree family is considerably more flexible. With appropriate specification of the prior parameters, the Pólya tree assigns probability 1 to the set of absolutely continuous distributions (Kraft, 1964; Ferguson, 1974). This property allows it to be used to directly model probability densities without the encumbrance of a mixing kernel.

The Pólya tree model allows tractable computation of the marginal likelihood, which has made it popular in Bayesian hypothesis testing of

nonparametric density models. Examples include Berger and Guglielmi (2001); Ma and Wong (2011); Chen and Hanson (2014); Holmes et al. (2015); Soriano and Ma (2017); Filippi and Holmes (2017). These approaches are unsatisfactory when estimation and prediction rather than formal hypothesis testing are of primary interest. Statisticians have recognized the benefits of partial shrinkage as far back as Stein’s shrinkage estimator for the mean of a multivariate normal distribution (Stein, 1956). We may expect that related samples will have similar but not identical distributions. In this situation, neither independence nor a single common distribution are appropriate models. A partial shrinkage model allows borrowing of information across samples while preserving cross-sample variation. The Polya tree has also been used as a building block for modeling dependent densities. Zhao and Hanson (2011) describe a spatially dependent Polya tree with an autoregressive structure. When the spatial dependence is removed, a partially exchangeable model similar to the one we consider here arises. Schörgendorfer and Branscum (2013) and Nieto-Barajas and Quintana (2016) describe dependent Polya tree models with autoregressive structures inducing dependence among densities over time. Jara and Hanson (2011) transform Gaussian processes in the framework of tail-free processes to create a regression model for dependent densities.

Within the Bayesian inference framework, partial shrinkage is naturally achieved with a hierarchical model. While a simple hierarchy of Pólya trees is possible and is described herein, we build on the richer Adaptive Pólya tree model of Ma (2017). This model allows us to learn the concentration parameters of the Pólya tree, rather than fixing them in the prior. In contrast to Dirichlet process-based models, the richer structure of the Pólya tree allows us to model both the means and the cross-sample variation of the densities in a fully nonparametric manner. This not only improves the estimation of the distributions, but also allows us to perform inference on how the variation across samples differs over the sample space.

In Section 2 we describe the model and contrast it with several existing models. We discuss Bayesian posterior inference and computation in Section 3. In Section 4 we give theoretical grounding for the appropriateness of the method. Section 5 describes two ways in which the model allows us to go beyond the capabilities of existing models. We show simulation results in Section 6 and demonstrate application to real data in Section 7. Finally, we conclude with some discussion in Section 8.

## 2. Model

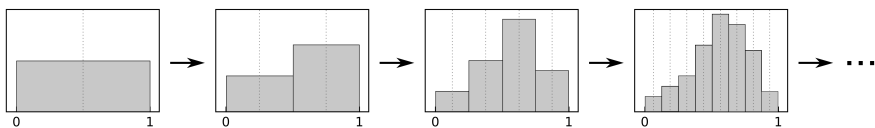
### 2.1. Reviewing the Pólya tree construction

We begin with a brief sketch of the Pólya tree; the reader interested in the mathematical details should refer to Mauldin et al. (1992); Lavine (1992, 1994). The Pólya tree consists of an infinite recursive partition  $\mathcal{A}$  of the sample space and a corresponding infinite sequence of Beta-distributed random variables which assign mass to the various regions  $A \in \mathcal{A}$  of the partition. Figure 1 illustrates the partitioning sequentially. In this illustration our sample space is the interval  $(0, 1]$  on the real line, and our prior mean, shown in the first pane, is the uniform distribution on that interval. At each level of the recursive partition we cut each region in half at the midpoint. Although arbitrary partitions may be used, the dyadic partition described here is convenient and is often used as a default partition in the absence of a compelling reason to use a different one. Other partitions may be more convenient in the presence, for example, of censored data (Muliere and Walker, 1997). The second pane shows the result after the first cut and mass allocation, in this case the majority of the mass having been allocated to the right-hand side. In the next step we cut each of the two regions from the second pane in half again, and assign the probability mass to each to its children according to a Beta-distributed random variable. This results in four regions, shown in the third pane, which are again cut and mass distributed in the fourth pane. The process continues indefinitely.

We denote the Pólya tree prior as  $Q \sim \text{PT}(Q_0, \nu)$ , where  $Q_0$  is the centering distribution and  $\nu$  is an (infinite-dimensional) concentration parameter describing how much  $Q$  is expected to vary from  $Q_0$ . The parameters of the sequence of Beta distributions from which the mass allocations are drawn are derived very simply from  $Q_0$  and  $\nu$ . For an arbitrary region  $A$  belonging to the recursive partition  $\mathcal{A}$ , the fraction of the mass allocated to the left child  $A_\ell$  of  $A$  is given by the random variable

$$\theta(A) \sim \text{Beta}(\theta_0(A)\nu(A), (1 - \theta_0(A))\nu(A)),$$

where  $\theta_0(A) = Q_0(A_\ell)/Q_0(A)$ , with the remainder of the mass allocated to the right child  $A_r$ .  $Q_0$  thus determines the mean of the mass allocations (and hence the expectation of the resulting density), while  $\nu = \{\nu(A) : A \in \mathcal{A}\}$  controls the variation of the mass allocations, and hence the dispersion of  $Q$  around  $Q_0$ . Alternatively,  $\nu$  controls the strength of the shrinkage of the posterior mean density from the empirical process towards  $Q_0$ . With an appropriate choice of  $\nu$ , the Pólya tree prior almost surely generates an absolutely continuous distribution (Kraft, 1964).



**Fig. 1.** An illustration of the recursive partitioning and probability allocation of the standard Pólya tree.

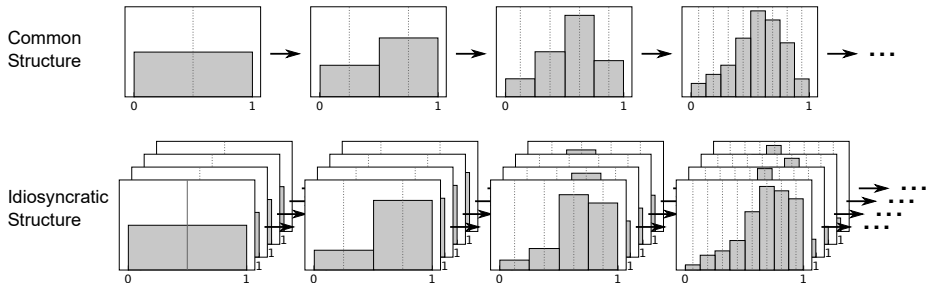
## 2.2. The Hierarchical Pólya Tree

It is conceptually straightforward to extend the Pólya tree to a hierarchical model. Let  $\mathbf{X}_1, \dots, \mathbf{X}_k$  be  $k$  samples arising from related distributions on a complete, separable space  $\Omega$ . For ease of exposition we again use  $\Omega = (0, 1]$ , though like the Pólya tree the model can be applied to more general sample spaces. We model these samples as coming from  $k$  exchangeable distributions  $Q_i$ , which are centered at a common underlying measure  $Q$ , itself unknown. Applying Pólya tree priors to both  $Q$  and the  $Q_i$ , all with identical partition structures, gives us the hierarchical model

$$\begin{aligned} X_{ij} | Q_i &\stackrel{\text{ind}}{\sim} Q_i \\ Q_i | Q &\stackrel{\text{iid}}{\sim} \text{PT}(Q, \boldsymbol{\tau}) \\ Q &\sim \text{PT}(Q_0, \boldsymbol{\nu}). \end{aligned} \tag{1}$$

Here  $Q_0$  is the overall prior mean,  $\boldsymbol{\tau}$  controls the variation across samples around the common structure  $Q$ , and  $\boldsymbol{\nu}$  controls the variation of the common structure from  $Q_0$ , which determines the smoothness of  $Q$ . This model, which we call the Hierarchical Pólya tree, allows nonparametric estimation of the sample distributions  $Q_i$  and the common structure  $Q$ . The Hierarchical Pólya tree model is illustrated in Figure 2. The first row shows the upper level of the hierarchy, which like the basic Pólya tree illustrated in Figure 1 is centered at the uniform distribution on  $(0, 1]$ . The second row shows the lower level of the hierarchy, the individual sample distributions  $Q_i$  conditioned on  $Q$ . They follow exactly the same Pólya tree construction, but each cut is centered on the corresponding cut from  $Q$ , rather than on the uniform distribution.  $Q$  captures the common structure of the samples, while the deviation of  $Q_i$  from  $Q$  captures the idiosyncratic structure of each sample.

Because the partition structures are identical, the hierarchy of Pólya trees translates directly to the decomposed space as a hierarchical model for



**Fig. 2.** An illustration of the Hierarchical Pólya tree.

Beta random variables. For an arbitrary region  $A \in \mathcal{A}$ , we have

$$\begin{aligned} \theta_i(A) \mid \theta(A) &\stackrel{iid}{\sim} \text{Beta}(\theta(A)\tau(A), (1 - \theta(A))\tau(A)) \\ \theta(A) &\sim \text{Beta}(\theta_0(A)\nu(A), (1 - \theta_0(A))\nu(A)). \end{aligned} \quad (2)$$

The representation of the hierarchy of Pólya trees as a hierarchy of Beta random variables allows tractable posterior inference, as described in Section 3.

