MOX-Report No. 58/2020

Spatially dependent mixture models via the Logistic Multivariate CAR prior

Beraha, M.; Pegoraro, M.; Peli, R.; Guglielmi, A

Spatially dependent mixture models via the Logistic Multivariate CAR prior

Mario Beraha*†, Matteo Pegoraro‡, Riccardo Peli‡ and Alessandra Guglielmi*

29 July 2020

Abstract

We consider the problem of spatially dependent areal data, where for each area independent observations are available, and propose to model the density of each area through a finite mixture of Gaussian distributions. The spatial dependence is introduced via a novel joint distribution for a collection of vectors in the simplex, that we term logisticMCAR. We show that salient features of the logisticMCAR distribution can be described analytically, and that a suitable augmentation scheme based on the Pólya-Gamma identity allows to derive an efficient Markov Chain Monte Carlo algorithm. When compared to competitors, our model has proved to better estimate densities in different (disconnected) areal locations when they have different characteristics. We discuss an application on a real dataset of Airbnb listings in the city of Amsterdam, also showing how to easily incorporate for additional covariate information in the model.

1 Introduction

In spatial statistics, it is often assumed that data in neighboring locations are likely to behave more similarly than those that are far away. Thus, inference and prediction methods have been developed to take into account spatial dependence (Cressie, 1992). Spatial data are usually classified into two main categories: geostatistical data, for which an exact location is known for each observation, and areal (or lattice) data, when each observation is associated to a specific area or node in a lattice. Examples of the former are environmental applications (see Webster and Oliver, 2007, for many applications and datasets of soil properties) and geological reservoir characterization for oil and gas recovery (see Pyrcz and Deutsch, 2014, for examples). A recent review paper on statistical models for areal data is Banerjee (2016), which focuses on disease mapping and spatial survival analysis. See also the textbook by Banerjee et al. (2014) for data classification, applications and statistical models and techniques for spatially dependent data.

---

*Department of Mathematics, Politecnico di Milano
†Department of Computer Science, Università di Bologna
‡MOX - Department of Mathematics, Politecnico di Milano

In this work, we consider the problem of modeling data from I different groups, where each group corresponds to a specific areal location. In particular we assume that the spatial domain Ω is divided into I areas and, for each area, there is a vector of observations $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iN_i})$ from the same variable, each corresponding to a different subject in area i. The goal of this manuscript is the proposal of a statistical model, for data $\{\boldsymbol{y}_i, i = 1, \ldots, I\}$, accounting for dependence arising from spatial proximity while being flexible enough to model data that do not fit standard parametric distributions. We assume that data, within each areal unit i, are independent and identically distributed (i.i.d.) from an area-specific density f_i; relaxing the assumption to independent observations within each area is straightforward in the regression context, i.e. when covariates for each unit are available. In short, the problem we address can be described as the joint estimation of spatially dependent densities. We assume the Bayesian approach and we specify a prior for dependent densities $(f_1, \ldots, f_I)$ that encourages distributions associated to areas that are spatially close to be more similar than those associated to areas that are far away.

As motivating application, we consider publicly available data on Airbnb listings in the city of Amsterdam (NL). Airbnb is the largest vacation rental marketplace. In recent years it has been debated that Airbnb has deeply transformed the social structure of major touristic cities, as Amsterdam (Van Der Zee, 2016), Barcelona (Garcia-Ayllon, 2018) and several US cities (Wachsmuth and Weisler, 2018), driving up property prices and disrupting communities. The application dataset consists of more than $20,000$ listings spread over neighborhoods in Amsterdam. Our goal is to predict the nightly price of a new listing, with information given by covariates, taking into account the spatial dependence.

More broadly, spatial density estimation is a very current topic and we believe that the methodology proposed in this work can be applied to address other important questions, from climate change to epidemiology, from mobility in metropolitan areas to mobile networks usage, whenever the data are grouped into areal locations.

Mixtures are a popular class of models that have been successfully used in many applications, as they apply to different data structures and in different contexts. The full versatility of mixtures as a modelling tool has been particularly useful in density estimation and clustering problems. For a recent review of the topic, see Fruhwirth-Schnatter et al. (2019). In the Bayesian nonparametric setting, starting from MacEachern (2000) a great effort has been dedicated to modelling a set of related, though not identical, distributions. Dealing with spatial processes, Gelfand et al. (2005) and Duan et al. (2007) developed a spatial dependent Dirichlet process as random-effects distribution in the context of point-reference data. The stick-breaking representation of the Dirichlet process allows all the models built from it to be considered as infinite mixture models. Starting from the stick-breaking representation of the dependent Dirichlet process in the particular case of *single atoms* (atoms not indexed by covariates), Dunson and Park (2008) proposed the kernel stick-breaking process mixtures; spatial extensions of these type of mixtures have been developed to accommodate for general covariates and spatial locations for geostatistical data, such as, e.g., Rodriguez and Dunson (2011) and Ren et al. (2011).

Despite the theoretical properties of Bayesian nonparametric mixtures, computing the posterior inference in this setting may yield computational issues. In fact, typical MCMC algorithms here would need to marginalize out the infinite dimensional distribution from

the joint distribution of data and parameters, which might not be possible for models exhibiting a complicate dependence structure such as those mentioned above. As an alternative, finite-dimensional approximations of the infinite stick-breaking representation are typically used in the MCMC algorithms. However, as recently pointed out by Lijoi et al. (2020), the truncation procedure, for some models, might yield unwanted assumptions on the prior distribution of the number of clusters.

An alternative consists in considering a finite mixture model, where the number H of components is fixed, as we do here. Finite mixtures are particularly suited for the problem of modelling areal densities because (i) they adapt capturing the spatial dependence more than nonparametric mixtures, mainly because the weights of the finite mixtures are not forced to decrease exponentially fast to 0 as in many Bayesian nonparametric mixture models, and (ii) posterior inference under finite mixtures is extremely simple and admits efficient parallel code (unlike nonparametric models), thus helping our model scaling up as the size of the dataset increases. See Frühwirth-Schnatter (2006) and Celeux et al. (2019) for more insights on finite mixtures.

The first contribution of this work is the introduction of a joint distribution for a collection $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I$ of I vectors in the simplex S^H, reflecting the areal proximity structure in the distribution, through a logistic transformation of Gaussian multivariate conditional autoregressive (CAR) models. This distribution has been termed here the logistic MCAR distribution. Other authors have considered similar tricks, e.g. Jo et al. (2017), who build on the CAR model by Clayton and Kaldor (1987).

CAR models are a special case of Markov random fields and they were introduced in Besag (1974); they have been widely used for modelling data with spatial dependencies. Recent contribution to the Bayesian literature on CAR models for areal data includes Li et al. (2015), where the authors propose an area-dependent Dirichlet process that can also formally identify boundaries between areas. Also Zhou et al. (2015) use the trick of normalization of CAR distributions to time-varying weights in a rather complex application with focus on estimation of ambulance demand.

A second contribution of this work is the proposal of a finite Gaussian mixture model for each of the I area-related densities, keeping in mind the flexibility of the Gaussian mixtures to accurately approximate smooth densities. We let all the mixtures share the same set of atoms, while introducing similarity between the different mixtures through the novel conditional autoregressive model for vectors on the simplex, that we use as a prior for the weights of the mixtures. Through simulated data examples and the Airbnb application we show how specific features of the proposed model include (i) a sparse mixture specification as meant in Malsiner-Walli et al. (2016) and (ii) densities corresponding to areal units which belong to two different connected components in the proximity graph may behave differently. We discuss this last particular point in our data illustrations.

A third contribution of this paper is that we show how the full conditionals of the mixture weights can be sampled using a Gibbs sampler based on the Pólya-Gamma distribution, without resorting to Metropolis-Hastings steps, by exploiting a data augmentation scheme. As discussed in Polson et al. (2013), this update can lead to major improvements in the mixing of the chain. Our examples focus on continuous responses and the Gaussian kernel, though extensions to different kernels can be straightforwardly accommodated in

our framework.

The rest of this article is organized as follows. Section 2 gives background on finite mixture models and the geometry on the finite-dimensional simplex. Section 3 illustrates the definition and properties of the joint distribution of a collection of I vectors in the simplex, taking into account the underlying spatial proximity matrix. Our area-dependent mixture model is illustrated in Section 4.1, and the sparse mixture specification is detailed in Section 4.2; Section 4.3 discusses on the differences between our spatial prior and that in Jo et al. (2017). Section 5 sketches the Gibbs sampler to compute the posterior. Section 6 presents results from two simulation studies, and the application to Airbnb Amsterdam is discussed in Section 7. We conclude in Section 8 with final comments and discussion. The Appendix collects the proofs for the theoretical results, Monte Carlo simulations from the joint distribution of the I vectors in the simplex, full description of the Gibbs sampler, as well as additional plots and tables for the examples.

2 Preliminaries

2.1 Mixture Models

As mentioned in the Introduction, we consider the context where we have N_i data $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iN_i})$ for each areal unit $i = 1, \ldots, I$. We assume that the data, within each areal unit, are independent and identically distributed (i.i.d.) from an area-specific density f_i, and each y_{ij} has values in $\mathbb{Y} \subset \mathbb{R}^p$. In this paper, we assume $p = 1$, but multivariate responses can be straightforwardly accommodated in our context. A flexible model for the density in each area can be constructed by assuming a finite mixture, specifically

$$y_{ij} \mid \boldsymbol{w}_i, \boldsymbol{\tau}_i \overset{\text{iid}}{\sim} f_i(\cdot) = \sum_{h=1}^{H} w_{ih} k(\cdot \mid \tau_{ih}) \quad j = 1, \ldots, N_i \tag{1}$$

where $k(\cdot \mid \tau)$ is a density on $\mathbb{Y}$ for any $\tau \in \Theta$, and Θ is the parameter space. Each vector $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iH})^T$, the weights of the mixture (1), belongs to the $H - 1$ dimensional simplex S^H, where

$$S^H := \{(z_1, \ldots, z_H) \in \mathbb{R}^H : 0 \leq z_h \leq 1, h = 1, \ldots, H, \ \sum_{h=1}^{H} z_h = 1\} \tag{2}$$

and $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iH})^T$ are parameters in Θ^H. In this paper, we refer to $\boldsymbol{\tau}_i$ and $\boldsymbol{w}_i$ as the *atoms* and the *weights* of the mixture f_i.

Our goal is to introduce dependence between mixtures such that data in neighboring areas are more likely to be modeled with similar distributions than data in far areal units. A general mixture model like (1) would require to model jointly both the atoms and the weights of all the mixtures, in order to obtain a dependence structure suitable for spatial applications, which can be a challenging task in general, unless we consider a very specific application. In our approach instead, borrowing ideas from the single atom dependent Dirichlet processes, we constrain all the atoms across the different areas to be

4

equal, i.e. $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2, \ldots = \boldsymbol{\tau}_I = \boldsymbol{\tau}$, and focus only on the weights of the mixtures. In this way, a sufficient condition for two different mixtures to be similar is to have similar weights. In general, it is more difficult to define mixtures with area-dependent weights than generalizing to area-dependent weights and atoms, since simulation algorithms for models based on standard mixture models can usually be adapted with few modifications to dependent atoms.

When the goal of the inference is cluster estimation, the choice of H might become crucial. An alternative consists in assuming H random, including it in the state space of the MCMC algorithm; see, for instance, Nobile (1994). However, inference in this setting can be computationally intensive as it needs to rely either on specifically designed trans-dimensional MCMC moves (see Green, 1995; Richardson and Green, 1997), or to numerically evaluate infinite series, as in Miller and Harrison (2018) and in the marginal sampler in Argiento and De Iorio (2019).

On the other hand, sparse mixture models, as meant in Malsiner-Walli et al. (2016), assume a large value for H, larger than needed, and a prior assigning large mass to configurations where the weights of the superfluous components assume values close to zero. This implies that the prior number of *non-empty* components (i.e. components where at least one observation is allocated to) is significantly smaller than H.

In Section 3 we propose a prior distribution for $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I)$, in such a way that weights associated to close areas are more similar than weights associated to areas farther away, by constructing a Markov random field for random vectors with bounded sum. Moreover, by assuming a prior on the hyperparameters, we also show that this prior can induce sparsity in the mixture (see Section 4.2).

2.2   Geometry on the simplex S^H

The simplex $S^H \subset \mathbb{R}^H$ defined in (2) is not a vector subspace of $\mathbb{R}^H$. However, S^H is a vector space when equipped with the so-called Aitchison geometry, that defines the operation of *perturbation* (analogous of addition), *powering* (analogous of multiplication by scalar) and *inner product*. If $\boldsymbol{w}, \boldsymbol{w}_1, \boldsymbol{w}_2 \in S^H$, $\alpha \in \mathbb{R}$ we have

$$\boldsymbol{w}_1 \oplus \boldsymbol{w}_2 = \mathcal{C}(w_{11}w_{21}, \ldots, w_{1H}w_{2H}) := \left(\frac{w_{11}w_{21}}{\sum_{i=1}^{H} w_{1i}w_{2i}} \cdots \frac{w_{1H}w_{2H}}{\sum_{i=1}^{H} w_{1i}w_{2i}} \right)$$

$$\alpha \odot \boldsymbol{w} = \mathcal{C}(w_1^\alpha, \ldots, w_H^\alpha) \qquad \langle \boldsymbol{w}_1, \boldsymbol{w}_2 \rangle = \frac{1}{2H} \sum_{i,j=1}^{H} \log \frac{w_{1i}}{w_{1j}} \log \frac{w_{2i}}{w_{2j}}$$

where $\mathcal{C}$ denotes the *closure*, or normalization (i.e. dividing each element by the sum of all the elements) of a vector in $\mathbb{R}^H$. The symbols $\oplus$, $\odot$ and $\langle \cdot, \cdot \rangle$ denote perturbation, powering and inner product, respectively.

Many maps from S^H to $\mathbb{R}^{H-1}$ are available in the literature. For our purpose we focus

5

on the bijective additive log-ratio transformation (alr), defined by alr : $\boldsymbol{w} \mapsto \widetilde{\boldsymbol{w}}$:

$$\widetilde{w}_j = \log \frac{w_j}{w_H}, \quad j = 1, \dots, H-1$$

and its inverse, $\boldsymbol{w} = \text{alr}^{-1}(\widetilde{\boldsymbol{w}}) := \mathcal{C}(\text{e}^{\widetilde{w}_1}, \dots, \text{e}^{\widetilde{w}_{H-1}}, 1)$, that is

$$w_j = \frac{\text{e}^{\widetilde{w}_j}}{1 + \sum_{h=1}^{H-1} \text{e}^{\widetilde{w}_h}}, \ j = 1, \dots, H-1, \quad w_H = 1 - \sum_{h=1}^{H-1} w_h = \frac{1}{1 + \sum_{h=1}^{H-1} \text{e}^{\widetilde{w}_h}} \ . \quad (3)$$

Observe that both maps are linear, i.e., for any $\boldsymbol{w}_1, \boldsymbol{w}_2 \in S^H$, $\tilde{\boldsymbol{w}}_1, \tilde{\boldsymbol{w}}_2 \in \mathbb{R}^{H-1}$, $\alpha \in \mathbb{R}$,

$$\text{alr}(\boldsymbol{w}_1 \oplus \boldsymbol{w}_2) = \text{alr}(\boldsymbol{w}_1) + \text{alr}(\boldsymbol{w}_2), \quad \text{alr}(\alpha \odot \boldsymbol{w}_1) = \alpha \, \text{alr}(\boldsymbol{w}_1)$$
$$\text{alr}^{-1}(\tilde{\boldsymbol{w}}_1 + \tilde{\boldsymbol{w}}_2) = \text{alr}^{-1}(\tilde{\boldsymbol{w}}_1) + \text{alr}^{-1}(\tilde{\boldsymbol{w}}_2), \quad \text{alr}^{-1}(\alpha \tilde{\boldsymbol{w}}_1) = \alpha \odot \text{alr}^{-1}(\tilde{\boldsymbol{w}}_1).$$

This map was used in Atchison and Shen (1980) to define a new distribution on the simplex, the logistic-normal distribution. Formally, we say that $\boldsymbol{w} = (w_1, \dots, w_{H-1}, w_H := 1 - \sum_{h=1}^{H-1} w_h)^T \in S^H$ follows the logistic-normal distribution of parameters $\boldsymbol{\mu}, \Sigma$ for $\boldsymbol{\mu} \in \mathbb{R}^{H-1}$, and Σ a positive definite $(H-1) \times (H-1)$ matrix if

$$\widetilde{\boldsymbol{w}} = \text{alr}(\boldsymbol{w}) = \left(\log \frac{w_1}{w_H}, \dots, \log \frac{w_{H-1}}{w_H} \right)^T \sim \mathcal{N}_{H-1}(\boldsymbol{\mu}, \Sigma)$$

where $\mathcal{N}_{H-1}(\boldsymbol{\mu}, \Sigma)$ denotes the $(H-1)$-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. The logistic-normal distribution offers a rich way to model data embedded on the simplex and is particularly suited for our application. Although moments of this distribution exist, their expression is not available analytically. However, when modeling data in the simplex, one is usually more interested in the pairwise ratios of the components than on the values of the components themselves. In turn, these expected values and covariances are available analytically and given by

$$\mathbb{E}\left[\log \frac{w_i}{w_j}\right] = \mu_i - \mu_j, \quad \text{Cov}\left(\log \frac{w_i}{w_j}, \log \frac{w_l}{w_k}\right) = \Sigma_{il} + \Sigma_{jk} - \Sigma_{ik} - \Sigma_{jl}$$

where Σ_{il} denotes the (i, l)-element of the matrix Σ.

