

Optional Pólya Trees and Bayesian Inference

Wing H. Wong and Li Ma 2010

Presented by Benedetta Bruni

16/11/2021

Pólya Trees Definition

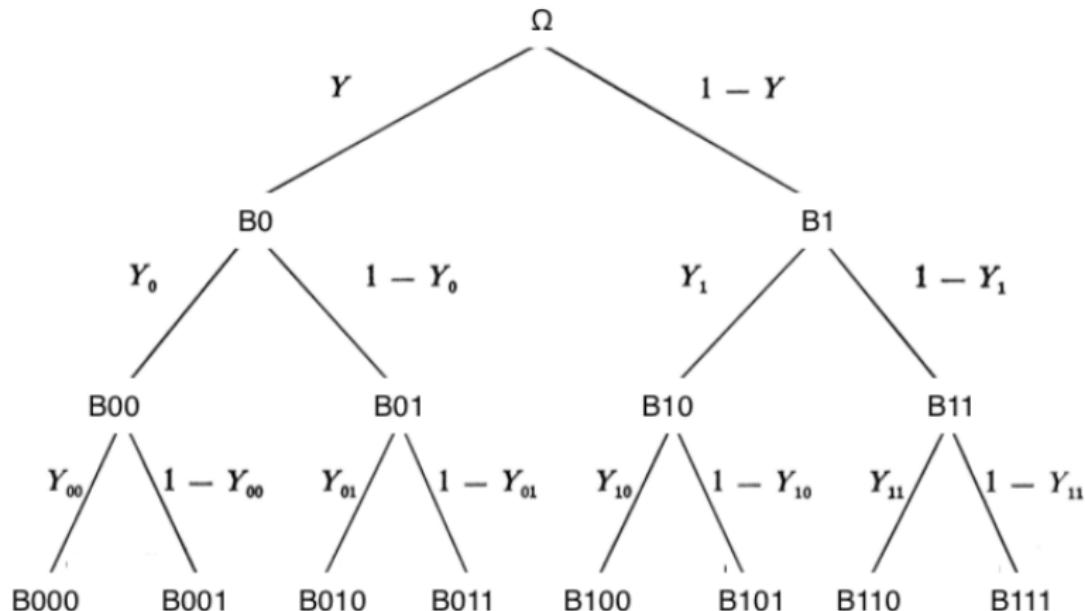
Definition

A random probability measure \mathcal{P} on Ω is said to have a Polya tree distribution, or a Polya tree prior, with parameter (Π, \mathcal{A}) , written $\mathcal{P} \sim PT(\Pi, \mathcal{A})$, if there exist nonnegative numbers $\mathcal{A} = \{\alpha_\epsilon : \epsilon \in E^*\}$ and random variables $\mathcal{Y} = \{Y_\epsilon : \epsilon \in E^*\}$ such that the following hold:

- (i) all the random variables in \mathcal{Y} are independent;
- (ii) for every $\epsilon \in E^*$, Y_ϵ has a beta distribution with parameters $\alpha_{\epsilon 0}$ and $\alpha_{\epsilon 1}$;
- (iii) for every $m = 1, 2, \dots$ and every $\epsilon \in E^m$,

$$\mathcal{P}(B_{\epsilon_1, \dots, \epsilon_m}) = \prod_{j=1; \epsilon_j=0}^m Y_{\epsilon_1 \dots \epsilon_{j-1}} \prod_{j=1; \epsilon_j=1}^m (1 - Y_{\epsilon_1 \dots \epsilon_{j-1}})$$

Setup



Why Optional Pólya Trees?

CONFLICT BETWEEN SMOOTHNESS AND

FAITHFULNESS: When constructing Polya Trees the choice of α_ϵ is relevant.

- The α_ϵ control how quickly the updated predictive distribution moves from the prior predictive distribution to the sample distribution.
- The α_ϵ s influence the smoothness of \mathcal{P} : the alphas should increase "rapidly" with respect to the number of partitioning steps m for the prior to generate absolutely continuous measures.

Pólya Trees and Optional Pólya Trees

- Pólya Trees: The choice of large parameters for the Betas constrain the allocation of conditional probability to represent faithfully the data distribution within small intervals;
- **Optional Stopping:** does not require large increase in Beta parameters;
- **Randomized Partitioning:** partitioning: the posterior will give more weights to partitions that better fit the data.

