JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 119, NO. 545, 189–201: Theory and Methods
https://doi.org/10.1080/01621459.2022.2105223

Hidden Markov Pólya Trees for High-Dimensional Distributions

Naoki Awaya and Li Ma[a]

Department of Statistical Science, Duke University, Durham, NC

ABSTRACT

The Pólya tree (PT) process is a general-purpose Bayesian nonparametric model that has found wide application in a range of inference problems. It has a simple analytic form and the posterior computation boils down to beta-binomial conjugate updates along a partition tree over the sample space. Recent development in PT models shows that performance of these models can be substantially improved by (i) allowing the partition tree to adapt to the structure of the underlying distributions and (ii) incorporating latent state variables that characterize local features of the underlying distributions. However, important limitations of the PT remain, including (i) the sensitivity in the posterior inference with respect to the choice of the partition tree, and (ii) the lack of scalability with respect to dimensionality of the sample space. We consider a modeling strategy for PT models that incorporates a flexible prior on the partition tree along with latent states with Markov dependency. We introduce a hybrid algorithm combining sequential Monte Carlo (SMC) and recursive message passing for posterior sampling that can scale up to 100 dimensions. While our description of the algorithm assumes a single computer environment, it has the potential to be implemented on distributed systems to further enhance the scalability. Moreover, we investigate the large sample properties of the tree structures and latent states under the posterior model. We carry out extensive numerical experiments in density estimation and two-group comparison, which show that flexible partitioning can substantially improve the performance of PT models in both inference tasks. We demonstrate an application to a mass cytometry dataset with 19 dimensions and over 200,000 observations. Supplementary Materials for this article are available online.

ARTICLE HISTORY
Received November 2020
Accepted July 2022

KEYWORDS
Bayesian nonparametrics;
Latent variable models;
Multi-scale inference;
Multivariate analysis;
Recursive partitioning

1. Introduction

The Pólya tree (PT) (Freedman 1963; Ferguson 1974; Lavine 1992) is a stochastic process that generates random probability measures and is introduced as a prior for Bayesian nonparametric inference. While the PT generalizes the Dirichlet process (DP) (Ferguson 1973) as it yields the DP under certain hyperparameters (Ferguson 1974), the statistical properties and practical applications of the PT are very different. While the DP is most frequently used as a mixing distribution that induces clustering structures, the PT is often adopted for directly modeling probability densities.

The PT is defined generatively on a recursive partition—or a partition tree—over the sample space through coarse-to-fine sequential probability assignment at each tree split. In a classical (univariate) PT, the tree is dyadic and the conditional probability assigned to the two children nodes at each tree split arises from independent beta priors, which leads to analytic simplicity and ease in computing the posterior. Obtaining the posterior is straightforward from beta-binomial conjugacy and incurs a computational budget that scales only linearly with the sample size, making the PT one of the few nonparametric models applicable to data with massive sample sizes. Moreover, the posterior computation is embassingly parallelizable across the tree nodes.

The PT has been applied in various contexts beyond the original application of density estimation. A far-from-exhaustive list includes survival analysis (Muliere and Walker 1997; Neath 2003), imputing missing values (Paddock 2002), goodness-of-fit tests (Berger and Guglielmi 2001), two-group comparison (Ma and Wong 2011; Chen and Hanson 2014; Holmes et al. 2015; Soriano and Ma 2017), density regression (Jara and Hanson 2011), ANOVA (Ma and Soriano 2018), testing independence (Filippi and Holmes 2017), and hierarchical modeling (Christensen and Ma 2020). The PT has also been used in semi-parametric analyses such as in (generalized) linear models (Walker et al. 1999; Walker and Mallick 1997; Hanson and Johnson 2002).

Early developments of the PT are based on an a priori fixed partition tree on the sample space. The resulting inference can be sensitive to the choice of the partition points defining the tree. In particular, the resulting process, both a priori and a posteriori, can be jumpy at these points. In hypothesis testing and model choice, this sensitivity is also reflected in the sometimes substantial change in the marginal likelihood/Bayes factor when the partition points are slightly varied. To remedy the issue, Paddock et al. (2003) modified the PT model so that observations are generated from the PT model with slightly different partition points. Hanson and Johnson (2002) and Hanson (2006) proposed a mixture of PTs by defining partition

CONTACT Li Ma li.ma@duke.edu Department of Statistical Science, Duke University, Durham, NC 27708.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

points along quantiles of a parametric model endowed with a hyperprior to allow model averaging on the partition points. This strategy does not allow individual partition points to adapt to local features of the distribution but only the whole set of points to the global structure of the distribution, and is most effective when the underlying density is close to the specified parametric model. Nieto-Barajas and Müller (2012) took a different approach by modeling the probability assignments within each level of the tree in a correlated manner to smooth out the random measure over the boundaries of partitioning. While these approaches alleviate the sensitivity to partition points in low-dimensional settings, they are not easily applicable (though in principle possible) to problems with even just a handful of dimensions. Moreover, Bayesian inference with these models generally require Markov Chain Monte Carlo (MCMC), whose effectiveness can (in fact often does) still suffer from the sensitivity with respect to the partition points.

Another related issue regarding the partitioning scheme of the PT is that in multivariate problems, traditionally the partition tree is constructed by dividing all dimensions of the sample space at each split. For example, for a d-dimensional sample space, each time a tree node is divided, it is split into 2^d children nodes, and probability assignment over these 2^d child nodes is modeled by independent Dirichlet priors. Wong and Ma (2010) noted that such a "symmetric" partition scheme is undesirable as the dimensionality increases, in which case due to the exponential growth of the partition blocks, the vast majority of the blocks are barely, if at all, populated by data. As such they propose to incorporate adaptivity into the partitioning strategy with respect to the structure of the underlying distribution through adopting a Bayesian CART-like prior (Chipman, George, and McCulloch 1998) on the space of dyadic partition trees.

However, in order to maintain the analytic simplicity of the posterior and achieving MCMC-free exact Bayesian inference with a linear computational budget, the Bayesian CART-like prior has to be restricted to only divide at the middle point (or otherwise a predetermined fixed point) on one of the dimensions at each tree split. Not only does this hamper the model's ability to adapt to distributional structures, but it makes the model suffer from the same sensitivity with respect to the partition points. Moreover, even with this restriction, the inference algorithm (based on recursive message passing) is only computationally practical for up to about 10 dimensions on continuous sample spaces.

In a different vein, recent developments have demonstrated that aside from enhancing the partitioning strategy, the PT can also be substantially improved by adopting more flexible priors (as opposed to independent betas) on the probability assignment at each tree split (Jara and Hanson 2011; Nieto-Barajas and Müller 2012; Ma 2017). One strategy for enriching the PT in this regard is by introducing latent state variables at each tree split and adopt priors on the probability assignment *given* these states. When the latent states are discrete and modeled with Markov dependency, analytical simplicity is preserved and exact Bayesian inference can proceed through recursive message passing that maintains the linear computational budget (Ma 2017).

Given these developments, we are motivated by the following questions: Is it possible to incorporate into the PT a very flexible partition tree prior, such as the general Bayesian CART (i.e., without the restriction to partition at middle points), that will (i) enhance its adaptivity to distributional structures in multivariate settings; (ii) resolve its sensitivity to the choice of partition points; and (iii) allow a tractable form of the joint posterior and a posterior inference algorithm that is scalable to moderately high-dimensional problems (e.g., up to 100 dimensions)? Moreover, should such a strategy exist, can the resulting model and inference algorithm be made compatible with incorporating (possibly Markov dependent) latent states on the tree nodes?

The goals of making the partition tree prior more flexible while enhancing the computational scalability appear at odds with each other. Large tree spaces are well known to be very hard to compute over. In moderate- to high-dimensional settings exact inference involving flexible tree structures is beyond reach and even the most advanced MCMC approaches tailor-made for trees encounter substantial difficulty due to the pervasive multimodality of distributions in such spaces. Recent advances in sequential Monte Carlo (SMC) for regression tree models (Lakshminarayanan, Roy, and Teh 2013; Lu, Jiang, and Wong 2013), however, suggest that efficient inference is possible in moderately high-dimensional settings (up to about 100 dimensions). Moreover, once the partition tree is sampled, the conditional posterior for the rest of the model can be computed analytically through recursive message passing. We will therefore exploit a hybrid strategy that uses a new SMC sampler to efficiently sample from the marginal posterior of the partition tree structure, along with recursive message passing to compute the exact conditional posterior of the latent state variables given the tree.

Beyond the methodological development, we will also investigate the theoretical properties of the posterior on the partition tree and the latent states. Previous theoretical literature on the PT and related models have mostly focused on establishing the posterior consistency and the contraction rate of the random measure induced under these models (Castillo 2017; Castillo and Randrianarisoa 2021). In multivariate settings, however, the partition tree itself is highly informative about the underlying distribution. Moreover, in applications involving model choice and hypothesis testing, it is often the latent states, not the random measures, that are of direct interest. As such, we focus on studying the asymptotic behavior of the marginal posterior on the partition tree and latent states, establishing consistency results on their convergence toward the trees and states that most closely characterize the underlying truth.

The rest of the article is organized as follows. In Section 2 we describe a flexible prior on the partition tree structure that relaxes the restriction of "dividing in the middle" on partition points and present a general form of PT models that adopt this prior along with latent states associated with the tree nodes with a Markov dependency structure. In Section 3, we present our hybrid computational strategy that can work effectively up to 100 dimensions consisting of an SMC algorithm for sampling on the marginal posterior of the partition tree and a recursive message passing algorithm for obtaining the exact conditional posterior of the latent states and the random measure given the sampled trees. In Section A we investigate the asymptotic properties of the tree structures and latent states identified under the posterior model. In Section 5, we carry out extensive numerical experiments to examine the performance of our method in the

context of two important applications of PTs—density estimation and two-group comparison, followed by an application to a dataset from mass cytometry in Section 6. In Section 7 we conclude with a brief discussion. All proofs are provided in Supplementary Materials E.

