Bayesian Analysis (2017) **12**, Number 3, pp. 779–805

Adaptive Shrinkage in Pólya Tree Type Models

Li Ma[*,†]

Abstract. We introduce a hierarchical generalization to the Pólya tree that incorporates locally adaptive shrinkage to data features of different scales, while maintaining analytical simplicity and computational efficiency. Inference under the new model proceeds efficiently using general recipes for conjugate hierarchical models, and can be completed extremely efficiently for data sets with large numbers of observations. We illustrate in density estimation that the achieved adaptive shrinkage results in proper smoothing and substantially improves inference. We evaluate the performance of the model through simulation under several schematic scenarios carefully designed to be representative of a variety of applications. We compare its performance to that of the Pólya tree, the optional Pólya tree, and the Dirichlet process mixture. We then apply our method to a flow cytometry data with 455,472 observations to achieve fast estimation of a large number of univariate and multivariate densities, and investigate the computational properties of our method in that context. In addition, we establish theoretical guarantees for the model including absolute continuity, full nonparametricity, and posterior consistency. All proofs are given in the Supplementary Material (Ma, 2016).

MSC 2010 subject classifications: primary 62F15, 62G99; secondary 62G07.

Keywords: multi-scale modeling, Bayesian nonparametrics, hierarchical models, density estimation.

1 Introduction

In his seminal works that jump-started modern Bayesian nonparametric inference, Ferguson (1973, 1974) formalized the notion of a Dirichlet process (DP) and introduced a tail-free process that contains the DP as a special case. This tail-free process was later named the Pólya tree (PT) due to its relationship to the Pólya urn (Mauldin et al., 1992), and was popularized in the 1990s by a sequence of works (Lavine, 1992; Mauldin et al., 1992; Lavine, 1994) that investigated its various theoretical properties. Recently, several authors have proposed extensions of the PT, improving its statistical performance and furthering its applicability. Some examples are mixtures of finite PTs (Hanson and Johnson, 2002; Hanson, 2006; Jara et al., 2009), the optional PT (Wong and Ma, 2010), the multivariate/dependent PT (Trippa et al., 2011), covariate-dependent tail-free processes (Jara and Hanson, 2011), and the rubbery PT (Nieto-Barajas and Müller, 2012). The PT and its relatives—which we refer to generally as "PT-type models"—have been applied to a variety of problems including density estimation, regression, and hypothesis testing. See for example Walker et al. (1999); Berger and Guglielmi (2001); Paddock et al. (2003); Holmes et al. (2015); Ma and Wong (2011); Chen and Hanson (2014); Tansey et al. (2015) among others.

[*] Department of Statistical Science, Duke University, Durham, NC 27708, USA, li.ma@duke.edu

[†] Supported by NSF grants DMS-1309057 and DMS-1612889, and a Google Faculty Research Award.

DOI: 10.1214/16-BA1021

The PT produces probability measures through a multi-resolution generative procedure. The sample space is recursively bisected into smaller and smaller sets, and for each set A that arises during the partition, the probability assigned to A is randomly split between its two children A_l and A_r based on a Beta variable—which we shall refer to (and define later) as the probability assignment coefficient (PAC) on A—corresponding to the proportion of mass assigned to A_l. (The PACs are sometimes also referred to as "branching probabilities" by other authors (Nieto-Barajas and Müller, 2012).)

Inference under the PT adopts a "divide-and-conquer" strategy: the original nonparametric problem is transformed into a collection of independent parametric ones—each PAC is inferred through a simple Beta-Binomial conjugate model. This results in extremely fast computation, often completed within a fraction of a second even for large data sets, which makes the PT an excellent choice for problems involving large quantities of data and/or require fast output compared to other Bayesian nonparametric models, which typically rely on Markov Chain Monte Carlo (MCMC) sampling.

In addition to computational efficiency, the conjugate nature of the PT allows intuitive interpretation on how the model incorporates prior knowledge and empirical evidence in drawing inference—in terms of "shrinkage" of the empirical distribution toward a prior mean distribution. The PT allows the amount of shrinkage to vary for empirical features of different scales. This will be more fully explained in Section 2.2.

Despite its computational efficiency and interpretability, however, the PT can often perform poorly in comparison to kernel mixture-based methods such as the Dirichlet process mixture (DPM), and in particular when the underlying density contains structures of varying scales and/or smoothness. As will be explained more in this work, this is largely because the Beta-Binomial model that the PT adopts for inferring each PAC imposes a predetermined, fixed amount of shrinkage, allowing no adaptivity to the (possibly varying) smoothness and scales of the underlying distributional features. We illustrate this through a simple example.

Example 1. We simulate 750 i.i.d. data from the following mixture distribution on $[0,1]$

$$0.1\,\mathrm{U}(0,1) + 0.3\,\mathrm{U}(0.25,0.5) + 0.4\,\mathrm{Beta}_{(0.25,0.5)}(2,2) + 0.2\,\mathrm{Beta}(6000,4000),$$

where $\mathrm{Beta}_{(0.25,0.5)}(2,2)$ represents a Beta$(2,2)$ translated and scaled to be supported on the interval (0.25,0.5)—that is, the distribution with density $8(4x-1)(1-2x)$ on $(0.25,0.5)$. Figure 1 illustrates the pdf (red dashed). The "hump" on the interval $(0.25,0.5)$ constitutes a large-scale, smooth distributional structure, while the mode given by $\mathrm{Beta}(6000,4000)$ constitutes a small-scale, spiky feature.

Let us place a PT prior on the underlying distribution corresponding to a $\mathrm{Beta}(k^2,k^2)$ prior on the PACs at level k of the partition sequence (Section 2.1 gives a brief review on PTs), which is the most common specification recommended by many authors in applying the PT. (See Lavine (1992); Walker et al. (1999); Hanson and Johnson (2002); Hanson (2006); Holmes et al. (2015) for example.) The gray solid curves in Figure 1 show the posterior predictive density (PPD) of the PT. The middle and right plots give

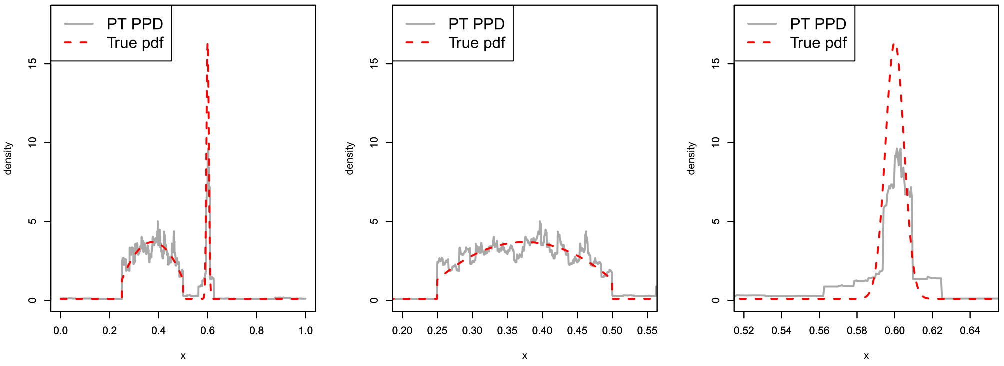


Figure 1: The true density (red dashed) and the posterior predictive density (PPD) of the PT (gray solid) for Example 1. The middle and right plots give the zoom-in views of the low- and high-resolution features.

the zoom-in views of the large-scale and small-scale features respectively. We see that overall the PT induces a decent amount of shrinkage for capturing the large-scale feature (middle), as the general shape of that feature is adequately recovered in the PPD. For the spiky feature, however, it results in too much shrinkage and thus underestimation of the mode (right). Interestingly, if we zoom into higher resolutions within the large-scale feature (middle), the PPD shows jumpy patterns of overfitting, indicating that more shrinkage is needed there at higher resolutions to ensure proper smoothness.

The above example represents a typical situation—the appropriate amount of shrinkage and smoothness *varies across locations and scales*. The PT is incapable of adjusting to the proper levels of shrinkage in a locally adaptive manner, which greatly hampers its usefulness in modern applications. Given the extremely desirable analytical and computational properties of the PT, a natural question is whether it is possible to install such adaptivity in the PT framework, but doing so in a manner that maintains the analytical simplicity and computational efficiency of the model. We will show that this is achievable in a principled hierarchical Bayesian approach through constructing hyperpriors on the Beta variances of the PACs.

The rest of the work is organized as follows. In Section 2, we construct a hyperprior on the variance of the PACs that induces a stochastically varying, adaptive rate of shrinkage across the sample space. We derive efficient inference recipe and investigate theoretical properties of the proposed model. In Section 3 we demonstrate our method in density estimation and evaluate its performance under various scenarios where the underlying density possesses a variety of features. We compare our method to two PT-type models—the PT and the optional Pólya tree (OPT) (Wong and Ma, 2010)—as well as to the very popular Dirichlet process mixture (DPM) of Gaussians (Escobar and West, 1995). Then in Section 4 we apply our method to analyzing data from a flow cytometry experiment. We conclude in Section 5 with brief remarks.

2 Method

2.1 Multi-resolution representation of probability distributions

We start by introducing some basic concepts and definitions that form the building blocks for PT-type multi-resolution modeling. Throughout this work, we let Ω denote the sample space, which can be finite or a (possibly unbounded) Euclidean rectangle such as an interval in $\mathbb{R}$ or a rectangle in $\mathbb{R}^p$. Let μ be the natural measure associated with Ω, i.e., the counting measure if Ω is finite and the Lebesgue measure if Ω is Euclidean.

Let $\mathcal{A}^1, \mathcal{A}^2, \ldots, \mathcal{A}^k, \ldots$ be a *sequence of nested dyadic partitions* of Ω. That is, each $\mathcal{A}^k = \{A_{k,1}, A_{k,2}, \ldots, A_{k,2^k}\}$ and it satisfies (i) $\Omega = \cup_{m=1}^{2^k} A_{k,m}$, (ii) $A_{k,m_1} \cap A_{k,m_2} = \emptyset$ for all $m_1 \neq m_2$, and (iii) $A_{k,m} = A_{k+1,2m-1} \cup A_{k+1,2m}$ for all $k = 1, 2, \ldots$ and $m = 1, 2, \ldots, 2^k$. In other words, starting from $\mathcal{A}^1 = \Omega$, the partition $\mathcal{A}^{k+1}$ is obtained by dividing each $A_{k,m}$ in $\mathcal{A}^k$ into two children, the *left child* $A_{k+1,2m-1}$ and the *right child* $A_{k+1,2m}$. We shall call $\mathcal{A}^k$ the partition at *resolution (or scale/level) k*. Also, we let $\mathcal{A}^{(\infty)} = \cup_{k=1}^{\infty} \mathcal{A}^k$, the totality of all partition sets that arise in all resolution levels. The partition sets form a bifurcating tree, so from now on we shall refer to $\mathcal{A}^{(\infty)}$ as the *partition tree*, and each A in $\mathcal{A}^{(\infty)}$ as a *node*.

For example, when the sample space Ω is a (possibly unbounded) one-dimensional interval, a simple strategy for constructing such a partition tree is by sequentially dividing an interval $A_{k,m} = (a, b]$—the two bounds can either be open or closed—into $A_{k+1,2m-1} = (a, c]$ and $A_{k+1,2m} = (c, b]$. A general way of division that applies to both bounded and unbounded intervals is to let $c = H^{-1}((H(a) + H(b))/2)$ for a cumulative distribution function H supported on Ω. In this case, the nodes in $\mathcal{A}^k$ are given by $(H^{-1}((m-1)/2^k), H^{-1}(m/2^k)]$ for $m = 1, 2, \ldots, 2^k$.

