# Chapter 4

# Pólya Trees

### 4.1. Definition

An intuitively attractive way to construct RPMs is as a random histogram, with a fixed set of bins and random probability mass associated with each bin (for example, see Loredo, 2011). In anticipation of the upcoming discussion, we assume that the bins define $2^m$ partitioning subsets and index the subsets by an $m$-digit binary number $\epsilon = e_1 \cdots e_m$, $e_j \in \{0, 1\}$. Let $\Pi_m = \{B_\epsilon,\ \epsilon = e_1 e_2 \cdots e_m\}$ denote this partition of the sample space into $2^m$ bins or partitioning subsets. A random histogram could define an RPM $G$ by defining the joint distribution $p(G(B_\epsilon);\ \ B_\epsilon \in \Pi_m)$.

Pólya trees arise as an extension of this idea where the size of the bins is made sequentially smaller. More specifically, consider sequentially refining a partition $\Pi_m$ to $\Pi_{m+1}$ by splitting $B_{e_1 \cdots e_m}$ into $B_{e_1 \cdots e_m} = B_{e_1 \cdots e_m 0} \cup B_{e_1 \cdots e_m 1}$ (see Figure 4.1.). The problem now arises to define $G(B_\epsilon)$ coherently across nested partitions, with $G(B_\epsilon) = G(B_{\epsilon 0}) + G(B_{\epsilon 1})$. The elegant solution is to define the random $G(B_\epsilon)$ through a sequence of conditional probabilities as

$$G(B_\epsilon) = \prod_{k=1}^{m} G(B_{e_1 \cdots e_{k-1} e_k} \mid B_{e_1 \cdots e_{k-1}}),$$

with the understanding that $B_\emptyset$ denotes the entire sample space. An RPM $p(G)$ is then defined by specifying a prior for the random splitting probabilities $Y_\epsilon = G(B_{\epsilon 0} \mid B_\epsilon)$ for any $m$-digit index $\epsilon = e_1 \cdots e_m$, $m > 0$.

The resulting construction is called a Pólya tree (PT) prior (Lavine, 1992, 1994; Mauldin *et al.*, 1992). The PT model assumes that $Y_\epsilon \sim \mathsf{Beta}(a_{\epsilon 0}, a_{\epsilon 1})$, independently across $m$. In general, an RPM with independent splitting probabilities $Y_\epsilon$ is known as tail-free with respect to $\Pi$. The PT can thus be characterized as a tail-free process with respect to a nested partition sequence $\Pi$ and beta distributed random splitting probabilities.

Notice the similarities between the PT prior the class of neutral to the right (NTR) priors introduced in §1.2.7. Recall that NTR refers to independence of the normalized increments $G(t_{i-1}, t_i]/G(-\infty, t_i]$ for a partition with partition boundaries $t_0 = -\infty < t_1 \ldots < t_n < \infty$. In contrast, the tailfree property of the PT refers to independence across two sets of partitions.

In summary, the PT prior defines an RPM by assigning any partitioning subset $B_\epsilon$ the random probability

$$G(B_{e_1 \cdots e_m}) = \underbrace{\prod_{j=1; e_j=0}^{m} Y_{e_1 \cdots e_{j-1} 0}}_{\text{all } \textit{left} \text{ splits}} \underbrace{\prod_{j=1; e_j=1}^{m} (1 - Y_{e_1 \cdots e_{j-1} 0})}_{\text{all } \textit{right} \text{ splits}},$$

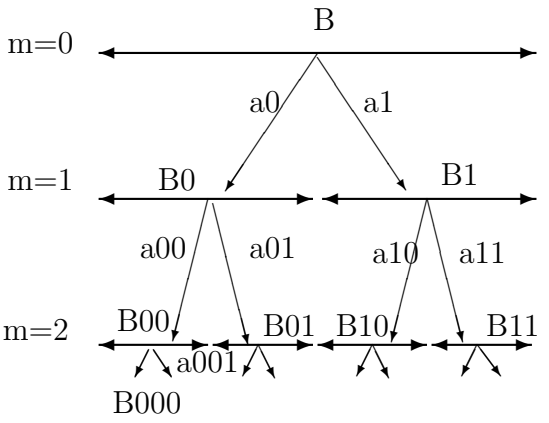


FIG 4.1. *PT: The diagram shows the nested sequence of partitions* $\Pi_m = \{B_\epsilon,\ \epsilon = e_1 \cdots e_m\}$ *with* $e_j \in \{0,1\}$. *The PT is defined by random splitting probabilities* $Y_{\epsilon 0} = G(B_{e_1 \cdots e_m 0} \mid B_{e_1 \cdots e_m})$ *with* $Y_{\epsilon 0} \sim \mathsf{Beta}(a_{e_1 \cdots e_m 0}, a_{e_1 \cdots e_m 1})$.

