Statistica Sinica **26** (2016), 1175-1195
doi:http://dx.doi.org/10.5705/ss.202015.0163

# MULTISCALE BERNSTEIN POLYNOMIALS FOR DENSITIES

Antonio Canale[1,2] and David B. Dunson[3]

[1] *University of Turin,* [2] *Collegio Carlo Alberto and* [3] *Duke University*

*Abstract:* Our focus is on constructing a multiscale nonparametric prior for densities. The Bayes density estimation literature is dominated by single scale methods, with the exception of Polya trees, which favor overly-spiky densities even when the truth is smooth. We propose a multiscale Bernstein polynomial family of priors, which produce smooth realizations that do not rely on hard partitioning of the support. At each level in an infinitely-deep binary tree, we place a beta dictionary density; within a scale the densities are equivalent to Bernstein polynomials. Using a stick-breaking characterization, stochastically decreasing weights are allocated to the finer scale dictionary elements. A slice sampler is used for posterior computation, and properties are described. The method characterizes densities with locally-varying smoothness, and can produce a sequence of coarse to fine density estimates. An extension for Bayesian testing of group differences is introduced and applied to DNA methylation array data.

*Key words and phrases:* Density estimation, multiresolution, multiscale clustering, multiscale testing, nonparametric Bayes, Polya tree, stick-breaking, wavelets.

## 1. Introduction

Multiscale estimators have well-known advantages, including the ability to characterize abrupt local changes and to provide a compressed estimate to a desired level of resolution. Such advantages have lead to the enormous popularity of wavelets, which are routinely used in signal and image processing, and have had attention in the literature on density estimation. Donoho et al. (1996) developed a wavelet thresholding approach for density estimation, and there is a literature developing modifications for deconvolution problems (Pensky and Vidakovic (1999)), censored data (Niu (2012)), time series (Garcia-Trevino and Barria (2012)) and other settings. Locke and Peter (2013) proposed an approach, that can better characterize local symmetry and other features commonly observed in practice, using multiwavelets. Chen et al. (2012) instead use geometric multiresolution analysis methods related to wavelets to obtain estimates of high-dimensional distributions having low-dimensional support.

Although there is a rich Bayesian literature on multiscale function estimation (Abramovich, Sapatinas, and Silverman (1998); Clyde, Parmigiani, and Vidakovic (1998); Clyde and George (2000); Wang, Ray, and Mallick (2007)), there has been limited consideration of Bayesian multiscale density estimation. Popular methods for Bayes density estimation rely on kernel mixtures. For example, Dirichlet process mixtures are applied routinely. By using location-scale mixtures, one can accommodate varying smoothness, with the density being flat in certain regions and concentrated in others. However, Dirichlet processes lack the appealing multiscale structure. Polya trees provide a multiscale alternative (Mauldin, Sudderth, and Williams (1992); Lavine (1992a,b)), but have practical disadvantages; they tend to produce highly spiky density estimates even when the true density is smooth, and have sensitivity to a pre-specified partition sequence. This sensitivity can be ameliorated by mixing Polya trees (Hanson and Johnson (2002)), but at the expense of more difficult computation.

Our focus is on developing a new approach for Bayesian multiscale density estimation, that inherits many of the advantages of Dirichlet process mixtures while avoiding the key disadvantages of Polya trees. We want a framework that is easily computable, has desirable multiscale approximation properties, allows centering on an initial guess at the density, and can be extended in a straightforward manner to include covariates and allow embedding within larger models. We accomplish this using a multiscale extension of mixtures of Bernstein polynomials (Petrone (1999a,b)).

## 2. Multiscale Priors for Densities

### 2.1. Proposed model

Let $x \in \mathcal{X} \subset \Re$ be a random variable having density $g$ with respect to Lebesgue measure. Assume that $g_0$ is a prior guess for $g$, with $G_0$ and $G_0^{-1}$ the corresponding cumulative distribution function and inverse cumulative distribution function, respectively. We induce a prior $g \sim \Pi$ centered on $g_0$ through a prior for the density $f$ of $y = G_0(x) \in (0,1)$. The cumulative distribution functions $F$ and $G$ corresponding to the densities $f$ and $g$, respectively, have the relationship

$$G(x) = F\{G_0(x)\}, x \in \mathcal{X}, \quad F(y) = G\{G_0^{-1}(y)\}, y \in (0,1). \tag{2.1}$$

We assume that $f$ is a multiscale mixture of Bernstein polynomials,

$$f(y) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} \mathrm{Be}(y; h, 2^s - h + 1), \tag{2.2}$$

where $\mathrm{Be}(a,\ b)$ denotes the beta density with mean $a/(a+b)$, and $\{\pi_{s,h}\}$ are random weights drawn from a suitable stochastic process. We introduce an infinite sequence of scales $s = 0, 1, \ldots, \infty$. At scale $s$, we include $2^s$ Bernstein

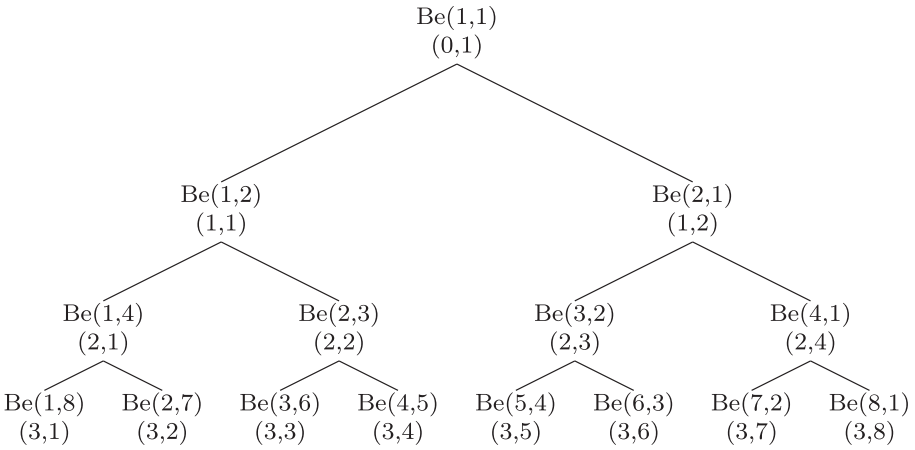


Figure 1. Binary tree with beta kernels at each node $(s, h)$, where $s$ is the scale level and $h$ is the index within the scale.

polynomial basis densities. The framework can be represented as a binary tree in which each layer is indexed by a scale and each node is a suitable beta density. For example, at the root node, we have the Be(1,1) density which generates two daughters Be(1,2) and Be(2,1) and so on. In general, let $s$ denote the scale and $h$ the polynomial within the scale. The node $(s, h)$ in the tree is related to the $\mathrm{Be}(h, 2^s - h + 1)$ density. A cartoon of the binary tree is reported in Figure 1.

A prior measure for the multiscale mixture (2.2) is obtained by specifying a stochastic process for the infinite dimensional set of weights $\{\pi_{s,h}\}$. To this end we introduce, for each scale $s$ and node $h$ within the scale, independent random variables

$$S_{s,h} \sim \mathrm{Be}(1, a), \quad R_{s,h} \sim \mathrm{Be}(b, b), \tag{2.3}$$

corresponding to the probability of stopping and taking the right path conditionally on not stopping, respectively. Define the weights as

$$\pi_{s,h} = S_{s,h} \prod_{r<s} (1 - S_{r,g_{shr}}) T_{shr}, \tag{2.4}$$

where $g_{shr} = \lceil h/2^{s-r} \rceil$ is the node traveled through at scale $r$ on the way to node $h$ at scale $s$, $T_{shr} = R_{r,g_{shr}}$ if $(r+1, g_{shr+1})$ is the right daughter of node $(r, g_{shr})$, and $T_{shr} = 1 - R_{r,g_{shr}}$ if $(r+1, g_{shr+1})$ is the left daughter of $(r, g_{shr})$. A cartoon of the weights construction is reported in Figure 2. For binary trees, there is a unique path leading from the root node to node $(s, h)$, and $\mathcal{T}$ denotes the infinite deep binary tree of the weights (2.4). We refer to the prior resulting from (2.2)−(2.4) as a multiscale Bernstein polynomial (msBP) prior and we write $f \sim \mathrm{msBP}(a, b)$. Choices for the hyperparameters are discussed in the next section.