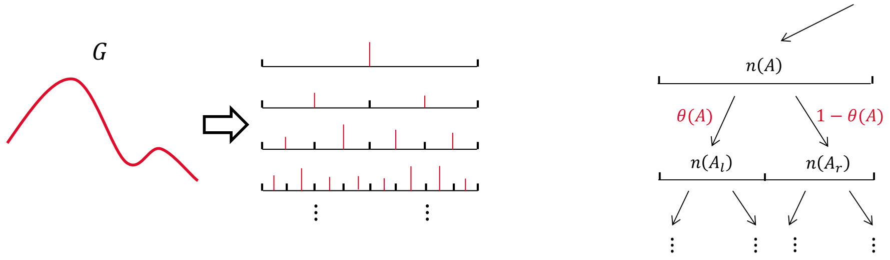
Given a partition tree $\mathcal{A}^{(\infty)}$ that generates the Borel σ-algebra, one can describe a probability distribution G by specifying how probability mass is split between the left and right children on each node $A \in \mathcal{A}^{(\infty)}$. Let A_l and A_r be the left and right children of a node A. We define the *probability assignment coefficient* (PAC) for A to be the proportion of probability mass assigned to A_l, and denote it as $\theta(A)$. So if the total probability mass on A is $G(A)$ then those assigned to the children are $G(A_l) = G(A)\theta(A)$ and $G(A_r) = G(A)(1 - \theta(A))$.

Lemma 1. *Given a partition tree $\mathcal{A}^{(\infty)}$ that generates the Borel σ-algebra, every probability distribution G can be mapped to a collection of PACs $\{\theta(A) : A \in \mathcal{A}^{(\infty)}\}$, and the mapping is unique on all A's such that $G(A) > 0$.*

Remark. A collection of PACs corresponding to G is given by $\theta(A) = G(A_l)/G(A)$ for all A with $G(A) > 0$ and $\theta(A) = 0$ otherwise.

Figure 2(a) illustrates the transformation of a distribution into PACs. Each PAC specifies the structure of the distribution at a given scale and location.

The lemma implies that inference on a distribution can be achieved by inferring the PACs, which motivates a "divide-and-conquer" strategy for nonparametric inference. More specifically, the multi-resolution transformation of G into PACs induces



(a) Transforming a distribution into PACs (red ticks)

(b) Local binomial experiment

Figure 2: The divide-and-conquer inference schema.

a corresponding decomposition of the statistical experiment that generates an i.i.d. sample of size n from G. In particular, the experiment is divided into a collection of "local" binomial experiments, carried out sequentially from low to high resolution: $\{n(A_l) \sim \text{Binomial}(n(A), \theta(A)) : A \in \mathcal{A}^k\}$ for $k = 1, 2, \ldots$ where $n(A)$ is the number of data points in A arising from the binomial experiment on A's parent in the previous resolution level $k-1$, except that for $k = 1$, $n(A) = n$ by design. See Figure 2(b) for an illustration of the local binomial experiment.

Accordingly, inferring a distribution through the PACs is divided into inference on the success probabilities of a collection of sequential binomial experiments. Viewed this way, the PT model provides a simple solution to this problem—it places independent conjugate Beta priors on the success probabilities. The posterior conjugacy of the PT follows immediately from the Beta-Binomial conjugacy. Thus inference under the PT is analytically tractable and computationally efficient.

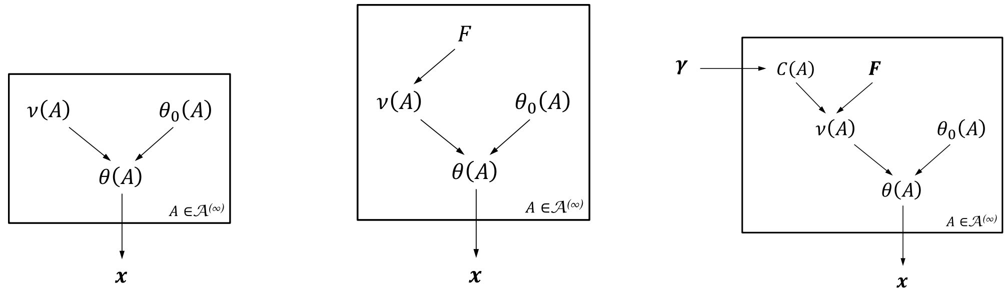
2.2 Adaptive shrinkage in the Pólya tree

One can now understand how the PT applies shrinkage toward the prior mean by viewing how each of the Binomial experiments does so. Under the PT model, $\theta(A) \sim \text{Beta}(\alpha_l(A), \alpha_r(A))$ for all $A \in \mathcal{A}^{(\infty)}$. (A popular specification has $\alpha_l(A) = \alpha_r(A) = k^2$ for $A \in \mathcal{A}^k$ as mentioned previously.) We shall prefer an alternative parametrization of Beta distributions in terms of its mean $\theta_0(A) = \alpha_l(A)/(\alpha_l(A) + \alpha_r(A))$ and precision $\nu(A) = \alpha_l(A) + \alpha_r(A)$. The posterior distribution of $\theta(A)$ is still Beta with mean

$$\theta_1(A) = \mathrm{E}(\theta(A)|\boldsymbol{x}) = \theta_0(A) \cdot \frac{\nu(A)}{\nu(A) + n(A)} + \frac{n(A_l)}{n(A)} \cdot \frac{n(A)}{\nu(A) + n(A)}$$

and precision $\tilde{\nu}(A) = \nu(A) + n(A)$. The posterior mean is a weighted average between $\theta_0(A)$, or the prior mean, and $n(A_l)/n(A)$, or the PAC on A of the empirical distribution. The level of shrinkage for $\theta(A)$ is controlled by the precision parameter $\nu(A)$, and thus we shall refer to $\nu(A)$ also as the *shrinkage parameter.*

The prior mean of the PT is the probability distribution given by the PACs $\{\theta_0(A) : A \in \mathcal{A}^{(\infty)}\}$, which we call Q_0. The posterior mean of the PT is the distribution given



(a) Pólya tree (b) Adaptive Pólya tree (c) Markov adaptive Pólya tree

Figure 3: Graphical representations of the PT, APT, and Markov-APT.

by the PACs $\{\theta_1(A) : A \in \mathcal{A}^{(\infty)}\}$. Intuitively, the posterior mean of a PT is a weighted average between Q_0 and the empirical distribution, but the weighting can be scale and location dependent. Figure 3(a) provides a graphical model representation of the PT.

From a hierarchical Bayesian perspective, we can incorporate data-adaptivity into the shrinkage by placing a hyperprior F_A on $\nu(A)$ thereby allowing the appropriate level of shrinkage for $\theta(A)$ to be inferred from the data. The subscript "A" in F_A indicates that the hyperprior can be specified differently across A's. In most applications, however, one typically does not have the prior knowledge to differentially specify the precision parameter for each A. Instead, we shall recommend a default choice for F_A that does not depend on A (details later). As such, we shall from now on suppress the notation, and use F to indicate the hyperprior on each $\nu(A)$.

This leads to a generative hierarchical model specified by F and Q_0. Using $\boldsymbol{\phi} = (F, Q_0)$ to represent the totality of all hyperparameters, we can write the model as

$$\begin{aligned} \nu(A) \,|\, \boldsymbol{\phi} &\sim F \\ \theta(A) \,|\, \boldsymbol{\phi}, \boldsymbol{\nu} &\sim \text{Beta}(\theta_0(A)\nu(A), (1-\theta_0(A))\nu(A)), \end{aligned}$$

for all $A \in \mathcal{A}^{(\infty)}$, where $\boldsymbol{\nu} = \{\nu(A) : A \in \mathcal{A}^{(\infty)}\}$, the collection of all shrinkage parameters. Figure 3(b) provides a graphical representation of this model.

We consider priors F supported on $(0, \infty]$. Note that we allow $\nu(A) = \infty$, in which case $\text{Beta}(\theta_0(A)\nu(A), (1-\theta_0(A))\nu(A))$ is a point mass at $\theta_0(A)$, corresponding to complete shrinkage of $\theta(A)$ to the prior mean $\theta_0(A)$. A first necessary question to address is whether this model always generates well-defined probability measures. In other words, given a set of PACs $\{\theta(A) : A \in \mathcal{A}^{(\infty)}\}$ arising from the above model, does there (almost surely) exist a distribution G such that $G(A_l|A) = \theta(A)$ for all $A \in \mathcal{A}^{(\infty)}$? The answer is positive by Theorem 3.3.2 in Ghosh and Ramamoorthi (2003). Hence we can define this model as a distribution on probability measures.

Definition 1 (Adaptive Pólya tree)**.** A probability measure Q is said to have an *adaptive Pólya tree* (APT) distribution with parameters $\boldsymbol{\phi} = (F, Q_0)$ if the corresponding PACs of Q, $\{\theta(A) : A \in \mathcal{A}^{(\infty)}\}$, are generated from the above hierarchical model. We write $Q \sim \text{APT}(F, Q_0)$, or equivalently $\text{APT}(F, \boldsymbol{\theta}_0)$.

2.3 Stochastically increasing shrinkage and adaptive smoothing

In many applications such as density estimation, a reasonable assumption adopted (explicitly or implicitly) in all statistical methods is the smoothness of the underlying density. Indeed, even "jumpy" densities—those with sharp changes—must be assumed to eventually smooth out at high enough resolutions as opposed to infinitely oscillating in arbitrarily small regions, because otherwise reliable estimation is infeasible. In PT-type models, it is well-understood that smoothing translates into an *increase in shrinkage or decrease in the prior variance of the PACs for higher resolutions* (Kraft, 1964). This is exactly the motivation for increasing Beta parameters with the level (such as k^2) in the PT (Lavine, 1992). Generally, the faster the shrinkage increases with the resolution level, the smoother the induced process.

However, to effectively model densities involving structures of various scales and varying smoothness across the sample space, it is necessary to incorporate a potentially heterogeneous rate of increasing shrinkage across the sample space into PT-type models. Because the proper varying rate of increasing shrinkage across the sample space is unknown *a priori*, we propose to infer it from the data.

To this end, next we shall construct a generative model for the priors on the Beta variances in an APT to allow data-adaptive rates of increasing shrinkage for higher resolutions across the sample space. First, we introduce a latent mixture representation for F whose mixing components represent different extents of shrinkage. Let us specify F using a mixture of I component distributions in a monotone increasing stochastic order

$$F_1 \prec F_2 \prec \ldots \prec F_I.$$

In particular, we can choose these components to have non-overlapping supports. For example F_i may be supported on an interval $[a(i), a(i+1))$ where $a(i)$ is increasing sequence in i. For now we shall treat the number of components I and each F_i as given, but will provide guidelines on choosing them in Section 2.6. We let $\boldsymbol{F} = (F_1, F_2, \ldots, F_I)$ denote the totality of all component distributions.

Moreover, we introduce a latent state variable $C(A)$ for each $A \in \mathcal{A}^{(\infty)}$ that indicates the mixture component $\nu(A)$ comes from:

$$\nu(A) \,|\, C(A) = i \;\sim\; F_i,$$

for $i = 1, 2, \ldots, I$. We refer to $C(A)$ as the *shrinkage state* on A, and let $\mathcal{C} = \{C(A) : A \in \mathcal{A}^{(\infty)}\}$ be the collection of all shrinkage states.