3   The logistic MCAR distribution

In this section, we introduce and describe a joint distribution for a collection of vectors in the simplex $\boldsymbol{w}_1, \dots, \boldsymbol{w}_I \in S^H$, reflecting the areal proximity structure in the distribution. For each pair of areas i and j, $g_{ij} \in [0, 1]$ indicates the amount of spatial proximity between them. In the rest of the paper we assume $g_{ij} = 1$ if i and j are neighbours, i.e. the areas share at least a border, and $g_{ij} = 0$ otherwise, but we could consider more general settings. By definition, $g_{ii} = 0$ for all i. The matrix $G = [g_{ij}]_{i,j=1}^I$ is called the proximity matrix and we assume it known. It will be useful, for our analyses, to identify the matrix G with a graph, whose nodes are denoted by indexes $1, \dots, I$ and the links are given by the g_{ij}'s, i.e. there is a link between nodes i and j if, and only if, $g_{ij} = 1$. We define the joint

distribution of $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I$ introducing the transformed vectors $\widetilde{\boldsymbol{w}}_i := \mathrm{alr}(\boldsymbol{w}_i)$, $i = 1, \ldots, I$ and assuming a joint Gaussian conditional autoregressive distribution for $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$.

Conditionally autoregressive (CAR) models are a special case of Markov random fields. In general, if $\{X_1, \ldots, X_n\}$, with $X_i \in \mathbb{R}$, is a set of random variables, to define a CAR model over $X_1, \ldots, X_n$, one usually starts by assigning the conditional distribution of each X_i given all the others $X_{-i} := \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_n\}$. The set of conditional distributions, under assumptions, identifies the unique joint distribution of $(X_1, \ldots, X_n)$. The class of CAR models is large; see further detail in Besag (1974), Cressie (1992), Cressie (1993), Kaiser and Cressie (2000), Cressie and Wikle (2015) and references therein, just to include a few papers.

We refer to Gelfand and Vounatsou (2003) for the definition of a CAR model for vector-valued random variables. We assume the following multivariate conditionally autoregressive (MCAR) model, originally proposed in Leroux et al. (2000) in the case of univariate responses:

$$\widetilde{\boldsymbol{w}}_i \mid \widetilde{\boldsymbol{w}}_{-i}, \Sigma, \rho \sim \mathcal{N}_{H-1} \left(\frac{\rho \sum_{j=1}^I g_{ij} \widetilde{\boldsymbol{w}}_j + (1-\rho)\widetilde{\boldsymbol{m}}_i}{\rho \sum_{j=1}^I g_{ij} + 1 - \rho}, \frac{\Sigma}{\rho \sum_{j=1}^I g_{ij} + 1 - \rho} \right), \ i = 1, \ldots, I, \tag{4}$$

where Σ is a definite positive $(H-1) \times (H-1)$ matrix and $\widetilde{\boldsymbol{m}}_i \in \mathbb{R}^{H-1}$ for all i. If $\rho \in (-1, 1)$, the joint distribution is well defined and unique (Gelfand and Vounatsou, 2003). From (4) we have that $\widetilde{\boldsymbol{w}} = vec(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$, the vectorization of the weights, is such that

$$\widetilde{\boldsymbol{w}} \sim \mathcal{N}_{I(H-1)} \left(\widetilde{\boldsymbol{m}}, \left((F - \rho G) \otimes \Sigma^{-1} \right)^{-1} \right) \tag{5}$$

where $\otimes$ denotes the Kronecker product, $\widetilde{\boldsymbol{m}} = vec(\widetilde{\boldsymbol{m}}_1, \ldots, \widetilde{\boldsymbol{m}}_I)$ and $F = \mathrm{diag}(\rho \sum_j g_{1j} + 1 - \rho, \ldots, \rho \sum_j g_{Ij} + 1 - \rho)$. The matrix $A^{-1}(G, \rho) := (F - \rho G) = \rho \left(diag(G\mathbf{1}_I) - G \right) + (1-\rho)\mathbb{I}_I$ in (5), where $\mathbf{1}_I \in \mathbb{R}^I$ denotes the vector of ones and $\mathbb{I}_I$ denotes the $I \times I$ identity matrix, have a key role here. When $\rho = 1$, (4) reduces to the intrinsic CAR model, and the joint density of $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$ is improper. If $\rho = 0$, the $\widetilde{\boldsymbol{w}}_i$'s are independent. See below for further properties of $A(G, \rho)$.

We say that the sequence of vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I$ follows a logistic multivariate CAR distribution of parameters ρ and Σ on a graph G if the transformed variables $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$, $\widetilde{\boldsymbol{w}}_i = \mathrm{alr}(\boldsymbol{w}_i)$, follow the MCAR model in (4) (or (5)). We write $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim \mathrm{logisticMCAR}(\widetilde{\boldsymbol{m}}, \rho, \Sigma; G)$.

One key aspect is the relation that (4) induces over the vectors on the simplex rather than on their transformation. This is made clear by the following proposition.

Proposition 1 *If $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim logisticMCAR(\widetilde{\boldsymbol{m}}, \rho, \Sigma; G)$, then, for any $i = 1, \ldots, I$,*

$$\mathbb{E}\left[\log \frac{w_{il}}{w_{ik}} \mid \boldsymbol{w}_{-i} \right] = \log \left(\left(\frac{m_{il}}{m_{ik}} \right)^{1-\rho} \prod_{j \in U_i} \left(\frac{w_{jl}}{w_{jk}} \right)^{\rho} \right)^{\frac{1}{\rho|U_i|+1-\rho}} \quad l, k = 1, \ldots, H \tag{6}$$

where $|U_i| = \sum_j g_{ij}$ and $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{iH})$, with $\boldsymbol{m}_i = alr^{-1}(\widetilde{\boldsymbol{m}}_i)$.

7

Proof: see Appendix A.

There are several immediate but interesting properties of (6). First of all, if $\rho = 1$, (6) means that the expected value of (the logarithm of) the ratios between the components of $\boldsymbol{w}_i$ is equal to (the logarithm of) the geometric mean of the corresponding ratios of the components of the vectors $\boldsymbol{w}_j$ nearby. If $\rho = 0$, the right handside of (6) reduces to $\log(m_{il}/m_{ik})$, which is to be expected since, in this case, the $\boldsymbol{w}_i$'s would not be spatially correlated. Instead, in case $0 < \rho < 1$, which we assume throughout the paper (see Section 4.1), we can interpret the right handside of (6) as a weighted mean on the simplex, according to Aitchison geometry, of two components: the first component m_{il}/m_{ik} corresponding to the mean $\boldsymbol{m}$ and the second $\prod_{j \in U_i} (w_{jl}/w_{jk})$ taking into account the spatial dependence. In other words, Proposition 1 provides the same interpretation of (4) but for ratios between components of the anti-transformed vectors in the simplex, if we look at them through the Aitchison geometry.

Starting from the joint distribution in (5), we can also study the marginal covariance of $\boldsymbol{w}_i, \boldsymbol{w}_j$ in $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim \text{logisticMCAR}(\widetilde{\boldsymbol{m}}, \rho, \Sigma; G)$ for $i \neq j$. We point out that the matrix $A(G, \rho)^{-1}$, introduced above, is a strictly diagonal dominant matrix (i.e. for each row, the absolute value of the diagonal entry is larger than or equal to the sum of the absolute values of the off-diagonal entries in that row) with negative off-diagonal entries, and, hence, its inverse $A(G, \rho)$ has elements which are all positive.

Proposition 2 *If* $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim \text{logisticMCAR}(\widetilde{\boldsymbol{m}}, \rho, \Sigma; G)$, *then*

$$\text{Cov}\left(\log \frac{w_{il}}{w_{im}}, \log \frac{w_{jl}}{w_{jm}}\right) = A_{ij}\left(\Sigma_{ll} - 2\Sigma_{lm} + \Sigma_{mm}\right) \quad i, j = 1, \ldots, I \quad l, m = 1, \ldots, H - 1$$

$$\text{Cov}\left(\log \frac{w_{il}}{w_{iH}}, \log \frac{w_{jl}}{w_{jH}}\right) = A_{ij}\Sigma_{ll} \quad i, j = 1, \ldots, I \quad l = 1, \ldots, H - 1$$

In particular, $\text{Cov}\left(\log \frac{w_{il}}{w_{im}}, \log \frac{w_{jl}}{w_{im}}\right) = 0$ *if areas i and j belong to different connected graph components of the graph G.*

Proof: see Appendix A.

Observe that the logisticMCAR distribution of $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I)$ is not exchangeable, i.e. it is not true that $\mathcal{L}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I)$ and $\mathcal{L}(\boldsymbol{w}_{\pi(\{1\})}, \ldots, \boldsymbol{w}_{\pi(\{I\})})$ are equal for any $(\pi(\{1\}), \ldots, \pi(\{I\}))$ permutation of $(1, \ldots, I)$. Here, as in the rest of the paper, the distribution of a random element y is denoted by $\mathcal{L}(y)$. Nonetheless, the logisticMCAR distribution induces exchangeable priors on all the fully connected components of the graph G.

The logistiMCAR distribution shares the same analytical limitations of the logistic-normal one, that make theoretical analysis arduous. In Appendix B we report an extensive Monte Carlo (MC) simulation where we compute the covariance between different components of the vectors of weights and we draw a comparison between the logisticMCAR and the Dirichlet distributions.

4 Spatially dependent mixture models

We return to the problem of formalizing a Bayesian model for I groups of data $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I)$, $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iN_i})$, $i = 1, \ldots, I$. As mentioned at the beginning of Section 3, we assume that each vector $\boldsymbol{y}_i$ is associated to an area i and that for each pair of areas i and j, $g_{ij} = 1$ if i and j are neighbours and $g_{ij} = 0$ otherwise.

4.1 The finite mixture model with spatially dependent weights

Let the proximity matrix $G = [g_{ij}]_{i,j=1}^I$ be fixed. We assume that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_I$, conditioning to $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I$ and $\boldsymbol{\tau}$, are independent and that, for each $i = 1, \ldots, I$,

$$y_{ij} \mid \boldsymbol{w}_i, \boldsymbol{\tau} \overset{\text{iid}}{\sim} \sum_{h=1}^{H} w_{ih} \mathcal{N}(\cdot \mid \tau_h) \quad j = 1, \ldots, N_i, \tag{7}$$

$$\tau_h \overset{\text{iid}}{\sim} P_0 \quad h = 1, \ldots, H \tag{8}$$

$$(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \mid \rho, \Sigma \sim \text{logisticMCAR}(\widetilde{\boldsymbol{m}}, \rho, \Sigma; G) \tag{9}$$

$$\Sigma \sim \text{Inv-Wishart}(\nu, V) \tag{10}$$

$$\rho \sim \pi(\rho) \tag{11}$$

where $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iH})^T \in S^{H-1}$ (see (2)) and $\widetilde{\boldsymbol{m}} = vec(\widetilde{\boldsymbol{m}}_1, \ldots, \widetilde{\boldsymbol{m}}_I) \in \mathbb{R}^{I(H-1)}$. As often considered, we study the case where the kernel in the mixture (7) is the Gaussian density with mean μ and variance σ^2, so that $\tau = (\mu, \sigma^2)$ and P_0 is a probability distribution over $\Theta = \mathbb{R} \times \mathbb{R}^+$. We consider independent marginal priors for ρ and Σ. Moreover, the support of the prior of ρ is typically assumed to be $(0, 1)$ to induce the similarity of spatial neighbours (see, for instance, Gelfand and Vounatsou, 2003, Section 4).

Model (7) - (11) assumes that each group of data $\boldsymbol{y}_i$ is modelled as a (finite) mixture of Gaussian kernels. Specifically, observations within each group are i.i.d given the weights and the atoms of the mixtures, while conditionally to all the mixture weights $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I$, observations in different groups are independent. All the I mixtures share the same set of atoms $\tau_1, \ldots, \tau_H$, which are assumed i.i.d from the base measure P_0, a continuous distribution on $\Theta = \mathbb{R} \times \mathbb{R}^+$. The dependence between mixtures in different areal units is induced only by the prior on the mixture weights.

As mentioned in the Introduction, we define a prior for $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I)$, allowing weights associated to close areas to be more similar than weights associated to areas farther away, through the logistic transformation of a Gaussian CAR model. The idea is not new in the literature, and the prior for the mixture weights of area-dependent densities in Jo et al. (2017) is closely related to our prior. We discuss the differences between the two priors in Section 4.3 and compare their features by fitting simulated data to the two models in Section 6.1.

In order to derive a Gibbs sampler for our model, we introduce the latent variables s_{ij}, one for each observation, indicating to which component of the mixture observations

are allocated to, and rewrite (7) as

$$y_{ij} \mid s_{ij} = h, \tau_h \overset{\text{ind}}{\sim} \mathcal{N}(\cdot \mid \tau_h) \quad j = 1, \dots, N_i, \ i = 1, \dots, I \tag{12}$$

$$p(s_{ij} = h \mid \boldsymbol{w}_i) = w_{ih} \quad h = 1, \dots, H \tag{13}$$

A component in the mixture is said *empty* if it has not been allocated to any observation. Here, and in the whole paper, *cluster* denotes any allocated component and the number of clusters is the number of allocated components. It is clear from (12)-(13) that the allocated and empty components, as well as the number of clusters, are random variables, with marginal prior distributions induced by our model.

We complete the specification of our model by adopting a marginal prior on $\widetilde{\boldsymbol{m}}$ that encourages sparsity in the mixtures. We discuss this choice in detail in the next Section 4.2.

4.2 Sparse mixtures via a prior on $\widetilde{\boldsymbol{m}}$

Generally, a sparse mixture is obtained when the number of clusters is smaller than the total number of components H. There are two well-known strategies to obtain sparse mixtures in the Bayesian context. The first one assigns a prior on the weights that forces them to be stochastically decreasing, so that the "last" weights are extremely small and the corresponding mixture components are seldom allocated. The alternative strategy consists in assigning a prior for the weights that concentrates its mass around the edges of the simplex in a symmetric way, as it is the case of the sparse Dirichlet distribution, i.e. a Dirichlet distribution with all the parameters equal to α, with $0 < \alpha < 1$. In the latter case, there is no preferential ordering of the weights and any mixture component could, in principle, be allocated. However, the first approach might not fit spatial applications, in particular when the proximity graph G has disconnected components, since assuming decreasing weights for all the mixtures would force data from two disconnected components to be always sampled from the few components with larger weights, and hence to behave always similarly.

