Statistica Sinica **13**(2003), 443-460

RANDOMIZED POLYA TREE MODELS FOR NONPARAMETRIC BAYESIAN INFERENCE

Susan M. Paddock, Fabrizio Ruggeri, Michael Lavine and Mike West

RAND, CNR-IATMI, Duke University and Duke University

Abstract: Like other partition-based models, Polya trees suffer the problem of partition dependence. We develop Randomized Polya Trees to address this limitation. This new framework inherits the structure of Polya trees but "jitters" partition points and as a result smooths discontinuities in predictive distributions. Some of the theoretical aspects of the new framework are developed, followed by discussion of methodological and computational issues arising in implementation. Examples of data analyses and prediction problems are provided to highlight issues of Bayesian inference in this context.

Key words and phrases: Bayesian nonparametrics, Bayesian trees, partitioning, Polya tree prior, Randomized Polya tree.

1. Introduction

The last decade or so has witnessed substantial interest in Bayesian nonparametric modeling, partly driven by the developing capacity to fit more flexible models due to advances in computation. Bayesian methods based on Polya tree models, however, have remained relatively undeveloped, though the field is now giving more attention to Polya tree based frameworks. The interest in tree models is in part due to the need to move Bayesian nonparametrics more forcefully into problems of moderate and high-dimensional data sets. Also, the computational challenges with more standard approaches, such as the "gold standard" of Dirichlet process mixtures, push researchers to look for alternative approaches. This paper is concerned with introducing and developing the new modeling framework of Randomized Polya trees that addresses certain practical shortcomings of more standard Polya trees encountered in data modeling.

One drawback of Polya tree models is the dependence of the model, and the inferences resulting from the model analysis, on the sample space partitions required to specify tree structures. This has been noted by authors such as Ferguson (1974) and Lavine (1992). Basically, the partition specified for a Polya tree can adversely affect inference by introducing artifacts into posterior distributions; Ferguson (1974) points out that a Polya tree can be used to construct a density with respect to Lebesgue measure that exists with probability one, but that the

density will have discontinuities at all partition endpoints with probability one. Mixtures of Polya trees (Lavine (1992) and (1994), Hanson and Johnson (2002)) can reduce the influence of the partition on inference; however, the modeling required for mixtures of Polya trees may be inappropriate for some scenarios. The problem of partition dependence is not restricted to Polya trees; almost any statistical method relying upon a partitioning of the sample space can be subject to this criticism. For example, Hartigan (1996) mentions concerns about lack of "amalgamation" of observations in neighboring bins of a Bayesian histogram built from a fixed set of candidate partition points; Coifman and Donoho (1995) use cycle-spinning to average over various "shifts" of the data in order to reduce the artifacts induced by the discontinuities of wavelet basis functions; when modeling time series count data, Kolaczyk (1999) averages over all possible sums of adjacent observations across all levels of a recursive dyadic partitioning tree to reduce the effect of the partition on inference in Bayesian multiscale modeling. Unlike the latter two approaches, which focus on shifting observations or signal over a fixed partition, we reduce partition dependence in the Polya tree modeling framework by developing the Randomized Polya tree, which is a class of hierarchical Bayesian models in which a Polya tree is the population hyperparameter, and the hierarchical structure induces a modest degree of additional smoothing that neatly reduces the problem of jumps in predictive densities. We do this by introducing observation-specific parameters that slightly "jitter" the tree partition.

In the following section, Polya tree priors are reviewed. We then discuss the issues of partition dependence and introduce the Randomized Polya tree framework. This is explored and developed, and we introduce simulation-based methods for model fitting and analysis. Implications of parameter choice in the randomized tree are assessed, and a small simulation study comparing the randomized tree and Polya tree follows. We illustrate with an example of a data analysis that motivated this research which highlights the impact of using Randomized Polya trees versus Polya trees for modeling, and we conclude with a summary discussion.

2. Polya Trees

We first present the definition of a Polya tree as described by Lavine (1992) and then give a brief overview of the specification and structure of a Polya tree. Let $E = \{0, 1\}$, $E^0 = \emptyset$, E^t be the $t-$fold product $E \times E \times \cdots \times E$ and $E^* = \cup_0^\infty E^t$. Let Ω be a separable measurable space, $\pi_0 = \Omega$ and $\Pi = \{\pi_t; t = 0, 1, \ldots\}$ be a separating binary tree of partitions of Ω; that is, $\pi_0, \pi_1, \ldots$ is a sequence of partitions such that $\cup_0^\infty \pi_t$ generates the measurable sets and such that every $B \in \pi_{m+1}$ is obtained by splitting some $B' \in \pi_m$ into two pieces. Let $B_\emptyset = \Omega$ and, for all $\epsilon = \epsilon_1 \ldots \epsilon_m$ (also denoted $\epsilon_{1:m}$) $\in E^*$, let $B_{\epsilon 0}$ and $B_{\epsilon 1}$ be the two

pieces into which B_ϵ is split. Then, the random probability measure F on Ω has a *Polya tree prior* (Lavine (1992, 1994), Mauldin, Sudderth and Williams (1992)) with parameter $(\Pi, \mathcal{A})$ if there exist nonnegative parameters $\mathcal{A} = \{\alpha_\epsilon : \epsilon \in E^*\}$ and a collection of random parameters $\mathcal{Y} = \{Y_\epsilon : \epsilon \in E^*\}$ such that

(i) the random variables in $\mathcal{Y}$ are independent;
(ii) for every $\epsilon \in E^*$, $Y_\epsilon \sim Beta(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$;
(iii) for every $m = 1, 2, \ldots$ and every $\epsilon_{1:m}$, $F(B_{\epsilon_{1:m}}) = \prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_{1:j-1}} \prod_{j=1;\epsilon_j=1}^{m} (1 - Y_{\epsilon_{1:j-1}})$, where the first term in each product is interpreted as $Y_\emptyset$ or $1 - Y_\emptyset$.