Now we can enforce a stochastic rate of increasing shrinkage along each branch of $\mathcal{A}^{(\infty)}$ by specifying a joint prior on $\mathcal{C}$ that prevents the shrinkage state from moving lower in any branch. That is, if A_p is A's parent in $\mathcal{A}^{(\infty)}$, then we require $C(A) \geq C(A_p)$. A simple stochastic model for $\mathcal{C}$ that can help us impose such a constraint is the Markov tree (MT) (Crouse et al., 1998), which links the $C(A)$'s using a Markov process such that the shrinkage state $C(A)$ depends on that of A_p through Markov transition. The Markov process is initiated on the root, Ω, as follows

$$P(C(\Omega) = i) = \gamma_i(\Omega) \quad \text{for } i \in \{1, 2, \ldots, I\},$$

where the $\gamma_i(\Omega)$'s are called the *initial state probabilities*, and can be put into a vector

$$\boldsymbol{\gamma}(\Omega) = (\gamma_1(\Omega), \gamma_2(\Omega), \ldots, \gamma_I(\Omega)).$$

Then for each $A \neq \Omega$, $C(A)$ is determined sequentially based on its parent according to

$$P(C(A) = i' \,|\, C(A_p) = i) = \gamma_{i,i'}(A) \quad \text{for } i, i' \in \{1, 2, \ldots, I\},$$

where $\gamma_{i,i'}(A)$ is called the *state transition probability*, which can be organized into a *transition probability matrix*

$$\boldsymbol{\gamma}(A) = \begin{pmatrix} \gamma_{1,1}(A) & \gamma_{1,2}(A) & \cdots & \gamma_{1,I}(A) \\ \gamma_{2,1}(A) & \gamma_{2,2}(A) & \cdots & \gamma_{2,I}(A) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{I,1}(A) & \gamma_{I,2}(A) & \cdots & \gamma_{I,I}(A) \end{pmatrix}.$$

Note that we set this transition probability matrix without reference to A_p—$\gamma_{i,i'}(A)$ gives the probability for $C(A) = i'$ given that $C(A_p) = i$, regardless of what A_p is.

The desired stochastically increasing shrinkage is achieved when the transition matrices are all upper-triangular. That is, $\gamma_{i,i'}(A) = 0$ if $i > i'$ for all $A \in \mathcal{A}^{(\infty)}$. From now on, we shall use $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}(A) : A \in \mathcal{A}^{(\infty)}\}$ to denote the collection of all initial state probabilities and transition probability matrices needed for specifying the MT. In Section 2.6 we present recommended choices of these hyperparameters. In particular, without additional information, we shall recommend a simple choice of $\boldsymbol{\gamma}(A)$ that is common for all A, in which case the "(A)" notation is unnecessary. We keep the "(A)" notation because it maintains the generality of the model and will be needed in describing the corresponding posterior.

Putting the pieces together, now we have the following hierarchical model for a probability distribution with hyperparameters $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \boldsymbol{F}, Q_0)$:

$$\mathcal{C} \,|\, \boldsymbol{\phi} \sim \mathrm{MT}(\boldsymbol{\gamma}), \qquad \nu(A) \,|\, \boldsymbol{\phi}, \mathcal{C} \sim \sum_{i=1}^{I} F_i \cdot \mathbf{1}_{C(A)=i},$$
$$\theta(A) \,|\, \boldsymbol{\phi}, \boldsymbol{\nu}, \mathcal{C} \sim \mathrm{Beta}(\theta_0(A)\nu(A), (1-\theta_0(A))\nu(A)),$$

for all $A \in \mathcal{A}^{(\infty)}$. A graphical representation of this model is given in Figure 3(c).

Because this hierarchical model also generates a probability measure with probability 1, one can again define it formally as a distribution on probability measures.

Definition 2 (Markov adaptive Pólya tree)**.** A probability measure Q is said to have a *Markov adaptive Pólya tree* (Markov-APT) distribution with parameters $\boldsymbol{\phi} = (\boldsymbol{\gamma}, \boldsymbol{F}, Q_0)$ if Q corresponds to the collection of PACs $\{\theta(A) : A \in \mathcal{A}^{(\infty)}\}$ generated from the above hierarchical model. We write $Q \sim \text{Markov-APT}(\boldsymbol{\gamma}, \boldsymbol{F}, Q_0)$. When $\boldsymbol{\gamma}(A)$ is upper-triangular for all $A \in \mathcal{A}^{(\infty)}$, we say that the Markov-APT is *stochastically increasing.*

In practice, it is typically sufficient to infer Q up to some finite maximum resolution K (Hanson, 2006). Such a "truncated" or "finite" Markov-APT can be constructed by simply setting the conditional measure $Q(\cdot|A) = Q_0(\cdot|A)$ for all A at level K. A related question is how one may choose the maximum resolution K in applications. Some recent works (Hanson, 2006; Watson et al., 2014) have investigated how well finite PT-type models can characterize the underlying distribution according to some global measure of distributional properties (such as L_1 distance and Kullback–Leibler divergence), and have proposed rules for setting K based on these results. One limitation of using global distances for choosing K is that they are insensitive to highly local structures, and the rules derived based on these results can be overly conservative when one is interested in characterizing local distributional features, which we believe is the situation when PT-type models are most favorable compared to other nonparametric models. (See our simulation results for example.) As such, we recommend setting the maximum resolution at what the available computational resources allow, and/or truncate the partition whenever the number of sample size is too few (e.g., < 5) on a node, in which case reliable inference on higher resolutions becomes infeasible anyway.

2.4 Bayesian inference with the Markov-APT

Next we address how to carry out posterior inference for the Markov-APT. We show that the full posterior can be derived following a general recipe for hierarchical models (Gelman et al., 2013, Sec. 5.3) and inference can proceed in a usual manner through drawing a sample from the posterior and/or computing some summary statistics such as the posterior mean. In particular, the posterior is available analytically (up to some numerical approximation) and so no MCMC is needed. In particular, the full posterior $\pi(\mathcal{C}, \boldsymbol{\nu}, \boldsymbol{\theta} \,|\, \boldsymbol{\phi}, \boldsymbol{x})$ can be described in three pieces: (i) $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}, \boldsymbol{\nu}, \mathcal{C}, \boldsymbol{x})$, (ii) $\pi(\boldsymbol{\nu} \,|\, \boldsymbol{\phi}, \mathcal{C}, \boldsymbol{x})$, and (iii) $\pi(\mathcal{C} \,|\, \boldsymbol{\phi}, \boldsymbol{x})$ as follows.

(i) $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}, \boldsymbol{\nu}, \mathcal{C}, \boldsymbol{x})$. This conditional posterior follows directly from the Beta-Binomial conjugacy. Specifically, we have

$$\theta(A) \,|\, \boldsymbol{\phi}, \boldsymbol{\nu}, \boldsymbol{x} \sim \text{Beta}(\theta_1(A)\tilde{\nu}(A), (1-\theta_1(A))\tilde{\nu}(A)) \quad \text{for all } A \in \mathcal{A}^{(\infty)},$$

where as before $\tilde{\nu}(A) = \nu(A) + n(A)$ and $\theta_1(A) = (\theta_0(A)\nu(A) + n(A_l))/\tilde{\nu}(A)$.

(ii) $\pi(\boldsymbol{\nu} \,|\, \boldsymbol{\phi}, \mathcal{C}, \boldsymbol{x})$. Due to the conjugacy of finite mixture models, the conditional posterior for $\nu(A)$ is still an I-component mixture

$$\nu(A) \,|\, \boldsymbol{\phi}, \mathcal{C}, \boldsymbol{x} \sim \sum_{i=1}^{I} \tilde{F}_{A,i} \cdot \mathbf{1}_{C(A)=i},$$

where each new mixture component distribution $\tilde{F}_{A,i}$ (which depends on A) has density

$$d\tilde{F}_{A,i}(\nu) = dF_i(\nu) \cdot M_A(\boldsymbol{\theta}_0, \nu)/M_{A,i}(\boldsymbol{\theta}_0),$$

for $i = 1, 2, \ldots, I$, where

$$M_A(\boldsymbol{\theta}_0, \nu) = \frac{\Gamma(\theta_0(A)\nu + n(A_l))\Gamma((1-\theta_0(A))\nu + n(A_r))\Gamma(\nu)}{\Gamma(\nu + n(A))\Gamma(\theta_0(A)\nu)\Gamma((1-\theta_0(A))\nu)}$$

is the marginal likelihood of the local Binomial experiment on A given ν, and

$$M_{A,i}(\boldsymbol{\theta}_0) = \int M_A(\boldsymbol{\theta}_0, \nu) dF_i(\nu)$$

is the marginal likelihood of the local Binomial experiment on A given $\theta_0(A)$ and $C(A) = i$, which can be numerically evaluated using standard strategies for one-dimensional integration. We describe a simple grid-based finite Riemann method in Supplementary Material S2 (Ma, 2016). Note that when $\nu = \infty$, $M_A(\boldsymbol{\theta}_0, \infty) := \lim_{\nu\uparrow\infty} M_A(\boldsymbol{\theta}_0, \nu) = \theta_0(A)^{n(A_l)}(1-\theta_0(A))^{n(A_r)}$.

(iii) $\pi(\mathcal{C} \,|\, \boldsymbol{\phi}, \boldsymbol{x})$. The last piece is the marginal posterior on $\mathcal{C}$, which follows again from the conjugacy of finite mixtures (of which the MT is a special case) to be an MT. The initial and transition probabilities of the posterior MT are computable analytically through a forward–backward algorithm, which is a dynamic programming strategy that can be used for evaluating marginal posterior state probabilities of Markov models. See Liu (2001) for more background. The forward step, or the summation step, is a bottom–up (leaf-to-root) recursion on the partition tree and the backward step, or the sampling step, a top–down recursion.

To describe the algorithm, we first define a mapping ξ_A for each $A \in \mathcal{A}^{(\infty)}$ as follows

$$\xi_A(i, \boldsymbol{\phi}) := \begin{cases} \int q(\boldsymbol{x}|A)\pi(dq \,|\, \boldsymbol{\phi}, C(A_p) = i) & \text{if } A \in \mathcal{A}^{(\infty)} \backslash \{\Omega\} \\ \int q(\boldsymbol{x}|A)\pi(dq \,|\, \boldsymbol{\phi}) & \text{if } A = \Omega, \end{cases}$$

for $i = 1, 2, \ldots, I$, where $q(\boldsymbol{x}|A) := \prod_{x \in A} \frac{q(x)}{Q(A)}$ with $q = dQ/d\mu$, and A_p is the parent of A in $\mathcal{A}^{(\infty)}$. Intuitively, $\xi_A(i, \boldsymbol{\phi})$ gives the marginal likelihood of the "submodel" on A—the Markov-APT with A being the sample space—given that the shrinkage state on A_p is i. Note that when $A = \Omega$, it does not have a parent, and $\xi_\Omega(i, \boldsymbol{\phi})$ is equal for all i to the overall marginal likelihood of the Markov-APT. The following lemma provides a bottom–up recursive recipe for computing $\xi_A(i, \boldsymbol{\phi})$.

Lemma 2 (Forward-summation)**.** *For $A \in \mathcal{A}^{(\infty)} \backslash \{\Omega\}$,*

$$\xi_A(i, \boldsymbol{\phi}) = \begin{cases} \sum_{i'=1}^{I} \gamma_{i,i'}(A) \cdot M_{A,i'}(\boldsymbol{\theta}_0) \cdot \xi_{A_l}(i', \boldsymbol{\phi}) \xi_{A_r}(i', \boldsymbol{\phi}) & \text{if } n(A) > 1 \\ q_0(\boldsymbol{x}|A) & \text{if } n(A) = 1 \text{ or } A \text{ has no children} \\ 1 & \text{if } n(A) = 0, \end{cases}$$

where $q_0(\boldsymbol{x}|A) = \prod_{x_i \in A} q_0(x_i)/Q_0(A)$ with $q_0 = dQ_0/d\mu$. For $A = \Omega$, we simply replace $\gamma_{i,i'}(A)$ with $\gamma_{i'}(\Omega)$ in the above equation. In particular, $\xi_\Omega(1, \boldsymbol{\phi})$ is the overall marginal likelihood, as a function of the hyperparameters $\boldsymbol{\phi}$.

