Bayesian nonparametric multiscale mixture models via Hilbert-curve partitioning

Daniele Zago

Department of Statistical Sciences, University of Padova

Joint work with Antonio Canale, University of Padova & Marco Stefanucci, Sapienza Università di Roma

UNIVERSITÀ DEGLI STUDI DI PADOVA

Abstract

We consider the problem of flexible **nonparametric density estimation** using mixtures of densities.

We are motivated by **astrological applications**, where galaxies might be clustered based on their colour spectrum.

We rely on a **multiscale mixture model for the density** in order to **cluster observations at different resolutions**

The multiscale structure is described by using a **sequence of Hilbert curves** in order to **map the multivariate space to a binary tree**

The resulting mixture is **flexible** and can **adapt** very well **to the underlying smoothness** of the data

Background

Bayesian nonparametric univariate multiscale models

Univariate multiscale models define a mixture of kernels indexed by a binary tree,

$$f(y) = \sum_{s=0}^{\infty}\sum_{h=1}^{2^s} \pi_{s,h}\mathcal{K}(y; \mu_{s,h}, \Omega_{s,h}),$$

where (s, h) **corresponds to a node of a binary tree** (Figure 1) and $\mathcal{K}$ is a scale and location kernel.

$\mathcal{K}$ is **increasingly concentrated** as s increases, and the location parameter $\mu_{s,h}$ should **span the whole space** as h moves between the values $1, \dots, 2^s$.

A nonparametric prior distribution for the mixture weights $\pi_{s,h}$ has been developed by Canale and Dunson [2016], whereas a Bayesian nonparametric univariate mixture of kernels has been introduced by Stefanucci and Canale [2021].

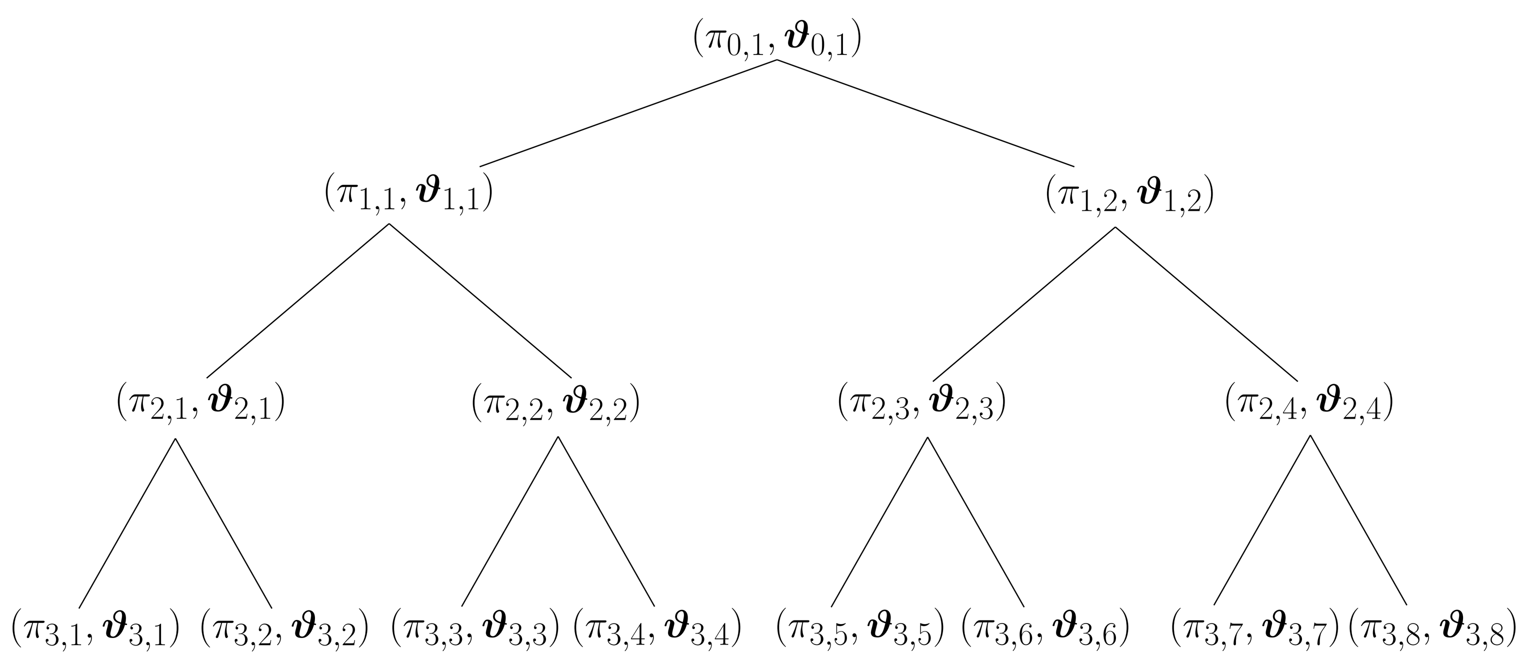

Figure 1: Binary tree associated to the multiscale model.