A "canonical" partition (Lavine (1992)) for the case in which $\Omega = (0, 1]$ is the dyadic rational binary partition, where $\Pi = \{(k2^{-t}, (k+1)2^{-t}]\}$ for $k = 0, 1, \ldots, 2^t - 1$ and every $t = 0, 1, \ldots$. For example, Ω would be partitioned into $B_0 = (0, 0.5]$ and $B_1 = (0.5, 1]$ at level 1; B_0 into $(0, 0.25]$ and $(0.25, 0.5]$, and B_1 into $(0.5, 0.75]$ and $(0.75, 1]$ at level 2; and so on.

The Polya tree prior can be centered about a continuous distribution function G by taking $\Pi = \{(G^{-1}(k2^{-t}), G^{-1}((k+1)2^{-t})]\}$. The strength of prior belief in the choice of G can be quantified by the selection of the parameters, $\mathcal{A}$. This specification assumes symmetric Beta priors with common hyperparameters $\alpha_{\epsilon_{1:j}} = a_j,\ j = 1, 2, \ldots$, so that $E(Y_{\epsilon_{1:(j-1)}}) = 0.5$. In the canonical Polya tree, conditions indicating that F is absolutely continuous (with probability one) require the α parameters to increase rapidly (i.e., variance to reduce rapidly) as one descends the tree. In the symmetric case, a sufficient condition is that $a_j = cj^2$ for some constant $c > 0$. The constant, c, can be regarded as a tuning parameter; the choice of c might depend on considerations about the amount of prior information relative to sample size. See Lavine (1992) and Ferguson (1974) for further details on the choice and implications of a given α.

In theory, the above construction can be continued indefinitely, but for the purposes of this paper it is impossible to compute an infinite-level tree. Therefore, the distribution F will be approximated with a finite Polya tree (Lavine (1994)) by specifying and updating the Polya tree to a prespecified, finite level, m. The choice of m could be based on the precision of the data – e.g., how many observations fall into each partition element at level m. The conditional densities on the resulting partition elements within each partition element at level m can be chosen randomly according to the base measure G; computationally, this entails following an observation x down the tree to level m, and drawing x as a random variate according to G restricted to the partition element $B_{\epsilon_{1:m}}$ which contains x.

Posterior updating of the Beta distribution parameters is straightforward, since Polya trees are conjugate. Each of the Beta parameters, $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$, is equal to its prior value plus a count of the number of observations falling into

the corresponding partition elements $B_{\epsilon 0}$ and $B_{\epsilon 1}$, respectively. This is clearly spelled out in Lavine (1992).

3. Theory and Construction

3.1. Rationale

One drawback of Polya trees is that resulting inferences may depend in important and, sometimes, undesirable ways on the specified sample space partition. Ferguson (1974) discusses partition dependence induced by recursive dyadic rational partitioning. In higher-dimensional spaces, the partition can play an even more dramatic role in posterior inference, as prior choice may more strongly affect posterior inference than in lower dimensional spaces.

The partition can be chosen for convenience; Lavine (1992) suggests using a dyadic rational recursive partition when there is no pressing reason to select the partition in any special manner. Even for other modeling approaches that require a choice of partition, such as Bayesian histograms, the ease of using a dyadic rational partition might outweigh concerns about partition dependence (Hartigan (1996)). The inferential focus of many Polya tree applications is not on how a particular partition affects posterior inference, which makes the strong influence of the partition on the posterior particularly disconcerting. It is therefore desirable to develop a new method which reduces the role of the partition on posterior distributions, as pointed out by Lavine (1992).

3.2. Random partition models

A new strategy for reducing the effect of partition dependence is to employ a Randomized Polya tree approach. The basic idea is that, in generating an observation from a tree model, the prespecified partition is subject to a modest amount of random "jitter" to induce smoothing of the discontinuities that result from using a fixed partition. This jitter is induced by constructing a function which "distorts" the quantile function corresponding to F. It is similar but not identical to one of the random functions studied by Dubins and Freedman (1967) and Mauldin and Williams (1990). The partitioning in the randomized tree developed here is based on the dyadic rational partition specification, but it is conceivable that many other variants on this construction could be developed as well. The randomized tree is constructed here by building upon the recursive dyadic partition as follows. At each level of the tree, a bit of randomness is added to the selection of partition element cut points. To partition $(0, 1]$, a partition cut point is selected to be "near" the dyadic rational, 0.5, at level 1; within each of the two subintervals at the second level, the subsequent cut points are selected to divide the subintervals "nearly" in half, with "nearly" represented by small random deviations from an equal splitting of the interval. Partitioning at subsequent levels of the tree follows this pattern.