### 2.3. The Stochastically Increasing Shrinkage prior on dispersion

The Pólya tree’s concentration parameter is traditionally set to increase with depth at a predetermined rate to ensure absolute continuity, with a constant multiplicative term to control the overall level of variation which may be treated as a tuning parameter (as in Berger and Guglielmi (2001)) or have a prior placed on it. Hanson (2006) discusses some of the necessary considerations when placing a prior on this parameter. More recently, Ma (2017) shows that putting a flexible nonparametric prior on the concentration parameter allows the Pólya tree to learn the true distribution more accurately, particularly when the smoothness of the underlying density varies over the sample space. We can extend the Hierarchical Pólya tree model by placing priors on both concentration parameters  $\tau$  and  $\nu$ . In addition to more accurate inference on  $Q$  and the  $Q_i$ , putting a prior distribution on  $\tau$  allows us to learn the variation across samples in a nonparametric way. That is, we can estimate a posterior *dispersion function* which summarizes the variability across sample densities at any given point in the sample space. Dispersion can be measured in a variety of ways; in Section 5.1 we show how to derive the posterior mean variance of the densities and interpret a standardized version using the coefficient of variation to correct for the height of the density. Nonparametric inference on the variation across samples over the sample space is

made possible by the flexibility of the Pólya tree model. We contrast how several other models treat cross-sample variation in Section 2.5.

A simple approach is to put independent priors on the variance of each Beta random variable. However, we expect spatial structure in the dispersion function—locations near each other in the sample space are likely to have similar levels of variation. The variance of the Beta distributions is generally expected to be smaller at deeper levels of the partition, but the decay in the variance may be heterogeneous over the sample space depending on the local smoothness of the densities. While the recursive partitioning allows independent priors to capture some spatial structure, we can do better by introducing dependency between regions in the partition. Ma (2017) introduces Markov dependency on the concentration parameter, following the tree topology. The Markov dependency among the Beta variance parameters allow the Beta variables to stochastically transition into lower prior variance (i.e., higher shrinkage) along each branch of the partition tree, but at potentially different rates, thereby allowing spatially heterogeneous prior variability in the random densities. Instead of constructing a continuous state-space model on the prior Beta variance directly, which would make posterior computation very challenging, a discrete latent state variable that characterizes a finite number of different levels of prior variance is introduced for each Beta variable, and the Markov dependency is then imposed on these latent state variables. This discretization strategy maintains the computational tractability of the Pólya tree model when equipped with more flexible prior on the variance parameters.

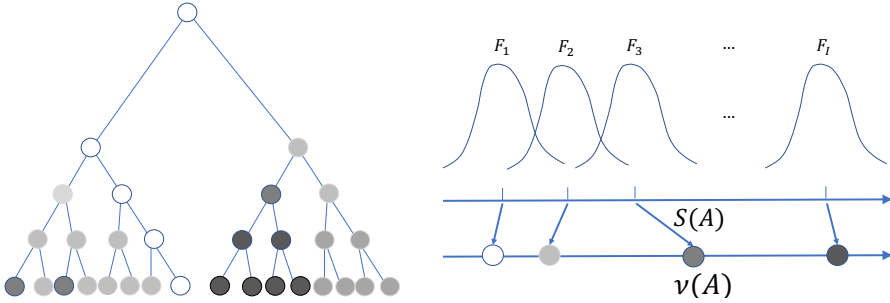
More specifically, the Stochastically Increasing Shrinkage (SIS) prior introduces a state variable  $S(A)$  supported on a finite set of integers  $1, \dots, I$ , corresponding to decreasing prior variance and increasing shrinkage for the Beta random variables. In particular, the last state  $I$  corresponds to complete shrinkage or zero variance, which is achieved through fixing  $\nu(A)$  at  $\infty$  when  $S(A) = I$ . For example, we may have  $S(A) \in \{1, 2, 3, 4\}$  with  $S(A) = i$  implying  $\nu(A) | S(A) = i \sim F_i$  with the  $F_i$  stochastically ordered  $F_1 \prec F_2 \prec F_3 \prec F_4$  and  $F_4$  being a point mass at infinity, corresponding to zero variance. The number of states and the corresponding distributions can be chosen to balance the flexibility and computational complexity of the model. A simple way to enforce such a stochastic ordering is through partitioning the support of the concentration parameter  $\nu(A)$  into disjoint intervals. Given these latent states, the SIS prior adopts a transition probability matrix  $\Gamma(A)$  for  $S(A) | S(\text{Par}(A))$ , where  $\text{Par}(A)$  represents the parent node of  $A$  in the tree. Ma (2017) discusses several prior possibilities for this transition matrix. We adopt the recommendation given there to use an exponential kernel for the transition probabilities. The resulting transition matrix can

be given as

$$\Gamma(A) = \begin{bmatrix} \frac{1}{\sum_{i=0}^{I-1} e^{\beta \cdot i}} & \frac{e^{\beta}}{\sum_{i=0}^{I-1} e^{\beta \cdot i}} & \cdots & \frac{e^{\beta(I-1)}}{\sum_{i=0}^{I-1} e^{\beta \cdot i}} \\ 0 & \frac{1}{\sum_{i=0}^{I-2} e^{\beta \cdot i}} & \cdots & \frac{e^{\beta(I-2)}}{\sum_{i=0}^{I-2} e^{\beta \cdot i}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

This upper-triangular transition probability matrix induces stochastically increasing shrinkage (or decreasing prior Beta variance) along each branch in the partition tree (see Figure 3), and ensures that the model generates absolutely continuous densities (see Theorem 1). The parameter  $\beta$  controls the stickiness of the transition; that is, larger values of beta correspond to a stronger dependence in the shrinkage state between adjacent nodes in the Pólya Tree.  $\beta$  can be set equal to zero, in which case the transition is uniform over the states with shrinkage no less than the current node. We denote the SIS prior

$$\nu \sim \text{SIS}(\Gamma).$$



**Fig. 3.** Illustration of the SIS prior. The shrinkage states increase as you follow the tree down to finer scales—indicated by darker shades of gray—at potentially different rates across the space, eventually reaching complete shrinkage—but allowing less shrinkage where the data dictates such.

Figure 3 illustrates the SIS prior in action. The gray-scale in each node indicates the value of the Beta variance, with darker gray indicates less variance and higher shrinkage. These Beta variances are determined by a set of latent state variables, each corresponding to a conditional prior  $F_i$  on  $\nu(A)$ , represented in the right panel of the figure. As you move down to finer resolutions the shrinkage state tends to increase, eventually



reaching complete shrinkage, indicated by black nodes, though the rates at which the shrinkage increases along different branches of the tree are different and are determined stochastically by the underlying top-down Markov model linking the latent state variables. This model allows the amount of shrinkage or prior variance to vary over the sample space to capture large-scale smooth features in one part of the sample space and smaller scale features in another part. This allows the resulting density to have heterogeneous smoothness and variability around the mean across the sample space.

## 2.4. The Hierarchical Adaptive Pólya Tree

Having described the SIS prior, we can adopt this prior for the concentration parameters  $\tau$  and  $\nu$ , and write the complete model as follows:

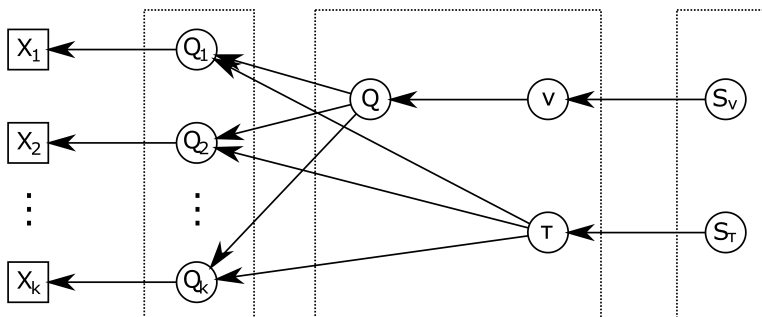
$$\begin{aligned} X_{ij} \mid Q_i &\overset{ind}{\sim} Q_i \\ Q_i \mid Q, \tau &\overset{iid}{\sim} \text{PT}(Q, \tau) \\ Q \mid \nu &\sim \text{PT}(Q_0, \nu) \\ \tau &\sim \text{SIS}(\Gamma_\tau) \\ \nu &\sim \text{SIS}(\Gamma_\nu). \end{aligned}$$

We call this model the Hierarchical Adaptive Pólya Tree (HAPT). In contrast to existing models, this specification allows fully nonparametric inference on both the densities and the variation across densities. The inclusion of the SIS priors allows the model to adapt the level of shrinkage or information borrowing in different parts of the sample space to more accurately capture the density of each sample, rather than using fixed uniform shrinkage. Figure 4 gives a graphical representation of the model, showing the conditional relationships between parameters. The three boxes indicate the sets of parameters that are handled using three different computational strategies in the calculation of the posterior; see Section 3.1 for details.

### 2.4.1. Choice of prior parameters

The Hierarchical Adaptive Pólya Tree has three prior parameters for which values must be chosen. The most straightforward is the prior mean measure  $Q_0$ , which we typically chose from a parametric family. Some consideration must be given to the choice of the SIS prior parameters  $\Gamma_\tau$  and  $\Gamma_\nu$ .