☺ Pro: flexibility and adaptability to data smoothness
☹ Con: how do we generalize the binary tree to multivariate kernels?

Problem

Generalizing the model to the multivariate case $\mathbb{R}^d$ seems hard, because the computational complexity of a d-tree will explode very quickly.

Solution: Hilbert curve partitioning

We partition the location space $\Theta_{\boldsymbol{\mu}}$ so that we span the whole space in h and the partitions are nested,

$$\Theta_{\boldsymbol{\mu}} = \bigcup_{h=1}^{2^s} \Theta_{\boldsymbol{\mu};s,h}, \quad \Theta_{\boldsymbol{\mu};s,h} = \Theta_{\boldsymbol{\mu};s+1,2h-1} \cup \Theta_{\boldsymbol{\mu};s+1,2h}.$$

The partition is done using the following two-stage procedure based on the Hilbert curve [Hilbert, 1891]:

– Partition the cube $[0, 1]^d$ using the Hilbert curve (Figure 2).
– Apply conditional quantiles of G_0 to the extremes of each subcube to obtain the partition.

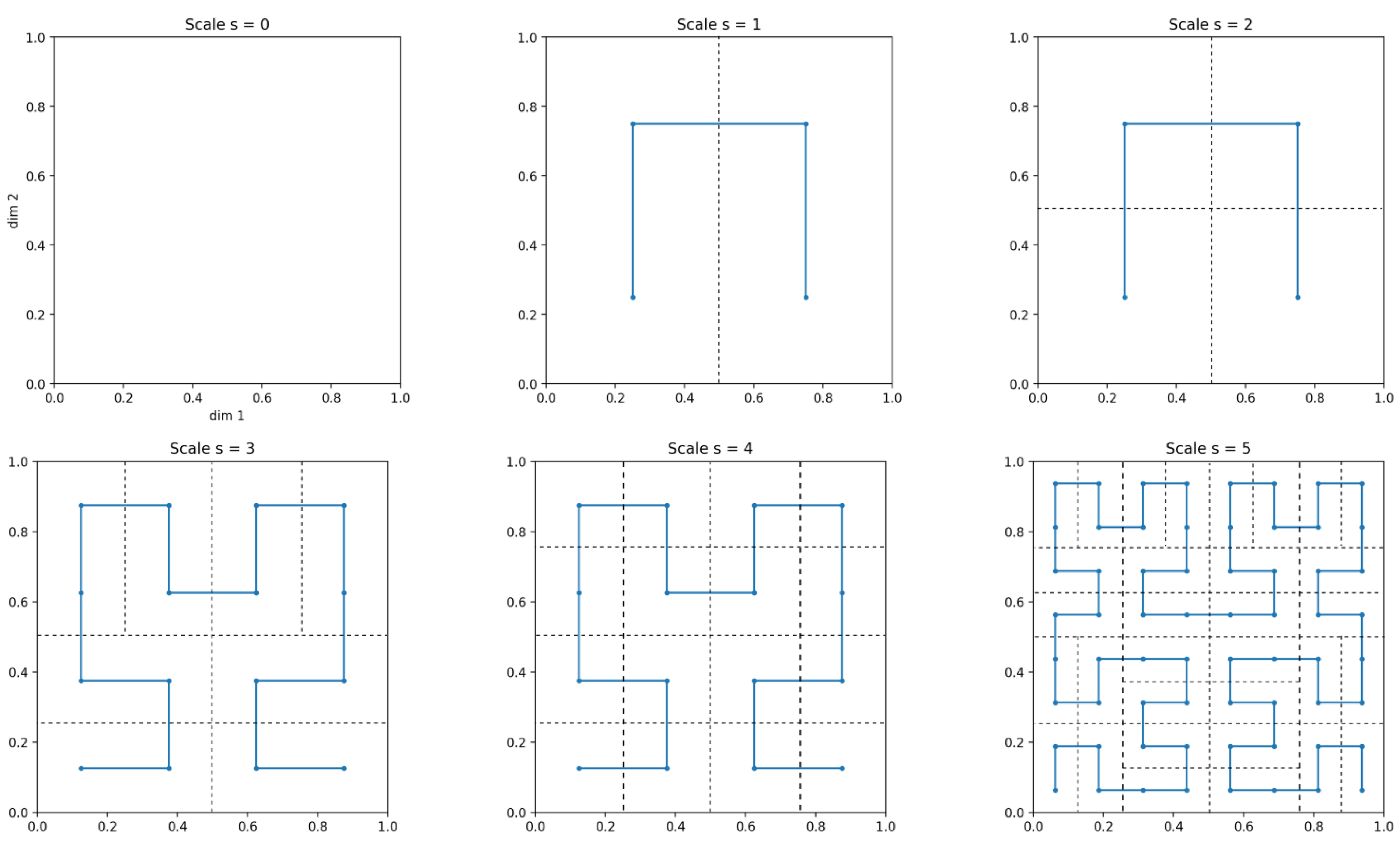

Figure 2: Dyadic partition of $[0, 1]^2$ obtained by the application of the Hilbert curve, for $s = 1, \dots, 5$.

Sampling at each node from G_0 truncated to $\Theta_{\boldsymbol{\mu};s,h}$ yields the **prior for the location parameter**,

$$\boldsymbol{\mu}_{s,h} \sim G_0 \cdot \mathbb{1}_{\Theta_{\boldsymbol{\mu};s,h}}.$$

The scale parameters $\Omega_{s,h}$ are sampled from a distribution H_0 **scaled by a deterministic monotone decreasing sequence** in s,

