Rubbery Polya Tree

Luis E. Nieto-Barajas[1,2] and Peter Mueller[1]

[1]*Department of Biostatistics, M.D. Anderson Cancer Center, USA*

[2]*Department of Statistics, ITAM, Mexico*

Abstract

Polya trees (PT) are random probability measures (RPM) which can assign probability one to the set of continuous distributions for certain specifications of the hyperparameters. This feature distinguishes the PT from the popular Dirichlet process (DP) model which assigns probability one to the set of discrete distributions. However, the PT is not nearly as widely used as the DP prior. Probably the main reason is an awkward dependence of posterior inference on the choice of the partitioning subsets in the definition of the PT. We propose a generalization of the PT prior that mitigates this undesirable dependence on the partition structure, by allowing the branching probabilities to be dependent within the same level. The proposed new process is not a PT anymore. However, it is still a tail free process and many of the prior properties remain the same as those for the PT.

Key words: Bayes nonparametrics, Markov beta process, partition model, Polya tree, random probability measure, tail free distribution.

1 Introduction

Since Ferguson (1973) introduced the Dirichlet process (DP) prior model it has become the by far most popular model in nonparametric Bayesian inference. Nonparametric Bayesian inference implements statistical inference with minimal assumptions similar to classical nonparametric methods. The DP prior is used as a prior model for unknown distributions. It allows inference for unknown distributions without the restriction to parametric families. See for example Walker et al. (1999) for an overview of nonparametric Bayesian methods.

However, a critical limitation of the DP is the restriction to the space of discrete distributions, complicating the use for applications with continuous data. Antoniak (1974)

considered mixtures of DP models by defining mixtures with respect to the hyperparameters of the centering measure (mixture of DP). In a different approach towards overcoming discreteness, Lo (1984) and Escobar and West (1995) used the DP as a mixing distribution to convolute a continuous (usually normal) kernel and introduced the DP mixture model (DPM). Since then many authors have developed applications in a variety of fields. Examples are Kottas et al. (2005) or Do et al. (2005), among many others.

In contrast, the Polya tree (PT) which could be arguably considered the simplest RPM for continuous data, has not seen much use since the early papers by Lavine (1992, 1994) who studied the properties of the Polya tree. Perhaps the main reason for the limited use of the model is the dependence of inference on the arbitrarily chosen partitioning subsets which are required in the definition of the PT prior. The density of the posterior estimated RPM is discontinuous at the boundaries of the partitioning subsets. To overcome this awkward dependence on the partitions, Lavine (1992, 1994), Hanson and Johnson (2002) and Hanson (2006) have considered a mixture of Polya trees by mixing over the centering measure that defines the tree (and thus the partitions). With the same objective, Paddock et al. (2003) considered a randomized Polya tree allowing the partitions to be jittered. More recent applications of PT models include, among others, Branscum et al. (2008) for inference with ROC curves, Branscum and Hanson (2008) for meta-analysis, Hanson and Johnson (2002) for regression residuals, Li et al. (2008) for genetic association studies, Hanson and Yang (2007) for survival data, Zhang et al. (2009) for survival data with longitudinal covariates, Yang et al. (2010) for repeated measurement data, Zhao and Hanson (2010) for spatially dependent survival data, Paddock (2002) for multiple imputation in missing data problems, and Jara et al. (2009) for multivariate PTs in mixed effects models. But the model is nowhere near as commonly used as the DP prior.

Polya tree priors are members of a more general class of tail free processes (Freedman, 1963; Fabius, 1964). In words, PTs are essentially random histograms with the bins determined by recursive binary splits of the sample space. Figure 1 illustrates the nested binary partitions created by these splits. Starting with the sample space B, a tree of nested parti-

tions is defined by $B_{\epsilon_1 \cdots \epsilon_m} = B_{\epsilon_1 \cdots \epsilon_m 0} \cup B_{\epsilon_1 \cdots \epsilon_m 1}$. The partitioning sets for the partition at leve m of the tree are indexed by binary sequences $(\epsilon_1 \cdots \epsilon_m)$, i.e., $\pi_m = \{\epsilon_1 \cdots \epsilon_m; \quad \epsilon_j \in \{0,1\}\}$ is a partition of the sample space B.

For a formal definition, let $\Pi = \{\pi_m; m = 1, 2, \ldots\}$ be a tree of measurable partitions of $(\mathbb{R}, \mathcal{B})$; that is, let $\pi_1, \pi_2, \ldots$ be a sequence of measurable partitions such that π_{m+1} is a refinement of π_m for each $m = 1, 2, \ldots$, and $\cup_{m=1}^{\infty} \pi_m$ generates $\mathcal{B}$. Let $E = \cup_{m=1}^{\infty} \{0,1\}^m$ be the infinite set of binary sequences, such that, if $\epsilon = \epsilon_1 \cdots \epsilon_m \in E$ then B_ϵ defines a set at level m, i.e. $B_\epsilon \in \Pi_m$. Without loss of generality we assume binary partitions, i.e., a set $B_\epsilon \in \pi_m$ is partitioned as $B_\epsilon = B_{\epsilon 0} \cup B_{\epsilon 1}$ in π_{m+1}. Partitioning subsets in π_m are indexed by dyadic sequences $\epsilon = \epsilon_1 \cdots \epsilon_m$.

Definition 1 (Ferguson, 1974). A random probability measure P on $(\mathbb{R}, \mathcal{B})$ is said to have a *tailfree distribution* with respect to Π if there exists a family of non-negative random variables $\mathcal{Y} = \{Y_\epsilon; \epsilon \in E\}$ such that

i. The families

$\mathcal{Y}_1 = \{Y_0\}, \mathcal{Y}_2 = \{Y_{\epsilon_1 0}\}, \ldots, \mathcal{Y}_m = \{Y_{\epsilon_1 \cdots \epsilon_{m-1} 0}\}, \ldots$ are independent, and;

ii. For every $m = 1, 2, \ldots$ and every $\epsilon = \epsilon_1 \cdots \epsilon_m$,

$$\mathrm{P}(B_{\epsilon_1 \cdots \epsilon_m}) = \prod_{j=1}^{m} Y_{\epsilon_1 \cdots \epsilon_j},$$

where $Y_{\epsilon_1 \cdots \epsilon_{j-1} 1} = 1 - Y_{\epsilon_1 \cdots \epsilon_{j-1} 0}$.

In words, Y_ϵ are random branching probabilities $P(B_{\epsilon 0} \mid B_\epsilon)$ in the tree of nested partitions.

If we further assume that the random variables in $\mathcal{Y}_m$ are all independent with beta distributions then the random probability measure (RPM) P has a Polya tree distribution (Lavine, 1992).

We propose a generalization of the PT that reduces the undesirable sensitivity to the choice of Π. In order to reduce the impact of the partition on statistical inference, we allow

the random variables $\mathcal{Y}$ to be dependent within the same level m, but keeping the independence assumption between different levels. This defines a new RPM that still belongs to the class of tailfree processes, and thus inference will still depend on the choice of partitions. But the random probabilities $P(B_{\epsilon_1 \cdots \epsilon_m})$ of the partitioning subsets vary in a smoother fashion across the sets in each level of the partition tree. The only tailfree process invariant to the choice of partitions is the DP. To keep the new prior comparable to the original Polya tree we continue to use beta distributions as the marginal distributions for each Y_ϵ. This is achieved by considering a stochastic process with beta stationary distribution as prior for $\mathcal{Y}_m$. The construction of the process is defined in such a way that for a specific choice of the hyper-parameters we recover independence of the Y_ϵ's within the same level and thus having the regular Polya tree as particular case.

It is convenient to introduce a new notation for indexing of the partitioning subsets. This and the actual definition of the RPM is introduced in Section 2. The properties of the rubbery Polya tree are studied in Section 3. Posterior inference is discussed in Section 4. Section 5 includes some simulation studies and comparisons with the Polya tree. Finally Section 6 contains a discussion and concluding remarks.

Throughout we use $[x]$ and $[x \mid y]$ to generically indicate the distribution of a random variable x and the conditional distribution of x given y.

2 The Rubbery Polya Tree

2.1 The rPT model

As in the PT, the proposed prior relies on a binary partition tree of the sample space. For simplicity of exposition we consider $(\mathbb{R}, \mathcal{B})$ as our measurable space with $\mathbb{R}$ the real line and $\mathcal{B}$ the Borel sigma algebra of subsets of $\mathbb{R}$. The binary partition tree is denoted by $\Pi = \{B_{mj}\}$, where the index m denotes the level in the tree and j the location of the partitioning subset within the level, with $j = 1, \ldots, 2^m$ and $m = 1, 2, \ldots$. The sets at level 1 are denoted by (B_{11}, B_{12}); the partitioning subsets of B_{11} are (B_{21}, B_{22}), and $B_{12} = B_{23} \cup B_{24}$, such that $(B_{21}, B_{22}, B_{23}, B_{24})$ denote the sets at level 2. In general, at level m, the set B_{mj} splits into

$(B_{m+1,2j-1}, B_{m+1,2j})$, where $B_{m+1,2j-1} \cap B_{m+1,2j} = \emptyset$ and $B_{m+1,2j-1} \cup B_{m+1,2j} = B_{mj}$.

Like in the Polya tree, we associate random branching probabilities Y_{mj} with every set B_{mj}. Let P denote the RPM. We define $Y_{m+1,2j-1} = P(B_{m+1,2j-1} \mid B_{mj})$, and $Y_{m+1,2j} = 1 - Y_{m+1,2j-1} = P(B_{m+1,2j} \mid B_{mj})$. We denote by $\mathcal{Y} = \{Y_{mj}\}$ the set of random branching probabilities associated with the elements of Π. Instead of independent, as in the Polya tree, we assume them to be positively correlated within the same level. Specifically, at level m, the set of variables $\mathcal{Y}_m = \{Y_{m1}, Y_{m3}, \ldots, Y_{m,2^m-1}\}$ follow a Markov beta process, similar to the one introduced in Nieto-Barajas and Walker (2002). This process is defined through a latent process $\mathcal{Z}_m = \{Z_{mj}\}$ in such a way that we have the Markov structure

$$Y_{m1} \stackrel{Z_{m,1}}{\longrightarrow} Y_{m3} \stackrel{Z_{m,3}}{\longrightarrow} Y_{m5} \stackrel{Z_{m,5}}{\longrightarrow} \cdots \stackrel{Z_{m,2^m-3}}{\longrightarrow} Y_{m,2^m-1},$$

where

$$Y_{m1} \sim \text{Be}(\alpha_{m,1}, \alpha_{m,2}),$$

and for $j = 1, 2, \ldots, 2^{m-1} - 1$

$$Z_{m,2j-1}|Y_{m,2j-1} \sim \text{Bin}(\delta_{m,2j-1}, Y_{m,2j-1})$$

and

$$Y_{m,2j+1}|Z_{m,2j-1} \sim \text{Be}(\alpha_{m,2j+1} + Z_{m,2j-1}, \alpha_{m,2j+2} + \delta_{m,2j-1} - Z_{m,2j-1}).$$

Let $\mathcal{A}_m = \{\alpha_{mj},\ j = 1, \ldots, 2^m\}$ and $\mathcal{D}_m = \{\delta_{m,2j-1},\ j = 1, \ldots, 2^{m-1}\}$. We say that $(\mathcal{Y}_m, \mathcal{Z}_m) \sim \text{BeP}(\mathcal{A}_m, \mathcal{D}_m)$ is a Markov beta process with parameters $(\mathcal{A}_m, \mathcal{D}_m)$. The binomial sample size parameters δ_{mj} determine the degree of dependence between the Y_{mj}'s. In particular, if $\delta_{mj} = 0$ for all j then Z_{mj}=0 w.p.1. Therefore the Y_{mj}'s in the set $\mathcal{Y}_m$ become independent. Moreover, if $\alpha_{mj} = \alpha_m$ for all j then the process $\mathcal{Y}_m$ becomes strictly stationary with $Y_{m,2j+1} \sim \text{Be}(\alpha_m, \alpha_m)$ marginally. With these definitions we are now ready to define the proposed RPM.

