Normalized Latent Measure Factor Models

Mario Beraha

Department of Economics and Statistics University of Torino. Torino, Italy

E-mail: mario.beraha@unito.it

Jim E. Griffin

Department of Statistical Science, University College London. London, United Kingdom

E-mail: j.griffin@ucl.ac.uk

Summary.
We propose a methodology for modeling and comparing probability distributions within a Bayesian nonparametric framework. Building on dependent normalized random measures, we consider a prior distribution for a collection of discrete random measures where each measure is a linear combination of a set of *latent* measures, interpretable as characteristic traits shared by different distributions, with positive random weights. The model is non-identified and a method for post-processing posterior samples to achieve identified inference is developed. This uses Riemannian optimization to solve a non-trivial optimization problem over a Lie group of matrices. The effectiveness of our approach is validated on simulated data and in two applications to two real-world data sets: school student test scores and personal incomes in California. Our approach leads to interesting insights for populations and easily interpretable posterior inference.

Keywords: Comparing probability distributions; Dependent random measures; Latent factor models; Normalized random measures; Riemannian optimization

1. Introduction

Modeling a set of related probability measures is a common task in Bayesian statistics, the most common example being when covariates are associated with each observation. In this work, we consider the case of a single discrete-valued covariate, which might be regarded as a group indicator, that is, when data are naturally divided into subpopulations or groups. One of the main motivations for these kinds of analyses is combining data from different sources or experiments, where, for each source, a set of observations is collected: pooling together all the data could ignore important differences across populations while modeling each group separately might result in poor performance especially if the number of observations in each group is small. Applications range from population genetics (Elliott et al., 2019) to healthcare (Müller et al., 2004; Rodríguez et al., 2008) and text mining (Teh et al., 2006).

Within this setting, our goal is to propose a flexible model that, in addition to combining heterogeneous sources of data, gives an efficient way of representing the difference in distribution across populations. Consider for example Figure 1, which displays the distribution of the personal annual income (on the log scale) in four different geographic areas of California: two in Los Angeles and two in San Francisco. In this case, similarities and differences between the distributions can be easily spotted by eye: the two areas in Los Angeles are associated with (much) lower incomes than the areas in San Francisco. When the number of groups increases, it is not possible to carry out these comparisons by eye. Our model provides a way to decompose the area-specific densities into a linear combination of "common traits", which are themselves probability measures. In Section 6.2, we

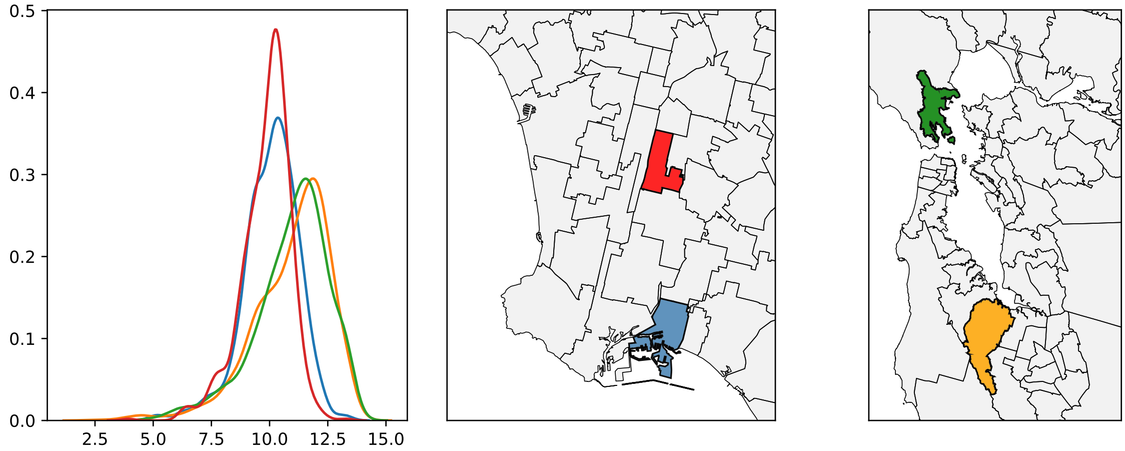

Fig. 1: Kernel density estimates of the (log) personal incomes in four areas in California (left plot): two in Los Angeles (middle plot) and two in San Francisco (right plot).

provide a thorough analysis of the Californian income data, finding four common traits, associated with an average distribution of income, and a prevalence of low, medium, and high incomes respectively. By looking at the weights (of the linear combination of common traits) associated with the four groups in Figure 1, we easily spot differences between the Los Angeles and San Francisco areas: the weight associated to the low-income trait is large in the first areas and low in the second two; vice versa for the weight associated to the high-income trait. See Figure 8 for more details.