2. Method

We first review the PT model (Ferguson 1973; Lavine 1992) on a dyadic recursive partition in Section 2.1. The model, while defined on a general multivariate sample space, differs from a traditional multivariate PT which adopts a multi-way symmetric recursive partitioning. Then we introduce a new class of PT models that incorporates both the flexible partition prior and latent states with Markov dependency.

2.1. Pólya Trees Defined on Recursive Dyadic Partitions

Without loss of generality, we consider a continuous sample space represented as a d-dimensional rectangle $\Omega = (0,1]^d$. For unbounded sample spaces such as $\mathbb{R}^d$, one can transform each margin to (0,1] by applying, say, a cumulative distribution function transform or by standardizing the data based on its observed range of values. We use μ to denote the Lebesgue measure on Ω. A (dyadic) recursive partitioning T on Ω is a sequence of partitions of Ω such that the partition blocks at each level of the partitioning are obtained by dividing each block in the previous level into two children blocks. Formally, we can write $T = \bigcup_{k=0}^{\infty} \mathcal{A}^k$, where $\mathcal{A}^k$ is a partition of Ω in the kth level. More specifically, $\mathcal{A}^0 = \{\Omega\}$, and $A \in \mathcal{A}^k$ $(k = 0, 1, 2, \ldots)$ is divided into A_l and A_r, which satisfy $A_l, A_r \in \mathcal{A}^{k+1}$, $A_l \cup A_r = A$, and $A_l \cap A_r = \emptyset$. (Throughout the article, a subscript "l" or "r" on a node indicates the left or right child node.) For example, when $d = 1$ and if the tree is recursively divided at the middle point of each node, then nodes in level k are of the form $(l/2^k, (l+1)/2^k]$ for some $l \in \{0, \ldots, 2^k - 1\}$. Another common strategy is to define the tree based on the quantiles of a probability measure F so that $A \in \mathcal{A}^k$ is of the form $A = (F^{-1}(\frac{l}{2^k}), F^{-1}(\frac{l+1}{2^k})]$ for $l \in 0, \ldots, 2^k - 1$.

Given a partition tree T, we can define a random measure Q by putting a prior on the conditional probability $\theta(A) = Q(A_l|A) = 1 - Q(A_r|A)$ at each $A \in T$. Under the PT prior, the parameters $\theta(A)$ follow independent beta distributions $\text{Beta}(\alpha_l(A), \alpha_r(A))$, where $\alpha_l(A)$ and $\alpha_r(A)$ are hyperparameters. The corresponding posterior, given an iid sample $x_1, \ldots, x_n$ from Q, is again a PT with a simple conjugate update on the conditional proabilities:

$$\theta(A) \mid x_1, \ldots, x_n \sim \text{Beta}(\alpha_l(A) + n(A_l), \alpha_r(A) + n(A_r)),$$

where $n(A)$ represents the number of observations in a set $A \subset \Omega$. Though the tree needs to be infinitely deep to ensure full support of the PT, for practical purposes, one typically sets a sufficiently large maximum depth (or resolution) of T and compute the posteriors of $\theta(A)$'s defined on this finite tree structure (Hanson and Johnson 2002). We shall refer to a node in the deepest level as a "leaf" or "terminal node." On a leaf, the conditional distribution can be set to a baseline $F(\cdot|A)$, such as the uniform distribution $\mu(\cdot|A)$. In Section 3 when we present inference algorithms, we shall adopt this practical strategy and assume T is finite and use $\mathcal{N}(T)$ and $\mathcal{L}(T)$ to denote the collection of the nonterminal nodes and the leaf nodes, respectively.

2.2. Incorporating Flexible Partition Points

We incorporate a Bayesian-CART like prior on T by randomizing both the dimension in which to divide a node and the location to divide. Our prior relaxes the "always-divide-in-the-middle" restriction imposed in Wong and Ma (2010). This prior on the partition tree T differs from that in the mixture of PTs of Hanson (2006), which does not randomize over the dimension to divide, but generates the boundaries of the tree nodes jointly using quantiles of a parametric family.

Our prior can be described iteratively as a generative process that recursively divides the sample space. Specifically, suppose we have a node A in the rectangular form, $A = (a_1, b_1] \times \cdots \times (a_d, b_d]$. We divide A into two rectangular children by cutting along a randomly chosen dimension at a random location. The dimension to divide $D(A) \in \{1, 2, \ldots, d\}$, and the (relative) location to divide $L(A) \in (0, 1)$ are given independent priors of the following forms:

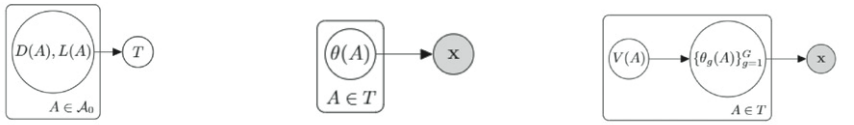
$$\begin{aligned} D(A) &\sim \text{Mult}(\lambda_1(A), \ldots, \lambda_d(A)) \quad \text{and} \\ L(A) &\sim \sum_{l=1}^{N_L - 1} \beta_l(A) \delta_{l/N_L}(\cdot), \end{aligned} \tag{1}$$

where $\delta_x(\cdot)$ represents the unit point mass at x, and $N_L - 1$ is the total number of grid points along $(0, 1)$. Both $\{\lambda_i(A)\}_{i=1,\ldots,d}$ and $\{\beta_l(A)\}_{l=1,\ldots,N_L-1}$ sum to 1. In the above, we have adopted a uniform grid over $(0, 1)$ for notational simplicity, but it does not have to be as such. With $D(A) = j$ and $L(A) = l/N_L$, the two children nodes A_l and A_r are

$$\begin{aligned} A_l &= (a_1, b_1] \times \cdots \times (a_j, a_j + l/N_L \cdot (b_j - a_j)] \\ &\quad \times \cdots \times (a_d, b_d], \\ A_r &= (a_1, b_1] \times \cdots \times (a_j + l/N_L \cdot (b_j - a_j), b_j] \\ &\quad \times \cdots \times (a_d, b_d]. \end{aligned}$$

In principle one could adopt a continuous prior on the partition location $L(A)$. A discretized prior is helpful, however, because it will substantially simplify posterior computation. In practice, as long as the grid is dense enough, the discrete prior will be practically just as flexible. Indeed we have verified in extensive numerical experiments that when N_L is large enough (more than 30–50) over a uniform grid, posterior inference no longer improves in any noticeable way.

For the prior on $D(A)$, we set $\lambda_j(A) = 1/d$ for all nodes A as a default choice. When $L(A)$ is given a weak prior widely spread over $(0, 1)$, the resulting inference can be sensitive to the "tail" behavior of the distribution in the node, resulting in high posteriors of $L(A)$ near the extreme values 0 and 1. A detailed discussion on this phenomenon will be provided in Section 5.1.1. This issue can be effectively addressed by making the prior of $L(A)$ depend on the sample size $n(A)$ so that it encourages more balanced divisions at large sample sizes. More



(a) Partition tree generation ($\mathcal{A}_0$: All potential nodes) (b) PT without latent states (c) PT with latent states

Figure 1. A graphical representation of the PT models with hyperparameters hidden.

specifically, we adopt the following prior with an exponentially decaying tail

$$P(L(A) = l/N_L) = \beta_l \propto \exp\left[-\eta n(A) f(|l/N_L - 0.5|)\right], \quad l = 1, \ldots, N_L - 1, \tag{2}$$

where $\eta \geq 0$ is a hyperparameter and $f : [0, \infty) \to [0, \infty)$ is an increasing function with $f(0) = 0$. In the following, we shall use a function $f(x) = x$, and so our prior on $L(A)$ is a (discretized) Laplace distribution. We provide theoretical justification for adopting this prior with exponential tails in Supplementary Materials A.

Another generalization of the prior on $L(A)$ is to incorporate a spike-and-slab set-up with a spike at the middle point 1/2. In particular, one can adopt a dependent spike prior among the nodes such that once a node A is divided exactly at the middle point, so are its descendants. This generalization will substantially reduce the amount of computation in regions of the sample space where the data are either sparse or lack interesting structure, for example, close to the uniform distribution. We implement the spike-and-slab in our software but defer the details of this generalization to Supplementary Materials A to avoid distracting the reader from the main ideas.

Given the tree prior, our PT model now consists of two components—tree generation and conditional probability assignment. Figures 1(a) and (b) present a graphical model representation for each.

2.3. Hidden Markov Pólya Tree Models

2.3.1. General Framework

Next we extend the above model to accommodate two recent developments in the PT literature: (i) incorporating latent state variables along the tree structure and (ii) joint modeling of multiple groups of observations. Incorporating latent variables allows more flexibly characterizing distributional features through adding prior dependency. As in recent literature, we consider incorporating discrete state variables that follow a Markov process along the tree structure. Because the description in this section always pertains to the model *given* the randomly generated partition tree T, for brevity we shall not keep stating "given T."

We generalize our notation to allow observing one or more groups of iid observations. Let G be the number of groups of iid observations. For the gth group ($g = 1, 2, \ldots, G$), let Q_g be the sampling measure for that group. Let $\mathbf{Q}$ denote the collection of all G sampling measures. That is, $\mathbf{Q} = \{Q_g\}_{g=1}^G$. Let $\boldsymbol{x}_g = (x_{g,1}, \ldots, x_{g,n_g})$ denote the observations in the gth group, which are iid given Q_g, where n_g, the sample size for the group, is allowed to differ across the groups. We use $\mathbf{x} = \{\boldsymbol{x}_g\}_{g=1}^G$ to denote the collection of all observations from all groups.