Each observation x_i in a data set of n observations is given its own partition via this construction, with each partition being a small, random perturbation of the underlying dyadic rational partition. The resulting n partitions can be intuitively regarded to be "centered" about the dyadic partition. A variant of the approach is to jitter partition elements for just few of the levels close to the root of the tree; this is appealing because the discontinuities that are induced by a fixed partitioning at the top levels of the tree are the most influential of the set of all discontinuities that result from partition choice. It is straightforward to accommodate this special case by modifying the methods presented in this paper.

3.3. Partition construction for Randomized Polya tree

Just as the partition for the Polya tree is constructed recursively, so is the partition for the randomized tree. We illustrate the construction on $(0, 1]$, but the framework can be readily extended to multidimensional Polya trees; for details, see Paddock (1999).

The randomized tree is constructed by altering the choice of the canonical partition as follows. For a single observation x_i, $(0, 1]$ will be split not into halves but into pieces of sizes β_{i_1} and $(1 - \beta_{i_1})$, so that $B_{i_0} = (0, \beta_{i_1}]$ and $B_{i_1} = (\beta_{i_1}, 1]$. Then, B_{i_0} and B_{i_1} will each be further split into two pieces of size determined by proportion β_{i_2} (or $1 - \beta_{i_2}$) of their length: $B_{i_0} = (0, \beta_{i_1}]$ becomes $B_{i_0} = (0, \beta_{i_1}\beta_{i_2}] \cup (\beta_{i_1}\beta_{i_2}, \beta_{i_1}]$, and similarly $B_{i_1} = (\beta_{i_1}, \beta_{i_1} + (1 - \beta_{i_1})\beta_{i_2}] \cup (\beta_{i_1} + (1 - \beta_{i_1})\beta_{i_2}, 1]$. At level 3, a new parameter β_{i_3} is introduced, and the partition elements of level 2 are split according to proportion β_{i_3}. This method of recursive partitioning occurs at all subsequent levels of the tree, as a parameter β_{i_j} is introduced at each level j of the tree. Further, the $\{\beta_{i_j}\}$ are taken to be independent. The partition element end points at level k are determined by $\beta_{i_1}, \ldots, \beta_{i_k}$.

The prior distributions for $\{\beta_{i_j}\}$ will determine how much "jitter" will be induced. The $\{\beta_{i_j}\}$ will be independent *a priori* given τ, and will come from a common uniform distribution which is concentrated about 0.5 along the interval $(0, 1)$:

$$\beta_{i_j} \sim U(0.5 - \tau, 0.5 + \tau), \qquad j = 1, 2, \ldots, \tag{1}$$

where τ is selected to be in $(0, 0.5)$. Guidance on how to choose τ follows in the next section.

3.4. Interpretation

The collection of the independent terms $\{\beta_{i_j}\}$ implies a function $\Lambda_i(\cdot)$, which determines the partition of $(0, 1]$ for the observation x_i. For any $z_i \in (0, 1]$

with $z_i = \sum_{j=1}^{\infty} \epsilon_{i_j}/2^j$, consider unique dyadic expansions of z_i by selecting ϵ_{i_j} according to the rule that $\epsilon_{i_j} = 0$ if $z_i \leq \sum_{k=1}^{j} \epsilon_{i_k}^{I\{j\neq k\}} 2^{-k}$, and $\epsilon_{i_j} = 1$ otherwise. By this rule we select only unique dyadic expansions; for instance, it would be impossible to represent $z_i = 0.5$ by both $(0111\ldots)$ and $(1000\ldots)$. The collection of independent terms $\{\beta_{i_j}\}$ determines the map $\Lambda_i(\cdot)$ from $(0,1]$ to $(0,1]$ such that $x_i = \Lambda_i(z_i) = \Lambda_i\left(\sum_{j=1}^{\infty} \epsilon_{i_j}/2^j\right) = \sum_{j=1}^{\infty} \epsilon_{i_j}\beta_{i_j}H_{i_j}$, where $H_{i_j} = \prod_{l=1}^{j-1} \beta_{i_l}^{1-\epsilon_{i_l}}(1-\beta_{i_l})^{\epsilon_{i_l}}$ and $H_{i_1} = 1$. Hence, Λ_i maps the partition for a value z_i to the perturbed partition for x_i. The common Polya tree distribution F, and the randomized distributions F_i, share the same set of parameters $\mathcal{A}$.

The function $\Lambda_i(\cdot)$ so defined is invertible. The structure of the tree can be exploited as follows to verify that $\Lambda_i(\cdot)$ does indeed have an inverse. Select any $x_i \in (0,1]$ according to the rule above that yields a unique dyadic expansion for x_i. The function Λ_i defines a partition of $(0,1]$ which guides us to finding $z_i = \Lambda_i^{-1}(x_i)$ via finding the vector $(\epsilon_{i_1}, \epsilon_{i_2}, \ldots)$, which represents the unique binary expansion of z_i. To find this vector, follow x_i down the tree: at each level m, set $\epsilon_{i_m} = 1$ if $\sum_{j=1}^{m} \beta_{i_j}\epsilon_{i_j}^{I\{j\neq m\}} \prod_{l=1}^{j-1} \beta_{i_j}^{1-\epsilon_{i_l}}(1-\beta_{i_l})^{\epsilon_{i_l}} < x_i$; otherwise $\epsilon_{i_m} = 0$. After finding $(\epsilon_{i_1}, \epsilon_{i_2}, \ldots)$, set $z_i = \sum_{j=1}^{\infty} \epsilon_{i_j} 2^{-j}$. To see that this z_i is unique and satisfies $\Lambda_i(z_i) = x_i$, note that by construction it is not possible to select the ϵ_{i_j} in such a way that there are two binary representations of the same number. To see that $\Lambda_i(z_i) = x_i$ is indeed true, notice that for all m,

