Bayesian nonparametric multiscale mixture models via Hilbert-curve partitioning

Daniele Zago[a,*], Marco Stefanucci[a,b], Antonio Canale[a]

[a] *Department of Statistical Sciences, University of Padova, Padova, Italy*
[b] *Department of Economic, Business, Mathematical and Statistical Sciences, University of Trieste, Trieste, Italy*

Abstract

Bayesian nonparametric multivariate density estimation typically relies on mixture specifications, exceptions made for Pólya tree constructions. Herein, we develop a multivariate mixture model exploiting the multiscale stick-breaking prior recently proposed by Stefanucci and Canale (2021). The building block of the proposed approach is a base measure defined exploiting the Hilbert space-filling curve which allows to adapt a simple partitioning of a univariate parameter space to the multivariate case with minor adjustments. Alongside the theoretical discussion, we illustrate the model's performance by analyzing both synthetic and real datasets. The results suggest that the proposed multiscale model achieves competitive performance with respect to state-of-the-art Bayesian nonparametric methods both in scenarios presenting single- and multi-scale features.

Keywords Bayesian nonparametrics; multiscale models; density estimation; Dirichlet process; Pitman-Yor process

1. Introduction

Nonparametric models typically characterize the unknown data distribution F through an infinite-dimensional statistical model. Bayesian nonparametric models (BNP) require the definition of a prior distribution over infinite-dimensional probability spaces in order to obtain an inferential procedure which is robust to model misspecifications.

2. Multiscale mixture models

Let $Y \in \mathcal{Y} \subseteq \mathbb{R}^d$ be a random variable. We assume a multiscale representation for the density f of Y,

*Corresponding author
Email addresses: daniele.zago.1@studenti.unipd.it (Daniele Zago), stefanucci@stat.unipd.it (Marco Stefanucci), canale@stat.unipd.it (Antonio Canale)

$$f(y) = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} \mathcal{K}(y; \boldsymbol{\vartheta}_{s,h}), \tag{1}$$

where $\mathcal{K}(\cdot; \boldsymbol{\vartheta})$ is a kernel function defined on $\mathcal{Y}$ and parametrized by $\boldsymbol{\vartheta}$. $\{\pi_{s,h}\}$ is a sequence of mixture weights such that $\sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} = 1$, and $\{\boldsymbol{\vartheta}_{s,h}\} \subseteq \Theta$ is a sequence of kernel parameters. This construction is reminiscent of the stick-breaking representation of the Dirichlet process mixture model (**ferguson1983**), and it allows us to relate the representation of the density to an infinitely deep binary tree, whose nodes are indexed by a pair of indices (s, h). Thus, each node is uniquely identified by the set of parameters $\{(\pi_{s,h}, \boldsymbol{\vartheta}_{s,h})\}$. Figure 1 shows a truncation of such a binary tree at depth $s = 3$.

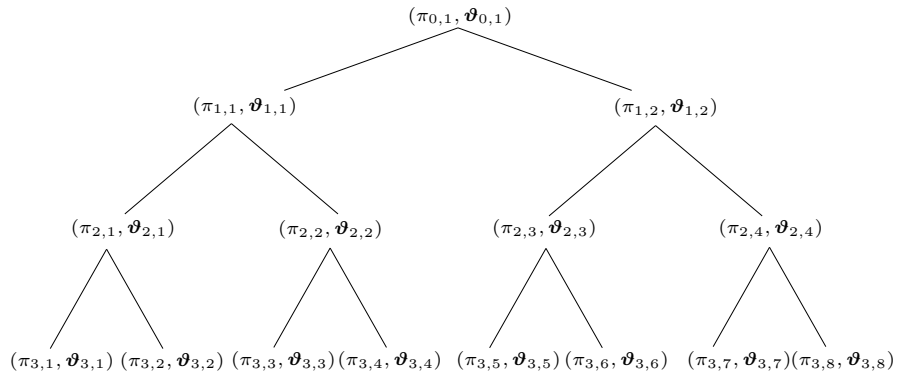

Figure 1: Binary tree representation of the mixture model, truncated at scale $s = 3$. At each node, $\boldsymbol{\vartheta}_{s,h}$ and $\pi_{s,h}$ are the associated kernel parameters and mixture weight, respectively.