Remark I. There are two situations in which a node A has no children. The first is when it is atomic—i.e., $A = \{a\}$ for a single value a—and so cannot be further divided. In this case, $q_0(\cdot|A)$ is a unit mass at the value a and so $q_0(\boldsymbol{x}|A) = 1$ as well. The other situation is when the partition is truncated at some maximum level K, and so all $A \in \mathcal{A}^K$ have no children. In this case $q(\boldsymbol{x}|A) = q_0(\boldsymbol{x}|A)$ by construction.

Remark II. This lemma shows that one can compute the mapping for A based on those for its children, A_l and A_r (hence bottom–up).

Remark III. Given that the data are arising from some true absolutely continuous distribution P_0, with P_0 probability 1, all nodes at deep enough levels of $\mathcal{A}^{(\infty)}$ will either have no children or contain ≤ 1 observation, and hence the recursive computation can be applied with or without setting a maximum resolution. One can start the recursion from those A's such that $n(A) \leq 1$ but $n(A_p) \geq 2$, because all descendants of such A's have no more than one data point and so the mapping is known there.

After computing $\{\xi_A(i, \boldsymbol{\phi}) : A \in \mathcal{A}^{(\infty)}$ and $i = 1, 2, \ldots, I\}$, we can then carry out a backward (top–down) recursion to derive the marginal posterior of $\mathcal{C}$.

Theorem 1 (Backward-sampling)**.** *The marginal posterior of the shrinkage states is*

$$\mathcal{C} \,|\, \boldsymbol{\phi}, \boldsymbol{x} \sim \mathrm{MT}(\tilde{\boldsymbol{\gamma}})$$

whose initial state and transition probabilities $\tilde{\boldsymbol{\gamma}} = \{\tilde{\boldsymbol{\gamma}}(A) : A \in \mathcal{A}^{(\infty)}\}$ *are as follows.*

- *The initial state probability vector:* $\tilde{\boldsymbol{\gamma}}(\Omega) = \boldsymbol{\gamma}(\Omega)\boldsymbol{D}''(\Omega)/\xi_\Omega(1, \boldsymbol{\phi})$*;*
- *The state transition probability matrix:* $\tilde{\boldsymbol{\gamma}}(A) = \boldsymbol{D}'(A)^{-1}\boldsymbol{\gamma}(A)\boldsymbol{D}''(A)$ *for all* $A \in \mathcal{A}^{(\infty)} \backslash \{\Omega\}$*,*

where for all $A \in \mathcal{A}^{(\infty)}$*,* $\boldsymbol{D}'(A)$ *is the* $I \times I$ *diagonal matrix with the diagonal elements being* $\xi_A(i, \boldsymbol{\phi})$ *for* $i = 1, 2, \ldots, I$*, and* $\boldsymbol{D}''(A)$ *is the* $I \times I$ *diagonal matrix with the diagonal elements being* $M_{A,i}(\boldsymbol{\theta}_0)\xi_{A_l}(i, \boldsymbol{\phi})\xi_{A_r}(i, \boldsymbol{\phi})$ *for* $i = 1, 2, \ldots, I$ *if* A *has children and* $q_0(\boldsymbol{x}|A)$ *if not.*

Remark. In particular, for any A with $n(A) \leq 1$, by the theorem $\tilde{\boldsymbol{\gamma}}(A) = \boldsymbol{\gamma}(A)$. So *a posteriori* the MT on A with no more than one observation is the same as the prior MT.

We have completely described the three components of the full posterior for $(\boldsymbol{\theta}, \boldsymbol{\nu}, \mathcal{C})$. One can draw from the joint posterior by sampling in the order of $\pi(\mathcal{C} \,|\, \boldsymbol{\phi}, \boldsymbol{x})$, $\pi(\boldsymbol{\nu} \,|\, \boldsymbol{\phi}, \mathcal{C}, \boldsymbol{x})$, and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}, \boldsymbol{\nu}, \mathcal{C}, \boldsymbol{x})$. The forward–backward recursion given in Lemma 2 and Theorem 1 is the most computationally intense step in the posterior inference. Fortunately, for any given data set and prior specification, this recursion only needs to be carried out *once and for all*, because $\tilde{\boldsymbol{\gamma}}$ stays the same for all the posterior draws.

With posterior draws $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\nu}^{(1)}, \mathcal{C}^{(1)}), (\boldsymbol{\theta}^{(2)}, \boldsymbol{\nu}^{(2)}, \mathcal{C}^{(2)}), \ldots, (\boldsymbol{\theta}^{(B)}, \boldsymbol{\nu}^{(B)}, \mathcal{C}^{(B)})$, one can carry out Bayesian inference in the usual manner. In particular, when one is interested in the unknown distribution Q, one can use $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(B)}$ to obtain a posterior sample $Q^{(1)}, Q^{(2)}, \ldots, Q^{(B)}$ while discarding the other variables. Sampling from the posterior is extremely efficient. We illustrate inference based on such a posterior sample in our analysis of a flow cytometry data set in Section 4 and evaluate the computation efficiency of the method in the context of that analysis.

Moreover, some posterior summaries can be evaluated analytically without resorting to posterior sampling at all. In particular, the PPD at any $x^* \in \Omega$ is equal to $\xi^*_\Omega(1,\boldsymbol{\phi})/\xi_\Omega(1,\boldsymbol{\phi})$, where $\xi^*_\Omega(1,\boldsymbol{\phi})$ is the overall marginal likelihood computed according to Lemma 2 for a data set that contains the original data plus an additional point at x^*. This is particularly useful in applications such as density estimation because it avoids Monte Carlo errors in computing an estimator. In our numerical examples in Section 3, we compute all PPDs this way. Note that after computing the $\xi_A(i,\boldsymbol{\phi})$ mappings for the original data set, the corresponding mappings $\xi^*_A(i,\boldsymbol{\phi})$ for the new data set can be obtained by updating only the branch of partition tree in which the new data point x^* falls, because the mapping stays the same on all other nodes.

2.5 Theoretical properties

We shall establish four basic but important theoretical properties of the Markov-APT model. The first two properties will guide us in prior specification. The latter two provide theoretical guarantees in terms of prior support and posterior consistency, which are deemed necessary for Bayesian nonparametric models.

Theorem 2 (Prior centering)**.** *The mean of a Markov-APT is Q_0. That is, for any Borel set $B \subset \Omega$, a random measure Q with a Markov-APT distribution satisfies $\mathrm{E}\,Q(B) = Q_0(B)$.*

We shall focus on Markov-APTs as a tool for modeling densities. In order to model a density, we must ensure that a random measure Q generated from a Markov-APT has a density. Earlier works in the literature have established two general approaches for achieving absolute continuity for PT-type priors, which could both be adopted for the Markov-APT. The first is to force $\nu(A)$ to increase with the level of A at a sufficiently fast rate (Lavine, 1992). For the Markov-APT, this can be achieved by choosing $\boldsymbol{F}$ such that for any $A \in \mathcal{A}^k$, $\nu(A) > l(k)$ with probability 1 where $l(k)$ is a positive function in k that satisfies $\sum_{k=1}^{\infty} 1/l(k) < \infty$ (Kraft, 1964). But this would impose a minimum amount of prespecified shrinkage homogeneously across the sample space, which is exactly the undesirable feature of the PT that we wish to avoid through adaptive shrinkage. Hence we prefer an alternative strategy for ensuring absolute continuity (Wong and Ma, 2010), which is to include a “complete shrinkage” state as follows.

Theorem 3 (Absolute continuity)**.** *Suppose Q has a stochastically increasing Markov-APT distribution for which $F_I = \mathbf{1}_\infty$, a point mass at ∞, and there exists $\delta > 0$ such that for all large enough k, the state transition probabilities for any $A \in \mathcal{A}^k$ satisfies*

$$\gamma_{i,I}(A) > \delta \quad \textit{for } i = 1, 2, \ldots, I-1,$$

then with probability 1, $Q \ll Q_0$ (i.e., Q is absolutely continuous with respect to Q_0). In particular, if $Q_0 \ll \mu$, then $Q \ll \mu$.

Remark. The complete shrinkage state eliminates the need for increasing lower bound on the support of $\nu(A)$ to ensure absolute continuity.

Next we establish two theoretical guarantees for inference using the Markov-APT model. The first regards the flexibility of the model—it shows that Markov-APT enjoys full prior support. Thus inference with the Markov-APT is fully nonparametric.

Theorem 4 (Large prior support)**.** *Suppose $Q \sim$ Markov-APT$(\boldsymbol{\gamma}, \boldsymbol{F}, Q_0)$ and the conditions in Theorem 3 are satisfied. In addition, suppose (i) $Q_0 \ll \mu$, (ii) $I \geq 2$, (iii) $\exists \delta' > 0$ such that for all large enough k, $\gamma_{i,I}(A) < 1 - \delta'$ for all $A \in \mathcal{A}^k$ and $i = 1, 2, \ldots, I-1$, and (iv) $\exists \epsilon > 0$ and $N > 0$ such that $F_i\big((0, N]\big) > \epsilon$ for all $i = 1, 2, \ldots, I-1$ and all $A \in \mathcal{A}^{(\infty)}$. Then for any distribution $G \ll Q_0$ and any $\tau > 0$, we have*

$$\pi\left(Q : \int |q - g| d\mu < \tau\right) > 0,$$

where $q = dQ/d\mu$ and $g = dG/d\mu$ are the corresponding densities with respect to μ.

The next result regards the asymptotic consistency of inference using the Markov-APT and it guarantees that with more and more data, the posterior will eventually concentrate into any weak neighborhood of the true density. For any probability measure P_0 on Ω, a *weak neighborhood* U of P_0 is a set of probability measures on Ω of the form

$$U = \left\{Q : \Big| \int f_i(\cdot) dQ - \int f_i(\cdot) dP_0 \Big| < \epsilon_i, \text{ for } i = 1, 2, \ldots, K\right\},$$

for any bounded continuous functions f_i's and non-negative constants ϵ_i's.

Theorem 5 (Posterior consistency under weak topology)**.** *Suppose $X_1, X_2, \ldots, X_n, \ldots$ are i.i.d. data from Q, and let $\pi(\cdot)$ be a stochastically increasing Markov-APT prior on Q that satisfies the conditions in Theorem 4 and let $\pi(\cdot|X_1, X_2, \ldots, X_n)$ be the corresponding posterior. Then for any $P_0 \ll Q_0$ with bounded density dP_0/dQ_0, we have*

$$\pi(U \,|\, X_1, X_2, \ldots, X_n) \longrightarrow 1 \quad \text{as } n \to \infty,$$

with $P_0^{(\infty)}$ probability 1 for any weak neighborhood U of P_0.

2.6 Prior specification for Markov-APTs

Next we provide guidelines for specifying a Markov-APT, i.e., for choosing the hyperparameters $\{(\theta_0(A), \boldsymbol{\gamma}(A), \boldsymbol{F}) : A \in \mathcal{A}^{(\infty)}\}$, in the context of density estimation. The objective is to balance robustness (for a variety of distributional features) and parsimony—involving only a small number of hyperparameters (i.e., the tuning parameters). We give an empirical Bayes strategy to set the hyperparameters adaptively.

