Spatially dependent mixture models via the logistic multivariate CAR prior

Mario Beraha [a,b,*], Matteo Pegoraro [c], Riccardo Peli [c], Alessandra Guglielmi [a]

[a] *Department of Mathematics, Politecnico di Milano, Piazza L. Da Vinci 32, Milano, 20133, Italy*
[b] *Department of Computer Science, Università di Bologna, Mura A. Zamboni 7, Bologna, 40126, Italy*
[c] *MOX, Department of Mathematics, Politecnico di Milano, Piazza L. Da Vinci 32, Milano, 20133, Italy*

ARTICLE INFO

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.

© 2021 Elsevier B.V. All rights reserved.

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

* Corresponding author at: Department of Mathematics, Politecnico di Milano, Piazza L. Da Vinci 32, Milano, 20133, Italy.
E-mail addresses: mario.beraha@polimi.it (M. Beraha), matteo.pegoraro@polimi.it (M. Pegoraro), riccardo.peli@polimi.it (R. Peli), alessandra.guglielmi@polimi.it (A. Guglielmi).

developed to take into account spatial dependence. Spatial data are classified into three main categories, according to Cressie (1992): geostatistical data, for which an exact location is known for each observation, areal (or lattice) data, when each observation is associated to a specific area or node in a lattice, and point patterns, where the object of the inference is the event location. Examples of the first are environmental applications (see Webster and Oliver, 2007) 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. Point patterns are often employed in ecology, as described in Velázquez et al. (2016). See also the textbook by Banerjee et al. (2014) for data classification, applications and statistical models and techniques for spatially dependent data.

1.1. Setup

We focus on areal data, and, in particular, we consider the problem of modeling data from I different groups, where each group corresponds to a specific areal location. More in detail, 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 value y_{ij} corresponding to a different subject j 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 further assume that data, within each areal unit i, are independent and identically distributed (i.i.d.) from an area-specific density f_i; the problem we address is the joint estimation of spatially dependent densities $f_1, \ldots, f_I$. We take the Bayesian viewpoint 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. Relaxing the assumption of identically distributed observations within each area is straightforward in the regression context, i.e. when covariates for each subject are available.

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 17,000 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. Such a model can be of interest to a "new" lessor wishing to rent their house or flat on Airbnb. The area-specific estimate of the density might allow the lessor to understand the full market of renting apartments in his/her neighborhood, unlike a simple point estimate of the average price. The lessor might also understand if it is worth making home improvements in order to get a higher rent or assessing, for instance, the posterior predictive probability of the rent being above some threshold.

A peculiar feature of the municipality of Amsterdam is that three neighborhoods are not connected to the rest of the city but among themselves (see, for instance, Fig. 6), i.e. there are two different connected components in the adjacency graph of neighborhoods. It is likely that the nightly prices exhibit substantially different behavior when comparing one component to the other. Hence, we want to build a model that encourages sharing of information across neighboring areas, but does not force densities belonging to different components to be similar a priori.

Compared to more traditional spatial regression techniques such as eigenvector spatial filtering (see Griffith et al., 2019, for a review), geographically weighted regression (Brunsdon et al., 1998) or the models in the R package *CARBayes* (Lee, 2013), our approach does not make distributional assumptions (such as assuming Gaussian-distributed responses) and our focus here is on density modeling and estimation and density regression via mixture models.

1.2. Previous work on Bayesian spatial density modeling

To model our distributions we resort to the well established class of mixture models (Fruhwirth-Schnatter et al., 2019), that are a classical tool for density estimation. In the Bayesian nonparametric setting, since (MacEachern, 2000), a great effort has been dedicated to modeling 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). Jo et al. (2017) considered mixture models based on species sampling priors where the spatial dependence is introduced through a Gaussian multivariate conditional autoregressive (CAR, Besag, 1974) model on a suitable transformation of the weights. Despite their focus being on point-referenced data, their model can be easily extended to areal data, as we do in Sections 4.3 and 6 for a comparison with our approach. The idea of building spatial dependence in mixture models through a CAR distribution on latent variables is also shared by Li et al. (2015), where the authors propose an area-dependent Dirichlet process that can also formally identify boundaries between areas, and Zhou et al. (2015), that use the trick of normalization of CAR distributions to time-varying weights in a rather complex application with focus on estimation of ambulance demand.

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 mixture 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.

1.3. Our contribution and outline

In this paper, we consider a finite mixture model, where the number H of components is fixed. Finite mixtures are particularly suited for the problem of modeling 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 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).

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 logistic MCAR distribution, 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 and Section 6 presents results from two simulation studies with comparison with competitor models. The application to Airbnb Amsterdam is discussed in Section 7, where we propose two generalizations of our area-dependent mixture model to include subject-specific covariates and relaxing the identity in distribution assumption within each area. 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. Codes of our MCMC algorithm for simulated data and Airbnb Amsterdam application has been implemented C++ and Python and is available at https://github.com/mberaha/spatial_mixtures.

2. Preliminaries

2.1. Mixture models

For any areal unit $i = 1, \ldots, I$ and subject $j = 1, \ldots, N_i$, we assume observation $y_{ij} \in \mathbb{Y} \subset \mathbb{R}^p$. In this paper, we fix $p = 1$, 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

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

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 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) as in Malsiner-Walli et al. (2016).

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

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

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

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

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

$$\text{alr}(\boldsymbol{w}_1 \oplus \boldsymbol{w}_2) = \text{alr}(\boldsymbol{w}_1) + \text{alr}(\boldsymbol{w}_2), \quad \text{alr}(\alpha \odot \boldsymbol{w}_1) = \alpha \, \text{alr}(\boldsymbol{w}_1)$$

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

The alr transformation is often applied in the context of compositional data analysis, where statistical inference for data in the simplex has been pioneered by Aitchison (1986). In particular, this map was used in Aitchison and Shen (1980) to define a new distribution on the simplex, the logistic-normal distribution. Formally, we say that $\boldsymbol{w} = (w_1, \ldots, w_{H-1}, w_H := 1 - \sum_{h=1}^{H-1} w_h)^T \in S^H$

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

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

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

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

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

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

We generalize the univariate CAR model in Leroux et al. (2000) assuming the following multivariate conditionally autoregressive (MCAR) model:

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

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. When $H-1=1$, (4) gives the prior proposed in Leroux et al. (2000).

Denoting with $A \otimes B$ the Kronecker product between the $m \times n$ matrix A and the $p \times q$ matrix B, i.e., $A \otimes B$ is the $pm \times qn$ matrix with entries $(A \otimes B)_{pr+v, qs+w} = A_{r,s} B_{v,w}$, the following proposition guarantees that the joint distribution of $\widetilde{\boldsymbol{w}} = vec(\widetilde{\boldsymbol{w}}_1, \dots, \widetilde{\boldsymbol{w}}_I)$, the vectorization of the weights, is well defined and unique.

Proposition 1. *Assume that $\rho \in (-1, 1)$ and $\widetilde{\boldsymbol{m}}_i = \widetilde{\boldsymbol{m}}_j$ if areas i and j belong to the same connected component of the graph G. Then the set of full conditionals in* (4) *defines a unique joint probability*

distribution for $\widetilde{\boldsymbol{w}} = vec(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$, given by

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

where $\widetilde{\boldsymbol{m}} = vec(\widetilde{\boldsymbol{m}}_1, \ldots, \widetilde{\boldsymbol{m}}_I)$ and $F = diag(\rho \sum_j g_{1j} + 1 - \rho, \ldots, \rho \sum_j g_{Ij} + 1 - \rho)$.

Proof. See Appendix A.

The matrix $A^{-1}(G, \rho) := (F - \rho G) = \rho \left(diag(G\boldsymbol{1}_I) - G \right) + (1 - \rho)\mathbb{I}_I$ in (5), where $\boldsymbol{1}_I \in \mathbb{R}^I$ denotes the vector of ones and $\mathbb{I}_I$ denotes the $I \times I$ identity matrix, has 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 $(\widetilde{\boldsymbol{w}}_1, \ldots, \widetilde{\boldsymbol{w}}_I)$, $\widetilde{\boldsymbol{w}}_i = alr(\boldsymbol{w}_i)$, follow the MCAR model in (4) (or (5)). We write $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim$ 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 alr-transformation. This is made clear by the following proposition.

Proposition 2. If $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim$ logisticMCAR($\widetilde{\boldsymbol{m}}, \rho, \Sigma; G$), then, for any $i = 1, \ldots, I$,

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

where $U_i = \{j : g_{ij} > 0\}$, $|U_i| = \sum_j g_{ij}$ and $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{iH})$, with $\boldsymbol{m}_i = alr^{-1}(\widetilde{\boldsymbol{m}}_i)$.

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 hand side 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 hand side 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 2 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 $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim$ 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 3. If $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_I) \sim$ 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, if areas i and j belong to different connected graph components of the graph G, $\text{Cov}\left(\log \frac{w_{il}}{w_{im}}, \log \frac{w_{jl}}{w_{im}} \right) = 0$ and $\widetilde{\boldsymbol{w}}_i, \widetilde{\boldsymbol{w}}_j$ are independent, conditioning to parameters $\widetilde{\boldsymbol{m}}, \rho, \Sigma$.