The density in (1) can be equivalently rewritten as a Lebesgue integral of the kernel function with respect to a discrete mixing measure P,

$$f(y) = \int \mathcal{K}(y; \boldsymbol{\vartheta}) \, dP(\boldsymbol{\vartheta}), \quad P = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} \delta_{\boldsymbol{\vartheta}_{s,h}}, \tag{2}$$

where $\delta_{\boldsymbol{\vartheta}_0}$ denotes the Dirac delta function centered in $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$.

We can place a nonparametric prior distribution on f by specifying a stochastic process that generates the mixing measure P or, equivalently, the sequence of parameters $\{(\pi_{s,h}, \boldsymbol{\vartheta}_{s,h})\}$. We will therefore begin the specification of this stochastic process by discussing the generation of the prior mixing weights $\{\pi_{s,h}\}$.

2.1. Mixture weights

The sequence of mixture weights $\{\pi_{s,h}\}$ is generated by the stochastic process introduced in Canale and Dunson (2016). The construction closely resembles the

stick-breaking construction of the Dirichlet process (Sethuraman 1994) and has been adapted in order to be employed in a binary tree setting. More specifically, let $S_{s,h}$ and $R_{s,h}$ be random variables that take values in $(0, 1)$. $S_{s,h}$ denotes the probability of stopping at node (s, h), whereas $R_{s,h}$ denotes the probability of taking the right path from node (s, h) to node $(s + 1, 2h)$, conditionally on not stopping on node (s, h). We also introduce an auxiliary variable T_{shr} such that

$$T_{shr} = \begin{cases} R_{r,\lceil h2^{r-s}\rceil} & \text{if } \lceil h2^{r-2+1}\rceil = 2h \\ 1 - R_{r,\lceil h2^{r-s}\rceil} & \text{otherwise} \end{cases}.$$

The mixture weights are then defined as

$$\pi_{s,h} = S_{s,h} \prod_{r<s} \big(1 - S_{r,\lceil h2^{r-s}\rceil}\big) T_{shr}. \tag{3}$$

Equation (3) describes a two-stage stick-breaking process that generates the mixture weights: the stick of length one is initially broken according to the distribution of $S_{0,1}$, and its value is given to the first weight $\pi_{0,1}$. The remainder of the stick, whose length is now $1 - S_{0,1}$, is then randomly split into two parts according to the distribution of $R_{0,1}$. The two parts are then assigned to the children of the parent node, in this case $(1, 1)$ and $(1, 2)$, and the process is repeated. The auxiliary variable T_{shr} is useful to keep track of the differing stick lengths between the left and right children of each node.

The resulting algorithm allows us to generate all the mixture weights and guarantees that the sequence of weights $\{\pi_{s,h}\}$ is such that (Canale and Dunson 2016)

$$\sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} = 1 \quad \text{almost surely.}$$

By way of analogy to the stick-breaking construction of Ishwaran and James (2001), the random variables $S_{s,h}$ and $R_{s,h}$ are given a prior distribution

$$S_{s,h} \sim \text{Beta}(a_s, b_s), \quad R_{s,h} \sim \text{Beta}(c_s, c_s),$$

where the parameters a_s, b_s, c_s define the behaviour of the weight-generating process. This is similar to the way the behaviour of the Dirichlet process and its generalization, the Pitman-Yor process, depend on the choice of the parameters of the underlying Beta distribution. Indeed, for $\delta \in [0, 1)$ and $\alpha > -\delta$, we can choose the distributions

$$S_{s,h} \sim \text{Beta}\big(1 - \delta, \alpha + \delta(s + 1)\big), \quad R_{s,h} \sim \text{Beta}(\beta, \beta), \tag{4}$$

in order to mimic the stick-breaking representation of the Pitman-Yor process. Stefanucci and Canale (2021) provide an interpretation of the stick-breaking distributions in (4), by observing that δ plays a similar role to the discount parameter σ of the stick-breaking representation Pitman-Yor process. In this