Prior choice of $\theta_0(A)$. $\boldsymbol{\theta}_0$ is the PACs corresponding to Q_0, the prior mean of the model. That is, $\theta_0(A) = Q_0(A_l)/Q_0(A)$ for all $A \in \mathcal{A}^{(\infty)}$. Depending on the application, one may or may not have relevant prior knowledge for setting the prior mean. In lack of such knowledge, a simple default is to let Q_0 be uniform over a wide enough interval. Another common situation is that one wants to center the Markov-APT around some parametric family such as the Gaussian location-scale family as in Berger and Guglielmi (2001); Hanson (2006), without specifying which member in that family Q_0 is. This can be achieved by placing another layer of hierarchical prior on Q_0, e.g. on the location and scale parameters (Hanson, 2006), forming a mixture of Markov-APTs.

Prior choice of $\boldsymbol{\gamma}(A)$. In density estimation, we shall choose $\boldsymbol{\gamma}(A)$ to be upper-triangular to achieve adaptive smoothing through stochastically increasing shrinkage. The most simple choice adopts the same transition probabilities for all A. For example,

$$\gamma_{i,i'}(A) = \begin{cases} 1/(I-i) & \text{if } i \leq i' \\ 0 & \text{if } i > i', \end{cases}$$

for all $i, i' \in \{1, 2, \ldots, I\}$ and $A \in \mathcal{A}^{(\infty)} \backslash \{\Omega\}$. That is, given A's parent is in shrinkage state i, A can take any higher shrinkage state (including i) with equal probability. This "uniform" transition probability specification is a special case of a more flexible kernel specification. Specifically, we can choose a kernel function $k(i, i')$ such that $k(i, i')$ is non-increasing in $|i - i'|$, and set the transition probability

$$\gamma_{i,i'}(A) \propto k(i, i')\mathbf{1}_{i \leq i'}.$$

A kernel $k(i, i')$ strictly decreasing in $|i-i'|$ will introduce "stickiness" into the shrinkage levels between a node and its parent (and thus also with its siblings and other relatives to a lesser degree). It encourages the shrinkage level to change gradually among nearby nodes in the partition tree, which is particularly useful when the smoothness of the underlying distribution tends to be similar for places closed in the sample space.

Of course, typically one does not know *a priori* whether and to what extent such sticky shrinkage is needed. Thus it is useful to allow the stickiness to be adaptively determined. One natural way to achieving this additional adaptivity is to choose a kernel that contains a tuning parameter for the stickiness, and use empirical Bayes to choose its value. For example, consider the exponential kernel

$$k(i, i') = e^{-\beta|i-i'|},$$

where $\beta \geq 0$ is the stickiness parameter. Note that $\beta = 0$ corresponds to the uniform transition probabilities described above, while a large positive value of β corresponds to strong stickiness in shrinkage. Finally, the initial state probabilities can be set to $\gamma_1(\Omega) = \gamma_2(\Omega) = \cdots = \gamma_I(\Omega) = 1/I$.

Prior choice of $\boldsymbol{F}$. Following a common practice (Gelman et al., 2013, Sec. 5.3) in Bayesian hierarchical modeling for Beta-binomial experiments, we specify prior on $\nu(A)$ on the log scale. First, we determine a *global* support for $\log_{10} \nu(A)$, i.e. the union of the supports of all F_i. A convenient choice of the global support, aside from the complete shrinkage state, is a finite interval $[L, U]$.

A simple and robust strategy for choosing the interval is to choose a wide enough range that covers all reasonable shrinkage levels and yet not so wide as to induce excessive prior probability in extremely strong or weak shrinkage levels. We recommend setting $[L, U] = [-1, 4]$. On one end, $\log_{10} \nu(A) = -1$ corresponds to a prior sample size of $10^{-1} = 0.1$, enforcing little shrinkage, while on the other $\log_{10} \nu(A) = 4$ corresponds to shrinkage equivalent to about 10,000 prior "observations" for the local binomial experiment, resulting in very strong shrinkage. We have experimented with treating L and U as tuning parameters and choosing their values in a data dependent fashion using empirical Bayes (described below), but that resulted in little improvement over the very simple choice of $[-1, 4]$ in all of the numerical scenarios we investigated.

Given the global support of $\log_{10} \nu(A)$, $[L, U] \cup \{\infty\}$, where ∞ is included for the complete shrinkage state, we now divide this support into I non-overlapping intervals. Specifically, we let the first $(I-1)$ intervals evenly divide $[L, U]$ and the last being $\{\infty\}$:

$$[a(1), a(2)), \;\; [a(2), a(3)), \;\; \ldots, \;\; [a(I-1), a(I)), \;\; \{\infty\},$$

where $a(i) = L + (i-1) \cdot (U - L)/(I-1)$. Then we let

$$F_i : \;\; \log_{10} \nu(A) \sim \text{Uniform}(a(i), a(i+1)),$$

for $i = 1, 2, \ldots, I-1$ and F_I being a point mass at ∞.

Choosing the tuning hyperparameters by empirical Bayes. Our default prior specification is parsimonious in that it reduces the number of free parameters down to two—the number of shrinkage states I and the stickiness parameter β for the transition kernel. One can set them in a data-adaptive manner by empirical Bayes. Lemma 2 provides the recipe for computing $\xi_\Omega(1, \boldsymbol{\phi}(I, \beta))$, the overall marginal likelihood. Maximizing this likelihood over a grid of allowed values of (I, β) produces the maximum marginal likelihood estimate (MMLE) $(\hat{I}, \hat{\beta})$, which one can then keep fixed in the inference.

2.7 Adaptive partitioning in multi-dimensional cases

So far the proposed model relies on a pre-determined, fixed bifurcating partition tree $\mathcal{A}^{(\infty)}$ on Ω that generates the Borel σ-algebra. In multivariate cases, however, there are a multitude of different dyadic partition sequences that all can generate the Borel σ-algebra, and not all partition trees are equally effective for characterizing the underlying distribution. In particular, an efficient partition tree should divide more frequently along the dimensions in which the underlying distribution changes sharply than in those the density is relatively flat. While the proper partitioning is not known *a priori*, one can infer it under the Bayesian hierarchical framework by treating $\mathcal{A}^{(\infty)}$ as a parameter and placing a hyperprior on it. Ideally, the hyperprior on $\mathcal{A}^{(\infty)}$ should be so chosen that it is flexible enough to cover a variety of partition sequences, but it must also maintain the analytical simplicity and computational efficiency of the resulting model. Otherwise a critical advantage for PT-type models would be lost.

One class of hyperpriors on $\mathcal{A}^{(\infty)}$ that satisfy these criteria is a dyadic recursive partitioning (RP) prior (Wong and Ma, 2010; Ma, 2013), which can be described inductively. Starting from $A = \Omega$, we consider all possible axis-aligned dyadic partitions of A, and divide A by randomly drawing one of these possible ways of partitioning from a uniform distribution. Specifically, suppose there are a total of $N(A)$

dimensions that we can divide A along. We draw $J(A) \in \{1, 2, \ldots, N(A)\}$ such that $P(J(A) = j) = \lambda_j(A) = 1/N(A)$ for all $j = 1, 2, \ldots, N(A)$. Then if $J(A) = j$, we divide A into two halves along the jth dimension to two children A_l^j and A_r^j. Applying this random partitioning on each child node and proceed iteratively in this manner generates the bifurcating partition sequence $\mathcal{A}^{(\infty)}$. With this hyperprior on $\mathcal{A}^{(\infty)}$, we arrive at the following hierarchical model:

$$\mathcal{A}^{(\infty)} \sim \mathrm{RP}, \qquad \mathcal{C} \,|\, \boldsymbol{\phi}, \mathcal{A}^{(\infty)} \sim \mathrm{MT}(\boldsymbol{\gamma}), \qquad \nu(A) \,|\, \boldsymbol{\phi}, \mathcal{C}, \mathcal{A}^{(\infty)} \sim \sum_{i=1}^{I} F_i \cdot \mathbf{1}_{C(A)=i},$$
$$\theta(A) \,|\, \boldsymbol{\phi}, \boldsymbol{\nu}, \mathcal{C}, \mathcal{A}^{(\infty)} \sim \mathrm{Beta}(\theta_0(A)\nu(A), (1 - \theta_0(A))\nu(A)),$$

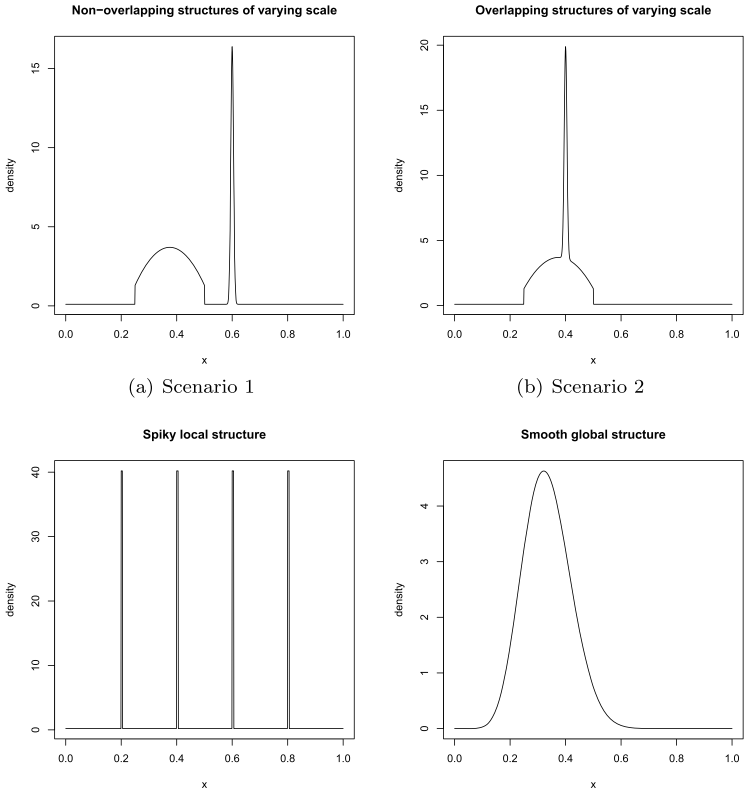
for all $A \in \mathcal{A}^{(\infty)}$. One can show that the four properties enjoyed by the Markov-APT remain true for this new model. Moreover, Bayesian inference for this model can still proceed in a similar fashion as before—through direct posterior sampling from three pieces in the following order $\pi(\mathcal{C}, \mathcal{A}^{(\infty)} \,|\, \boldsymbol{\phi}, \boldsymbol{x})$, $\pi(\boldsymbol{\nu} \,|\, \boldsymbol{\phi}, \mathcal{C}, \mathcal{A}^{(\infty)}, \boldsymbol{x})$ and $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{\phi}, \boldsymbol{\nu}, \mathcal{C}, \mathcal{A}^{(\infty)}, \boldsymbol{x})$. The last two pieces are essentially the same as before with a fixed $\mathcal{A}^{(\infty)}$. The first piece, the joint posterior of $(\mathcal{C}, \mathcal{A}^{(\infty)})$, is an inductive generative procedure that can be directly sampled from using forward–backward recursion. Details are given in Supplementary Material S3 (Ma, 2016). We apply this model in the flow cytometry example.