$$\Omega_{s,h} = \operatorname{diag}\left(c(s), \dots, c(s)\right) W_{s,h}, \quad W_{s,h} \overset{\text{iid}}{\sim} H_0.$$

Interpretation

- Nodes higher in the tree correspond to **coarser kernels** whereas deeper nodes correspond to **more localized kernels**. The posterior adapts the kernels to the smoothness of the data.
- We generalize a property of Stefanucci and Canale [2021] to show that the random location measure $G = \sum_{s=0}^{\infty}\sum_{h=1}^{2^s} \pi_{s,h}\delta_{\boldsymbol{\mu}_{s,h}}$ is **centered around** G_0 ***a priori***,

$$\mathbb{E}[G(A)] = G_0(A) \quad \text{for all } A \subseteq \Theta_{\boldsymbol{\mu}}.$$

- Also, we have control on the **prior average depth** of the binary tree by solving a

Performance in simulated datasets

We use the multiscale model with Gaussian prior on the location parameters and inverse Wishart prior on the scale parameter to have efficient Gibbs sampling updates [Kalli et al., 2011].

- Scenarios: 1) mixture of normals/skew normals, 2) multiscale mixture.
- Competitor: Pitman-Yor mixture model.
- Mean posterior predictive distribution for the multiscale model in Figure 3.

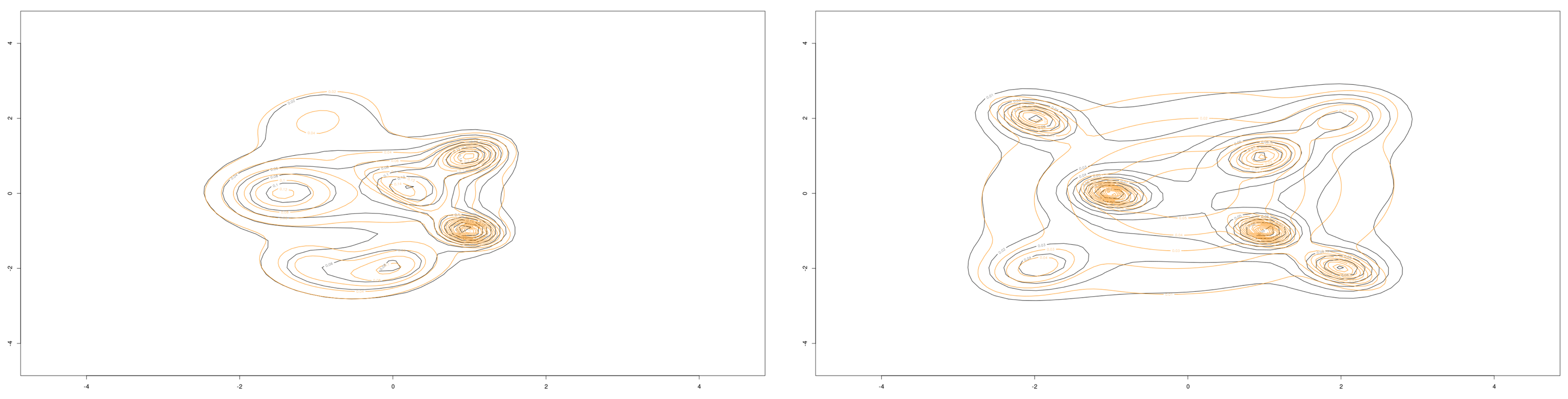

Figure 3: True density function (*black*) and mean posterior predictive distribution (*orange*) for the multiscale model.