Here, we show how we can mimic the sparse Dirichlet distribution for the weights, by assuming a suitable prior on parameters $\widetilde{\boldsymbol{m}}_i$'s. We start by observing that in the mixture model (7) for the i-th area, values in the vector $\widetilde{\boldsymbol{m}}_i$ in (9) that are very different would force some components h in (7) to be more often allocated than others, being their weights larger than the others (in mean). Hence, we induce "symmetric" sparsity in our marginal prior for the weights by assuming $\widetilde{\boldsymbol{m}}_i \sim \mathcal{N}_{H-1}(\boldsymbol{0}, \eta^2 \mathbb{I})$. Observe that, since the distribution of $\widetilde{\boldsymbol{m}}_i$ is centered in $\boldsymbol{0}$ and isotropic, we are not forcing, marginally, any specific ordering on the weights. Moreover large values of the η^2 above imply that $\widetilde{\boldsymbol{m}}_i$ has components with very large values, some positive and some negative and this pattern of values will likely be preserved in $\widetilde{\boldsymbol{w}}_i$. Thus, by the inverse alr transformation, we would obtain that components of the vector $\boldsymbol{w}_i$ associated to negative values in $\widetilde{\boldsymbol{w}}_i$ will be close to zero and those associated to large positive values would be far from zero. Hence η^2 can be interpreted as a sparsity-tuning parameter. In addition, since for large values of η^2 a small number of components in $\widetilde{\boldsymbol{m}}_i$ will be significantly larger than all the others, under the marginal logisticMCAR prior for the weights $\boldsymbol{w}_i$ in a single area i, the number of clusters will have a prior distribution that is highly peaked on values near 1; see also

Figure 2(b). On the other hand, when η^2 is very small, all the components in $\widetilde{\boldsymbol{m}}_i$ are close to 0, implying that all the weights in (7) are close to $1/H$ (see (3)) and consequently a large prior number of clusters in each finite mixture.

Assuming each $\widetilde{\boldsymbol{m}}_i$ in the prior specification (4) to be random and i.i.d would imply that, a priori, the mixture weights could be extremely different in neighboring areas. However, this is not what we aim at modeling in applications, where we typically assume that data in close areal units follow similar distributions. On the other hand, assuming $\widetilde{\boldsymbol{m}}_1 = \ldots = \widetilde{\boldsymbol{m}}_I$ seems overly restrictive, since we would loose the property that two connected graph components in $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$ are independent under CAR distributions when marginalizing out the only shared parameter $\widetilde{\boldsymbol{m}}_1$. Hence we propose to extend the logisticMCAR$(\widetilde{\boldsymbol{m}}, \rho, \Sigma; G)$ in (9) assuming

$$\{\widetilde{\boldsymbol{m}}_i = \widetilde{\boldsymbol{m}}_j = \widetilde{\boldsymbol{m}}_{C_m} \text{ iff } i, j \in C_m \text{ for some } m\} \tag{14}$$

where $C_1, \ldots, C_k$ denote the connected components of graph G, i.e. all the parameters $\widetilde{\boldsymbol{m}}_i$s are assumed common within each connected component. For such parameters, we assume

$$\widetilde{\boldsymbol{m}}_{C_1}, \ldots \widetilde{\boldsymbol{m}}_{C_k} \overset{\text{iid}}{\sim} \mathcal{N}_{H-1}(\boldsymbol{0}, \eta^2 \mathbb{I}). \tag{15}$$

4.3   Comparison with competitor models

We briefly introduce the class of spatially dependent species sampling mixtures in Jo et al. (2017), to be used as a competitor model for our applications. Jo et al. (2017) define the weights in the mixtures to be spatially dependent, modeling them from a Gaussian CAR distribution, as we do. For the case of areal data, they resort to the CAR specification in Clayton and Kaldor (1987). Here we compare the two marginal priors for the weights in each area, whereas in Section 6 we compare the mixture models on simulated datasets. We have shown in Section 4.2 that our marginal prior can mimic the sparse Dirichlet distribution by assuming $\widetilde{\boldsymbol{m}}$ in the logisticMCAR$(\widetilde{\boldsymbol{m}}, \rho, \Sigma; G)$ to be random, unlike Jo et al. (2017). Moreover, we also show that this sparsity characteristic is obtained by our model without forcing the different connected graph components in G to behave similarly as under the model in Jo et al. (2017).

In the species sampling mixture in Jo et al. (2017), in the case of $I = 1$ area, after truncation of the infinite-dimensional species sampling model, the marginal prior on the weights $\boldsymbol{w}$ (dropping out the index i) is defined as

$$\widetilde{\boldsymbol{w}} = (\widetilde{w}_1, \ldots, \widetilde{w}_H)^T \sim \mathcal{N}_H(\widetilde{\boldsymbol{\theta}}, \tau^2 \mathbb{I}), \quad \boldsymbol{w} = \left(\frac{\mathrm{e}^{\widetilde{w}_1}}{\sum_h \mathrm{e}^{\widetilde{w}_h}}, \ldots, \frac{\mathrm{e}^{\widetilde{w}_H}}{\sum_h \mathrm{e}^{\widetilde{w}_H}} \right). \tag{16}$$

Moreover, they assume $\widetilde{\theta}_h$, the h-component of $\widetilde{\boldsymbol{\theta}}$, to be equal to $\log\{1 - (1 + \mathrm{e}^{b-ah})^{-1}\}$, a and b being positive. Note also that (16) makes $\boldsymbol{w}$ non-identifiable.

Figure 1 shows a comparison of the marginal priors of the weights in the mixture (7) with $H = 3$, under our logisticMCAR prior and (16) introduced in Jo et al. (2017). In particular, for our logisticMCAR$(\widetilde{\boldsymbol{m}}, \rho, \Sigma; G)$ prior we have assumed $\widetilde{\boldsymbol{m}} \sim \mathcal{N}_2(\boldsymbol{0}, 9\mathbb{I})$, $\widetilde{\boldsymbol{w}} \sim \mathcal{N}_2(\widetilde{\boldsymbol{m}}, \mathbb{I})$, that is (5) with $\Sigma = \mathbb{I}$, while we fix $\widetilde{\theta}_h = \log\{1-(1+\mathrm{e}^{1-h})^{-1}\}$, $h = 1, \ldots, H$,
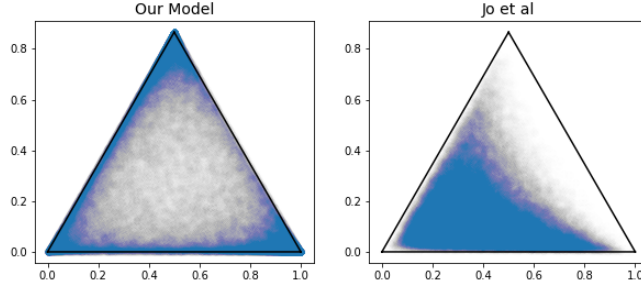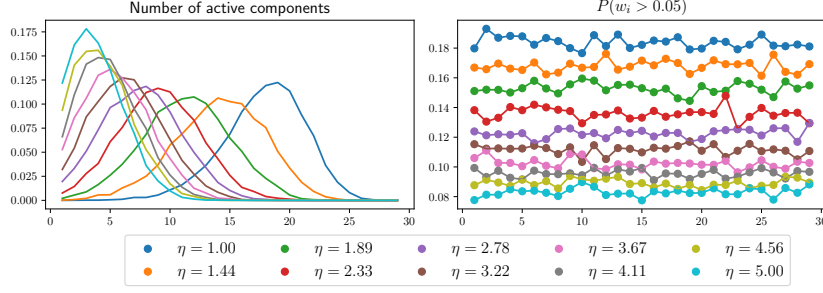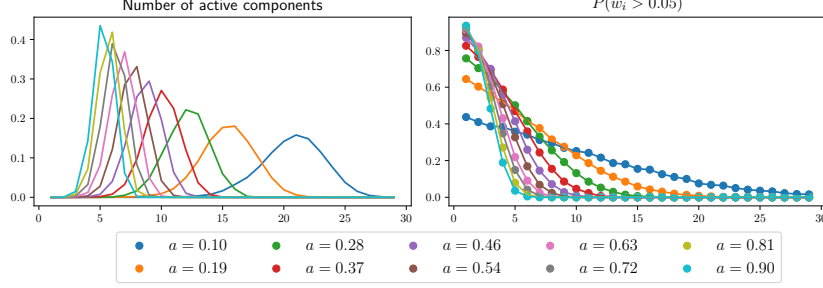
Figure 1: Scatterplots of $N = 100,000$ MC draws (in blue) from the marginal priors of the weights in one single area with $H = 3$, under our logisticMCAR prior (left) and (16) in Jo et al. (2017) (right). White/gray areas represents low-density zones and dark blue zones high-density ones.

in (16) as in Jo et al. (2017) ($a = b = 1$) and $\tau^2 = 1$.

We compare the priors via $N = 100,000$ MC draws. Figure 1 shows the scatterplots of the draws of the two marginal priors. In particular, the draws from our prior (left panel) recover the "sparse" symmetric Dirichlet prior with all parameters equal to $\alpha < 1$; the draws are symmetrically concentrated around the edges of the simplex, and give significant mass to locations near the vertexes. On the other hand, the draws from the marginal prior in Jo et al. (2017) clearly show asymmetry in favor of the first component, also giving negligible mass to neighborhoods of the vertices. When the number H of components in the mixture (7) is larger, we can compare the priors via two functionals by computing (i) the number of active components ($H^{(a)}$), that we define as the components associated to weights greater than 0.01, i.e. the cardinality of the set $\{h : w_h > 0.01\}$, and (ii) the probability for each component of the vector $\boldsymbol{w} \in S^H$ to be grater than the threshold 0.05. We fix $H = 30$ and, simulating $N = 10,000$ MC draws as before, we plot the marginal prior distributions of these functionals under our logisticMCAR prior (Figure 2(a)) and (16) in Jo et al. (2017) (Figure 2(b)), for different values of the hyperparameters in the priors. From both left panels, displaying the marginal priors of $H^{(a)}$ (as continuous lines to help seeing the differences), it is clear that both models may induce different types of prior behaviors. However, when considering the right panels, displaying, for each index $h = 1, \ldots, H$, the probability that $w_h > 0.05$, it is clear that, while under our prior, for each degree of sparsity η^2, the probability of inclusion of a single component h is rather uniform on $\{1, \ldots, H\}$, this probability decreases with h under the marginal prior by Jo et al. (2017). Going back to the prior in (15), observe how this model specification gives a major difference with the mixture model in Jo et al. (2017). Indeed, if we aim at considering the context where areal units are connected through the graph G, but there are at least two different connected components, as we will have in the application in Section 7, the prior in Jo et al. (2017) would still force the different connected components in the graph to behave similarly, because of the parameter $\widetilde{\boldsymbol{\theta}}$ shared by all the mixtures; see (16). This effect becomes increasingly more evident as the sparsity in each mixture is increased, as shown in Figure 2b. On the other hand, our model allows for the required level of sparsity in each mixture without forcing the different connected components in

12

(a) Our prior for different values of η^2

(b) The prior in Jo et al. (2017) for different values of a and $b = 0.5$

Figure 2: Prior distribution of the number of active components (left) and the probability for w_h to be greater than 0.05 (right) under our prior (top row) and prior (16) in Jo et al. (2017) (bottom row). Here $H = 30$.

the graph to behave similarly. For this reason, we believe we have introduced a more flexible model for jointly estimate spatially dependent densities than Jo et al. (2017), at least for applications where different connected components in the graph should exhibit different behaviors.

We will provide comparison also with the Hierarchical Dirichlet Process (HDP) mixture model in Teh et al. (2006) in Section 6. To keep the paper self-contained as much as possible, we report the HDP mixture model as follows

$$y_{ij} \mid F_i \overset{\text{iid}}{\sim} \int_{\Theta} k(y_{ij} \mid \tau) F_i(d\tau), \quad \{F_i\}_{i=1}^I \mid G \overset{\text{iid}}{\sim} \mathcal{D}_{\alpha G} \quad G \sim \mathcal{D}_{\beta P_0} \tag{17}$$

where $\mathcal{D}_{\beta P_0}$ denotes the Dirichlet measure, i.e. the distribution of a random probability measure that is the Dirichlet process with measure parameter βP_0. We assume the kernel $k(\cdot \mid \tau)$ as the Gaussian density on $\mathbb{Y}$ for $\tau = (\mu, \sigma^2)$ as in (7). Thanks to the stick-breaking representation of the Dirichlet process, it is possible to rewrite the likelihood in (17) as

$$y_{ij} \mid \{w_{ih}\}_{h=1}^\infty, \{\phi_{ih}^*\}_{h=1}^\infty \overset{\text{iid}}{\sim} \sum_{h=1}^\infty w_{ih} k(y_{ij} \mid \phi_{ih}^*)$$

where $\phi_{ih}^* \mid G \overset{\text{iid}}{\sim} G$ in (17) and $\{w_{ih}\}$ for each i are a sequence of non-negative weights summing to 1. Moreover, since each F_i, conditionally to G, is an independent draw from the Dirichlet process prior with discrete base measure G, this yields that all the atoms are

shared across all populations. This means that the set of the unique values in $\{\phi^*_{ih}\}_{h=1}^\infty$ is equal to the set of unique values in $\{\phi^*_{jh}\}_{h=1}^\infty$ for $j \neq i$ and coincides with the set of atoms in G. Hence, denoting by $\{\tau_h\}_{h=1}^\infty$ the atoms in G, the HDP mixture model defines a joint probability distribution for random probability measures with the same support points, as in our model. This is the motivation to consider the HDP mixture model as the "natural competitor" of ours.

5   The Gibbs Sampler

We illustrate a MCMC algorithm to sample from the posterior distribution of our model (7)-(11) and (14) - (15). The state is described by parameters $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_H)$, $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$, where $\widetilde{\boldsymbol{w}}_i = \mathrm{alr}(\boldsymbol{w}_i)$, $i = 1, \ldots, I$, $\{s_{ij}\}_{ij}$ $(j = 1, \ldots, N_i)$ in (12)-(13) and $\widetilde{\boldsymbol{m}}_{C_1}, \ldots \widetilde{\boldsymbol{m}}_{C_k}$ in (15).

We use the following notation: given the sequence of vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I$, we denote by $\boldsymbol{w}_{-i}$ the same sequence where the i–th vector has been removed. Given a single vector $\boldsymbol{w}_i$, we denote by $\boldsymbol{w}_{i,-h}$ the same vector where the h-component as been removed. Finally, for the matrix Σ, let Σ_{ij} denote the (i, j)-element; moreover, Σ_i denotes its i–th row (as a vector) so that $\Sigma_{i,-j}$ denotes the i–th row where the j–th element has been removed and $\Sigma_{-h,-k}$ denotes the $(H - 2) \times (H - 2)$ matrix there the h–th row and k–th column have been removed.

There are two "non-standard" steps in the Gibbs sampler: the update of the transformed weights $\widetilde{\boldsymbol{w}}_i$ and the update of their means $\widetilde{\boldsymbol{m}}_{C_1}, \ldots \widetilde{\boldsymbol{m}}_{C_k}$. Here, we only describe the full conditionals of each $\widetilde{\boldsymbol{w}}_i$, as the update of $\widetilde{\boldsymbol{m}}_{C_i}$ is rather technical and is deferred to Appendix C together with the other standard full conditionals.