Definition 2 Let $\mathcal{A}_m = \{\alpha_{mj},\ j = 1, \ldots, 2^m\}$ be non-negative real numbers, and $\mathcal{D}_m = \{\delta_{m,2j-1},\ j = 1, \ldots, 2^{m-1}\}$ be non-negative integers for each m, $m = 1, 2, \ldots$, and let $\mathcal{A} =$

$\bigcup \mathcal{A}_m$ and $\mathcal{D} = \bigcup \mathcal{D}_m$. A random probability measure P on $(\mathbb{R}, \mathcal{B})$ is said to have a *rubbery Polya tree prior* with parameters $(\Pi, \mathcal{A}, \mathcal{D})$, if for $m = 1, 2, \ldots$ there exist random variables $\mathcal{Y}_m = \{Y_{m,2j-1}\}$ and $\mathcal{Z}_m = \{Z_{m,2j-1}\}$ for $j = 1, \ldots, 2^{m-1}$, such that the following hold:

i. The sets of random variables $(\mathcal{Y}_1), (\mathcal{Y}_2, \mathcal{Z}_2), \ldots$ are independent across levels m.

ii. $\mathcal{Y}_1 = Y_{11} \sim \text{Be}(\alpha_{11}, \alpha_{12})$, and $(\mathcal{Y}_m, \mathcal{Z}_m) \sim \text{BeP}(\mathcal{A}_m, \mathcal{D}_m)$ for $m = 2, 3 \ldots$.

iii. For every $m = 1, 2, \ldots$ and every $j = 1, \ldots, 2^m$

$$\text{P}(B_{mj}) = \prod_{k=1}^{m} Y_{m-k+1, r(m-k+1)},$$

where $r(k-1) = \lceil r(k)/2 \rceil$ is a recursive decreasing formula, whose initial value is $r(m) = j$, that locates the set B_{mj} with its ancestors upwards in the tree. $\lceil \cdot \rceil$ denotes the ceiling function, and $Y_{m,2j} = 1 - Y_{m,2j-1}$ for $j = 1, \ldots, 2^{m-1}$.

The Y_{mj} are random branching probabilities. The Z_{mj} are latent (conditionally binomial) random variables that induce the desired dependence. We write $\text{P} \sim \text{rPT}(\Pi, \mathcal{A}, \mathcal{D})$.

Comparing Definitions 1 and 2, it is straightforward to verify that the rubbery Polya tree is a tailfree distribution with respect to the partition Π. We recall that tailfree processes are conjugate, in the sense that if P is tailfree with respect to Π then is $\text{P}|\,\mathbf{x}$ (Ferguson, 1974). Moreover, being a tailfree distribution is a condition for posterior consistency (Freedman, 1963; Fabius, 1964). A special case of the rubbery Polya tree is obtained when setting $\delta_{mj} = 0$ for all m and all j, reducing the prior to a Polya tree. In short, $\text{P} \sim \text{rPT}(\Pi, \mathcal{A}, \mathbf{0}) \equiv \text{PT}(\Pi, \mathcal{A})$.

The Markov beta process $[\mathcal{Y}_m, \mathcal{Z}_m]$ can be characterised by the conditional distribution $[\mathcal{Y}_m | \mathcal{Z}_m]$, and the marginal distribution of the latent process $[\mathcal{Z}_m]$. It can be shown that $Y_{m1}, \ldots, Y_{m,2^m-1}$ are conditionally independent given $\mathcal{Z}_m$ with beta distributions that only depend on the neighboring latent variables, that is,

$$Y_{m,2j+1} | Z_{m,2j-1}, Z_{m,2j+1} \sim \text{Be}(\alpha_{m,2j+1} + Z_{m,2j-1} + Z_{m,2j+1}\,, \\ \alpha_{m,2j+2} + \delta_{m,2j-1} - Z_{m,2j-1} + \delta_{m,2j+1} - Z_{m,2j+1}), \quad (1)$$

for $j = 0, 1, \ldots, 2^{m-1} - 1$, with $\delta_{m,-1} = 0$ and $Z_{m,-1} = 0$ w.p.1. Furthermore, the marginal distribution of the latent process $\mathcal{Z}_m$ is another Markov process with

$$Z_{m1} \sim \text{BeBin}(\delta_{m1}, \alpha_{m1}, \alpha_{m2}),$$

and beta-binomial transition distributions for $j = 1, \ldots, 2^{m-1} - 1$:

$$Z_{m,2j+1} | Z_{m,2j-1} \sim \text{BeBin}(\delta_{m,2j+1},\ \alpha_{m,2j+1} + Z_{m,2j-1},\ \alpha_{m,2j+2} + \delta_{m,2j-1} - Z_{m,2j-1}). \qquad (2)$$

The above characterisation of the Markov beta process implies that if $\text{P} \sim \text{rPT}(\Pi, \mathcal{A}, \mathcal{D})$, then conditionally on $\mathcal{Z}$, P is a Polya tree $\text{PT}(\Pi, \mathcal{A}^{\mathcal{Z}})$ with the parameters $\mathcal{A}^{\mathcal{Z}}$ being a function of Z_{mj}. Therefore the rubbery Polya tree can be seen as a particular $\mathcal{A}^{\mathcal{Z}}-$mixture of Polya trees with the mixing distribution determined by the law of $\mathcal{Z}$. In other words,

$$\text{P} \sim \int \text{PT}(\Pi, \mathcal{A}^{\mathcal{Z}}) \mathcal{L}(\text{d}\mathcal{Z})$$

where $\mathcal{A}^{\mathcal{Z}} = \{\alpha^z_{mj}\}$ such that $\alpha^z_{m,2j+1} = \alpha_{m,2j+1} + Z_{m,2j-1} + Z_{m,2j+1}$ and $\alpha^z_{m,2j+2} = \alpha_{m,2j+2} + \delta_{m,2j-1} - Z_{m,2j-1} + \delta_{m,2j+1} - Z_{m,2j+1}$, for $j = 0, 1, \ldots, 2^{m-1} - 1$ and $m = 1, 2, \ldots$. This is in contrast with mixtures of Polya trees defined in Lavine (1992) or Hanson and Johnson (2002) where the mixing is with respect to the partition Π rather than $\mathcal{A}$, and thus produce processes which are not tail free anymore. Even though the nature of the mixture is different, the general theory for mixtures, presented in Lavine (1992) and Lavine (1994), remains valid.

2.2 Finite tree

For practical purposes, inference with a tree-based prior can be simplified if we consider a finite or partially specified tree (Lavine, 1994; Hanson and Johnson, 2002). A *finite rubbery Polya tree* is defined by stopping the nested partitions at level M. We write $\text{P} \sim \text{rPT}(\Pi_M, \mathcal{A}_M, \mathcal{D}_M)$.

Lavine (1994) suggests to choose the level M, in a PT, to achieve a specified precision in the posterior predictive distribution. Alternatively, Hanson and Johnson (2002) recommend the rule of thumb $M \doteq \log_2 n$ such that the number of partitioning sets at level M is

approximately equal to the sample size n, trying to avoid empty sets in the updated tree. We will use the latter to determine the level M for defining a finite rubbery Polya tree.

Finally, for the sets in level M, we may consider P to be either uniform (on bounded sets) or to follow P_0 restricted to the set. It is worth noting that the latter option has to be used if it is desired to center the tree on P_0.

3 Centering the prior and further properties

For statistical inference, it is desirable to center the process around a given (usually parametric) distribution. Centering the process frees the researcher from the need to explicitly specify $\mathcal{A}$ and $\mathcal{D}$ element by element and is usually sufficient to represent available prior information. Walker et al. (1999) discuss several ways of centering a Polya tree process. The simplest and most used method (Hanson and Johnson, 2002) consists of matching the partition with the dyadic quantiles of the desired centering measure and keeping α_{mj} constant within each level m.

More explicitly, let P_0 be the desired centering measure on $\mathbb{R}$ with cdf $F_0(x)$. At each level m we take

$$B_{mj} = \left(F_0^{-1}\left(\frac{j-1}{2^m}\right), F_0^{-1}\left(\frac{j}{2^m}\right)\right], \tag{3}$$

for $j = 1, \ldots, 2^m$, with $F_0^{-1}(0) = -\infty$ and $F_0^{-1}(1) = \infty$. If we further take $\alpha_{mj} = \alpha_m$ for $j = 1, \ldots, 2^m$, and for any value of δ_{mj}, we get $\mathrm{E}\{\mathrm{P}(B_{mj})\} = \prod_{k=1}^{m} \mathrm{E}(Y_{m-k+1,r(m-k+1)}) = 2^{-m} = P_0(B_{mj})$.

The proof is straightforward. If we fix the parameters $\alpha_{mj} \equiv \alpha_m$ constant within each level m, then we are in the stationary setting of the Markov beta process (for any choice of $\mathcal{D}$). As mentioned in the previous section, $Y_{mj} \sim \mathrm{Be}(\alpha_m, \alpha_m)$ marginally for all m and all j, and therefore $\mathrm{E}(Y_{mj}) = 1/2$. This leads us to an interesting property of the proposed prior.