Ma (2017) recommends an empirical Bayes approach to setting the  $\beta$  parameters, and where computationally feasible we may also recommend



**Fig. 4.** A graphical representation of the HAPT model. The boxes outline the parts of the model whose posteriors are computed with each of the three strategies described in Section 3.1. From left to right: The posterior of the  $Q_i$  conditional on other parameters is conjugate and can be integrated out numerically; the posterior of  $Q, \nu$ , and  $\tau$  conditional on  $S_\nu$  and  $S_\tau$  is approximated using quadrature; and the posterior of  $S_\nu$  and  $S_\tau$  is computed using HMM methods.

this approach. When the dataset is quite large or HAPT is embedded in a larger algorithm, as in Section 5.2, empirical Bayes estimation may not be practical. Based on our experience the model fit is not very sensitive to reasonable variation in this parameter; we suggest default values of  $\beta = 1$  for both  $\Gamma_\tau$  and  $\Gamma_\nu$ .

## 2.5. Comparison to existing models

The most prominent existing nonparametric models for estimation of related distributions are based on the Dirichlet process, including the Hierarchical Dirichlet process (HDP) (Teh et al., 2006), Nested Dirichlet process (NDP) (Rodríguez et al., 2008; Rodríguez and Dunson, 2014), and the hierarchical mixture of common and idiosyncratic Dirichlet process model of Müller, Quintana, and Rosner (MQR) (Müller et al., 2004). The Hierarchical Adaptive Pólya tree enjoys several advantages over these methods:

- (a) **Nonparametric estimation of cross-sample variation.** The HDP and NDP have a single scalar concentration parameter that controls the dependence across samples. MQR has one scalar parameter per sample that controls what proportion of the sample is explained by common structure and how much by idiosyncratic structure.

In contrast, because the concentration parameter  $\tau = \{\tau(A) : A \in \mathcal{A}\}$  in the HAPT is infinite-dimensional, HAPT places a highly flexible prior on the variation across samples, which allows it to learn spatially heterogeneous variation across samples. Indeed, the variation between samples at different locations of the sample space may be of primary interest in some scientific applications: learning where common structure is largely preserved between samples and where distributions vary widely may point the way to understanding important underlying phenomena.

- (b) **Computation.** HDP, NDP, and MQR all rely on MCMC methods to draw from the posterior. The Hierarchical Adaptive Pólya tree is not fully conjugate, but the necessary integration can be split into low-dimensional integrals and approximated extremely quickly using adaptive quadrature methods, without concerns about Markov chain convergence. See Section 3.1 for details. Note that this is not to say that one will never use MCMC in the presence of HAPT in any Bayesian model. In some inference tasks, one may embed HAPT into a more complex hierarchical model, whose other components may require MCMC for inference. In such cases, the computational tractability of HAPT implies that one can Rao-Blackwellize (i.e., marginalize out) the HAPT part within that MCMC algorithm for the more complex model. This will substantially simplify the MCMC sampler. We present one such application of HAPT in Section 5.2.
- (c) **Interpretability.** HAPT provides an easily interpretable estimate of the common structure: The posterior estimate of  $Q$  is both the estimate of the mean density across samples, and the posterior predictive distribution for a new sample. In contrast, the HDP estimates a discrete instead of continuous distribution, and the NDP does not provide any estimate of common structure. MQR provides an estimate of common structure, but it is neither the mean of sample distributions nor a posterior predictive estimate. Interpretation of the common structure in the MQR model is most straightforward if variation between samples involves contamination of an underlying distribution by an idiosyncratic process for each sample.

Other existing models such as the Hierarchical Pitman-Yor model (Teh, 2006) and those based on Normalized Random Measures (Regazzini et al., 2003) or more generally Poisson-Kingman models (Pitman, 2003) are subject, to various extents, to limitations similar to those of the models based on the Dirichlet process. Recent works such as Griffin et al. (2013); Griffin and Leisen (2017); Camerlenghi et al. (2017); Camerlenghi et al. (2018) provide more flexible means to modeling the dependency among multiple samples than the classical models, but the

aforementioned benefits of the Pólya tree remain even in view of these state-of-the-art models.

### 3. Bayesian inference and computation

The HAPT model is partially conjugate: the conditional posterior for  $Q_i \mid Q, \tau, \nu, \mathbf{S}_\tau, \mathbf{S}_\nu$  is a standard Pólya tree. Though not fully conjugate, the conditional posterior for  $Q, \tau, \nu \mid \mathbf{S}_\tau, \mathbf{S}_\nu$  can be reliably approximated using adaptive quadrature methods. The computational strategies used are described in Section 3.1

To derive the posterior we use the representation of the Pólya trees  $Q$  and  $Q_i$  in terms of Beta-distributed random variables  $\theta(A)$  and  $\theta_i(A)$  for each node  $A$  of the tree. With this notation, The second and third lines in (1) can be written in terms of the  $\theta$  and  $\theta_i$  as in Equation 2:

$$\begin{aligned}\theta_i(A) \mid \theta(A) &\stackrel{iid}{\sim} \text{Beta}(\theta(A)\tau(A), (1 - \theta(A))\tau(A)) \\ \theta(A) &\sim \text{Beta}(\theta_0(A)\nu(A), (1 - \theta_0(A))\nu(A)).\end{aligned}$$

Including the concentration parameters, we can write the posterior for the parameters of a region  $A$  in the following form conditional on the state parameters  $S_\tau(A), S_\nu(A)$ :

$$\begin{aligned}\pi(\theta(A), \tau(A), \nu(A) \mid S_\tau(A), S_\nu(A), \mathbf{X}) &\propto \\ \theta(A)^{\theta_0(A)\nu(A)-1} (1 - \theta(A))^{(1-\theta_0(A))\nu(A)-1} &\times \\ [B(\theta(A)\tau(A), (1 - \theta(A))\tau(A))]^{-k} &\times \\ \prod_{i=1}^k B(\theta(A)\tau(A) + n_i(A_\ell), (1 - \theta(A))\tau(A) + n_i(A_r)) &\times \\ \pi(\tau(A) \mid S_\tau(A))\pi(\nu(A) \mid S_\nu(A)). &\end{aligned}\tag{3}$$

where  $B(\cdot, \cdot)$  is the Beta function and the function  $n_i(\cdot)$  counts the number of observations from the  $i$ th sample contained in a region. The full posterior is the summation of Equation (3) over the possible states of  $\mathbf{S}_\tau$  and  $\mathbf{S}_\nu$ , with their respective priors factored in. The derivation of this posterior is given in the supplementary material.

#### 3.1. Computation

Posterior computation of the HAPT model requires three distinct computational strategies. We split the model (see Figure 4) into three parts, each of which requires a different approach. Each part of the model is conditioned on all parameters which are further to the right in Figure 4.

We first describe how to integrate out each of the first two parts. This reduces the problem to evaluating the posterior probabilities of all possible combinations of the state variables  $\mathbf{S}_\tau, \mathbf{S}_\nu$ , which can be accomplished using a forward-backward algorithm for hidden Markov models.

- (a)  $\pi(Q_i | Q, \tau, \nu, \mathbf{S}_\tau, \mathbf{S}_\nu, \mathbf{X})$ : The individual sample densities  $Q_i$ , conditional on all other parameters, are *a priori* distributed according to a standard Pólya tree. The corresponding conditional posterior is therefore also a Pólya tree. This allows us to analytically integrate out the  $Q_i$  when computing the posterior. If individual sample densities are of inferential interest their posteriors can easily be reconstructed after the main posterior computation is completed.
- (b)  $\pi(Q, \tau, \nu | \mathbf{S}_\tau, \mathbf{S}_\nu, \mathbf{X})$ : The remaining continuous parts of the joint model, namely the common structure  $Q$  and the continuous concentration parameters  $\tau$  and  $\nu$ , conditioned on the discrete state parameters  $\mathbf{S}_\tau$  and  $\mathbf{S}_\nu$ , are not conjugate and must be integrated numerically. Because all parameter dependence across nodes in the Pólya tree topology is restricted to the discrete state parameters, by conditioning on those parameters we are able to compute the posterior of the remaining parameters for each node of the tree independently.

This has two significant implications. First, rather than tackling a very high-dimensional integral over the product space of the parameters for all nodes, we have a much more tractable collection of low-dimensional integrals: we need only integrate the three-dimensional joint posterior of  $\theta(A), \tau(A), \nu(A)$  for each region  $A$  in the Pólya tree. Each of these integrals is tractable using standard quadrature techniques. Second, these integrals can be computed in parallel.

An additional observation allows us to further accelerate the adaptive quadrature. We note that the joint posterior distribution for  $\theta(A), \tau(A), \nu(A)$  conditional on  $S_\tau(A)$  and  $S_\nu(A)$ , with the other parameters integrated out, can be factored as

$$\pi(\theta(A), \tau(A), \nu(A) | S_\nu(A), S_\tau(A), \mathbf{X}) = g(\theta(A), \tau(A)) \times h(\theta(A), \nu(A)).$$

This allows us to factor the three dimensional integral:

$$\begin{aligned} & \iiint \pi(\theta(A), \tau(A), \nu(A) | S_\nu(A), S_\tau(A), x) d\tau(A) d\nu(A) d\theta(A) \\ &= \int \left[ \int g(\theta(A), \tau(A)) d\tau(A) \right] \left[ \int h(\theta(A), \nu(A)) d\nu(A) \right] d\theta(A), \end{aligned}$$

This factorization effectively reduces the dimensionality of the integral: rather than evaluating the unnormalized posterior at points

throughout a 3-dimensional space, we need only evaluate it on the union of two 2-dimensional spaces.