case, δ controls the prior expected scale at which an observation falls, $\mathbb{E}\big[\tilde{S}\big] = \sum_{s=0}^{\infty} s\mathbb{E}\big[\pi_{s,h}\big]$. More formally, we observe that the expected value of $S_{s,h}$ is heavily influenced by δ, since

$$\mathbb{E}\big[S_{s,h}\big] = \frac{1-\delta}{\alpha + \delta(s+1)}.$$

We can easily see that increasing δ results in a lower expected value for $S_{s,h}$; this, in turn means that the unitary stick is, on average, broken by smaller pieces at each scale. Therefore, a larger value of δ favours *a priori* deeper trees over shallower ones. Following the discussion in Stefanucci and Canale (2021), we can numerically compute suitable values of α and δ in order to guarantee a desired prior expected tree depth $\mathbb{E}\big[\tilde{S}\big]$.

2.2. Kernel parameters

In this section we will present a multivariate generalization of the stochastic processes developed in Stefanucci and Canale (2021) for mixtures of univariate kernels. In their exposition, they consider a parameter space of the form $\Theta = \Theta_\mu \times \Theta_\omega \subseteq \mathbb{R} \times \mathbb{R}^+$, so that $\{\boldsymbol{\vartheta}_{s,h}\} = \{(\mu_{s,h}, \omega_{s,h})\} \subseteq \Theta$ is a sequence of location and scale parameters. The resulting model is a mixture of univariate location and scale kernels $\mathcal{K}(\cdot; \mu_{s,h}, \omega_{s,h})$, for which they provide a suitable stochastic process that generates each component separately. We will now consider the multivariate generalization of this location and scale setting, where we denote by $\mathcal{K}(\cdot; \boldsymbol{\mu}_{s,h}, \Omega_{s,h})$ the d-dimensional kernel defined on the sample space $\mathcal{Y}$. In this case, the parameter space is $\Theta = \Theta_{\boldsymbol{\mu}} \times \Theta_\Omega \subseteq \mathbb{R}^d \times \mathcal{M}_d$, where $\mathcal{M}_d$ is the set of positive-definite square matrices of dimension $d \times d$. In the general case, we are interested in constructing suitable stochastic processes to separately generate the sequence of location parameters $\{\boldsymbol{\mu}_{s,h}\}$ and scale parameters $\{\Omega_{s,h}\}$.

2.2.1. Scale parameters

We assume the kernel parameters $\{\Omega_{s,h}\}$ to be proportional to the variance-covariance matrix of the kernel. Under this assumption we sample each parameter from

$$\Omega_{s,h} = C(s)W_{s,h}, \quad W_{s,h} \overset{i.i.d.}{\sim} H_0. \tag{5}$$

In Equation (5), $C(s) = \text{diag}\left(c^{(1)}(s), c^{(2)}(s), \ldots, c^{(d)}(s)\right)$ is a $d \times d$ matrix such that each sequence $c^{(j)}(s)$, $j = 1, 2, \ldots, d$ is monotone decreasing in s. The resulting sequence of multivariate scale parameters is componentwise stochastically decreasing. Furthermore, in absence of prior information we have a natural choice of $C(s)$ given by specifying $c^{(j)}(s) = 2^{-s}$ for all j. On the other hand, under prior information we could choose different sequences in order to obtain a different behaviour for each variable separately. Using the specification in Equation (5) we have that, for each diagonal component $\omega_{s,h}^{(j,j)}$ of $\Omega_{s,h}$,

$$\mathbb{E}_{H_0}\big[\omega_{s+1,h}^{(j,j)}\big] \leq \mathbb{E}_{H_0}\big[\omega_{s,h}^{(j,j)}\big], \quad \mathbb{V}_{H_0}\big[\omega_{s+1,h}^{(j,j)}\big] \leq \mathbb{V}_{H_0}\big[\omega_{s,h}^{(j,j)}\big].$$