3 Performance evaluation in density estimation

We next evaluate the performance of the Markov-APT in density estimation under four schematic simulation scenarios (Figure 4). Each scenario corresponds to an underlying density with a particular type of structure commonly encountered in real applications. In particular, in Scenarios 1 and 2 the underlying densities contain structures of varying scales (both spiky and smooth). In Scenario 3 the density contains only spiky structures and in Scenario 4 only smooth structures. In each scenario, the underlying density is supported on the interval [0,1]. Our proposed model does not require the support to be a bounded interval. This choice is to simplify the numerical evaluation of the L_1 loss as a performance measure. This causes no loss of generality because any density on $\mathbb{R}$ can be transformed onto [0,1] after applying, say, a CDF transformation. For each scenario, we simulate data sets of six different sample sizes—125, 250, 500, 750, 1,000, and 1,250.

We compare the performance of Markov-APT to that of three other models—the PT, the OPT, and the DPM of normals (Escobar and West, 1995). For the PT, we use the standard specification with $\theta(A) \sim \mathrm{Beta}(ck^2, ck^2)$ for all $A \in \mathcal{A}^k$, with the global smoothing parameter c chosen by empirical Bayes through MMLE. For the OPT, $\theta(A) \sim \mathrm{Beta}(0.5, 0.5)$ for all $A \in \mathcal{A}^{(\infty)}$ as recommended in Wong and Ma (2010), and set the stopping probability on each A to a common value ρ_0, chosen through empirical Bayes (MMLE). We fit the DPM in `R` using the library `DPpackage` (Jara, 2007; Jara et al., 2011). Details on the specification of the DPM are in Supplementary Material S4 (Ma, 2016).

For each method, we use the PPD, denoted by $\hat{f}$, as an estimator. To measure performance, we adopt the L_1 loss, i.e. the L_1 distance between $\hat{f}$ and the true density



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

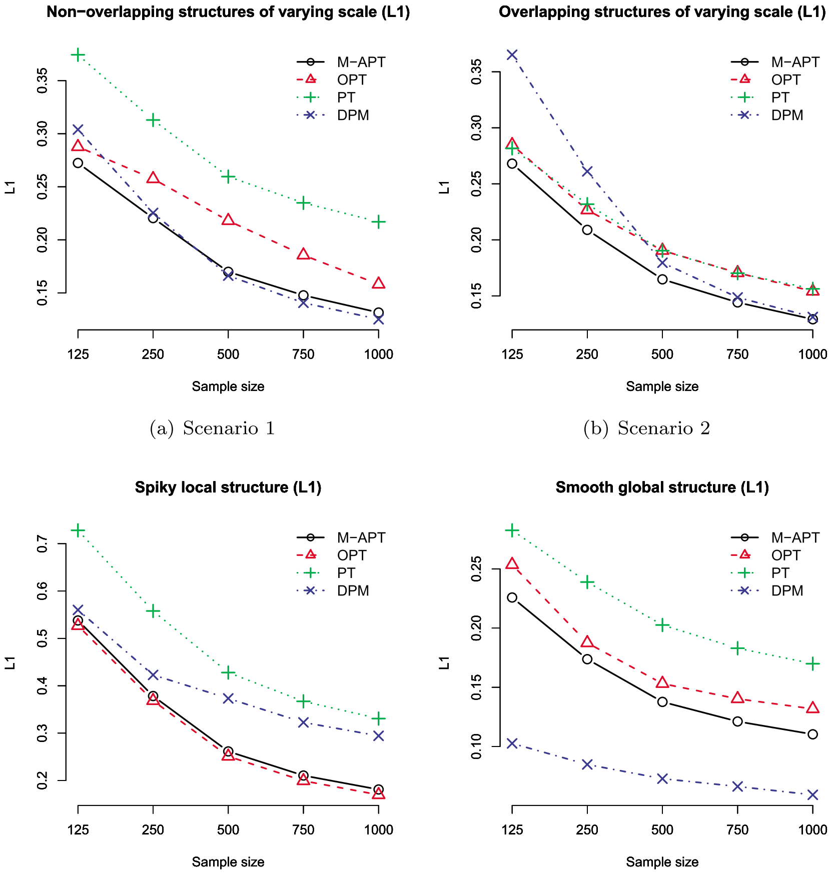
(d) Scenario 4

Figure 4: True densities of the four simulation scenarios.

f_0, $||\hat{f} - f_0||_1 = \int |\hat{f} - f_0| d\mu$. For each simulation scenario and sample size, we generate 200 data sets, with $\hat{f}^{(k)}$ being the estimate for the kth data set. We numerically calculate $||\hat{f}^{(k)} - f_0||_1$ using Riemann integral for all four methods. To make a comparison, for each of the competitors—PT, OPT, and DPM—we compute the percentage difference between its L_1 loss and that of the Markov-APT:

$$\frac{||\hat{f}^{(k)}_{\text{Competitor}} - f_0||_1 - ||\hat{f}^{(k)}_{\text{Markov-APT}} - f_0||_1}{||\hat{f}^{(k)}_{\text{Markov-APT}} - f_0||_1} \times 100\%.$$

A positive value indicates outperformance of the Markov-APT over the competitor, with larger values indicating more gain of the Markov-APT. Computing this metric for each simulation allows us to evaluate both the average performance gain and the variability in the improvement across repeated experiments. In addition, we also estimate the average L_1 loss, i.e. the L_1 risk, $R_{L_1}(f_0, \hat{f}) = \mathrm{E}_{f_0}||\hat{f} - f_0||_1$ for each method under each simulation setting using the Monte Carlo average $\widehat{R}_{L_1}(f_0, \hat{f}) = \frac{1}{200}\sum_{k=1}^{200} ||\hat{f}^{(k)} - f_0||_1$.



(a) Scenario 1

(b) Scenario 2

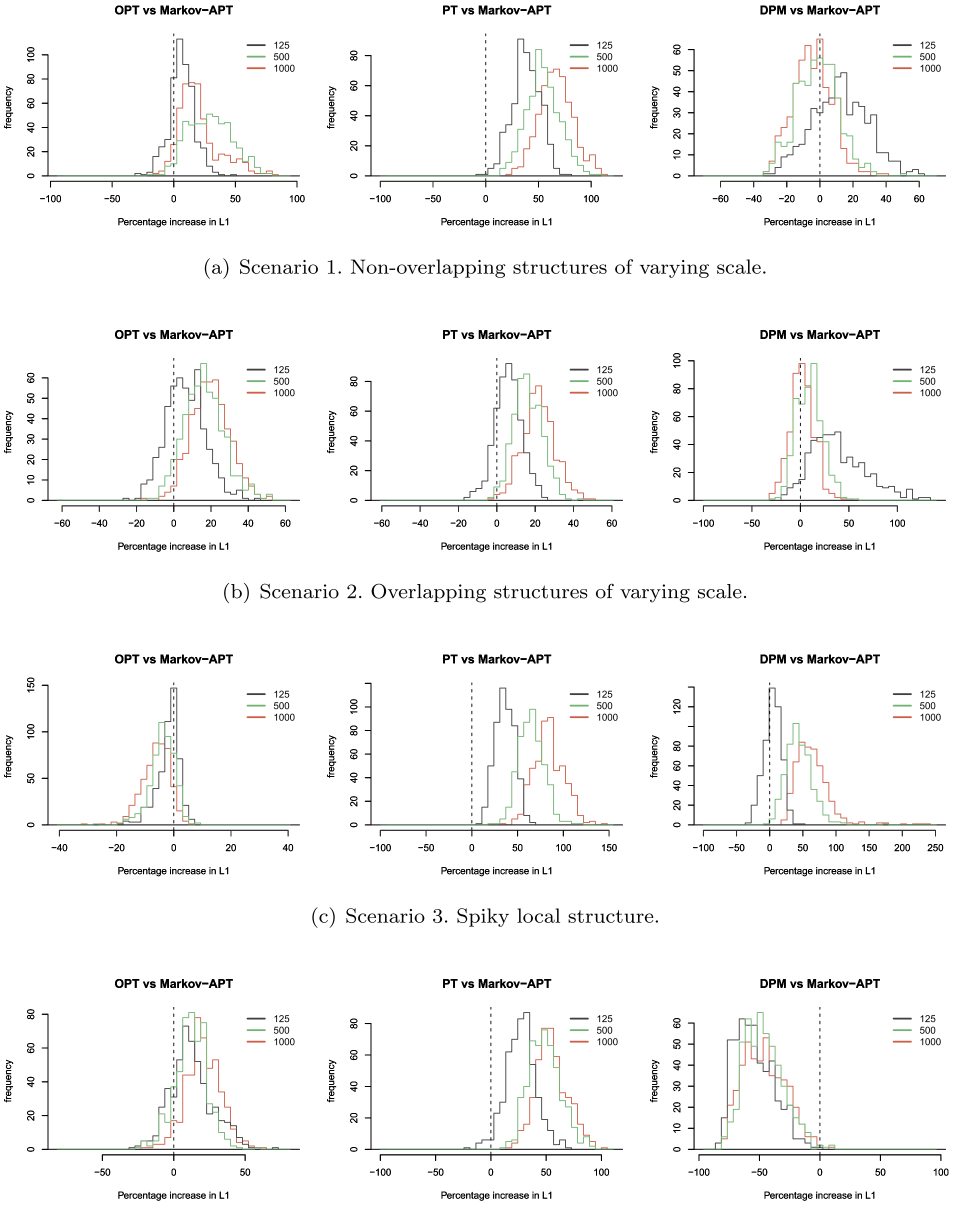
(c) Scenario 3

(d) Scenario 4

Figure 5: Estimated L_1 risk by sample size for the four simulation scenarios.

In all simulation settings, we adopt the prior specification recommended in Section 2.6 with an exponential transition kernel, and use empirical Bayes to set the tuning parameters (I, β). The range of tuning parameter values over which we maximize the marginal likelihood is $\{2, 3, \ldots, 11\} \times [0, 2]$. The OPT also involves a tuning parameter ρ_0, the prior "stopping" probability (Wong and Ma, 2010), which we set by empirical Bayes using MMLE over $[0, 1]$. We implement Markov-APT, PT, and OPT up to the 12th level in the partition tree. Deeper partitioning results in little numerical difference.

Figure 5 presents the L_1 risks for all methods and scenarios versus sample size. Figure 6 presents histograms of the percentage increase in L_1 loss for the three competitor methods relative to Markov-APT for each simulation setting. For easier comparison, we overlay the histograms for three different sample sizes—125 (small), 500 (medium), and 1,000 (large)—to show how the relative performance changes for different sample sizes. (We only show the three sample sizes in this figure rather than all six sample



(a) Scenario 1. Non-overlapping structures of varying scale.

(b) Scenario 2. Overlapping structures of varying scale.

(c) Scenario 3. Spiky local structure.

(d) Scenario 4. Smooth global structure.

Figure 6: Histograms of percentage increase (or lag) in L_1 risk in 500 simulations for three methods compared to Markov-APT over three sample sizes—125, 500, and 1,000.

sizes because overlaying six histograms makes the plot illegible while using the three sample sizes is sufficient to convey the main finding.) Finally, to help understand why each method performs well or poorly in each scenario, in Figure 7 we plot a typical PPD for each model under each scenario for a sample size that well differentiates the performance of the methods. We discuss the results for each scenario below.

- *Scenario 1: Non-overlapping structures of different scales.* The true distribution is

$$0.1\,\mathrm{U}(0,1) + 0.3\,\mathrm{U}(0.25,0.5) + 0.4\,\mathrm{Beta}_{(0.25,0.5)}(2,2) + 0.2\,\mathrm{Beta}(6000,4000).$$

See Figure 4(a) for the true density. This is the scenario given earlier in Example 1. The underlying density has two bumps of different scales—one large and the other small—that are not overlapping with each other. This is a favorable scenario for kernel mixture methods such as the DPM. By allowing the local kernel to have varying variance, the DPM is also able to adapt to the different scales of the two bumps. Thus one would expect the DPM to perform well.