Application to the Galaxy dataset

- The galaxy marginal distributions (Figure 4) are correctly estimated, except for the redshift which seems more problematic.
- The binary tree with nodes proportional to the estimated posterior median weight distribution (Figure 5) gives **information** on the **location of the clusters within the space**.

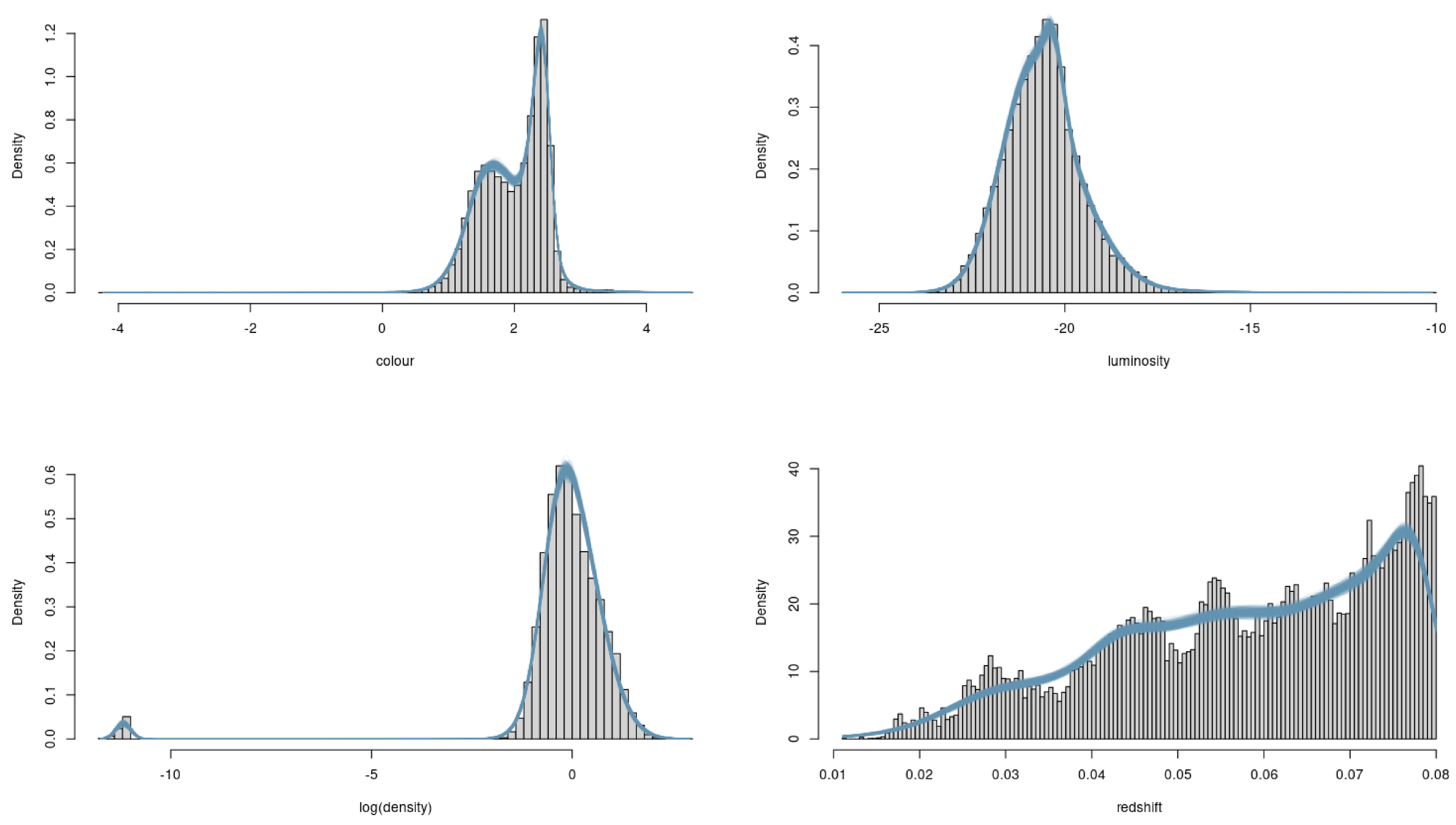

Figure 4: Posterior predictive distribution for the marginal distributions of the Galaxy dataset.