We begin by writing the full conditional for $\widetilde{w}_{ih}$, for each i and h, as

$$\mathcal{L}(\widetilde{w}_{ih} \mid \widetilde{\boldsymbol{w}}_{-i}, \widetilde{\boldsymbol{w}}_{i,-h}, rest) \propto \pi(\widetilde{w}_{ih} \mid \widetilde{\boldsymbol{w}}_{-i}, \widetilde{\boldsymbol{w}}_{i,-h}, \rho, \Sigma) \, \mathcal{L}(\widetilde{w}_{ih} \mid \boldsymbol{s}_i, \widetilde{\boldsymbol{w}}_{i,-h}) \qquad (18)$$

where $\boldsymbol{s}_i = (s_{i1}, \ldots, s_{iH})^T$. The conditional prior $\pi(\widetilde{w}_{ih} \mid \widetilde{\boldsymbol{w}}_{-i}, \widetilde{\boldsymbol{w}}_{i,-h}, \rho, \Sigma)$ can be derived from (5) conditioning with respect to the other components of the vector $\widetilde{\boldsymbol{w}}_i$; we find

$$\pi(\widetilde{w}_{ih} \mid \widetilde{\boldsymbol{w}}_{-i}, \widetilde{\boldsymbol{w}}_{i,-h}, \rho, \Sigma) = \mathcal{N}(\mu^*_{ih}, \Sigma^*_{ih}),$$

where $\mu^*_{ih} = \mu_{ih} + \Sigma_{h,-h}\Sigma^{-1}_{-h,-h}(\widetilde{\boldsymbol{w}}_{i,-h} - \boldsymbol{\mu}_{i,-h})$ and $\Sigma^*_{ih} = (\rho \sum_{j=1}^I g_{ij} + 1 - \rho)^{-1} (\Sigma_{h,h} - \Sigma_{h,-h}\Sigma^{-1}_{-h,-h}\Sigma_{-h,h})$ by standard properties of the normal distribution, with $\boldsymbol{\mu}_i = (\rho \sum_{j=1}^I g_{ij} + 1 - \rho)^{-1}(\rho \sum_{j=1}^I g_{ij}\widetilde{\boldsymbol{w}}_j + (1-\rho)\widetilde{\boldsymbol{m}}_i)$. Moreover, using the same data augmentation scheme proposed in Holmes and Held (2006), we write the term $\mathcal{L}(\widetilde{w}_{ih} \mid \boldsymbol{s}_i, \widetilde{\boldsymbol{w}}_{i,-h})$ as

$$\mathcal{L}(\widetilde{w}_{ih} \mid \boldsymbol{s}_i, \widetilde{\boldsymbol{w}}_{i,-h}) = \left(\frac{e^{\eta_{ih}}}{1 + e^{\eta_{ih}}}\right)^{N_{ih}} \left(\frac{1}{1 + e^{\eta_{ih}}}\right)^{N_i - N_{ih}}$$

where $\eta_{ih} = \widetilde{w}_{ih} - C_{ih}$, $C_{ih} = \log \sum_{k \neq h} e^{\widetilde{w}_{ik}}$ (with $\widetilde{w}_{iH} := 0$) and N_{ih} is the number of observations in area i assigned to component h.

To be able to sample from the full conditional of $\widetilde{w}_{ih}$, we express $\mathcal{L}(\widetilde{w}_{ih} \mid \boldsymbol{s}_i, \boldsymbol{\widetilde{w}}_{i,-h})$ using an augmentation technique, based on the Pólya-Gamma distribution. The trick is analogous to that in Polson et al. (2013), in this case without covariates. We describe it in detail in the next paragraphs.

We denote by $\omega \sim PG(b, c)$ a random variable with a Pólya-Gamma distribution with parameters b and c, i.e.

$$\omega = \frac{1}{2\pi^2} \sum_{k=1}^{+\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)} \tag{19}$$

where $g_k \overset{\text{iid}}{\sim} Gamma(b, 1)$ and $b, c > 0$. The data-augmentation technique based on the Pólya-Gamma distribution relies on the following integral identity:

$$\frac{(\mathrm{e}^\eta)^a}{(1 + \mathrm{e}^\eta)^b} = 2^{-b} \mathrm{e}^{(a - b/2)\eta} \int_0^{+\infty} \mathrm{e}^{-\omega\eta^2/2} p(\omega) d\omega$$

where $p(\omega)$ is the density of the $PG(b, 0)$ random variable. Taking advantage from the above equality, when introducing the latent variable $\omega_{ih} \sim PG(N_i, 0)$, we can derive the following full conditional for $\widetilde{w}_{ih}$:

$$\mathcal{L}(\widetilde{w}_{ih} \mid \boldsymbol{\widetilde{w}}_{-i}, \boldsymbol{\widetilde{w}}_{i,-h}, \boldsymbol{s}_i, \rho, \Sigma, \omega_{ih}) = N(\hat{\mu}_{ih}, \hat{\Sigma}_{ih}) \tag{20}$$

where

$$\hat{\mu}_{ih} = \left(\frac{\mu_{ih}^*}{\Sigma_h^*} + N_{ih} - N_i/2 + \omega_{ih} C_{ih} \right) \left(\frac{1}{\Sigma_{ih}^*} + \omega_{ih} \right)^{-1} \qquad \hat{\Sigma}_{ih} = \left(\frac{1}{\Sigma_{ih}^*} + \omega_{ih} \right)^{-1}.$$

Moreover, the full conditional of ω_{ih} can be expressed as

$$\mathcal{L}(\omega_{ih} \mid \boldsymbol{\widetilde{w}}_i) = PG\left(N_i, \widetilde{w}_{ih} - \log \sum_{k \neq h} e^{\widetilde{w}_{ik}} \right). \tag{21}$$

See Appendix C for the proof of Equations (20)-(21).

These equations give a two steps Gibbs update for the variable $\widetilde{w}_{ih}$. Indeed, one can first sample ω_{ih} from (21) (which depends on $\widetilde{w}_{ih}$) and secondly update $\widetilde{w}_{ih}$ from (20) (which depends on ω_{ih}). In this way, we are able to make two Gibbs steps in an augmented state space instead of a single Metropolis Hastings step. There are two reasons why one should prefer the former algorithm to the latter. First, the two-Gibbs-steps simulation avoids the choice of a proposal density for the update, that can be difficult due to the shape of the logistic transformation applied to the weights. Moreover, using the Pólya Gamma augmentation trick can be helpful in settings where the number of observations in a single area is not significantly greater than the number of components in the mixture, as we consider in Section 6.1, scenario II; see Section S6.3 of the supplementary material in Polson et al. (2013) for an explanation of this statement.

6 Simulated data

We consider two simulation studies to illustrate the flexibility of our model; in particular we will see that the model is able to exploit spatial dependence between densities corresponding to close areas. In the first example, we compare our model (SPMIX) with the Clayton-Kaldor Species Sampling Model of Jo et al. (2017) (CK-SSM) and the HDP mixture model (see (17)), that we use as a sort of black-box model for density estimation of grouped data. In the second example we generate data from spatially dependent densities and we check if our model is flexible enough to recover such dependence.

We implemented the Gibbs sampler for our model (7)-(11) together with the prior specification (14) - (15), which can be found in Appendix C, and the *direct sampler* for the HDP-mixture model in Teh et al. (2006). In addition, we also implemented the Clayton-Kaldor Species Sampling Mixture Model in `Stan` (Stan Development Team, 2018) with the prior (16). All the MCMC chains were run for 10,000 iterations after discarding the first 10,000 iterations as burn-in, keeping one every five iterations, resulting in a final sample size of 2,000, unless otherwise specified. In all cases, convergence was checked using both visual inspection of the chains and standard diagnostics available in the CODA package.

The base measures for our model, for the HDP-mixture and for the CK-SSM mixture are assumed all equal (and denoted by P_0) to match the models under comparison. Unless otherwise stated, we assume P_0 equal to the Normal-inverse-gamma distribution with parameters $\mu_0 = 0, a = b = 2, \lambda = 0.1$, i.e. $\mu \mid \sigma^2 \sim \mathcal{N}\left(\mu_0, \lambda^{-1}\sigma^2\right)$, $\sigma^2 \sim IG(a, b)$ and the prior in (11) as $\rho \sim Beta(1, 1)$. For the HDP, the total mass parameters α and β are fixed and equal to 1. For our model, we set the prior hyperparameters for the marginal prior (10) of Σ as $\nu = 100$ and $V = \mathbb{I}$ for all the simulated examples. For the CK-SSM, we followed the hyperparameter tuning outlined in their paper, except for the parameters a and b that we fix to $a = 0.1$ and $b = 0.5$.

As metrics to compare the density estimates, i.e. the posterior mean of the density evaluated on a fixed grid, we use the Kullback-Leibler divergence and the Hellinger distance between the estimated density and the true one.

6.1 Non-Gaussian simulated data

We consider three scenarios. In each scenario we generate, for $I = 6$ different areas, an i.i.d. sample from a density that is not Gaussian: namely t-student (t), skew-normal (SN), chi-squared (χ^2) and Cauchy (Ca). The matrix G is fixed and represents a graph with only three connected components $\{1, 2\}$, $\{3, 4\}$, $\{5, 6\}$. The three scenarios differ in the number of data in each area and in the data generating densities, as reported in Table 1: $t(\nu, \mu, \sigma)$ denotes the Student's t distribution with ν degrees of freedom, centered in μ and scaled by a factor σ; $SN(\xi, \omega, \alpha)$ denotes the Skew normal distribution with mean $\xi + \omega\alpha/\sqrt{1 + \alpha^2}\sqrt{2/\pi}$, $\chi^2(k, 0, 1)$ denotes the standard chi-squared distribution with k degrees of freedom and $Ca(0, 1)$ the Cauchy distribution. They cover extremely different cases: in Scenario I a large number of data is available in each area, so that borrowing strength from nearby areas would be superfluous; we actually expect our model to perform worse than the HDP-mixture, being the latter fully nonparametric. On the other hand, in

	Area	1	2	3	4	5	6
Scenario I	Density	$t(6,-4,1)$	$t(6,-4,1)$	$SN(4,4,1)$	$SN(4,4,1)$	$\chi^2(3,0,1)$	$\chi^2(3,0,1)$
	N_i	1000	1000	1000	1000	1000	1000
Scenario II	Density	$t(6,-4,1)$	$t(6,-4,1)$	$SN(4,4,1)$	$SN(4,4,1)$	$\chi^2(3,0,1)$	$\chi^2(3,0,1)$
	N_i	1000	10	1000	10	1000	10
Scenario III	Density	$t(6,-4,1)$	$t(6,-4,1)$	$SN(4,4,1)$	$SN(4,4,1)$	$Ca(0,1)$	$Ca(0,1)$
	N_i	100	100	100	100	100	100

Table 1: Non-Gaussian simulated data: true densities and sample sizes for each area under all scenarios

	Model	1	2	3	4	5	6
Scenario I	SPMIX	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.01	0.02 ± 0.01
	HDP	0.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.01	0.02 ± 0.01
	CK-SSM	0.92 ± 0.46	0.92 ± 0.46	0.97 ± 0.16	0.98 ± 0.16	1.10 ± 0.31	1.10 ± 0.31
Scenario II	SPMIX	0.02 ± 0.00	0.04 ± 0.04	0.02 ± 0.01	0.02 ± 0.07	0.03 ± 0.01	0.03 ± 0.10
	HDP	0.01 ± 0.00	0.13 ± 0.04	0.03 ± 0.01	0.21 ± 0.07	0.03 ± 0.01	0.32 ± 0.10
	CK-SSM	0.91 ± 0.40	0.90 ± 0.40	0.97 ± 0.17	0.97 ± 0.17	1.22 ± 0.45	1.23 ± 0.44
Scenario III	SPMIX	0.15 ± 0.19	0.15 ± 0.18	0.09 ± 0.25	0.09 ± 0.25	0.06 ± 0.12	0.06 ± 0.12
	HDP	0.16 ± 0.19	0.16 ± 0.18	0.26 ± 0.25	0.26 ± 0.25	0.13 ± 0.12	0.13 ± 0.12
	CK-SSM	0.86 ± 0.33	0.86 ± 0.34	1.25 ± 0.29	1.25 ± 0.29	0.86 ± 0.41	0.86 ± 0.42

Table 2: Kullback Leibler divergences between the true densities and the estimated ones, aggregated over 100 simulated datasets with $\pm$ one standard deviation

Scenario II there are three areas (2, 4 and 6) with few data points (only 10). In this case, we expect our model to express its strength and give a better density estimate than the HDP-mixture, especially in those areas where few data are present. Finally, Scenario III is an in-between condition, where not so many observations as in Scenario I are available in each area. We also compare the results obtained with the CK-SSM mixture.

In order to make a fair comparison between our model, CK-SSM and the HDP models, we fixed the number of components H in our mixtures and in the CK-SSM to 10. This choice was made by looking at the posterior distribution of the number of components under the HDP-mixture in the different scenarios; we found that the number of clusters ranges between 3 and 10. For each scenario we repeatedly simulated 100 independent datasets. Table 2 shows the KL-divergence between the true density and the estimate under the three models. We average those values over the 100 simulated datasets, also considering $\pm$ one empirical standard deviation of the 100 values obtained. Table 4 in Appendix D, reports the same values for the Hellinger distance. From both tables, we can see that in all the three scenarios, the CK-SSM has the worst performance in recovering the true data generating density. This reflects what we discussed in Section 4.3: the prior of such model forces mixture weights to be too similar across different connected components in the graph. We can clearly see this for example from Figure 3, where the density estimates for areas 4 and 6 (not connected in G) are close under the CK-SSM but not under our model. The HDP-mixture gives overall better estimates than those under our model in Scenario I. In this case, both density estimates are close enough to the true densities; see Figure 3. As we expect, under Scenario II, our model gives a better density estimate (than the HDP-mixture) in areas 2, 4 and 6, where only 10 data points are available; see Figure 4. Indeed, our model retrieves information from the neighbouring
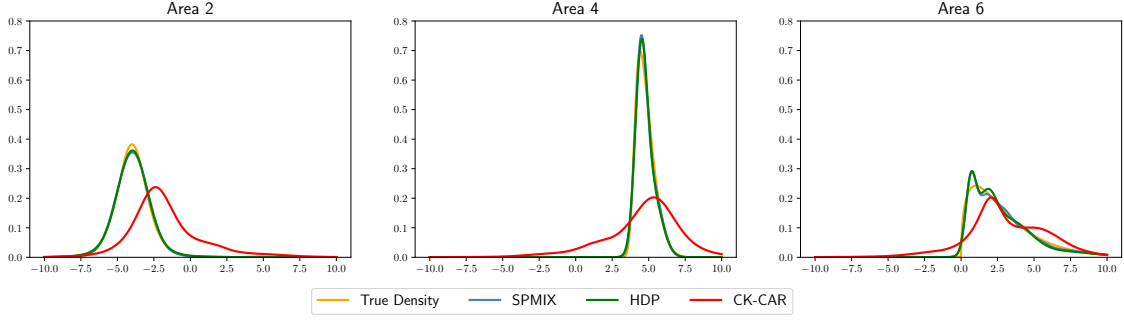
Figure 3: Non-Gaussian simulated data, Scenario I: true densities for areas 2, 4 and 6 and the corresponding density estimates under the three different models.
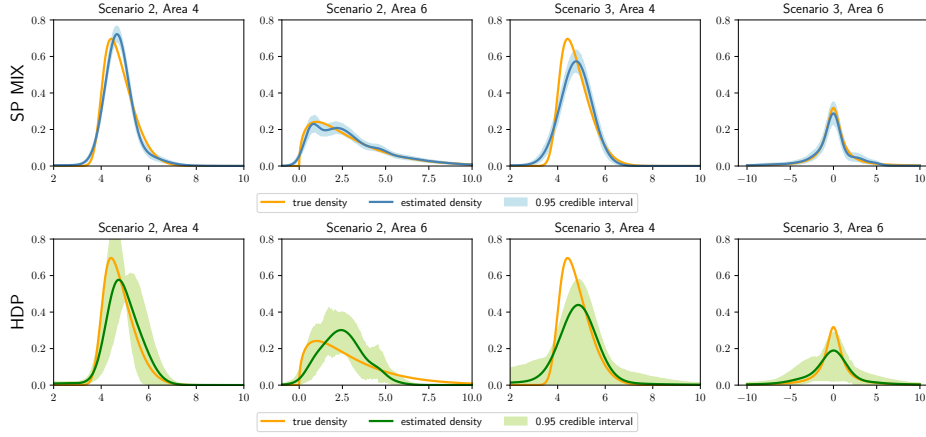

Figure 4: Non-Gaussian simulated data, Scenario II and III: estimated and true densities for areas 4 and 6 under our (top row) and HDP mixture (bottom row) models.

areas, overcoming the lack of data in some of the areas. Interestingly, our model performs better in areas 3-6, and similarly in areas 1 and 2, under Scenario III, probably because of "extreme" data in areas 5 and 6, where we generate data from a Cauchy distribution. This behavior is evident from Figure 4, being the 95% point-wise credible interval of the posterior distribution of the density much wider in HDP than in our approach. Finally, it is clear that our model fits data well also when the true density is highly non-symmetric, such as in locations 3-6 in Scenarios I and II.