- (c) The posterior distribution of the state parameters  $\mathbf{S}_\tau$  and  $\mathbf{S}_\nu$  at first appears to be the most intimidating part of the model: It is a distribution over the product space of a large number of discrete parameters, resulting in an enormous number of level combinations. The naive computation of the joint posterior,

$$P\left(\bigcap_{A \in \mathcal{A}} S_\tau(A) = i_A, S_\nu(A) = j_A\right)$$

is straightforward but needs to be repeated for every possible combination of  $S_\tau(A)$  and  $S_\nu(A)$  for every node in the tree, which is computationally prohibitive. Here the Markov dependency structure comes to our rescue. The shrinkage states constitute a Hidden Markov Model following the tree topology (Crouse et al., 1998), and we can factor the joint distribution and calculate the posterior probabilities using a forward-backward algorithm in a manner analogous to inference strategies for linear Hidden Markov Models.

During the forward-backward algorithm we can compute expectations of any function that can be expressed in the form

$$f(\cdot) = \prod_A f^*(\theta(A), \tau(A), \nu(A)),$$

where  $f^*(\cdot)$  is an arbitrary function in  $L_1$ . This includes the marginal likelihood, the expected value of the estimated common density  $q(\cdot)$  or any individual sample density  $q_i(\cdot)$  at any given point, expectations of random variables  $Y \sim Q$  or  $Y_i \sim Q_i$ , and a wide variety of other functions, such as the variance function described in Section 5.1.

This computation is recursive, and we give a brief example of how it is carried out for the marginal likelihood. The previous two computational strategies give us the ability to calculate (up to quadrature approximation) the marginal likelihood, within a given node, of all remaining parameters conditional on  $\mathbf{S}_\tau$  and  $\mathbf{S}_\nu$ . Combining this with the prior transition parameters specified in  $\mathbf{\Gamma}$  we are able to calculate the posterior probabilities for  $S_\tau(A)$  and  $S_\nu(A)$ , and the overall marginal likelihood of the distribution on  $A$ , by considering the recursively-calculated marginal likelihoods of the child regions of  $A$  under each possible state.

Obviously this recursion requires a stopping point. The simplest method is to truncate the tree at a predetermined depth. Hanson

(2006) offers some guidance on how to choose the depth of a truncated Pólya tree based on sample size and other considerations. We recommend using as large a tree as is computationally feasible in order to minimize approximation errors due to truncation. If the data deviate very strongly from the prior distribution—as is common, for example, in high-dimensional settings—a more sophisticated approach may be required, such as truncating a branch of the tree when it reaches a depth where the node contains only a few data points.

## 4. Theoretical results

We describe several desirable theoretical properties of the HAPT model. Proofs are given in the supplementary material.

**THEOREM 1.** (*Absolute Continuity*) *Let  $Q, Q_1, \dots, Q_k$  be random measures distributed according to a HAPT model with base measure  $Q_0$ . If the SIS priors on  $\tau$  and  $\nu$  each include a complete shrinkage state that absorbs all possible chains in a finite number of steps with probability 1, then  $Q, Q_1, \dots, Q_k \ll Q_0$  with probability 1.*

**REMARK 1.** *A sufficient condition for the complete shrinkage state to absorb all chains in a finite number of steps with probability 1 is that the transition probability from every other state to the complete shrinkage state is bounded away from zero, which is satisfied by our choice of  $\Gamma$ .*

**REMARK 2.** *The absorbent, complete shrinkage state in the SIS prior is needed to ensure the absolute continuity of the random distributions. Wong and Ma (2010) showed that in lieu of decreasing the Beta variances along branches of the partition tree at a sufficiently fast fixed rate (Kraft, 1964), one can also ensure the absolute continuity of random distributions from the Pólya tree as long as the Beta variables will with probability 1 eventually have zero variance at deep enough levels almost everywhere on the sample space. The absorbent, complete shrinkage state in the SIS prior ensures this condition.*

**THEOREM 2.** (*Prior support*) *Let  $f, f_1, \dots, f_k$  be the probability density functions corresponding to arbitrary distributions that are absolutely continuous with respect to a measure  $\mu$  on  $\Omega$ . Let  $q, q_1, \dots, q_k$  be corresponding densities from a HAPT model satisfying*

- (a)  $Q_0$  and  $\mu$  are equivalent measures, that is  $Q_0 \ll \mu$  and  $\mu \ll Q_0$ ;
- (b) There are at least two shrinkage states, including the complete shrinkage state, at each level.

Then  $f, f_1, \dots, f_k$  are in the  $L_1$  prior support.

Our final result gives posterior consistency in the case where the samples have equal sample sizes:

**THEOREM 3. (Posterior consistency)** *Let  $\mathbf{D}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_k\}$  be observed data consisting of  $k$  independent samples, each of size  $n$ , from absolutely continuous distributions  $P_1, \dots, P_k$ . Let  $\pi(\cdot)$  be a Hierarchical Adaptive Pólya tree model on the  $k$  densities with overall prior mean  $Q_0^*$ , and let  $\pi(\cdot | \mathbf{D}_n)$  be the corresponding posterior. If  $P_i \ll Q_0^*$  for all  $i$  then we have posterior consistency under the weak topology. That is,*

$$\pi(U | \mathbf{D}_n) \rightarrow 1 \text{ as } n \rightarrow \infty$$

for any weak neighborhood  $U$  of the product measure  $P_1 \times \dots \times P_k$ .

## 5. Methodological applications of the HAPT

We present two ways in which the HAPT model can be applied to infer structures that existing models have not been able to capture. The first application is the ability of HAPT to model the “dispersion function” (defined below) on the sample space; we show how to calculate the posterior dispersion function from the  $\tau$  parameter. According to our knowledge, at the time of the writing, no existing model permits inference on the variation across sample densities in this manner. The second application is clustering samples based on their distributions, while allowing for within-cluster variation. While the Nested Dirichlet process clusters distributions, it allows no variation among the underlying sampling distributions within each cluster. The importance of allowing for variation within clusters was first pointed out by MacEachern (2008), who described a dependent Dirichlet process which would incorporate within-cluster variation.

### 5.1. Inferring the cross-sample dispersion function

The primary target of inference in problems with multiple samples is often the variation across samples. It is this inference, for example, which lends ANOVA its name, though the model is typically presented in terms of the overall and sample means.

In the HAPT model, variation across samples is captured by  $\tau$ , an infinite-dimensional parameter which characterizes variation at different locations and scales. Rather than trying to provide guidance on how to interpret the multiscale structure in  $\tau$ , we show how to recast it into an estimate of the variation across samples at any given point in the sample



space, giving us a posterior *dispersion function* analogous to the posterior mean function.

Let  $q_\star$  be density for a new sample drawn from the HAPT model, with corresponding Beta-distributed random variables  $\theta_\star(A)$  for each region  $A$  in the recursive partition. Since  $q_\star$  is random, we can estimate a “variance function”  $v : \Omega \rightarrow \mathbb{R}^+$  which gives, for any point  $x$  in the sample space, the expected variance of  $q_\star(x)$  conditional on the density of the common structure,  $q(\cdot)$ .

PROPOSITION 1. *The variance function is given by*

$$\begin{aligned} \mathbb{E}_q [\text{Var}(q_\star(x) \mid q)] &= q_0(x)^2 \times \\ &\mathbb{E}_{\mathbf{S}_\tau} \left[ \prod_{A \ni x} \|A\|^2 \mathbb{E}_{\theta(A), \tau(A)} \left( \frac{\theta(A)(\theta(A)\tau(A) + 1)}{\|A_\ell\|^2(\tau(A) + 1)} \middle| S_\tau(A) \right)^{\mathbf{1}(x \in A_\ell)} \right. \\ &\quad \mathbb{E}_{\theta(A), \tau(A)} \left( \frac{(1 - \theta(A))((1 - \theta(A))\tau(A) + 1)}{\|A_r\|^2(\tau(A) + 1)} \middle| S_\tau(A) \right)^{\mathbf{1}(x \in A_r)} - \\ &\quad \prod_{A \ni x} \|A\|^2 \mathbb{E}_{\theta(A), \tau(A)} \left( \frac{\theta(A)^2}{\|A_\ell\|^2} \middle| S_\tau(A) \right)^{\mathbf{1}(x \in A_\ell)} \\ &\quad \left. \mathbb{E}_{\theta(A), \tau(A)} \left( \frac{(1 - \theta(A))^2}{\|A_r\|^2} \middle| S_\tau(A) \right)^{\mathbf{1}(x \in A_r)} \right]. \end{aligned}$$

The derivation of this result is given in the supplementary material. The expectations with respect to  $\theta(A)$  and  $\tau(A)$  can be estimated during the same quadrature routines used to compute the posterior distributions of  $\theta(A)$ , described in Section 3.1. The expectation with respect to  $\mathbf{S}_\tau$  can then be calculated during the forward-backward routine used to calculate the posterior distribution of  $\mathbf{S}_\tau$ , as described in the same section.

We naturally expect more absolute variation across samples in areas where the densities of all the samples are higher, so we also introduce a standardized dispersion function measuring the coefficient of variation. The posterior mean coefficient of variation of  $q_\star$  at any given point is not analytically tractable; we can obtain an estimate by taking the square root of the variance function and dividing by the mean density function. We illustrate the application of the dispersion functions in Section 6.2 and Section 7.

## 5.2. Dirichlet Process Mixture of HAPT

Not only can the HAPT be used as a standard-alone Bayesian model, but it can also be embedded as a component in more complex hierarchical models to address a variety of inference tasks. Of course, in such cases

some parts of the larger hierarchical model may require MCMC for inference, but even then, one can numerically marginalize out the HAPT portion of the model, just like one could integrate out any other conjugate model component. This often results in very simple MCMC samplers for very sophisticated nonparametric models. We provide one illustration of such an inference task in this subsection.