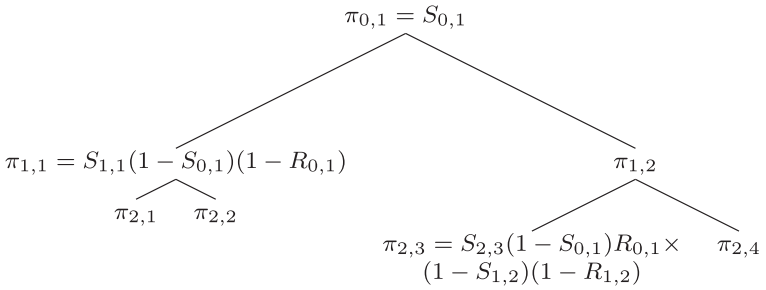


Figure 2. Examples of weights construction.

The infinite tree of probability weights is generated from a generalization of the stick-breaking process representation of the Dirichlet process (Sethuraman (1994)). Each time the stick is broken, it is consequently randomly divided in two parts (one for the probability of going right, the remainder for the probability of going left) before the next break. An alternative tree stick-breaking process is proposed by Adams, Ghahramani, and Jordan (2010) where a first stick-breaking process defines the vertical growth of an infinitely wide tree and a second puts weights on the infinite number of descendant nodes.

Sampling a random variable $y$ from a random density, which is generated from a msBP prior, can be described as follows. At node $(s, h)$, generate a random probability $S_{s,h} \sim \text{Be}(1, a)$ corresponding to the probability of stopping at that node given you passed through that node, and $R_{s,h} \sim \text{Be}(b, b)$ corresponding to the probability of taking the right path in the tree in moving to the next finer scale given you did not stop at node $(s, h)$. Conditionally on being at the node $(s, h)$ we assume that $y \sim \text{Be}(y; h, 2^s - h + 1)$.

### 2.2. Basic properties

In this section we study basic properties of the proposed prior. A first requirement is that the construction leads to a meaningful sequence of weights.

**Lemma 1.** *If $\pi_{s,h}$ is an infinite sequence of weights defined as in* (2.3)−(2.4),

$$\sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} = 1 \tag{2.5}$$

*almost surely for any $a, b > 0$.*

The total weight placed on a scale $s$ is controlled by the prior for $S_{s,h}$. The expected probability allocated to node $h$ at scale $s$ can be expressed as

$$E(\pi_{s,h}) = E\left\{ S_s \prod_{l=0}^{s-1} (1 - S_l) \prod_{l=1}^{s} T_l \right\}$$

$$= \left(\frac{1}{1+a}\right)\left(\frac{a}{1+a}\right)^s\left(\frac{1}{2}\right)^s = \frac{1}{1+a}\left(\frac{a}{2+2a}\right)^s, \tag{2.6}$$

where we discard the $h$ subscript on $S_l \sim \text{Be}(1,a)$ and $T_l \sim \text{Be}(b,b)$ for ease in notation. This does not impact the calculation because any path taken up to scale $s$ has the same probability *a priori* and the random variables in (2.3) have the same distribution regardless of the path that is taken. Similarly

$$E(\pi_{s,h}^2) = E\left\{S_s^2\prod_{l=0}^{s-1}(1-S_l)^2\prod_{l=1}^{s}T_l^2\right\} = \frac{2}{(1+a)(2+a)}\left(\frac{a}{2+a}\right)^s\left\{\frac{b+1}{2(2b+1)}\right\}^s.$$

Hence at scale $s=0$ the variance is $\text{var}(\pi_{0,1}) = a/\{(2+a)(1+a)^2\}$, while for $s>0$

$$\text{var}(\pi_{s,h}) = \frac{2}{(1+a)(2+a)}\left(\frac{a}{2+a}\right)^s\left\{\frac{b+1}{2(2b+1)}\right\}^s - \left\{\frac{1}{1+a}\left(\frac{a}{2+2a}\right)^s\right\}^2. \tag{2.7}$$

In the Supplementary material, we report the prior expectation and 95% prior credible intervals of the total weight assigned to each scale for some hyperparameter values.

We can additionally verify that our prior for the cumulative distribution function $G$ is centered on the chosen $G_0$. Letting $F(A) = \int_A f$, we obtain $E\{F(A)\} = \lambda(A)$, where $\lambda(A)$ is the Lebesgue measure over the set $A$. Details are reported in the Appendix. Hence, the prior for the density of $y$ is automatically centered on a uniform density on $[0,1]$. This is the desired behavior as $y \sim \text{Unif}(0,1)$ with $x = G_0^{-1}(y)$ implies that $x \sim g_0$, which is our prior guess for the observed data density. In addition, from (2.1), $E\{F(y)\} = y$ implies

$$E\{G(x)\} = E[G\{G_0^{-1}(y)\}] = y = G_0(x),$$

so that the prior expectation for the cumulative distribution function $G$ is $G_0$ as desired.

From (2.6) and (2.7), the hyperparameter $a$ controls the decline in probabilities over scales. In general, letting $S^{(i)}$ denote the scale at which the $i$th observation falls, we have

$$E(S^{(i)}) = \sum_{s=0}^{\infty} s\frac{1}{1+a}\left(\frac{a}{2+2a}\right)^s = a.$$

Hence, the value of $a$ is the expected scale from which observations are drawn. For small $a$, high probability is placed on coarse scales, leading to smoother densities, with $a \to 0$ inducing $\pi_{0,1} = 1$ and hence $f(y)$ uniform. As $a$ increases, finer scale densities will be weighted higher, leading to spikier realizations. To

illustrate this, in the Supplementary materials we show realizations from the prior for different $a$ values.

An appealing aspect of the proposed formulation is that individuals sampled from a distribution that is assigned a msBP prior are allocated to clusters in a multiscale fashion. In particular, two individuals having similar observations may have the same cluster allocation up to some scale $s$, but perhaps are not clustered on finer scales. Clustering is intrinsically a scale dependent notion, and our model is the first to our knowledge to formalize multiscale clustering in a model based probabilistic manner. Under the above structure, the probability that two individuals $i$ and $i'$ are assigned to the same scale $s$ cluster is one for $s = 0$ and for $s > 0$, is *a priori* equal to

$$2^s E\left\{\prod_{l=0}^{s-1}(1-S_l)^2 T_l^2\right\} = 2^s\left(\frac{a}{a+2}\right)^s\left(\frac{1}{2}\right)^s\left(\frac{b+1}{2b+1}\right)^s = \left\{\left(\frac{a}{a+2}\right)\left(\frac{b+1}{2b+1}\right)\right\}^s.$$

This is derived by calculating the expected probability that two individuals travel though node $h$ at scale $s$ and multiplying by the number of nodes in scale $s$. This form is intuitive. As $b \to 0$, the $\mathrm{Be}(b, b)$ density degenerates to $0.5\delta_0 + 0.5\delta_1$, so that variability among subjects in the chosen paths through the tree decreases and all subjects take a common path chosen completely at random via unbiased coin flips at each node. In such a limiting case, $(b + 1)/(2b + 1) \to 1$ and the probability of clustering subjects at scale $s$ is simply the probability of surviving to that scale and not being allocated to a coarser scale component. At the other extreme, as $b \to \infty$ each subject independently flips an unbiased coin in deciding to go right or left at each node of the tree, and

$$\frac{b+1}{2b+1} \to \frac{1}{2}.$$

Hyperpriors can be chosen for $a$ and $b$ to allow the data to inform about these tuning parameters; we find that choosing a hyperprior for $a$ is particularly important, with $b = 1$ as a default.