with independent beta priors $Y_{\epsilon 0} \sim \mathsf{Beta}(a_{\epsilon 0}, a_{\epsilon 1})$. The PT model is indexed with two sets of parameters, the nested sequence of partitions $\Pi = \{\Pi_m\}$ and the parameters $\mathcal{A} = \{a_\epsilon\}$ for the beta-distributed random splitting probabilities. Hence, we write $G \sim \mathsf{PT}(\Pi, \mathcal{A})$.

One of the important features of the PT prior is that it can generate continuous probability measures. A random probability measures $G \sim \mathsf{PT}(\mathcal{A}, \Pi)$ is absolutely continous with probability 1 when the $\alpha_{e_1 \cdot e_m}$ parameters increase suffiently fast with $m$. A popular choice is $\alpha_{\epsilon_1 \ldots \epsilon_m} = c\, m^2$. On the other hand, for decreasing $\alpha$, the random probability measure can also be almost surely discrete. For example, for $\alpha_{e_1 \cdots e_m} = c/2^m$ the PT prior reduces to the special case of the DP prior.

One of the attractions of the PT model is the ease of centering the model at any desired prior mean $G_0$. One way to accomplish this centering is to fix the partitioning subsets $B_\epsilon$ as the dyadic quantiles of $G_0$. More specifically, let $z_\epsilon = \sum_{j=1}^m 2^{-e_j}$ and define $B_{e_1 \cdots c_m} = (G_0^{-1}(z_\epsilon), G_0^{-1}(z_\epsilon + 2^{-m})]$. At $m = 1$, the two subsets $\{B_0, B_1\}$ are simply below and above the median of $G_0$, at $m = 2$, the partitioning subsets are determined by the quartiles, etc. If the $a_\epsilon$ parameters are chosen to be symmetric, $a_{\epsilon 0} = a_{\epsilon 1}$, then it is easy to show that $E(G(B)) = G_0(B)$, i.e., the RPM is centered at $G_0$, as desired. Alternatively, the same centering can be achieved with an arbitrary nested partition sequence $\Pi$ by taking $a_\epsilon = c_m G_0(B_\epsilon)$ for some sequence $(c_m)$ (for example, $c_m = m^2$). This second method of prior centering might be preferable when $G_0 = G_{0,\eta}$ includes some unknown hyper-parameters $\eta$. It would be computationally awkward if one had to change the partitioning sequence each time a different value of $\eta$ is being considered and, in most implementations of posterior inference, it is easier to change the parameters $a_\epsilon$. By a slight abuse of notation we write $\mathsf{PT}(\mathcal{A}, G_0)$ to indicate a PT prior with partitioning subsets determined to achieve a desired prior mean $G_0$, and similarly we write $\mathsf{PT}(G_0, \Pi)$ for a PT prior with the beta parameters chosen to match a desired prior centering.

Figure 4.9 shows some random realizations from a finite PT prior. More specifically, we plot the random probabilities $G(B_{e_1 e_2})$ up to level $m = 2$ under a PT prior $G \sim \mathsf{PT}$. The plot highlights that realizations from a PT prior are discontinous at the partition boundaries, a feature that is often considered as a limitation of this class of priors. However, the effect of the discrete partition boundaries can trivially

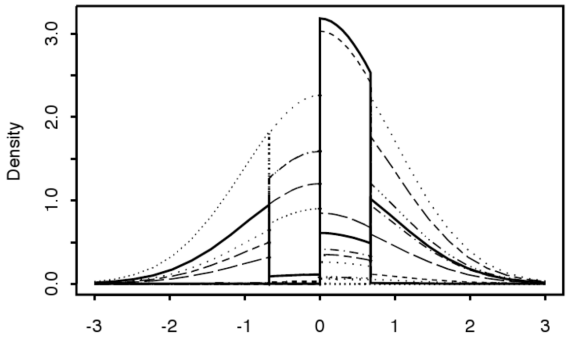


FIG 4.2. *Prior simulation of 10 random realizations* $G \sim \mathsf{PT}$ *with a standard Gaussian centering measure. Note the discontinuities at the boundaries of the partitioning sets.*

be removed. The most commonly used approach is to add a mixture with respect to $\Pi$. We will return to this problem §4.4.