6.2 Simulation from spatially dependent weights

In the second simulation example we apply our model to estimate spatially dependent densities in contiguous areas, placed in a squared grid with a total number I of areas, in a unit area squared domain; we study three different scenarios, choosing $I = 16, 64, 256$. In the i-th area, we simulate observations as follows:

$$y_{ij} \overset{\text{iid}}{\sim} w_{i1}\mathcal{N}(-5, 1) + w_{i2}\mathcal{N}(0, 1) + w_{i3}\mathcal{N}(5, 1) \quad j = 1, \ldots, 25 \tag{22}$$

where the weights are chosen as $alr^{-1}(\tilde{\boldsymbol{w}}_i)$ and the transformed weights $\tilde{\boldsymbol{w}}_i$ are given by

$$\tilde{w}_{i1} = 3(x_i - \bar{x}) + 3(y_i - \bar{y}) \quad \tilde{w}_{i2} = -3(x_i - \bar{x}) - 3(y_i - \bar{y}) \tag{23}$$

where (x_i, y_i) are the coordinates of the center of area i and $(\bar{x}, \bar{y})$ the coordinates of the grid center. In this way, we introduce strong spatial dependence, induced by (23), among the weights of different areas, as we observe in Figures 5a and 5b, where we plot the weights of the first two components $\{w_{i1}\}$ and $\{w_{i2}\}$, for the scenario $I = 64$; of course $w_{i3} = 1 - w_{i1} - w_{i2}$.
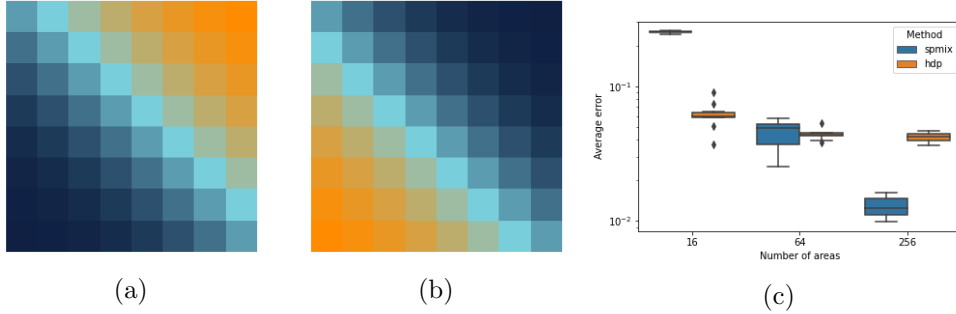

(a)　　　　　　　(b)　　　　　　　(c)

Figure 5: Simulation from spatially dependent weights in Sect. 6.2: plots of $\{w_{i1}\}$ (a), $\{w_{i2}\}$ (b) when $I = 64$; boxplots (c) of the Kullback-Leibler divergence between true density (22) and estimated one under our model (spmix) and the HDP-mixture model (hdp) for each simulation, averaged over the areas, for $I = 16, 64, 256$, in logarithmic scale.

In our model, we consider areas i and j to be neighbors if they share an edge, setting $g_{ij} = 1$ in (9) in this case, and $g_{ij} = 0$ otherwise. For each scenario, we simulated 10 independent datasets, sampling 25 observations per area, and then we compare the posterior estimates of the densities with the true ones via Kullback-Leibler divergence. We compare our model with the HDP-mixture model, reporting in Figure 5c the errors, averaged over all areas, for the ten repetitions. Though when $I = 16$ HPD gives much better estimates, our model outperforms the HDP-mixture, when the number of areas is sufficiently large, with consistent results using the Hellinger distance to measure the errors, as shown in Figure 12 in Appendix D.

The median execution times, over the 10 datasets, of the code corresponding to our model was 25.28, 118.14 and 616.41 seconds for $I = 16, 64, 256$, respectively, whereas, for fitting data for the HPD-mixtures was 18.39, 72.59, 207.46 seconds. Based on our implementation, HDP is slightly faster, but our model still exhibits competitive computational times, paired with lower errors when the number of areas is large. Simulations were performed on a machine equipped with a 4x Xeon E5-2640 v4 @2.4GHz processor and 64 GB of RAM. In order to provide a fair comparison, the implementation for our model ran on a single core since the sampler for the HDP is inherently sequential. However, the Gibbs sampler we proposed can be straightforwardly parallelized and this could greatly decrease the runtimes of our model.

7 Airbnb Amsterdam

We consider the Airbnb listings dataset for the city of Amsterdam (Netherlands), publicly available at `http://insideairbnb.com/get-the-data.html`. The dataset consists of more than $20,000$ listings spread over Amsterdam, grouped by the neighborhood. Our goal is to predict the nightly price of a new listing, with information given by covariates, and taking into account the spatial dependence. For this reason, we generalize the model adding one more level to the likelihood (a regression model). Specifically, we assume for responses y_{ij} (i.e. the nightly price of accommodation j in neighborhood i), conditional to d covariates $\boldsymbol{x}_{ij} = (x_{ij1}, \dots, x_{ijd})^T$ and parameters, the same model (7)-(11), where τ_h in (7) are such that $\tau_h = (\mu_h + \boldsymbol{\beta}^T \boldsymbol{x}_{ij}, \sigma_h^2)$, $h = 1, \dots, H$. The prior specification is completed assuming

$$\boldsymbol{\beta} \sim \mathcal{N}_d(\mathbf{0}, \sigma_\beta^2 \mathbb{I}_d) \tag{24}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^T$ is the vector of regression parameters common to all the areas.

7.1 Data description and exploratory data analysis

We consider as predictors characteristics of the house such as: (i) `accommodates`, the number of guests that can be hosted, (ii) `bathrooms`, the number of bathrooms, (iii) `bedrooms`, the number of bedrooms; together with two indicators of popularity of the listing: (iv) `number_of_reviews`, the number of reviews present for that listing, (v) `review_scores_rating` the average rating of the reviews. Finally, we complete the set of covariates with two binary variables: (vi) `instant_bookable` which equals 1 if the listing can be booked instantly from the user and 0 if, instead, the request must go through an acceptance procedure from the host; (vii) `host_is_superhost` that is 1 if the host is classified as a *superhost* by Airbnb and 0 otherwise. The *superhost* badge can be obtained once a host has a sufficient number of reviews with a rating above a certain threshold. These binary variables were included since the user, while searching for an accommodation, can reduce her/his search only to instant bookable listings and/or only to *superhosts*.

As preprocessing, we used the following steps: we removed the listings for which at least one predictor is missing, as well as listings whose nightly price is below two euros or above one thousand euros; then we transformed `number_of_reviews` by taking the natural logarithm and `review_scores_rating` by the Box-Cox transformation $x_i^{(\lambda)} = (x_i^\lambda - 1)/\lambda$ (Box and Cox, 1964) with $\lambda = 12$.

Each listing is assigned to one of the twenty-two Amsterdam neighborhoods provided at the dataset web page, so that $I = 22$. Figure 6(a) shows sample sizes for each neighborhood (in the log scale): of course, city center is the area with the largest number of observations. Furthermore, in Figure 6, panels (b) and (c), we report sample means and standard deviations of the nightly price in each neighborhood; the plots motivate the modeling of the spatial dependence, as close neighborhoods tend to have similar distributions, at least in terms of mean and standard deviation. It is clear from Figure 6 that there are two distinct graph connected components, one made only by three areal units. As far as covariates are concerned, Figure 13 in Appendix D, shows empirical correlations among

(a) No. of listings in the log-scale

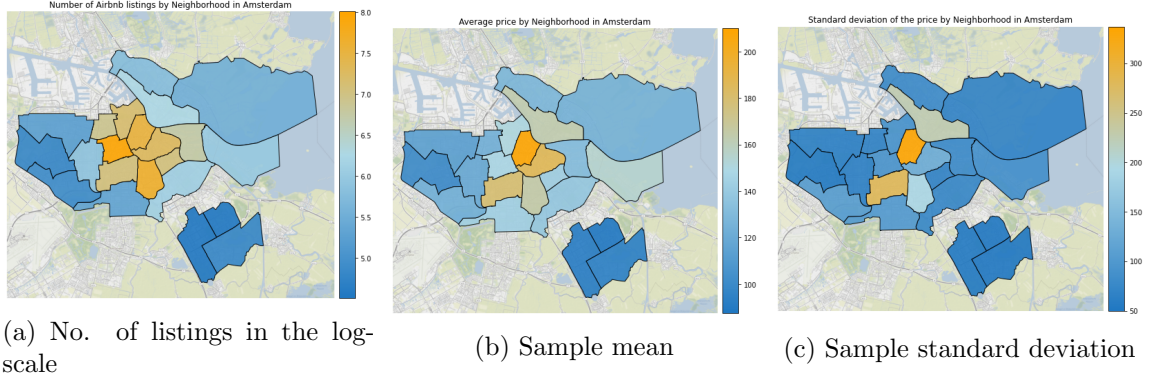(b) Sample mean

(c) Sample standard deviation

Figure 6: Number of listings, sample means and standard deviations of the nightly price for each neighborhood in the Airbnb dataset, after preprocessing.

the predictors and, in the last column, between predictors and the response. Figure 14 in in Appendix D, displays scatterplots of the response price versus numerical predictors and boxplots for categorical predictors. We note that only `accommodates`, `bathrooms`, `bedrooms` exhibit a significant linear correlation with the price, which is confirmed by the scatter plots, but they are strongly correlated among them. On the other hand, there is no empirical effect of `instant_bookable` and `host_is_superhost` on the nightly price.

7.2 Posterior inference and findings

We fit model (7) - (11) and (14)-(15) with the prior (24), including all the covariates described in the previous section. We first standardize all the numerical predictors, subtracting the sample mean and dividing by the standard deviation of each predictor; we also center the responses on the overall sample mean. We modified the original Gibbs sampler by adding a further step where we sample from the full conditional of $\boldsymbol{\beta}$ (see Appendix C) .

We fix hyperparameters in P_0 as follows: $\mu_0 = 0$, $a = 0.5$, $b = 2.0$, $\lambda = 0.5$. We checked that posterior inference is robust to the choice of all the hyper-parameters in the prior distribution, but the number H of components in the mixture. Hence, we perform a sensitivity analysis by comparing the prediction error for different values of H, through a 10-fold cross-validation. The cross-validation is stratified according to the areas, so that, each time, approximately 10% of the data is missing from each neighborhood. More in detail, each time we selected 90% of the dataset as "training set" (to compute the posterior) and run the MCMC algorithm to compute the mean of the predictive distribution corresponding to data in the "testing set", when $H = 5, 10, 15, 20$. Observe that the exact same datapoints are shared across all values of H, both for training and for testing. Then we computed the mean square prediction error (MSE) on the test set, i.e. the quantity $\sum_{i=1}^{m} (y_i - \hat{y}_i)^2 / m$, where m is the cardinality of the test set and $\hat{y}_i$ is the mean of the predictive density for covariates x_i. The values of the MSE were very robust with respect to H values. Hence, for all subsequent analysis, we fixed $H = 5$ for a more parsimonious model.

Figure 7(b) reports 95% posterior credibility intervals for the regression parameters. All the covariates, except for `host_is_superhost`, seem to be significant, if we assume hard shrinkage as a criterion for significance, i.e. the marginal credibility interval does not include 0. It is interesting to observe how the coefficients associated to `number_of_reviews` and `instant_bookable` are negative. This might indicate that hosts that receive many reviews and many reservations could tend to lower their prices in order be more attractive. On the other hand, as one would expect, all the other coefficients are positive, the one furthest right being associated to `accommodates`, i.e. the number of guests that can be hosted. In Figure 7(c) we show the density estimates in the four neighborhoods highlighted
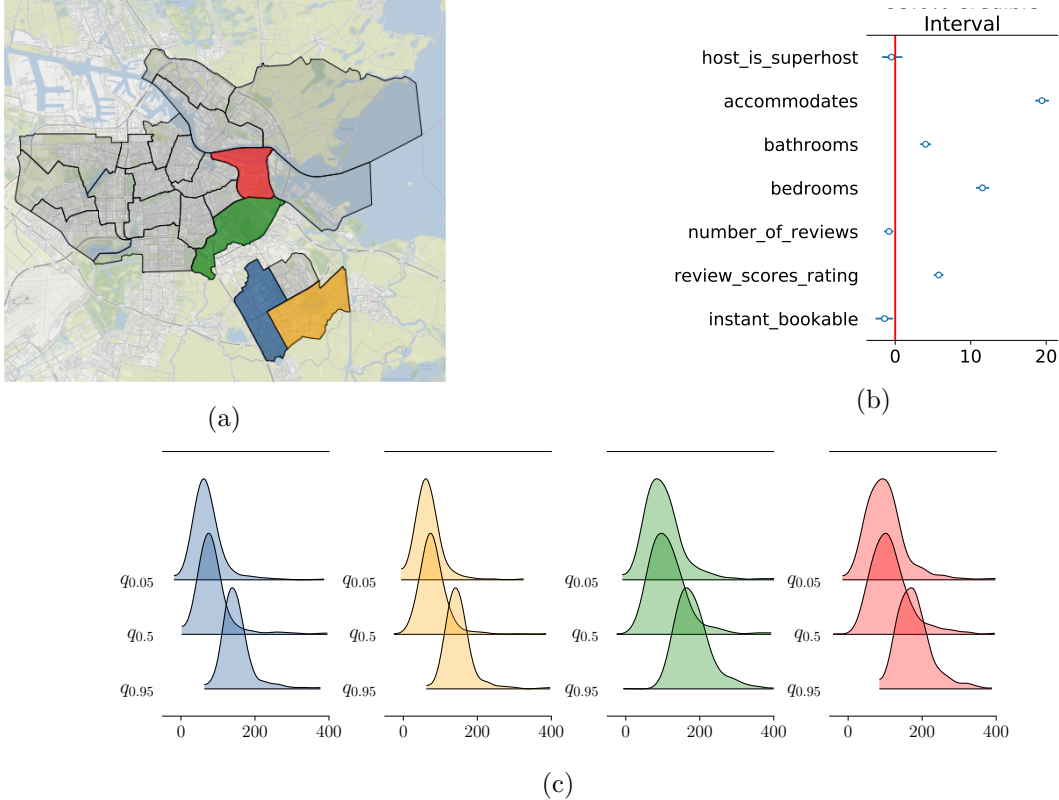

(a)

(b)

(c)

Figure 7: (a): Map of the city of Amsterdam with neighborhoods *Bijlmer-Centrum* in blue, *Gaasperdam - Driemond* in orange, *Oostelijk Havengebied - Indische Buurt* in green and *Watergraafsmeer* in red. (b): 95% credibility intervals of the marginal posterior of the regression parameter $\boldsymbol{\beta}$. (c): Predictive densities for different neighborhoods, the colors match the ones of the map. In each plot, three lines represent three different values of the covariates `accomodates`, `number_of_bedrooms`, `number_of_bathrooms`) that take range in the first, second and third quantile, while the other numerical covariates are fixed to the empirical median.

in Figure 7(a). Each plot shows three density estimates, corresponding to different values of the covariates. In this case, the covariates were set to the empirical median except for `accomodates`, `number_of_bedrooms`, `number_of_bathrooms`. Since the marginal sample correlation between these three covariates is high (see Figure 13 in Appendix D), we fixed the three of them equal to 5%, 50% and 95% empirical quantiles. For instance, in each panel of Figure 7(c), the top plot corresponds to a vector of covariates in which

`accomodates`, `number_of_bedrooms`, `number_of_bathrooms` are fixed to their 5% sample quantile, respectively. It is clear from Figure 7(c) that the predictive densities in blue and in yellow are similar, as well as the lines in green and in red. However there are evident differences when comparing for instance the yellow densities with the green ones; indeed the green densities give substantial mass to the right tail, especially to values greater than 200, while the yellow densities do not. This behavior agrees with the marginal posterior of ρ, that is strongly concentrated near 1 ($\mathbb{E}(\rho|data) = 0.993$): in fact, the blue and yellow neighborhoods, as well as the red and green ones, are connected in the graph. However, blue and yellow predictive densities are different from the green and red estimates, since the neighborhoods belong to different connected components in the graph. Finally, as it is to be expected, in all the neighborhoods the listings price increases as the `accomodates`, `number_of_bedrooms`, `number_of_bathrooms`.