Approximations of the msBP process can be obtained fixing an upper bound $s$ for the depth of the tree. The truncation is applied by pruning $\mathcal{T}$ at scale $s$, setting $S_{s,h} = 1$ for each $h = 1, \ldots, 2^s$ as done in Ishwaran and James (2001) and related works in the single scale case. The truncations can be applied if one considers not scientifically relevant higher levels of resolution or for computational reasons. See the discussion in the next section. We denote the scale $s$ approximation as

$$f^s(y) = \sum_{l=0}^{s}\sum_{h=1}^{2^l}\tilde{\pi}_{l,h}\mathrm{Be}(y; h, 2^l - h + 1), \tag{2.8}$$

with $\tilde{\pi}_{l,h}$ identical to $\pi_{l,h}$ except that we set all the stopping probabilities at scale $s$ equal to one, $\tilde{\pi}_{l,h} = \pi_{l,h}$ for $l < s$ and

$$\tilde{\pi}_{s,h} = \prod_{r<s}(1 - S_{r,g_{shr}})T_{shr}.$$

This is made to ensure that the weights sum to one and that $f^s(y)$ is a valid probability density on $\mathcal{Y} = [0,1]$. Let $\mathcal{T}^s$ denote the pruned binary tree of weights. It is interesting to study the accuracy of the approximation of $f^s(y)$ to $f(y)$ as the scale $s$ changes.

**Lemma 2.** *Let $p^s(y)$, $p^\infty(y)$ be the marginal likelihood for $y$ under the $s$-truncated and full msBP mixture model, respectively. Then*

$$||p^s(y) - p^\infty(y)|| = 0,$$

*where $||p^s(y) - p^\infty(y)||$ denotes the $L_1$ distance.*

In Lemma 2, $p^s(y)$ and $p^\infty(y)$ can be thought of as the expected sampling density of $y$ under the prior truncated to $s$ levels and under the full msBP prior, respectively. The $L_1$ distance between these expected densities provides a measure of prior bias induced by the truncation. Lemma 2 shows that the prior is calibrated so that this bias is exactly zero; it is somewhat surprising that this is possible, and we expected to instead obtain a bound on the $L_1$ distance that decreases exponentially with $s$ similar to Theorem 2 in Ishwaran and James (2001). Such bounds are obtained in Lemma 3, which instead focuses on the total variation distance between the truncated and exact random measures $P^s$ and $P$.

**Lemma 3.** *The expectation and variance of the total variation distance between $P^s(B)$ and $P(B)$ are*

$$E\left\{d_{TV}(P_s, P)\right\} \leq \left(\frac{a}{a+1}\right)^{s+1}, \quad \text{var}\left\{d_{TV}(P_s, P)\right\} \leq 2\left(\frac{a}{a+1}\right)^{s}.$$

## 3. Posterior Computation

In this section we demonstrate that a straightforward Markov chain Monte Carlo algorithm can be constructed to perform posterior inference under the msBP prior. The algorithm consists of two primary steps: first, allocate each observation to a multiscale cluster, conditionally on the current values of the probabilities $\{\pi_{s,h}\}$, and second, conditionally on the cluster allocations, update the probabilities.

Suppose subject $i$ is assigned to node $(s_i, h_i)$, with $s_i$ the scale and $h_i$ the node within scale. Conditionally on $\{\pi_{s,h}\}$, the posterior probability of subject $i$ belonging to node $(s,h)$ is simply

$$\text{pr}(s_i = s, h_i = h \mid y_i, \pi_{s,h}) \propto \pi_{s,h}\text{Be}(y; h, 2^s - h + 1).$$

Consider the total mass assigned at scale $s$, defined as $\pi_s = \sum_{h=1}^{2^s} \pi_{s,h}$, and let $\bar{\pi}_{s,h} = \pi_{s,h}/\pi_s$. Under this notation, we can rewrite (2.2) as

$$f(y) = \sum_{s=0}^{\infty} \pi_s \sum_{h=1}^{2^s} \bar{\pi}_{s,h} \mathrm{Be}(y; h, 2^s - h + 1).$$

To allocate each subject to a multiscale cluster, we rely on a multiscale modification of the slice sampler of Kalli, Griffin, and Walker (2011). Consider the joint density

$$f(y_i, u_i, s_i) \propto \mathbb{1}(u_i < \pi_{s_i}) \sum_{h=1}^{2^{s_i}} \bar{\pi}_{s_i,h} \mathrm{Be}(y_i; h, 2^{s_i} - h + 1).$$

The full conditional posterior distributions are

$$u_i \mid y_i, s_i \sim \mathrm{Unif}(0, \pi_{s_i}), \tag{3.1}$$

$$\mathrm{pr}(s_i = s \mid u_i, y_i) \propto \mathbb{1}(s : \pi_s > u_i) \sum_{h=1}^{2^s} \bar{\pi}_{s,h} \mathrm{Be}(y_i; h, 2^s - h + 1), \tag{3.2}$$

$$\mathrm{pr}(h_i = h \mid u_i, y_i, s_i) \propto \bar{\pi}_{s_i,h} \mathrm{Be}(y_i; h, 2^{s_i} - h + 1). \tag{3.3}$$

Even with an infinite resolution level, (3.2) implies that observations are assigned to a finite number of scales and there are a finite number of probabilities to evaluate. Conditionally on the scale, (3.3) induces a simple multinomial sampling, which allocates a subject to a particular node within that scale. Algorithm 1 summarizes the posterior cluster allocation step. The tree is grown and shrunk adaptively, up to the needed level of resolution, so that it is not necessary to use the truncation described at the end of the previous section for computation.

for each scale $s$
  calculate $\pi_s = \sum_{h=1}^{2^s} \pi_{s,h}$;
simulate $u_i \mid y_i, s_i \sim U(0, \pi_{s_i})$;
for each scale $s$
  If $\pi_s > u_i$, for $h = 1, \dots 2^s$
    $\bar{\pi}_{s,h} \leftarrow \pi_{s,h}/\pi_s$
    $\mathrm{pr}(s_i = s \mid u_i, y_i) \propto \sum_{h=1}^{2^s} \bar{\pi}_{s,h} \mathrm{Be}(y_i; h, 2^s - h + 1)$
  else
    $\mathrm{pr}(s_i = s \mid u_i, y_i) = 0$;
sample $s_i$ with probability $\mathrm{pr}(s_i = s \mid u_i, y_i)$;
sample $h_i$ with probability $\mathrm{pr}(h_i = h \mid y_i, s_i) \propto \bar{\pi}_{s_i,h} \mathrm{Be}(y_i; h, 2^{s_i} - h + 1)$.

**Algorithm 1:** Multiscale cluster posterior allocation for $i$th subject.

Conditionally on cluster allocations, we sample all the stopping and descending-right probabilities from their full conditional posterior distributions:

$$S_{s,h} \sim \text{Be}(1+n_{s,h}, a+v_{s,h}-n_{s,h}), \quad R_{s,h} \sim \text{Be}(b+r_{s,h}, b+v_{s,h}-n_{s,h}-r_{s,h}), \tag{3.4}$$

where $v_{s,h}$ is the number of subjects passing through node $(s,h)$, $n_{s,h}$ is the number of subjects stopping at node $(s,h)$, and $r_{s,h}$ is the number of subjects that continue to the right after passing through node $(s,h)$. Calculation of $v_{s,h}$ and $r_{s,h}$ can be performed via parallel computing due to the binary tree structure, improving efficiency.