For later reference we note that the split at each level of the nested partition sequence need not be binary. In general, each partitioning subset $B_\epsilon$ could be split into $q$ subsets $B_{e_1 \cdots e_{m-1}} = \biguplus_{e_m=0}^{q-1} B_{e_1 \cdots e_{m-1} e_m}$, at the next level of the partitioning sequence. The digits $e_j$ of the index $\epsilon$ are $e_j \in \{0, \ldots, q-1\}$ and the beta prior for the random splitting probability $G(B_{\epsilon 0} \mid B_\epsilon)$ is replaced by a Dirichlet distribution for the $q-$way splitting probabilities $(G(B_{\epsilon 0}), \ldots, G(B_{\epsilon,q-1}) \mid G(B_\epsilon)) \sim \mathsf{Dir}(\alpha_{\epsilon 0}, \ldots, \alpha_{\epsilon,q-1})$.

## 4.2. Posterior Inference

The PT is conjugate under i.i.d. sampling. Assume $x|G \sim G$ with a PT prior, $G \sim \mathsf{PT}(\mathcal{B}, \mathcal{A})$. Then the posterior on the unknown probability measure $G$ is again a PT, $(G|x) \sim \mathsf{PT}(\mathcal{B}, \mathcal{A}^*)$ with

$$\alpha_\epsilon^* = \begin{cases} \alpha_\epsilon + 1 & \text{if } x \in B_\epsilon \\ \alpha_\epsilon & \text{otherwise.} \end{cases} \tag{4.1}$$

The $\alpha_\epsilon$ parameters corresponding to the partitioning subsets are incremented by one for each subset $B_\epsilon$ that contains $x$. In practice, if a finite tree with $T$ levels is used, (typically with $T \approx 7$) equation (4.1) allows for straightforward posterior updating.

Figure 4.3 shows a simple example of posterior updating for a PT prior. The prior model in the example is centered at a standard normal. The figure shows posterior inference conditional on observed data. Posterior updating for censored data introduces no additional difficulties if the partition boundaries are chosen to match the censoring times (Muliere and Walker, 1997).

The nature of the posterior PT also leads to straightforward posterior predictive simulation. To draw a new, future observation $x_{n+1}$ from $G$ conditional on observed data, $x_1, \ldots, x_n$ also drawn from $G$ we only need to follow the posterior updating over finitely many levels. First generate an indicator $e_1 = I(x_{n+1} \in B_0)$ to determine whether the new observation $x_{n+1}$ falls into $B_0$. To determine $e_1$ we generate $y_0 = G(B_0)$. The earlier discussion of the posterior process implies $p(y_0 \mid x_1, \ldots, x_n) = \mathsf{Beta}(\alpha_0^\star, \alpha_1^\star)$. Next let $e_2 = I(x_{n+1} \in B_{e_1 0})$. Again, $p(e_2 = 1 \mid$

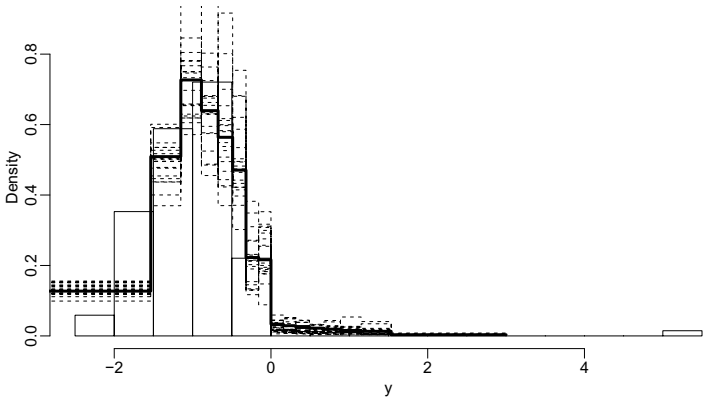


FIG 4.3. *Data (histogram), posterior mean $E(G \mid \boldsymbol{y})$ (thick line), random posterior draws $G \sim p(G \mid \boldsymbol{y})$ (dashed lines). The prior mean was $G_0 = \mathsf{N}(0, 1)$.*