SETUP

- Space (Ω, μ) .
- Ω finite and μ counting measure;
- Ω bounded rectangle \mathbb{R}^p and μ Lebesgue measure.
- For any level- k elementary region A , $M(A)$ ways to partition it: for $j = 1, 2, \dots, M(A)$,

$$A = \bigcup_{k=1}^{K^j(A)} A_k^j$$

- \mathcal{A}^k all possible level- k elementary regions; $\mathcal{A}^{(k)} = \bigcup_{l=1}^k \mathcal{A}^l$

Recursive Partitioning Process

- Recursive partition of depth k : $J^{(k)} = (J_1, J_2, \dots, J_k)$;
- $\Omega = T_0^k \cup T_1^k$ with $T_0^k = \bigcup_{i=1}^I A_i$ (with $A_i \in \mathcal{A}^{(k-1)}$ disjoint) and $T_1^k = \bigcup_{i=1}^{I'} A'_i$ (with $A'_i \in \mathcal{A}^k$ disjoint);
- Random probability measure $Q^{(k)}$ on Ω uniformly distributed within each region in T_0^k and T_1^k .

Step $(k+1)$:

- For each elementary region A in T_1^k , generate $S \sim \text{Bernoulli}(\rho)$. If $S = 1$ stop partitioning A ;
- If $S = 0$, draw a partition of A : draw $J \in \{1, 2, \dots, M(A)\}$ according to $\lambda(A) = (\lambda_1, \dots, \lambda_{M(A)})$ with $P(J=j) = \lambda_j$ and $\sum_{l=1}^{M(A)} \lambda_l = 1$,

Recursive Partitioning Process

- If $J = j$, $A = \bigcup_{l=1}^K A_l^j$;
- Set $Q^{(k+1)}(A_l^j) = Q^{(k)}(A)\theta_l^j$ where $\theta^j = (\theta_1^j, \dots, \theta_k^j)$ is generated from a Dirichlet distribution with parameter $(\alpha_1^j, \dots, \alpha_k^j)$;
- $\Omega = T_0^{k+1} \cup T_1^{k+1}$

New measure $Q^{(k+1)}$:

- For $B \subset T_0^{(k+1)}$, $Q^{(k+1)}(B) = Q^{(k)}(B)$;
- For $B \subset T_1^{(k+1)}$ with T_1^{k+1} partitioned as $T_1^{k+1} = \bigcup_{i=1}^{I'} A'_i$ (with $A'_i \in \mathcal{A}^{k+1}$ disjoint), set

$$Q^{(k+1)}(B) = \sum_{i=1}^J Q^{(k+1)}(A_i) \left(\frac{\mu(A_i \cap B)}{\mu(A_i)} \right)$$

Definition of Optional Pólya Trees

Theorem 1

Suppose there is a $\delta > 0$ such that with probability 1,
 $1 - \delta > \rho(A) > \delta$ for any region A generated during any step in
the recursive partitioning process. Then with probability 1, $Q^{(k)}$
converges in variational distance to a probability measure Q that is
absolutely continuous with respect to μ .

Definition

The random probability measure Q defined in Theorem 1 is said to
have an Optional Pólya tree distribution with parameters λ
(selection probabilities vectors), α (assignment-weight vectors) and
stopping rule ρ .

Optional Pólya Trees - Property

Theorem 2

Let Ω be a bounded rectangle in \mathbb{R}^P . Suppose that the condition of Theorem 1 holds and that the selection probabilities $\lambda_i(A)$, the assignment probabilities $\alpha_i^j(A) / \sum_j(\alpha_i^j(A))$, for all i, j and $A \in \mathcal{A}^{(\infty)}$ are uniformly bounded away from 0 and 1. Let $q = dQ/d\mu$; then for any density f and any $\tau > 0$, we have

$$P\left(\int |q(x) - f(x)|d\mu > \tau\right) > 0.$$

Bayesian Inference with Optional Pólya Tree prior

- Suppose we have observed $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ with x_i independent draws from a probability measure Q with Optional Pólya Tree distribution;
- Call $\pi()$ the prior for $q = \frac{\partial Q}{\partial \mu}$;
- For any $A \in \Omega$ define $\mathbf{x}(A) = \{x_i \in \mathbf{x} : x_i \in A\}$;
- Let $q(x) = \frac{\partial Q}{\partial \mu}(x)$ for $x \in \Omega$ and $q(x|A) = \frac{q(x)}{Q(A)}$ for $x \in A$;

The likelihood is given by:

$$P(\mathbf{x}|Q) = \prod_{i=1}^n q(x_i) = q(\mathbf{x})$$

Bayesian Inference with Optional Pólya Tree prior

Assume independent stopping rule for Q . By how Ω is partitioned and how probabilities are assigned:

$$q(\mathbf{x}) = Su(\mathbf{x}) + (1 - S) \left(\prod_{i=1}^{K^j} (\theta_i^j)^{n_i^j} \right) q(\mathbf{x} | \mathbf{N}^j = \mathbf{n}^j)$$

with

$$q(\mathbf{x} | \mathbf{N}^j = \mathbf{n}^j) = \prod_{i=1}^{K^j} q(\mathbf{x}(\Omega_i^j) | \Omega_i^j).$$