If hyperpriors for $a$ and $b$ are assumed, additional sampling steps are required. Assuming $a \sim \text{Ga}(\beta, \gamma)$, where $\text{Ga}(k, \theta)$ is the gamma density with mean $k/\theta$ and variance $k/\theta^2$, its full conditional posterior is

$$a \mid - \sim \text{Ga}\Big(\beta + 2^{s'+1} - 1, \gamma - \sum_{s=0}^{s'} \sum_{h=1}^{2^s} \log(1 - S_{s,h})\Big), \tag{3.5}$$

where the symbol $|-$ stands for "conditionally on the the rest of the parameters". If $b \sim \text{Ga}(\delta, \lambda)$ its full conditional posterior is proportional to

$$\frac{b^{\delta-1}}{B(b,b)^{2^{s'+1}-1}} \exp\Big(b\Big[\sum_{s=0}^{s'} \sum_{h=1}^{2^s} \log\{R_{s,h}(1-R_{s,h})\} - \lambda\Big]\Big), \tag{3.6}$$

where $s'$ is the maximum occupied scale and $B(p,q)$ is the Beta function. To sample from the latter distribution, a Metropolis-Hastings step is required. The Gibbs sampler iterates the steps outlined in Algorithm 2.

for $i = 1, \ldots, n$
  assign observation $i$ to a cluster $(s_i, h_i)$ as in Algorithm 1
for $s = 0, \ldots, s_{\text{MAX}}$
  for $h = 1, \ldots, 2^s$
    update $S_{s,h} \sim \text{Be}(1 + n_{s,h}, a + v_{s,h} - n_{s,h})$;
    update $R_{s,h} \sim \text{Be}(b + r_{s,h}, b + v_{s,h} - n_{s,h} - r_{s,h})$;
update $a$ from (3.5);
update $b$ from (3.6).

**Algorithm 2:** Gibbs sampler steps for posterior computation under msBP prior.

## 4. Simulation Study

We compared our msBP method to standard Bayesian nonparametric techniques including Dirichlet process location-scale mixtures of Gaussians, Dirichlet process mixtures of Bernstein polynomials, and mixtures of Polya trees (Hanson

(2006)), all using a default implementation of the R package `DPpackage`. In addition, we implemented a frequentist wavelet density estimator using the package `WaveThresh`, and a simple frequentist kernel estimator. Several simulations have been run under different simulation settings leading to qualitatively similar results. We report the results for four scenarios. Scenario 1 simulated data from a mixture of betas, $0.6\text{Be}(3, 3) + 0.4\text{Be}(21, 5)$; Scenario 2 used a mixture of Gaussians, $0.5N(0, 4) + 0.3N(2, 1) + 0.2N(1.5, 0.25)$; Scenario 3 generated data from a density supported on the positive real line, a mixture of a gamma and a left truncated normal, $0.9\text{Ga}(2, 2) + 0.1N_{\text{LT}}(4, 0.4)$; Scenario 4 generated data from a symmetric density with two spiky modes, $0.7N(0, 4) + 0.1N(0.5, 0.01) + 0.2N(1.5, 0.4)$.

For each case, we generated sample sizes of $n = 25, 50, 100, 250$. Each of the approaches were applied to 200 replicated data sets under each scenario. The methods were compared based on a Monte Carlo approximation to the mean Kolmogorov-Smirnov distance (KS), $L_1$, and $L_2$ distances.

To implement Algorithm 2, we exploit the binary tree structure of our modelling framework using efficient C++ code embedded into R functions. Code is freely available on CRAN in the `msBP` package. In implementing the Gibbs sampler, the first 1,000 iterations were discarded as a burn-in and the next 2,000 samples were used to calculate the posterior mean of the density on a fine grid of points. To center our prior, using a default empirical Bayes approach, we set $g_0$ equal to a kernel estimate. For the hyperparameters we fixed $b = 1$ and let $a \sim \text{Ga}(5, 0.5)$. We truncated the depth of the binary tree to the sixth scale. The values of the density for a wide variety of points in the domain were monitored to gauge rates of convergence and mixing. The trace plots showed excellent mixing and the Geweke's diagnostic suggested that convergence is reached within a few hundred iterations.

The results of the simulation are reported in the Supplementary material. The proposed method performs better or equally to the best competitor in almost all scenarios and sample sizes. The worst performance in each case is obtained for mixtures of Polya trees, with overly-spiky density estimates leading to higher distances from the truth. In Scenario 1 the msBP approach beats all the competitors, except in large sample sizes when single-scale Dirichlet process mixtures of Bernstein polynomials are comparable. In Scenario 2 the msBP approach is comparable to the frequentist kernel smoother estimator. In scenario 3 the msBP approach is comparable to Dirichlet process location-scale mixtures and in Scenario 4 our multiscale approach is clearly performing better than any other method.

Although it is difficult to compare execution times fairly given differences in implementation, we report computation times for all methods in Table 3 of the Supplementary material. In general, our implementation of msBP was comparable to but faster than the other Bayesian nonparametric competitors (with

the exception of Polya trees), averaging only a couple of seconds per simulated data set to run on a MacBook Pro with 2.8 GHz Intel Quadcore i7 CPU and 16 GB of RAM. Our approach is motivated by settings in which the true density has both fine and coarse scale features, so that a multiscale density estimator may be needed. However, we designed the prior and computational algorithm to also perform well when the density is well approximated by a single scale basis. This adaptivity was illustrated in the overall good performance in the simulations across cases. We would nonetheless say that in large sample sizes when the entire goal is to produce a single density estimate, the proposed approach may have limited motivation relative to simpler methods. However, a substantial motivation is the ease in which msBP can be generalized, as we discuss in the next Section.

## 5. Extensions

An appealing aspect of the proposed method is ease of generalization to include predictors, hierarchical dependence, time series, spatial structure and so on. To incorporate additional structure, one can replace model (2.2) for the stopping and right path probabilities with an appropriate variant. Similar extensions have been proposed for single resolution mixture models by replacing the beta random variables in a stick-breaking construction with probit regressions (Chung and Dunson (2009); Rodriguez and Dunson (2011)), logistic regressions (Ren et al. (2011)) or broader stochastic processes (Pati, Dunson, and Tokdar (2013)). We focus here on one interesting extension to the under-studied problem of Bayesian multiscale inferences on differences between groups.

### 5.1. Multiscale testing of group differences

Motivated by epigenetic data, we propose Bayesian multiscale hypothesis tests of group differences using multiscale Bernstein polynomials. DNA methylation arrays collect data on epigenetic modifications at a large number of CpG sites. Let $y_i = (y_{i1}, \ldots, y_{ip})^T$ denote the DNA methylation data for patient $i$ at $p$ different sites, with $d_i \in \{0, 1\}$ denoting the patient's disease status, either $d_i = 0$ for controls or $d_i = 1$ for cases. Current standard analyses rely on independent screening using $t$-tests to assess differences between cases and control at each site. However, DNA methylation data are constrained to $y_{ij} \in (0, 1)$ and tend to have a complex distribution having local spikes and varying smoothness.

As illustration we focus on nonparametric independent screening; the approach is easily adapted to accommodate dependence across sites. We center our prior on the uniform as a default. The density of $y_{ij}$ given $d_i = 0$ is modeled as in previous sections. Let $H_0 : f_0 = f_1$ denote the *global* null hypothesis of no difference between groups, with $H_1 : f_0 \neq f_1$ denoting the alternative. Using

a msBP representation, $f_0 = f_1$ if the groups share weights over the dictionary of beta densities. If $f_0 \neq f_1$, we may have the same weights on the dictionary elements up to a given scale, so that the densities are equivalent up to that scale but not at finer scales. With this in mind, let $H_0^s : f_0^s = f_1^s$ denote the null hypothesis of no differences between groups at scale $s$, and $H_1^s : f_0^s \neq f_1^s$ the alternative. As $H_0^0$ is true with probability one, we set $S_{0,1} = 0$ and concentrate on $H_0^s$ for $s \geq 1$.