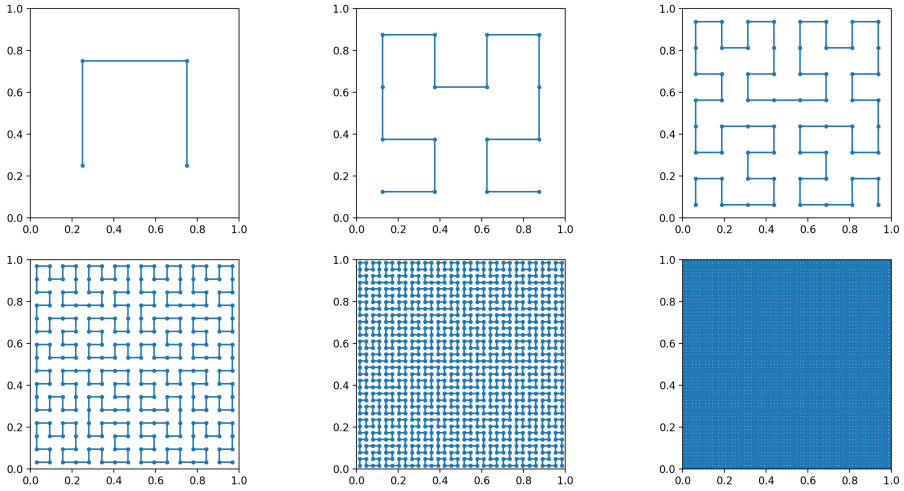

Figure 2: First six orders of the 2-dimensional approximate Hilbert curve, from H_1 to H_6.

Recalling that $\Theta_{s,h}$ is proportional to the variance-covariance matrix of the kernel, we find that the kernels become on average more concentrated both in mean as well as in variability as s increases.

2.2.2. Location parameters

Consistently with the construction of Stefanucci and Canale (2021), we consider a partition of the parameter space $\Theta_{\boldsymbol{\mu}} \subseteq \mathbb{R}^d$ such that, for any scale s of the binary tree, the whole space is spanned by the nodes at that scale, i.e.

$$\Theta_{\boldsymbol{\mu}} = \bigcup_{h=1}^{2^s} \Theta_{\boldsymbol{\mu};s,h}, \tag{6}$$

and

$$\Theta_{\boldsymbol{\mu};s,h} = \Theta_{\boldsymbol{\mu};s+1,2h-1} \cup \Theta_{\boldsymbol{\mu};s+1,2h}. \tag{7}$$

Although there is an infinite number of binary partitions of $\Theta_{\boldsymbol{\mu}}$, we are interested in partitions which allows us to accurately relate the binary tree representation to the estimated model. We choose such a binary partition by introducing the Hilbert curve (Hilbert 1891), a space-filling curve $H : [0,1] \to [0,1]^d$ which is surjective and displays a high degree of *locality* (Moon et al. 2001). Indeed, given two points $x, x' \in [0,1]$ which are close to each other, then their image $H(x)$ and $H(x')$ are close as well on the codomain $[0,1]^d$. The function H is obtained by computing the limit of a sequence of approximate Hilbert curves $(H_m)_{m\in\mathbb{N}}$ as $m \to +\infty$. Figure 2 illustrates the first six approximate Hilbert curves in two dimensions.

Let $\mathcal{I}_m^d$ be the set of consecutive subintervals of $[0,1]$ of length $1/2^{dm}$,

$$\begin{aligned}\mathcal{I}_m^d &= \left\{I_m^d(k), k = 0, 1, \ldots, 2^{dm} - 1\right\} \\ &= \left\{\left[\frac{k}{2^{dm}}, \frac{k+1}{2^{dm}}\right], k = 0, 1, \ldots, 2^{dm} - 1\right\},\end{aligned}$$

then we have that $\left\{H_m\big(I_m^d(k)\big)\right\}$ is a sequence of subcubes of $[0,1]^d$ which have the following properties:

- (*Bijection*): Each subinterval is mapped to a different section of the space,

$$H_m\big(I_m^d(k)\big) \neq H_m\big(I_m^d(k')\big) \quad \text{for } k \neq k'$$

- (*Adjacency*): For any k, the two successive subcubes $H_m\big(I_m^d(k)\big)$ and $H_m\big(I_m^d(k+1)\big)$ have exactly a $(d-1)$-dimensional face in common.