To formalize the discussion above, let $(\boldsymbol{y}_1, \dots, \boldsymbol{y}_g)$ denote a sample of observations divided into g groups where $\boldsymbol{y}_j = (y_{j1}, \dots, y_{jn_j})$. A common assumption is that data are exchangeable in each group, but exchangeability might not hold across different groups. In particular, by de Finetti's theorem, this is tantamount to assuming that there is a vector of random probability measures $(p_1, \dots, p_g) \sim Q$ such that, in each group, $y_{j1}, \dots, y_{jn_j} \overset{\text{iid}}{\sim} p_j$ and that independence, conditionally on $p_1, \dots, p_g$, holds across groups. We focus here on mixture models of the kind $p_j(y) = \int_\Theta f(y \mid \theta) \widetilde{p}_j(\mathrm{d}\theta)$.

The construction of a flexible prior Q that can suitably model heterogeneity while borrowing information across different groups has been thoroughly studied in Bayesian nonparametrics. See Quintana et al. (2022) for a recent review of such constructions. Previously proposed approaches consider constructing $\widetilde{p}_1, \dots, \widetilde{p}_g$ in a hierarchical model fashion (Teh et al., 2006; Camerlenghi et al., 2019; Bassetti et al., 2020; Argiento et al., 2019), considering convex combinations of shared and group-specific random measures (Müller et al., 2004), and starting from additive processes (Griffin et al., 2013; Lijoi et al., 2014). Another fruitful approach to borrow strength across different groups is to cluster the $\widetilde{p}_j$'s by building Q using nested processes, such as is the case of the nested Dirichlet process (Rodríguez et al., 2008), of which several extensions and modifications have been proposed in the last few years starting from Camerlenghi et al. (2019), see, for instance, Denti et al. (2021); Beraha et al. (2021); Lijoi et al. (2022). As previously mentioned, the focus of the present paper is slightly different. First of all, we are interested in the situation when the number of groups g is large relative to the sample size in each group n_j. Then, it is likely that the dataset cannot inform the huge number of parameters that are associated with extremely flexible models and we advocate for a more parsimonious model where substantial sharing of information is encouraged across different groups of data. Moreover, in addition to modeling the densities $p_1, \dots, p_g$, we also want to identify the main differences in distribution of the data across groups. To the best of our knowledge, this question has not been addressed systematically in the Bayesian nonparametric literature. In the frequentist one, several approaches to principal component analysis for probability distribution have been proposed, see for instance Pegoraro and Beraha (2022) and the references therein.

The setting "large g, small n_j" is somewhat reminiscent of high-dimensional data analysis, where the dimension of each observation is large relative to the sample size. In this case, latent factor models (see, e.g., Arminger and Muthén, 1998) provide a powerful tool. In a latent factor model, it is assumed that each observation $x_i \in \mathbb{R}^p$ is a linear combination of a set of H d-dimensional latent factors weighted by observation-specific scores, plus an isotropic error term. We follow this analogy and propose *normalized latent measure factor models*, a class of prior distributions for a vector of random probability measures $\widetilde{p}_1, \ldots, \widetilde{p}_g$. Informally, our model amounts to considering $\widetilde{p}_j$ as a convex combination of a set of latent random probability measures, see Section 2.

Our construction shares similarities with Griffin et al. (2013) and Lijoi et al. (2014). There, the authors assume each $\widetilde{p}_j$ as the normalization of a random measure obtained by superposing several completely random measures. Essentially, this is analogous to our approach if we let all the scores (before some normalization step, see Section 2) be zero or one. The main difference is that, since their scores are binary, they usually assume that the number of latent factors H is larger than the number of groups g. This leads to posterior simulation algorithms that can scale and/or mix poorly with g. Moreover, they do not consider the problem of decomposing the populations' distribution into interpretable common traits, which necessarily requires H to be much smaller than g.