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 logisticMCAR distribution shares the same limitation as the logistic-normal one, i.e. moments are not available in closed-form expressions. 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 neighbors 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 notation $\overset{\text{iid}}{\sim}$ denotes independent and identically distributed random variables, $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{iH})^T \in S^{H-1}$ (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_h and variance σ_h^2, so that $\tau_h = (\mu_h, \sigma_h^2)$ and P_0 is a probability distribution over $\Theta = \mathbb{R} \times \mathbb{R}^+$. Specific choices of P_0 are discussed in Sections 6 and 7. 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 neighbors (see, for instance, Gelfand and Vounatsou, 2003, Section 4).

Model (7)–(11) assumes that each group of data $\boldsymbol{y}_i$ is modeled 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.

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

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

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

where notation $\overset{\text{ind}}{\sim}$ is used to represent independent random variables (from different distributions). 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 $\widetilde{m}$ 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 $\widetilde{m}$

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 very 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 be allocated. We think that 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 $\widetilde{m}_i$'s in our model. We start by observing that in the mixture model (7) for the ith area, if coordinate values in the vector $\widetilde{m}_i$ in (9) are very different among each other, this 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 $\widetilde{m}_i \sim \mathcal{N}_{H-1}(\mathbf{0}, \eta^2 \mathbb{I})$. Observe that, since the distribution of $\widetilde{m}_i$ is centered in $\mathbf{0}$ and isotropic, we are not forcing, marginally, any specific ordering on the weights.

To understand the role of η^2, let us consider an illustrative example when $H = 3$ and $I = 1$. Let $\mathbf{m}_1 = \text{alr}^{-1}\widetilde{m}_1$ and consider $d_{12} = (\log(m_{11}/m_{1H}) - \log(m_{12}/m_{1H}))^2$, which corresponds to the distance between m_{11} and m_{12} in the Aitchison geometry. We may consider d_{12} as a plug-in estimator of the distance between w_{11} and w_{12}. The largest values of d_{12} are obtained when one between m_{11}, m_{12} and m_{1H} is approximately 1 and the others are close to zero. Moreover, from $\widetilde{m}_1 \sim \mathcal{N}_2(\mathbf{0}, \eta^2 \mathbb{I})$, we have that $d_{12}/(2\eta^2)$ has chi-squared distribution with one degree of freedom. Hence the random variable d_{12} is stochastically increasing with η^2.

This feature holds also for larger values of H as shown in Fig. 2, where the behavior of w_i, for different values of η^2, is illustrated. We conclude that η^2 is a sparsity tuning parameter and sparsity of the w_i's is obtained for larger values of η^2. Note that this is the opposite behavior of other sparsity priors, such as the double exponential or the horseshoe (Bhadra et al., 2019), where a distribution with significant mass near zero is assumed. Because our parameters are transformed through the logistic map (3), assuming a prior concentrated in zero for $\widetilde{m}$ would result in a prior concentrated on $(1/H, \ldots, 1/H)$ for w, of course this being far from sparsity.

Moreover, observe that Proposition 1 requires

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

where $C_1, \ldots, C_k$ denote the connected components of graph G, i.e. all the parameters $\widetilde{m}_i$s are assumed common within each connected component. This condition is obviously met if all the $\widetilde{m}_i$'s are equal. However, this seems overly restrictive, since we would loose the property that two connected graph components in $(\widetilde{w}_1, \ldots, \widetilde{w}_l)$ are independent under CAR distributions when marginalizing out the only shared parameter $\widetilde{m}_1$. Hence, we propose to extend the logisticMCAR$(\widetilde{m}, \rho, \Sigma; G)$ in (9) assuming

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

4.3. Comparison with competitor models

As mentioned in the Introduction, we have defined 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 this section and we further compare their features by fitting simulated data to the two models in Section 6.1.

We briefly introduce the class of spatially dependent species sampling mixtures in Jo et al. (2017), who define the weights in the mixtures to be spatially dependent, modeling them from a Gaussian CAR distribution, as we do. Their focus is on geo-referenced data (with multiple observations in each geographic location), and they propose two different CAR specifications, namely the Mercer CAR and the Clayton–Kaldor CAR (Clayton and Kaldor, 1987) priors. Since it is not straightforward to extend the Mercer CAR formulation to areal data as it requires the computation of a geographical distance rather than defining a proximity matrix, we only consider the Clayton–Kaldor CAR species sampling model in Jo et al. (2017) for comparison. 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. Below, we discuss how sparsity is obtained also in the spatially dependent species sampling model in Jo et al. (2017), but only in some sort of "asymmetric" manner, and how this impacts the modeling of different connected components in the graph G.

In the following, we refer to the prior in Jo et al. (2017) as *CK-SSM*. Instead of jointly modeling the transformed weights in each location, Jo et al. (2017) assume independent univariate CAR model for (a transformation of) the weights associated to each component of the mixture in the different areas. Recall that they assume a mixture with infinite components, i.e., $h = 1, 2, \ldots$. With our notation, let $\boldsymbol{v}_h = (w_{1h}, \ldots, w_{Ih})$, then the CK-SSM prior for $\boldsymbol{v}_h$ is

$$\widetilde{\boldsymbol{v}}_h \overset{\text{ind}}{\sim} \mathcal{N}_I(\widetilde{\theta}_h, \tau^2(I - \rho G)) \ \ h = 1, 2 \ldots, \quad v_{ih} = w_{ih} = \frac{e^{\widetilde{v}_{hi}}}{\sum_j e^{\widetilde{v}_{ji}}} \ \ i = 1, \ldots, I \tag{16}$$

In order to guarantee that the denominator of the fraction in (16) is finite, Jo et al. (2017) assume that $\widetilde{\theta}_h$ is a vector with all components equal to $\log\{1 - (1 + e^{b - ah})^{-1}\}$, a and b being positive hyperparameters. In force of that, the weights $\widetilde{v}_{hi}$ are stochastically decreasing with h for each area i. This ordering is preserved by the exponential and normalization transformations, so that w_{ih} will be stochastically decreasing with h as well, for each i. Note also that (16) makes $\boldsymbol{w}$ non-identifiable.

We start by considering $I = 1$ area and drop the subscript i. As in Jo et al. (2017), we truncate (16) to the first H terms for computation. Fig. 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 $\theta_h = \log\{1 - (1 + e^{1-h})^{-1}\}$, $h = 1, \ldots, H$, 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. Fig. 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 CK-SSM prior 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 greater 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 (Fig. 2(a)) and (16) in Jo et al. (2017) (Fig. 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 the two models may induce different types of prior behaviors. However, when considering the right panels,
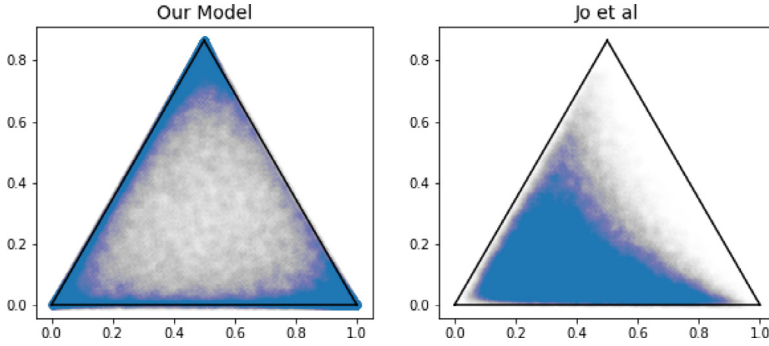
Fig. 1. Scatterplots of $N = 100,000$ MC draws (in blue) from the marginal priors of the weights in one single area with $H = 3$, under our logisticMCAR prior (left) and (16) in Jo et al. (2017) (right). White/gray areas represents low-density zones and dark blue zones high-density ones.
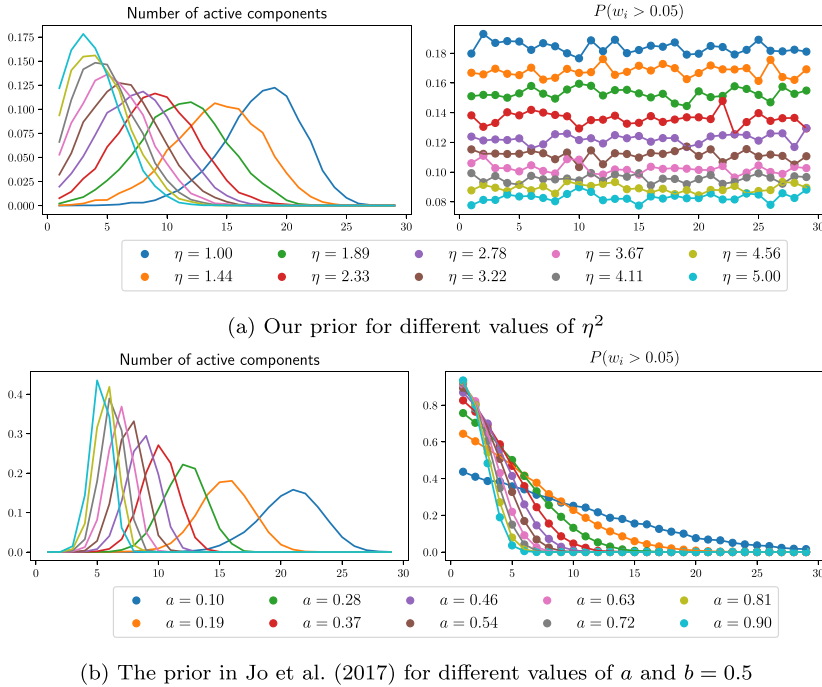

(a) Our prior for different values of η^2

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

Fig. 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 $H = 30$.