Proposition 1 *Let $P' \sim rPT(\Pi, \mathcal{A}, \mathcal{D})$ and $P^* \sim PT(\Pi, \mathcal{A})$ be a rubbery Polya tree and a Polya tree, respectively, with common partitions Π and common set of parameters $\mathcal{A}$. If for each level $m = 1, 2, \ldots$, we take $\alpha_{mj} = \alpha_m$ for $j = 1, \ldots, 2^m$, then for any measurable set*

$B_{mj} \in \Pi$,

$$P'(B_{mj}) \stackrel{d}{=} P^*(B_{mj}),$$

where $\stackrel{d}{=}$ *denotes equality in distribution.*

Proof. The result follows from part (iii) of Definition 2. Note that the product only involves one variable Y_{mj} from each level m and exploit stationarity to conclude the claim. $\diamond$

Proposition 1 says that the processes rPT and PT share the same marginal distribution. For the default choice of the $\mathcal{A}$ parameters being equal within each row, we can marginally generate the same measure for single sets with both processes. However, the joint distribution of the measure on two disjoint sets, say $(P(B_{mj}), P(B_{mj'}))$, for $j \neq j'$ is different under the rPT and PT. The following two corollaries provide more interesting properties of our prior.

Corollary 1 *Let* $P \sim rPT(\Pi, \mathcal{A}, \mathcal{D})$ *be a rubbery Polya tree with* $\alpha_{mj} = \alpha_m$*, for* $j = 1, \ldots, 2^m$*, and* $m = 1, 2, \ldots$*. All the conditions on the* $\mathcal{A}$ *parameters needed for a Polya tree to be a.s. continuous are inherited by the rubbery Polya tree. That is,* $\sum_{m=1}^{\infty} \alpha_m^{-1} < \infty$ *implies that* P *is absolutely continuous a.s.*

Proof. We write $f_m(x) = \{\prod_{k=1}^{m} Y_{m-k+1,r(m,x)}\} 2^m f_0(x)$, where $r(m,x) = j$ if $x \in B_{mj}$. Noting that this product involves only one Y_{mj} from each level m and when $\alpha_{mj} = \alpha_m$ then $Y_{mj} \sim \text{Be}(\alpha_m, \alpha_m)$ marginally. Therefore by taking the limit when $m \to \infty$ we can use the Theorem from Kraft (1964) and its corollary to obtain the result. $\diamond$

In particular, $\alpha_m = a/2^m$ defines an a.s. discrete measure, whereas $\alpha_m = am^2$ an a.s. continuous measure. Alternative choices of α_m can also be used to define continuity. For instance, Berger and Guglielmi (2001) considered $\alpha_m = am^3$, $a2^m$, $a4^m$, $a8^m$. In all these choices the parameter a controls the dispersion of P around P_0. A small value of a implies a large variance and thus a weak prior belief, i.e., a plays the role of a precision parameter.

Next we consider posterior consistency. Denote by $\mathcal{K}(f, g)$ the Kullback-Leibler divergence measure for densities f and g as $\mathcal{K}(f, g) = \int f(x) \log\{f(x)/g(x)\} \text{d}x$. Assume an i.i.d.

sample $X_i|\mathrm{P} \sim \mathrm{P}$, $i = 1, \ldots, n$ with an rPT prior $\mathrm{P} \sim \mathrm{rPT}(\Pi, \mathcal{A}, \mathcal{D})$. The following result states conditions for posterior consistency as $n \to \infty$ when data is generated from an assumed fixed model $f^\star$.

Corollary 2 *Let X_i, $i = 1, \ldots, n$ be i.i.d. observations from $f^\star$. We assume that $X_i \mid P \overset{iid}{\sim} P$, where $P \sim rPT(\Pi, \mathcal{A}, \mathcal{D})$ is a rubbery Polya tree centered at P_0 (with density f_0) with partitions as in (3) and with $\alpha_{mj} = \alpha_m$, for $j = 1, \ldots, 2^m$, and $m = 1, 2, \ldots$. If $\mathcal{K}(f^\star, f_0) < \infty$ and $\sum_{m=1}^{\infty} \alpha_m^{-1/2} < \infty$, then as $n \to \infty$ P achieves weak posterior consistency. Furthermore, if α_m increases at a rate at least as large as 8^m then P achieves posterior strong consistency.*

Proof. From Cororally 1, the softer condition $\sum_{m=1}^{\infty} \alpha_m^{-1} < \infty$ implies the existence of a density f of the RPM P, that is, $f(x) = \lim_{m\to\infty}\{\prod_{k=1}^{m} Y_{m-k+1,r(m,x)}\}$, where $r(m, x) = j$ if $x \in B_{mj}$. By the martingale convergence theorem, there also exists a collection of numbers $\{y_{m-k+1,r(m,x)}\} \in [0, 1]$ such that w.p.1 $f^\star(x) = \lim_{m\to\infty}\{\prod_{k=1}^{m} y_{m-k+1,r(m,x)}\}$. Now, since $Y_{mj} \sim \mathrm{Be}(\alpha_m, \alpha_m)$ marginally, then resorting to the proof of Theorem 3.1 in Ghosal et al. (1999) we obtain the weak consistency result. As for the strong consistency, we rely on the same derivations from Section 3.2 in Barron et al. (1999) to obtain the result. $\diamond$

In Barron et al. (1999)'s terminology, Corollary 2 ensures that the rPT is posterior consistent as long as the prior predictive density f_0 is not infinitely away from the true density f^*.

In the previous paragraphs we have presented several properties that are similar to the PT, however an important question remains unanswered. What is the impact of introducing dependence in the random variables within the levels of the tree? To respond to this question we study the correlation in the induced measures for two different sets in the same level, say $\mathrm{P}(B_{mj})$ and $\mathrm{P}(B_{mj'})$. For that we consider a finite tree $\mathrm{rPT}(\Pi_2, \mathcal{A}_2, \mathcal{D}_2)$ that consists of only $M = 2$ levels, say

$$\begin{array}{ccc} B_{11} & | & B_{12} \\ B_{21} \mid B_{22} & | & B_{23} \mid B_{24} \end{array} \tag{4}$$

For each B_{mj} there is a random variable Y_{mj}, which in the stationary case are defined by

1) $Y_{11} \sim \text{Be}(\alpha_1, \alpha_1)$, $Y_{12} = 1 - Y_{11}$, for level 1, and

2) $Y_{21} \sim \text{Be}(\alpha_2, \alpha_2)$, $Y_{22} = 1 - Y_{21}$, $Z_{21}|Y_{21} \sim \text{Bin}(\delta_{21}, Y_{21})$, $Y_{23}|Z_{21} \sim \text{Be}(\alpha_2 + Z_{21}, \alpha_2 + \delta_{21} - Z_{21})$, $Y_{24} = 1 - Y_{23}$, for level 2.

The marginal variance for the random measure of any partitioning sets at level 2 is the same, that is, $\text{Var}\{\text{P}(B_{2j})\} = \{2(\alpha_1+\alpha_2)+3\}/\{16(2\alpha_1+1)(2\alpha_2+1)\}$ for all $j = 1, \ldots, 4$. It is straightforward to show that the correlation of the measures assigned to two sets at level 2 are:

$$\begin{aligned}
\rho_{12} &= \text{Corr}\{\text{P}(B_{21}), \text{P}(B_{22})\} = \frac{2(\alpha_2 - \alpha_1) - 1}{2(\alpha_1 + \alpha_2) + 3}, \\
\rho_{13} &= \text{Corr}\{\text{P}(B_{21}), \text{P}(B_{23})\} = \frac{\delta_{21}(2\alpha_1 - 1) - 2\alpha_2(2\alpha_2 + \delta_{21} + 1)}{(2\alpha_2 + \delta_{21})\{2(\alpha_1 + \alpha_2) + 3\}} \quad \text{and} \\
\rho_{14} &= \text{Corr}\{\text{P}(B_{21}), \text{P}(B_{24})\} = \frac{-\delta_{21}(2\alpha_1 + 1) - 2\alpha_2(2\alpha_2 + \delta_{21} + 1)}{(2\alpha_2 + \delta_{21})\{2(\alpha_1 + \alpha_2) + 3\}}.
\end{aligned}$$

Finally due to symmetry in the construction, $\text{Corr}\{\text{P}(B_{22}), \text{P}(B_{23})\} = \rho_{14}$, $\text{Corr}\{\text{P}(B_{22}), \text{P}(B_{24})\} = \rho_{13}$ and $\text{Corr}\{\text{P}(B_{23}), \text{P}(B_{24})\} = \rho_{12}$.

For illustration we concentrate on two cases for α_m, namely $a/2^m$ to define a discrete measure, and am^2 to define a continuous measure for $m = 1, 2$. In both cases $a > 0$. We write $\rho_{ij}(a, \delta_{21})$ to highlight the dependence on a and δ_{21}.

Figure 2 depicts the correlation function $\rho_{ij}(a, \delta_{21})$ for $a \in (0, 20)$ and $\delta_{21} = 0, 1, 10, 100$. The panels in the first row correspond to the case when the rubbery Polya tree defines a discrete measure. The solid line in the three panels, obtained by taking $\delta_{21} = 0$, shows the correlation in a Dirichlet process, this turns out to be constant for all $a > 0$ and takes the value of $-1/3$. Starting from this Dirichlet case, as we increase the value of δ_{21} we see that the correlation ρ_{13} (first row, middle panel) increases as δ_{21} increases, even becoming positive for $\delta_{21} = 10, 100$ and for approximately $a > 3$, whereas ρ_{14} becomes more negative as δ_{21} increases. This complementary effect between ρ_{13} and ρ_{14} simply reflects the fact that the measures $P(B_{2j})$, $j = 1, \ldots, 4$ need to add up to one. Note that there is only one (solid) line in the two panels of the first column. This is because the two measures for this case, only involve one variable Y_{mj} for each row and the value of δ_{21} does not change the marginal

$\text{Be}(\alpha_m, \alpha_m)$ distribution in the stationary case.

The second row in Figure 2 corresponds to a continuous rubbery Polya tree. If we take $\delta_{21} = 0$ (solid line in the three panels) the correlation in the rPT correspond to that of a continuous PT. We can see that the second and third panels show negative correlation functions, whereas the first panel (ρ_{12}) presents a positive correlation except for values of a close to zero. In this continuous setting (second row) the effect of $\delta_{21} > 0$ on the correlations is not as evident as in the discrete case (first row). However, a similar behavior as in the discrete case exists. A larger value of δ_{21} increases the correlation ρ_{13}, making it less negative, and decreases the correlation ρ_{14}, making it more negative. This effect is stronger for smaller values of a.

Differences in the correlations of random probabilities over the first two levels are key to understand the differences in posterior inference under the PT and the rPT. In the continuous case with $\alpha_m = am^2$, the correlation ρ_{12} in a PT can take positive values for certain values of the parameter a (see bottom left panel in Figure 2). In fact $\rho_{12} > 0$ for $a > 1/6$. This is in contrast to the DP, for which the correlation is always negative between any disjoint pair of sets.