- (*Nesting*): Increasing the curve order from m to $m+1$ corresponds to partitioning of each cube $H_m\big(I_m^d(k)\big)$ into 2^d subcubes. More specifically, for $k_i = 2^d k + i$, $i = 0, 1, \ldots, 2^d - 1$, the union of the 2^d successive subcubes $H_{m+1}\big(I_{m+1}^d(k_i)\big)$ yields the cube $H_m\big(I_m^d(k)\big)$.

In order to obtain a binary partition of $[0,1]^d$ at each scale s, we set $m = \lceil s/d \rceil$ and we join together every $2^{|d-s| \bmod d}$ subcubes. This produces a valid dyadic partition by virtue of the bijection, adjacency, and nesting properties of the Hilbert curve. Moreover, since successive subcubes are adjacent, we partition the subrectangles along the dimension that each pairwise group of $2^{|d-s| \bmod d}$ subcubes do not share. Figure 3 shows the proposed binary partition scheme in the two-dimensional case. Following the ordering of the centers of the approximate Hilbert curve, we obtain an ordering in the subrectangles of $[0,1]^d$. At each fixed depth s, we can assign each subrectangle from left to right to the binary nodes in Figure 1. This allows us to construct a binary tree partition of $[0,1]^d$ such that properties (6) and (7) are both satisfied.

Let now G_0 be a base probability measure on $\Theta_{\boldsymbol{\mu}}$. Once the space $[0,1]^d$ has been partitioned into subrectangles, we can obtain a suitable partition of $\mathbb{R}^d$ simply by applying the quantile function of each conditional distribution of G_0 to the extremes of the interval $[\boldsymbol{a}_{s,h}, \boldsymbol{b}_{s,h}]$, i.e.

$$\Theta_{\mu^{(j)};s,h} = [q^{(j)}_{a^{(j)}_{s,h}}, q^{(j)}_{b^{(j)}_{s,h}}], \quad j = 1, \ldots, d, \tag{8}$$

where $q^{(j)}(\cdot)$ is the conditional quantile function of G_0 for the j-th component. Finally, we obtain the complete space partition by considering the Cartesian product

$$\Theta_{\boldsymbol{\mu};s,h} = \Theta_{\mu^{(1)};s,h} \times \Theta_{\mu^{(2)};s,h} \times \cdots \times \Theta_{\mu^{(d)};s,h}.$$

With this choice of partition, the locality property of the Hilbert curve ensures that nodes which are close on the binary tree refer to partitions which are

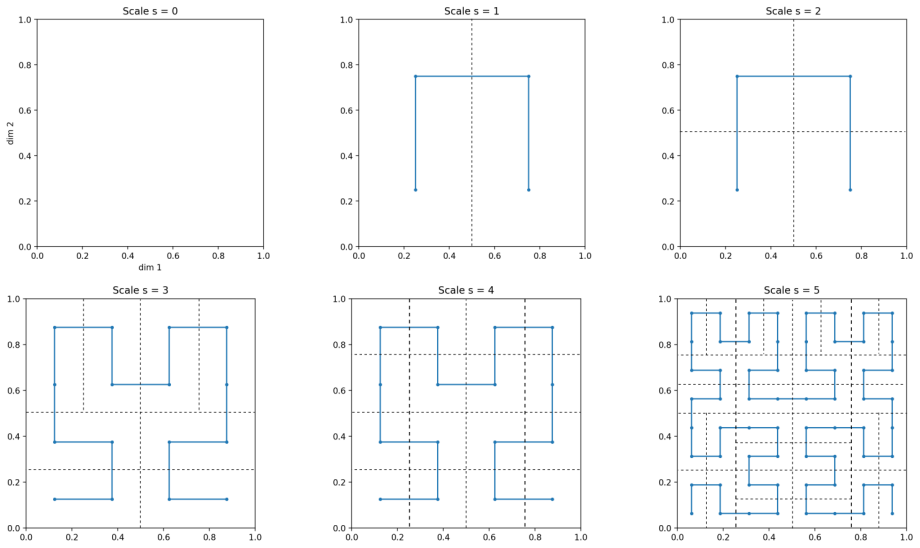

Figure 3: Dyadic partition of the two-dimensional square obtained by the application of the Hilbert curve, for the first six scales of the binary tree.

close on $\mathbb{R}^d$. This allows us to inspect the posterior distribution of the weights of the binary tree to infer the shape of the estimated multidimensional density.