Next we specify a prior on $\mathbf{Q}$ in terms of a joint prior on the conditional probabilities on each $A \in T$, $\theta_g(A) = Q_g(A_l \mid A) = 1 - Q_g(A_r \mid A)$. We use latent variable modeling to incorporate prior dependency among the tree nodes. Specifically, let $\{V(A) : A \in T\}$ denote a collection of latent state variables, one for each A, and without loss of generality, assume that $V(A)$ takes discrete values from $\{1, \ldots, I\}$. (In practice, the number of states I can differ among A.) Joint priors of $\theta_g(A)$ for all g and A are then defined conditionally on these latent states.

Existing literature has exploited these latent states to characterize both the within-group structure of each distribution Q_g and the between-group relationship among the Q_g. An example of within-sample structures is the smoothness of each underlying distribution, which is explored in the context of density estimation (Ma 2017). An example of between-group structures is the difference between two (or more) distributions (Soriano and Ma 2017).

Dependent modeling of the latent states over the partition tree is desirable as a priori one would expect interesting structures (both within-group and between-group) to exhibit themselves in a correlated manner over the sample space. For example, functions tend to have similar smoothness over adjacent locations, and two-group difference tend to be clustered in space. A computationally efficient strategy for modeling such dependency over the tree is by a hidden Markov process along the tree (Crouse and Baraniuk 1997), which starts from the root node, $A = \Omega$, and sequentially generates the latent states in a coarse-to-fine fashion according to (possibly node-specific) transition matrices $\boldsymbol{\xi}(A)$ whose (i, i')th element is

$$\boldsymbol{\xi}_{i,i'}(A) = P(V(A) = i' \mid V(A^p) = i) \quad \text{for } i, i' \in \{1, \ldots, I\},$$

where A^p denotes A's parent. (We shall use superscript "p" to indicate the parent of a node in T.) For $A = \Omega$, since Ω has no parent, we can simply let $\boldsymbol{\xi}_{i,i'}(\Omega)$ be constant over i, representing the initial state probabilities on Ω.

Given the $V(A)$'s, $\{\theta_g(A)\}_{g=1,\ldots,G}$ can then be modeled as conditionally independent a priori. Figure 1(c) presents a graphical model representation for the latent state modeling on G probability distributions by PTs given T, which along with our generalized prior on the partition tree T presented in Figure 1(a) forms the most general version of the model we consider in this work. The specific choices of these conditional priors are problem-dependent. We give two examples below.

Example 1: Density Estimation with Adaptive Smoothness

An example of within-group structures that the latent state $V(A)$ can characterize is the smoothness of the density functions

for the random measures. For example, Ma (2017) proposed the adaptive Pólya tree (APT) model which incorporates latent states to allow different levels of local smoothness in the underlying distribution. This is achieved by modeling the $\theta_g(A)$'s as Beta$(m(A)\nu(A), (1-m(A))\nu(A))$, where $m(A)$ is the prior mean and $\nu(A)$ the precision parameter which characterizes the smoothness of the random measure with larger $\nu(A)$ corresponding to more smoothness. Then we model the precision parameters conditional on the latent state $V(A)$, which follows a Markov process along a tree. The detail of the APT model is provided in Supplementary Materials C.

Example 2: Two-Group Comparison

In two-group comparison, we are interested in testing and identifying differences between two measures $\mathbf{Q} = \{Q_g\}_{g=1,2}$ based on an iid sample from each. The "global" testing problem can be formulated as testing the following null and alternative hypotheses: $H_0 : Q_1 = Q_2$ versus $H_1 : Q_1 \neq Q_2$. Noting that two-group differences may exist in parts of the sample space and not others, the coupling OPT (Ma and Wong 2011) and the multi-resolution scanning (MRS) model (Soriano and Ma 2017) are PT-based models that allow the measures to differ on some nodes $A \in T$ and not others. This more "local" perspective on two-group comparison enables these models to not only test for H_0 versus H_1, but to identify regions on which the two measures differ. To achieve this, these models incorporate state variables that characterize whether the conditional probabilities on each A are equal:

$$\begin{aligned} V(A) = 1 &\Leftrightarrow Q_1(A_l \mid A) \neq Q_2(A_l \mid A), \\ V(A) = 2 &\Leftrightarrow Q_1(A_l \mid A) = Q_2(A_l \mid A). \end{aligned} \tag{3}$$

When $V(A) = 1$, the two corresponding conditional probabilities are given independent beta priors, whereas if $V(A) = 2$, they are tied and given a single beta prior. Markov dependency among the states on different nodes are incorporated to induce the desired spatial correlation of cross-group differences. The MRS model also incorporates "an absorbing state" $V(A) = 3$ with which we can ignore uninteresting regions, as detailed in Supplementary Materials D.

3. Bayesian Inference

In sum, the models we consider all share a common structure consisting of the following components: (i) the partition tree T defined by the dimension and location variables D's and L's, which follow the priors given in Equation (1); (ii) the latent state variables $V(A)$ given T which follow a Markov prior; (iii) the conditional probabilities along the given tree T, $\{\theta_g(A)\}_{g=1}^G$, whose joint prior are specified independently across the nodes on T given the latent states; and finally (iv) given the random measures Q_g defined by T and $\theta_g(A)$'s, we observe an iid sample $\boldsymbol{x}_g$ from each Q_g, independently across g.

We shall refer to this general model class as the Hidden Markov Pólya tree, or HMPT, and summarize it below:

$$\begin{aligned} T \mid \boldsymbol{\lambda}, \boldsymbol{\beta}, \eta &\sim p(T \mid \boldsymbol{\lambda}, \boldsymbol{\beta}, \eta) \\ \{V(A) : A \in T\} \mid \boldsymbol{\xi}, T &\sim \text{Markov}(\boldsymbol{\xi}) \\ (\theta_1(A), \ldots, \theta_G(A)) \mid V(A), T &\overset{\text{ind}}{\sim} p(\theta_1(A), \ldots, \theta_G(A) \mid V(A)) \\ &\qquad \text{for } A \in T \\ \boldsymbol{x}_g = (x_{g,1}, x_{g,2}, \ldots, x_{g,n_g}) \mid Q_g &\overset{\text{iid}}{\sim} Q_g \text{ for } g = 1, 2, \ldots, G. \end{aligned}$$

The key to Bayesian inference is the ability to either compute or sample from the joint posterior $(T, \mathbf{V}, \boldsymbol{\theta})$ given all data $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_G)$, where $\mathbf{V}$ and $\boldsymbol{\theta}$ represent the totality of all latent states and conditional probabilities given T, respectively. While in some problems such as density estimation one may mainly be interested in just the marginal posterior of the Q_g's, in others such as two-group comparison where one wants to characterize the between-group relationships among the distributions, the latent states (along with T), which characterizes such relationships, are often of prime interest. In multivariate problems, the tree structure T is also of great interest as it sheds light on the underlying structures in the distributions.

To this end, we shall take advantage of recent developments in sequential Monte Carlo (SMC) sampling for tree-based models (Lakshminarayanan, Roy, and Teh 2013; Lu, Jiang, and Wong 2013) and advances in message passing algorithms for PT models with Markov dependency (Ma 2017). We introduce a hybrid algorithm that combines these two computational strategies to effectively sample from the joint posterior in high-dimensional spaces. Overall, the hybrid algorithm consists of two stages:

1. *Sampling from the marginal posterior of the partition tree*
   We design an SMC sampler—that is, a particle filter—to sample a collection of tree structures $T^1, \ldots, T^M$ by growing each tree from coarse to fine scales. It uses one-step look-ahead message passing to construct proposal distributions for $D(A)$ and $L(A)$, one node at a time.
2. *Computing the conditional posterior given the sampled trees*
   Given each tree sampled by the SMC, the conditional model essentially becomes a hidden Markov process, for which we can analytically compute the exact conditional posteriors of $V(A)$'s and $\theta(A)$'s using recursive message passing.

3.1. SMC to Sample from Tree Posterior

In the SMC stage to sample the trees, each particle stores a realized form of a finite tree structure, and one node of each tree is divided at each step of the SMC sampling. Suppose T_t is the finite tree obtained after dividing the sample space t times in a particle, and for this tree we define the target distribution

$$\pi_t(T_t) = P(T_t \mid \mathbf{x}) \propto P(T_t) P(\mathbf{x} \mid T_t).$$

Here $P(T_t)$ is the joint prior of the variables $D(A)$'s and $L(A)$'s for the non-leaf nodes of T_t, and $P(\mathbf{x} \mid T_t)$ is the marginal likelihood given the tree T_t, in which $\mathbf{V}$ and $\boldsymbol{\theta}$ are integrated out. To sample from this target distribution, we sequentially construct a set of M particles $\{T_t^m, W_t^m\}_{m=1}^M$, where T_t^m is a realized tree and W_t^m is the associated importance weight for the mth particle. Examples of generated trees are given in Figure 2, where the sample space has been divided three times, and in the next step, new partition boundaries will be added in the gray nodes.