$$\begin{aligned} &\sum_{j=1}^{m} \beta_{i_j}\epsilon_{i_j} \prod_{l=1}^{j-1} \beta_{i_l}^{1-\epsilon_{i_l}}(1-\beta_{i_l})^{\epsilon_{i_l}} < x_i \\ \leq &\sum_{j=1}^{m} \beta_{i_j}\epsilon_{i_j} \prod_{l=1}^{j-1} \beta_{i_l}^{1-\epsilon_{i_l}}(1-\beta_{i_l})^{\epsilon_{i_l}} + \prod_{l=1}^{m} \beta_{i_l}^{1-\epsilon_{i_l}}(1-\beta_{i_l})^{\epsilon_{i_l}}. \end{aligned}$$

The difference between the lower and upper bounds on the interval containing x_i converges to 0 almost surely as $m \to \infty$.

Suppose F follows a Polya tree distribution on a binary tree created via dyadic partitioning. Conditional on Λ_i and F, observation x_i has distribution F_i: $(x_i|\Lambda_i, F) \sim F_i$. Further, the x_i are conditionally independent given the Λ_i and F. By construction, $F(z_i) = F_i(\Lambda_i(z_i))$ for $z_i \in (0,1]$; in other words, if $z_i \sim F(\cdot)$, then F_i is the distribution implied by $x_i = \Lambda_i(z_i)$. The invertibility of $\Lambda_i(\cdot)$ allows for $\Lambda_i^{-1}(x_i) = z_i$ to hold, and an explicit expression of F_i in terms of Λ_i and F is $F(\Lambda_i^{-1}(x_i)) = F_i(x_i)$.

Note that F_i can now be seen to be a random perturbation of F but on the quantile scale. For any value x_i set $u = F_i(x_i) = F(\Lambda_i^{-1}(x_i))$, where u is a random variate in $(0,1)$. By invertibility of F and F_i, it follows that

$F^{-1}(u) = \Lambda_i^{-1}(x_i) = \Lambda_i^{-1}(F_i^{-1}(u))$, implying $\Lambda_i(F^{-1}(u)) = F_i^{-1}(u)$ or, equivalently, $\Lambda_i(Q(u)) = Q_i(u)$, where Q is the quantile function of F and Q_i is the quantile function of F_i. Thus $\Lambda_i(Q(\cdot))$ is the quantile function corresponding to F_i. The nature of the randomized tree is clear here: the Λ_i randomly "distorts" Q, the quantile function corresponding to F.

4. Randomized Polya Trees: Analysis

4.1. Posterior distributions

Combining the likelihood and prior distributions yields the joint posterior distribution:

$$p(\{\beta_{i_1}, \ldots, \beta_{i_m}\}_{i=1}^n, \mathcal{Y}^m | x_1, \ldots, x_n) \propto \left\{ \prod_{i=1}^n \frac{1}{\nu(B_{\epsilon_{i_{1:m}}})} \prod_{j=1}^m Y_{\epsilon_{i_{1:j-1}}}^{1-\epsilon_{i_{1:j}}} (1 - Y_{\epsilon_{i_{1:j-1}}})^{\epsilon_{i_{1:j}}} p(\beta_{i_j}) \right\} \times p(\mathcal{Y}), \qquad (2)$$

where $\nu(\cdot)$ is the Lebesgue measure of $B_{\epsilon_{i_{1:m}}}$. The term $p(\beta_{i_j})$ comes from the prior, while the other terms in the braces of (2) come from the likelihood, and the component outside of the braces is the prior for $\mathcal{Y}$. Exact computation of the marginal posterior distributions of model parameters given data is not possible. Simulation of these marginal posterior distributions by implementing Markov chain Monte Carlo (MCMC) (Gelfand and Smith (1990), Tierney (1994), Gilks, Richardson, and Spiegelhalter (1996)) to obtain draws from the full conditional posterior distributions of the parameters of interest is done.

4.2. Sampling the posterior

The conditional posterior distribution for $Y_{\epsilon_{1:j-1}}$ given $\{\beta_{i_j}\}$ and $\{x_i\}$ is a Beta distribution, as in the usual Polya tree model; just as before, the Beta parameters are equal to the sum of their prior values plus the number of observations falling in the partition element corresponding to that Beta parameter. We sample from this posterior distribution using Gibbs sampling.