In the remaining of this section let us concentrate in the continuous case $\alpha_m = am^2$. Consider a PT in levels beyond $m = 2$, and focus on the sibling pair of partitioning subsets $\{B_{m,2j-1}, B_{m,2j}\}$ of a parent set $B_{m-1,j}$, for $j = 1, \ldots, 2^{m-1}$. It is straightforward to show that the covariance in the random measures for any two sibling subsets is the same for all pairs in the same level m, and is given by

$$\sigma_{\text{sib}}^{(m)} = \text{Cov}\{\text{P}(B_{m,2j-1}), \text{P}(B_{m,2j})\} = \frac{\alpha_m}{2(2\alpha_m + 1)} \prod_{k=1}^{m-1} \frac{\alpha_k + 1}{2(2\alpha_k + 1)} - \left(\frac{1}{2}\right)^{2m}, \qquad (5)$$

for $m > 1$. It is not difficult to prove that if $a > 1/(4m - 2)$ then $\sigma_{\text{sib}}^{(m)} > 0$. In other words, the correlation between sibling subsets is positive for sufficiently large precision parameter a. Moreover, from levels $m = 3$ onwards the correlation in the measure for any two pair of subsets under the same first partition (either B_{11} or B_{12}) is positive, regardless of whether or not the sets are siblings. In complement, the correlation between any two sets in the same

level, one descendant of B_{11} and the other descendant of B_{12}, is negative.

Since the rPT does not change the covariance between sibling subsets, $\sigma^{(m)}_{\text{sib}}$ in (5) remains valid for the rPT implying that the correlation between siblings is positive. Focus now on the two sets at level m that are next to each other at the right and left boundaries of B_{11} and B_{12} respectively, in notation $B_{m,2^{m-1}}$ and $B_{m,2^{m-1}+1}$. In the PT the correlation in the measures assigned to these two sets is always negative for all $m \geq 1$. In the rPT this correlation becomes more negative. If we now concentrate on the left neighbor of the set $B_{m,2^{m-1}}$, say $B_{m,2^{m-1}-1}$, and the set $B_{m,2^{m-1}+1}$, their correlation induced under the PT is also negative for all $m \geq 1$, however, under the rPT the same correlation is less negative only in the sets at level $m = 2$ (see ρ_{13} in Figure 2), for $m \geq 3$ the correlation is more negative. Therefore, the sets B_{21} and B_{23} have increased (less negative) correlation. Furthermore, two descendants on the same level of the tree, one from B_{21} and another from B_{23} have also increased correlation. Something similar happens between B_{22} and B_{24} and all its descendants. In general, as mentioned before, descendants of B_{11} (or B_{12}) on the same level, have positive correlation under the PT for sufficiently large a, and in the rPT the correlation is increased between every other set (1-3, 2-4, etc.) and the correlation is slightly decreased, otherwise. This differences between continuous PT and rPT are summarized in Figure 3. This elaborated correlation structure leads to the desired smoothing across random probabilities.

In summary, the PT and the rPT share many prior properties as random probability measures. However, the rubbery Polya tree imposes a correlation of the random branching probabilities that induce a tighter structure within each level. In contrast, the PT prior assumes independence. The positive correlation in the pair $(Y_{m,j}, Y_{m,j+2})$ is achieved by adding the latent variables $\mathcal{Z}$ which allow for borrowing of information across partitioning subsets within each level. An increment in Y_{mj} favors a corresponding increment in $Y_{m,j+2}$. This in turn smooths the abrupt changes in probabilities in neighboring partition sets. We will later discuss the details of the posterior updating in the following Section. In particular, Figure 4 illustrates this desired effect.

4 Posterior inference

We illustrate the advantages, from a data analysis perspective, of the rubbery Polya tree compared to the simple Polya tree. Recall from Section 1 that the rPT can be characterized as an $\mathcal{A}-$mixture of Polya trees with the mixing distribution given by the law of the latent process $\mathcal{Z}$. Thus, posterior inference for the proposed process is as simple as for the Polya tree model.

4.1 Updating the rubbery Polya tree

Let $X_1, \ldots, X_n$ be a sample of size n such that $X_i|\mathrm{P} \overset{\text{iid}}{\sim} \mathrm{P}$ and $\mathrm{P} \sim \mathrm{rPT}(\Pi, \mathcal{A}, \mathcal{D})$. Then the likelihood for $\mathcal{Y}_1, \mathcal{Y}_2, \ldots, \mathcal{Y}_M$, given the sample $\mathbf{x}$, is:

$$\prod_{i=1}^{n} \prod_{m=1}^{M} \prod_{j=1}^{2^m} Y_{mj} I(x_i \in B_{mj}) = \prod_{m=1}^{M} \prod_{j=1}^{2^m} Y_{mj}^{N_{mj}},$$

where $N_{mj} = \sum_{i=1}^{n} I(x_i \in B_{mj})$ for $j = 1, \ldots, 2^m$.

By Theorem 2 of Ferguson (1974), the posterior distribution $\mathrm{P}|\,\mathbf{x}$ is again tailfree. The updated process $[(\mathcal{Y}_m, \mathcal{Z}_m)|\mathbf{x}]$ is not a Markov beta process as in equations (1) and (2). It is a new Markov beta process with a different distribution for the latent process $\mathcal{Z}_m$. This posterior distribution can be characterized by the conditional posterior distribution $[\mathcal{Y}_m|\mathcal{Z}_m, \mathbf{x}]$ and the marginal posterior distribution $[\mathcal{Z}_m|\mathbf{x}]$. As in the prior process, $Y_{m1}, \ldots, Y_{m,2^m-1}$ are conditionally independent given $\mathcal{Z}_m$ and $\mathbf{x}$ with (updated) beta distributions given by

$$Y_{m,2j+1}|Z_{m,2j-1}, Z_{m,2j+1}, \mathbf{x} \sim \mathrm{Be}(\alpha_{m,2j+1} + Z_{m,2j-1} + Z_{m,2j+1} + N_{m,2j+1}\,,$$

$$\alpha_{m,2j+2} + \delta_{m,2j-1} - Z_{m,2j-1} + \delta_{m,2j+1} - Z_{m,2j+1} + N_{m,2j+2}), \quad (6)$$

for $j = 0, \ldots, 2^{m-1} - 1$. The conditional independence structure of the model implies that also the posterior latent process $\mathcal{Z}_m$ follows again another Markov process. Unfortunately the appropriate transition probabilities for $\mathcal{Z}_m$ are not easily seen. This makes it impossible to exploit the representation as Markov beta process for posterior simulation.

Instead, a straightforward Gibbs sampling posterior simulation scheme (Smith and Roberts, 1993) can be implemented. For that we require the conditional distribution $[\mathcal{Y}_m|\mathcal{Z}_m, \mathbf{x}]$

given in (6) together with the conditional distribution $[\mathcal{Z}_m|\mathcal{Y}_m, \mathbf{x}]$. Since the likelihood does not involve $\mathcal{Z}_m$, the latter full conditional does not depend on the data. Moreover, the $Z_{m,1}, \ldots, Z_{m,2^{m-1}-3}$ are conditionally independent given $\mathcal{Y}_m$ with probabilities given by

$$Z_{m,2j-1}|Y_{m,2j-1}, Y_{m,2j+1} \sim \text{BBB}(\alpha_{m,2j+1}, \alpha_{m,2j+2}, \delta_{m,2j-1}, p_{m,2j-1}), \tag{7}$$

with $p_{m,2j-1} = y_{m,2j-1}y_{m,2j+1}/\{(1-y_{m,2j-1})(1-y_{m,2j+1})\}$ for $j = 1, \ldots, 2^{m-1}-1$. BBB stands for a new discrete distribution called Beta-Beta-Binomial whose probability mass function is given by

$$\text{BBB}(z|\alpha_1, \alpha_2, \delta, p) = \frac{\Gamma(\delta+1)\Gamma(\alpha_1)\Gamma(\alpha_2+\delta)}{{}_2H_1(-\delta, -\delta+1-\alpha_2; \alpha_1; p)} \frac{p^z I_{\{0,1,\ldots,\delta\}}(z)}{\Gamma(1+z)\Gamma(1+\delta-z)\Gamma(\alpha_1+z)\Gamma(\alpha_2+\delta-z)},$$

where ${}_2H_1(-\delta, -\delta+1-\alpha_2; \alpha_1; p) = \sum_{k=0}^{\delta}(p^k/k!)(-\delta)_k(-\delta+1-\alpha_2)_k/(\alpha_1)_k$ is the hypergeometric function, which can be evaluated in most statistical software packages, and $(\alpha)_k$ is the pochhammer number.

The conditional distribution (6) is of a standard form. The conditional distribution (7) is finite discrete. Therefore sampling is straightforward. We illustrate the proposed prior by considering two small examples:

Example 1.

As a first example we consider an extreme case with only one observation, say $X = -2$. For the prior specification we centered the rPT at $\text{P}_0 = \text{N}(0, 1)$, with the partitions B_{mj} defined as in (3). We use $\alpha_{mj} = \alpha_m = am^2$, and set $a = 0.1$. The parameters $\mathcal{D}$ were taken to be constant across the tree, that is, $\delta_{mj} = \delta$ for all m and j. A range of values for δ was used for illustration.

We considered a finite rPT with $M = 4$ levels for illustration and defined P to be uniform within sets B_{4j} for $j = 1, \ldots, 2^4$. The partitioning sets were bounded to lie in $(-3, 3)$. A Gibbs sampler was run for 10,000 iterations with a burn-in of 1,000. Figure 4 presents the posterior predictive distributions, that is, posterior means for the probability assigned to the elements of the partition at level 4 divided by the length of the set. The top left graph ($\delta = 0$) corresponds to the posterior estimates obtained by a PT prior. The choice of $\delta > 0$ in the rPT clearly makes the posterior estimates to be a lot smoother. In particular for $\delta = 10$ (bottom

right graph) the mass has been shifted to the left towards the observed point producing a smooth density (histogram). The counterintuitive seesaw pattern following the partition boundaries in the PT has disappeared. The extreme outlier in this example exacerbated the differences between the two models.

Example 2.

As a second illustration we consider a simulated data set of size $n = 30$ taken from a normal distribution with mean -0.5 and standard deviation 0.5. We used a rPT process with prior mean $\mathrm{P}_0 = \mathrm{N}(0, 1)$. The parameters satisfy $\alpha_{mj} = am^2$ and $\delta_{mj} = \delta$ for all m, j, and we used $a = 0.01, 0.1, 1$ and $\delta = 0, 20$ for comparison. Since $\log_2(30) = 4.90$, a finite tree with $M = 5$ levels is used. The measure P is distributed uniformly within the sets at level 5. The Gibbs sampler was run for 20,000 iterations with a burn-in of 2,000.

Figure 5 shows summaries of the posterior distribution. For the graphs in the first column we took $\delta = 0$, which corresponds to a PT, and for the second column we took $\delta = 20$. For the first, second and third rows we took $a = 0.01, 0.1, 1$ respectively. The solid line corresponds to the posterior predictive density, the dotted lines are 95% posterior probability intervals and the dashed line corresponds to the $N(-0.5, 0.5^2)$ simulation truth.