8   Discussion

In this work, we have considered the problem of the joint estimation of spatially dependent densities in the context of repeated areal measurements. We have presented a finite mixture model to represent the density in each area; assuming that all the mixtures share the same set of atoms, the spatial dependency has been introduced through a novel joint distribution for I vectors in the simplex as a prior for the mixture weights. This distribution, that we termed logisticMCAR, was built as a logistic transformation of a specification of the multivariate CAR model. When compared to alternatives proposed in the literature, the logisticMCAR distribution showed to have a higher degree of interpretability, as we were able to derive the analytic expression for the expected values of ratios of components and their covariances, via the Aitchison geometry. Moreover, we also showed as the logisticMCAR can be used to accurately model sparse mixtures.

Posterior simulation has been carried out by means of a Gibbs sampler scheme. In particular the update of the mixture weights was performed by introducing a data augmentation scheme based on the Pólya-Gamma identity, which avoids the tedious tuning of the proposal distribution.

In the simulation studies and the real application included in this paper, our model has shown to be able to represent a wide range of different behaviours. In particular, we argue that when different connected graph components are present, and heterogeneous behaviour is observed across these components, our model should be preferred as it does not force the densities in different graph component to behave too similarly. Moreover, as in the case of the Airbnb Amsterdam application, our model can be easily extended to include additional covariate information. Although not our target here, a sub-product of the approach is the prior induced on the partition of the subjects in the sample, which in this case, has a spatial connotation; relations with spatial product partition models (Page and Quintana, 2016) could be further investigated.

Another point that we did not address here, and will be focus of future study, is an extension to boundary detection problems, i.e. when the proximity matrix G is unknown, but its elements depend on dissimilarity metrics available for each pair of neighboring

areas. This is an extremely interesting problem, widely studied in the context of one single response per area; see, for instance Lu et al. (2007) and Lee and Mitchell (2012). However preliminary investigation showed how the non-identifiability of overfitted mixtures might produce erroneous results. Possible extensions of our model to account for boundary detection might then include either a prior on the number of components or a repulsive prior distribution on the atoms, or both, to reduce the impact of non-identifiability.

A Proofs

Proof of Proposition 1

From equation (4) we have that

$$\mathbb{E}\left[\widetilde{\boldsymbol{w}}_i \mid \widetilde{\boldsymbol{w}}_{-i}\right] = \frac{\rho\sum_{j\in U_i}\widetilde{\boldsymbol{w}}_j + (1-\rho)\widetilde{\boldsymbol{m}}_i}{\rho|U_i|+1-\rho} = \frac{\rho\sum_{j\in U_i} alr(\boldsymbol{w}_j) + (1-\rho)alr(\boldsymbol{m}_i)}{\rho|U_i|+1-\rho}$$

$$= \frac{1}{\rho|U_i|+1-\rho}\left(\log\frac{\prod_{j\in U_i}w_{j1}^\rho m_{i1}^{1-\rho}}{\prod_{j\in U_i}w_{jH}^\rho m_{iH}^{1-\rho}},\ldots,\log\frac{\prod_{j\in U_i}w_{jH-1}^\rho m_{iH-1}^{1-\rho}}{\prod_{j\in U_i}w_{jH}^\rho m_{iH}^{1-\rho}}\right)$$

$$= \frac{1}{\rho|U_i|+1-\rho}\left(\sum_{j\in U_i}\log(w_{j1}^\rho m_{i1}^{1-\rho}),\ldots,\sum_{j\in U_i}\log(w_{jH-1}^\rho m_{iH-1}^{1-\rho})\right)$$

$$-\sum_{j\in U_i}\log(w_{jH}^\rho m_{iH}^{1-\rho})$$

where the last subtraction is meant elementwise. Hence we have that

$$\mathbb{E}\left[\log\frac{w_{il}}{w_{ik}}\mid \boldsymbol{w}_{-i}\right] = \mathbb{E}\left[\widetilde{w}_{il}-\widetilde{w}_{ik}\mid \boldsymbol{w}_{-i}\right]$$

$$= \frac{1}{\rho\,|U_i|+1-\rho}\left(\sum_{j\in U_i}\log(w_{jl}^\rho m_{il}^{1-\rho})-\sum_{j\in U_i}\log(w_{jk}^\rho m_{ik}^{1-\rho})\right)$$

$$= \log\left(\left(\frac{m_{il}}{m_{ik}}\right)^{1-\rho}\prod_{j\in U_i}\left(\frac{w_{jl}}{w_{jk}}\right)^\rho\right)^{\frac{1}{\rho|U_i|+1-\rho}}$$

which proves the proposition. $\qquad\square$

Proof of Proposition 2

The (marginal) joint distribution of two different components of $\widetilde{\boldsymbol{w}}_i,\widetilde{\boldsymbol{w}}_j$, with $i,j = 1,\ldots,I$, $i\neq j$ can be easily derived from (5):

$$\begin{pmatrix}\widetilde{w}_{il}\\\widetilde{w}_{jm}\end{pmatrix}\sim\mathcal{N}_2\left(\boldsymbol{0},\begin{bmatrix}A_{ii}\Sigma_{ll} & A_{ij}\Sigma_{lm}\\A_{ji}\Sigma_{ml} & A_{jj}\Sigma_{mm}\end{bmatrix}\right)\quad l,m=1,\ldots,H-1$$

Hence, we compute the covariance of the log ratios of different components as

$$\mathrm{Cov}\left(\log\frac{w_{il}}{w_{im}},\log\frac{w_{jl}}{w_{jm}}\right) = \mathrm{Cov}\left(\widetilde{w}_{il}-\widetilde{w}_{im},\ \widetilde{w}_{jl}-\widetilde{w}_{jm}\right)$$

$$= \mathrm{Cov}\left(\widetilde{w}_{il},\widetilde{w}_{jl}\right)+\mathrm{Cov}\left(\widetilde{w}_{il},\widetilde{w}_{jm}\right)++\mathrm{Cov}\left(\widetilde{w}_{im},\widetilde{w}_{jl}\right)+\mathrm{Cov}\left(\widetilde{w}_{im},\widetilde{w}_{jm}\right)$$

$$= A_{ij}\left(\Sigma_{ll}-2\Sigma_{lm}+\Sigma_{mm}\right)$$

whereas, for the last component,

$$\text{Cov}\left(\log\frac{w_{il}}{w_{iH}}, \log\frac{w_{jl}}{w_{jH}}\right) = \text{Cov}\left(\widetilde{w}_{il}, \widetilde{w}_{jl}\right) = A_{ij}\Sigma_{ll}$$

which proves the formula in the proposition.

It is possible to rearrange the indices $1, \ldots, I$ in order for $(F - \rho G)$ to be a block diagonal matrix, where each block corresponds to a connected graph component according to the neighbouring structure; this will not affect the joint law. By the properties of strictly diagonally dominated matrices, the same pattern of blocks is preserved in the inverse matrix A. Hence $A_{ij} = 0$ if i and j belong to two non-connected graph components, proving the proposition. □

B MC simulations from the logisticMCAR distribution

In Section 3 we have pointed out that the theoretical analysis of the logisticMCAR distribution is limited by its analytic intractability. Here we compute covariances between different components of the vectors of weights and Euclidean distances between the vectors themselves through Monte Carlo simulation. Specifically, we simulate from (5) and then obtain draws from the logisticMCAR distribution through the transformation alr^{-1}.

In particular, we fix $I = 5$, $H = 3$, $\widetilde{\boldsymbol{m}}_i = 0$ for all i and the covariance matrix Σ

$$\Sigma = \begin{bmatrix} 1 & \Sigma_{12} \\ \Sigma_{12} & 1 \end{bmatrix}$$

where Σ_{12} denotes the covariance, but also the correlation since $\Sigma_{11} = \Sigma_{22} = 1$, between $\widetilde{w}_{i1}$ and $\widetilde{w}_{i2}$. We fix the proximity matrix G such that $g_{12} = g_{13} = g_{23} = 1$ and $g_{45} = 1$. This corresponds assuming that areal units/nodes 1, 2 and 3 are connected to each other, and 4 and 5 are connected to each other, though separated from the others.
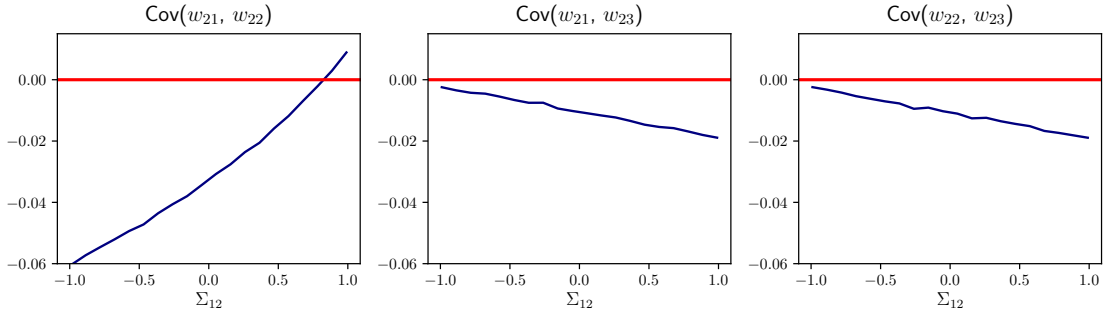

Figure 8: Pairwise covariance values of the components of $\boldsymbol{w}_2 = (w_{21}, w_{22}, w_{23})$ as a function of the correlation parameter Σ_{12}. The horizontal red line indicates the value 0.

Figure 8 shows the covariance between the three components of $\boldsymbol{w}_2 = (w_{21}, w_{22}, w_{23})$

as a function of the correlation parameter Σ_{12} in the matrix Σ in (5), having simulated $N = 10,000$ MC draws. Note that, unlike the finite-dimensional Dirichlet distribution, the logistic-normal distribution may have positive covariance among the components.
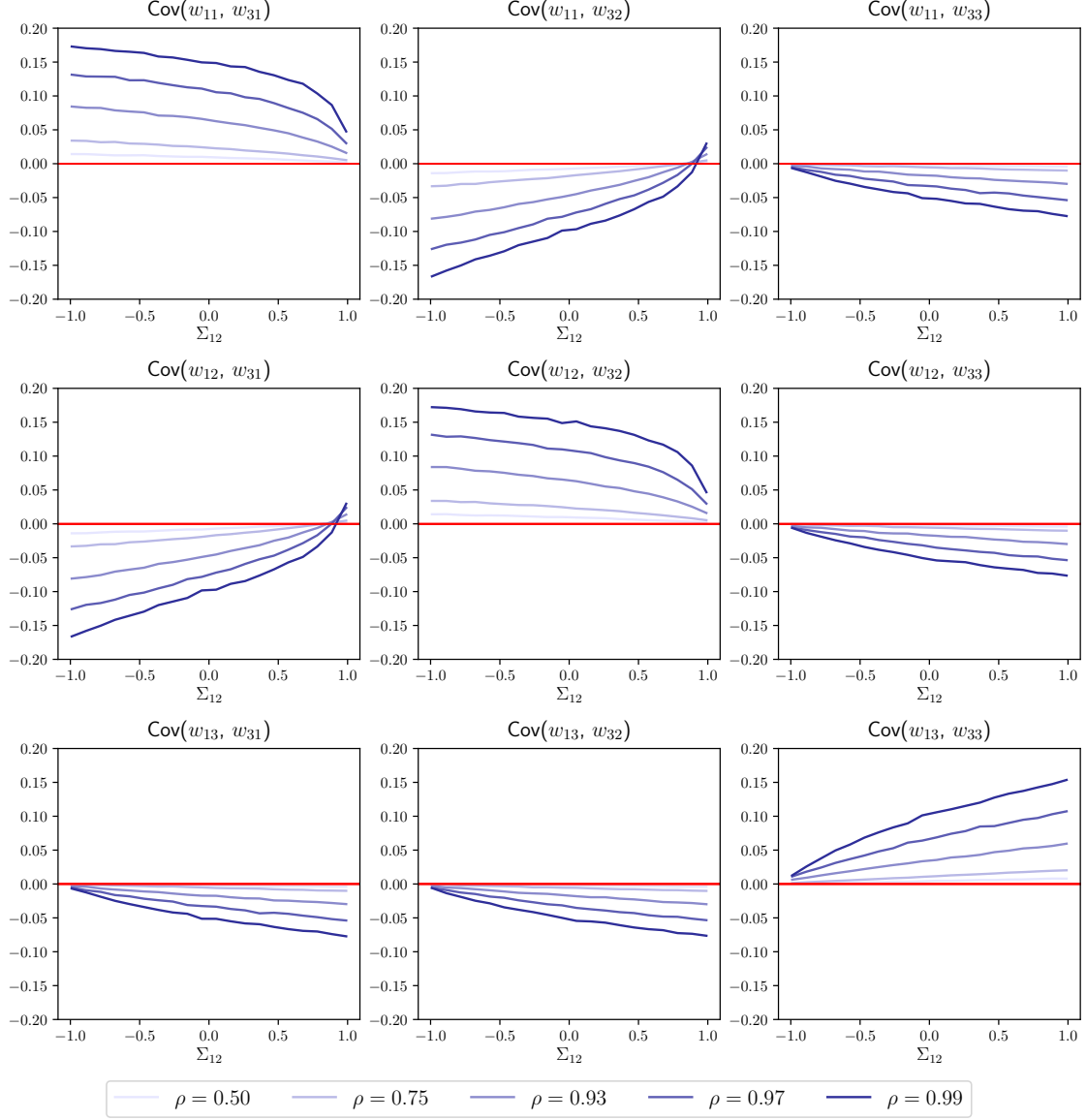

Figure 9: Pairwise covariance values between components of $\boldsymbol{w}_1$ and $\boldsymbol{w}_3$, as a function of the correlation parameter Σ_{12}, for different values of ρ, when $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_5)$ has the logisticMCAR distribution. The horizontal red line indicates the value 0.

Figure 9 instead shows the covariance between all the possible pairs (w_{1j}, w_{3m}) for $j, m = 1, 2, 3$, for different values of the parameter ρ. The covariances between corresponding entries, i.e. $(w_{1j}, w_{2j} \ j = 1, 2, 3)$ is always positive, as expected since the spatial correlation parameter ρ is always fixed to a positive value. The marginal prior for $\boldsymbol{w}_1, \boldsymbol{w}_3$ is exchangeable, since nodes 1 and 3 belong to the same connected component in G. This explains the symmetries in Figure 9.

	min	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	max
d_{12}	4×10^{-4}	0.10	0.18	0.27	0.86
d_{15}	0.01	0.33	0.55	0.77	1.36
d_γ	0.007	0.31	0.52	0.71	1.36

Table 3: Summary statistics of the marginal distributions of the distances d_{12}, d_{15}, d_γ, estimated from the MC samples; q_α denotes the α-quantile.