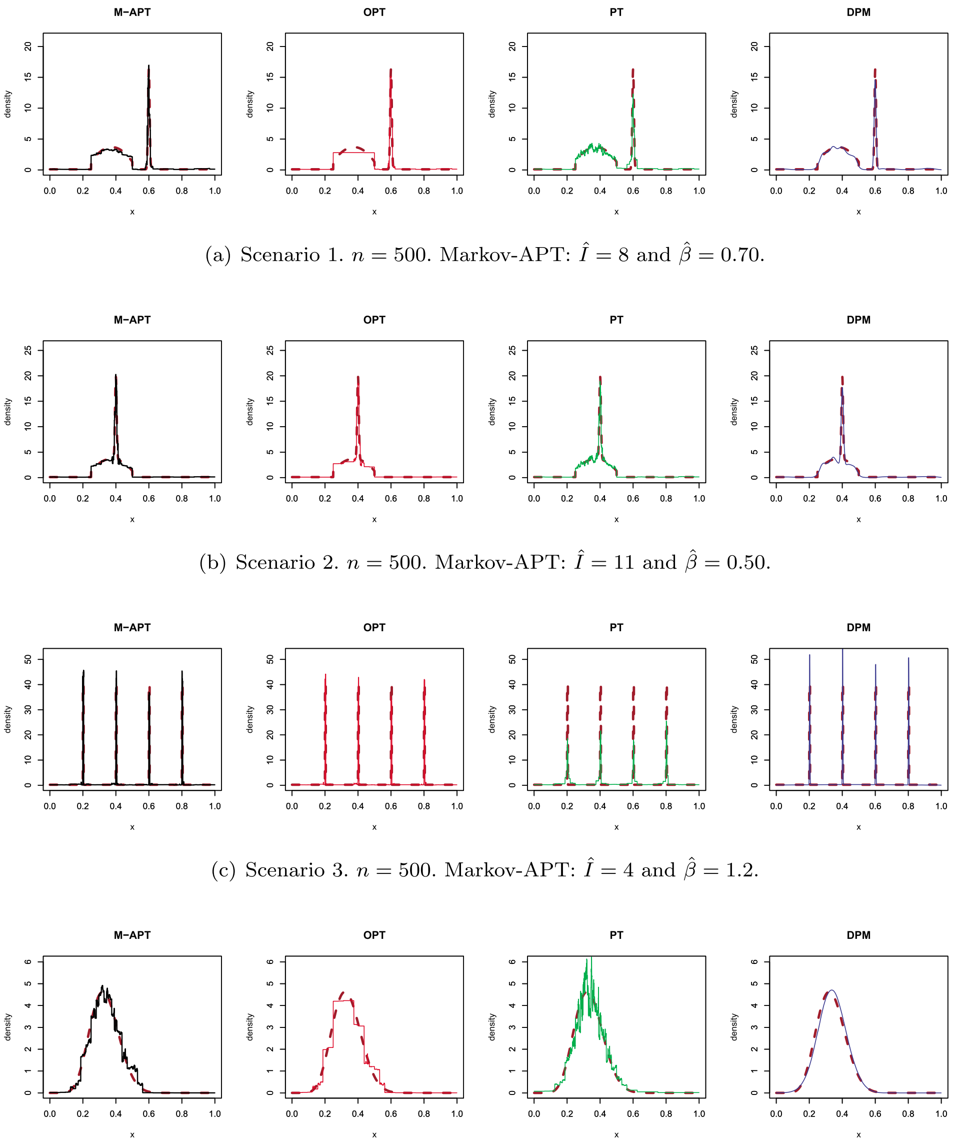
From Figure 5(a) and Figure 6(a), we see that the Markov-APT achieves better performance than the DPM at sample size 125, and comparable performance at larger sample sizes. The performance gain of Markov-APT over the other PT-type models—OPT and PT—is substantial at all sample sizes. Figure 7(a) shows that the OPT substantially oversmooths the large-scale feature while capturing the small-scale feature well. The reason is that OPT only allows two extremes of shrinkage—no shrinkage or complete shrinkage. While no shrinkage state for the small-scale feature in this example, for the large-scale feature, the complete shrinkage, which is the more appropriate of the two states, results in considerable oversmoothing. The PT oversmooths the small-scale feature while undersmooths at high-resolutions within the large-scale feature.

- *Scenario 2: Overlapping structures of different scales.* The true distribution is

$$0.1\,\mathrm{U}(0,1) + 0.3\,\mathrm{U}(0.25,0.5) + 0.4\,\mathrm{Beta}_{(0.25,0.5)}(2,2) + 0.2\,\mathrm{Beta}(4000,6000).$$

See Figure 4(b) for the true density. This case is similar to the previous except that now the spiky local structure lies inside the smooth large-scale structure. From Figure 5(b), we see that the performance of the Markov-APT is essentially unchanged from the case where the structures are non-overlapping. In contrast, Figure 5(b) and Figure 6(b) show that the other methods all perform quite differently in this scenario. In particular, there is a substantial decay in performance for the DPM at smaller sample sizes compared to the case with non-overlapping structures, whereas the OPT and PT perform better for the current scenario. The dramatic change in the performance of PT compared to that in Scenario 2 illustrates the importance of adaptivity in achieving robust inference.

The sudden improvement in PT's performance is readily explained in Figure 7(b). The estimation error from PT in this and the previous scenario comes from two sources—the under-smoothing (i.e. under-shrinkage) in the large-scale smooth structure and over-smoothing (i.e. over-shrinkage) in the local spiky structure. By moving the spiky structure into the smooth structure, the error that comes from the under-smoothing in the large-scale structure is reduced because the smooth portion of the large-scale structure now accounts for a smaller proportion of the total probability mass, while the error



(a) Scenario 1. $n = 500$. Markov-APT: $\hat{I} = 8$ and $\hat{\beta} = 0.70$.

(b) Scenario 2. $n = 500$. Markov-APT: $\hat{I} = 11$ and $\hat{\beta} = 0.50$.

(c) Scenario 3. $n = 500$. Markov-APT: $\hat{I} = 4$ and $\hat{\beta} = 1.2$.

(d) Scenario 4. $n = 500$. Markov-APT: $\hat{I} = 11$ and $\hat{\beta} = 0.60$.

Figure 7: Typical PPDs (solid) of four methods and true density (dashed) for the four simulation scenarios. Sample sizes and tuning parameters for Markov-APT chosen by MMLE are given.

that comes from over-smoothing the local structure is also reduced because now the local structure corresponds to more probability mass and hence more data. Note also in Figure 5(b) and Figure 6(b) that the relative performance gain of Markov-APT over

PT increases with the sample size.

- *Scenario 3: Spiky local structures.* In this case, the true distribution is

$$0.2\,\mathrm{U}(0,1)+0.2\,\mathrm{U}(0.2,0.205)+0.2\,\mathrm{U}(0.4,0.405)+0.2\,\mathrm{U}(0.6,0.605)+0.2\,\mathrm{U}(0.8,0.805).$$

See Figure 4(c) for the true density. This represents the case when the underlying distribution has a few spiky structures in the midst of a flat background. The key is to effectively determine the size (or height) of those spikes and pin down their boundaries.

As shown in Figure 5(c) and Figure 6(c), the Markov-APT and the OPT perform comparably and substantially better than the DPM, especially for medium and large sample sizes. It may first appear surprising that the PT, being a multi-resolution approach, performs the worst among all in capturing spiky structures, as multi-scale methods are known to be effective in characterizing such features. But this can be expected because the PT is unable to amply capture the height of the spikes due to over shrinkage at high-resolutions. In contrast, the amount of shrinkage under the Markov-APT is adaptive and thus automatically adjusts to low levels in and around the spikes. The OPT, which only allows no shrinkage or complete shrinkage, also performs very well. Because the underlying density consists of step functions, the only appropriate shrinkage levels are indeed no shrinkage and complete shrinkage. Therefore the OPT is in a sense an "oracle" in this case and one should expect it to perform the best. It is reassuring that the Markov-APT, while allowing more flexible shrinkage, did not lose much efficiency relative to the "oracle".

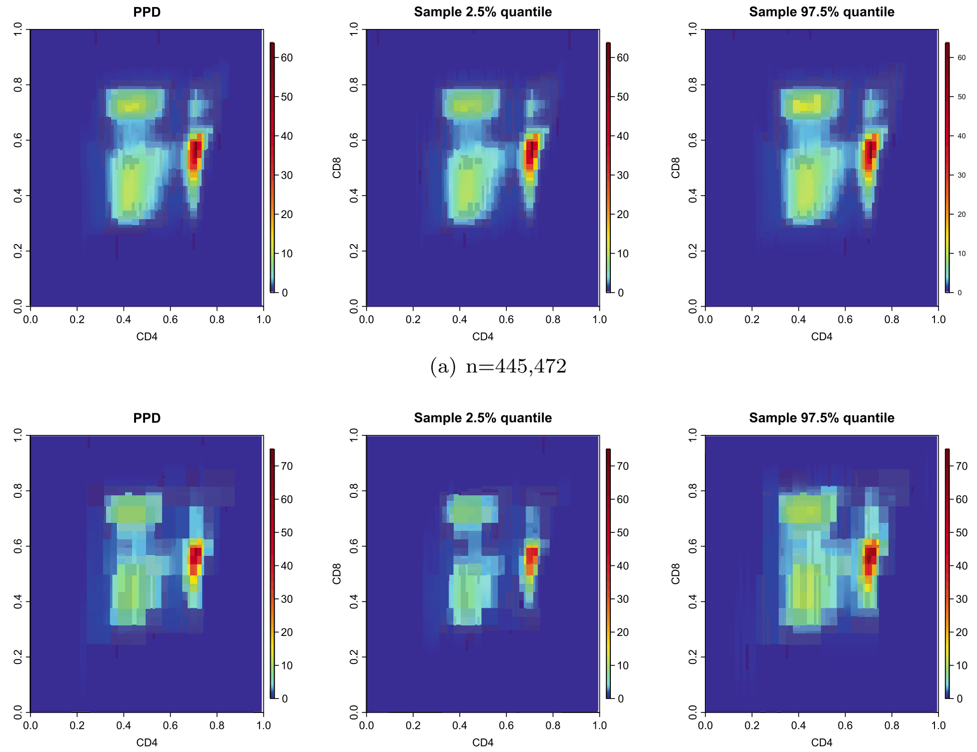
From the PPDs in Figure 7(c) we see that the DPM tends to overestimate the height of the spikes—this is likely because in order to characterize the sharp boundaries the DPM needs to "squeeze" the mixture component to have very thin tails, and thus making the mode of the mixture component much taller than the truth.

- *Scenario 4: Globally smooth structure.* The true distribution is $\mathrm{Beta}(10,20)$.

See Figure 4(d) for the true density, which is an approximately Gaussian, smooth distribution. DPM with a Gaussian kernel is essentially the true model for this scenario, and unsurprisingly performs the best. Figure 5(d) and Figure 6(d) show that for all sample sizes, the L_1 loss is on average about 50% smaller under the DPM vs that under the Markov-APT. Among the multi-resolution methods, the Markov-APT substantially outperforms the OPT and the PT, and the performance gain widens for larger sample sizes.