The scale in the right panels was kept the same as in the left panels to facilitate comparison. There are two aspects that can be seen from Figure 5. The predictive distribution (solid lines) obtained with the rPT smooths out the peaks when compared with that of the PT, and is closer to the true density (dashed line). Additionally, there is a huge gain in precision when using a rPT instead of a PT. This increment in precision is more marked for smaller values of a (first and second rows). The advantages of the rPT versus the PT can be explained by the borrowing of strength across partitioning subsets in the rPT. Of course, if the simulation truth were a highly irregular distribution with discontinuous density and other rough features, then the borrowing of strength across partitioning subsets could be inappropriate and would lead to a comparison that is less favorable for the rPT. See model number 5 in the simulation study reported later, in Section 5.1. for an example when

borrowing of strength might be undesireable.

From these examples we can see that the effect of δ in the rPT is to smooth the posterior probabilities and decrease the posterior variance.

4.2 Mixture of rubbery Polya trees

For actual data analysis, when more smoothness is desired, an additional mixture can be used to define an rPT mixture model. For example, let $N(x|\,\mu,\sigma^2)$ denote a normal kernel with moments (μ,σ^2). Consider $G(y) = \int \mathrm{N}(y|\,\mu,\sigma^2)\,\mathrm{dP}(\mu)$ with the rPT prior on P as before.

Alternatively, a mixture can be induced by assuming that the base measure that defines the partition is indexed with an unknown hyperparameter θ. Let P_θ denote the base measure and $\Pi_\theta = \{B^\theta_{mj}\}$ the corresponding sequence of partitions. A hyperprior $\theta \sim \pi(\theta)$ leads to a mixture of rPT's with respect to the partition Π_θ. If we consider a finite tree and define the measure on the sets at level M according to P_θ, then the posterior conditional distribution for θ has the form

$$[\theta|\mathcal{Y},\mathbf{x}] \propto \left\{\prod_{m=1}^{M}\prod_{j=1}^{2^m} Y_{mj}^{N^\theta_{mj}}\right\}\left\{\prod_{i=1}^{n} f_\theta(x_i)\right\}\pi(\theta),$$

where $N^\theta_{mj} = \sum_{i=1}^n I(x_i \in B^\theta_{mj})$ for $j = 1,\ldots,2^m$ and f_θ the density corresponding to P_θ. Sampling from this posterior conditional distribution can be achieved by implementing a Metropolis-Hastings step as suggested by Walker and Mallick (1997).

5 Numerical studies

In this section we carry out two numerical studies to further illustrate inference under the proposed rPT.

5.1 Simulation Study

We consider the set of mixtures of normal densities originally studied by Marron and Wand (1992), which are often used as benchmark examples for density estimation problems. The examples include unimodal, multimodal, symmetric and skewed densities. We concentrate

on the first ten of these benchmark examples. The ten densities are shown as the solid lines in Figure 6.

From each of the ten models we simulated $n = 50$ and 100 observations, and repeated this experiment fifty times. For all repetitions of the experiment and all models we assumed a rPT with prior specifications: $\mathrm{P}_0 = \mathrm{N}(0, 1)$, $\alpha_{mj} = am^2$, $\delta_{mj} = \delta$ for all m and j. Several choices for the precision and rubbery parameters were considered for comparison. Specifically, $a = 0.01, 0.1, 1$ and $\delta = 5, 20$. In this case, $\log_2(50) = 5.64$ and $\log_2(100) = 6.64$, so the rule of thumb suggests 6 or 7 levels to define a finite tree. We used $M = 6$ for both sample sizes.

For each experiment we computed (by Monte Carlo integration) the integrated L1 error defined as $L1 = \int |\hat{f}(x) - f(x)| \mathrm{d}x$, with $\hat{f}(x)$ the posterior mean density based on 20,000 iterations of a Gibbs sampler with a burn-in of 2,000, and $f(x)$ the true density that was used to simulate the data. The L1 error for the rPT was compared with that under a simple PT with the same prior specifications. The ratio of the integrated L1 errors (RL1) was then averaged over the 50 experiments. The mean RL1 and the numerical standard deviations are presented in Table 1.

The numbers reported in Table 1 highlight the differences in inference under the PT versus the rPT. The effect of the rubbery parameter δ is relative to the value of the precision parameter a. For smaller a, the rPT shows a better performance than the simple PT, except perhaps for density 5, which has a sharp spike around zero (see, Figure 6). For larger values of a, the effect of δ vanishes for most of the models, as the prior becomes increasingly more informative. The effect worsens for the spiked model 5 and the the well separated bimodal model 7. Regarding the sample size, the rPT performs slightly better for smaller sample sizes together with larger values of δ. This is explained by the fact that the latent process $\mathcal{Z}$ can be seen as additional latent data that compensate the lack of observations in some regions by borrowing strength from the neighbors.

The optimal degree of dependence (δ) varies across different data sets. One may therefore allow δ to be random by assigning a hyper-prior distribution, say $\pi(\delta)$, and let the data

determine the best value. The complete conditional posterior distribution for δ is

$$[\delta \mid \mathcal{Y}, \ldots] \propto \left[\prod_{m=1}^{M} \prod_{j=1}^{2^{m-1}-1} \frac{\Gamma(\delta+1)\Gamma(2\alpha_m+\delta)\left\{(1-Y_{m,2j-1})(1-Y_{m,2j+1})\right\}^{\delta}}{\Gamma(\delta-Z_{m,2j-1}+1)\Gamma(\alpha_m+\delta-Z_{m,2j-1})} \right] \pi(\delta) I(\delta \geq z^*),$$

where $z^* = \max\{Z_{m,2j-1} : j = 1, \ldots, 2^{m-1}-1, m = 1, \ldots, M\}$ and "..." behind the conditioning bar stands for all other parameters. In particular we propose a truncated geometric hyper prior distribution of the form $\pi(\delta) \propto p(1-p)^{\delta} I_{\{0,\ldots,20\}}(\delta)$. We implemented this extra step in the simulation study with $p = 0.5$ and concentrated on the case with $a = 0.01$ and $n = 100$. The relative integrated L1 errors with respect to the simple PT are shown in Table 2. As can be seen, the RL1 errors favor the rPT against the simple PT. The only exception is model 5, for which, if we consider the standard error, the performance is the same for both RPMs. Figure 6 shows the density estimates obtained with this setting of the rPT.

For actual data analysis, PT models are often replaced by mixture of PT models to reduce the dependence on the partitions. The mixture is with respect to the centering measure. We therefore include also mixture of PT models in the comparison and report relative L1 error, relative to a simple PT. The prior specifications for the mixture are: $P_\theta = N(\theta, 4)$ and $\theta \sim \mathrm{N}(0, 1)$. We ran a Gibbs sampler for 20,000 iterations with a burn in of 2,000. Again, we took samples from the 10 models and repeated the experiment fifty times with $n = 50, 100$ sample sizes.

The average RL1 together with their numerical standard deviations are reported in the last two columns of Table 1. As can be seen, for a small precision parameter, $a = 0.01$, mixtures of Polya trees present a similar error as the simple PT for small sample size ($n = 50$) and slightly better performance for $n = 100$. However, the rPT outperforms the mixture of Polya trees in 7 out of 10 models. On the other hand, the mixture of Polya trees compare favorably for larger values of the precision parameter, say $a = 1$.

In general, for small values of a, the posterior RPMs (PT, rPT or mixtures of PT's) depend almost entirely on the data, whereas setting a larger a would mean that the prior RPM is more informative and that there is more shrinkage to the centering measure P_θ. On the other hand, a small value of a implies a rough RPM, due to a larger variance. In the

latter case the relative advantage of the rPT comes to bear.

Finally, additional simulations (not reported here) show that the number of levels M in the finite tree prior has an important effect on the RL1 values. Larger values of M clearly benefit the rPT with respect to a simple PT.

5.2 Nuclear Waste Data

Draper (1999) presented an interesting discussion about what he called "the small print on Polya trees". He considered highly skewed data that were collected to assess the risk of underground storage of nuclear waste. The observations are radiologic doses for humans on the surface. There are $n = 136$ positive values, 134 of which ranging from 0 to 0.8522 with two outliers at 3.866 and 189.3. Since the complete original data is not available, we use simulated data that replicates all important features of the original data by including the two outliers with known values and simulating the remaining 134 observations from a lognormal distribution in such a way that they are mostly within the interval $(0, 0.8522)$, as in Draper (1999). That is, let $X_i = \exp(W_i)$ with $W_i \sim \mathrm{N}(-1, 0.5^2)$ for $i = 1, \ldots, 134$ together with $X_{135} = 3.866$ and $X_{136} = 189.3$. The simulated sample, on log-scale, is shown in Figure 7.

We analyzed this data with both the rPT and the PT models. We worked on the log-scale and centered the prior measures at $\mathrm{P}_0 = \mathrm{N}(0, 4)$ with the partitions defined by (3). We defined continuous measures with parameters $\alpha_{mj} = am^2$ and took $a = 0.1$ as in Draper (1999). Finite trees were defined for $M = 7$ and $M = 8$ levels. The former is the number of levels suggested by the sample size and the rule of thumb and the latter is the number of levels actually used by Draper (1999). The rubbery parameter was fixed at $\delta = 20$.

We ran the Gibbs sampler for 20,000 iterations with a burn-in of 2,000. We computed the logarithm of the pseudo-marginal likelihood (LPML) statistic to assess the goodness of fit for the models. The LPML is defined as the sum of the logarithm of the conditional predictive ordinate for each observation. See, for example, Gelfand et al. (1992). These and some other posterior summaries are presented in Table 3. The LPML statistics for the PT and rPT models have almost the same values for the same M, showing a minimally better fit

for the rPT. In general, models with $M = 8$ have better fit than those with $M = 7$. However, posterior inference for the quantities reported in Table 3 do not change much when going from 7 to 8 levels.

Posterior credible intervals were obtained for the mean radiologic dose μ_X. From Table 3 we can see that the posterior distribution of μ_X is narrower under the rPT prior, for both values of M, resulting in a shorter credible interval for μ_X. Perhaps the most important aspect in the context of the application, is the amount of mass assigned to large radiologic doses (upper tail), where the two outliers are present. We computed the posterior probability of the event $\{X > 1.65\}$, on the original scale. These probabilities, with $M = 8$, were estimated at 0.035 under the PT and 0.101 under the rPT priors, that is, the rPT is assigning considerably more probability to the possibility of an outlier than the PT.

We finish our study by comparing with inference under a mixture of PT and a mixture of rPT model. The mixture is with respect to the centering measure. In particular, the centering measure was $\mathrm{P}_\theta = N(\theta, 9)$ with $\theta \sim \mathrm{N}(0, 1)$. The rubbery parameter for the rPT was $\delta = 20$. Model comparison and posterior summaries are reported in the last block of rows in table 3. The additional mixture improves the model fit with a modest advantage for the mixture of rPT. As before, the 95% posterior credible interval for μ_X is narrower for the mixture of rPT, which also assigns a larger probability to the tail beyond 1.65, compared with the mixture of a simple PT.