As is usually the case for latent factor models, our model is not identifiable, due to two parameter matrices entering multiplicatively. To tackle this issue, we propose post-processing the MCMC chains to find an "optimal representative" for both parameters which leads to a non-trivial optimization problem. Indeed, taking into account the invariance to scaling of normalized random measures leads to formulating the optimization over a Riemannian manifold of matrices, specifically the special linear group (matrices whose determinant is equal to one). Moreover, additional constraints must be taken into account to ensure the positiveness of both parameters and we propose an iterative algorithm based on gradient descent. The first constraint (determinant equal to one) can be tackled by means of differential geometric tools: leveraging the differential structure of the special linear group, we use a variant of Riemannian gradient descent which ensures that all the intermediate points of the algorithm lie inside the special linear group. To take into account the positivity constraints, we propose to use the augmented Lagrangian multiplier method within the previously discussed Riemannian framework, leading to a Riemannian augmented Lagrangian multiplier method.

We consider two motivating applications. The first one is the scores on a mathematics test of approximately $40,000$ students in 1048 Italian high schools from the *invalsi* dataset. The median number of students taking the test in each high school is as few as 37, the minimum being 4 and the maximum 131. The second one comes from the US income survey. Here, the groups are represented by geographical units called *PUMAs*, which correspond to areas with roughly $100,000$ inhabitants. We show how our model can be adapted to induce correlation between the distribution of incomes in PUMAs that are geographically close, by assuming that the scores are distributed as a log Gaussian Markov random field. Compared to traditional spatial factor models, we introduce the spatial dependence in the loadings matrix instead of the latent factors.

The rest of the paper is organized as follows. Section 2 formalizes our model and discusses its statistical properties. Section 3 describes the MCMC algorithm for posterior inference and we present our post-processing algorithm in Section 4. Section 5 and Section 6 present numerical illustration on simulated data and real data, respectively. Finally, we discuss possible extensions of the proposed approach in Section 7. The Supplementary Material collects background material on Riemannian optimization and completely random measures, proofs of the theoretical results, and additional simulations. `Python` code implementing the MCMC and the post-processing algorithms is available at github.com/mberaha/nrmifactors.

2. The Model

For simplicity and specifity, we assume that each $y_{ji} \in \mathbb{R}^d$ and that $\Theta \subset \mathbb{R}^q$ for some d and q. The results can be easily extended to the case when y_{ji} are elements of a complete and separable (i.e., Polish) metric space and Θ is Polish as well.

To keep the discussion light, we defer all technical details and the proofs of the results to the Supplementary Material.

2.1. Preliminaries

Before presenting our model in detail, we give some background material on completely random measure and their normalization. This will constitute the backbone of our approach.

Let $(\Theta, \mathcal{B}(\Theta))$ be a complete and separable metric space endowed with its Borel σ-algebra. A random measure is a random element μ taking values in the space of probability measures over Θ, such that $\mu(B) < +\infty$ almost surely for all $B \in \mathcal{B}(\Theta)$. Such a measure is termed completely random by Kingman (1967) if, for pairwise disjoint $B_1, \ldots, B_n \in \mathcal{B}(\Theta)$, the random variables $\mu(B_j)$, $j = 1, \ldots, n$, are independent. For our purposes, it is sufficient to consider completely random measures of the kind $\mu(A) = \int_{\mathbb{R}_+ \times A} s N(\mathrm{d}s \, \mathrm{d}x)$, where N is a Poisson point process on $\Theta \times \mathbb{R}_+$ with base (intensity) measure $\rho(\mathrm{d}s \, \mathrm{d}x)$. We further assume $\rho(\mathrm{d}s \, \mathrm{d}x) = \nu(\mathrm{d}s) \, \alpha(\mathrm{d}x)$ where ν is a Lévy measure on the positive reals, α is a Borel measure on Θ. See, e.g., Kingman (1993) for a detailed account of random measures.

A fruitful approach to constructing random probability measures is by normalization of completely random measures, i.e., by setting $p(\cdot) = \mu(\cdot)/\mu(\Theta)$, which was originally introduced in Regazzini et al. (2003). For the random measure p to be well defined, one must ensure that $\mu(\Theta) > 0$ and $\mu(\Theta) < +\infty$ almost surely. As shown in Regazzini et al. (2003), sufficient conditions are $\int_{\mathbb{R}_+} \nu(\mathrm{d}s) = +\infty$ and $\int_{\mathbb{R}_+} \min\{1, s\} \, \nu(\mathrm{d}s) < +\infty$.