To sample the conditional posterior distribution for β_{i_j}, notice from inspection of the joint posterior distribution that the posterior factorizes; i.e., the β_{i_j} are conditionally independent over $i = 1, \ldots, n$. The conditional posterior distribution of β_{i_j} is

$$p(\beta_{i_1}, \ldots, \beta_{i_m} | \mathcal{Y}^m, x_i). \propto \nu(B_{\epsilon_{i_{1:m}}})^{-1} \prod_{j=1}^m p(\beta_{i_j}) Y_{\epsilon_{i_{1:j-1}}}^{1-\epsilon_{i_{1:j}}} (1 - Y_{\epsilon_{i_{1:j-1}}})^{\epsilon_{i_{1:j}}}. \qquad (3)$$

A complication in simulating from the conditional posterior distribution for β_{i_j} is that one does not know which $B_{\epsilon_{1:m}}$ contains x_i. Prior to sampling β_{i_j}, one

of the possible 2^m partition elements at level m, $B_{\epsilon_{1:m}}$ that is to contain x_i must be selected. In the MCMC framework, we select $B_{\epsilon_{1:m}}$ by implementing an independent Metropolis step for drawing β_{i_j}. We draw $\beta_{i_j}^{(t)}$ at iteration t from its proposal distribution, $q(\beta_{i_j}^{(t)}) = U(0.5-\tau, 0.5+\tau)$, and accept the draw with probability $\alpha(\beta_{i_j}^{(t)}, \beta_{i_j}^{(t-1)}) = min\{1, w(\beta_{i_j}^{(t)})/w(\beta_{i_j}^{(t-1)})\}$, where $w(\beta_{i_j})$, the weight function, is equal to the target density divided by the proposal density, $p(\beta_{i_j}|\mathcal{Y}^m, x_i)/q(\beta_{i_j})$.

5. Choosing τ

The objective of the randomized tree is to allow for some jitter of the dyadic partition to induce smoothing. It is not necessary to select Λ_i to be very noisy. The objective pursued by specifying a prior for β_{i_j} is to induce a reasonable amount of randomness to the partition by effectively centering $\Lambda_i(z_i)$ about the line $\Lambda_i(z_i) = z_i$ (which corresponds to the dyadic partition) by centering and concentrating the prior for β_{i_j} about 0.5.

To further guide the choice of τ, we illustrate the impact of τ on the implied prior predictive distribution of x_i

$$P(x_i|F) = \int F_i(x_i)dP(\Lambda_i) = \int F(\Lambda_i^{-1}(x_i))dP(\Lambda_i), \tag{4}$$

under the case where $z_i \sim U(0,1)$. The x_i are conditionally independent given F as the Λ_i are independent and are independent of F.

As the exact form of $P(x_i|F)$ above is difficult to interpret analytically, we explore this via simulation. The effect of various choices of prior distributions for β on $\Lambda_i(\cdot)$ is examined to assess how much variability is added to $P(x_i|F)$ by $\Lambda_i(\cdot)$. F is fixed to be a $U(0,1)$ distribution ($Y_\epsilon = 0.5$ for all ϵ, and $\Pi =$ dyadic rational partition), and computation of the tree is carried to 15 levels. Figure 1 displays 10^5 draws from the simulated distribution for x_i. The simulated distributions for $\tau = \{0.01, 0.05, 0.10, 0.25, 0.33, 0.5\}$ are displayed in Figure 1. For each τ, a histogram of the simulated distribution for x_1 and a normal quantile plot of a random subsample of 1000 of the 10^5 simulations of $\Phi^{-1}(x_1)$ plotted against quantiles of the standard normal distribution are presented. The lower the value of τ, the closer to uniformity of the simulated distribution. As τ increases, the median of the simulated distribution is sampled from a wider range, which pushes the mass of the distribution toward the edges of the unit interval. Another feature of the histograms of the samples in Figure 1 is that there are bumps visible at approximately 0.25 and 0.75, which are clear for $\tau = 0.25$ and $\tau = 0.33$. These bumps could be due to the effect of the variable median being replicated at all levels of the tree. Another explanation for the bumps is that

the Lebesgue measure of the partition element sizes plays a role in the predictive distribution; small values of the Lebesgue measure of a partition element will increase the predictive density, while comparatively larger values will decrease it. Note that for all values of τ examined in Figure 1, this effect does not seem to dominate the overall pattern of the simulated predictive distributions. The normal quantile plots reveal generally linear patterns to the quantiles of the simulated values for x_1, suggesting the distributions of $\Phi^{-1}(\cdot)$ might approximate normal distributions with variances greater than 1. Based on these simulations, we recommend selecting τ to be small enough so that β_{i_j} are "near" 0.5 to avoid adding excessive noise to the model. In our experience, as illustrated below, we have found that selecting τ to be between 0.01 and 0.05 usually suffices.