6 Discussion

We have introduced a new tail free random measure that improves the traditional Polya tree prior by allowing the branching probabilities to be dependent within the same level of the tree, defining a tightened structure in the tree. Our new prior retains the simplicity of the Polya tree for making nonparametric inference. Centering our prior around a parametric model is achieved in the same way as in the simple Polya tree. However, posterior estimates obtained by the rubbery Polya tree are improved by the borrowing of information within the levels which produce an spreading of information everywhere in the tree.

Although the rubbery Polya tree prior greatly reduces the mass jumps on neighboring partitions, inference for density estimation might still be desired to be even smoother. For example, the density estimates shown in Figure 5 might be unreasonable for a distribution that is known to be smoother. This could easily be addressed by adding an additional convolution in the sampling model. The resulting rPT mixture model generates much smoother random densities.

Another critical issue for implementations of PT and rPT models is the computational effort that is required to track the number of observations in each of many partitioning subsets and to update the random probabilities. The problem is exacerbated in higher dimensions when partitions become multivariate rectangles. This difficulty is not addressed by the rPT and remains exactly as in the PT. Hanson (2006) and Jara et al. (2009) propose efficient implementations of PT models for multivariate distributions. Using a marginalized version of the model, marginalizing with respect to the random probability measure, it is possible to implement efficient posterior simulation. However, these constructions can not be naturally used for the rPT. The rPT remains useful only for univariate distributions.

Acknowledgments

The research of the first author was partially supported by *The Fulbright-García Robles Program* and *Asociación Mexicana de Cultura, A. C.*.

References

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

Barron, A., Schervish, M. J., and Wasserman, L. (1999), "The consistency of posterior distributions in nonparametric problems," *Annals of Statistics*, 27, 536–561.

Berger, J. and Guglielmi, A. (2001), "Bayesian testing of a parametric model versus nonparametric alternatives," *Journal of the American Statistical Association*, 96, 174–184.

Branscum, A., Johnson, W., Hanson, T., and Gardner, I. (2008), "Bayesian semiparametric ROC curve estimation and disease risk assessment," *Statistics in Medicine*, in press.

Branscum, A. J. and Hanson, T. E. (2008), "Bayesian Nonparametric Meta-Analysis Using Polya Tree Mixture Models," *Biometrics*, 64, 825–833.

Do, K.-A., Müller, P., and Tang, F. (2005), "A Bayesian mixture model for differential gene expression," *Journal of the Royal Statistical Society, Series C*, 54, 627–644.

Draper, D. (1999), "Discussion on the paper: Bayesian nonparametric inference for random distributions and related functions," *Journal of the Royal Statistical Society, Series B*, 61, 510–513.

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

Fabius, J. (1964), "Asymptotic behavior of Bayes estimates," *Annals of Mathematical Statistics*, 35, 846–856.

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

— (1974), "Prior distributions on spaces of probability measures," *Annals of Statistics*, 2, 615–629.

Freedman, D. A. (1963), "On the asymptotic behaviour of Bayes estimates in the discrete case," *Annals of Mathematical Statistics*, 34, 1386–1403.

Gelfand, A., Dey, D., and Chang), H. (1992), "Model determination using predictive distributions with implementation via sampling based methods (with discussion)," in *Bayesian Statistics 4 – Proceedings of the Fourth Valencia International Meeting*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., pp. 147–167.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999), "Consistent semiparametric Bayesian inference about a location parameter," *Journal of Statistical Planning and Inference*, 77, 181–193.

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

Hanson, T. and Yang, M. (2007), "Bayesian semiparametric proportional odds models," *Biometrics*, 63, 88–95.

Hanson, T. E. (2006), "Inference for mixtures of finite Polya tree models," *Journal of the American Statistical Association*, 101, 1548–1564.

Jara, A., Hanson, T. E., and Lesaffre, E. (2009), "Robustifying Generalized Linear Mixed Models Using a New Class of Mixtures of Multivariate Polya Trees," *Journal of Computational and Graphical Statistics*, 18, 838–860.

Kottas, A., Müller, P., and Quintana, F. (2005), "Nonparametric Bayesian modeling for multivariate ordinal data," *Journal of Computational and Graphical Statistics*, 14, 610–625.

Lavine, M. (1992), "Some aspects of Polya tree distributions for statistical modelling," *Annals of Statistics*, 20, 1222–1235.

— (1994), "More aspects of Polya tree distributions for statistical modelling," *Annals of Statistics*, 22, 1161–1176.

Li, M., Reilly, C., and Hanson, T. (2008), "A semiparametric test to detect associations between quantitative traits and candidate genes in structured populations," *Bioinformatics*, 24, 2356–62.

Lo, A. Y. (1984), "On a class of Bayesian nonparametric estimates: I. Density estimates," *Annals of Statistics*, 12, 351–357.

Marron, J. S. and Wand, M. P. (1992), "Exact mean integrated square error," *Annals of Statistics*, 20, 712–736.

Nieto-Barajas, L. E. and Walker, S. G. (2002), "Markov beta and gamma processes for modelling hazard rates," *Scandinavian Journal of Statistics*, 29, 413–424.

Paddock, S., Ruggeri, F., Lavine, M., and West, M. (2003), "Randomised Polya tree models for nonparametric Bayesian inference," *Statistica Sinica*, 13, 443–460.

Paddock, S. M. (2002), "Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse ," *Biometrika*, 89, 529–538.

Smith, A. and Roberts, G. (1993), "Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Series B*, 55, 3–23.

Walker, S., Damien, P., Laud, P., and Smith, A. (1999), "Bayesian nonparametric inference for distributions and related functions (with discussion)," *Journal of the Royal Statistical Society, Series B*, 61, 485–527.

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.

Yang, Y., Müller, P., and Rosner, G. (2010), "Semiparametric Bayesian Inference for Repeated Fractional Measurement Data," *Chilean Journal of Statistics*, 1, 59–74.

Zhang, S., Müller, P., and Do, K.-A. (2009), "A Bayesian Semiparametric Method for Jointly Modeling a Primary Endpoint and Longitudinal Measurements," *Biometrics*.

Zhao, L. and Hanson, T. (2010), "Spatially dependent Polya tree modeling for survival data," *Biometrics*, to appear.

Table 1: Relative integrated L1 errors (RL1): rPT over PT for $\delta = 5, 20$, and MPT over PT. Fifty data sets of size $n = 50, 100$ were simulated for each of the 10 models in Marron and Wand (1992). Average over the 50 repetitions as well as the standard deviation in parenthesis are reported.

	rPT($\delta = 5$)		rPT($\delta = 20$)		MPT	
	$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$
Model	$a = 0.01$					
1	0.72 (0.06)	0.72 (0.04)	0.54 (0.09)	0.55 (0.06)	1.03 (0.13)	0.98 (0.12)
2	0.72 (0.06)	0.74 (0.05)	0.58 (0.10)	0.60 (0.07)	0.75 (0.13)	0.62 (0.15)
3	0.87 (0.12)	0.86 (0.11)	0.96 (0.14)	0.87 (0.13)	0.76 (0.19)	0.56 (0.11)
4	0.77 (0.07)	0.81 (0.06)	0.72 (0.08)	0.79 (0.09)	1.01 (0.19)	0.95 (0.21)
5	1.21 (0.17)	1.10 (0.13)	1.99 (0.30)	1.94 (0.36)	0.81 (0.38)	0.83 (0.47)
6	0.71 (0.07)	0.76 (0.05)	0.59 (0.08)	0.61 (0.07)	1.00 (0.15)	1.02 (0.16)
7	0.85 (0.07)	0.86 (0.05)	1.01 (0.11)	0.95 (0.08)	1.02 (0.19)	0.96 (0.10)
8	0.72 (0.06)	0.74 (0.05)	0.57 (0.08)	0.62 (0.07)	0.98 (0.12)	0.95 (0.14)
9	0.73 (0.07)	0.77 (0.06)	0.60 (0.07)	0.65 (0.08)	1.01 (0.12)	0.98 (0.14)
10	0.75 (0.07)	0.79 (0.05)	0.64 (0.08)	0.69 (0.08)	1.05 (0.17)	1.06 (0.20)
Model	$a = 0.1$					
1	0.93 (0.07)	0.93 (0.05)	0.82 (0.11)	0.81 (0.08)	0.83 (0.22)	0.86 (0.21)
2	0.94 (0.06)	0.94 (0.05)	0.86 (0.11)	0.85 (0.11)	0.56 (0.23)	0.63 (0.22)
3	1.03 (0.08)	1.00 (0.06)	1.22 (0.11)	1.23 (0.15)	0.55 (0.10)	0.45 (0.09)
4	1.04 (0.09)	1.01 (0.07)	1.15 (0.13)	1.13 (0.12)	0.77 (0.21)	0.69 (0.19)
5	1.55 (0.20)	1.32 (0.14)	2.45 (0.44)	2.35 (0.38)	0.36 (0.12)	0.56 (0.31)
6	0.95 (0.07)	0.96 (0.06)	0.93 (0.12)	0.89 (0.10)	0.92 (0.17)	0.96 (0.18)
7	1.10 (0.10)	1.03 (0.06)	1.46 (0.16)	1.29 (0.15)	0.97 (0.10)	0.95 (0.11)
8	0.95 (0.06)	0.96 (0.04)	0.89 (0.10)	0.88 (0.09)	0.79 (0.22)	0.88 (0.17)
9	0.99 (0.06)	0.96 (0.05)	0.96 (0.15)	0.92 (0.08)	0.96 (0.19)	0.95 (0.14)
10	0.99 (0.06)	0.98 (0.05)	1.02 (0.09)	1.00 (0.09)	0.85 (0.18)	0.96 (0.18)
Model	$a = 1$					
1	1.00 (0.05)	1.00 (0.03)	0.94 (0.14)	0.95 (0.08)	0.57 (0.19)	0.51 (0.23)
2	0.97 (0.03)	0.98 (0.03)	0.96 (0.07)	0.95 (0.06)	0.57 (0.29)	0.86 (0.26)
3	1.00 (0.01)	1.00 (0.01)	1.01 (0.01)	1.02 (0.02)	0.48 (0.10)	0.41 (0.06)
4	1.04 (0.03)	1.06 (0.03)	1.16 (0.09)	1.21 (0.08)	0.60 (0.13)	0.54 (0.08)
5	1.12 (0.02)	1.14 (0.02)	1.33 (0.05)	1.46 (0.05)	0.59 (0.09)	0.53 (0.11)
6	1.05 (0.04)	1.02 (0.03)	1.20 (0.09)	1.15 (0.09)	0.66 (0.20)	0.85 (0.21)
7	1.12 (0.02)	1.09 (0.02)	1.39 (0.07)	1.37 (0.06)	0.83 (0.07)	0.93 (0.04)
8	1.01 (0.04)	1.00 (0.04)	1.03 (0.10)	1.03 (0.08)	0.51 (0.27)	0.86 (0.12)
9	1.07 (0.06)	1.04 (0.03)	1.29 (0.12)	1.19 (0.10)	0.77 (0.24)	0.94 (0.12)
10	1.02 (0.02)	1.03 (0.01)	1.06 (0.04)	1.09 (0.04)	0.67 (0.17)	0.57 (0.12)

Table 2: Relative integrated L1 errors for the rPT over the PT for the 10 models of Marron and Wand (1992). Same specifications as in Table 1 but with $\pi(\delta) \propto (0.5)^{\delta+1} I_{\{0,\ldots,20\}}(\delta)$, $n = 100$ and $a = 0.01$.