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 does not show a preferential ordering, this probability decreases with h under the CK-SSM prior. 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).

This is particularly relevant 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. Intuitively from the discussion above, CK-SSM 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). We tested this scenario more in detail by considering a spatial domain subdivided in four areas with two connected components {1, 2}, {3, 4}. Fig. D.12 shows the total variation distance for $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ and $(\boldsymbol{w}_1, \boldsymbol{w}_4)$ under the logisticMCAR and CK-SSM priors, having fixed hyperparameters as above and $\rho = 0.95$. It is clear that, as sparsity increases, the distance between $\boldsymbol{w}_1$ and $\boldsymbol{w}_4$ increases under the logisticMCAR but decreases under the CK-SSM, showing how imposing a sparse behavior in the CK-SSM prior forces similar distributions in disconnected components of the graph. See also Fig. D.13 for a visual representation of draws from the prior distributions. This effect becomes more and more evident as the sparsity in each mixture is increased, as shown in Fig. 2b. On the other hand, our model allows for the required level of sparsity in each mixture without forcing the different connected components in 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

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

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

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

where $\phi_{ih}^* \mid G \overset{iid}{\sim} G$ 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 ith 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 ith row (as a vector) so that $\Sigma_{i,-j}$ denotes the ith row where the jth element has been removed and $\Sigma_{-h,-k}$ denotes the $(H - 2) \times (H - 2)$ matrix there the hth row and kth 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 $\widetilde{\boldsymbol{w}}_i$. The full conditional of $\widetilde{\boldsymbol{m}}_{C_i}$ is a multivariate Gaussian distribution. See Appendix C for more detail on it, together with the other standard full conditionals.

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

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

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

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

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

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

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 $\mathcal{L}(\widetilde{w}_{ih} \mid \mathbf{s}_i, \widetilde{\mathbf{w}}_{i,-h})$ 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 $g_k \overset{iid}{\sim} Gamma(b, 1)$ and $b, c > 0$. The data-augmentation technique based on the Pólya-Gamma distribution relies on the following integral identity:

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

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 $\widetilde{w}_{ih}$:

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

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

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

See Appendix C for the proof of Eqs. (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.

Table 1

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

	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

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 run the Gibbs sampler for our model (7)–(11) together with the prior specification (14)–(15) (see Section 5 and Appendix C), and the *direct sampler* for the HDP-mixture model in Teh et al. (2006). Both algorithms were coded in C++. In addition, we have also implemented the CK-SSM 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 2000, 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 Scenario II there are three areas (2, 4 and 6) with few data points (only 10). In this case, we

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

|	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
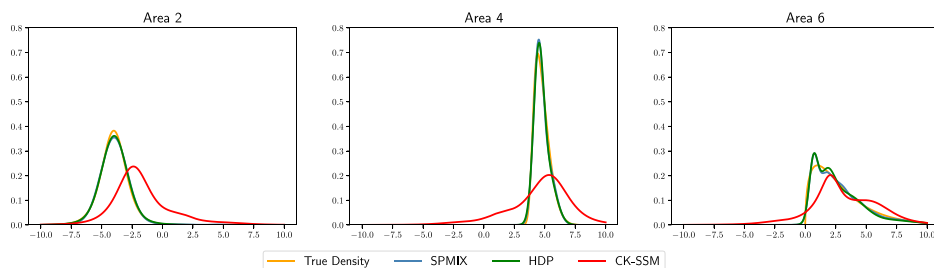
Fig. 3. Non-Gaussian simulated data, Scenario I: true densities for areas 2, 4 and 6 and the corresponding density estimates under the three different models.

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 D.6 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 Fig. 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 Fig. 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 Fig. 4. Indeed, our model retrieves information from the neighboring 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 Fig. 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.
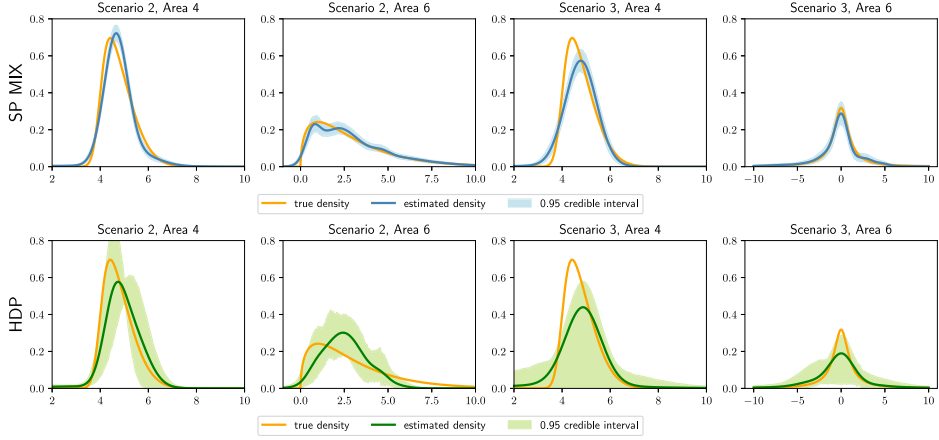
Fig. 4. Non-Gaussian simulated data, Scenario II and III: estimated and true densities for areas 4 and 6 under our (top row) and HDP mixture (bottom row) models.
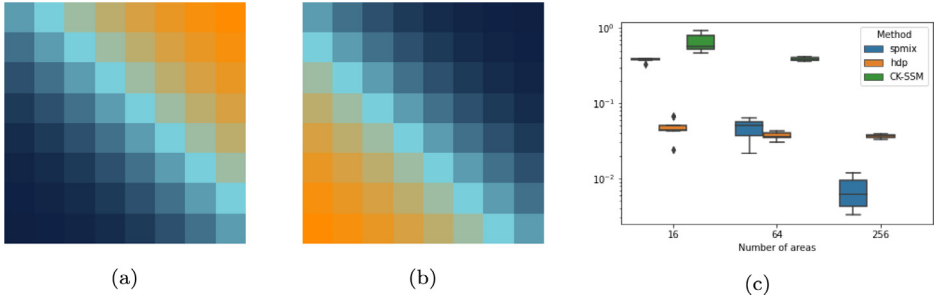

Fig. 5. Simulation from spatially dependent weights in Section 6.2: plots of $\{w_{i1}\}$ (a), $\{w_{i2}\}$ (b) when $I = 64$; boxplots (c) of the Kullback–Leibler divergence between true density (22) and estimated one under our model (spmix), the HDP-mixture model (hdp) and the CK-SSM for each simulation,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

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 ith area, we simulate observations as follows:

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

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 Figs. 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 $w_{i3} = 1 - w_{i1} - w_{i2}$.

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 and CK-SSM, reporting in Fig. 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 Fig. D.14 in Appendix D. On the other hand, CK-SSM shows the worst posterior estimates for $I = 16, 64$, whereas we do not present the case $I = 256$ for its high computational cost.

Concerning our model (spmix) and the HDP-mixtures, the median execution times, over the 10 datasets, of the code corresponding to spmix was 25.28, 118.14 and 616.41 s for $I = 16, 64, 256$, respectively, whereas, for fitting data for the HPD-mixtures was 18.39, 72.59, 207.46 s. 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 @2.4 GHz 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 (The 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 main goal is the prediction of the nightly price of a new listing, with information given by covariates, and taking into account the spatial dependence. As mentioned in the Introduction, (joint) density modeling and estimation, in this case, can give insight to landlords who need to take decision on where their flats should be positioned in the flat rental market. In fact, in this application, uncertainty quantification associated to the prediction, which can be easily derived from the full posterior density estimate, can lead to better informed decisions about the market strategy. We consider two generalizations of model (7) to account for covariates as follows. Denote responses as y_{ij} (i.e. the nightly price of accommodation j in neighborhood i) and covariates as $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijd})^T$. In the first model, denoted here $M1$, we assume τ_h in (7) such that $\tau_h = (\mu_h + \boldsymbol{\beta}^T \boldsymbol{x}_{ij}, \sigma_h^2)$, $h = 1, \ldots, H$. $M1$ can be understood as a linear regression model with component-specific intercept and variance. We further generalize $M1$ by assuming that all the regression coefficients are component-specific, i.e. $\tau_h = (\mu_h + \boldsymbol{\beta}_h^T \boldsymbol{x}_{ij}, \sigma_h^2)$, $h = 1, \ldots, H$, and denote it by $M2$. While model $M1$ assumes that the effect of the covariates on the pricing is shared across all neighborhoods, and the spatial effect can be represented by the only intercept, model $M2$ assumes that all the covariates have different effect on the pricing depending on the neighborhood.