In many applications we may not believe that the samples collected all share a single common structure. A more appropriate model may be that the samples are drawn from several latent populations, with samples being drawn from the same population having structure in common. In this case we may reconstruct the latent structure by clustering the samples. To learn the clustering of samples without fixing the number of clusters in advance, we add a Dirichlet process component to the model. We can write the model as follows:

$$\begin{aligned} X_{ij} &| Q_i \stackrel{iid}{\sim} Q_i \\ Q_i &| Q_i^*, \tau_i^* \stackrel{iid}{\sim} \text{PT}(Q_i^*, \tau_i^*) \\ (Q_i^*, \nu_i^*, \tau_i^*) &| G \stackrel{iid}{\sim} G \\ G &\sim \text{DP}(\alpha H(Q^*, \nu^*, \tau^*)), \end{aligned}$$

where the base measure can be factored as

$$H(Q^*, \nu^*, \tau^*) = \{\pi(Q^* | \nu^*) \times \pi(\nu^*)\} \times \pi(\tau^*)$$

with  $\pi(Q^* | \nu^*)$  being a  $\text{PT}(Q_0, \nu^*)$ ,  $\pi(\nu^*)$  a  $\text{SIS}(\mathbf{T}_\nu^*)$ , and  $\pi(\tau^*)$  a  $\text{SIS}(\mathbf{T}_\tau^*)$ . The Dirichlet process introduces clustering among the samples, so that some set of  $Q_i$ , belonging to the same cluster, have a cluster centroid  $Q_i^*$  and a cluster-specific dispersion parameter,  $\tau_i^*$ . The cluster centroids  $Q_i^*$  are also allowed to have different smoothness corresponding to the cluster-specific  $\nu_i^*$ . Conditional on the clustering structure, the model reduces to a collection of independent HAPT models. In other words, while the above model may first appear dauntingly complex, it is nothing but a Dirichlet process mixture model on the space of distributions using the HAPT as the mixing kernel along with the hyperprior on the kernel.

We call this model a Dirichlet Process Mixture of Hierarchical Adaptive Pólya Trees, or DPM-HAPT. It is comparable to the Nested Dirichlet process in the way it induces clustering among the samples, but is considerably more flexible. While the NDP requires that all samples in a cluster have identical distributions (MacEachern, 2008), DPM-HAPT allows the distributions within each cluster to vary according to the HAPT model. In addition, the advantages of the HAPT model discussed earlier, such as

flexible modeling of variation in different parts of the sample space, still apply.

Posterior computation for DPM-HAPT consists of a combination of standard Dirichlet Process methods and the HAPT posterior calculations described in Section 3.1. As noted above, conditional on the clustering structure, the model consists of a number of independent HAPT models. Although the HAPT model is not fully conjugate, our posterior computation strategy allows us to calculate the marginal likelihood to arbitrary precision. This allows the use of a Dirichlet process algorithm designed for conjugate mixture models. We use the Pólya urn representation to sample the clustering structure, using marginal likelihoods calculated from the HAPT model.

This algorithm requires the computation of

- (a) marginal likelihoods under the HAPT model for clusters including a single sample, and
- (b) marginal posterior predictive likelihoods under the HAPT model of one sample conditional on one or more other samples making up a cluster.

The first item is straightforward. The second is easily achieved by fitting the HAPT model twice, and calculating

$$f(\mathbf{X}_i \mid \mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_k}) = \frac{f(\mathbf{X}_i, \mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_k})}{f(\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_k})}.$$

## 6. Simulation results

### 6.1. Density estimation and comparison to MQR and a HDP mixture

In this section we use a simulation study to evaluate the performance of density estimation under the HAPT model and compare it to the MQR model (Müller et al., 2004), a Hierarchical Dirichlet Process Mixture of Beta distributions (HDPMBeta) (Teh et al., 2006), and a Linear Dependent Dirichlet Process Mixture model (LDDP) (Iorio et al., 2004). MQR and LDDP are available in the R package `DPpackage` (Jara et al., 2011) as the functions `HDPmdensity` and `LDDPdensity`, respectively. HDPMBeta is available in the R package `dirichletprocess` (Ross and Markwick, 2019) as the function `DirichletProcessHierarchicalBeta`.

MQR models  $k$  related densities with  $k + 1$  Dirichlet process mixtures. The density of each sample is modeled as a mixture of a common component and a unique idiosyncratic component, each with a Dirichlet process mixture prior. In all there are  $k$  Dirichlet process mixtures for the  $k$  idiosyncratic components and one for the common component. HDPMBeta

fits a mixture model using Beta kernels and a hierarchical Dirichlet Process as the mixing distribution. LDDP fits a mixture model using Normal kernels with a linear model on the kernel locations. In the case we consider, where we have no covariates, this results in sharing information by way of shrinking kernel locations across samples.

We evaluated HAPT, HDPMBeta, LDDP, and MQR under three simulation scenarios, with 6 or 10 random samples of 100 observations each. HAPT, HDPMBeta, and MQR were also compared in scenarios with 10 random samples of 100 observations each. (At this writing, the R function `LDDPdensity` for fitting LDDP becomes unstable as the number of samples increases beyond 6 in our simulation, and so we were not able to fit LDDP with 10 samples.) In each scenario the sample densities are mixtures of four components with random weights. The first scenario constructs a smooth density out of mixtures of Beta distributions. There are two modes with weights varying from sample to sample. This density is highly amenable to modeling with a mixture of Normal or Beta distributions as HDPMBeta, LDDP, and MQR do. The components are:

- (a) A Uniform(0,1) distribution, with expected weight 0.1;
- (b) A Beta(2,2) distribution, with expected weight 0.1;
- (c) A Beta(30,10) distribution, with expected weight 0.4;
- (d) A Beta(10,30) distribution, with expected weight 0.4.

The density under the mean weights is illustrated in Figure 5(a).

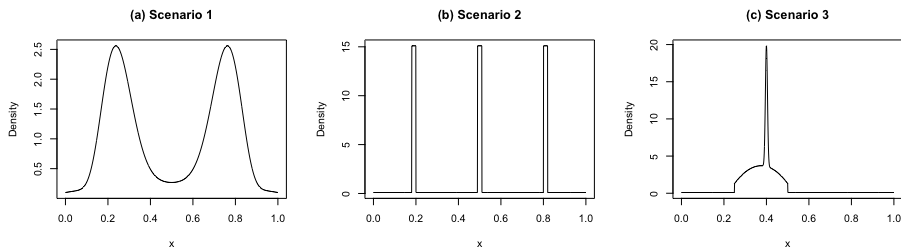
The second scenario has a low-weighted diffuse base distribution, with most of the mass being concentrated in 3 spikes. We expect HAPT to perform better than the other methods in this scenario, as a mixture of Normal or Beta distributions does not fit the narrow spikes well. The components are:

- (a) A Uniform(0,1) distribution, with expected weight 0.1;
- (b) A Uniform(0.18,0.20) distribution, with expected weight 0.3;
- (c) A Uniform(0.49,0.51) distribution, with expected weight 0.3;
- (d) A Uniform(0.80,0.82) distribution, with expected weight 0.3.

The density under the mean weights is illustrated in Figure 5(b)

The third scenario presents a mix of the two situations explored in the previous scenarios, with a larger-scale smooth component and a narrow spike. The components of this mixture are:

- (a) A Uniform(0,1) distribution, with expected weight 0.1;
- (b) A Uniform(.25,.5) distribution, with expected weight 0.3;
- (c) A scaled Beta(2,2) distribution, rescaled to cover the interval [.25, .5], with expected weight 0.4;



**Fig. 5.** The four-component mixtures used in the comparison between HAPT, HDPMBeta, LDDP, and MQR, under the expected weights. (a) shows the density for the first scenario (smooth structures easily modeled by mixtures of Normal distributions), (b) gives the density for the second scenario (narrow spikes not easily modeled by mixtures of Normal distributions), and (c) gives the density for the third scenario (both smooth structures and a narrow spike). Depending on the concentration of the Dirichlet distribution, the individual sample densities will vary more or less from this density.

(d) A Beta(4000,6000) distribution, with expected weight 0.2.

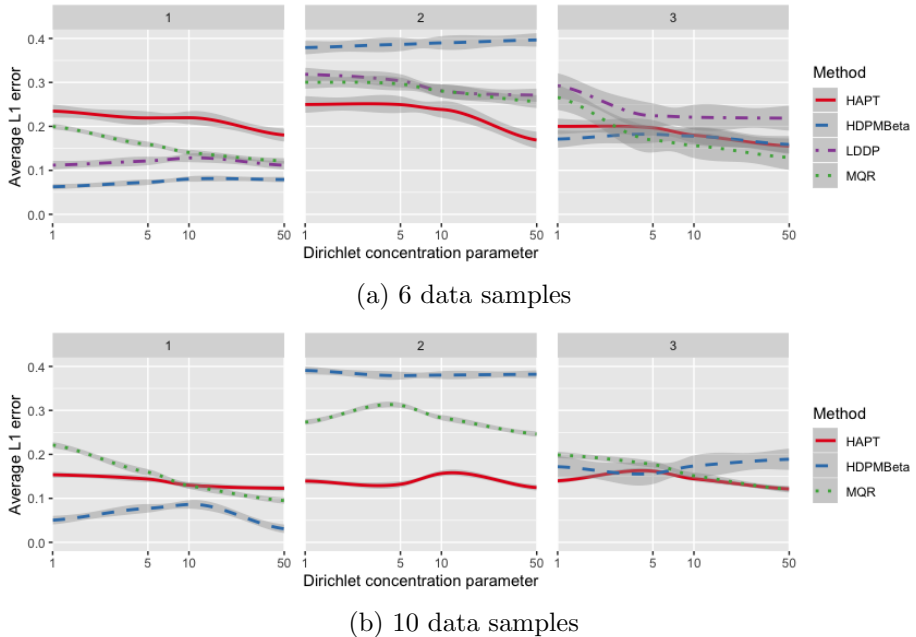
The density under the mean weights is illustrated in Figure 5(c).