Following Lakshminarayanan, Roy, and Teh (2013), we adopt in each step of the SMC a breadth-first tree-growth strategy by

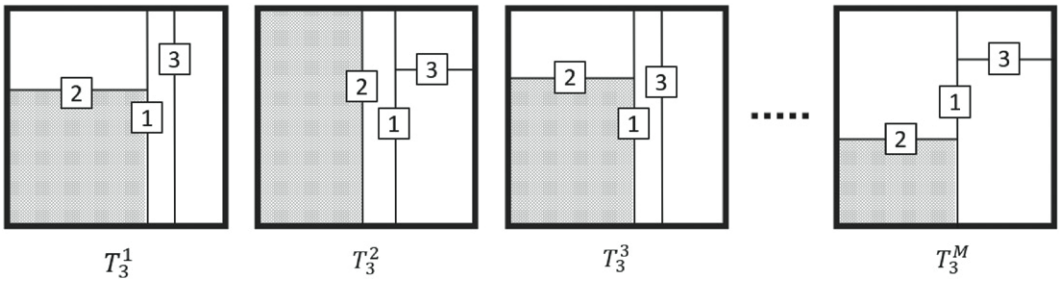


Figure 2. An example of realized finite trees in the particle system obtained after the step $t = 3$. The numbers in the squares indicate in which step the boundaries are drawn. Among the current leaf nodes, the nodes colored gray are the oldest nodes generated in the earliest step, so they are split in the next step.

dividing the "oldest active" leaf node—that is, the one generated in the earliest step and is yet to be terminated in division. Further division of a node is terminated once the number of observations in that node is below a preset threshold (e.g., 5 in our software implementation) to avoid excessive partitioning. (In Supplementary Materials G, we show through an additional experiment where data points accumulate around the boundaries of the sample space to confirm that this sample size thresholding indeed helps avoid excessive partitioning.) Otherwise a node is bisected along a boundary whose dimension and location are randomly drawn from a proposal distribution. For each particle, a finite tree T_t is formed by a sequence of decisions $\{J_s\}_{s=1}^t$, where $J_s = (D_s, L_s)$ correspond to all of the variables $D(A)$ and $L(A)$ at the sth step of the SMC.

As we will see in Proposition 3.1, the target distribution $\pi_t(T_t) = C_t\pi_{t-1}(T_{t-1})\pi_t(J_t \mid T_{t-1})w_t(T_{t-1})$, where C_t is a constant independent of T_t, $\pi_t(J_t \mid T_{t-1})$ a conditional distribution on J_t given T_{t-1}, and $w_t(T_{t-1})$ a function of T_{t-1}. We will choose $\pi_t(J_t \mid T_{t-1})$ as the proposal for J_t under which the corresponding importance weight will simply be $w_t(T_{t-1})$, independent of J_t.

More specifically, suppose at the current step t, we are to divide $A_t \in T_{t-1}$ into $A_{t,l}$ and $A_{t,r}$. Let $M_i(A_t \mid J_t)$ be the marginal likelihood on the node A_t under the decision J_t evaluated based on the observations in A_t. That is,

$$M_i(A_t \mid J_t) = \int \left[\prod_{g=1}^{G} \theta_g(A_t)^{n_g(A_{t,l})} (1-\theta_g(A_t))^{n_g(A_{t,r})} \right] dP(\theta_1(A_t), \ldots, \theta_G(A_t) | V(A_t) = i), \tag{4}$$

where $n_g(A)$ is the number of observations of the gth group included in A. To avoid cumbersome notation, we suppress in our notation the dependency of $M_i(A_t \mid J_t)$ on the observations $\mathbf{x}$. For example, if the $\{\theta_g(A_t)\}_{g=1}^G$ follow independent beta priors written as $\text{Beta}(\alpha_l^i(A_t), \alpha_r^i(A_t))$ given $V(A_t) = i$, then the marginal likelihood has the following expression $M_i(A_t \mid J_t) = \prod_{g=1}^G \frac{B(\alpha_l^i(A)+n_g(A_{t,l}), \alpha_r^i(A)+n_g(A_{t,r}))}{B(\alpha_l^i(A_t), \alpha_r^i(A_t))}$, where $B(\cdot,\cdot)$ is the beta function. Based on the values of $M_i(A_t \mid J_t)$, we can analytically compute the proposal and the importance weight using a general recursive algorithm, as described in the following proposition.

Proposition 3.1. For every possible decision J_t and states $i = 1, \ldots, I$, let $\varphi_i(A_t)$ be a function defined recursively:

$$\boldsymbol{\varphi}_i(A_t) = \begin{cases} \frac{\boldsymbol{\xi}_{1,i}(\Omega) M_i(\Omega | J_t)}{\sum_{j=1}^I \boldsymbol{\xi}_{1,j}(\Omega) M_j(\Omega | J_t)} & \text{if } A_t = \Omega \\ \frac{\sum_{j=1}^I \varphi_j(A_t^p) \boldsymbol{\xi}_{j,i}(A_t) M_i(A_t | J_t)}{\sum_{k=1}^I \sum_{j=1}^I \varphi_j(A_t^p) \boldsymbol{\xi}_{j,k}(A_t) M_i(A_t | J_t)} & \text{otherwise,} \end{cases} \tag{5}$$

where A_t^p is A_t's parent node. Also, let $h(J_t \mid A_t)$ be a function of J_t defined as

$$h(J_t \mid A_t) = \sum_{i=1}^I \left\{ \sum_{j=1}^I \boldsymbol{\varphi}_j(A_t^p) \boldsymbol{\xi}_{j,i}(A_t) \right\} M_i(A_t \mid J_t) \frac{\mu(A_{t,l})^{-n(A_{t,l})} \mu(A_{t,r})^{-n(A_{t,r})}}{\mu(A_t)^{-n(A_t)}}, \tag{6}$$

where $n(A)$ denotes the total number of observations included in a node A. Then the target distribution $\pi_t(T_t)$ can be expressed in terms of $\pi_{t-1}(T_{t-1})$ as

$$\pi_t(T_t) = C_t\pi_{t-1}(T_{t-1})\pi_t(J_t \mid T_{t-1})w_t(T_{t-1}),$$

where C_t is a constant and

$$\pi_t(J_t \mid T_{t-1}) = \frac{P(J_t)h(J_t \mid A_t)}{\sum_{j_t} P(j_t)h(j_t \mid A_t)},$$
$$w_t(T_{t-1}) = \sum_{j_t} P(j_t)h(j_t \mid A_t).$$

The summation over j_t is taken over all possible decisions.

Corollary 3.1. Let $h(J_t \mid A_t)$ be the function defined in Proposition 3.1. Then the proposal distribution $\pi_t(D_t \mid T_{t-1})$ is given by

$$\pi_t(D_t \mid T_{t-1}) = \pi_t(D_t \mid T_{t-1})\pi_t(L_t \mid D_t, T_{t-1}), \text{ where}$$

1. $\pi_t(D_t \mid T_{t-1})$ is $\text{Mult}(\widetilde{\lambda}_1(A_t), \ldots, \widetilde{\lambda}_d(A_t))$ with

$$\widetilde{\lambda}_j(A_t) \propto \sum_{l=1}^{N_L-1} \pi_t((j, l/N_L) \mid T_{t-1}) \propto \lambda_j(A_t) \sum_{l=1}^{N_L-1} \beta_l(A_t) h((j, l/N_L) \mid A_t).$$

2. Given $D(A_t) = j$, the conditional proposal of $L(A_t)$ is

$$\pi_t(L_t = l/N_L \mid D_t = j, T_{t-1}) = \sum_{l=1}^{N_L-1} \widetilde{\beta}_l(A_t)\delta_{l/N_L}(\cdot),$$

for $j = 1, 2, \ldots, I$ and $l = 1, \ldots, N_L - 1$ with

$$\widetilde{\beta}_l(A_t) \propto \beta(A_t)h(j, l/N_L \mid T_{t-1}).$$

We also have an analytical expression of the incremental weight:

$$w_t(T_{t-1}) = \sum_{j=1}^{d} \sum_{l=1}^{N_L-1} \lambda_j(A_t)\beta_l(A_t)h((j, l/N_L) \mid A_t).$$

Remark. The recursive function $\varphi_i(A_t)$ can be computed based on $\varphi_i(A_t^p)$ with the fixed computational cost. Hence, the optimal proposal $\pi_t(J_t \mid T_{t-1})$ and the incremental weight $w_t(T_{t-1})$, which are functions of $h(J_t \mid A_t)$, can be obtained at each step with constant computational cost with complexity $O(I^2 N_L d\, n(A_t))$. As such, our inference algorithm scales linearly in both the dimensionality and the sample size.

The pseudo-code of the new SMC algorithm that summarizes the discussion is provided in Supplementary Materials A. In the algorithm, we stop dividing A_t if either the depth of A_t is equal to a preset maximum resolution K (e.g., 15) or the number of observations in A_t is less than a preset threshold (e.g., 5). The SMC algorithm terminates when all the nodes of all the particles have been stopped. The maximum resolution K controls the level of local details that the HMPT model allows to infer, and larger values of K require more computational time. In a range of applications we have found that setting K beyond 15–20 leads to minimal changes in the resulting inference.

3.2. Posterior Computation Given Sampled Tree Structures

The second stage of our inference strategy is to compute the posterior distributions of the latent states $V(A)$ and the conditional probabilities $\theta_g(A)$ given each sampled tree. In this stage we first compute the marginal posterior of the latent states given the tree with a recursive message-massing algorithm as shown in Ma (2017) and Soriano and Ma (2017). The algorithm written in our notation is provided in Supplementary Materials A, and we note that this algorithm works for all models under consideration. We note that in this recursive algorithm we can compute the overall marginal likelihood given the tree T, $P(\mathbf{x} \mid T)$, which can be used to find the *maximum a posteriori* (MAP) tree among the sampled trees, that is, the sampled tree T^m that maximizes $P(T^m \mid \mathbf{x}) \propto P(T^m)P(\mathbf{x} \mid T^m)$. We can use this "representative" tree, along with the conditional posterior of the latent states given this tree, to visualize and summarize the posterior inference in an interpretable way.

Given both the tree and the latent states, the posterior of $\theta_g(A)$ boils down to the corresponding posterior of standard PT models on a dyadic tree, which is problem-specific as provided in the literature on each such model. We specifically use the two examples from Section 2.3.1 to demonstrate how one may use the output of the algorithm—namely the sampled trees along with the conditional posterior given the trees—to carry out inference. We note that the inference strategies for these quintessential examples are generalizable to a variety of other tasks.