$e_1, G) = y_{e_1 0}$ and $p(y_{e_1 0} \mid x_1, \ldots, x_n) = \mathsf{Beta}(\alpha^\star_{e_1 0}, \alpha^\star_{e_1 1})$, etc. The iteration ends when it reaches the first level $m$ with $\alpha_{e_1 \cdots e_m} = \alpha^\star_{e_1 \cdots e_m}$, i.e., when the new draw is imputed to fall within a partitioning subinterval without earlier data. At that moment we simply generate $x_{n+1}$ from the prior mean $G_0 = E(G)$ restricted to this subset, $x_{n+1} \sim G_0 I(x_{n+1} \in B_{e_1 \cdots e_m})$.

The described process of generating from the predictive distribution $p(x_{n+1} \mid x_1, \ldots, x_n)$ is beautifully illustrated by the following special case. Consider $n = 0$, i.e., marginal simulation for the first observation, assume that the sample space is the unit interval $B = [0, 1]$, the partition boundaries are the dyadic subintervales $[0, 1/2), [1/2, 1], [0, 1/4), [1/4, 1/2), \ldots$, and the centering measure is the uniform distribution. In that case the indicators $e_j$ are simply the digits of $x_{n+1}$ in a dyadic expansion and the process amounts to iteratively generating its digits.

**Example 12 (A Survival Model with a Longitudinal Covariate)** *Zhang* et al. *(2010) report a typical application of PT models in survival analysis. They discuss inference for data from a phase III trial of androgen ablation (AA) vs. chemohormonal (CH) therapy for patients with metastatic prostate cancer. Patients joined the trial with very diverse prior treatment histories, giving rise to a challenging statistical inference problem. The study enrolled $n = 286$ subjects, randomized to the two arms with $n_0 = 137$ patients assigned to the AA arm, and $n_1 = 149$ patients assigned to CH. The primary endpoint is time to progression (TTP) to androgen independent prostate cancer.*

*An important covariate for TTP is the change of prostate specific antigen (PSA) over time. Let $\boldsymbol{y}_i$ denote the longitudinal trajectory of PSA measurements over time for patient $i$, let $T_i$ denote the TTP, and let $x_i \in \{0, 1\}$ denote an indicator for assignment to AA (0) or CH (1). Figure 4.4 shows the data as Kaplan Meier plots arranged by treatment allocation.*

*We construct a joint probability model for $\boldsymbol{y}_i$ and $T_i$ for a patient in treatment group $x_i$ as a marginal model $G_{x_i}(T_i)$ and a conditional model $p(\boldsymbol{y}_i \mid T_i, x_i)$. This unusual factorization into a marginal model for TTP and a conditional model for the longitudinal covariate given TTP makes it easy to go nonparametric on the event time model. We assume $G_x \sim \mathsf{PT}(\mathcal{A}, \Pi)$, independently for $x = 0, 1$. Conditional on $T_i$ the model for $\boldsymbol{y}_i$ is a non-linear regression. Details of the regression mean*

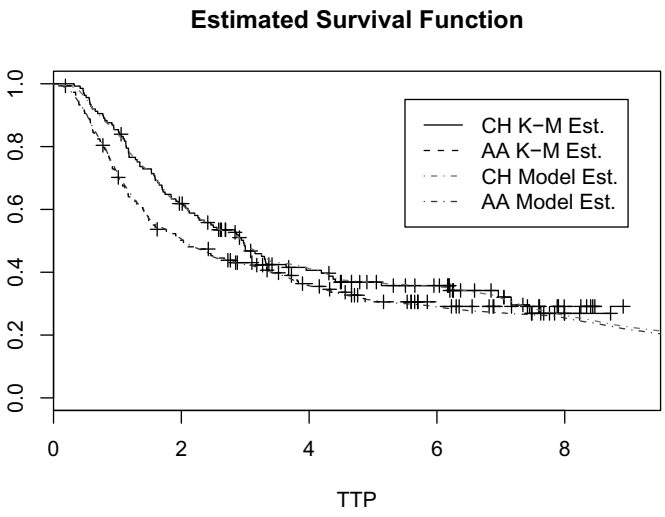


FIG 4.4. *Prostate cancer trial. Kaplan Meier plot of the data, arranged by the two treatment arms. The dot-dashed lines show the model based inference.*