Model	1	2	3	4	5	6	7	8	9	10
Ave. RL1	0.75	0.78	0.88	0.84	1.27	0.78	0.89	0.79	0.78	0.78
Std. RL1	(0.061)	(0.070)	(0.045)	(0.088)	(0.358)	(0.074)	(0.050)	(0.063)	(0.053)	(0.066)

Table 3: Posterior inference summaries for the PT model, rPT model with $\delta = 20$, and mixtures of PT and rPT with respect to the centering measure. In all cases $a = 0.1$.

Levels	Model	LPML	95% CI for μ_X	$P(X > 1.65\|\mathbf{x})$
$M = 7$	PT	-133.30	(0.48, 2.16)	0.030
	rPT	-132.61	(0.51, 1.63)	0.120
$M = 8$	PT	-130.28	(0.49,2.15)	0.035
	rPT	-129.91	(0.51,1.66)	0.101
$M = 7$	MPT	-126.42	(0.47, 1.97)	0.016
	MrPT	-124.91	(0.52, 1.71)	0.096

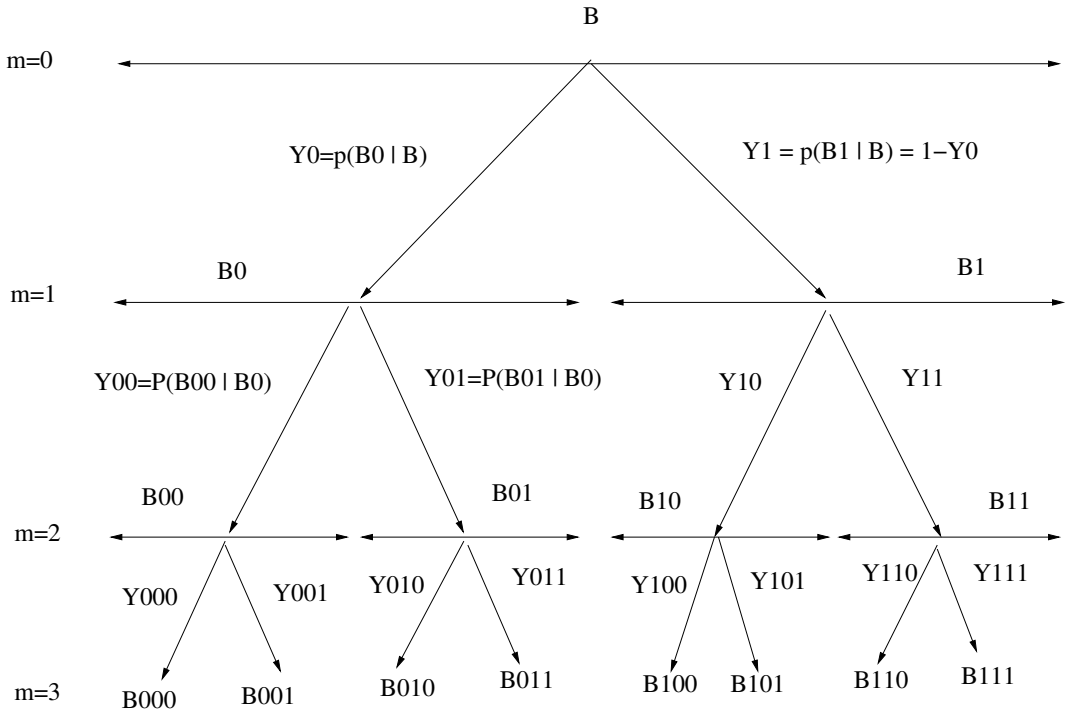

Figure 1: Nested partition of the sample space B into partitions $\pi_m = \{B_{\epsilon_1 \cdots \epsilon_m},\ \ \epsilon_j \in \{0, 1\}\}$, $m = 1, 2, \ldots$. The random variables $Y_{\epsilon_1 \cdots \epsilon_m}$ determine the random probabilities $P(B_{\epsilon_1 \cdots \epsilon_m} \mid B_{\epsilon_1 \cdots \epsilon_{m-1}})$ under a PT distributed RPM P.

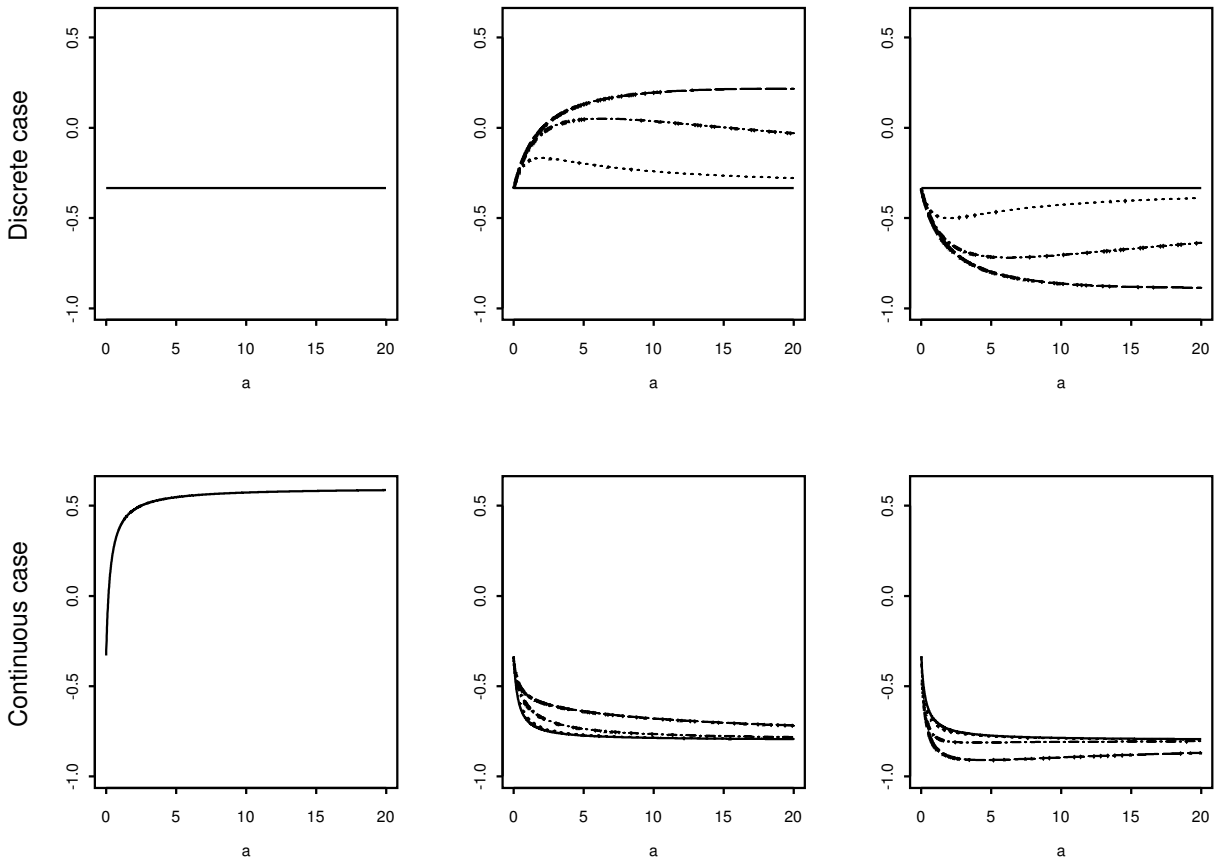

Figure 2: Correlation function $\rho_{ij}(a, \delta_{21})$: First row $\alpha_m = a/2^m$, second row $\alpha_m = am^2$, $m = 1, 2$. First column ρ_{12}, second column ρ_{13} and third column ρ_{14}. (—) $\delta_{21} = 0$, $(\cdots)$ $\delta_{21} = 1$, (-·-) $\delta_{21} = 10$, and (---) $\delta_{21} = 100$.

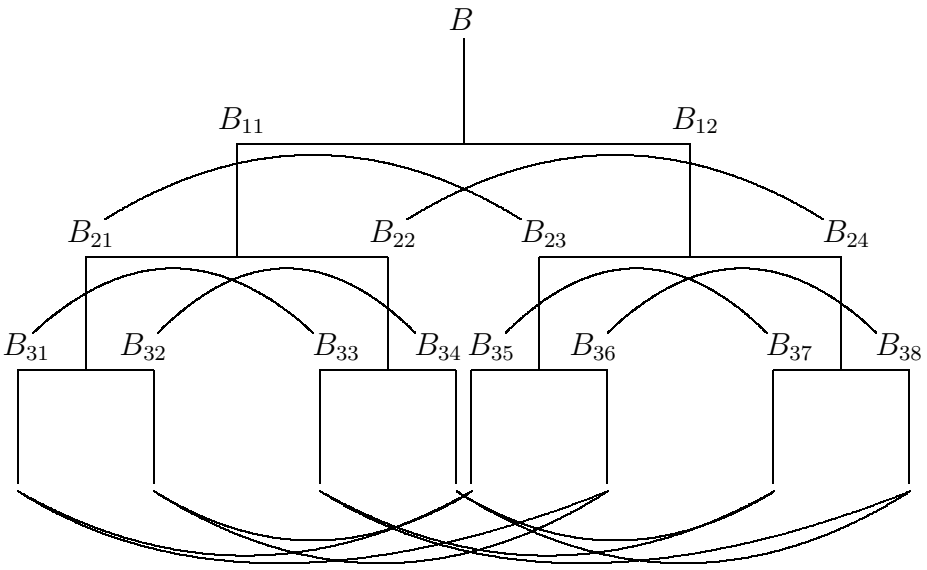

Figure 3: Changes in correlations $\mathrm{Corr}\{\mathrm{P}(B_{mj}), \mathrm{P}(B_{mj'})\}$ for some pairs of partitioning subsets from a PT prior to a corresponding rPT prior. A solid curve indicates increased correlation. From $m \geq 2$, any other pair of sets in the same level with no linking curve have decreased correlation.