4 Case study: density estimation in flow cytometry

In a flow cytometry experiment, a large number (typically thousands to millions) of cells are measured in terms of p parameters or markers. Each cell is thus an observation in $\mathbb{R}^p$ and the density of the underlying cell distribution can be used for automatically classifying the cells into various subtypes (Malek et al., 2015). Traditionally such clas-



(a) n=445,472

(b) n=5,000

Figure 8: The PPD of Markov-APT (left column) and the pointwise central 95% credible band (middle and right columns), estimated from 1,000 posterior samples, for the CD4/CD8 two-marker density for two sample sizes 455,472 (top) and 5,000 (bottom).

sification, so-called "gating", is done manually—human experts look at the marginal distributions of the data set over (typically pairs of) signature markers, and determine the grouping and clustering of the cells based on expert knowledge. This can be very time-consuming especially in cases with relatively large numbers of parameters. More recently, automatic gating methods based on estimated marginal density functions of the cell distribution have been proposed (Malek et al., 2015). Fast and accurate automatic gating under this approach hinges on the availability of high-performance density estimators that enjoy both statistical and computational efficiency in handling large numbers of observations. The density of the cells is often multi-modal and contains structures of different scales.

We apply the Markov-APT with adaptive partitioning as described in Section 2.7 (with maximum level 11) to a flow cytometry data set, consisting of $n = 455{,}472$ cells each with $p = 14$ marker measurements. The computational efficiency of our method allows us to very quickly estimate all 14 one-marker, 91 two-marker, and 364 three-marker marginal densities on a single desktop CPU core, each in a mat-

	# post. samples	$n = 1{,}000$	$n = 10{,}000$	$n = 100{,}000$	$n = 455{,}472$
1D	100	0.16(0.02)	0.24(0.02)	0.42(0.03)	0.82(0.04)
	500	0.7(0.1)	1.1(0.1)	1.6(0.1)	2.3(0.2)
	1,000	1.4(0.2)	2.1(0.2)	3.1(0.3)	4.2(0.5)
2D	100	0.5(0.1)	1.0(0.2)	2.0(0.2)	5.0(0.2)
	500	1.9(0.2)	3.2(0.5)	5.4(0.8)	9.3(0.9)
	1,000	3.5(0.4)	6.1(0.8)	9.8(1.4)	14.9(1.7)
3D	100	1.6(0.3)	3.1(0.6)	7.7(0.7)	22.8(0.8)
	500	3.4(0.5)	6.2(1.2)	12.1(1.5)	28.0(1.7)
	1,000	5.7(0.9)	10(1.9)	17.3(2.6)	34.6(2.8)

Table 1: Average computing time in seconds (standard error in parentheses) on a single 3.6 GHz Intel$^{\circledR}$ Core$^{\mathrm{TM}}$-i7 desktop core to draw posterior samples over four sample sizes using Markov-APT for 1D, 2D, and 3D marginal densities in the flow cytometry data.

ter of seconds (Table 1). In Figure 8(a) we plot as an example the PPD for a two-marker (CD4 and CD8) density, along with the pointwise 2.5% and 97.5% quantiles of 1,000 posterior draws. The large sample size allows us to pin down precisely the underlying density except in the tails, where most of the variability in the posterior variability is and where often is the biologically interesting region in identifying rare cell types. To further illustrate the nature of the posterior estimate from Markov-APT, we also present the corresponding PPD and quantiles for a randomly drawn subsample of 5,000 cells in Figure 8(b) where the posterior uncertainty is more apparent.

To investigate the computational properties of our method, we repeat our analysis at a range of different sample sizes, obtained through down-sampling the original data, along with different numbers of posterior draws. Table 1 presents the average computing time along with the standard error for all one-, two-, and three-marker densities. The computation is done on a single Intel$^{\circledR}$ Core$^{\mathrm{TM}}$-i7 3820 3.6 GHz CPU core without parallelization.

5 Concluding remarks

We have showed that inference under PT-type models can be understood from a shrinkage perspective, and have introduced a hierarchical Bayesian approach to incorporating multi-scale adaptive shrinkage into such models. This development maintains the analytic tractability and computational efficiency of PT-type models while substantially improving statistical performance. Through incorporating a combination of numerical integration and conditional sampling, inference under the model avoids MCMC sampling. We believe that due to their computational efficiency and principled probabilistic nature, PT-type methods have tremendous potential for applications where flexible nonparametric modeling is desired, but computational speed is critical and/or sample sizes are huge. Additional effort is worthwhile to study the theory and to further improve the statistical and computational performance of this class of methods.

Supplementary Material

Supplementary Material for "Adaptive Shrinkage in Pólya Tree Type Models" (DOI: 10.1214/16-BA1021SUPP; .pdf).

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(453): 174–184. MR1952730. doi: http://dx.doi.org/10.1198/016214501750333045. 779, 792

Chen, Y. and Hanson, T. E. (2014). "Bayesian nonparametric k-sample tests for censored and uncensored data." *Computational Statistics & Data Analysis*, 71(0): 335–346. http://www.sciencedirect.com/science/article/pii/S0167947312003945. MR3131974. doi: http://dx.doi.org/10.1016/j.csda.2012.11.003. 779

Crouse, M. S., Nowak, R. D., and Baraniuk, R. G. (1998). "Wavelet-based statistical signal processing using hidden Markov models." *Signal Processing, IEEE Transactions on*, 46(4): 886–902. http://dx.doi.org/10.1109/78.668544. MR1665651. doi: http://dx.doi.org/10.1109/78.668544. 785

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. MR1340510. 781, 794

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *Annals of Statistics*, 1: 209–230. MR0350949. 779

Ferguson, T. S. (1974). "Prior distributions on spaces of probability measures." *Annals of Statistics*, 2: 615–629. MR0438568. 779

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis*, Third Edition. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. https://encrypted.google.com/books?id=ZXL6AQAAQBAJ. MR3235677. 787, 792

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics. New York: Springer-Verlag. MR1992245. 784

Hanson, T. and Johnson, W. O. (2002). "Modeling regression error with a mixture of Pólya trees." *Journal of the American Statistical Association*, 97(460): pp. 1020–1033. http://www.jstor.org/stable/3085827. MR1951256. doi: http://dx.doi.org/10.1198/016214502388618843. 779, 780

Hanson, T. E. (2006). "Inference for mixtures of finite Pólya tree models." *Journal of the American Statistical Association*, 101(476): 1548–1565. MR2279479. doi: http://dx.doi.org/10.1198/016214506000000384. 779, 780, 786, 787, 792

Holmes, C. C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015). "Two-sample Bayesian nonparametric hypothesis testing." *Bayesian Analysis*, 10(2): 297–320. MR3420884. doi: http://dx.doi.org/10.1214/14-BA914. 779, 780

Jara, A. (2007). "Applied Bayesian non- and semi-parametric inference using DPpackage." *R News*, 7(3): 17–26. http://CRAN.R-project.org/doc/Rnews/. 794

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(5): 1–30. http://www.jstatsoft.org/v40/i05/. 794

Jara, A. and Hanson, T. E. (2011). "A class of mixtures of dependent tail-free processes." *Biometrika*, 98(3): 553–566. MR2836406. doi: http://dx.doi.org/10.1093/biomet/asq082. 779

Jara, A., Hanson, T. E., and Lesaffre, E. (2009). "Robustifying generalized linear mixed models using a new class of mixtures of multivariate Pólya trees." *Journal of Computational and Graphical Statistics*, 18(4): 838–860. http://amstat.tandfonline.com/doi/abs/10.1198/jcgs.2009.07062. MR2598032. doi: http://dx.doi.org/10.1198/jcgs.2009.07062. 779

Kraft, C. H. (1964). "A class of distribution function processes which have derivatives." *Journal of Applied Probability*, 1: 385–388. MR0171296. 785, 790

Lavine, M. (1992). "Some aspects of Pólya tree distributions for statistical modelling." *Annals of Statistics*, 20(3): 1222–1235. MR1186248. doi: http://dx.doi.org/10.1214/aos/1176348767. 779, 780, 785, 790

Lavine, M. (1994). "More aspects of Pólya tree distributions for statistical modelling." *Annals of Statistics*, 22(3): 1161–1176. MR1311970. doi: http://dx.doi.org/10.1214/aos/1176325623. 779

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer. MR1842342. 788

Ma, L. (2013). "Adaptive testing of conditional association through recursive mixture modeling." *Journal of the American Statistical Association*, 108(504): 1493–1505. MR3174724. doi: http://dx.doi.org/10.1080/01621459.2013.838899. 793

Ma, L. (2016). "Supplementary material for "Adaptive shrinkage in Pólya tree type models"." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1021SUPP. 779, 788, 794

Ma, L. and Wong, W. H. (2011). "Coupling optional Pólya trees and the two sample problem." *Journal of the American Statistical Association*, 106(496): 1553–1565. MR2896856. doi: http://dx.doi.org/10.1198/jasa.2011.tm10003. 779

Malek, M., Taghiyar, M. J., Chong, L., Finak, G., Gottardo, R., and Brinkman, R. R. (2015). "flowDensity: Reproducing manual gating of flow cytometry data by automated density-based cell population identification." *Bioinformatics*, 31(4): 606–607. doi: http://dx.doi.org/10.1093/bioinformatics/btu677. 800, 801

Mauldin, D. R., Sudderth, W. D., and Williams, S. C. (1992). "Pólya trees and random distributions." *Annals of Statistics*, 20(3): 1203–1221. MR1186247. doi: http://dx.doi.org/10.1214/aos/1176348766. 779

Nieto-Barajas, L. E. and Müller, P. (2012). "Rubbery Pólya tree." *Scandinavian Journal of Statistics*, 39(1): 166–184. MR2896797. doi: http://dx.doi.org/10.1111/j.1467-9469.2011.00761.x. 779, 780

Paddock, S. M., Ruggeri, F., Lavine, M., and West, M. (2003). "Randomized Pólya tree models for nonparametric Bayesian inference." *Statistica Sinica*, 13(2): 443–460. MR1977736. 779

Tansey, W., Athey, A., Reinhart, A., and Scott, J. G. (2015). "Multiscale spatial density smoothing: an application to large-scale radiological survey and anomaly detection." arXiv:1507.07271. 779

Trippa, L., Müller, P., and Johnson, W. (2011). "The multivariate beta process and an extension of the Pólya tree model." *Biometrika*, 98(1): 17–34. http://ideas.repec.org/a/oup/biomet/v98y2011i1p17-34.html. MR2804207. doi: http://dx.doi.org/10.1093/biomet/asq072. 779

Walker, S. G., Damien, P., Laud, P. W., and Smith, A. F. M. (1999). "Bayesian nonparametric inference for random distributions and related functions." *Journal of the Royal Statistical Society: Series B*, 61(3): 485–527. MR1707858. doi: http://dx.doi.org/10.1111/1467-9868.00190. 779, 780

Watson, J., Nieto-Barajas, L., and Holmes, C. (2014). "Characterising variation of nonparametric random probability measures using the Kullback–Leibler divergence." arXiv:1411.6578. 787

Wong, W. H. and Ma, L. (2010). "Optional Pólya tree and Bayesian inference." *Annals of Statistics*, 38(3): 1433–1459. http://projecteuclid.org/euclid.aos/1268056622. MR2662348. doi: http://dx.doi.org/10.1214/09-AOS755. 779, 781, 790, 793, 794, 796

Acknowledgments

The author would like to thank the referees and editors for very helpful comments and suggestions. The flow cytometry data set was provided by EQAPOL (HHSN272201000045C), an NIH/NIAID/DAIDS-sponsored, international resource that supports the development, implementation, and oversight of quality assurance programs (Sanchez PMC4138253).