In order to create samples with varying underlying true densities in each scenario, for each sample the weights of the mixture components were randomized according to a Dirichlet distribution with the expectations given above. The concentration of the Dirichlet distribution around the mean was varied across a wide range, with the sum of the parameters taking the values 1, 5, 10, and 50. For example, in the highest concentration setting, the weights of the four components were given by a Dirichlet(5,15,20,10) distribution.

We draw the multiple (6 or 10) underlying sample densities and then draw 100 points from each sample, for a total sample size of 600 or 1,000. After fitting all four models, we calculate the  $L_1$  distance between each sample's true density and the estimate of the sample density given by each of the two models. This is averaged across the six samples. We repeat this procedure twenty times for each value of the Dirichlet concentration parameter in order to gain a more accurate view of the performance of each model.

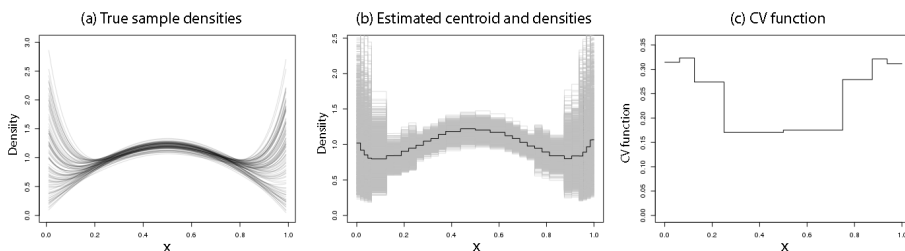
The results of this simulation are shown in Figure 6.1. The first scenario with its smooth Beta components is particularly friendly to HDPM-Beta (which is nearly the true model), LDDP, and MQR. In the second scenario HAPT unsurprisingly outperforms the other two methods by a wide margin across the board. In the third scenario the four methods are

more or less comparable in performance.



**Fig. 6.** Results of the simulation described in Section 6.1, comparing performance of HAPT, HDPMBeta, LDDP, and MQR across varying values of the Dirichlet concentration parameter. (a) presents the result when there are 6 data samples; (b) presents the results when there are 10 data samples. The LDDPdensity function available in R does not run when the data includes more than 6 samples, so (b) does not include results for LDDP. The performance of HAPT improves most substantially as the number of samples increases.

HDPMBeta and LDDP both show much more variable performance in the third scenario compared to the other two methods and the other two scenarios. This appears to be due to the models failing to capture the spike component when run on some of the simulated datasets, resulting in high L1 errors for those simulation runs. These simulations were conducted with the package’s default values for prior parameters, and it may be possible to avoid this issue with careful selection of the prior parameters. This would decrease the variability as well as improve the mean L1 score for HDPMBeta and LDDP on this scenario. Overall, this simulation demonstrates that HAPT compares favorably to HDPMBeta, LDDP, and MQR, particularly when the densities contain features not easily modeled by a mixture of smooth distributions. Moreover, the performance of HAPT improves much more substantially as the number of



**Fig. 7.** (a) One hundred sample densities from the simulation setting used in Section 6.2; (b) Estimated centroid and sample densities after fitting the HAPT model; (c) Estimated coefficient of variation function.

samples increases, indicating its ability to effectively borrow information across samples through the hierarchical structure.

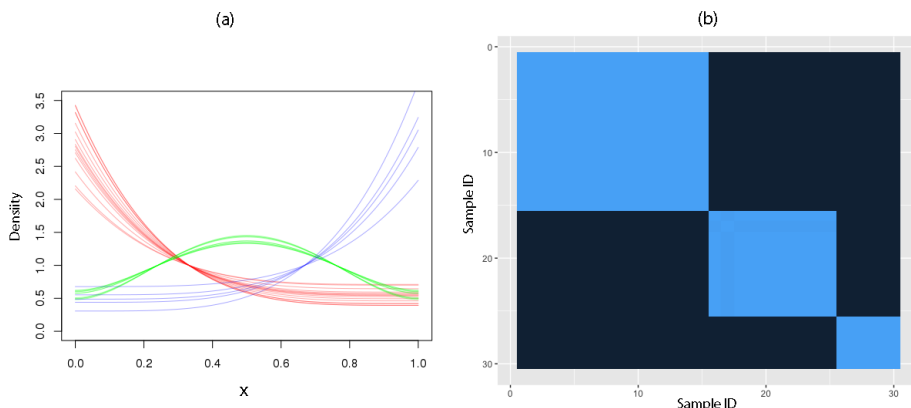
## 6.2. Estimation of the dispersion function

To simulate data with dispersion that varies across the sample space, we simulate from a mixture of three Beta distributions, such that the variation across sample densities is low in the middle of the space and much higher near zero and one. Specifically, each sample is a mixture of three components:  $\text{Beta}(2,2)$ ,  $\text{Beta}(1,12)$ , and  $\text{Beta}(12,1)$ . The corresponding weights  $w_1, w_2, w_3$  of the three components are drawn according to the following scheme. First we draw  $w_1 \sim \text{Beta}(80, 20)$ . Then we draw  $v \sim \text{Beta}(1, 1)$  and set  $w_2 = v(1 - w_1)$  and  $w_3 = (1 - v)(1 - w_1)$ . This results in the central part of the sample space having a small amount of variation between samples, while the edges on either side have much more variation. Sample densities are illustrated in Figure 7. One hundred sample densities are plotted in Figure 7(a). The variation in the dispersion of the sample densities can be seen clearly.

Figure 7(b) shows the estimated centroid and sample densities from fitting the HAPT model. The dispersion function, or estimated coefficient of variation is plotted in Figure 7(c). The dispersion function clearly shows how the variation across samples is low near the center of the space and high on either end.

## 6.3. DPM-HAPT Simulations

We simulate a simple 1-dimensional example to demonstrate the clustering behavior of the DPM-HAPT model. The simulation contains 30 samples belonging to three true clusters, with 15, 10, and 5 samples respectively. Each sample is drawn from a mixture of a  $\text{Uniform}(0,1)$  distri-



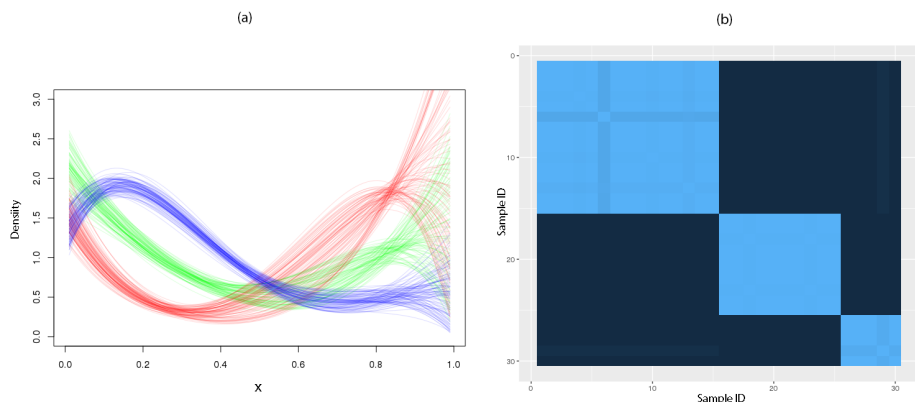
**Fig. 8.** (a) Sample densities for the clustering simulation, and (b) probability of two samples clustering together based on the DPM-HAPT model, which clearly identifies the three clusters that exist in this simulation, despite significant variation across samples within clusters.

bution and a Beta distribution, with the parameters of the Beta varying by cluster:  $\text{Beta}(1,5)$  for the first cluster,  $\text{Beta}(3,3)$  for the second cluster, and  $\text{Beta}(5,1)$  for the third cluster. The weights of the two components are randomized in each sample. The weight of one component is drawn from a  $\text{Beta}(10,10)$  distribution, which creates weights varying approximately between 0.3 and 0.7, with the actual observed proportions in realized samples varying more widely due to the additional Binomial variation. Sample densities are plotted in Figure 8(a). Each sample contains  $n = 300$  points.

An MCMC sampler is run using the Pólya urn scheme to sample the clustering structure. We summarize the results by looking at how often each of the 30 samples is clustered together with each other sample. These results are plotted in Figure 8(b). We can see in the figure that the DPM-HAPT model clearly identified the three clusters.

We now consider an example in which the variation across samples differs across the sample space. Samples with heterogeneous variation are drawn from mixtures of four Beta distributions, with varying parameters for each cluster. The parameter are chosen so that each cluster has much higher variation across sample densities on the right half of the space than on the left half. The data generating process is described in the supplementary material. We consider 30 samples belonging to three true clusters, as above. Figure 9 (a) shows one hundred draws from each of three clusters used in this simulation, to illustrate the variability between





**Fig. 9.** (a) One hundred draws from each of three clusters in the heterogeneous variation clustering example; (b) Probability of two samples clustering together based on 1,000 MCMC draws for the heterogeneous variance example. Despite the variation in the dispersion of the densities, HAPT-DPM clearly identifies the three true clusters.

and within clusters and the heterogeneity across the sample space.