2.2. Normalized Latent Measure Factor Models

As already mentioned in the Introduction, we assume

$$y_{j1}, \ldots, y_{jn_j} \mid \widetilde{p}_j \overset{\text{iid}}{\sim} p_j := \int_\Theta f(\cdot \mid \theta) \widetilde{p}_j(\mathrm{d}\theta)$$

and that each $\widetilde{p}_j$ is a normalized random measure, that is

$$\widetilde{p}_j(\cdot) = \frac{\widetilde{\mu}_j(\cdot)}{\widetilde{\mu}(\Theta)}, \qquad j = 1, \ldots, g.$$

Then, the model is specified by a choice of the mixture kernel $f(\cdot \mid \cdot)$ and a prior distribution for $(\widetilde{\mu}_1, \ldots, \widetilde{\mu}_g)$. Let $(\mu^*_1, \ldots, \mu^*_H)$ be a completely random vector (i.e., a vector of completely random measures). Let λ_{jh}, $j = 1, \ldots, g$, $h = 1, \ldots, H$ be a double sequence of almost surely positive random variables (specific choices of the distribution of the λ_{jh}'s are discussed later). We assume

$$\widetilde{\mu}_j(\cdot) = \sum_{h=1}^{H} \lambda_{jh} \, \mu^*_h(\cdot). \tag{1}$$

Note that (1) generalizes the construction in Griffin et al. (2013) and Lijoi et al. (2014). Specifically, we recover the GM-dependent Dirichlet process in Lijoi et al. (2014) by setting $g = 2$, $H = 3$, fixing $(\lambda_{1,1}, \lambda_{1,2}, \lambda_{1,3}) = (1, 0, 1)$ and $(\lambda_{2,1}, \lambda_{2,2}, \lambda_{2,3}) = (0, 1, 1)$, and assuming that the $(\mu_1^*, \mu_2^*, \mu_3^*)$ is a vector of independent Gamma processes. In the CNRMI process in Griffin et al. (2013), $H > g$ is generic and Λ is a $g \times H$ binary matrix. The random measures μ_h^*'s are independent completely random measures with Lévy intensity $M_h \nu(s)\mathrm{d}s\alpha(\mathrm{d}x)$. In the most general formulation, Griffin et al. (2013) set $H = 2^{g-1}$ and the h-th column of Λ is fixed equal to the binary representation of h. Then, the authors propose to perform variable selection on the columns of Λ by assuming a prior for the M_h's parameters that is a mixture of a point mass at 0 and a diffuse distribution on $\mathbb{R}_+$. Indeed, if $M_h = 0$ then $\mu_h^*(A) = 0$ for any measurable A, that is equivalent to removing the h-th column of Λ.

We could choose $(\mu_1^*, \ldots, \mu_H^*)$ to be independent and identically distributed random measures, i.e.

$$\mu_h^*(\cdot) = \sum_{k \geq 1} W_{hk}\, \delta_{\theta_{hk}^*}(\cdot)$$

where $\{W_{hk}, \theta_{hk}^*\}_{k=1}^{\infty}$ are the points of a Poisson point process on $[0, +\infty) \times \Theta$ with, for instance, intensity $\nu_h(\mathrm{d}s_h\, \mathrm{d}x_h) = \rho(s_h)\mathrm{d}s_h\, \alpha(\mathrm{d}x_h)$, i.e., all the intensities are equal. This choice leads to a particularly tractable model for $(\widetilde{\mu}_1, \ldots, \widetilde{\mu}_g)$ as we have that marginally, each $\widetilde{\mu}_j$ is a completely random measure as specified in the following proposition.

Proposition 1. *Let $\widetilde{\mu}_j = \sum_{h=1}^{H} \lambda_{jh}\mu_h^*$ where the μ_h^*'s are completely random measures with associated Lévy intensity $\nu_h^*(\mathrm{d}s_h, \mathrm{d}x_h) = \rho_h^*(s_h)\mathrm{d}s_h\, \alpha_h^*(\mathrm{d}x_h)$. Further, assume that the μ_h^*'s are independent. Then $\widetilde{\mu}_j$ is a completely random measure with Lévy intensity*

$$\nu_j(\mathrm{d}s, \mathrm{d}x) = \sum_{h=1}^{H} \frac{1}{\lambda_{jh}} \rho_h^*(s/\lambda_{jh}) \alpha_h^*(\mathrm{d}x)$$