7.1. Data description

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, and (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
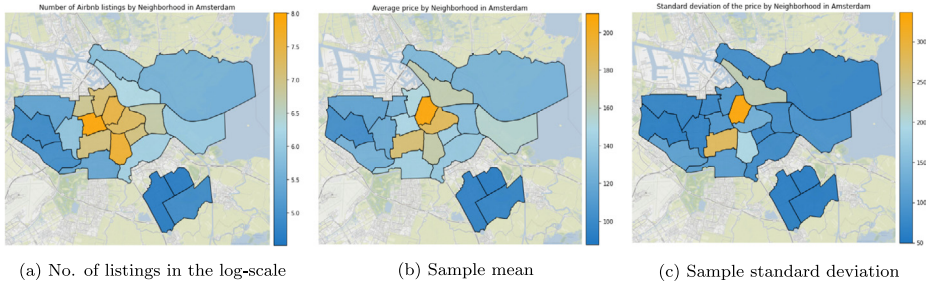
(a) No. of listings in the log-scale   (b) Sample mean   (c) Sample standard deviation

Fig. 6. Number of listings (in the log-scale), sample means and standard deviations of the nightly price in euros for each neighborhood in the Airbnb dataset, after preprocessing.

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 $\lambda = 12$, being this value automatically chosen by the Python package `scipy`.

Each listing is assigned to one of the twenty-two Amsterdam neighborhoods provided at the dataset web page, so that $I = 22$. The total number of observations considered for our analysis is $N_1 + \cdots + N_I = 17{,}201$. Fig. 6(a) shows sample sizes in the log-scale for each neighborhood; of course, city center is the area with the largest number of observations. Furthermore, in Fig. 6, panels (b) and (c), we report sample means and standard deviations of the nightly price in euros 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. Fig. 6 shows that there are two distinct graph connected components, one made only by three areal units; this agrees with official neighborhood maps of the city of Amsterdam. As far as covariates are concerned, Fig. D.15 in Appendix D shows empirical correlations among the predictors and between predictors and the response. Fig. D.16 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, while sample linear correlation between `accommodates` and `bathrooms` is 0.362, 0.730 between `accommodates` and `bedrooms`, 0.430 between `bathrooms` and `bedrooms`. However, when computing the variance inflation factor, we found 2.197, 1.243 and 2.334, respectively, values that suggest very mild multicollinearity. On the other hand, there is no evident empirical effect of `instant_bookable` and `host_is_superhost` on the nightly price as Fig. D.16 in Appendix D shows. In the next subsection, we consider both models $M1$ and $M2$ for the dataset, including all the covariates above described, i.e. $d = 7$. We standardized all numerical predictors, subtracting the sample mean and dividing by the sample standard deviation of each predictor; we also centered the response on the overall sample mean.

7.2. Posterior inference

We complete the prior for model $M1$ assuming

$$(\mu_h, \sigma_h^2) \overset{\text{iid}}{\sim} \mathcal{N}(\mu_h \mid 0, 2\sigma_h^2) \times IG(\sigma_h^2 \mid 2, 2), \quad h = 1, \dots, H$$

and $\boldsymbol{\beta} \sim \mathcal{N}_d(\mathbf{0}, \sigma_\beta^2 \mathbb{I}_d)$. For the prior of model $M2$ we assume

$$((\mu_h, \boldsymbol{\beta}_h), \sigma_h^2) \overset{\text{iid}}{\sim} \mathcal{N}((\mu_h, \boldsymbol{\beta}_h) \mid \mathbf{0}, 10\mathbb{I}_{d+1}) \times IG(\sigma_h^2 \mid 2, 2), \quad h = 1, \dots, H.$$

We need to change the Gibbs sampler in Section 5, adding two further steps to sample from the full conditional of $\boldsymbol{\beta}$ for model $M1$ or from the full conditional of (μ_h, β_h) for model $M2$. Both steps are standard updates in Bayesian linear regression models; see Appendix C for further details.

Table 3

LPML and WAIC for various choices of H under $M1$, $M2$ and CK-SSM.

H	$M1$			$M2$			CK-SSM		
	5	10	15	5	10	15	5	10	15
LPML	−92 619	−92 444	−92 441	−97 998	−97 836	−97 828	−98 751	−98 752	−98 755
WAIC	185 238	184 888	184 882	195 996	195 672	195 656	197 502	197 504	197 505

Posterior inference is robust to the choice of all the hyper-parameters in the prior distribution, but for the number H of components in the mixture, that is a key parameter for mixture models. For this reason, we compare $M1$ and $M2$ via predictive goodness-of-fit indexes such as the log-pseudo marginal likelihood (LPML, Geisser and Eddy, 1979) and the widely applicable information criterion (WAIC, Watanabe, 2013), when H varies in {5, 10, 15}. Better predictive performances are associated to higher LPML and lower WAIC. In this comparison, we also consider a generalization of the CK-SSM model in Jo et al. (2017) along the lines of model $M1$. Table 3 shows that the best model is $M1$, across all values of H and that CK-SSM does a worse job than $M1$ and $M2$. Given its superior predictive performance, in the following we consider only $M1$.

In particular, $M1$ with $H = 15$ gives slightly better values of LPML and WAIC, but the difference across all values of H seems negligible so that, to fix H, we consider also the predictive mean squared error computed 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 select 90% of the dataset as "training set" (to simulate from the relative posterior) and compute the mean of the predictive distribution corresponding to data in the "testing set". Observe that the same datapoints are shared across all values of H, both for training and for testing. Then we compute the predictive mean squared error (pMSE) on the testing set, i.e. $\sum_{i=1}^{m}(y_i - \hat{y}_i)^2/m$, where m is the size of the testing set and $\hat{y}_i$ is the mean of the posterior predictive density of the response corresponding to covariate $\boldsymbol{x}_i$. The average cross validation error $\pm$ one standard deviation is equal to 5468 ± 952, 5474 ± 850 and 5477 ± 956 for $H = 5, 10, 15$ respectively.

We have also considered the case $H = 1$, i.e., when all $M1$, $M2$ and CK-SSM models are equivalent to a standard Gaussian linear regression In this case, the predictive performance is much worse (LMPL and WAIC are approximately equal to -1.3×10^5 and 2.6×10^5 respectively), hence showing that a richer model with explicit modeling of the spatial dependence structure is needed to obtain better predictive performances.

Lastly, removing covariates `bathrooms` and `bedrooms`, correlated with `accommodates`, resulted in slightly worse predictive performance for all models tested; for instance, $M1$ showed a decrease in LMPL of 2.5%, while for $M2$ the decrease was around 1% across all values of H. Summing up, for the reasons above, including parsimony of the model, in the rest of the section, we consider only model $M1$ when $H = 5$.

Further comparisons are discussed in Appendix E. We first analyze this dataset by modeling connected components separately. Then we compare also with the fit given by geographically weighted regression. In both cases, the predictive performances are worse than the ones obtained with our model.

Fig. 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 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 Fig. 7(c) we show the density estimates in the four neighborhoods highlighted in Fig. 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 `accommodates`, `number_of_bedrooms`,
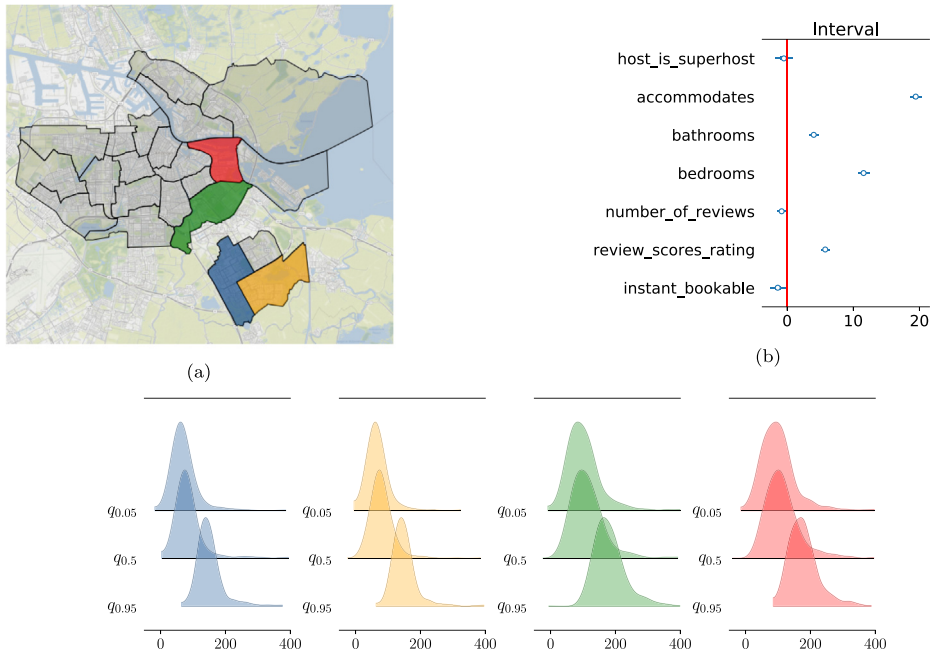
Fig. 7. (a) Map of the city of Amsterdam with neighborhoods *Bijlmer-Centrum* in blue, *Gaasperdam - Driemond* in orange, *Oostelijk Havengebied - Indische Buurt* in green and *Watergraafsmeer* in red. (b) 95% credibility intervals of the marginal posterior of the regression parameter $\boldsymbol{\beta}$. (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 accommodates, number_of_bedrooms, number_of_bathrooms), equal to their 5%, 50% and 95% sample quantiles, while the other numerical covariates are fixed to the empirical median. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

number_of_bathrooms. Since the marginal sample correlation between these three covariates is not negligible as mentioned in Section 7.1, we have fixed all their values simultaneously equal to 5%, 50% and 95% empirical quantiles, respectively. For instance, in each panel of Fig. 7(c), the top lines correspond to density estimates for a vector of covariates in which accommodates, number_of_bedrooms, number_of_bathrooms are fixed to their 5% sample quantile, respectively. It is clear from Fig. 7(c) that the predictive densities in blue (first panel from the left) 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 (second panel from the left) with the green ones (third panel from the left); indeed the green densities give substantial mass to the right tail, especially to values greater than 200 euros, 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. As expected, in all the neighborhoods the listings price increases as the accommodates, number_of_bedrooms, number_of_bathrooms increase as well.