For any $A \subset \bigcup_{k=1}^{\infty}$ we have an induced Pólya tree distribution $\pi_A(q)$ for the conditional density $q(\cdot | A)$ and we define (if $\mathbf{x}(A) \neq \emptyset$, otherwise $\Phi(A) = 1$):

$$\Phi(A) = \int q(\mathbf{x}(A) | A) d\pi_A(q)$$

Bayesian Inference with Optional Pólya Tree prior

Define also $\Phi_0(A) = u(\mathbf{x}(A)|A) = \prod_{x \in \mathbf{x}(A)} u(x|A)$ (and $\Phi_0(A) = 1$ if $\mathbf{x}(A) = \emptyset$). Then compute

$$\Phi(\Omega) = \rho\Phi_0(\Omega) + (1 - \rho) \sum_{j=1}^M \lambda_j \frac{D(\mathbf{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha})} \prod_{i=1}^{K^j} \Phi(\Omega_i^j)$$

with

$$q(\mathbf{x}|\mathbf{N}^j = \mathbf{n}^j) = \prod_{i=1}^{K^j} q(\mathbf{x}(\Omega_i^j)|\Omega_i^j)$$

$$\text{and } U(\mathbf{x}) = \prod_{i=1}^n u(x_i) \text{ and } u(x) = \frac{1}{\mu(\Omega)}$$

Bayesian Inference with Optional Pólya Tree prior

Theorem 3

Suppose $\mathbf{x} = (x_1, \dots, n)$ are independent observations from Q where Q has a prior distribution $\pi(\cdot)$ that is an optional Pólya tree with independent stopping rule and satisfying the condition of Theorem 2, the conditional distribution of Q given $\mathbf{X} = \mathbf{x}$ is also an optional Pólya tree where, for each $A \subset A^\infty$, the parameters are given as follows:

- Stopping probability: $\rho(A|\mathbf{x}) = \rho(A) \frac{\Phi_0(A)}{\Phi(A)}$;
- Selection Probabilities: $P(J = j|\mathbf{x}) \propto \lambda_j \frac{D(\mathbf{n}^j + \boldsymbol{\alpha}^j)}{D(\boldsymbol{\alpha}^j)} \prod_{i=1}^{K^j} \Phi(A_i^j)$, with $j = 1, \dots, M$;
- Allocation of probability to subregions: the probabilities θ_i^j for subregion A_i^j , $i = 1, \dots, K^j$ are drawn from Dirichlet $(\mathbf{n}^j + \boldsymbol{\alpha}^j)$.

In the above, it is understood that $M, K^j, \lambda_j, \mathbf{n}^j, \boldsymbol{\alpha}^j$ all depend on A .

Density estimation with Optional Pólya Tree Prior

I. Computation of the Posterior Mean Density

Very efficient in single-dimensional setting with unique splitting criterion for each elementary region (intuitively no overlap);

More difficult in multi-dimensional settings: multiple ways to split at each node (possible overlaps of the A^k). Possible solutions:

- Restrictions on how elementary regions can be split (alternate splitting rules). However, downside for the variability of the estimates.
- Estimate the posterior mean density at a point $x \in \Omega$ by $\Phi(\Omega|x, D)/\Phi(\Omega|D)$. However still computationally intensive and only for a specific point.

Density estimation with Optional Pólya Tree Prior

II. Hierarchical MAP Method

Two-stage approach:

- Learn a fixed tree topology representative of the structure of the distribution;
- Estimate the mean density function conditional on this tree topology.

Advantages:

- Reduction of the prior process from an infinite mixture of infinite trees to fixed finite tree;
- Easy to compute conditional mean density function (posterior probability mass of each node is just a product of Beta means and the distribution within stopped regions is uniform by construction);
- Representative partition over the state space is already informative of the structure of the distribution.

Density estimation with Optional Pólya Tree Prior

II. Hierarchical MAP method

Usually candidate tree topology for representing the data structure is the MAP (Maximum A Posteriori) topology, but biased over short tree branches.

Thus, top-down procedure.