Each of the $n$ subjects in the sample takes a path through the binary tree, stopping at a finite depth. Let $\mathcal{I}^s = \{i : s_i \geq s\}$ index the subjects *surviving* up to scale $s$ and let $\mathcal{N}^s$ denote the actions of these subjects at scale $s$, including stopping or progressing downward to the left or right for each of the nodes. Subscripts $(d)$ on $\mathcal{I}^s$ and $\mathcal{N}^s$ denote the restriction to subjects having $d_i = d$. Conditionally on $H_0^s$, the probabilities for each scale $s$ action are the same in the two groups and the likelihood of actions $\mathcal{N}^s$ is

$$\begin{aligned}
\mathrm{pr}(\mathcal{N}^s \mid H_0^s) &= \int_{\mathcal{T}} \mathrm{pr}(\mathcal{N}^s \mid \mathcal{T})\mathrm{pr}(\mathcal{T} \mid a, b)d\mathcal{T} \\
&= \left\{ \frac{\Gamma(a+1)}{\Gamma(a)} \frac{\Gamma(2b)}{\Gamma(b)^2} \right\}^{2^s} \int_{\mathcal{T}} \prod_{h=1}^{2^s} S_{s,h}^{n_{s,h}} (1 - S_{s,h})^{\hat{a}_{s,h}-1} R_{s,h}^{\hat{b}_{s,h}-1} (1 - R_{s,h})^{\hat{c}_{s,h}-1} d\mathcal{T} \\
&= \left\{ \frac{\Gamma(a+1)\Gamma(2b)}{\Gamma(a)\Gamma(b)^2} \right\}^{2^s} \prod_{h=1}^{2^s} \frac{\Gamma(1+n_{s,h})\Gamma(\hat{a}_{s,h})}{\Gamma(a+v_{s,h}+1)} \frac{\Gamma(\hat{b}_{s,h})\Gamma(\hat{c}_{s,h})}{\Gamma(2b+v_{s,h}-n_{s,h})}, \qquad (5.1)
\end{aligned}$$

where $\hat{a}_{s,h} = a + v_{s,h} - n_{s,h}$, $\hat{b}_{s,h} = b + r_{s,h}$, and $\hat{c}_{s,h} = b + v_{s,h} - n_{s,h} - r_{s,h}$. Similarly under $H_1$ we have

$$\begin{aligned}
\mathrm{pr}(\mathcal{N}^s \mid H_1^s) &= \mathrm{pr}(\mathcal{N}_{(0)}^s \mid H_1^s) \times \mathrm{pr}(\mathcal{N}_{(1)}^s \mid H_1^s) \\
&= \left\{ \frac{\Gamma(a+1)\Gamma(2b)}{\Gamma(a)\Gamma(b)^2} \right\}^{2^{2s}} \prod_{h=1}^{2^s} \frac{\Gamma(1+n_{s,h}^{(0)})\Gamma(\hat{a}_{s,h}^{(0)})}{\Gamma(a+v_{s,h}^{(0)}+1)} \frac{\Gamma(\hat{b}_{s,h}^{(0)})\Gamma(\hat{c}_{s,h}^{(0)})}{\Gamma(2b+v_{s,h}^{(0)}-n_{s,h}^{(0)})} \\
&\quad \times \prod_{h=1}^{2^s} \frac{\Gamma(1+n_{s,h}^{(1)})\Gamma(\hat{a}_{s,h}^{(1)})}{\Gamma(a+v_{s,h}^{(1)}+1)} \frac{\Gamma(\hat{b}_{s,h}^{(1)})\Gamma(\hat{c}_{s,h}^{(1)})}{\Gamma(2b+v_{s,h}^{(1)}-n_{s,h}^{(1)})}, \qquad (5.2)
\end{aligned}$$

where $v_{s,h}^{(d)}$ is the number of subjects passing through node $(s,h)$ in group $d$, $n_{s,h}^{(d)}$ is the number of subjects stopping at node $(s,h)$ in group $d$, and $r_{s,h}^{(d)}$ is the number of subjects that continue to the right after passing through node $(s,h)$ in group $d$, with $d = 0, 1$.

Combining (5.1)−(5.2) we can obtain a closed form for the posterior probability of $H_0$ being true at scale $s$, given $\mathcal{N}_{(0)}^s$ and $\mathcal{N}_{(1)}^s$:

$$\mathrm{pr}(H_0^s \mid \mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s) = \frac{P_0^s \mathrm{pr}(\mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s \mid H_0^s)}{P_0^s \mathrm{pr}(\mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s \mid H_0^s) + (1 - P_0^s)\mathrm{pr}(\mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s \mid H_1^s)}, \qquad (5.3)$$

where $P_0^s$ is our prior guess for the null being true at scale $s$. The global null will be the cumulative product of the $\text{pr}(H_0^s \mid \mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s)$ for each scale. An interesting feature of this formulation is to have a multiscale hypothesis testing setup. Indeed the posterior probability of $H_0$ up to scale $\tilde{s}$ will be $\prod_{s \leq \tilde{s}} \text{pr}(H_0^s \mid \mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s)$ and hence the hypothesis that two groups have the same distribution may have high posterior probability for coarse scales, but can be rejected for a finer scale.

### 5.2. Posterior computation

The conditional posterior probability for $H_0^s$ in (5.3) is simple, but not directly useful due to the dependence on the unknown $\mathcal{N}^s$ allocations. To marginalize out these allocations, we modify Algorithm 2. For node $h$ at scale $s$, let $\pi_{s,h}^{(0)}$ denote the weight under $H_0^s$ and $\pi_{s,h}^{(1,d)}$ for $d = 0, 1$ denote the group-specific weights under $H_1^s$. At each iteration, the allocation of subject $i$ of group $d$ will be made according to the tree of weights given by

$$\pi_{s,h}^{(d)} = \text{pr}(H_0^s \mid \mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s)\pi_{s,h}^{(0)} + \{1 - \text{pr}(H_0^s \mid \mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s)\}\pi_{s,h}^{(1,d)}. \tag{5.4}$$

Given the allocation one can calculate all the quantities in (5.1)−(5.2) and then update the stopping and descending probabilities under $H_0$ and $H_1$ following (3.4) and the posterior of the null following (5.3) up to a desired upper scale.

## 6. Application

We illustrate our approach on a methylation array dataset for $n = 597$ breast cancer samples registered at $p = 21{,}986$ CpG sites (Cancer Genome Atlas Network (2012)). We test for differences between tumors that are identified as basal-like ($n_0 = 112$) against those that are not ($n_1 = 485$) at each CpG site. This problem was considered in a single scale manner by Lock and Dunson (2015) using finite mixtures of truncated Gaussians.

We ran the Gibbs sampler reported in Algorithm 3 in the Appendix, assuming a uniform prior for $P_0^s$ for each scale $s$. We fixed the maximum scale to 4 as an upper bound, as finer scale tests were not thought to be interpretable. The sampler was run for 2,000 iterations after 1,000 burn-in iterations. The chains mixed well and converged quickly for all sites and all scales.