We find that a more suitable model for our applications arises when $\mu_1^*, \ldots, \mu_H^*$ share their support points. In particular, we will assume that $\mu_1^*, \ldots, \mu_H^*$ is a compound random measure (CoRM, Griffin and Leisen, 2017). That is,

$$\mu_h^*(\cdot) = \sum_{k \geq 1} m_{hk} J_k \delta_{\theta_k^*}(\cdot),$$

where m_{hk} are positive random variables such that $m_k = (m_{1k}, \ldots, m_{Hk})$, $k \geq 1$, are independent and identically distributed from a probability measure on $\mathbb{R}_+^H$, and $\eta = \sum_{k \geq 1} J_k \delta_{\theta_k^*}$ is a completely random measure with Lévy intensity $\nu^*(\mathrm{d}z)\alpha(\mathrm{d}x)$. We argue that a CoRM-based construction should be preferred to an independent CRMs-based one since (i) sharing atoms across all measures is linked to better predictive performance (Quintana et al., 2022), (ii) the number of parameters involved is much smaller, which ultimately leads to the possibility of fitting this model to large datasets, and (iii) each latent factor μ_h^* can be interpreted separately (through the post-processing algorithm presented in Section 4). The effectiveness of this model comes with a tradeoff in analytical tractability, since, as shown in the Supplementary Material, the random measure (2) is not completely random.

In this case we can write

$$\widetilde{\mu}_j(\cdot) = \sum_{k \geq 1} (\Lambda M)_{jk} J_k \delta_{\theta_k^*}(\cdot), \tag{2}$$

where Λ is the $J \times H$ matrix with entries λ_{jh}, M is a $H \times \infty$ matrix, so that $\Gamma = \Lambda M$ is a $g \times \infty$ matrix with entries γ_{jk}, $j = 1, \ldots, g$, $k \geq 1$. Note that, in analogy to CoRMs, our model includes shared weights J_k for all the measures $\widetilde{\mu}_j$. We find that the additional borrowing of strength obtained through the J_k's is useful in practice since, in our applications, the $\widetilde{\mu}_j$'s are usually similar.

Equations (1) and (2) share analogies to latent factor models, where the observed variable is $X \in \mathbb{R}^p$ and its ℓ-th entry is modeled as $X_\ell \approx \sum_{h=1}^H \omega_{\ell h} Z_h$, for $Z = (Z_1, \ldots, Z_H)$ an H-dimensional random variable. In particular, we could consider $\mu_1^*, \ldots, \mu_H^*$ to be measure-valued factor loadings and the λ_{jh}'s to be factor scores. This yields an interpretation similar to functional factor models (Montagna et al., 2012). On the other hand, we could consider the measure-valued vector $(\widetilde{\mu}_1, \ldots, \widetilde{\mu}_g)$ as a single high-dimensional observation, and model it as a linear combination of measure-valued factors with loadings λ_{jh}'s. Both interpretations make sense and lead to interesting analogies. We use the latter one and call Λ the loadings matrix and the μ_h^*'s the latent measures.

Prior elicitation is required to set the Lévy intensity ν^* of the CoRM, the distribution of the scores m_{hk}, and the distribution of Λ. Following Griffin and Leisen (2017), we assume that $m_{hk} \overset{\text{iid}}{\sim} \text{Ga}(\phi)$, where $\text{Ga}(\phi)$ denotes the law of a gamma random variable with shape parameter ϕ and rate parameter 1 (we will also use $\text{Ga}(\phi, \beta)$ to denote a gamma random variable with rate parameter $\beta \neq 1$). Therefore, the dependence across the $\widetilde{\mu}_j$'s depends on H, ν^*, and Λ.