*function are motivated by the typical PSA profiles and the nature of the intervention. See Zhang* et al. *(2010) for details.*

*Figure 4.5a shows the estimated distributions of TTP under the two treatments. For comparison, Figure 4.5b shows the same inference using two independent Weibull models for* $G_x$, $x = 0, 1$.

*An important feature of density estimation with nonparametric Bayesian models is the coherent nature of inference as a posterior probability model on the unknown probability measure. This enables us to report uncertainties on any event or summary of interest. For example, Figure 4.6 shows the implied uncertainty on the hazard function.*

### 4.3. The Marginal Model

The PT prior allows a closed form expression for the marginal distribution $p(x_1, \ldots, x_n)$ of a random sample $x_i \sim G$, i.i.d., under $G \sim \mathsf{PT}(\cdot)$. The PT shares this practically very useful property with the DP. For the DP prior the marginal model is determined by the Pólya urn in (7.1).

Lavine (1992) shows the expression of the marginal model for the PT prior. Let $\epsilon_m(x_i) = e_1 \ldots e_m$ denote the index of the level $m$ subset that contains $x_i$, i.e., $x_i \in B_{e_1 \cdot e_m}$. Also, let $m^*(x_i)$ denote the lowest level $m$ such that $x_i$ is the only data point in $B_{\epsilon_m(x_i)}$. Formally, $m^\star(x_i) = \min_m\{x_j \notin B_{\epsilon_m(x_i)},\ j \neq i\}$. Then

$$p(x_1, \ldots, x_n \mid \eta) = \prod_{i=1}^{n} G_0(x_i \mid \eta) \prod_{j=2}^{n} \prod_{m=1}^{m^*(x_j)} \frac{\alpha^\star_{\epsilon_m(x_j)}}{\alpha_{\epsilon_m(x_j)}} \cdot \frac{\alpha_{\epsilon_{m-1}0(x_j)} + \alpha_{\epsilon_{m-1}1(x_j)}}{\alpha^\star_{\epsilon_{m-1}0(x_j)} + \alpha^\star_{\epsilon_{m-1}1(x_j)}}.$$

See Berger and Guglielmi (2001) and Hanson and Johnson (2002) for more discussion. In particular, Berger and Guglielmi (2001) use the marginal model to evaluate Bayes factors.

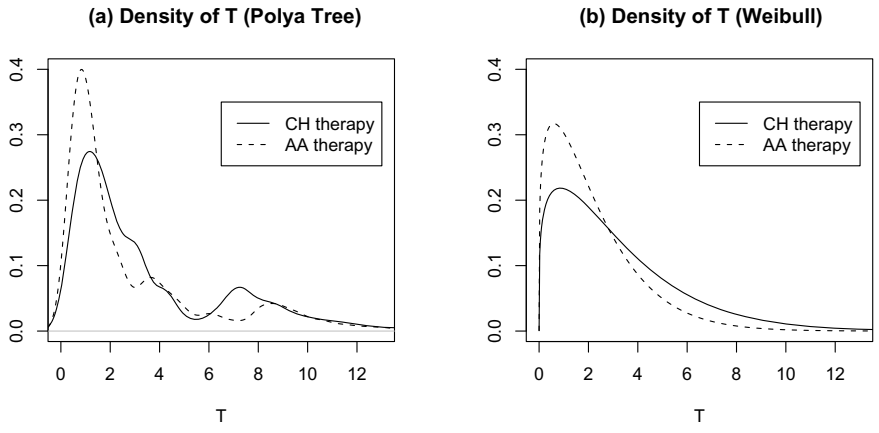


FIG 4.5. *Prostate cancer trial. Estimated $G_x(T)$ using PT (left) and Weibull (right).*