The posterior distribution of $1 - P_0^s$ for each scale provides a summary of the overall proportion of CpG sites for which there was a difference between the two groups. The estimated posterior means for these probabilities were 0.04, 0.07, 0.05 and 0.03, respectively, for scales $1, \ldots, 4$. This suggests that DNA methylation levels were different for a small minority of the CpG sites, which is as expected. Examining the posterior probabilities of $H_1^s$ across the 21,986 CpG sites, consistently with the estimates for $1 - P_0^s$, we find that scale-specific

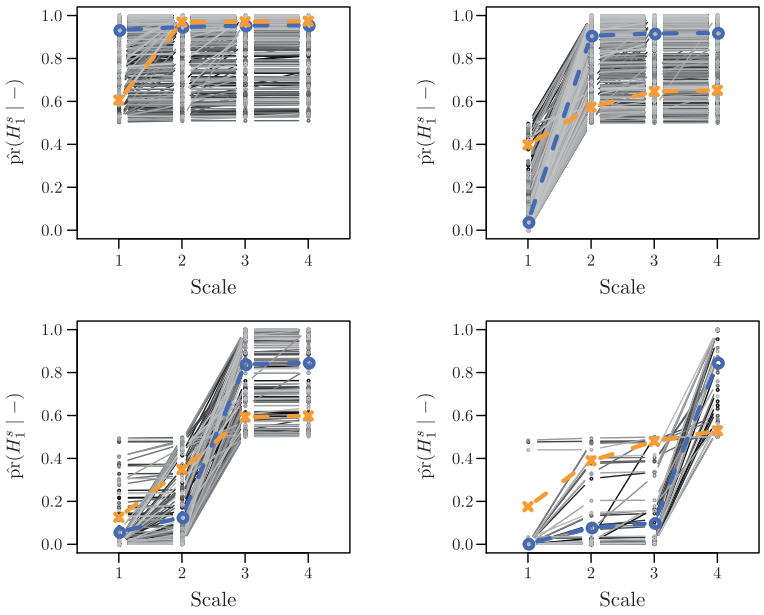


Figure 3. Posterior mean probabilities of $H_1$ depending on scale for the 1,696 sites, with some evidence of differences in the two groups, grouped in subplots by minimal scale showing $\hat{\text{pr}}(H_1^s \mid -) > 0.5$ for $s = 1, \dots 4$. Within each panel, the thick dashed lines represents the average between the sites in two clusters showing different patterns.

estimated posterior probabilities are close to zero for most sites. Focusing on the 1,696 sites for which the overall posterior probability of $H_1$ is greater than 0.5, we calculated the minimal scale showing evidence of a difference, $\min\{s : \hat{\text{pr}}(H_1^s \mid -) > 0.5\}$, with $\hat{\text{pr}}(H_1^s \mid -) = 1 - \prod_{l \leq s} \text{pr}(H_0^l \mid \mathcal{N}_{(0)}^l, \mathcal{N}_{(1)}^l)$ denoting the estimated posterior mean probability. The proportions of sites having minimal scale equal to $1, 2, 3, 4$ were $47\%, 43\%, 7\%, 3\%$, respectively.

Figure 3 shows $\hat{\text{pr}}(H_1^s \mid -)$ for these 1,696 sites. In the top left quadrant we report those sites having minimal scale equal to 1. Two patterns are evident: consistently high $\hat{\text{pr}}(H_1^s \mid -)$, with differences evident at the coarse scale, Site *cg00117172* is among those and its sample distribution is reported in panel (a) of Figure 4; moderate $\hat{\text{pr}}(H_1^s \mid -)$ for $s = 1$, with clear evidence at $s = 2$. Averages of the sites in these two groups are shown with thick dashed lines. The top right panel, representing sites having minimal scale equal to 2, presents two patterns: no differences at scale one but clear evidence of $H_1$ at scale two, Site *cg00186954* in panel (b) of Figure 4 has this behavior; moderately growing evidence for $H_1$ for increasing scale level. The bottom two panels show results for sites having minimal scale equal to 3 and 4, showing again two patterns: a group with mild or no evidence for $H_1$ up to scale 3 and 4, respectively (e.g.,

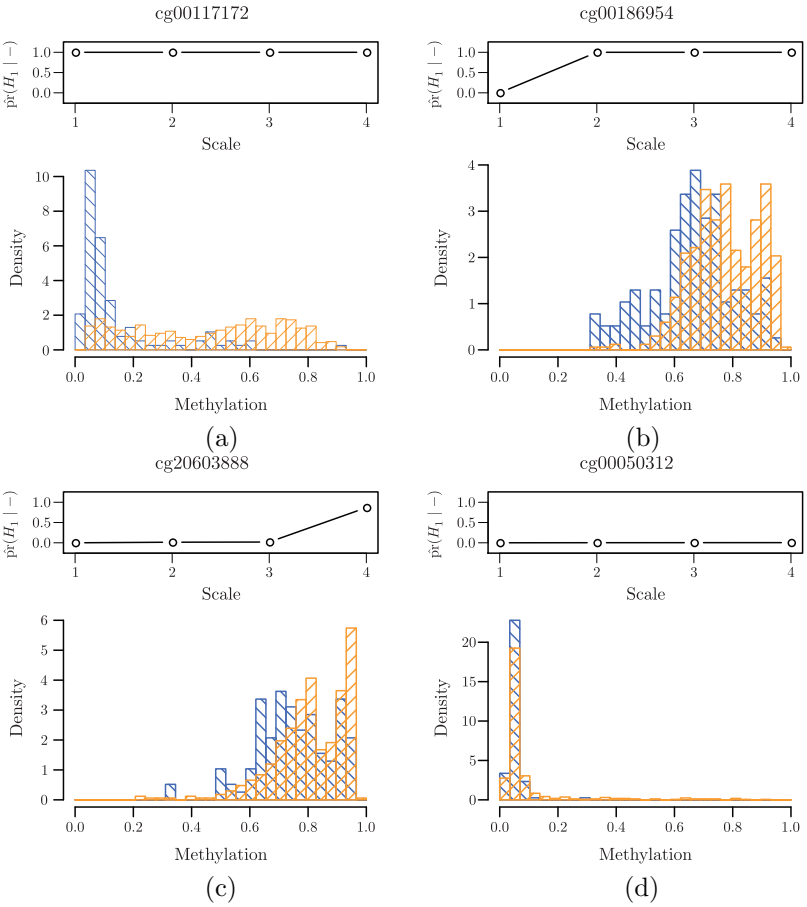


Figure 4. Histogram of the methylation for the basal (decreasing 45 degree angle shading) and non-basal (increasing 45 degree angle shading) samples for four CpG sites and posterior mean probabilities of $H_1$ in function of scale.

site *cg20603888* reported in panel (c) of Figure 4); another group with increasing evidence for increasing scale. These scale-specific significant tests are interesting in that coarser scale differences are more likely to be biologically significant, while very fine scale differences may represent local changes with minor impact. This is definitively a pro of our method, if compared to a single scale method. The latter indeed is not able to discriminate between coarse or finer differences.

## 7. Discussion

Existing Bayesian nonparametric multiscale tools for density estimation have unappealing characteristics, such as favouring overly-spiky densities. Our framework overcomes such limitations. We have demonstrated some practically appealing properties, including simplicity of formulation and ease of computation,

and proposed an extension for Bayesian multiscale hypothesis testing of group differences. Multiscale hypothesis testing is of considerable interest in itself, and provides a new view on the topic of nonparametric testing of group differences, with many interesting facets. For example, it can be argued that in large samples there will always be small local differences in the distributions between groups that may not be scientifically relevant. By allowing scale-specific tests, we accommodate the possibility of focusing inference on the range of relevant scales in an application, providing additional insight into the nature of the differences. We also accommodate scale-specific adaptive borrowing of information across groups in density estimation; extensions to include covariates and hierarchical structure are straightforward.