As in the previous example, we simulate three clusters with 15, 10 and 5 samples respectively. We draw 150 observations from each sample, and apply DPM-HAPT to cluster the resulting samples. We run the MCMC for 1,000 draws after burnin; estimated coclustering probabilities are plotted in Figure 9 (b). The three true clusters are clearly identified even in the presence of substantial heterogeneity.

## 7. Application: DNase-seq data

DNase sequencing (DNase-seq) is a method used to identify regulatory regions of the genome (Song and Crawford, 2010). DNA is treated with Deoxyribonuclease (DNase) I, an enzyme that cuts the DNA strand. The cut strands are then sequenced and the locations of the cuts are identified and tallied. The vulnerability of the DNA strand to DNase varies by location, resulting in a distribution of cut counts which is nonuniform. The density of this distribution is related to various biological factors of interest: for example, it tends to be high near potential binding sites for transcription factors, since these proteins require access to the DNA strand in much the same way as DNase I, but will be low if a transcription factor already bound at that site blocks access to the DNA strand.

We consider the problem of clustering DNase-seq profiles near potential transcription binding sites, identified by a specific genetic motif. Each

sample consists of observed counts in a range of 100 base pairs on either side of one occurrence of the motif. A single motif, consisting of 10–20 base pairs, may appear tens of thousands of times in the genome, with each occurrence presenting one sample for analysis. Many samples, however, have very few cuts observed. For analysis we restrict ourselves to samples which meet a minimum sample size threshold.

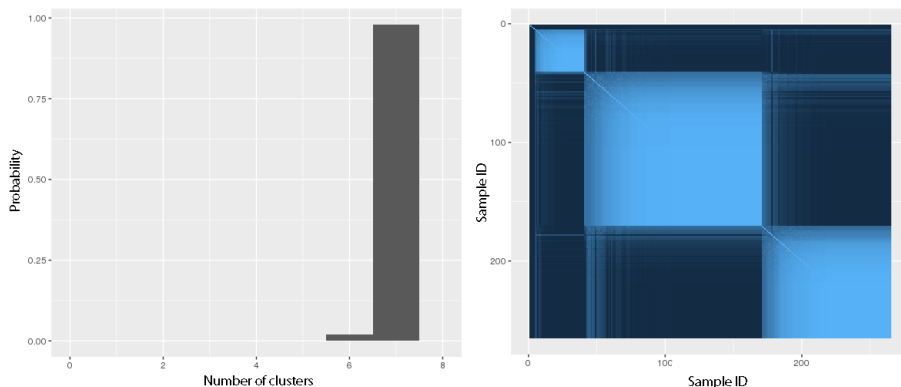
Different locations where the transcription factor motif of interest appears may be expected to show different DNase behavior in the region around the motif for a variety of reasons. This makes clustering a more appropriate approach to the problem than treating all the samples as having a single common structure. Identifying clusters of locations which have similar DNase-seq profiles may reveal previously unrecognized factors. We also expect within-cluster variation above and beyond sampling variation, which makes the Nested Dirichlet process unsuitable.

Here we present data from the ENCODE project (ENCODE Project Consortium, 2012) for locations surrounding a motif associated with the RE1-silencing transcription factor (REST). REST suppresses neuronal genes in non-neuronal cells (Chong et al., 1995). The data includes 48,549 locations where the REST motif appears in the genome. The motif consists of 21 base pairs, and the data includes an additional 100 base pairs on each side, for a total of 221 base pairs. In all, 922,704 cuts were tallied, an average of 19 per location. 468 locations have zero cuts observed. The number of cuts per location is heavily right skewed, with a median of 13 observations, first and third quartiles of 7 and 21 respectively, and a maximum of 2,099 cuts observed in a single sample.

For this analysis we restrict ourselves to locations which have at least 200 observations, a total of 265 samples. These samples include a total of 70,225 observations, an average of 330 observations per sample. The distribution is still quite skewed, with a minimum of 201 observations and a maximum of 2099. The median is 279 and the first and third quartiles are 232 and 366 observations. Histograms of 30 of the 265 locations are shown in the supplementary material.

We fit the DPM-HAPT clustering model to this data, using 100 post-burnin draws for inference. The model estimates 7 clusters with high probability (see Figure 10(a)), of which there are three large clusters and four singleton locations, each of which consists of a single large spike. One of the singleton locations occasionally joins one of the larger clusters, resulting in 6 clusters. A heatmap of the clustering structure is shown in Figure 10(b). The clusters are clearly differentiated and vary in size, with the largest cluster containing about 130 locations, though the cluster sizes vary from iteration to iteration due to uncertainty in the cluster assignment.

The estimated posterior mean densities of the three largest clusters are

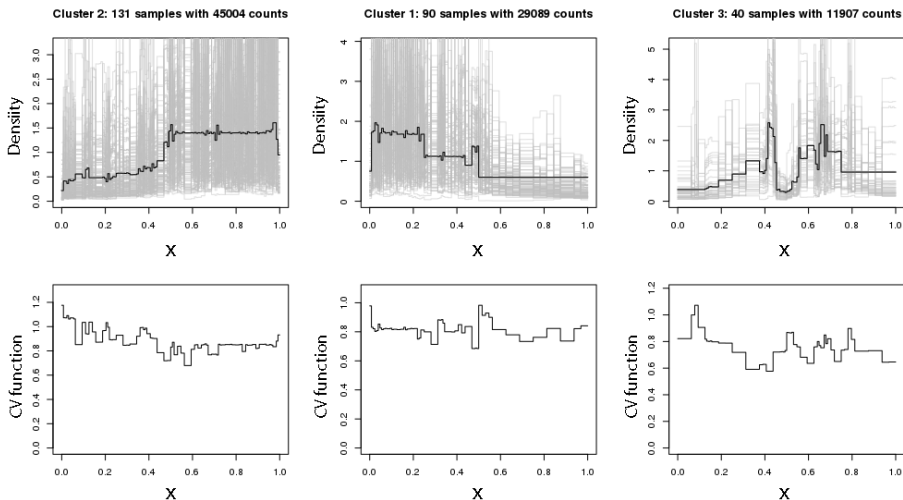


**Fig. 10.** (a) Posterior distribution of the number of clusters from the DPM-HAPT in the DNase-seq application; (b) the model shows clear clustering of the samples.

plotted in Figure 11. One of these clusters includes locations with cuts which are roughly symmetric around the transcription factor binding site with a dip in the middle. The central dip indicates that a transcription factor may be bound at that location, blocking access for the DNase I molecule. The other two largest clusters include locations which have cuts heaped up on one side or another of the binding site, suggestive of larger-scale patterns in the DNase I sensitivity due to factors like the folded structure of the DNA. The four singleton clusters show other densities which do not conform to the general patterns of the three largest clusters. The plots show substantial variation around the cluster centroid, much more than can be explained by sampling variation alone. The corresponding estimated dispersion functions are plotted in the second row, and are quite noisy; this is not surprising given that they are cross-sample dispersions (a second-order parameter) estimated with relatively few degrees of freedom.

## 8. Discussion

The HAPT model offers a compelling alternative to existing nonparametric models that share information across multiple samples. The Pólya tree’s ability to directly model the density of an absolutely continuous distribution frees us from the necessity of using mixture models to obtain densities, while the computational tractability of the posterior avoids the need to run MCMC chains for posterior inference. The addition of the



**Fig. 11.** Posterior mean densities and estimated dispersion functions of the three largest clusters in the DNase-seq example. The heavy black line shows the cluster centroid; light gray lines in the background show the estimated means of each sample in the cluster. The dispersion function is plotted below.

SIS prior allows us not only to more accurately model the densities of interest, but also to estimate a fully nonparametric dispersion function over the sample space. The model extends easily—both conceptually and computationally—to the setting where we do not believe all our samples have the same common structure, where DPM-HAPT allows us to learn both clustering structure and the distributional structure within each cluster.

Although we have presented HAPT in a one-dimensional space for the sake of clarity, adoption of the randomized recursive partitioning scheme first introduced in Wong and Ma (2010) allows the extension of HAPT to model densities in multidimensional spaces. Variables other than simple continuous ones can also be handled naturally—all that is needed is the definition of an appropriate recursive partition. This allows inclusion of categorical and ordinal-valued variables, as well as more exotic possibilities: a continuous variable that lives on the surface of a torus, a partially ordered categorical variable, or a zero-inflated variable with a point mass at zero and a continuous component on the positive halfline.

The Pólya tree’s decomposition of the density space into orthogonal Beta-distributed random variables, which extends to HAPT, is central to HAPT’s computational efficiency. It also allows the performance of quick

online updates in the HAPT model: when a new data point arrives we only need to update the nodes of the Pólya tree which contain the new data point, rather than recomputing the entire posterior. In a HAPT truncated at a depth of  $L$  levels, this means we need to reevaluate the posteriors of only  $L$  nodes, rather than  $2^L$ . HAPT may thus be used in streaming data settings where fast online updates are essential. This is one of our current developments.