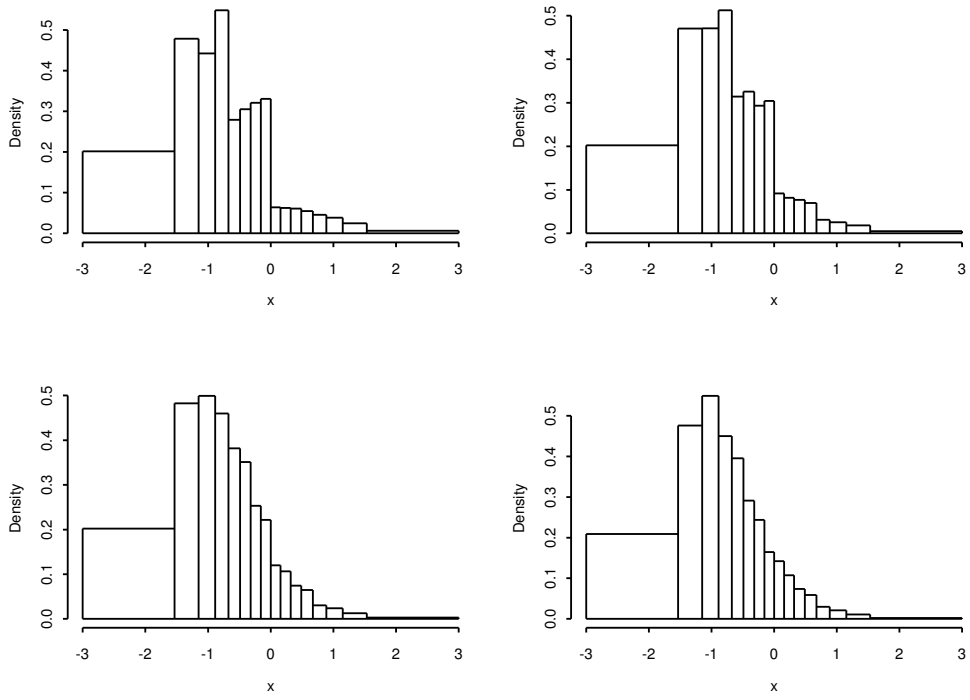

Figure 4: Posterior predictive distributions for a rPT with a sample of size 1 at $X = -2$: Top left $\delta = 0$, top right $\delta = 1$, bottom left $\delta = 5$ and bottom right $\delta = 10$.

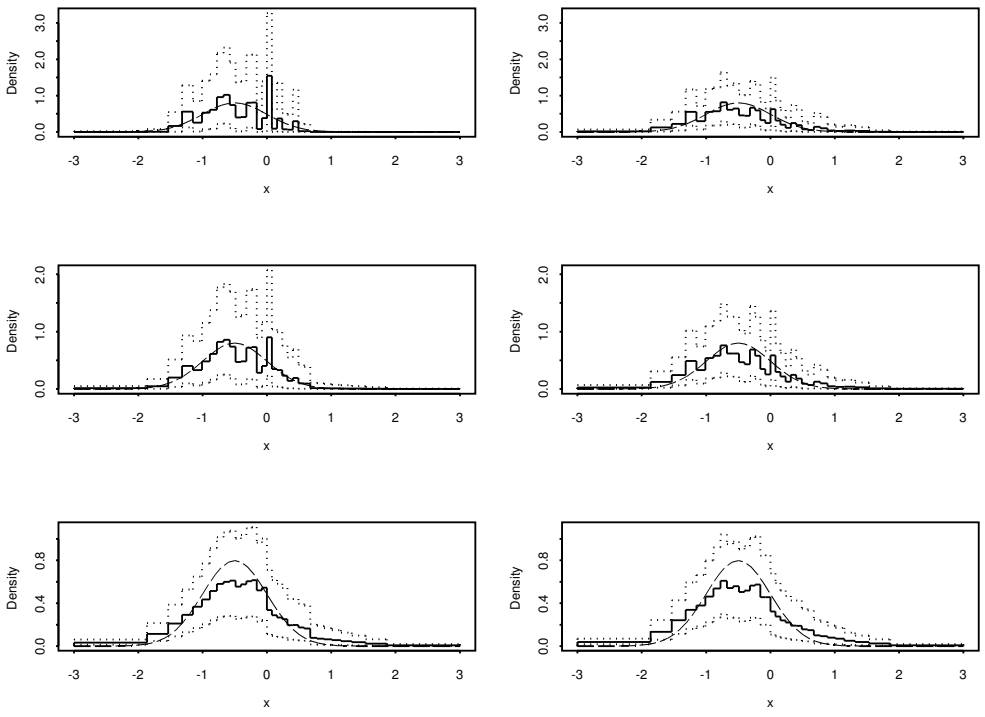

Figure 5: Posterior distributions from a rPT with $\delta = 0$ (first column) and $\delta = 20$ (second column) with 30 simulated data points from $\mathrm{N}(-0.5, 0.5^2)$. First row ($a = 0.01$), second row ($a = 0.1$) and third row ($a = 1$). Posterior predictive (solid line), 95% CI (dotted line) and true density (dotted line).

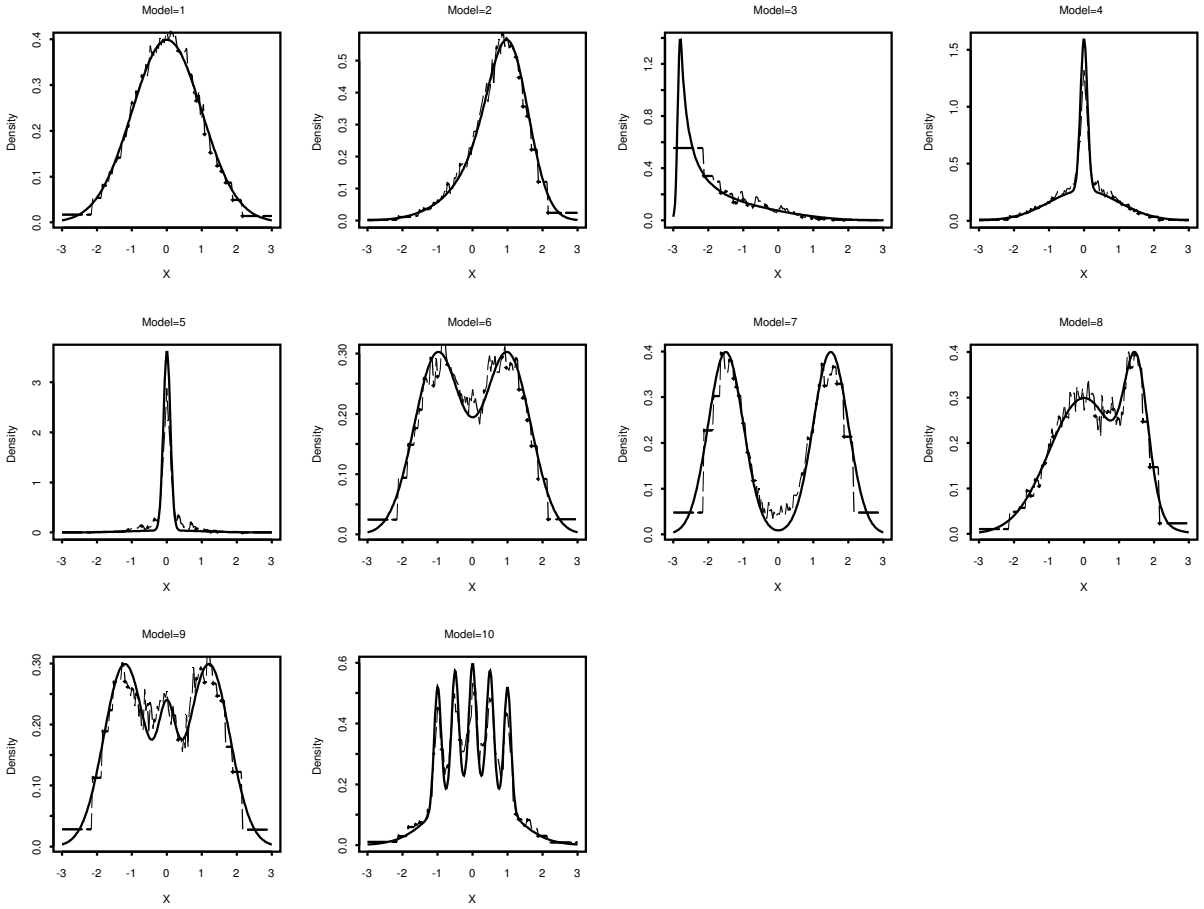

Figure 6: Benchmark models of Marron and Wand (1992). True density (solid line), rPT estimate with a hyper prior on δ (dashed line).

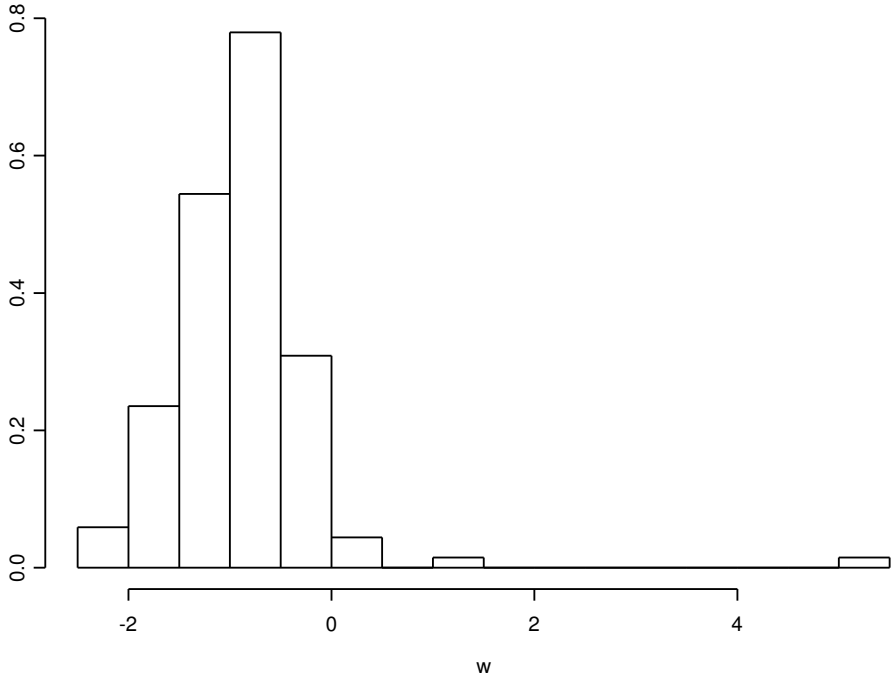

Figure 7: Histogram of the simulated data, mimicking Draper (1999). In logarithmic scale.