The prior for Λ allows us to address several interesting modeling questions. When no additional group-specific information is available, such as comparing the distribution of test results in different schools, a natural choice would be to assume the λ_{ij}'s i.i.d. from some probability distribution with support on $\mathbb{R}_+$, such as the gamma distribution. We find it more convenient to specify a *shrinkage* prior on Λ, to automatically select the number of latent factors H. This approach has received considerable attention in Gaussian latent factor models, see, for instance, Bhattacharya and Dunson (2011); Legramanti et al. (2020); Schiavon et al. (2022). In our example, we consider Λ distributed as the variances of a multiplicative gamma process (Bhattacharya and Dunson, 2011), i.e., we assume:

$$\lambda_{jh} = (\phi_{jh}\tau_h)^{-1}, \ \tau_h = \prod_{j=1}^{h} \theta_j, \ \theta_1 \sim \text{Ga}(a_1), \ \theta_2, \ldots \overset{\text{iid}}{\sim} \text{Ga}(a_2), \ \phi_{jh} \overset{\text{iid}}{\sim} \text{Ga}(\nu/2, \nu/2). \quad (3)$$

With an abuse of notation, in the rest of the paper, we say that Λ is distributed as a multiplicative gamma process if (3) holds. In Section 3 we propose a variant of the adaptive Gibbs sampler of Bhattacharya and Dunson (2011) to automatically select H in the first iterations of the MCMC algorithm.

If group-specific information, such as covariates, is available, we can model the finite-dimensional matrix Λ. For example, the PUMAs in the Californian income data are indexed by a specific areal location. This can be modelled using a $g \times g$ spatial proximity matrix denoted by W, where $W_{j\ell} = 1$ if areas j and ℓ share an edge and $W_{j\ell} = 0$ otherwise, but more general choices of proximity could be considered in other examples. Then, we can encourage spatial dependence between the $\widetilde{\mu}_j$'s by assuming

$$\log \boldsymbol{\lambda}^h \overset{\text{iid}}{\sim} \mathcal{N}_H\left(\mu, (\tau(F - \rho W))^{-1}\right), \qquad h = 1, \ldots, H \quad (4)$$

where $\boldsymbol{\lambda}^h = (\lambda_{1h}, \ldots, \lambda_{gh})$ is the h–th column of the matrix Λ, F is a diagonal matrix with entries $F_{ii} = \sum_j W_{ij}$, and $\rho \in (0, 1)$. We suggest setting $\mu = \log(1/H, \ldots, 1/H)$ in (4) to encourage a priori each $\widetilde{\mu}_j$ to be a convex combination of the μ_h^*'s with equal weights. The model could also be applied to geo-referenced data using a log Gaussian process,

$$\log \boldsymbol{\lambda}^h \overset{\text{iid}}{\sim} \mathcal{GP}(\mu, \mathcal{K}), \qquad h = 1, \ldots, H$$

where $\boldsymbol{\lambda}^h = (\lambda_{1h}, \ldots, \lambda_{gh})$ is the h–th column of the matrix Λ. In a similar fashion, if group-specific covariates $\boldsymbol{x}_j \in \mathbb{R}^q$ were available, these could also be included in our model by assuming

$$\log \lambda_{jh} \mid \boldsymbol{\beta}_h, s_h^2 \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_j^\top \boldsymbol{\beta}_h, s_h^2), \qquad j = 1, \dots, g, \; h = 1, \dots, H$$

and standard parametric priors could be assumed for $(\boldsymbol{\beta}_h, s_h^2)$. Of course, more complex functional relationships between the mean (and/or variance) of the $\log \lambda_{jh}$'s and the available covariates can be also assumed, together with standard priors on the associated parameters

2.3. Statistical Properties

In this section, we discuss some distributional properties of the measures $\widetilde{\mu}_1, \dots, \widetilde{\mu}_g$ in light of the prior assumption above. We assume that the λ_{jh}'s are independent of $\mu_1^*, \dots, \mu_H^*$. Firstly, it is clear that

$$\mathbb{E}[\widetilde{\mu}(A)] = \sum_{h=1}^{H} \mathbb{E}[\lambda_{jh}] \mathbb{E}[\mu_h^*(A)].$$

When we consider the normalized measures, the expression of the expected value is more complex.

Theorem 1. *Let $(\mu_1^*, \dots, \mu_H^*)$ be a CoRM with i.i.d. scores. Denote the Laplace transform of the scores' distribution by $\mathcal{L}_m(u) := \mathbb{E}[e^{-um}]$ and let $\kappa_m(u, n) := \mathbb{E}[e^{-um} m^n]$. Then for all measurable $A \subset \Theta$*

$$\mathbb{E}[\widetilde{p}_j(A)] = \alpha(A) \sum_{h=1}^{H} \int \mathbb{E}\left[\lambda_{jh} \psi_\rho(u\lambda_{j1}, \dots, u\lambda_{jH}) \int_{\mathbb{R}_+} z \prod_{k \neq h} \mathcal{L}_m(u\lambda_{jk} z) \kappa_m(u\lambda_{jh} z, 1) \nu^*(\mathrm{d}z) \right] \mathrm{d}u$$

where ψ_ρ is the Laplace functional of $(\mu_1^*, \dots, \mu_H^*)$ (evaluated at the constant functions $u\lambda_{j1}, \dots, u\lambda_{jH}$).