To improve understanding of the posterior predictive densities evaluated on a grid of points as in Fig. 7(c), which are, by definition, the posterior means of the likelihood function, we focus on different summary statistics of these distributions. In particular, for any of the selected areas as in Fig. 7, we focus on the posterior predictive mean, median, standard deviation of the nightly price and on the posterior predictive probability $P(y^\star > 200 \mid \boldsymbol{x}^\star, i)$ that the nightly price exceeds 200 euros, conditioning on the covariates selected above. Numerical values for the summary statistics are

Table 4

Summary statistics (mean, median, standard deviation and probability of exceeding 200 euros) of the posterior predictive distributions for the same neighborhoods and covariate choices as in Fig. 7.

	Bijlmer centrum			Gaasperdam driemond			Oostelijk havengebied indische buurt			Watergraafsmeer		
	$q_{0.05}$	$q_{0.50}$	$q_{0.95}$	$q_{0.05}$	$q_{0.50}$	$q_{0.95}$	$q_{0.05}$	$q_{0.50}$	$q_{0.95}$	$q_{0.05}$	$q_{0.50}$	$q_{0.95}$
mean	71.5	84.23	149.5	69.38	82.18	147.8	99.49	112.1	176.5	100.0	112.6	177.0
median	66.53	79.35	145.9	64.92	77.75	145.1	92.18	105.0	172.33	92.98	105.8	173.1
sd	40.73	40.38	39.93	37.59	37.30	36.86	52.85	52.67	53.30	52.89	52.711	53.37
P_{200}	0.02	0.02	0.06	0.01	0.01	0.05	0.05	0.06	0.26	0.05	0.06	0.27

reported in Table 4, where for each of the four selected neighborhoods we consider three different values of the covariates $\boldsymbol{x}^\star$, denoted by $q_{0.05}$, $q_{0.5}$ and $q_{0.95}$ (see the discussion above and the caption of Fig. 7 for their definition). It is clear that for each of the neighborhoods, the mean and median of the price increase as `accommodates`, `number_of_bedrooms`, `number_of_bathrooms` increase as well. Interestingly, $P(y^\star > 200 \mid \boldsymbol{x}^\star, i)$ increases only slightly when the covariates go from $q_{0.05}$ to $q_{0.50}$ but increases significantly when covariates are $q_{0.95}$. As expected from Fig. 7(c), for each value of covariates, summary statistics in Bijlmer Centrum and Gaasperdam Driemond assume close values. A similar comment applies to summary statistics in Oostelijk Havengebied Indische Buurt and Watergraafsmeer. Instead, larger differences are observed in Table 4 when comparing Bijlmer Centrum and Watergraafsmeer (corresponding to the blue and red densities in Fig. 7(c)).

To conclude, we believe that a new lessor could benefit from our analysis because Airbnb makes available only a handful of covariates which might not be suited to fully characterize the "right" nightly price for a listing. For instance, we expect that the presence of a balcony or garden might lead to higher prices. Hence, when deciding the price for a listing, the lessor could look at the predictive distribution from our model given the covariates and neighborhood of their house, and choose to place the listing in the right or left tail of the predictive distribution considering additional information not included in our model.

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 behaviors. In particular, we argue that when different connected graph components are present, and heterogeneous behavior 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 models where the *graph G* is not fixed, and should be learned by the data (and the prior). In particular, we aim at considering 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.

Appendix A. Proofs

Proof of Proposition 1. The proof proceeds along the lines of Section 2 in Mardia (1988) where a similar MCAR model is proposed. Briefly, we need only to show that the conditional distributions of $\widetilde{w}_i \mid \widetilde{w}_{-i}$ derived from the joint distribution (5) coincide with the one given by (4), since by the Hammersley–Clifford theorem the set of full conditionals will uniquely identify the joint distribution in our setting; see also Theorem 2.1 in Mardia (1988). We start by assuming that only one connected component is present in G and that, without loss of generality, $\widetilde{m}_i = \widetilde{m}_j = \mathbf{0}$ for each $i, j = 1, \ldots, I$. The more general case, i.e., when $\widetilde{m}_i$ is not zero (but $\widetilde{m}_i = \widetilde{m}_j$ if i and j belong to the same connected component), follows from the same calculations below replacing $\widetilde{w}_i$ with $\widetilde{w}_i - \widetilde{m}_i$. For ease of notation we focus here on the full conditional of $\widetilde{w}_1$ but the other distributions can be derived in a similar manner.

From (5), the joint density of $\widetilde{w}$ is proportional to:

$$\exp\left(-\frac{1}{2}\widetilde{w}^T\left((F - \rho G) \otimes \Sigma^{-1}\right)\widetilde{w}\right).$$

We rewrite the quadratic form above to highlight only what depends on $\widetilde{w}_1$:

$$(\widetilde{w}_1 \cdots \widetilde{w}_I)\begin{pmatrix} (F - \rho G)_{11}\Sigma^{-1} & \cdots\cdots & (F - \rho G)_{1I}\Sigma^{-1} \\ \vdots & & \vdots \\ (F - \rho G)_{I1}\Sigma^{-1} & \cdots\cdots & (F - \rho G)_{II}\Sigma^{-1} \end{pmatrix}\begin{pmatrix} \widetilde{w}_1 \\ \vdots \\ \widetilde{w}_I \end{pmatrix}$$

where $(\widetilde{w}_1 \cdots \widetilde{w}_I)$ corresponds to the vectorization of the $\widetilde{w}_i$'s. Thanks to this block structure we obtain

$$\exp\left(-\frac{1}{2}\widetilde{w}^T\left((F - \rho G) \otimes \Sigma^{-1}\right)\widetilde{w}\right)$$
$$= \exp\left\{ -\frac{1}{2}\left(\widetilde{w}_1^T(F - \rho G)_{11}\Sigma^{-1}\widetilde{w}_1 + \right.\right.$$
$$+ 2\widetilde{w}_1^T[(F - \rho G)_{12} \cdots (F - \rho G)_{1I}] \otimes \Sigma^{-1}[\widetilde{w}_2 \cdots \widetilde{w}_I]^T$$
$$\left.\left. + [\widetilde{w}_2 \cdots \widetilde{w}_I]K[\widetilde{w}_2 \cdots \widetilde{w}_I] \right)\right\} \tag{A.1}$$

for some matrix K, which is irrelevant for our purposes since it does not interact with $\widetilde{w}_1$ so that the term depending on K can be discarded. Hence, we are able to identify in (A.1) the quadratic form of a Gaussian distribution and conclude that the full conditional of $\widetilde{w}_1$ is Gaussian. In particular, the first addend in the right hand-side of (A.1) is the only term that is quadratic in $\widetilde{w}_1$, so that it must be equal to $\widetilde{w}_1 S^{-1}\widetilde{w}$ where S is the covariance matrix of the full conditional of $\widetilde{w}_1$. Similarly, the second addend can be expressed as $-2\widetilde{w}_1 S^{-1}\mu$ where μ is the mean of the Gaussian distribution.

Some linear algebra yields

$$\text{Var}[\widetilde{\boldsymbol{w}}_1 \mid \widetilde{\boldsymbol{w}}_{-1}] = S = \left((F - \rho G)_{11} \Sigma^{-1}\right)^{-1} = \frac{\Sigma}{\rho \sum_{j=1}^{I} g_{ij} + 1 - \rho}$$

$$\mathbb{E}[\widetilde{\boldsymbol{w}}_1 \mid \widetilde{\boldsymbol{w}}_{-1}] = \boldsymbol{\mu} = \frac{\rho \sum_{j=1}^{I} g_{ij} \widetilde{\boldsymbol{w}}_j}{\rho \sum_{j=1}^{I} g_{ij} + 1 - \rho}.$$

When there are at least two connected components in the graph, we note that full independence holds across each pair of $\widetilde{\boldsymbol{w}}_i$'s as long as the two vectors belong to different connected components. Hence, the same argument above can be carried out on each single connected component. The joint density still takes the same form because the matrix $(F - \rho G) \otimes \Sigma^{-1}$ is block diagonal in this case with each block corresponding to one of the connected components. □

Proof of Proposition 2. From Eq. (4) we have that

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

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

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

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

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

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

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

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

which proves the proposition. □

Proof of Proposition 3. 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

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

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

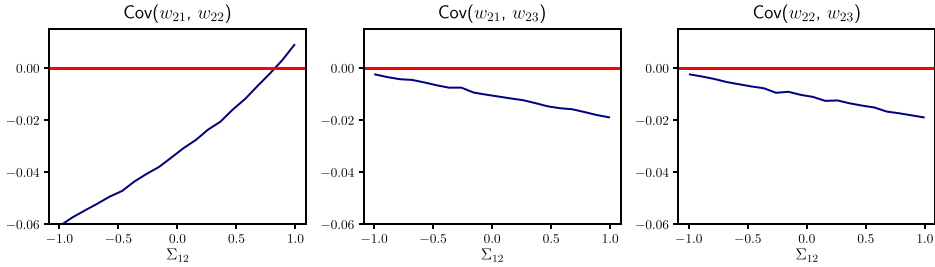$$= A_{ij}\left(\Sigma_{ll} - 2\Sigma_{lm} + \Sigma_{mm} \right)$$

Fig. B.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.