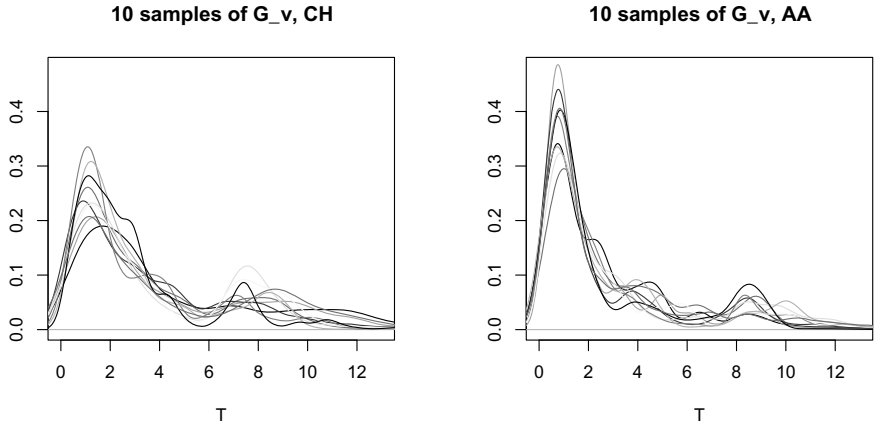


FIG 4.6. *Prostate cancer trial. Uncertainty $p(G(\cdot) \mid \boldsymbol{y})$ using PT (left) and Weibull (right).*

## 4.4. Mixtures of Pólya Trees

Figure 4.3 highlights a critical limitation of the PT prior, i.e., posterior draws $G \sim p(G \mid x_1, \ldots, x_m)$ as well as the posterior mean $\bar{G} = E(G \mid x_1, \ldots, x_m)$ show visible discontinuities at the partition boundaries. This sensitivity of posterior inference to the chosen partition is undesirable in most data analyses.

One possible fix is to consider PTs with random centering measures. Recall the discussion of the DP prior; a random probability measure which is assigned a DP prior, $G \sim \mathsf{DP}$, is almost surely discrete. The discrete nature of the DP greatly simplified many of the computational details of posterior simulation but is unappealing for most applications. In that context, we mitigated concerns related to the discrete nature of a DP random measure by convoluting $G$ with an additional smooth kernel to define DP mixture models. Similarly, we can mitigate the undesirable sensitivity of the PT prior to partition boundaries by introducing additional mixing with respect to the centering measure. Assume that the PT prior is centered by defining the nested partition sequence $\Pi$ to be determined by dyadic quantiles of

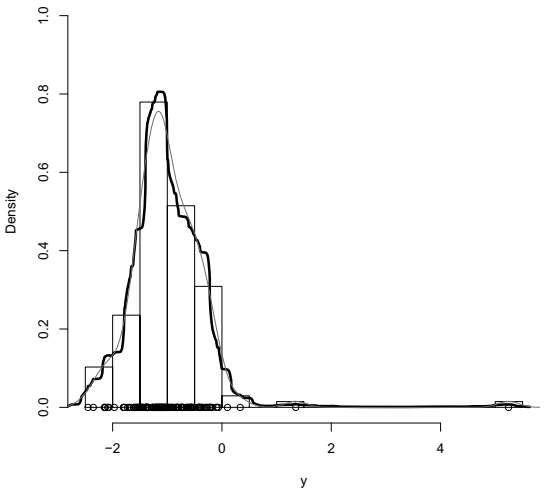


FIG 4.7. *Same data as in Figure 4.3, now with PT mixture (black). For comparison a kernel density estimate (light grey).*

a desired prior mean $G_0 = E(G)$. Hanson and Johnson (2002) propose to introduce additional hyperparameters $\eta$ to index the centering model $G_{0,\eta}$ and extend the model with a hyperprior on $\eta$.

$$G \mid \eta \sim \mathsf{PT}(\mathcal{A}, G_{0,\eta}), \qquad \eta \sim p(\eta). \tag{4.2}$$

Mixing with respect to a hyperprior $p(\eta)$ smoothes out the undesired dependence on partition boundaries that appears in Figure 4.3. We refer to model (4.2) as a mixture of PT model. Note the difference to the DP mixture model that defined a mixture of a kernel with respect to a DP prior; hence, a mixture of PT models is analogous to the MDP model discussed in §3.5. Figure 4.7 shows how posterior inference is improved under a PT mixture prior.