We sample each location parameter $\boldsymbol{\mu}_{s,h}$ proportionally to G_0 truncated to the set $\Theta_{\boldsymbol{\mu};s,h}$, so that the random probability measure

$$G = \sum_{s=0}^{\infty} \sum_{h=1}^{2^s} \pi_{s,h} \delta_{\boldsymbol{\mu}_{s,h}} \tag{9}$$

defined on $\Theta_{\boldsymbol{\mu}}$ is centered around G_0, as shown in Theorem 1.

Theorem 1. *Let G_0 be a base probability measure defined on $\Theta_{\boldsymbol{\mu}}$, and consider a dyadic partition of $\Theta_{\boldsymbol{\mu}}$ such as the one defined by Equations (6), (7), and (8). If $\boldsymbol{\mu}_{s,h}$ is randomly sampled proportionally to G_0 truncated to $\Theta_{\boldsymbol{\mu};s,h}$, then, for any subset $A \subseteq \Theta_{\boldsymbol{\mu}}$,*

$$\mathbb{E}\big[G(A)\big] = G_0(A). \tag{10}$$

Proof.

$$
\begin{aligned}
\mathbb{E}\big[G(A)\big] &= \mathbb{E}\Big[\sum_{s=0}^{\infty}\sum_{h=1}^{2^s}\pi_{s,h}\delta_{\boldsymbol{\mu}_{s,h}}(A)\Big] \\
&= \sum_{s=0}^{\infty}\sum_{h=1}^{2^{sz}}\mathbb{E}\big[\pi_{s,h}\big]G_0(A\cap\Theta_{\boldsymbol{\mu};s,h}) \\
&= \sum_{s=0}^{\infty}\frac{(1-\delta)\prod_{j=0}^{s-1}(\alpha+\delta(j+1))}{\prod_{j=0}^{s}(\alpha+\delta j+1)}\sum_{h=1}^{2^s}G_0(A\cap\Theta_{\boldsymbol{\mu};s,h}) \\
&= G_0(A)\sum_{s=0}^{\infty}\frac{(1-\delta)\prod_{j=0}^{s-1}(\alpha+\delta(j+1))}{\prod_{j=0}^{s}(\alpha+\delta j+1)} \\
&= G_0(A).
\end{aligned}
$$

□

3. Posterior computation

In this section we describe in detail the Markov Chain Monte Carlo algorithm we use to perform posterior inference for the multiscale model introduced in Section Section 2. The posterior computation for the multiscale stick-breaking mixture model is based on a blocked Gibbs sampler algorithm developed by Stefanucci and Canale (2021), where the kernels are assumed to be Gaussian for ease of derivation. In this section, we also focus on the particular case of a Gaussian location-scale mixture of kernels, for which an extension of the aforementioned Gibbs sampler is straightforward.

Let $\pi_s = \sum_{h=1}^{2^s}\pi_{s,h}$ be the total probability mass assigned to scale s and $\bar{\pi}_{s,h} = \pi_{s,h}/\pi_s$ the proportion of mass associated to node (s,h). For the i-th observation we can rewrite the likelihood as

$$
f(y_i) = \sum_{s=0}^{\infty}\pi_s\sum_{h=1}^{2^s}\bar{\pi}_{s,h}\mathcal{K}(y_i;\vartheta_{s,h}).
$$

Let now (s_i,h_i) be auxiliary variables that indicate the node to which subject i is allocated. Conditionally on s_i and h_i, we have that

$$
f(y_i;s_i,h_i)\propto\mathcal{K}(y_i;\vartheta_{s_i,h_i}).
$$

Following (Kalli, Griffin, and Walker 2011), we introduce an auxiliary random variable $u_i|y_i,s_i\sim\text{Unif}(0,\pi_{s_i})$, and write the joint density for (y_i,u_i,s_i) as follows:

$$
f(y_i,u_i,s_i)\propto\mathbb{1}_{(0,\pi_{s_i})}(u_i)\sum_{h=1}^{2^{s_i}}\bar{\pi}_{s_i,h}\mathcal{K}(y_i;\vartheta_{s_i,h_i}).
$$