Although it is not possible to evaluate the quantity in Theorem 1 analytically, a priori Monte Carlo simulation can be used to numerically estimate the expected value of $\widetilde{p}_j(A)$.

To characterize the dependence induced by the latent measure factor model, an intuitive measure is the covariance between two random measures.

Proposition 2. *The following expression holds.*

$$\mathrm{Cov}\left[\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(B)\right] = \sum_{h,k} \left\{ \mathbb{E}[\lambda_{jh}\lambda_{\ell k}] \mathrm{Cov}(\mu_h^*(A), \mu_k^*(B)) + \mathrm{Cov}(\lambda_{jh}, \lambda_{\ell k}) \mathbb{E}[\mu_h^*(A) \mu_k^*(B)] \right\} \quad (5)$$

If the λ_{jh}'s have the same marginal distribution, the μ_h^*'s have the same marginal distribution, $\lambda_j = (\lambda_{j1}, \dots, \lambda_{jH})$ and λ_ℓ (defined analogously) are independent, $\mathbb{E}[\lambda_{jh}\lambda_{\ell h}] = \kappa$, $\mathrm{Cov}(\lambda_{jh}, \lambda_{\ell h}) = \rho$ for all j, ℓ, h, then:

$$\mathrm{Cov}\left[\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(B)\right] = \mathrm{Cov}(\mu_1^*(A), \mu_1^*(B)) \kappa H + m_1^*(A) m_1^*(B) \rho H + \sum_{h \neq q} \bar{\lambda}_{11}^2 \mathrm{Cov}(\mu_h^*(A), \mu_k^*(B))$$

where $\bar{\lambda}_{jh} := \mathbb{E}[\lambda_{jh}]$ and $m_h^*(A) = \mathbb{E}[\mu_h^*(A)]$.

Finally, if in addition the μ_h^*'s are independent, the latter sum disappears

From (5), it is clear that $\text{Cov}\,[\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(B)]$ increases with: (i) the correlation of the measures at the latent level ($\text{Cov}(\mu^*_h(A), \mu^*_k(B))$ large), (ii) the correlation of the scores ($\text{Cov}(\lambda_{jh}, \lambda_{\ell k})$ large), (iii) large values in the scores ($\mathbb{E}[\lambda_{jh}\lambda_{\ell k}]$ large), (iv) random measures with large masses ($\mathbb{E}[\mu^*_h(A), \mu^*_k(B)]$ large), and (v) large values of H (more terms in the summation).

The correlation between $\widetilde{\mu}_j(A)$ and $\widetilde{\mu}_\ell(B)$ can be formally derived from (5) but its expression is not easily interpretable in general. The next proposition specialises it to the case $A = B$.

Proposition 3. *For any measurable A, let* $\text{Cov}(\mu^*_h(A), \mu^*_k(A)) = \text{Cov}(\mu^*_m(A), \mu^*_n(A)) =:$ c_A, $\mathbb{E}[\mu^*_h(A)] = \mathbb{E}[\mu^*_k(A)] =: m_A$. *Then*

$$\text{Cov}\,[\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(A)] = \mathbb{E}[\mu^*_1(A)^2] \left(\sum_{h=1}^{H} \mathbb{E}[\lambda_{jh}\lambda_{\ell h}]\right) + \\ (c_A + m_A^2)\left(\sum_{h \neq k} \mathbb{E}[\lambda_{jh}\lambda_{\ell k}]\right) - m_A^2 \left(\sum_{h,k} \bar{\lambda}_{jh}\bar{\lambda}_{\ell k}\right)$$

Let us specialize the above expression further. Consider first the case of independent scores $\lambda_{jh} \overset{\text{iid}}{\sim} \text{Ga}(\psi, 1)$. The correlation between $\widetilde{\mu}_j(A)$ and $\widetilde{\mu}_\ell(A)$ amounts to

$$\left(1 + \frac{m_A}{(\text{Var}[\mu^*_1(A)] + c_A(H-1))\,\psi}\right)^{-1} \tag{6}$$

which is an increasing function of H and ψ as expected. See Section B of the Supplementary Material for a proof.