## 4.5. Multivariate Pólya Trees

The definition of the PT prior is general, all we needed was a nested sequence of partitions and the beta priors for the random splitting probabilities. When $B = \mathbb{R}$, the partitions can be described by partition boundaries and can be naturally indexed by sequences of binary indicators for left versus right splits.

For higher dimensional sample spaces, it becomes awkward to specify and keep track of the nested partition sequence. For $B = \mathbb{R}^p$ this task becomes challenging, but not impossible. Jara *et al.* (2009) propose a possible construction that remains feasible even for moderately high dimensions, $p = 8$ and beyond. The construction works with a multivariate normal centering measure, $G_0 = \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ where $\Sigma = UU'$, along with a split of each partitioning subset into $2^p$ partitioning subsets. More specifically, the partitioning subsets $B_\epsilon$ are indexed with sequences of base $2^p$ digits, i.e., $\epsilon = e_1 \cdot e_m$, $e_j \in \{0, \ldots, 2^p - 1\}$. For example, in Figure 4.1, instead of splitting each $B_\epsilon$ into two daughters, each partitioning subset $B_\epsilon$ is split into $2^p$ nested sets. The definition of these partitioning subsets $B_\epsilon$ starts with $p$-dimensional rectangles defined by standard normal dyadic quantiles. Let $B_0(m, k)$ denote the

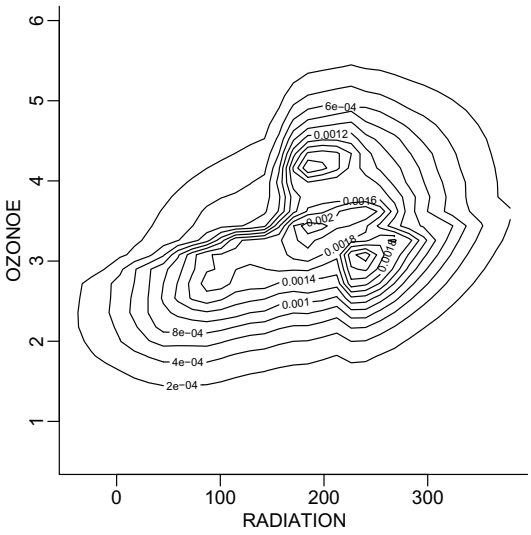


FIG 4.8. *Airquality data. Bivariate density estimate, using the R function* `PTdensity(·)` *from the R package* `DPpackage`. *The air quality data is available in R.*

$k$-th of $2^m$ dyadic (univariate) standard normal quantiles at level $m$. At level $m = 1$ the two subsets are defined by the partition boundary at 0, the standard normal median. At $m = 2$, the four subsets are defined by the partition boundaries at the 1/4, 1/2 and 3/4 quantiles, etc. Next define the $p$-dimensional product sets $\boldsymbol{B}_0(m, \boldsymbol{k}) = B_0(j, k_1) \times \ldots \times B_0(j, k_p)$. Finally, the partitioning subsets for the nested partition sequence $\Pi$ are defined by an affine transformation $\boldsymbol{B}(m, \boldsymbol{k}) = \{\boldsymbol{\mu} + Uz;\ z \in \boldsymbol{B}_0(j, \boldsymbol{k})\}$.

The construction was easily explained, but the reader might be reluctant to venture into an implementation. Fortunately inference for the multivariate PT is implemented in `DPpackage`. Figure 4.8 shows an example of output from the function `PTdensity(·)`.

## 4.6. Rubbery Pólya Tree

Recall that an RPM $G$ with PT prior includes discontinuities at the partition boundaries. These discontinuities can clearly be seen in Figure 4.9, and they persist in the posterior means. See, for example, see 4.3. This awkward property limits the use of the PT prior for many data analysis problems. In §4.4 we discussed a construction that can mitigate this awkward feature of the PT prior by adding uncertainty about the centering measure $G_{0\eta}$. Mixing over $\eta$ smears out the partition boundaries. Alternatively, Paddock *et al.* (2003) introduce additional randomness in the model by jittering the cutoff points in a dyadic nested partition; the discontinuities in the RPM are then removed by averaging with respect to this additional jittering.