Example 1: Density Estimation

The problem of estimating an unknown density corresponds to $G = 1$ and so we can drop the subscript g to simplify the notation. We shall use the posterior mean density, also called the predictive density—$\mathbb{E}[q(\cdot) \mid \mathbf{x}]$—as an estimate for the density $q = dQ/d\mu$. As shown in Wong and Ma (2010) and Ma (2017), given a tree T^m, we can use the marginal posterior of the latent states to compute the conditional predictive measure $\mathbb{E}[Q(\cdot) \mid \mathbf{x}, T^m]$ with a top-down recursive algorithm. The algorithm is described in our generic notation in Supplementary Materials C. Hence, given an SMC sample of M trees and weights, it is possible to integrate out the random trees and compute the posterior predictive density as follows:

$$\mathbb{E}[q(x) \mid \mathbf{x}] \approx \sum_{m=1}^{M} W^m \frac{\mathbb{E}[Q(B^m(x)) \mid \mathbf{x}, T^m]}{\mu(B^m(x))},$$

where $B^m(x) \in \mathcal{L}(T^m)$ the leaf node to which x belongs, and W^m is the final importance weight for T^m.

Example 2: Two-Group Comparison

To compare two groups of observations using generalizations to the PT models described in Section 2.3.1, we shall compute the posterior probability of the two hypotheses H_0 and H_1. For example, when $V(A)$ is defined as in Equation (3), the posterior probability of the "global" null hypothesis $H_0 : Q_1 = Q_2$ is given by

$$\begin{aligned} P(H_0 \mid \boldsymbol{x}) &= \sum_{T \in \mathcal{T}} P(V(A) \neq 1 \text{ for all } A \in \mathcal{N}(T) \mid T, x) P(T \mid \boldsymbol{x}) \\ &\approx \sum_{m=1}^{M} W^m P(V(A) \neq 1 \text{ for all } A \in \mathcal{N}(T^m) \mid T^m, \mathbf{x}), \end{aligned}$$

where the sum over $\mathcal{T}$ in the first row is over all finite trees with maximum resolution K and the quantity $P(V(A) \neq 1 \text{ for all } A \in \mathcal{N}(T^m) \mid T^m, \mathbf{x})$ again is available analytically by message passing (details given in Supplementary Materials D).

We can also detect where and how the underlying distributions differ by computing the "posterior marginal alternative probability" (PMAP) on each node A, along any sampled tree T^m:

$$P(\theta_1(A) \neq \theta_2(A) \mid T^m, \mathbf{x}) = P(V(A) = 1 \mid T^m, \mathbf{x}).$$

Reporting the PMAPs along a representative tree such as the MAP among the sampled trees can be a particularly useful visualizing tool to help understand the nature of the underlying difference. One can also report on each A the estimated magnitude of the difference using a notion of "effect size" based on the log-odds ratio (Soriano and Ma 2017), $\text{eff}(A) = \left|\log\left[\frac{\theta_1(A)}{1-\theta_1(A)}\right] - \log\left[\frac{\theta_2(A)}{1-\theta_2(A)}\right]\right|$. In particular, one can report the posterior expected effect size $\mathbb{E}[\text{eff}(A) \mid \mathbf{x}, T]$, which can be computed using a standard Monte Carlo (not MCMC) sample from the exact posterior given the representative tree. We will demonstrate this using a mass cytometry dataset in Section 6.

4. Theoretical Properties

Next we investigate the theoretical properties of the HMPT model. Previous theoretical analysis on the PT had mostly focused on establishing the marginal posterior consistency and contraction of the random measures Q_g with respect to an

unknown fixed truth (Walker and Hjort 2001; Castillo 2017). We, however, shall take a different perspective and instead provide asymptotic theorems regarding the following questions that are often of practical interest:

1. What tree structures does the marginal posterior of T concentrate around?
2. How does the posterior of the latent states given the tree behave?

These two questions have broad relevance in inference using PT models, and previously several authors have investigated the second question in the two-group comparison context for their variants of the PT model (Holmes et al. 2015; Soriano and Ma 2017). In addressing the second question more generally, we aim to provide results that encompass these previous analyses as special cases.

We will address each of the two questions in turn. Throughout this section, we consider finite PTs with maximum depth of the trees set to some value K. We use $\mathcal{T}^K$ to denote this collection of trees. Also, the asymptotic results are derived under the prior for $L(A)$ provided in Section 2.2 which can depend on the (finite) sample size. The case of an uniform prior on $L(A)$ independent of the sample size is included as a special case where the hyperparameter $\eta = 0$. Finally, we consider models and sample sizes that satisfy Assumptions 1 and 2. The models discussed in Section 2.3.1 all meet this requirement.

Assumption 1. For each group $g \in \{1, \ldots, G\}$, let n_g be the sample size and P_g the true probability measure from which the observations are generated. We assume that

(i) There exists $\zeta_g \in (0, 1]$ such that $\zeta_g = \lim_{n\to\infty} \frac{n_g}{n}$ for $g \in \{1, \ldots, G\}$, where $n = n_1 + \cdots + n_G$ is the total number of observations across all groups.
(ii) The sampling distribution P_g satisfies $P_g \ll \mu$, and the density $p_g = dP_g/d\mu$ is positive almost everywhere.

Additionally, given the tree T and the latent states, the parameters $\{\theta_g(A)\}_{g=1}^G$ are given one of the following priors (the model can adopt a mix of these priors for different combinations of A and $V(A)$ values):

Prior A : $\theta_g(A)$ independently follow a beta prior.
Prior B : $\theta_1(A) = \cdots = \theta_G(A)$ and follow a beta prior.
Prior C : $\theta_1(A) = \cdots = \theta_G(A) \equiv c(A)$, some constant in $(0, 1)$.

Establishing the theoretical properties also requires a condition on the latent states. In particular, under some states, the support of the prior of the parameters $\{\theta_g(A)\}_{g=1}^G$ needs to include the true conditional probabilities. To describe this requirement, given a tree $T \in \mathcal{T}^K$, let $S_i(A \mid T)$ be the support of the prior on $(\theta_1(A), \ldots, \theta_G(A))$ under the state $V(A) = i$. Then, let $\tau(A \mid T)$ denote the collection of "feasible states" on A. (A state is "feasible" if the true conditional probabilities are in the support of the corresponding prior given the state.) That is,

$$\tau(A \mid T) := \{i \in \{1, \ldots, I\} : (P_1(A_l \mid A), \ldots, P_G(A_l \mid A)) \in S_i(A \mid T)\}.$$

The next assumption states that the prior for the latent states must give positive probability for all the latent states to all simultaneously be feasible.

Assumption 2. For every $T \in \mathcal{T}^K$, $P(V(A) \in \tau(A \mid T)$ for all $A \in T) > 0$.

With these assumptions, we next derive asymptotic properties for the marginal posteriors for the tree and the state variables. In the following, we use the notation $\mathbf{x}_n$ instead of $\mathbf{x}$ for the data to indicate the total sample size.

In order to describe the posterior convergence of the partition trees, we introduce a notion for "tree-based approximation for probability measures." Let T be a finite tree and H a probability measure. Then the "tree-based approximation of H under T," denoted by $H|_T$, is defined as $H|_T(B) = \sum_{A\in\mathcal{L}(T)} H(A)\frac{\mu(B\cap A)}{\mu(A)}$. for any $B \in \mathcal{B}(\Omega)$. The following theorem then characterizes the trees the posterior concentrates on as the sample size grows.

Theorem 4.1. Let $\mathcal{T}_M^K$ be the collection of trees under which the tree-based approximation of the measures P_g minimizes the Kullback-Leibler divergence from the P_g's plus a penalty term on unbalanced splits. That is,

$$\mathcal{T}_M^K = \arg\min_{T\in\mathcal{T}^K} \sum_{g=1}^{G} \zeta_g \left\{\mathrm{KL}(P_g||P_g|_T) + \eta B_g(T)\right\}, \qquad (7)$$

where

$$B_g(T) = \sum_{A\in\mathcal{N}(T)} P_g(A) f\left(\left|\frac{\mu(A_l)}{\mu(A)} - 0.5\right|\right).$$

Then the marginal posterior of T concentrates on $\mathcal{T}_M^K$. That is, as $n \to \infty$,

$$P(T \in \mathcal{T}_M^K \mid \mathbf{x}_n) \xrightarrow{p} 1.$$

For the state variables, it is desirable that their posterior distribution concentrates on a collection of feasible states. Moreover, when multiple configurations of the states are feasible, it is desirable that the posterior concentrates around such configurations that provide the most parsimonious representation of the true distributions. For example, if the true conditional distribution on a node is uniform, a model that introduces a possible nonuniform structure on this node is feasible but redundant. White and Ghosal (2011) and Li and Ghosal (2014) showed that, in quite general settings of multi-resolution inference, the posterior probability of such redundant models tends to concentrate its mass on 0. By adapting their techniques, we show that the same property holds in the case of the HMPT model.

To formally describe the results, we need to define the complexity of the model specified by the latent states. Given the state $V(A) = i$, the complexity of the $\{\theta_g(A)\}_{g=1}^G$, in other words, the number of free parameters of the prior distribution under the ith state is denoted by $C_i(A)$. For example, for two-group comparison,

$$C_i(A) = \begin{cases} 1 & \text{if } \theta_1(A) = \theta_2(A) \\ 2 & \text{if } \theta_1(A) \neq \theta_2(A). \end{cases}$$

Next we introduce the complexity of a combination of states on the tree T. Given a tree T, let $\mathbf{V}$ denote a combination of the state variables $\{V(A)\}_{A\in\mathcal{N}(T)}$ and let $\mathbf{v} = \{\mathbf{v}(A)\}_{A\in\mathcal{N}(T)}$ ($\mathbf{v}(A) \in \{1,\ldots,I\}$) be one of the possible realizations of $\mathbf{V}$. Then we define the model complexity under $\mathbf{v}$ as follows:

$$C(\mathbf{v}) = \sum_{A\in\mathcal{N}(T)} C_{\mathbf{v}(A)}(A). \tag{8}$$

The next theorem shows that the posterior distribution of the states given the tree will concentrate on those that are feasible and most parsimonious.