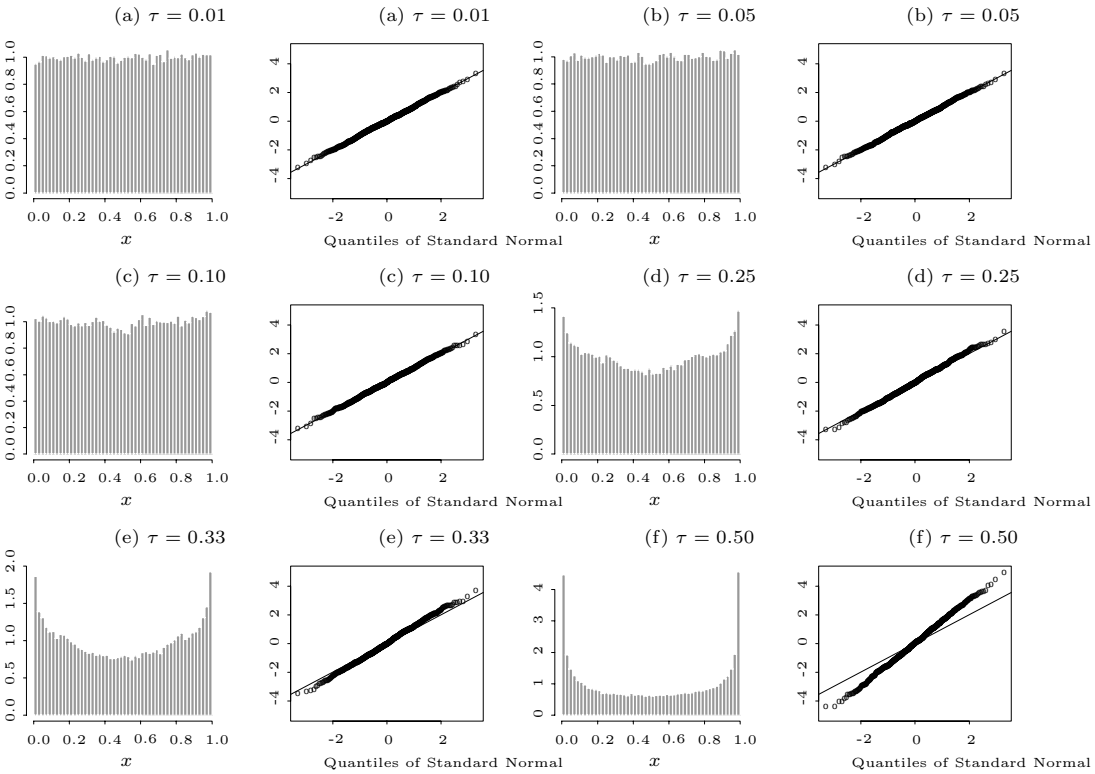

Figure 1. Simulated distribution under prior choices for τ when $F = U(0, 1)$.

To briefly illustrate how the choice of τ determines the degree of smoothing, suppose $F \sim PT(\Pi, \mathcal{A})$, where Π is the dyadic rational partition, and the parameters in $\mathcal{A}$, $\alpha_{\epsilon_{1:j}} = j^2$. Suppose that $x = 0.51$ is observed. Figure 2 shows histograms of draws from the posterior predictive distributions for the Polya tree prior (upper left) and the Randomized Polya tree for τ equal to $0.025, 0.05$ and

0.10. As is clear from the upper left subfigure of Figure 2, the choice of partition cutpoints, 0.5 and 0.75, plays a key role in determining the resulting distribution. The distributions yielded by the randomized trees are smoothed about the canonical partition cutpoints. Values for τ of 0.025 and 0.05 reasonably smooth the distribution, while $\tau = 0.10$ may oversmooth a little in this case, as evidenced by the slightly increased mass toward the endpoints 0 and 1 in the lower right histogram in Figure 2.

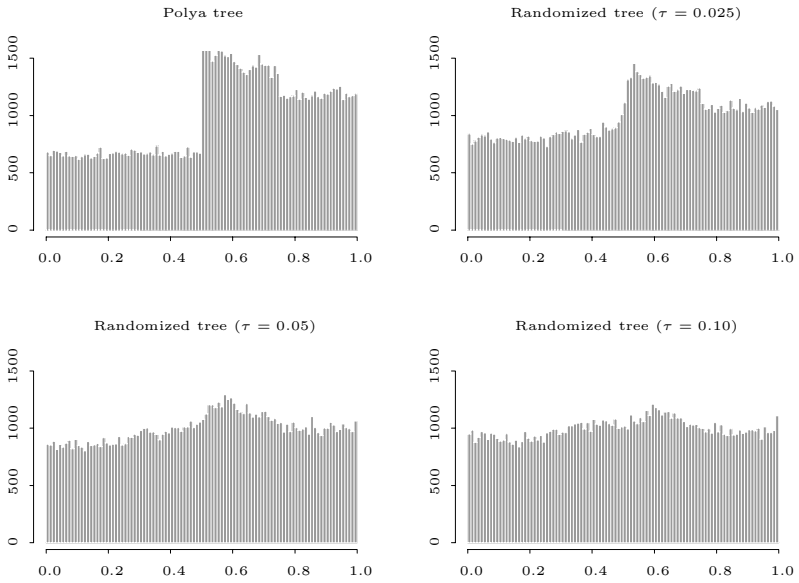

Figure 2. $x = 0.51$: Simulations of the posterior predictive distributions for the Polya tree prior (upper left) and Randomized Polya trees for $\tau \in \{0.025, 0.05, 0.10\}$.