Then, we can sample cluster indicators s_i and h_i by sampling from the following conditional distributions:

$$\mathbb{P}(s_i = s|u_i, y_i) \propto \mathbb{1}_{(u_i,1)}(\pi_s) \sum_{h=1}^{2^s} \bar{\pi}_{s,h} \mathcal{K}(y_i; \vartheta_{s,h}), \tag{11}$$

$$\mathbb{P}(h_i = h|u_i, y_i, s_i) \propto \bar{\pi}_{s_i,h} \mathcal{K}(y_i; \vartheta_{s,h}). \tag{12}$$

Conditionally on the sampled cluster indicators, the mixture weights are sampled from

$$S_{s,h}|- \sim \text{Beta}\big(1 - \delta + n_{s,h}, \alpha + \delta(s+1) + v_{s,h} - n_{s,h}\big), \tag{13}$$

$$R_{s,h}|- \sim \text{Beta}\big(\beta + r_{s,h}, \beta + v_{s,h} - n_{s,h} - r_{s,h}\big). \tag{14}$$

Where $n_{s,h}$ is the number of subjects allocated to node (s,h), $r_{s,h}$ is the number of subjects that continue to the right after passing through node (s,h), and $v_{s,h}$ is the number of subject passing through node (s,h) (Canale and Dunson 2016). Conditionally on cluster allocations, the update of kernel parameters depends on model specification. In this section, we propose two versions of the MSM of Gaussian densities, namely the *product kernel* and the *full-covariance kernel*. Whereas the former assumes a diagonal covariance matrix $\Omega_{s,h} = \text{diag}\left(\omega_{s,h}^{(j,j)}\right)$ for the Gaussian kernel, the latter allows a full covariance matrix.

3.1. Product kernel

Under the product kernel specification, we sample each component of the location parameter from

$$\mu_{s,h}^{(j)}|- \sim N_{\Theta_{\mu^{(j)};s,h}}\left(\frac{\mu_0 \omega_{s,h}^{(j,j)} + n_{s,h}\bar{y}_{s,h}^{(j)}\kappa^{(j)}}{n_{s,h}\kappa^{(j)} + \omega_{s,h}^{(j)}}, \frac{\omega_{s,h}^{(j,j)}\kappa^{(j)}}{n_{s,h}\kappa^{(j)} + \omega_{s,h}^{(j,j)}}\right). \tag{15}$$

where $\bar{y}_{s,h} = \left(\bar{y}_{s,h}^{(1)}, \bar{y}_{s,h}^{(2)}, \ldots, \bar{y}_{s,h}^{(d)}\right)$ is the sample mean of the observations assigned to node (s,h), and $N_A(m,v)$ denotes the Gaussian distribution with mean m and variance v truncated to the set A. Similarly, we update the scale parameters by sampling from

$$\omega_{s,h}^{(j,j)}|- \sim \text{IGa}\left(k + \frac{n_{s,h}}{2}, \frac{\lambda}{2^s} + \frac{\sum_{i:s_i=s,h_i=h}(y_i^{(j)} - \mu_{s,h}^{(j)})^2}{2}\right). \tag{16}$$

This specification of the multiscale model is simply an application of the original univariate multiscale model in (Stefanucci and Canale 2021) to each coordinate independently.

3.2. Full-covariance kernel

Under the full-covariance kernel specification, the application of conjugate Bayesian analysis yields straightforward update rules. More specifically, by assuming base measures $G_0 = MVN(\boldsymbol{\mu}_0, \Sigma_0)$ and $H_0 = \mathcal{IW}(\nu_0, \Psi)$ we sample the posterior scale parameters from

$$\Omega_{s,h}|- \sim \mathcal{IW}\left(\nu_0 + n_{s,h} + 2^s, \Psi + \sum_{\substack{i:\ s_i=s,\\ h_i=h}} (y_i - \boldsymbol{\mu}_{s,h})(y_i - \boldsymbol{\mu}_{s,h})^\top\right),$$