Theorem 4.2. For $T \in \mathcal{T}^K$, let $\mathcal{V}_T = \{\mathbf{v} : \mathbf{v}(A) \in \tau(A \mid T)$ for all $A \in \mathcal{N}(T)\}$.

Then $P\left(\{\mathbf{V} \in \mathcal{V}_T\} \cap \left\{C(\mathbf{V}) = \min_{\mathbf{v}\in\mathcal{V}_T} C(\mathbf{v})\right\} \mid T, \mathbf{x}_n\right) \xrightarrow{p} 1$.

Remark. Consistency results for several existing models are special cases of this theorem. For example, we derive the consistency of the MRS model for two-group comparison as a corollary in Supplementary Materials F.

5. Experiments

In this section, we carry out simulation studies to examine the performance of the HMPT model and inference algorithm. We again consider the two quintessential examples—(i) density estimation and (ii) the two-group comparison—for inferring within-group and between-group structures, respectively. Details such as the settings of hyper-parameters and simulated datasets are provided in Supplementary Materials H unless explicitly described in this section.

5.1. Density Estimation

We first consider two-dimensional examples to observe what kind of tree structures are obtained under the HMPT model and how prior specification in Equation (2) influences the performance. After that, we move to higher dimensional cases to examine the scalability of our new SMC method and the effect of incorporating the flexible partition. For this task we compare the HMPT model with the original APT model (Ma 2017) which also incorporates a prior on the dimension to divide but restricts partitioning at middle points. Its posterior computation is implemented by the `apt` function in the R package `PTT`.

5.1.1. Two-Dimensional Cases

Simulated data are generated from the three scenarios with the densities visualized in the first row of Figure 3. (Details on the simulation settings are provided in Supplementary Materials H.1.2.) Also presented in Figure 3 are examples of the posterior mean densities $\mathbb{E}[q \mid \mathbf{x}]$ as well as the partition blocks under the MAP tree. Note that the posterior mean is computed by integrating out the unknown tree, and the MAP tree is presented to visualize key distributional features. The results for the first scenario confirms that the HMPT model is much more effective in capturing the discontinuous boundaries of the true density. For the second scenario, our model tends to draw the boundaries that surround the true clusters. In the trees given the different values of η, however, we can see that fewer nodes were divided inside the clusters when $\eta = 0.01$. In contrast, when $\eta = 0.1$, the representative tree draws outlines of the clusters and divides regions inside of the clusters at the same time. A similar phenomenon is observed in the third scenario—under our model with flexible partitioning points, partition lines are formed around the region with high density, and when $\eta = 0.1$, the boundaries were also drawn within the high probability region. The quantitative comparison based on the KL divergence is provided in Figure 11 in Supplementary Materials I, which is consistent with the observations above.

5.1.2. Higher-Dimensional Cases

We generate d-dimensional iid observations from a density with independent pairs of margins, that is, $f(x_1, x_2, \ldots, x_d) = \prod_{j=1}^{d/2} f_j(x_{2j-1}, x_{2j})$ where

$$\begin{aligned} f_j(x_{2j-1}, x_{2j}) = p_j &\text{Beta}(x_{2j-1} \mid 0.25, 1) \times \text{Beta}(x_{2j} \mid 0.25, 1) \\ &+ (1 - p_j)\text{Beta}(x_{2j-1} \mid 50/j, 50/j) \\ &\times \text{Beta}(x_{2j} \mid 50/j, 50/j), \end{aligned}$$

with $p_j = 0.25 + 0.7/j$. We consider two different situations: (i) the dimension $d = 6$, and the sample size n changes from 5000 to 50,000; and (ii) the sample size $n = 10{,}000$ and the dimensionality changes from 10 to 100. For our method, the maximum depth K is set to 15. We show the comparison with the original APT, and also with the classical PT method, the kernel density estimation, and the Dirichlet process Gaussian mixture model (Escobar and West 1995; Müller, Erkanli, and West 1996).

Figure 4 presents the computational time for five different datasets. To obtain the result, we used a single-core environment using Intel Xeon Gold 6154 (3.00 GHz) CPU. The computational time is linear in both the sample size and the dimensionality.

The models are compared based on predictive scores, that is, the average of log-predictive densities, where as the predictive density the posterior mean of the density $\mathbb{E}[q \mid \mathbf{x}]$ is used. The size of the test and training sets is both n, and we repeat the computation for 50 pairs and take the average. The results, given in Figure 5, show that the HMPT model substantially outperforms the competitors by this criteria both when $d = 6$ with varying sample size and when n is fixed with varying dimensionality. It is worth noting that the poor performance of the APT in the $d = 6$ case is due to the fact that available software in the `apt` package, which does not uses SMC, cannot be fit for maximal resolution > 9. We also investigate the performance under sample sizes < 1000, and the results are similar (Figure 12 in Supplementary Materials I.)

5.2. Two-Group Comparison

Next we consider the two-group comparison problem, evaluate the performance of the HMPT model, and compare it to the original MRS with the "divide-in-the-middle" restriction. We use three scenarios ("Local location shift," "Local dispersion difference," and "Correlation") to generate 50-dimensional datasets. (The details are provided in Supplementary Materials H.2.2.) The first two scenarios involve two-group difference that lies in only parts of the sample space, or "local" differences,

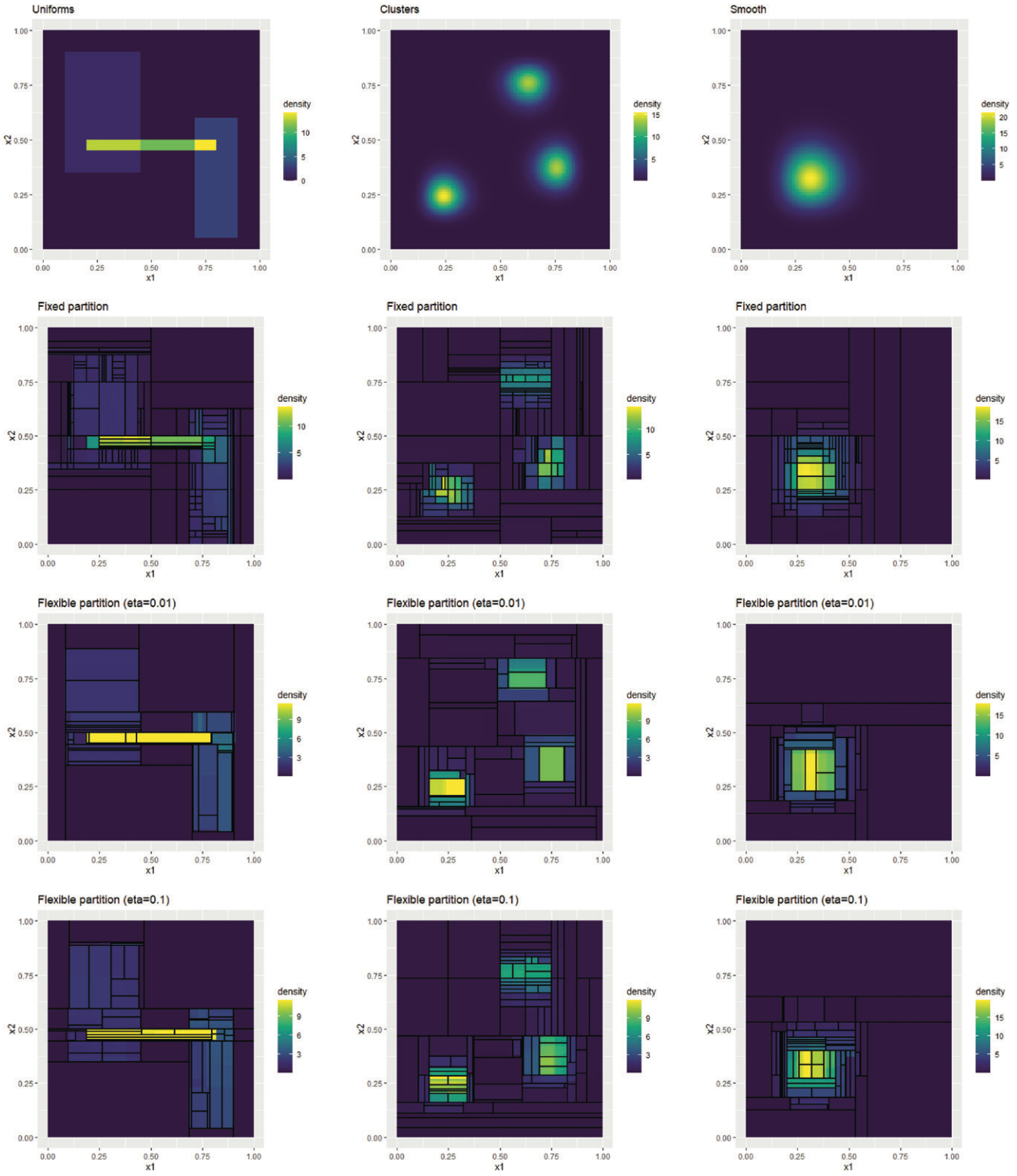


Figure 3. The posterior means of the densities and the representative trees obtained under $n = 1000$. Each column corresponds to a simulation scenario. The first row shows the true densities, the second row corresponds to the APT model (with fixed partition), and the third and fourth rows correspond to the HMPT model with flexible partitioning with parameters $\eta = 0.01$ and 0.1, respectively.

which will help demonstrate the usefulness of inferring the partition tree in identifying the nature of the differences. The sample size is $n_1 = n_2 = 2000$ in all scenarios.