A third approach that directly addresses the cause of the discontinuities in the PT mode is the rubbery PT introduced by Nieto-Barajas and Müller (2012). Recall the independent random splitting probabilities

$$Y_{\epsilon 0} = G(B_{\epsilon 0} \mid B_\epsilon).$$

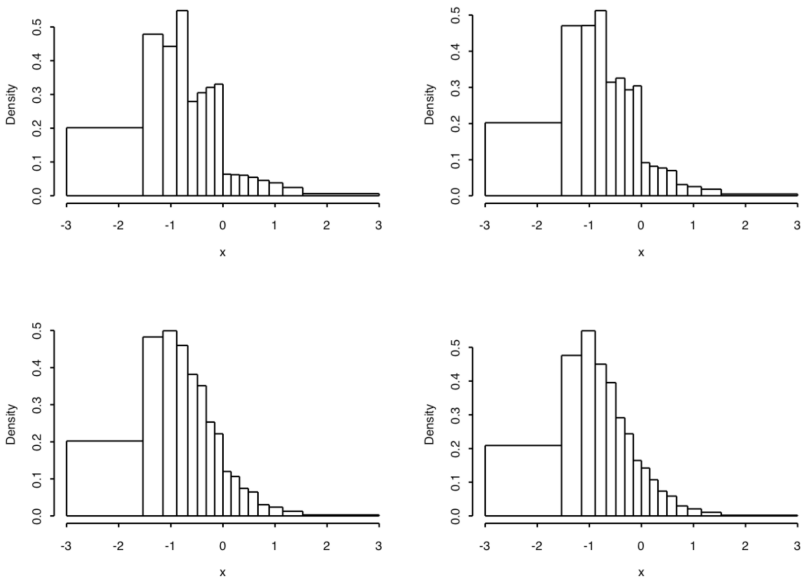


FIG 4.9. *Posterior predictive distributions for a rPT with a sample of size 1 at $x = -2$: Top left $\delta = 0$, top right $\delta = 1$, bottom left $\delta = 5$ and bottom right $\delta = 10$. The posterior predictive is identical to the posterior mean, $p(x_2 \mid x_1 = x) = E(G \mid x)$.*

The independence of these $Y_\epsilon$, across $\epsilon$ is the source of the discontinuities in $G$ at the partition boundaries; by introducing dependence among these probabilities it is possible to eliminate the discontinuities.

The construction of the rubbery PT is more easily explained in the context of a finite PT. For ease of exposition assume a $PT_2$ prior with only two levels and use decimal integers to index the partitioning subsets at each level:

$$\begin{array}{ccc} B_{11} & | & B_{12} \\ B_{21} \mid B_{22} & | & B_{23} \mid B_{24} \end{array} \tag{4.3}$$

A computationally easy way to introduce dependence between the two random probabilities $Y_{21}$ and $Y_{23}$ while leaving the marginal beta distribution unchanged is the use of a latent binomial variables $Z_{21}$ sandwiched between them. We leave the marginal distribution of $Y_{21}$ unchanged as $Y_{21} \sim \mathsf{Beta}(\alpha_2, \alpha_2)$ and $Y_{22} = 1 - Y_{21}$. The prior for $Y_{23}$ is changed to

$$Z_{21} \mid Y_{21} \sim \mathsf{Bin}(\delta_{21}, Y_{21}), \qquad Y_{23} \mid Z_{21} \sim \mathsf{Beta}(\alpha_2 + Z_{21}, \alpha_2 + \delta_{21} - Z_{21}),$$

and $Y_{24} = 1 - Y_{23}$. It is easily verified that the implied marginal prior $p(Y_{23})$ remained unchanged as a $\mathsf{Beta}(\alpha_2, \alpha_2)$ while introducing the desired dependence of $Y_{21}$ and $Y_{23}$. Also, the level 1 priors remain unchanged. The Binomial sample size parameter $\delta_{21}$ tunes the desired level of smoothing; large values imply more smoothing. We use $\mathsf{rPT}(\Pi, \mathcal{A}, \delta)$ to denote a rubbery PT with a sequence of latent binomial variables indexed by $\delta_{mj}$, $m > 1$, $j = 1, 3, \ldots$. Figure 4.9 shows posterior predictive distributions for a future observation for different choices of $\delta$.