and the location parameters from

$$\boldsymbol{\mu}_{s,h} \sim \mathrm{MVN}_{\Theta_{\boldsymbol{\mu};s,h}}\left(\boldsymbol{m}_{s,h}, \Sigma_{s,h}\right),$$

where $\mathrm{MVN}_A(\boldsymbol{m}, \Sigma)$ denotes the multivariate normal distribution of mean $\boldsymbol{m}$ and variance-covariance matrix Σ truncated to the set A, and

$$\begin{aligned}\boldsymbol{m}_{s,h} &= (\Sigma_0^{-1} + n_{s,h}\Omega_{s,h}^{-1})^{-1}(\Sigma_0^{-1}\mu_0 + n_{s,h}\Omega_{s,h}^{-1}\bar{y}_{s,h}),\\ \Sigma_{s,h} &= (\Sigma_0^{-1} + n_{s,h}\Omega_{s,h}^{-1})^{-1}.\end{aligned}$$

4. Simulation study

In this section we conduct a simulation study to investigate the performance of the proposed multiscale model with respect to the known underlying distribution. Moreover, we compare the multiscale method against its Bayesian nonparametric single-scale counterpart, the Pitman-Yor (PY) model (Pitman and Yor 1997). We generate samples $i = 1, \ldots, N$ with different sample sizes n from different true distributions, which are obtained as mixtures of multidimensional normal and skew-normal distributions. **??** shows a two-dimensional representation of the considered scenarios, which present an increasingly evident multiscale structure. ...

5. Discussion

In this article we generalized the multiscale stick-breaking mixture model introduced by Stefanucci and Canale (2021) to the multivariate setting. This generalization retains the original binary-tree structure even when considering d-dimensional kernels for which $d \geq 2$. We were able to do so by exploiting the ordering induced onto the d-dimensional space induced by the Hilbert curve in order to traverse the partitions of the location space $\Theta_{\boldsymbol{\mu}}$ by means of a binary tree. This allowed us to straightforwardly generalize the Gibbs sampler algorithm via conjugate arguments in the multivariate Gaussian specification of the multiscale model. When compared to a standard single-scale Bayesian nonparametric competitor such as the Pitman-Yor process, our proposed multiscale model showed competitive performance in terms of the LPML criterion, when equipped with a full-covariance kernel, especially when the density displayed an evident multiscale structure.

References

Canale, Antonio and David B. Dunson (2016). "Multiscale Bernstein Polynomials for Densities". In: *Statistica Sinica* 26.3, pp. 1175–1195. ISSN: 1017-0405.

Hilbert, David (Sept. 1891). "Über die stetige Abbildung einer Line auf ein Flächenstück". In: *Mathematische Annalen* 38.3, pp. 459–460. ISSN: 1432-1807. DOI: 10.1007/BF01199431.

Ishwaran, Hemant and Lancelot F. James (Mar. 2001). "Gibbs Sampling Methods for Stick-Breaking Priors". In: *Journal of the American Statistical Association* 96.453, pp. 161–173. ISSN: 0162-1459. DOI: 10.1198/016214501750332758.

Kalli, Maria, Jim E. Griffin, and Stephen G. Walker (Jan. 2011). "Slice Sampling Mixture Models". In: *Statistics and Computing* 21.1, pp. 93–105. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-009-9150-y.

Moon, B. et al. (Jan. 2001). "Analysis of the Clustering Properties of the Hilbert Space-Filling Curve". In: *IEEE Transactions on Knowledge and Data Engineering* 13.1, pp. 124–141. ISSN: 1558-2191. DOI: 10.1109/69.908985.

Pitman, Jim and Marc Yor (Apr. 1997). "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator". In: *The Annals of Probability* 25.2, pp. 855–900. ISSN: 0091-1798, 2168-894X. DOI: 10.1214/aop/1024404422.

Sethuraman, Jayaram (1994). "A Constructive Definition of Dirichlet Priors". In: *Statistica Sinica* 4.2, pp. 639–650. ISSN: 1017-0405.

Stefanucci, Marco and Antonio Canale (Jan. 2021). "Multiscale Stick-Breaking Mixture Models". In: *Statistics and Computing* 31.2. ISSN: 1573-1375. DOI: 10.1007/s11222-020-09991-1.