The original algorithm for inference under the MRS model by message passing, which is implemented by the `mrs` function in the R package `MRS`, is not scalable beyond about 10 dimensions even with fixed partition locations. Hence, we compute the posterior for both the HMPT model and the original MRS in all scenarios with our SMC and message passing hybrid algorithm with the maximum resolution fixed to 15. We compare the performance using receiver operating characteristic (ROC) curves computed based on 200 simulated datasets under each scenario.

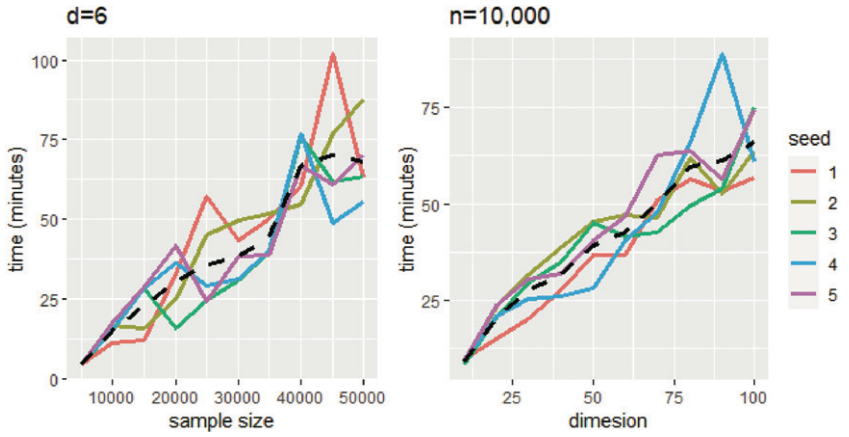


Figure 4. The wall time under five different datasets. The HMPT model with $\eta = 0.01$ is used. The black dashed lines indicate the average times.

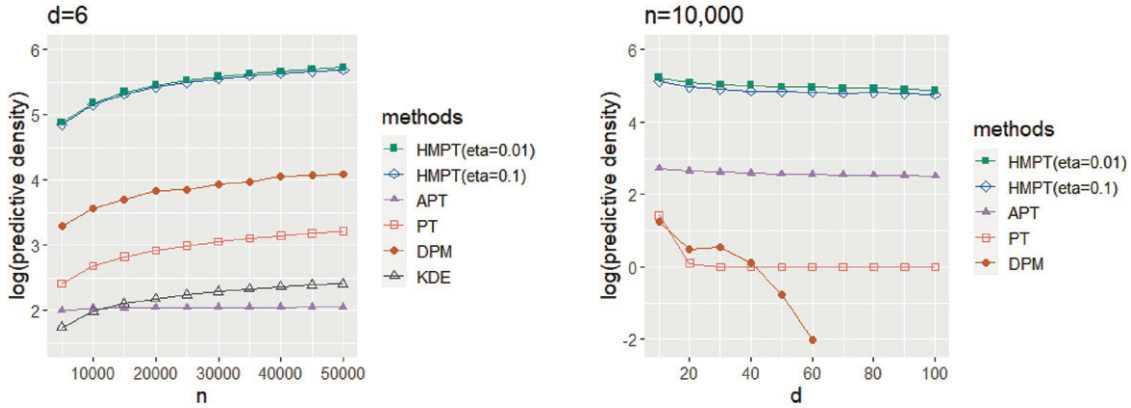


Figure 5. Predictive performance of six methods. Each point corresponds to the average of the predictive score based on 50 datasets. Each interval is formed by adding and subtracting the standard deviation. In the right plot, the predictive scores of the DPM model for the over 60-dimensional cases are below the displayed range.

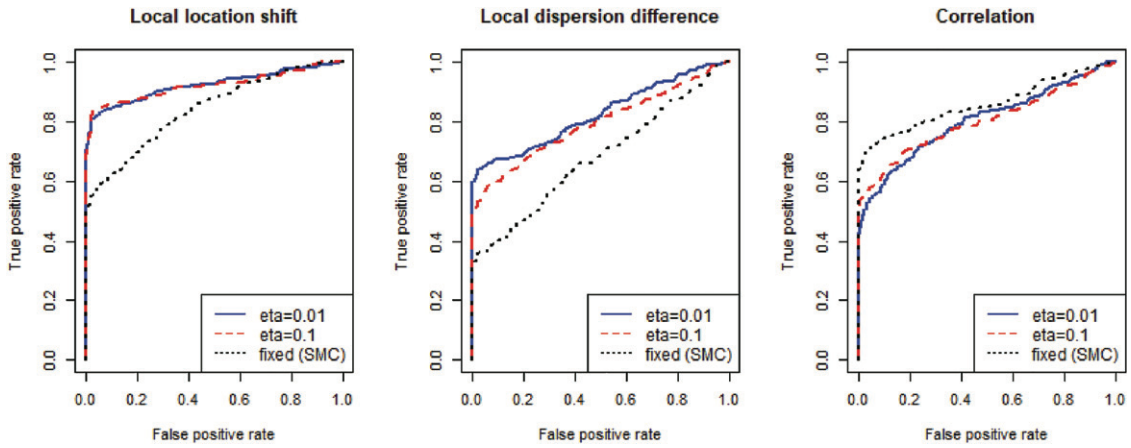


Figure 6. The ROC curves for the 50-dimensional examples.

Figure 6 presents the ROC curves. For the location shift and dispersion differences, the HMPT model with flexible partitioning results in substantially higher sensitivity. For the correlation scenario, the model with fixed partitioning locations performed slightly better. This is not surprising since in this scenario the difference exists smoothly over entire ranges of the dimensions without natural "optimal" division points, and so the performance gap is the cost for searching over more possible partition locations, none of which improves the model fit than the middle point. We again note that while the model with fixed partitioning performs well here, it works only with our new computational algorithm for data of such dimensionality.

To demonstrate how the posterior model can help understand the nature of the differences, we present under each scenario the node with the highest PMAP, or $P(V(A) = 1 \mid \boldsymbol{x}) = P(\theta_1(A) \neq \theta_2(A) \mid \boldsymbol{x})$, in Figure 7. In the location shift and dispersion difference scenarios the boundaries are away from the middle point to characterize the difference, which partly explains the sensitivity gain in adopting the flexible tree prior.

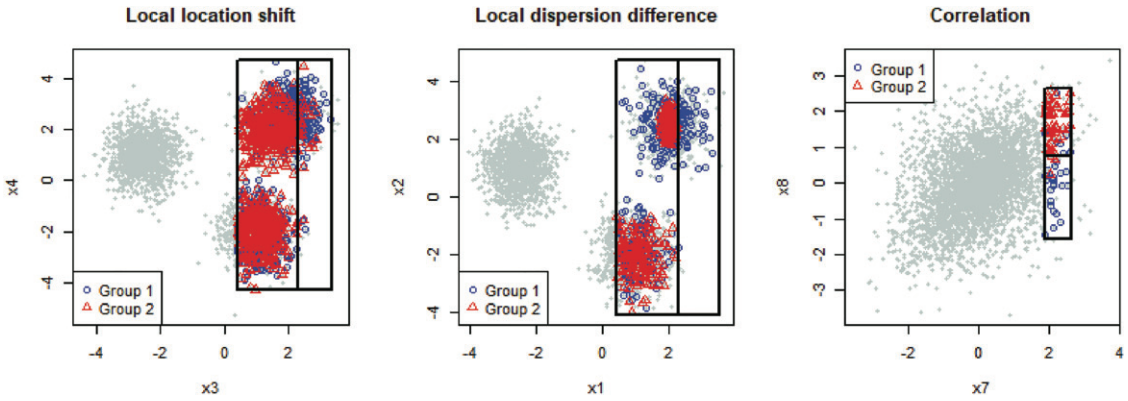


Figure 7. Examples of the node with the highest PMAP under the three scenarios for two-group comparison, under the MRS with flexible partitioning and $\eta = 0.1$. The solid lines mark the boundaries of the nodes and the partition line that divides them into the two children nodes. The red triangles and the blue circle are the observations from the two groups in the node. Gray points are the observations outside the node.

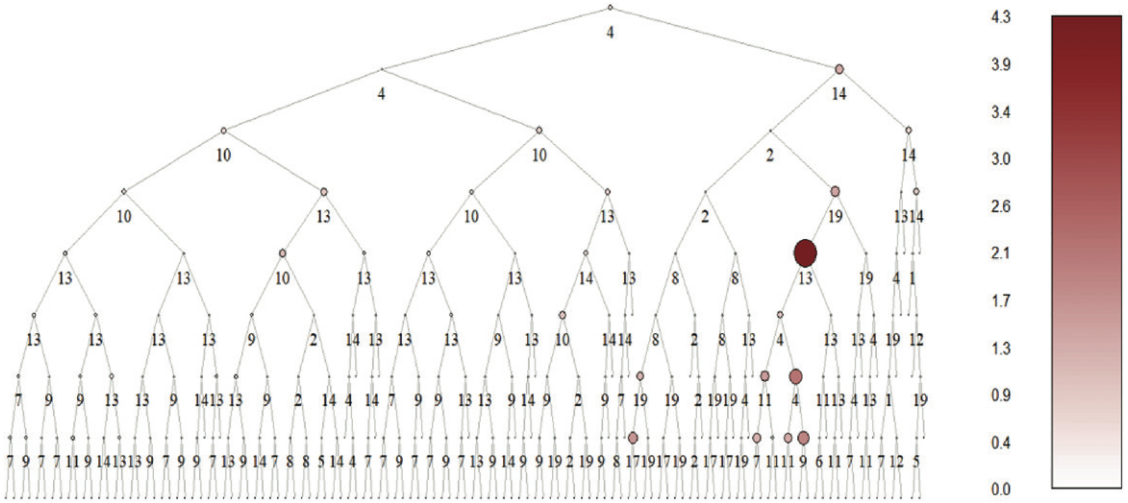


Figure 8. The MAP tree for the mass cytometry dataset visualized up to the ninth level. The size and the color of each node indicate the estimated eff(A), and the number above a node indicates dimension in which it is split. Only the nodes with more than 50 observations are shown.