Although the focus has been on the univariate case, multivariate extensions are possible. A simple solution consists in substituting the beta dictionary densities with a suitable multivariate multiscale basis. In high-dimensional applications, it is not feasible to specify the basis in advance, and fully Bayesian approaches for learning the basis may be computationally infeasible. Wang, Canale, and Dunson (2016) modified the proposed msBP method to learn a basis of multivariate multiscale densities in advance using geometric multiresolution analysis. This approach had excellent performance in practice relative to competitors. Other multivariate and hierarchical extensions of the proposed msBP model are of interest in future research.

## Supplementary Materials

Supplementary materials contain additional plots and tables for Sections 2 and 4.

## Acknowledgements

The authors thanks Eric Lock for helpful comments on Section 5 and Roberto Vigo for comments on the code implementation. Comments and suggestions of the referees are gratefully acknowledged. This work was support by Award Number R01ES017436 from the National Institute of Environmental Health Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences or the National Institutes of Health.

## Appendix

*Detail on moments of $F(A)$.* The expectation of $F(A)$ is simply

$$E[F(A)] = E\left[\sum_{s=0}^{\infty}\sum_{h=1}^{2^s}\pi_{s,h}\int_A \text{Be}(y; h, 2^s - h + 1)\right]$$

$$= \sum_{s=0}^{\infty} \frac{1}{1+a} \left(\frac{a}{1+a}\right)^s \frac{1}{2^s} \sum_{h=1}^{2^s} \int_A \mathrm{Be}(y; h, 2^s - h + 1)$$
$$= \sum_{s=0}^{\infty} \frac{1}{1+a} \left(\frac{a}{1+a}\right)^s \lambda(A) = \lambda(A) \sum_{s=0}^{\infty} \frac{1}{1+a} \left(\frac{a}{1+a}\right)^s = \lambda(A),$$

where the third equality follows from the fact that the average measure over scale $s$ beta dictionary densities of any region $A$ equals the Lebesgue measure of $A$.

**Proof of Lemma 1.** For finite $N$, if $\Delta_N = 1 - \sum_{s=0}^{N} \sum_{h=1}^{2^s} \pi_{s,h}$,

$$\Delta_N = \sum_{h=1}^{2^N} \prod_{r \le N} (1 - S_{r,g_{Nhr}}) T_{Nh(r-1)} \le 2^N \max_{h=1,\ldots,2^N} \prod_{r \le N} (1 - S_{r,g_{Nhr}}) T_{Nh(r-1)}. \tag{A.1}$$

To establish (2.5), it is sufficient to take the limit of $\Delta_N$ for $N \to \infty$ and show that it converges to 0 a.s. To this end, take the logarithm of the right hand side of (A.1),

$$\log(\Delta_N) \le \max_{h=1,\ldots,2^N} \sum_{r \le N} \log\left\{2^N (1 - S_{r,g_{Nhr}}) T_{Nh(r-1)}\right\}, \tag{A.2}$$

and notice that for each $h = 1, \ldots, 2^N$ we have

$$E\left\{2^N (1 - S_{r,g_{Nhr}}) T_{Nh(r-1)}\right\} = 2^N \left(\frac{a}{a+1}\right) \frac{1}{2^N} = \frac{a}{a+1}. \tag{A.3}$$

Therefore taking $N \to \infty$, by Kolmogorov's Three Series Theorem and Jensen's Inequality, the argument of the maximum of (A.2), converges to $-\infty$ a.s. for each $h$. Thus $\Delta_N$ converges to 0 a.s. which concludes the proof.

**Proof of Lemma 2.** The $L_1$ distance can be written as

$$||p^s(y) - p^\infty(y)||$$
$$= \int \Big| \sum_{l=0}^{\infty} \sum_{h=1}^{2^l} E\left(\tilde{\pi}_{l,h} - \pi_{l,h}\right) \mathrm{Be}(y; h, 2^l - h + 1) \Big| dy$$
$$= \int \Big| \sum_{h=1}^{2^s} E[\tilde{\pi}_{s,h} - \pi_{s,h}] \mathrm{Be}(y; h, 2^s - h + 1) - \sum_{l=s+1}^{\infty} \sum_{h=1}^{2^l} E[\pi_{l,h}] \mathrm{Be}(y; h, 2^l - h + 1) \Big| dy$$
$$= \left(\frac{a}{1+a}\right)^{s+1} - \sum_{l=s+1}^{\infty} \frac{1}{1+a} \left(\frac{a}{1+a}\right)^l$$
$$= \left(\frac{a}{1+a}\right)^{s+1} - \left(\frac{a}{1+a}\right)^{s+1} = 0.$$

**Proof of Lemma 3.** First note that twice the total variation distance between two measures $P^s$ and $P$ equals the $L_1$ distance between the densities $f^s$ and $f$. The $L_1$ distance can be written as

$$\begin{aligned}
&\int \left| \sum_{l=0}^{s} \sum_{h=1}^{2^l} \tilde{\pi}_{l,h} \mathrm{Be}(y; h, 2^l - h + 1) - \sum_{l=0}^{\infty} \sum_{h=1}^{2^l} \pi_{l,h} \mathrm{Be}(y; h, 2^l - h + 1) \right| dy \\
&= \int \left| \sum_{l=0}^{\infty} \sum_{h=1}^{2^l} (\tilde{\pi}_{l,h} - \pi_{l,h}) \, \mathrm{Be}(y; h, 2^l - h + 1) \right| dy \\
&\leq \int \sum_{l=0}^{\infty} \sum_{h=1}^{2^l} \left| (\tilde{\pi}_{l,h} - \pi_{l,h}) \, \mathrm{Be}(y; h, 2^l - h + 1) \right| dy \\
&= \sum_{l=0}^{\infty} \sum_{h=1}^{2^l} |(\tilde{\pi}_{l,h} - \pi_{l,h})| \int \mathrm{Be}(y; h, 2^l - h + 1) dy \\
&= \sum_{l=0}^{\infty} \sum_{h=1}^{2^l} |(\tilde{\pi}_{l,h} - \pi_{l,h})| = \sum_{h=1}^{2^s} (\tilde{\pi}_{s,h} - \pi_{s,h}) + \sum_{l=s+1}^{\infty} \sum_{h=1}^{2^l} \pi_{l,h},
\end{aligned}$$

where the inequality holds since for each $y$ the absolute values of the sum is less than the sum of the absolute values, and since for each $h = 1 \ldots, 2^s$, $\tilde{\pi}_{s,h} \geq \pi_{s,h}$. Now taking the expectation of the above,

$$\begin{aligned}
E\Big[ \int \Big| f^s(y) - f(y) \Big| dy \Big] &\leq E\Big[ \sum_{h=1}^{2^s} (\tilde{\pi}_{s,h} - \pi_{s,h}) + \sum_{l=s+1}^{\infty} \sum_{h=1}^{2^l} \pi_{l,h}, \Big] \\
&= \sum_{h=1}^{2^s} E[\tilde{\pi}_{s,h} - \pi_{s,h}] + \sum_{l=s+1}^{\infty} \sum_{h=1}^{2^l} E[\pi_{l,h}] \\
&= \left( \frac{a}{1+a} \right)^{s+1} + \sum_{l=s+1}^{\infty} \frac{1}{1+a} \left( \frac{a}{1+a} \right)^{l} = 2 \left( \frac{a}{1+a} \right)^{s+1},
\end{aligned}$$

which leads to $E\left\{ d_{TV}(P_s, P) \right\} \leq \left\{ a/(a+1) \right\}^{s+1}$. Consider the second moment

$$\begin{aligned}
E\left[ \left\{ \int |f^s(y) - f(y)| \, dy \right\}^2 \right] &\leq E\Big[ \Big( \sum_{h=1}^{2^s} (\tilde{\pi}_{s,h} - \pi_{s,h}) + \sum_{l=s+1}^{\infty} \sum_{h=1}^{2^l} \pi_{s,h} \Big)^2 \Big] \\
&\leq 2E\Big[ \Big( \sum_{h=1}^{2^s} (\tilde{\pi}_{s,h} - \pi_{s,h}) \Big)^2 + \Big( \sum_{l=s+1}^{\infty} \sum_{h=1}^{2^l} \pi_{s,h} \Big)^2 \Big].
\end{aligned}$$