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 neighboring 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 thanks to equivalence between uncorrelation and independence of Gaussian random variables. $\quad\square$

Appendix 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.

Fig. B.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.

Fig. B.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 Fig. B.9.

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 particular, we simulated $N = 10,000$ draws from the full joint logisticMCAR distribution of $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_5)$ with parameters as above, fixing $\Sigma_{12} = 0.5$, and computed the Euclidean distances $d_{12} = \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$ and $d_{15} =$
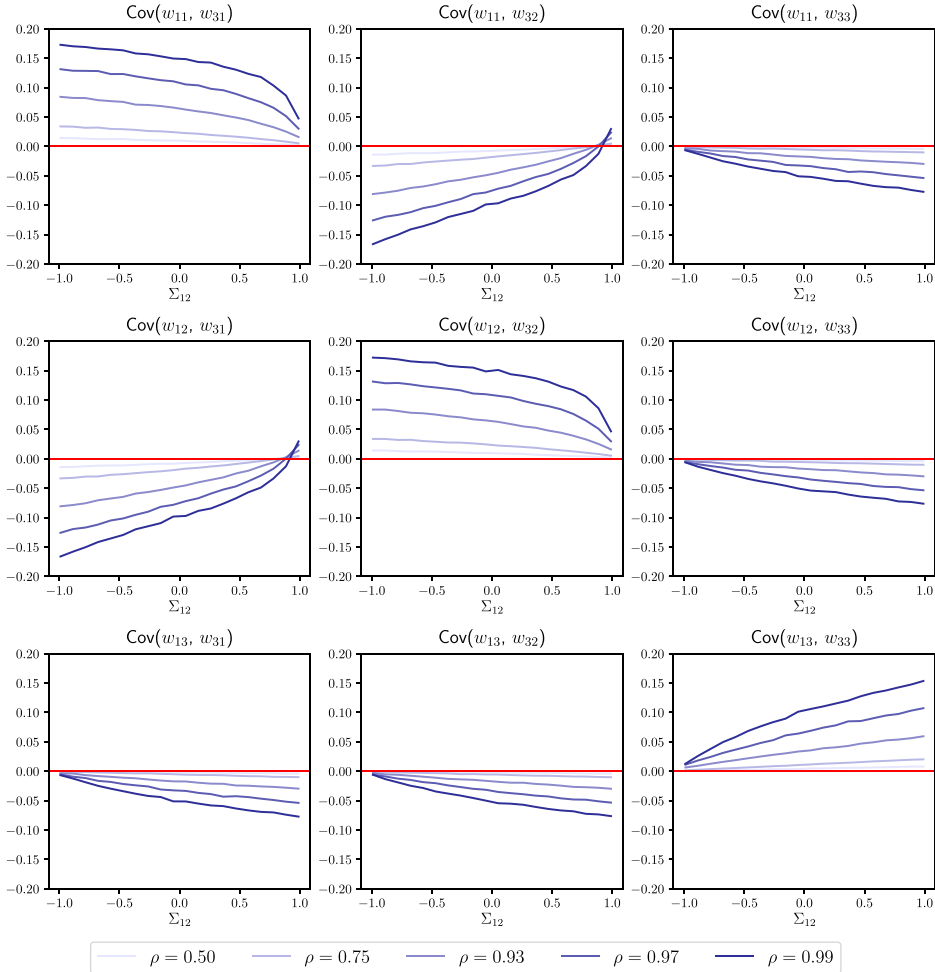
Fig. B.9. Pairwise covariance values between components of $\boldsymbol{w}_1$ and $\boldsymbol{w}_3$, as a function of the correlation parameter Σ_{12}, for different values of ρ, when $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_5)$ has the logisticMCAR distribution. The horizontal red line indicates the value 0.

$\|\boldsymbol{w}_1 - \boldsymbol{w}_5\|$. 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.

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

and computed their Euclidean distance as well, that we denote by d^γ. Fig. B.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 B.5.
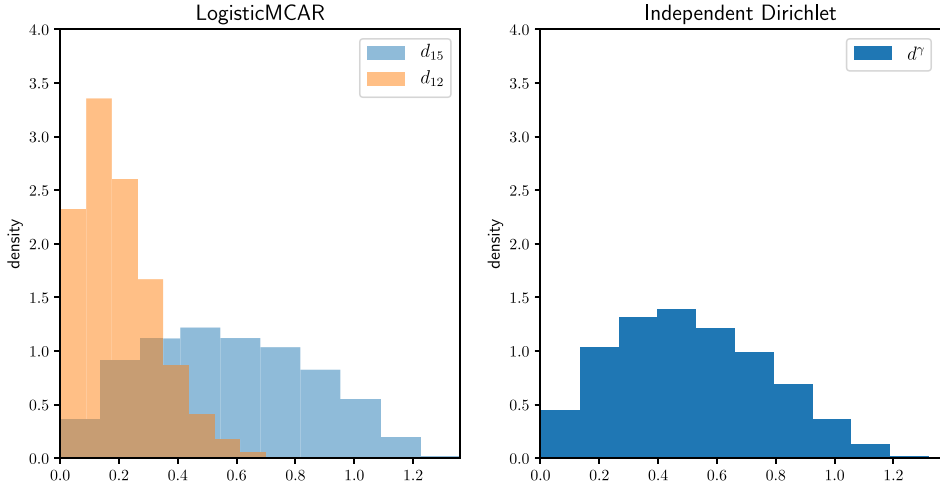
Fig. B.10. Histogram of MC draws from the marginal distributions of d_{12} (orange) and d_{15} (light blue) on the left and from the marginal distribution of d^γ on the right. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table B.5
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.

	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

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 Fig. B.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 $\mathcal{L}(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$.

Appendix C. The Gibbs sampler

Proof of Eqs. (20)–(21). We start by writing the full conditional distribution (18) as follows:

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

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

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

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

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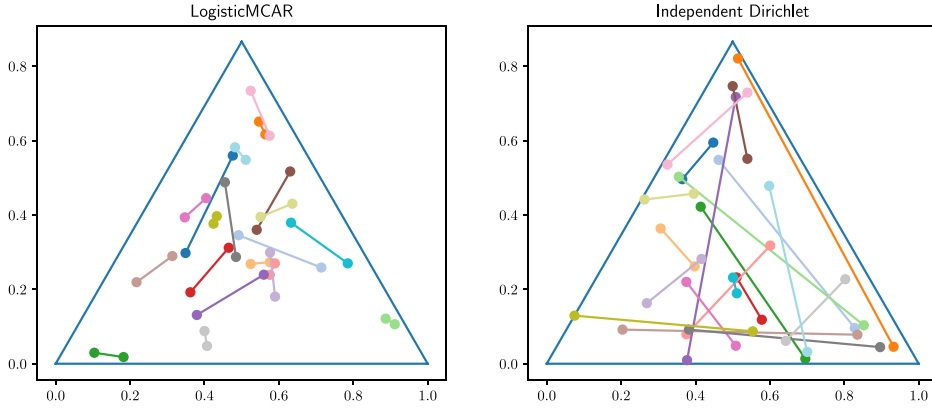
$$\propto e^{-\frac{E}{2}}$$

Fig. B.11. Plots of $N = 20$ MC draws from the logisticMCAR distribution (left) and of $N = 20$ MC draws from Dir($\mathbf{1}$)$\times$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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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

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

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 Eq. (19), with parameters $b = N_i$, $c = \widetilde{w}_{ih} - \log \sum_{k \neq h} \exp(\widetilde{w}_{ik})$. □

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 $\widetilde{\boldsymbol{m}}_{C_1}, \ldots \widetilde{\boldsymbol{m}}_{C_k}$. 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 hand side 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 $\Sigma \mid rest \sim$ Inv-Wishart(ν_p, V_p)
- Sample ρ from its full conditional:

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

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 straightwardly 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:

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

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

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 $M1$ as in Section 7, the full-conditional of the regression coefficients $\boldsymbol{\beta}$ is given by

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

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 $V_{k,k} = \sigma^2_{c_k}$. 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".