In order to measure the association induced by our logisticMCAR prior, we simulate the distances (Euclidean) of two vectors drawn from the joint distribution. In partic-
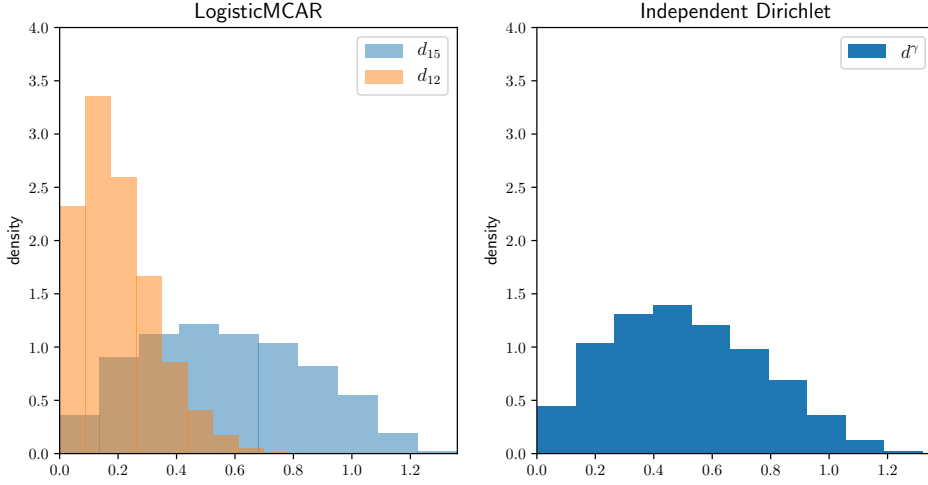

Figure 10: Histogram of MC draws from the marginal distributions of d_{12} (orange) and d_{15} (light blue) on the left and from the marginal distribution of d^γ on the right.

ular, we simulated $N = 10,000$ draws from the full joint logisticMCAR distribution of $(\boldsymbol{w}_1, \dots, \boldsymbol{w}_5)$ with parameters as above, fixing $\Sigma_{12} = 0.5$, and computed the Euclidean distances $d_{12} = ||\boldsymbol{w}_1 - \boldsymbol{w}_2||$ and $d_{15} = ||\boldsymbol{w}_1 - \boldsymbol{w}_5||$. As $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ belong to the same connected graph component while $\boldsymbol{w}_5$ belongs to another component, we expect $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ to be more similar than $\boldsymbol{w}_1$ and $\boldsymbol{w}_5$ belonging to separate components. Hence the distance d_{12} should be smaller than d_{15}. Moreover, for comparison, we also simulated $N = 10,000$ draws from the joint distribution of two independent finite-dimensional Dirichlet random variables, i.e.

$$(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)_i \overset{\text{iid}}{\sim} \text{Dir}(\mathbf{1}) \times \text{Dir}(\mathbf{1}) \quad i = 1, \dots N$$

and computed their Euclidean distance as well, that we denote by d^γ. Figure 10 reports the histograms of the marginal distributions of d_{12}, d_{15} on the left and d^γ on the right. It is clear that d_{12} is substantially smaller than d_{15}, as expected. Moreover, by comparing d_{15} and d^γ, we see that their marginal distributions are very similar. See also the summary statistics of these marginal distributions in Table 3.

For more insight, we report a subsample of size $N = 20$ of the MC simulated values from the marginal distributions of $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ and $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$, plotted on the two dimensional

projection of the simplex S^3 in Figure 11. Each pair is denoted by two points inside the triangle and a line connecting them. It is clear that simulated values from $\mathcal{L}(\boldsymbol{w}_1, \boldsymbol{w}_2)_i$ are much closer each other than those from $\mathcal{L}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$.
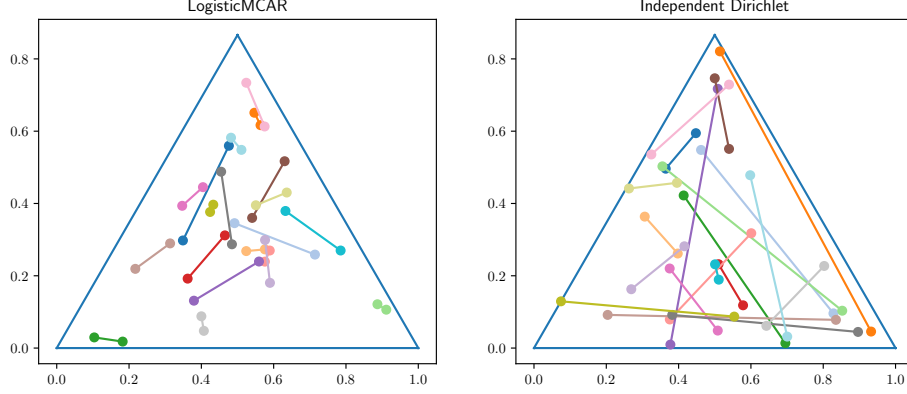

Figure 11: Plots of $N = 20$ MC draws from the logisticMCAR distribution (left) and of $N = 20$ MC draws from $\mathrm{Dir}(\mathbf{1}) \times \mathrm{Dir}(\mathbf{1})$ (right). Each draws is represented with two dots (the values of the two random vectors) together with a colored line connecting them for visual purposes. Different colors correspond to independent draws.

C   The Gibbs sampler

Proof of Equations (20) **-** (21)

We start by writing the full conditional distribution (18) as follows:

$$\mathcal{L}(\widetilde{w}_{ih} \mid \widetilde{\boldsymbol{w}}_{-i}, \widetilde{\boldsymbol{w}}_{i,-h}, \boldsymbol{s}_i, \rho, \Sigma) \propto \mathcal{N}(\widetilde{w}_{ih} \mid \mu_{ih}^*, \Sigma_{ih}^*) \left(\frac{\mathrm{e}^{\eta_{ih}}}{1 + \mathrm{e}^{\eta_{ih}}}\right)^{N_{ih}} \left(\frac{1}{1 + \mathrm{e}^{\eta_{ih}}}\right)^{N_i - N_{ih}}$$

$$\propto \mathcal{N}(\widetilde{w}_{ih} \mid \mu_{ih}^*, \Sigma_{ih}^*) \frac{(\mathrm{e}^{\eta_{ih}})^{N_{ih}}}{(1 + \mathrm{e}^{\eta_{ih}})^{N_i}}$$

$$\propto \mathcal{N}(\widetilde{w}_{ih} \mid \mu_{ih}^*, \Sigma_{ih}^*) \, \mathrm{e}^{(N_{ih} - N_i/2)\eta_{ih}} \int_0^\infty \mathrm{e}^{-\omega_{ih}\eta_{ih}^2/2} p(\omega_{ih}) d\omega_{ih}$$

where $\omega_{ih} \sim PG(N_i, 0)$. We now include the latent variable ω_{ih} and derive the conditional distribution of $\widetilde{w}_{ih}$, conditioning also to ω_{ih}. We have

$$\mathcal{L}(\widetilde{w}_{ih} \mid \widetilde{\boldsymbol{w}}_{-i}, \widetilde{\boldsymbol{w}}_{i,-h}, \boldsymbol{s}_i, \rho, \Sigma, \omega_{ih}) \propto \mathcal{N}(\widetilde{w}_{ih} \mid \mu_{ih}^*, \Sigma_{ih}^*) \, \mathrm{e}^{(N_{ih} - N_i/2)\eta_{ih}} \mathrm{e}^{-\omega_{ih}\eta_{ih}^2/2}$$

$$\propto \mathrm{e}^{-\frac{E}{2}}$$

where

$$E = \frac{(\widetilde{w}_{ih} - \mu_{ih}^*)^2}{\Sigma_{ih}^*} - (2N_{ih} - N_i)(\widetilde{w}_{ih} - C_{ih}) + \omega_{ih}(\widetilde{w}_{ih} - C_{ih})^2$$

$$\propto \widetilde{w}_{ih}^2 \left(\frac{1}{\Sigma_{ih}^*} + \omega_{ih} \right) - 2\widetilde{w}_{ih} \left(\frac{\mu_{ih}^*}{\Sigma_{ih}^*} + N_{ih} - N_i/2 + \omega_{ih} C_{ih} \right)$$

$$\propto \left(\frac{1}{\Sigma_{ih}^*} + \omega_{ih} \right) \left(\widetilde{w}_{ih}^2 - 2\widetilde{w}_{ih} \left(\frac{\mu_{ih}^*}{\Sigma_{ih}^*} + N_{ih} - N_i/2 + \omega_{ih} C_{ih} \right) \left(\frac{1}{\Sigma_{ih}^*} + \omega_{ih} \right)^{-1} \right)$$

Thus

$$\mathcal{L}(\widetilde{w}_{ih} \mid \widetilde{\boldsymbol{w}}_{-i}, \widetilde{\boldsymbol{w}}_{i,-h}, \boldsymbol{s}_i, \rho, \Sigma, \omega_{ih}) \sim \mathcal{N}(\hat{\mu}_{ih}, \hat{\Sigma}_{ih})$$

where

$$\hat{\mu}_{ih} = \left(\frac{\mu_{ih}^*}{\Sigma_h^*} + N_{ih} - N_i/2 + \omega_{ih} C_{ih} \right) \left(\frac{1}{\Sigma_h^*} + \omega_{ih} \right)^{-1} \quad \hat{\Sigma}_{ih} = \left(\frac{1}{\Sigma_h^*} + \omega_{ih} \right)^{-1}$$

For the full conditional of ω_{ih} instead, it is sufficient to apply Theorem 1 in Polson et al. (2013) with $\psi = \eta_{ih}$ to obtain that the law of ω_{ih}, conditional to $\widetilde{\boldsymbol{w}}_i$ is a Pólya-Gamma distribution, i.e. the density of ω_{ih} can be expressed as in Equation (19), with parameters $b = N_i$, $c = \widetilde{w}_{ih} - \log \sum_{k \neq h} \exp(\widetilde{w}_{ik})$.

$\square$

Detailed description of the Gibbs sampler

The state of the MCMC sampler is made of $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_H)$, $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$, where $\widetilde{\boldsymbol{w}}_i = \text{alr}(\boldsymbol{w}_i)$, $\{s_{ij}\}_{ij}$ and $\widetilde{m}_{C_1}, \ldots \widetilde{\boldsymbol{m}}_{C_k}$. The Gibbs sampler is obtained repeatedly sampling from the following conditional distributions:

- For any $i = 1, \ldots, I$ and $j = 1, \ldots, N_i$, independently update the cluster allocation variables from

$$p(s_{ij} = h \mid rest) \propto \text{alr}^{-1}(\widetilde{w}_{ih}) \, k(y_{ij} \mid \tau_h) \quad h = 1, \ldots, H$$

- Independently update the atoms of the mixture from

$$\mathcal{L}(\tau_h \mid rest) \propto P_0(\tau_h) \prod_{ij: s_{ij} = h} k(y_{ij} \mid \tau_h) \quad h = 1, \ldots, H$$

- Sample Σ from

$$\mathcal{L}(\Sigma \mid rest) \propto \mathcal{L}(\widetilde{\boldsymbol{w}} \mid rest) \mathcal{L}(\Sigma)$$

We show that the full conditional of Σ is still an inverse-Wishart distribution. To

see this, write the right handside as follows

$$\mathcal{L}(\Sigma \mid rest) \propto |(F - \rho G)^{-1} \otimes \Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{m}})^T \left((F - \rho G) \otimes \Sigma^{-1}\right)(\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{m}})\right) \times$$

$$|\Sigma|^{-(\nu + (H-1)+1)/2} \exp\left(-\frac{1}{2}tr(V\Sigma^{-1})\right)$$

Now $|(F - \rho G)^{-1} \otimes \Sigma| = |(F - \rho G)^{-1}|^{H-1} \times |\Sigma|^I$, so that the degrees of freedom in the full conditional are $\nu_p = \nu + I$. Working on the exponent, write the quadratic form involving the Kronecker product as follows

$$(\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{m}})^T \left((F - \rho G) \otimes \Sigma^{-1}\right)(\widetilde{\boldsymbol{w}} - \widetilde{\boldsymbol{m}}) = \sum_{i,j=1}^I (F - \rho G)_{ij}(\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{m}}_i)^T \Sigma^{-1}(\widetilde{\boldsymbol{w}}_j - \widetilde{\boldsymbol{m}}_j)$$

By exploiting multiple times the linearity of the trace operator and its cyclic property, the scale matrix V_p can be seen to equal

$$V_p = \sum_{i,j=1}^I (F - \rho G)_{ij}(\widetilde{\boldsymbol{w}}_j - \widetilde{\boldsymbol{m}}_j)(\widetilde{\boldsymbol{w}}_i - \widetilde{\boldsymbol{m}}_i)^T + V$$

and we conclude that $\Sigma \mid rest \sim \text{Inv-Wishart}(\nu_p, V_p)$

- Sample ρ from its full conditional:

$$\mathcal{L}(\rho \mid rest) \propto \pi(\rho)\mathcal{N}(vec(\widetilde{w}_1, \ldots, \widetilde{w}_I) \mid \boldsymbol{0}, (F - \rho G)^{-1} \otimes \Sigma)$$

This distribution does not have a closed form analytic expression because the support of ρ is $(0, 1)$ and hence we resort to a Metropolis Hastings step. The proposal distribution is a truncated normal (with support on $(0, 1)$) centered in the current value of ρ with standard deviation 0.1. Sampling from the truncated normal is performed by rejection sampling, whereas the computation of the acceptance rate for the Metropolis Hastings step is obtained by exploiting the law of the matrix normal distribution, which does not require to factorize the matrix $(F - \rho G)^{-1} \otimes \Sigma$. To improve the mixing of the chain, we resort to an Adaptive Metropolis Hastings move as in Roberts and Rosenthal (2009) to automatically tune variance of the normal proposal distribution.

- For each $i = 1, \ldots, I$ and each $h = 1, \ldots H$, independently sample $\widetilde{w}_{ih}$ as follows:

  - Sample the latent variable ω_{ih} from

  $$\mathcal{L}(\omega_{ih} \mid \widetilde{\boldsymbol{w}}_i) = PG(N_i, \eta_{ih}) = PG\left(N_i, \widetilde{w}_{ih} - \log\sum_{k \neq h} e^{\widetilde{w}_{ik}}\right)$$

  - Sample the transformed weight $\widetilde{w}_{ih}$ from

  $$\mathcal{L}(\widetilde{w}_{ih} \mid \widetilde{\boldsymbol{w}}_{-i}, \widetilde{\boldsymbol{w}}_{i,-h}, \boldsymbol{s}_i, \rho, \Sigma, \omega_{ih}) = N(\hat{\mu}_{ih}, \hat{\Sigma}_{ih}).$$

- for each connected component m of the graph we sample from

$$\mathcal{L}(\widetilde{\boldsymbol{m}}_{C_m} \mid rest) = \mathcal{N}(\boldsymbol{m}_{C_m}, \Lambda_{C_m})$$

For ease of notation, we show how to obtain expression of $\boldsymbol{m}_{C_m}$ and Λ_{C_m} in the case where is only one connected component in the graph. However the general update can be straightforwardly recovered since $\widetilde{m}_{C_1}, \ldots \widetilde{\boldsymbol{m}}_{C_k}$ corresponding to connected components in the graph are conditionally independent a priori. In case of one single connected component in the graph, we rewrite (5), letting all the $\widetilde{\boldsymbol{m}}_i$s to be equal to $\widetilde{\boldsymbol{m}}_1$, as

$$\widetilde{\boldsymbol{w}} \sim \mathcal{N}_{I(H-1)} \left(\mathbf{1}_I \otimes \mathbb{I}_{H-1} \widetilde{\boldsymbol{m}}_1, \left((F - \rho G) \otimes \Sigma^{-1}\right)^{-1} \right)$$

where $\mathbf{1}_I$ is the vector of ones of length I and $\mathbb{I}_{H-1}$ is the $(H-1) \times (H-1)$ identity matrix. Then if $\Lambda := diag(\sigma^2, \ldots, \sigma^2)$ and writing $\mathbb{I}^* = \mathbf{1}_I \otimes I_{H-1}$, $Q = (F - \rho G) \otimes \Sigma^{-1}$, we can write the full conditional of $\widetilde{\boldsymbol{m}}_1$ as follows:

$$\mathcal{L}(\widetilde{\boldsymbol{m}}_1 \mid rest) \propto \exp\left(-0.5\left(\widetilde{\boldsymbol{w}} - \mathbb{I}^*\widetilde{\boldsymbol{m}}_1\right)^T Q\left(\widetilde{\boldsymbol{w}} - \mathbb{I}^*\widetilde{\boldsymbol{m}}_1\right)^T + \widetilde{\boldsymbol{m}}_1^T \Lambda^{-1}\widetilde{\boldsymbol{m}}_1 \right) \right)$$

$$\propto \exp\left(-0.5\widetilde{\boldsymbol{m}}_1^T \left(\mathbb{I}^{*T}Q\mathbb{I}^*\right)\widetilde{\boldsymbol{m}} + \widetilde{\boldsymbol{m}}_1^T\Lambda^{-1}\widetilde{\boldsymbol{m}}_1 + -2\widetilde{\boldsymbol{m}}_1^T\left(\mathbb{I}^{*T}Q\widetilde{\boldsymbol{w}}\right) \right)$$

This is the kernel of a multivariate normal distribution with covariance matrix $\Lambda_C = \left(\mathbb{I}^{*T}Q\mathbb{I}^* + \Lambda^{-1}\right)^{-1}$ and mean $\boldsymbol{m}_C = \Lambda_C\left(\mathbb{I}^{*T}Q\widetilde{\boldsymbol{w}}\right)$.