6. Comparison of Randomized and Polya Trees

To understand how the smoothing of discontinuities using the randomized tree affects the estimation of posterior predictive distributions relative to the Polya tree, we present a simulation study in which we compare the sum of squared error loss (SSEL) of the randomized tree and Polya tree posterior predictive distributions with respect to the true distribution from which the data were generated. For each of 100 simulations, Π = dyadic rational partition on $(0, 1]$ and $c\alpha_{\epsilon_{1:j}} = cj^2$ $(j = 1, \ldots, 8)$ *a priori,* where $c > 0$ is a scalar. Data were drawn from a Uniform$(0.25, 0.55)$ distribution. This allows us to compare not just overall performance of the randomized and Polya trees but also performance on two regions of the space – $(0.25, 0.50]$, where the bulk of the posterior predictive distribution mass falls, and $(0.50, 0.55)$, where a much smaller portion of

it lies. Given the dyadic rational partition, we would expect to see differences between the two methods with respect to how they distribute mass on these two parts of the support. We drew 5000 MCMC draws from the conditional posterior distributions of model parameters, discarding a burn-in of 500. We varied the sample size ($N = 50, 100, 500, 1000$), scalar ($c = 0.1, 1, 10$), and τ for the randomized tree ($\tau = 0.01, 0.025, 0.05, 0.10, 0.25$) across the simulations. The simulation parameters and results are available in Table 1. The overall SSEL for the randomized tree relative to that of the Polya tree on $(0, 1]$ is given in column 4, while SSEL(RT)/SSEL(PT) on the regions $(0.25, 0.50]$ and $(0.50, 0.55]$ are given in columns 5 and 6, respectively. The overall comparison of the randomized and Polya trees (column 4) shows that only for the extreme value of $\tau = 0.25$, does the randomized tree add enough noise to make it worse than the Polya tree in terms of SSEL. Despite this, note that column 6, the comparison of the randomized tree to the Polya tree on $(0.50, 0.55)$, shows that the randomized tree does better on this region than the Polya tree for almost all cases of $\tau = 0.25$. However, performance is sacrificed on the larger portion on the space, $(0.25, 0.50]$. For the other values of τ, the posterior predictive distribution generated from the randomized tree has a smaller loss than that coming from the Polya tree for Sets (a)$-$(b); both models are equivalent for Sets (c)$-$(d); and the Polya tree has smaller loss generally for Sets (e)$-$(f).

We also examined the integrated squared error loss (ISEL) of the posterior predictive distributions generated from the randomized and Polya trees with respect to the true distribution function of the data (Table 2). Again, there is variation across parameterizations and sample sizes with respect to the relative performance of the randomized tree; but it appears that the randomized tree outperforms or has comparable performance to the Polya tree overall (column 4) for $\tau = 0.01, 0.025, 0.05$.

7. Example

We illustrate the above development with an application to a data set from Chambers and Hastie (1992); the data are included in the S-Plus software (S-PLUS (1999)) (data frame "air"). The data set contains $n = 111$ complete observations on daily air quality measurements for the New York Metropolitan area; these data were previously analyzed by Müller, Erkanli and West (1996). Variables are ozone (parts per billion), solar radiation, average wind speed in miles per hour, and maximum daily temperature. Like Müller, Erkanli and West (1996) and Chambers and Hastie (1992), we transform the ozone measurement by taking its cube root. The goal of our analyses is to explore the relationships amongst the measured variables via conditional predictive simulation and

to highlight how randomized trees smooth partition artifacts in predictive distributions. Since the data are multidimensional, we extend the Randomized Polya tree framework which was described in Section 2 by substituting Dirichlet for Beta distributions in the specification of the prior on Y_ϵ, as described by Mauldin, Sudderth and Williams (1992) and Paddock (1999).

Table 1. Sum of squared error loss with respect to $U(0.25, 0.55)$ of the posterior predictive distributions generated from the randomized relative to that of the Polya tree for various sample sizes N and parameter values c and τ. Column 4: Overall comparison on $(0, 1]$. Column 5: Comparison on $(0.25, 0.50]$. Column 6: Comparison on $(0.50, 0.55)$.

Set	(1) N	(2) c	(3) τ	(4) SSEL(RT)/ SSEL(PT) overall	(5) SSEL(RT)/ SSEL(PT) on (0.25,0.50]	(6) SSEL(RT)/ SSEL(PT) on (0.50,0.55)
	50	0.10	0.010	0.766	0.653	1.071
	50	0.10	0.025	0.687	0.459	1.383
(a)	50	0.10	0.050	0.708	0.486	1.313
	50	0.10	0.100	0.685	0.393	0.878
	50	0.10	0.250	2.810	1.760	0.722
	50	1.00	0.010	1.007	0.929	1.047
	50	1.00	0.025	0.998	0.886	1.064
(b)	50	1.00	0.050	0.912	0.989	0.886
	50	1.00	0.100	0.943	1.324	0.571
	50	1.00	0.250	2.403	4.221	0.641
	100	0.10	0.010	0.715	0.604	1.036
	100	0.10	0.025	0.739	0.463	1.539
(c)	100	0.10	0.050	0.849	0.548	1.796
	100	0.10	0.100	0.844	0.454	0.968
	100	0.10	0.250	4.000	2.357	0.889
	100	1.00	0.010	1.018	0.948	1.065
	100	1.00	0.025	1.112	0.992	1.224
(d)	100	1.00	0.050	1.046	1.192	1.044
	100	1.00	0.100	1.064	1.590	0.626
	100	1.00	0.250	3.056	5.235	0.704
	500	1.00	0.010	1.136	0.938	1.380
	500	1.00	0.025	1.645	1.215	2.260
(e)	500	1.00	0.050	1.997	2.059	2.262
	500	1.00	0.100	2.080	2.330	1.423
	500	1.00	0.250	9.840	10.526	1.090
	1000	10.00	0.010	1.076	1.094	1.118
	1000	10.00	0.025	1.207	1.394	1.247
(f)	1000	10.00	0.050	1.208	2.443	1.022
	1000	10.00	0.100	1.201	2.971	0.618
	1000	10.00	0.250	3.281	8.992	0.698

Table 2. Integrated squared error loss with respect to $U(0.25, 0.55)$ of the posterior predictive distributions generated from the randomized relative to that of the Polya tree for various sample sizes N and parameter values c and τ. Column 4: Overall comparison on $(0, 1]$. Column 5: Comparison on $(0.25, 0.50]$. Column 6: Comparison on $(0.50, 0.55)$.