In case of model $M2$, as in Section 7, the full-conditional of each regression coefficients $\left((\mu_h, \boldsymbol{\beta}_h), \sigma^2_h \right)$ is straightforwardly computed from standard Bayesian linear regression, considering only observations that are allocated to component h. In particular, if we denote by $\boldsymbol{y}_h$ the vector of $\{y_{ij} : s_{ij} = h\}$, by X_h the matrix with rows $\{x_{ij} : s_{ij} = h\}$, and by n_h the size of $\boldsymbol{y}_h$, then we have

$$\sigma^2_h \mid rest \sim IG(a_{ph}, b_{ph})$$
$$(\mu_h, \boldsymbol{\beta}_h) \mid \sigma^2_h, rest \sim \mathcal{N}(\boldsymbol{\mu}_{ph}, \Delta_{ph})$$

where

$$\Delta_{ph} = X_h^T X_h + 10 \mathbb{I}_{d+1}$$
$$\boldsymbol{\mu}_{ph} = \Delta_{ph}^{-1} X_h^T \boldsymbol{y}_h$$
$$a_{ph} = 2 + n_h/2$$
$$b_{ph} = 2 + \frac{1}{2} (\boldsymbol{y}_h^T \boldsymbol{y}_h - \boldsymbol{\mu}_{ph}^T \Delta_{ph} \boldsymbol{\mu}_{ph})$$

Appendix D. Additional plots and tables

- Fig. D.12 shows the total variation distance for $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ and $(\boldsymbol{w}_1, \boldsymbol{w}_4)$ under the logisticMCAR and the prior in Jo et al. (2017) with parameters as in Section 4.3. Observe how the distance between $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ decreases as the sparsity increases under both priors. This is expected since areas 1 and 2 are neighbors. However, the distance between $(\boldsymbol{w}_1, \boldsymbol{w}_4)$ increases with sparsity under the logisticMCAR but decreases under CK-SSM. Hence, under CK-SSM, forcing sparsity in the mixture model results in imposing similar behaviors to different connected components.
- Fig. D.13 shows draws from the prior mixture density corresponding to parameters sampled from the prior under the logisticMCAR and CK-SSM, having fixed the atoms to have means $\mu_1 = -5$, $\mu_2 = -3.33, \ldots, \mu_6 = 5$ and equal variances 0.25^2 and remaining hyperparameters as in Section 4.3. It is clear that the logisticMCAR prior allows great variety among disconnected components as well as across different independent samples. Instead, CK-SSM shows that only the first 2/3 components have a nonzero weight, so that the densities across different areas and coming from independent samples are also similar.
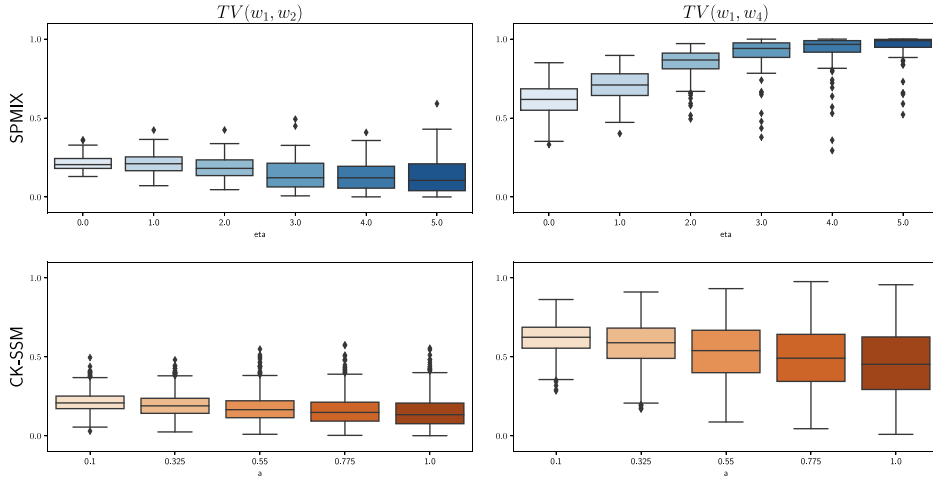
Fig. D.12. Total variation distances between the vectors $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ and $(\boldsymbol{w}_1, \boldsymbol{w}_4)$ under the logisticMCAR distribution (first row) and the CK-SSM(second row). Each plot shows the boxplots of 1000 independents simulations, for different values of the sparsity-tuning parameters (sparsity is increasing from left to right in each plot). The remaining hyperparameters and the adjacency matrix are as discussed in Section 4.3.

- Table D.6 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 ± one empirical standard deviation of the 100 values obtained.
- Fig. D.14 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.
- Fig. D.15 displays empirical correlations among the predictors and, in the last column, between predictors and the response for the Airbnb Amsterdam dataset in Section 7.
- Fig. D.16 shows the scatterplots of the response price versus numerical predictors and boxplots for categorical predictors for the Airbnb Amsterdam dataset in Section 7.
- Fig. D.17 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 reasons for this. First, 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.
- Table 4 shows the posterior predictive probability $P(y^\star > 200 \mid x^\star, i)$ for the same neighborhoods and values of $x^\star$ considered in Fig. 7(c).

Appendix E. Further comparisons on the Airbnb Amsterdam dataset

Below, we fit two additional models on the Airbnb dataset, and compare them with our proposal.

E.1. Modeling disconnected components separately

As pointed out by one of the reviewers, modeling disconnected components jointly might result in a less flexible model since the parameters ρ and Σ are shared across the spatial components.
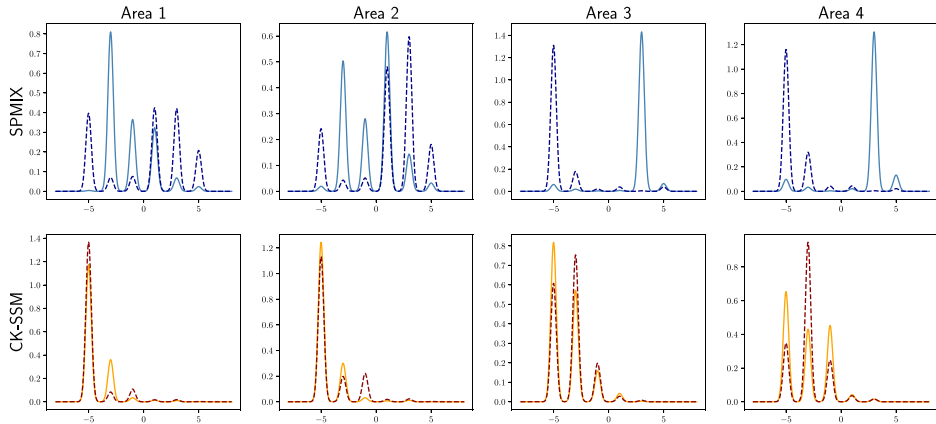
Fig. D.13. Samples from the prior distribution with $H = 6$ under the logisticMCAR (first row, with $\eta = 5.0$) and CK-SSM in Jo et al. (2017) (second row, with $b = 0.5, a = 1.0$). The remaining hyperparameters and the adjacency matrix are as discussed in Section 4.3. Each plot shows the mixture density in one particular area and different line types/colors represent independent draws from the prior. Here the means of the mixture model are fixed as $\mu_1 = -5$, $\mu_2 = -3.33, \ldots, \mu_6 = 5$ and the variances are all equal to 0.25^2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table D.6
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.

|	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
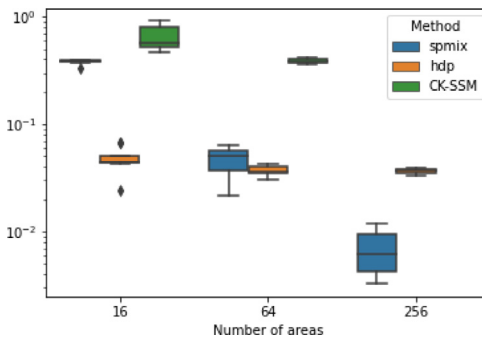
Fig. D.14. Boxplots of the Hellinger distance between true density (22) and estimated one under our model (spmix), the HDP-mixture model (hdp) and the CK-SSM for each simulation, averaged over the areas, for $I = 16, 64, 256$, in logarithmic scale, in Section 6.2.
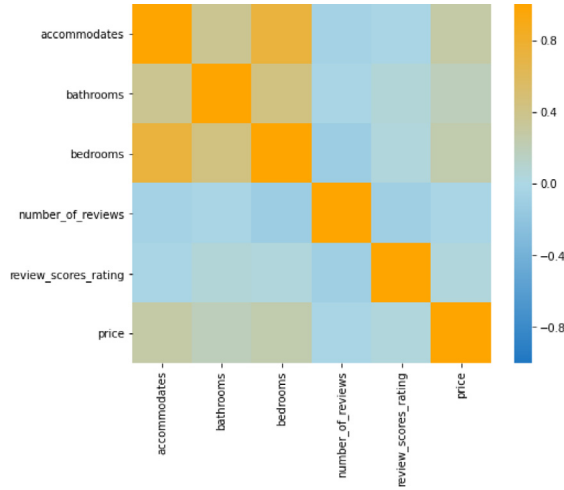
Fig. D.15. Correlation matrix between numerical predictors and response for Airbnb Amsterdam.
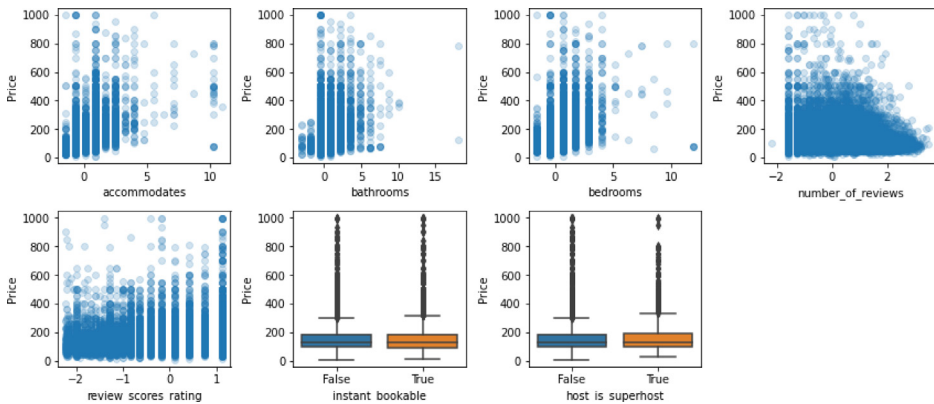

Fig. D.16. Scatterplots and boxplots of the nightly price versus predictors for Airbnb Amsterdam. Numerical predictors have been standardized.