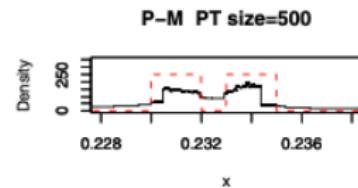
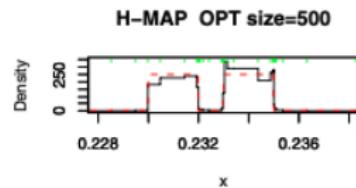
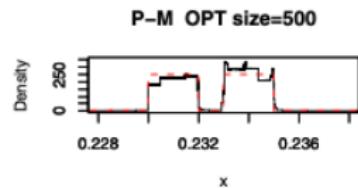
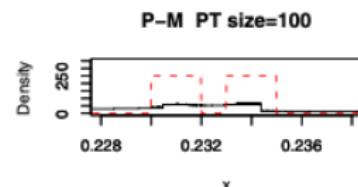
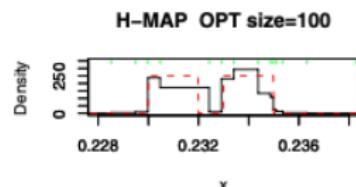
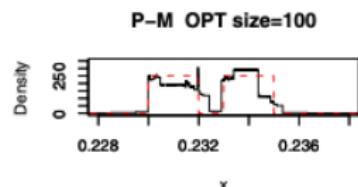
- From the root node, if the posterior $\rho > 0.5$ stop the tree;
- If not, divide the tree into direction k with largest λ_k ;
- Repeat for any A_k^j until all branches have been stopped.

Numerical Examples

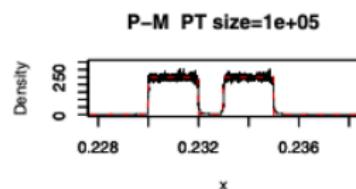
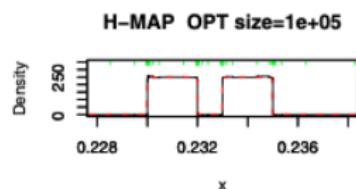
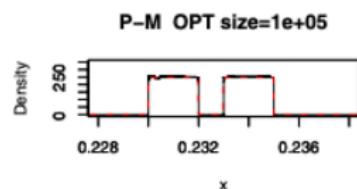
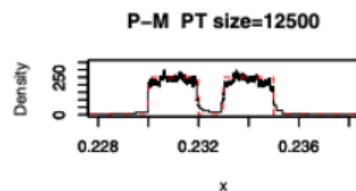
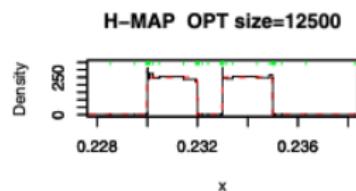
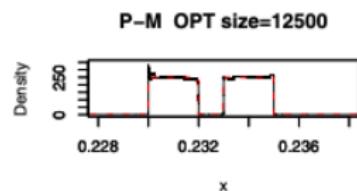
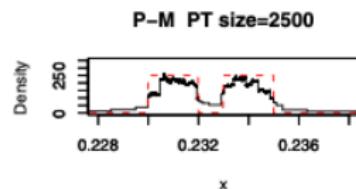
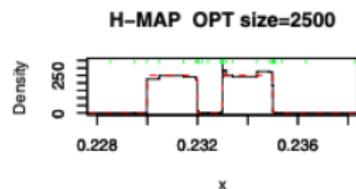
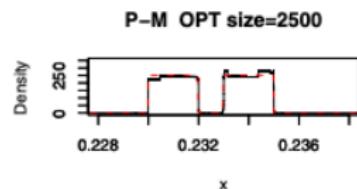
- Setting is $\Omega = [0, 1]$;
- Cutting point is midpoint of its range for the corresponding elementary region;
- Optional Pólya Tree priors: prior $\rho = 0.5$, $\alpha = 0.5$ for all elementary regions;
- Pólya Tree priors: $\alpha = \text{depth}^2$;
- Simulate data from mixture of two uniforms:

$$0.5U(0.23, 0.232) + 0.5U(0.233, 0.235)$$

Mixture of two Uniforms



Mixture of two Uniforms



Conclusions

- Optional Pólya Trees already with sample size of 500 capture the boundaries and modes of the distribution;
- Optional Pólya Trees estimate become smoother with increase in sample size, while Pólya Tree is not.
- Hierarchical MAP performs as well as P-M, but less computation and memory;
- The partition learned in the hierarchical MAP approach reflects the structure of the distribution.

References

-  Billingsley: "Probability and Measure" 3rd ed.,
Wiley-Interscience Publication, 3rd ed., 1995.
-  Ferguson: "A Bayesian Analysis of some Nonparametric Problems", *The Annals of Statistics*, vol. 1, 209-230, 1973.
-  Lavine: "Some Aspects of Polya Tree Distributions for Statistical Modelling", *The Annals of Statistics*, Vol. 20, 1222-1235, 1992.
-  Wong, Ma: "Optional Pólya Trees and Bayesian Inference", *The Annals of Statistics*, vol. 38, No.3, 1433-1459, 2010.