	(1)	(2)	(3)	(4) ISEL(RT)/ ISEL(PT)	(5) ISEL(RT)/ ISEL(PT)	(6) ISEL(RT)/ ISEL(PT)
Set	N	c	τ	overall	on (0.25,0.50]	on (0.50,0.55)
	50	0.10	0.010	1.019	1.016	1.033
	50	0.10	0.025	0.981	0.971	0.875
(a)	50	0.10	0.050	1.095	1.115	1.382
	50	0.10	0.100	1.359	1.115	1.351
	50	0.10	0.250	42.202	20.307	92.800
	50	1.00	0.010	0.945	1.066	0.788
	50	1.00	0.025	0.885	1.278	0.512
(b)	50	1.00	0.050	0.808	1.957	0.181
	50	1.00	0.100	1.081	2.526	0.114
	50	1.00	0.250	3.822	6.184	2.651
	100	0.10	0.010	1.028	1.016	1.036
	100	0.10	0.025	1.064	0.999	1.192
(c)	100	0.10	0.050	1.580	1.432	3.250
	100	0.10	0.100	1.946	1.376	2.580
	100	0.10	0.250	85.665	39.573	269.955
	100	1.00	0.010	0.960	1.088	0.794
	100	1.00	0.025	0.863	1.331	0.460
(d)	100	1.00	0.050	0.906	2.388	0.391
	100	1.00	0.100	1.165	3.150	0.124
	100	1.00	0.250	7.671	9.289	9.056
	500	1.00	0.010	1.012	1.057	0.858
	500	1.00	0.025	1.264	1.560	0.887
(e)	500	1.00	0.050	2.293	3.597	2.777
	500	1.00	0.100	2.880	4.742	0.506
	500	1.00	0.250	152.207	155.649	188.209
	1000	10.00	0.010	0.919	1.194	0.722
	1000	10.00	0.025	0.890	1.938	0.412
(f)	1000	10.00	0.050	0.984	4.553	0.477
	1000	10.00	0.100	1.301	5.642	0.132
	1000	10.00	0.250	7.834	22.050	7.052

We perform two randomized tree analyses, one with the value of τ fixed at 0.01 and another with $\tau = 0.05$. For both analyses, the prior parameters are set to $\alpha_{\epsilon_{1:j}} = 0.01j^2$ $(j = 1, \ldots, 10)$ and G_h is a uniform on the range of the

variables on the h^{th} axis of the four-dimensional hypercube ($h = 1, \ldots, 4$); we select this "flattening" prior (Clogg et al. (1991), Paddock (2002)) in the absence of having prior information about the modality of the distribution. The randomized tree prior is centered about the Polya tree prior, with the partition created by implementing the canonical partition along each axis of the hypercube. For our MCMC sampling, 5000 draws are retained for analysis following a burn-in of 10000 iterations. MCMC diagnostic procedures such as analysis of autocorrelations and cross-correlations of MCMC output, visual inspection of trace plots, and some diagnostics (not shown) from the BOA (Smith (1999)) or CODA (Best, Cowles, and Vines (1995)) packages indicate no troubles with convergence. The acceptance rates for the β_i, which are sampled by an independence Metropolis step, are about $50-77\%$, with almost all of the β_i having acceptance rates around $65 - 75\%$.

The first column of Figure 3 displays four of the 2-D scatterplots of the data, the second column displays draws from the corresponding bivariate marginals using a non-Randomized Polya tree, column 3 displays the bivariate marginals obtained with the randomized tree ($\tau = 0.01$), and column 4 displays the same for $\tau = 0.05$. Examination of these scatterplots of draws from the bivariate marginal distributions of $F(\cdot)$ highlight the effect of τ. Partition artifacts can clearly be seen as "blocks" in the plots of the Polya tree prior (column 2). While the choice of $\tau = 0.01$ (column 3) induces some smoothing of these artifacts, the use of $\tau = 0.05$ (column 4) induces much more. This comparison is even more interesting in light of the 2-D marginal plots of the data (column 1). Take for instance the bivariate marginal distribution of the cubed root of ozone and radiation; given the data and the continuous nature of the variables ozone and radiation, it would be fair to conclude that the sudden jumps in the probability at the cube root of ozone around ≈ 2.5, as well as that at radiation ≈ 175, are due in large part to the partitioning. Similar conclusions can be drawn from comparing the other bivariate marginal plots.

Figure 4 shows a subset of the results of the conditional predictive simulation of the cube root of ozone given various values of wind speed, temperature and radiation. The smoothing effect that is clear from Figure 3 is also clear from this figure. The effect of τ on the results is apparent by comparing the corresponding histograms for $\tau = 0.01$ and $\tau = 0.05$; the larger value of τ smooths out some of the jumps in the simulated conditional predictive distributions that are attributable to the partition.