On the other hand, modeling different connected components independently might result in an overparametrized model with poorer predictive performance.

We compare model $M1$ in Section 7 with another model obtained by applying $M1$ separately to the two connected components. By computing LPML and WAIC, we observe that this second model results in a worse predictive error (with a decrease of LMPL and an increase of WAIC of about 0.1%, for all the values of H considered), showing that the joint estimation of the shared parameters across the disjoint connected components, such as ρ, Σ and τ_h, is beneficial in this scenario.

E.2. Geographically weighted regression

Another comparison was performed by fitting geographically weighted regression (GWR, Brunsdon et al., 1998) to this dataset. Observe that while our model estimates an entire predictive distribution for each observation, from which different point estimates can be easily derived as discussed in Section 7, GWR is a frequentist model that can only provide a point estimate for the mean of the response.
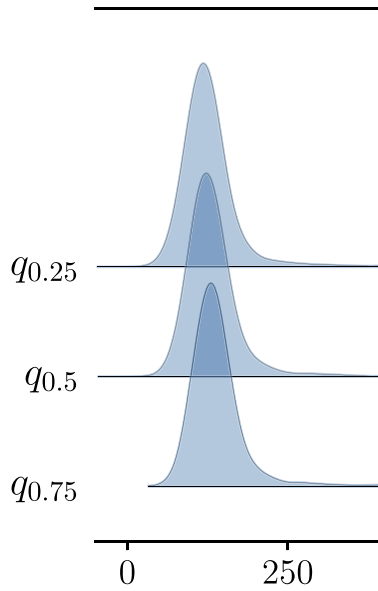
Fig. D.17. Predictive density for a new listing in *Bijlmer-Centrum* with all numerical covariates fixed to the empirical median of the dataset except reviews_scores_rating that ranges in the values $q_{0.25}, q_{0.5}, q_{0.75}$, where q_α denotes the empirical quantile of order α. Each line corresponds to one of these values, from top to bottom.

GWR considers observations $\{(y_j, \boldsymbol{x}_j)\}_{j=1}^{J}$ associated to spatial locations s_j. In our case, the observations are the nightly price (y_j) and the covaraites $(\boldsymbol{x}_j \in \mathbb{R}^d)$ of an Airbnb listing, while the spatial locations are the neighborhoods. The GWR model is then

$$y_j = \boldsymbol{\beta}_{s_j}^t \boldsymbol{x}_j + \varepsilon_j$$

We can assume that the ε_j's are independent and have mean equal to zero. The coefficients $\boldsymbol{\beta}_{s_j}$ (observe that in our case there is one for each neighborhood), are estimated by solving a weighted least-squares optimization problem so that:

$$\boldsymbol{\beta}_{s_j} = \left(X^T W_{s(y_j)} X\right)^{-1} \left(X^T W_{s_j} Y\right)$$

where X is the $J \times d$ matrix with rows x_j, Y is the vector with entries y_j and W_{s_j} is a diagonal matrix whose elements are a function of the distance between location s_j and the location associated to observation y_i. With an abuse of notation, denote with $s(y_i)$ the neighborhood observation y_i belongs to, then $(W_{s_j})_{ii} = g(d(s_j, s(y_i)))$.

As in Brunsdon et al. (1998), we let the weights $g(d(s_j, s(y_i))) = \exp(-\gamma d(s_j, s(y_i)))$. For two neighborhoods s_i and s_j we let $d(s_i, s_j)$ be the shortest path distance on the graph G between neighborhood i and the neighborhood which observation j belongs to.

Performance is measured by means of pMSE through a 10-fold cross-validation. In each fold, γ is chosen among the values $\{0.01, 0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$ via another (nested) cross-validation.

The resulting pMSE is much higher (roughly five times) than the one from $M1$, for all values of H. While GWR results to be not competitive with our model in this case, we expect that it might yield better performances in other scenarios with a finer spatial scale (e.g., geolocation of each observation). In such cases, GWR could be preferred to provide a point estimate.

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London.

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

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

Banerjee, S., 2016. Spatial data analysis. Annu. Rev. Public Health 37, 47–60.

Banerjee, S., Carlin, B.P., 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. J. R. Stat. Soc. Ser. B Stat. Methodol. 36 (2), 192–225.

Bhadra, A., Datta, J., Polson, N.G., Willard, B., et al., 2019. Lasso meets horseshoe: A survey. Statist. Sci. 34 (3), 405–427.

Box, G.E., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B Stat. Methodol. 26 (2), 211–243.

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically weighted regression. J. R. Stat. Soc. D 47 (3), 431–443.

Celeux, G., Frühwirth-Schnatter, S., Robert, C.P., 2019. Model selection for mixture models-perspectives and strategies. In: Handbook of Mixture Analysis. Chapman & Hall/CRC, pp. 117–154.

Clayton, D., 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.

Cressie, N., 1993. Statistics for Spatial Data. Wiley.

Cressie, N., Wikle, C.K., 2015. Statistics for Spatio-Temporal Data. John Wiley & Sons.

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

Dunson, D.B., 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., 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.

Geisser, S., Eddy, W.F., 1979. A predictive approach to model selection. J. Amer. Statist. Assoc. 74 (365), 153–160.

Gelfand, A.E., Kottas, A., MacEachern, S.N., 2005. Bayesian nonparametric spatial modeling with Dirichlet process mixing. J. Amer. Statist. Assoc. 100 (471), 1021–1035.

Gelfand, A.E., 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.

Griffith, D., Chun, Y., Li, B., 2019. Spatial Regression Analysis using Eigenvector Spatial Filtering. Academic Press.

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

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

Kaiser, M.S., Cressie, N., 2000. The construction of multivariate distributions from Markov random fields. J. Multivariate Anal. 73 (2), 199–220.

Lee, D., 2013. CARBayes: An R Package for Bayesian spatial modeling with conditional autoregressive priors. J. Stat. Softw. 55 (13), 1–24, URL https://www.jstatsoft.org/v55/i13/.

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

Leroux, B.G., Lei, X., 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. Springer, pp. 179–191.

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

Lijoi, A., Prünster, I., Rigon, T., 2020. The pitman–yor multinomial process for mixture modeling. Biometrika 107 (4), 891–906.

Lu, H., Reilly, C.S., Banerjee, S., Carlin, B.P., 2007. Bayesian areal wombling via adjacency modeling. Environ. Ecol. Stat. 14 (4), 433–452.

MacEachern, S.N., 2000. Dependent Dirichlet Processes. Department of Statistics, the Ohio State University, pp. 1–40, Unpublished Manuscript.

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

Mardia, K., 1988. Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. J. Multivariate Anal. 24 (2), 265–284.

Miller, J.W., Harrison, M.T., 2018. Mixture models with a prior on the number of components. J. Amer. Statist. Assoc. 113 (521), 340–356.

Nobile, A., 1994. Bayesian Analysis of Finite Mixture Distributions (PhD thesis). Ph.D. thesis, Carnegie Mellon University, Pittsburgh.

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

Polson, N.G., Scott, J.G., Windle, J., 2013. Bayesian inference for logistic models using Pólya–Gamma latent variables. J. Amer. Statist. Assoc. 108 (504), 1339–1349.

Pyrcz, M.J., Deutsch, C.V., 2014. Geostatistical Reservoir Modeling. Oxford University Press.

Ren, L., Du, L., Carin, L., Dunson, D., 2011. Logistic stick-breaking process. J. Mach. Learn. Res. 12 (Jan), 203–239.

Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). J. R. Stat. Soc. Ser. B Stat. Methodol. 59 (4), 731–792.

Roberts, G.O., Rosenthal, J.S., 2009. Examples of adaptive MCMC. J. Comput. Graph. Statist. 18 (2), 349–367.

Rodriguez, A., Dunson, D.B., 2011. Nonparametric Bayesian models through probit stick-breaking processes. Bayesian Anal. 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., Blei, D.M., 2006. Hierarchical Dirichlet processes. J. Amer. Statist. Assoc. 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. Guardian 6 October 2016.

Velázquez, E., Martínez, I., Getzin, S., Moloney, K.A., Wiegand, T., 2016. An evaluation of the state of spatial point pattern analysis in ecology. Ecography 39 (11), 1042–1055.

Wachsmuth, D., Weisler, A., 2018. Airbnb and the rent gap: Gentrification through the sharing economy. Environ. Plan. A 50 (6), 1147–1170.

Watanabe, S., 2013. A widely applicable Bayesian information criterion. J. Mach. Learn. Res. 14 (Mar), 867–897.

Webster, R., Oliver, M.A., 2007. Geostatistics for Environmental Scientists. John Wiley & Sons.

Zhou, Z., Matteson, D.S., Woodard, D.B., Henderson, S.G., Micheas, A.C., 2015. A spatio-temporal point process model for ambulance demand. J. Amer. Statist. Assoc. 110 (509), 6–15.