The DPM-HAPT model illustrates the flexibility of HAPT as a component of larger hierarchical models as well as the computational tractability resulting from the “almost” conjugacy of the HAPT. Another such example is a mixture of HAPT model. Specifically, one well-recognized limitation of the Pólya tree and related models is its dependence on the choice of a particular partition tree on the sample space, which is typically determined by the quantiles of a given base measure  $Q_0$ . Hanson (2006) shows that by treating the base measure, and thus the partition tree, as an unknown and place a hyperprior on it, one can arrive at the so-called mixture of Pólya trees, which overcomes the sensitivity to the choice of the partition, and moreover, very simple MCMC strategies can be used for such models because the Pólya tree can be analytically marginalized out given each base measure. We note that exactly the same extension can be applied to the HAPT by placing hyperpriors on the base measure. The same samplers that work for the mixture of Pólya trees will work for the mixture of HAPT as the HAPT also be integrated out numerically.

A common phenomenon in applying mixture models for clustering is that when the variance or dispersion parameter of the kernel is specified to be too large, the model will reduce down to a single massive cluster *a posteriori*. This can be a concern especially when the mixing kernel is as flexible as a Bayesian nonparametric model such as the PT. As in classical mixture models, what allows the DPM-HAPT to properly identify clusters is the constraints it imposes on the underlying within cluster dispersion. As such the choice of the prior on the dispersion play an important role. In particular, this prior should allow the model to place more prior mass around specification of clusters in which densities share general common structures than around massive clusters without much similarity. We have adopted the SIS prior for this purpose as a general choice, but in specific applications, some modification to the prior that incorporates the background knowledge might be desirable.

Finally, we note that the DPM-HAPT model may also be easily extended by replacing the Dirichlet process with any cluster-inducing process which admits inference given the marginal likelihoods and conditional probabilities of the clusters. This includes, for example, the Pitman-Yor process. This allows the properties of the clustering process to be adapted if the clustering assumptions implicit in the Dirichlet process are not ap-

appropriate.

## Software

An R package, `HAPT-package`, which implements the HAPT model is available at <https://github.com/MaStatLab/HAPT-package>. We used the R package `DPpackage` (Jara et al., 2011) to fit the MQR and LDDP models, and used the R package `dirichletprocess` to fit the HDPMBeta model. For readers interested in fitting the nested DP model, we note that though no R packages currently implements this model, the R code for a recent paper (Zuanetti et al., 2018) provides an implementation.

## Acknowledgment

LM’s research is supported by NSF grants DMS-1612889 and DMS-1749789. We would like to thank the editors and three referees for very helpful comments.

## References

- Anderson, T. W. (1962) On the Distribution of the Two-Sample Cramer-von Mises Criterion. *The Annals of Mathematical Statistics*, **33**, 1148–1159.
- Barrios, E., Lijoi, A., Nieto-Barajas, L. E. and Prünster, I. (2013) Modeling with Normalized Random Measure Mixture Models. *Statistical Science*, **28**, 313–334.
- Beal, M. J., Ghahramani, Z. and Rasmussen, C. E. (2002) The Infinite Hidden Markov Model. In *Advances in Neural Information Processing Systems 14*. MIT Press.
- Berger, J. O. and Guglielmi, A. (2001) Bayesian and conditional frequentist testing of a parametric model versus nonparametric alternatives. *Journal of the American Statistical Association*, **96**, 174–184.
- Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I. and Rodríguez, A. (2018) Latent nested nonparametric priors. *ArXiv e-prints*.
- Camerlenghi, F., Lijoi, A. and Prünster, I. (2017) Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis*, **156**, 18–28.
- Chen, Y. and Hanson, T. E. (2014) Bayesian nonparametric k-sample tests for censored and uncensored data. *Computational Statistics & Data Analysis*, **71**, 335 – 346.

- Chong, J. A., Tapia-Ramirez, J., Kim, S., Toledo-Aral, J. J., Zheng, Y., Boutros, M. C., Altshuller, Y. M., Frohman, M. A., Kraner, S. D. and Mandel, G. (1995) REST: A mammalian silencer protein that restricts sodium channel gene expression to neurons. *Cell*, **80**, 949–957.
- Crouse, M. S., Nowak, R. D. and Baraniuk, R. G. (1998) Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing*, **46**, 886–902.
- ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
- Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, **1**, 209–230.
- (1974) Prior distributions on spaces of probability measures. *Ann. Statist.*, **2**, 615–629.
- Filippi, S. and Holmes, C. C. (2017) A Bayesian nonparametric approach to testing for dependence between random variables. *Bayesian Anal.*, **12**, 919–938.
- Freedman, D. A. (1963) On the asymptotic behavior of bayes’ estimates in the discrete case. *Ann. Math. Statist.*, **34**, 1386–1403.
- Griffin, J. E., Kolossiatis, M. and Steel, M. F. J. (2013) Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **75**, 499–529.
- Griffin, J. E. and Leisen, F. (2017) Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 525–545.
- Hanson, T. E. (2006) Inference for Mixtures of Finite Pólya Tree Models. *Journal of the American Statistical Association*, **101**, 1548–1565.
- Holmes, C. C., Caron, F., Griffin, J. E. and Stephens, D. A. (2015) Two-sample Bayesian nonparametric hypothesis testing. *Bayesian Anal.*, **10**, 297–320.
- Iorio, M. D., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004) An anova model for dependent random measures. *Journal of the American Statistical Association*, **99**, 205–215. URL: <https://doi.org/10.1198/016214504000000205>.
- Jara, A., Hanson, T., Quintana, F., Müller, P. and Rosner, G. (2011) DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software*, **40**, 1–30. URL: <http://www.jstatsoft.org/v40/i05/>.

- Jara, A. and Hanson, T. E. (2011) A class of mixtures of dependent tail-free processes. *Biometrika*, **98**, 553–566. URL: <http://dx.doi.org/10.1093/biomet/asq082>.
- Kolmogorov, A. (1933) Sulla Determinazione Empirica di una Legge di Distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 83–91.
- Kraft, C. H. (1964) A class of distribution function processes which have derivatives. *Journal of Applied Probability*, **1**, 385–388.
- Lavine, M. (1992) Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, **20**, 1222–1235.
- (1994) More aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.*, **22**, 1161–1176.
- Ma, L. (2017) Adaptive Shrinkage in Pólya Tree Type Models. *Bayesian Analysis*, **12**, 779–805.
- Ma, L. and Wong, W. H. (2011) Coupling optional Pólya trees and the two sample problem. *Journal of the American Statistical Association*, **106**, 1553–1565.
- MacEachern, S. N. (1999) Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999.
- (2008) Discussion of "The nested Dirichlet process" by A.E. Gelfand, D.B. Dunson and A. Rodriguez. *Journal of the American Statistical Association*, **103**, 1149–1151.
- Mauldin, R. D., Sudderth, W. D. and Williams, S. C. (1992) Pólya trees and random distributions. *Ann. Statist.*, **20**, 1203–1221.
- Muliere, P. and Walker, S. (1997) A Bayesian Non-parametric Approach to Survival Analysis Using Pólya Trees. *Scandinavian Journal of Statistics*, **24**, 331–340.
- Müller, P., Quintana, F. and Rosner, G. (2004) A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 735–749.
- Nieto-Barajas, L. E. and Quintana, F. A. (2016) A bayesian non-parametric dynamic ar model for multiple time series analysis. *Journal of Time Series Analysis*, **37**, 675–689.
- Pitman, J. (2003) *Poisson-Kingman partitions*, vol. Volume 40 of *Lecture Notes–Monograph Series*, 1–34. Beachwood, OH: Institute of Mathematical Statistics.



- Regazzini, E., Lijoi, A. and Prünster, I. (2003) Distributional results for means of normalized random measures with independent increments. *Ann. Statist.*, **31**, 560–585.
- Rodriguez, A. and Dunson, D. B. (2014) Functional clustering in nested designs: Modeling variability in reproductive epidemiology studies. *Ann. Appl. Stat.*, **8**, 1416–1442.
- Rodríguez, A., Dunson, D. B. and Gelfand, A. E. (2008) The nested Dirichlet process. *Journal of the American Statistical Association*, **103**, 1131–1154.
- Ross, G. J. and Markwick, D. (2019) *dirichletprocess: Build Dirichlet Process Objects for Bayesian Modelling*. URL: <https://CRAN.R-project.org/package=dirichletprocess>. R package version 0.3.0.
- Schörgendorfer, A. and Branscum, A. J. (2013) Regression analysis using dependent polya trees. *Statistics in medicine*, **32**, 4679–4695.
- Song, L. and Crawford, G. E. (2010) DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols*, **2010**, pdb.prot5384.
- Soriano, J. and Ma, L. (2017) Probabilistic multi-resolution scanning for two-sample differences. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**, 547–572.
- Stein, C. (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 197–206. Berkeley, Calif.: University of California Press.
- Teh, Y. W. (2006) 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*, ACL-44, 985–992. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Teh, Y. W. and Jordan, M. I. (2010) *Hierarchical Bayesian nonparametric models with applications*, 158–207. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006) Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**, 1566–1581.
- Tomlinson, G. and Escobar, M. (1999) Analysis of densities.

- Westenberg, J. (1948) Significance test for median and interquartile range in samples from continuous populations of any form. *Proceedings Koninklijke Nederlandse Akademie van Wetenschappen*, **51**, 252–261.
- Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, **1**, 80–83.
- Wong, W. H. and Ma, L. (2010) Optional Pólya tree and Bayesian inference. *Ann. Statist.*, **38**, 1433–1459.
- Zhao, L. and Hanson, T. E. (2011) Spatially dependent polya tree modeling for survival data. *Biometrics*, **67**, 391–403.
- Zuanetti, D. A., Müller, P., Zhu, Y., Yang, S. and Ji, Y. (2018) Clustering distributions with the marginalized nested dirichlet process. *Biometrics*, **74**, 584–594. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12778>.