We study separately the expectations of these two summands. Again for each $h = 1 \dots, 2^s$, $\tilde{\pi}_{s,h} \geq \pi_{s,h}$, thus the first expectation is

$$\begin{aligned} E\Big\{\Big(\sum_{h=1}^{2^s} \tilde{\pi}_{s,h} - \sum_{h=1}^{2^s} \pi_{s,h}\Big)^2\Big\} &\leq E\Big\{\Big(\sum_{h=1}^{2^s} \tilde{\pi}_{s,h}\Big)^2 + \Big(\sum_{h=1}^{2^s} \pi_{s,h}\Big)^2\Big\} \\ &\leq E\Big(\sum_{h=1}^{2^s} \tilde{\pi}_{s,h} + \sum_{h=1}^{2^s} \pi_{s,h}\Big) \\ &= \left(\frac{a}{1+a}\right)^s + \frac{1}{1+a}\left(\frac{a}{1+a}\right)^s, \end{aligned}$$

where the first inequality holds removing minus twice the cross product, and the second since the quantities are strictly less than one. The second expectation is simply

$$E\Big\{\Big(\sum_{l=s+1}^{\infty}\sum_{h=1}^{2^l} \pi_{s,h}\Big)^2\Big\} \leq E\Big(\sum_{l=s+1}^{\infty}\sum_{h=1}^{2^l} \pi_{s,h}\Big) = \left(\frac{a}{1+a}\right)^{s+1}.$$

It follows that the second moment of the $L_1$ distance between $f^s$ and $f$ is less than $4\{a/(1+a)\}^s$ and thus that $\operatorname{var}\{d_{TV}(P_s, P)\} \leq 2\{a/(a+1)\}^s$.

for $j = 1, \dots, p$
  compute the trees for the node allocation according to (5.4);
  for $i = 1, \dots, n$
    assign observation $i$ at site $j$ to cluster $(s_i, h_i)$ as in Algo. 1;
  compute $n_{s,h}$, $v_{s,h}$, and $r_{s,h}$ and $n_{s,h}^{(j)}$, $v_{s,h}^{(j)}$, and $r_{s,h}^{(j)}$ for $j = 0, 1$;
  let $s_{\text{MAX}}$ the maximum occupied scale;
  for $s = 0, \dots, s_{\text{MAX}}$
    for $h = 1, \dots, 2^s$
      $S_{s,h} \sim \text{Be}(1 + n_{s,h}, a + v_{s,h} - n_{s,h})$, $R_{s,h} \sim \text{Be}(b + r_{s,h}, b + v_{s,h} - n_{s,h} - r_{s,h})$;
      $S_{s,h}^{(0)} \sim \text{Be}(1 + n_{s,h}^{(0)}, a + v_{s,h}^{(0)} - n_{s,h}^{(0)})$, $R_{s,h}^{(0)} \sim \text{Be}(b + r_{s,h}^{(0)}, b + v_{s,h}^{(0)} - n_{s,h}^{(0)} - r_{s,h}^{(0)})$;
      $S_{s,h}^{(1)} \sim \text{Be}(1 + n_{s,h}^{(1)}, a + v_{s,h}^{(1)} - n_{s,h}^{(1)})$, $R_{s,h}^{(1)} \sim \text{Be}(b + r_{s,h}^{(1)}, b + v_{s,h}^{(1)} - n_{s,h}^{(1)} - r_{s,h}^{(1)})$;
  compute the trees of weights under $H_0$ and $H_1$ for the two groups
  for $s = 0, \dots, s_{\text{MAX}}$
    compute $P_m^s = \text{pr}(H_0^s | \mathcal{N}_{(0)}^s, \mathcal{N}_{(1)}^s)$ as in (5.3);
draw $P_0^s \sim \text{Be}(1 + \sum_{j=1}^p P_j^s, 1 + p - \sum_{j=1}^p P_j^s)$.

**Algorithm 3:** Gibbs sampler steps for posterior computation for multiscale hypothesis testing of group differences using msBP prior.

## References

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc. Ser. B* **60**, 725-749.

Adams, R. P., Ghahramani, Z. and Jordan, M. I. (2010). Tree-structured stick breaking for hierarchical data. *Adv. Neural Infor. Process. Systems* **23**, 19-27.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **460**, 61-70.

Chen, G. L., Iwen, M., Chin, S. and Maggioni, M. (2012). A fast multiscale framework for data in high-dimensions: Measure estimation, anomaly detection and compressive measurements. *IEEE Visual Comm. Image Process.*, 1-12.

Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *J. Amer. Statist. Assoc.* **104**, 1646-1660.

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. Roy. Statist. Soc. Ser. B* **62**, 681-698.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391-401.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24**, 508-539.

Garcia-Trevino, E. S. and Barria, J. A. (2012). Online wavelet-based density estimation for non-stationary streaming data. *Comput. Statist. Data Anal.* **56**, 327-344.

Hanson, T. E. and Johnson, W. O. (2002). Modeling regression error with a mixture of Polya trees. *J. Amer. Statist. Assoc.* **97**, 1020-1033.

Hanson, T. E. (2006). Inference for mixtures of finite Polya tree models. *J. Amer. Statist. Assoc.* **101**, 1548-1565.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick breaking priors. *J. Amer. Statist. Assoc.* **96**, 161-173.

Kalli, M., Griffin, J. E. and Walker, S. G. (2011). Slice sampling mixture models. *Stat. Comput.* **21**, 93-105.

Lavine, M. (1992a). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **20**, 1222-1235.

Lavine, M. (1992b). More aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **22**, 1161-1176.

Lock, E. F. and Dunson, D. B. (2015). Shared kernel Bayesian screening. *Biometrika* **102**, 829-842.

Locke, J. B. and Peter, A. M. (2013). Multiwavelet density estimation. *Appl. Math. Comput.* **219**, 6002-6015.

Mauldin, D., Sudderth, W. D. and Williams, S. C. (1992). Polya trees and random distributions. *Ann. Statist.* **20**, 1203-1203.

Niu, S. L. (2012). Nonlinear wavelet density estimation with censored dependent data. *Math. Methods Appl. Sci.* **35**, 293-306.

Pati, D., Dunson, D. B. and Tokdar, S. (2013). Posterior consistency in conditional distribution estimation. *J. Multivariate Anal.* **116**, 456-472.

Pensky, M. and Vidakovic, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.* **27**, 2033-2053.

Petrone, S. (1999a). Bayesian density estimation using Bernstein polynomials. *Canad. J. Statist.* **27**, 105-126.

Petrone, S. (1999b). Random Bernstein polynomials. *Scand. J. Statist.* **26**, 373-393.

Ren, L., Du, L., Carin, L. and Dunson, D. B. (2011). Logistic stick-breaking process. *J. Machine Learn. Res.* **12**, 203-239.

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal.* **6**, 145-177.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 639-650.

Wang, X., Ray, S. and Mallick, B. K. (2007). Bayesian curve classification using wavelets. *J. Amer. Statist. Assoc.* **102**, 962-973.

Wang, Y., Canale, A. and Dunson, D. B. (2016). Scalable multiscale density estimation. In *Artificial Intelligence and Statistics* (AISTATS).

Department of Economics and Statistics, University of Turin, Turin 10134, Italy.

Collegio Carlo Alberto, Moncalieri, Italy.

E-mail: antonio.canale@unito.it

Department of Statistical Sciences, Duke University, Durham, NC 27708, USA.

E-mail: dunson@stat.duke.edu

(Received May 2015; accepted September 2015)