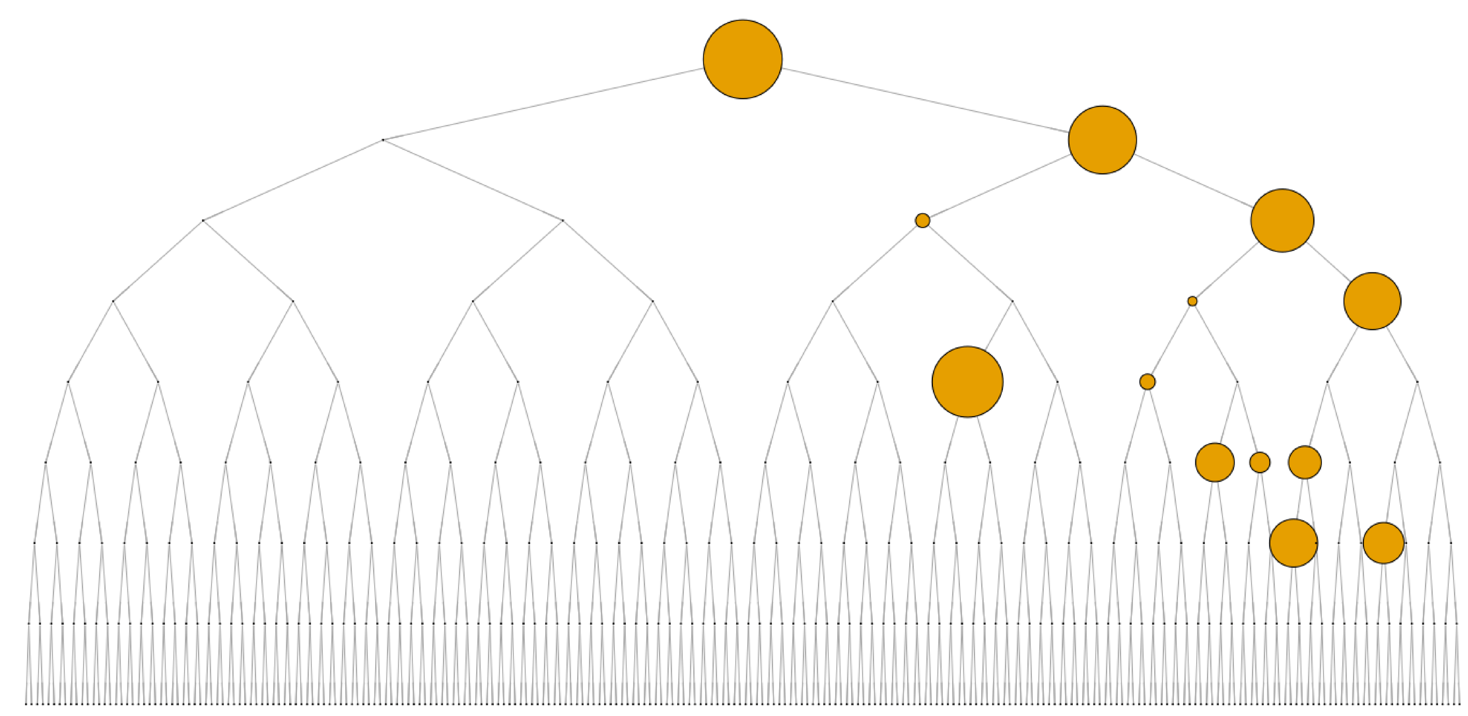

Figure 5: Multiscale binary tree with circles drawn proportionally to the median of the posterior distribution of the weights $\pi_{s,h}$.

References

Antonio Canale and David B. Dunson. Multiscale Bernstein polynomials for densities. *Statistica Sinica*, 26(3):1175–1195, 2016. ISSN 1017-0405.

David Hilbert. Über die stetige Abbildung einer Line auf ein Flächenstück. *Mathematische Annalen*, 38(3):459–460, September 1891. ISSN 1432-1807. doi: 10.1007/BF01199431.

Maria Kalli, Jim E. Griffin, and Stephen G. Walker. Slice sampling mixture models. *Statistics and Computing*, 21(1):93–105, January 2011. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-009-9150-y.

Marco Stefanucci and Antonio Canale. Multiscale stick-breaking mixture models. *Statistics and Computing*, 31(2), January 2021. ISSN 1573-1375. doi: 10.1007/s11222-020-09991-1.