- If there are covariates in the model as in Section 7, the full-conditional of the regression coefficients $\boldsymbol{\beta}$ is given by

$$\mathcal{L}(\boldsymbol{\beta} \mid rest) = \mathcal{N}_d\left((\Sigma^{-1} + X^TVX)^{-1}(X^TV(\boldsymbol{y} - \boldsymbol{\mu})), (\Sigma^{-1} + X^TVX)^{-1}\right)$$

where V is an $N \times N$ diagonal matrix and $N = \sum N_i$. Denoting by $\boldsymbol{c} = (c_1, \ldots, c_N)$, the vectorization of the sequence of latent vectors $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_I$ in (12)-(13), then one has $V_{k,k} = \sigma_{c_k}^2$. The formula above can be derived by standard posterior updates in the Bayesian linear regression, when the mixture model (7)-(11) is the model for the "regression error".

D   Additional plots and tables

- Table 4 shows the Hellinger distance between the true density and the estimate under the three models under comparison in Section 6.1 for the three simulated scenarios in Table 1 for 100 repeatedly simulated datasets. We average these values over the simulated datasets, also considering $\pm$ one empirical standard deviation of the 100 values obtained.

- Figure 12 shows errors, measured with the Hellinger distance, under our model (spmix) and the HDP-mixture model (hdp) for each simulation, averaged over the areas, for $I = 4, 64, 256$, in Section 6.2.

- Figure 13 displays empirical correlations among the predictors and, in the last col-

umn, between predictors and the response for the Airbnb Amsterdam dataset in Section 7.

- Figure 14 shows the scatterplots of the response price versus numerical predictors and boxplots for categorical predictors for the Airbnb Amsterdam dataset in Section 7.

- Figure 15 shows the predictive densities in area Bijlmer-Centrum, corresponding to different covariate specifications: all the covariates are fixed to their empirical median except for `reviews_scores_rating`, which assumes values equal to the empirical quartiles $q_{0.05}, q_{0.5}, q_{0.95}$. It is clear that the three densities overlap almost perfectly. There are two reason for this. First, the the empirical distribution of this covariate is it is highly concentrated around high values, as people tend to give mostly positive reviews. Second, the coefficient associated to this covariate, despite significant, has a very small absolute value.

	Model	1	2	3	4	5	6
Scenario I	SPMIX	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	0.09 ± 0.01	0.09 ± 0.01
	HDP	0.03 ± 0.01	0.03 ± 0.01	0.06 ± 0.01	0.06 ± 0.01	0.09 ± 0.01	0.09 ± 0.01
	CK-SSM	0.44 ± 0.06	0.44 ± 0.06	0.53 ± 0.03	0.53 ± 0.03	0.44 ± 0.03	0.44 ± 0.03
Scenario II	SPMIX	0.08 ± 0.01	0.11 ± 0.02	0.07 ± 0.01	0.08 ± 0.03	0.11 ± 0.00	0.11 ± 0.03
	HDP	0.04 ± 0.01	0.19 ± 0.02	0.09 ± 0.01	0.24 ± 0.03	0.10 ± 0.00	0.27 ± 0.03
	CK-SSM	0.44 ± 0.06	0.43 ± 0.06	0.53 ± 0.03	0.53 ± 0.03	0.45 ± 0.05	0.45 ± 0.05
Scenario III	SPMIX	0.20 ± 0.07	0.20 ± 0.07	0.16 ± 0.06	0.16 ± 0.06	0.11 ± 0.05	0.11 ± 0.05
	HDP	0.12 ± 0.07	0.12 ± 0.07	0.21 ± 0.06	0.21 ± 0.06	0.13 ± 0.05	0.13 ± 0.05
	CK-SSM	0.42 ± 0.06	0.42 ± 0.06	0.59 ± 0.03	0.59 ± 0.03	0.38 ± 0.07	0.38 ± 0.07

Table 4: Hellinger distances between the true densities and the estimated ones, aggregated over 100 simulated datasets with $\pm$ one standard deviation for the simulated data in Section 6.1
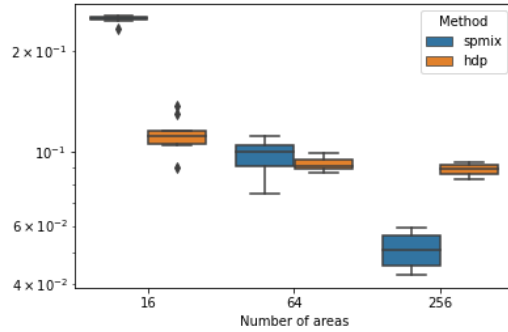

Figure 12: Boxplots of the Hellinger distance between true density (22) and estimated one under our model (spmix) and the HDP-mixture model (hdp) for each simulation, averaged over the areas, for $I = 16, 64, 256$, , in logarithmic scale, in Section 6.2.
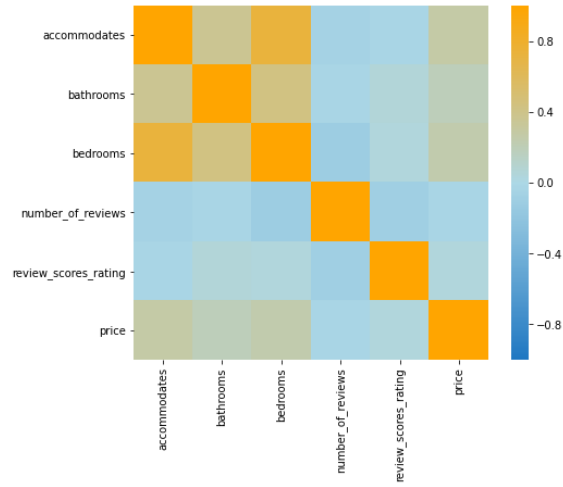
Figure 13: Correlation matrix between numerical predictors and response for Airbnb Amsterdam
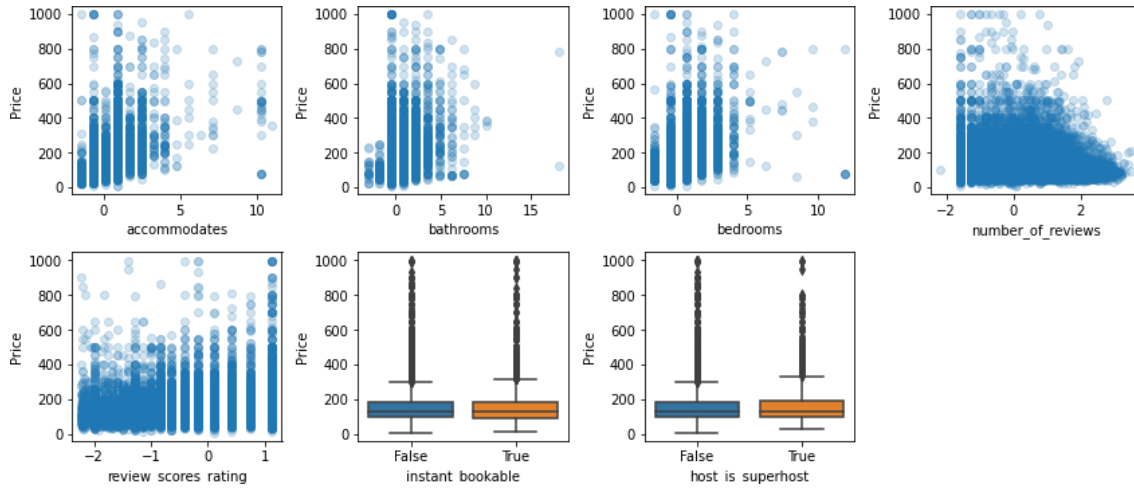

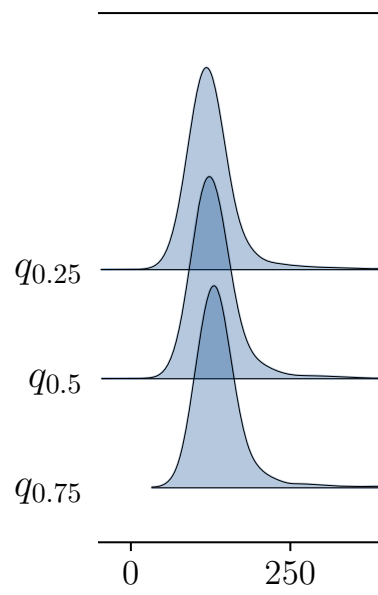Figure 14: Scatterplots and boxplots of the nightly price versus predictors for Airbnb Amsterdam

Figure 15: Predictive density for a new listing in *Bijlmer-Centrum* with all numerical covariates fixed to the empirical median of the dataset except `reviews_scores_rating` that ranges in the values $q_{0.25}, q_{0.5}, q_{0.75}$, where q_α denotes the empirical quantile of order α. Each line corresponds to one of these values, from top to bottom.

References

Argiento, R. and De Iorio, M. (2019). "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models." *arXiv preprint arXiv:1904.09733*.

Atchison, J. and Shen, S. M. (1980). "Logistic-normal distributions: Some properties and uses." *Biometrika*, 67(2), 261–272.

Banerjee, S. (2016). "Spatial data analysis." *Annual review of public health*, 37, 47–60.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press.

Besag, J. (1974). "Spatial interaction and the statistical analysis of lattice systems." *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 192–225.

Box, G. E. and Cox, D. R. (1964). "An analysis of transformations." *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–243.

Celeux, G., Frühwirth-Schnatter, S., and Robert, C. P. (2019). "Model selection for mixture models-perspectives and strategies." In *Handbook of mixture analysis*, 117–154. Chapman & Hall/CRC.

Clayton, D. and Kaldor, J. (1987). "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping." *Biometrics*, 671–681.

Cressie, N. (1992). "Statistics for spatial data." *Terra Nova*, 4(5), 613–617.

— (1993). *Statistics for spatial data*. Wiley.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). "Generalized spatial Dirichlet process models." *Biometrika*, 94(4), 809–825.

Dunson, D. B. and Park, J.-H. (2008). "Kernel stick-breaking processes." *Biometrika*, 95(2), 307–323.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Fruhwirth-Schnatter, S., Celeux, G., and Robert, C. P. (2019). *Handbook of mixture analysis*. Chapman and Hall/CRC.

Garcia-Ayllon, S. (2018). "Urban transformations as an indicator of unsustainability in the P2P mass tourism phenomenon: The Airbnb case in Spain through three case studies." *Sustainability*, 10(8), 2933.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association*, 100(471), 1021–1035.

Gelfand, A. E. and Vounatsou, P. (2003). "Proper multivariate conditional autoregressive models for spatial data analysis." *Biostatistics*, 4(1), 11–15.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82(4), 711–732.

Holmes, C. C. and Held, L. (2006). "Bayesian auxiliary variable models for binary and multinomial regression." *Bayesian analysis*, 1(1), 145–168.

Jo, S., Lee, J., Müller, P., Quintana, F. A., and Trippa, L. (2017). "Dependent species sampling models for spatial density estimation." *Bayesian Analysis*, 12(2), 379–406.

Kaiser, M. S. and Cressie, N. (2000). "The construction of multivariate distributions from Markov random fields." *Journal of Multivariate Analysis*, 73(2), 199–220.

Lee, D. and Mitchell, R. (2012). "Boundary detection in disease mapping studies." *Biostatistics*, 13(3), 415–426.

Leroux, B. G., Lei, X., and Breslow, N. (2000). "Estimation of disease rates in small areas: a new mixed model for spatial dependence." In *Statistical models in epidemiology, the environment, and clinical trials*, 179–191. Springer.

Li, P., Banerjee, S., Hanson, T. A., and McBean, A. M. (2015). "Bayesian models for detecting difference boundaries in areal data." *Statistica Sinica*, 25(1), 385.

Lijoi, A., Prünster, I., and Rigon, T. (2020). "The Pitman–Yor multinomial process for mixture modeling." *Biometrika*.

Lu, H., Reilly, C. S., Banerjee, S., and Carlin, B. P. (2007). "Bayesian areal wombling via adjacency modeling." *Environmental and ecological statistics*, 14(4), 433–452.

MacEachern, S. N. (2000). "Dependent dirichlet processes." *Unpublished manuscript, Department of Statistics, The Ohio State University*, 1–40.

Malsiner-Walli, G., Frühwirth-Schnatter, S., and Grün, B. (2016). "Model-based clustering based on sparse finite Gaussian mixtures." *Statistics and computing*, 26(1-2), 303–324.

Miller, J. W. and Harrison, M. T. (2018). "Mixture models with a prior on the number of components." *Journal of the American Statistical Association*, 113(521), 340–356.

Nobile, A. (1994). "Bayesian analysis of finite mixture distributions." Ph.D. thesis, PhD Thesis. Carnegie Mellon University, Pittsburgh.

Page, G. L. and Quintana, F. A. (2016). "Spatial product partition models." *Bayesian Analysis*, 11(1), 265–298.

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association*, 108(504), 1339–1349.

Pyrcz, M. J. and Deutsch, C. V. (2014). *Geostatistical reservoir modeling*. Oxford university press.

Ren, L., Du, L., Carin, L., and Dunson, D. (2011). "Logistic stick-breaking process." *Journal of Machine Learning Research*, 12(Jan), 203–239.

Richardson, S. and Green, P. J. (1997). "On Bayesian analysis of mixtures with an unknown number of components (with discussion)." *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4), 731–792.

Roberts, G. O. and Rosenthal, J. S. (2009). "Examples of adaptive MCMC." *Journal of Computational and Graphical Statistics*, 18(2), 349–367.

Rodriguez, A. and Dunson, D. B. (2011). "Nonparametric Bayesian models through probit stick-breaking processes." *Bayesian analysis*, 6(1), 145–178.

Stan Development Team (2018). "Stan Modeling Language Users Guide and Reference Manual."
URL http://mc-stan.org

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, 101(476), 1566–1581.

Van Der Zee, R. (2016). "The 'Airbnb effect': Is it real, and what is it doing to a city like Amsterdam." *The Guardian*, 6 October 2016.

Wachsmuth, D. and Weisler, A. (2018). "Airbnb and the rent gap: Gentrification through the sharing economy." *Environment and Planning A: Economy and Space*, 50(6), 1147–1170.

Webster, R. and Oliver, M. A. (2007). *Geostatistics for environmental scientists*. John Wiley & Sons.

Zhou, Z., Matteson, D. S., Woodard, D. B., Henderson, S. G., and Micheas, A. C. (2015). "A spatio-temporal point process model for ambulance demand." *Journal of the American Statistical Association*, 110(509), 6–15.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

55/2020   Botti, M.; Castanon Quiroz, D.; Di Pietro, D.A.; Harnist, A.
A Hybrid High-Order method for creeping flows of non-Newtonian fluids

56/2020   Botti, L.; Botti, M.; Di Pietro, D. A.;
A Hybrid High-Order method for multiple-network poroelasticity

57/2020   Regazzoni, F.; Quarteroni, A.
An oscillation-free fully partitioned scheme for the numerical modeling of cardiac active mechanics

54/2020   Arnone, E.; Bernardi, M. S.; Sangalli, L. M.; Secchi, P.
Analysis of Telecom Italia mobile phone data by space-time regression with differential regularization

53/2020   Arnone, E.; Kneip, A.; Nobile, F.; Sangalli, L. M.
Some numerical test on the convergence rates of regression with differential regularization

52/2020   Arnone, E.; Kneip, A.; Nobile, F.; Sangalli, L. M.
Some first results on the consistency of spatial regression with partial differential equation regularization

51/2020   Ferraccioli, F.; Sangalli, L. M.; Arnone, E.,; Finos, L.
A functional data analysis approach to the estimation of densities over complex regions

50/2020   Bonaventura,L.; Gomez Marmol, M.
The TR-BDF2 method for second order problems in structural mechanics

49/2020   Bonaventura,L.; Garres Diaz,J.
Flexible and efficient discretizations of multilayer models with variable density

44/2020   Masci, C.; Ieva, F.; Paganoni A.M.
EM algorithm for semiparametric multinomial mixed-effects models