6. Application to a Mass Cytometry Dataset

Finally, we apply our model for two-group comparison to a mass cytometry dataset collected by Kleinsteuber et al. (2016). The dataset records 19 different measurements including physical measurements and biomarkers on single cells in blood samples from a group of HIV patients as well as in reference samples from healthy donors. For demonstration, we compare the sample from an individual patient sample (Patient #1) to that from a healthy donor to identify differences in immune cell profiles. The sample sizes are 29,226 for the health donor and 228,498 for the patient, with each observation corresponding to a cell. We set $\eta = 0.1$ and the maximum depth K to 25.

Given the large sample sizes, the posterior probability for the global alternative $P(Q_1 \neq Q_2 \mid \mathbf{x})$ is almost 1 and so is of less interest. Our focus is instead on identifying the cell subsets on which the samples differ and on quantifying such differences. To this end, we report the effect size eff(A) defined in Section 3.2 on each node in a representative tree—the MAP among the sampled trees.

The estimated eff(A)'s on the MAP tree up to level 9 is visualized in Figure 8. The full tree and the nodes with large eff(A) are provided in Supplementary Materials I. We note that the nodes on which there is significant evidence for two-group differences, as well as those with large estimated effect sizes tend to be nested or clustered in subbranches of the tree, which is consistent with our intuition that there is spatial correlation in the two-group differences, and justifies the hidden Markov structure embedded in the MRS model.

7. Concluding Remarks

We have proposed a general framework for the PT model that incorporates a flexible prior on the partition tree and can accommodate latent state variables with Markov dependency along the partition tree. We have proposed a sampling algorithm that combines SMC and recursive message passing that can scale up to moderately high-dimensional ($\sim$ 100-dim) problems. Our numerical experiments confirm that our sampling algorithm scales linearly in the sample as well as the dimensionality size and the flexible partitioning tree prior can result in substantial gain in performance in some settings. Though we have mainly used two inference tasks—namely density estimation and two-group comparison—to demonstrate the HMPT model and algorithm, our approach can be readily applied to other PT models with a hidden Markov structure.

Our proposed algorithm is currently designed to be run in a single computer environment, so though the computational

cost is linear to the sample size n, direct application to problems involving huge n (e.g., $> 10^9$) is not yet feasible. It is of future interest to develop versions of the algorithm for distributed systems, which could explore either the parallel structure over nodes or parallel SMC algorithms.

Supplementary Materials

Supplementary Materials for this article are available online. Supplementary Materials A provides details on the posterior computation. Supplementary Materials B describes the spike-and-slab prior used for the location variable. The details on density estimation with APT model and two-sample comparison with the MRS model can be found in Supplementary Materials C and D, respectively. All proofs are included in Supplementary Materials E. Supplementary Materials F details the proof of consistency for the MRS model with the flexible partitioning. Supplementary Materials G presents the result of an additional experiment for evaluating the performance of our HMPT model in density estimation. The details of the numerical experiments are provided in Supplementary Materials H, and the additional figures can be found in Supplementary Materials I.

Acknowledgments

We thank the associate editor and the referees for very helpful comments. LM's research is partly supported by NSF grants DMS-2013930 and DMS-1749789. NA is supported by a fellowship from the Nakajima Foundation. The authors report there are no competing interests to declare.

Software and Data Availability Statement

An R package for our method is available at *https://github.com/MaStatLab/HMPT* and the code for the examples is at *https://github.com/MaStatLab/HMPT_experiments*.

ORCID

Li Ma http://orcid.org/0000-0002-0159-3296

References

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. [189]

Castillo, I. (2017), "Pólya Tree Posterior Distributions on Densities," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53, 2074–2102. [190,196]

Castillo, I., and Randrianarisoa, T. (2021), "Optional Pólya Trees: Posterior Rates and Uncertainty Quantification." arXiv preprint arXiv:2110.05265. [190]

Chen, Y., and Hanson, T. E. (2014), "Bayesian Nonparametric k-sample Tests for Censored and Uncensored Data," *Computational Statistics & Data Analysis*, 71, 335–346. [189]

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998), "Bayesian Cart Model Search," *Journal of the American Statistical Association*, 93, 935–948. [190]

Christensen, J., and Ma, L. (2020), "A Bayesian Hierarchical Model for Related Densities by Using Pólya Trees," *Journal of the Royal Statistical Society*, Series B, 82, 127–153. [189]

Crouse, M. S., and Baraniuk, R. G. (1997), "Contextual Hidden Markov Models for Wavelet-Domain Signal Processing," in *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems and Computers*, Vol. 1, pp. 95–100. [192]

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference using Mixtures," *Journal of the American Statistical Association*, 90, 577–588. [197]

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230. [189,191]

——— (1974), "Prior Distributions on Spaces of Probability Measures," *Annals of Statistics*, 2, 615–629. [189]

Filippi, S., and Holmes, C. C. (2017), "A Bayesian Nonparametric Approach to Testing for Dependence between Random Variables," *Bayesian Analysis*, 12, 919–938. [189]

Freedman, D. A. (1963), "On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case," *Annals of Mathematical Statistics*, 34, 1386–1403. [189]

Hanson, T., and Johnson, W. O. (2002), "Modeling Regression Error with a Mixture of Polya Trees," *Journal of the American Statistical Association*, 97, 1020–1033. [189,191]

Hanson, T. E. (2006), "Inference for Mixtures of Finite Polya Tree Models," *Journal of the American Statistical Association*, 101, 1548–1565. [189,191]

Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015), "Two-sample Bayesian Nonparametric Hypothesis Testing," *Bayesian Analysis*, 10, 297–320. [189,196]

Jara, A., and Hanson, T. E. (2011), "A Class of Mixtures of Dependent Tail-Free Processes," *Biometrika*, 98, 553–566. [189,190]

Kleinsteuber, K., Corleis, B., Rashidi, N., Nchinda, N., Lisanti, A., Cho, J. L., Medoff, B. D., Kwon, D., and Walker, B. D. (2016), "Standardization and Quality Control for High-Dimensional Mass Cytometry Studies of Human Samples," *Cytometry Part A*, 89, 903–913. [200]

Lakshminarayanan, B., Roy, D., and Teh, Y. W. (2013), "Top-Down Particle Filtering for Bayesian Decision Trees," in *International Conference on Machine Learning*, pp. 280–288. [190,193]

Lavine, M. (1992), "Some Aspects of Polya Tree Distributions for Statistical Modelling," *Annals of Statistics*, 20, 1222–1235. [189,191]

Li, M., and Ghosal, S. (2014), "Bayesian Multiscale Smoothing of Gaussian Noised Images," *Bayesian Analysis*, 9, 733–758. [196]

Lu, L., Jiang, H., and Wong, W. H. (2013), "Multivariate Density Estimation by Bayesian Sequential Partitioning," *Journal of the American Statistical Association*, 108, 1402–1410. [190,193]

Ma, L. (2017), "Adaptive Shrinkage in Pólya Tree Type Models," *Bayesian Analysis*, 12, 779–805. [190,192,193,195,197]

Ma, L., and Soriano, J. (2018), "Analysis of Distributional Variation through Graphical Multi-Scale Beta-Binomial Models," *Journal of Computational and Graphical Statistics*, 27, 529–541. [189]

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. [189,193]

Muliere, P., and Walker, S. (1997), "A Bayesian Non-parametric Approach to Survival Analysis using Polya Trees," *Scandinavian Journal of Statistics*, 24, 331–340. [189]

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting using Multivariate Normal Mixtures," *Biometrika*, 83, 67–79. [197]

Neath, A. A. (2003), "Polya Tree Distributions for Statistical Modeling of Censored Data," *Advances in Decision Sciences*, 7, 175–186. [189]

Nieto-Barajas, L. E., and Müller, P. (2012), "Rubbery Polya Tree," *Scandinavian Journal of Statistics*, 39, 166–184. [190]

Paddock, S. M. (2002), Bayesian Nonparametric Multiple Imputation of Partially Observed Data with Ignorable Nonresponse," *Biometrika*, 89, 529–538. [189]

Paddock, S. M., Ruggeri, F., Lavine, M., and West, M. (2003), "Randomized Polya Tree Models for Nonparametric Bayesian Inference," *Statistica Sinica*, 13, 443–460. [189]

Soriano, J., and Ma, L. (2017), "Probabilistic Multi-Resolution Scanning for Two-Sample Differences," *Journal of the Royal Statistical Society*, Series B, 79, 547–572. [189,192,193,195,196]

Walker, S., and Hjort, N. L. (2001), "On Bayesian Consistency," *Journal of the Royal Statistical Society*, Series B, 63, 811–821. [196]

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. (1999), "Bayesian Nonparametric Inference for Random Distributions and Related Functions," *Journal of the Royal Statistical Society*, Series B, 61, 485–527. [189]

Walker, S. G., and Mallick, B. K. (1997), "Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing," *Journal of the Royal Statistical Society*, Series B, 59, 845–860. [189]

White, J. T., and Ghosal, S. (2011), "Bayesian Smoothing of Photon-Limited Images with Applications in Astronomy," *Journal of the Royal Statistical Society*, Series B, 73, 579–599. [196]

Wong, W. H., and Ma, L. (2010), "Optional Pólya Tree and Bayesian Inference," *Annals of Statistics*, 38, 1433–1459. [190,191,195]