To evaluate m_A, and c_A we use the following result.

Proposition 4. *Consider a CoRM with Ga(ϕ) distributed scores and gamma process marginals (i.e., each μ^*_h is distributed as a gamma process). Then for any measurable A:*

(a) $\mathbb{E}[\mu^*_h(A)] = \alpha(A)$,

(b) $\mathbb{E}[\mu^*_h(A)\mu^*_k(A)] = (\alpha(A) + \alpha(A)^2)\phi^2(B(1,\phi))^2 3/2$, *where $B(a,b)$ denotes the Beta function.*

Consider now the case when Λ is distributed as a multiplicative gamma process, see (3). In this case, we don't have an interpretable expression for the correlation between $\widetilde{\mu}_j(A)$ and $\widetilde{\mu}_\ell(A)$. In Section B of the Supplementary Material we report the expressions for $\text{Cov}\,[\widetilde{\mu}_j(A), \widetilde{\mu}_\ell(A)]$ and $\text{Var}[\widetilde{\mu}_j(A)]$ which might be used to numerically compute the desired correlation. Figure 2 displays the correlation between $\widetilde{\mu}_j(A)$ and $\widetilde{\mu}_\ell(A)$ for a set A such that $\alpha(A) = 0.5$. We notice that when the CoRM has gamma process marginals, the parameter ϕ has little effect on the correlation between the $\widetilde{\mu}_j$'s. On the contrary, there is a strong interaction between a_2, ν, and H. For smaller values of ν, larger values of H imply a higher correlation. When ν is sufficiently large (e.g. larger than 6), the effect of H is less evident. Moreover, larger values of a_2 imply a weaker correlation. This is expected as it essentially reduces the number of active latent measures. In Figure E.1 in the Supplementary Material, we show the correlation between $\widetilde{\mu}_j(A)$ and $\widetilde{\mu}_\ell(A)$ under prior (4) for different choices of areas j and ℓ, as a function fo τ and ρ.

Since the atoms are shared across all the measures $\widetilde{\mu}_j$'s, another possible way of characterizing the dependence between two measures is to consider the ratio of weights associated to the k–th atom in $\widetilde{\mu}_j$ and $\widetilde{\mu}_\ell$,

$$r^k_{j\ell} := \frac{(\Lambda M)_{jk}}{(\Lambda M)_{\ell k}} = \frac{\sum_{h=1}^{H} \lambda_{jh} m_{hk}}{\sum_{h=1}^{H} \lambda_{\ell h} m_{hk}}. \tag{7}$$

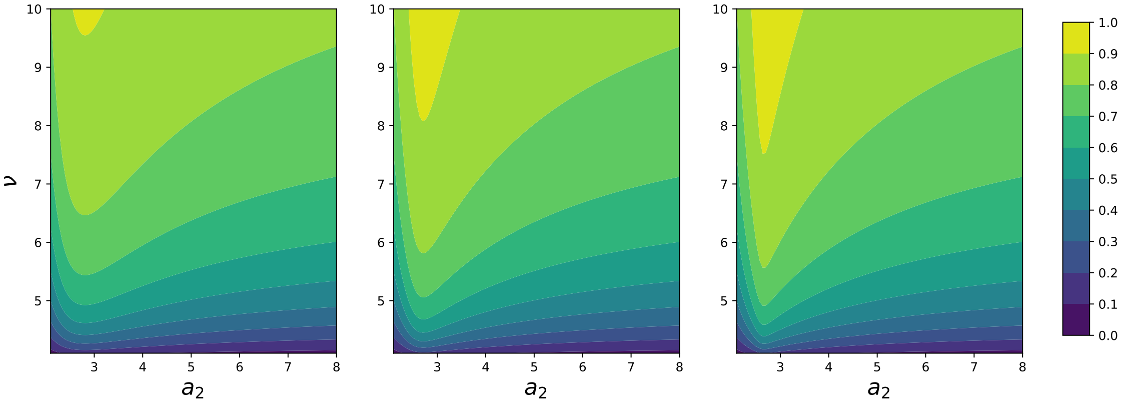

Fig. 2: Correlation between $\widetilde{\mu}_j(A)$ and $\widetilde{\mu}_\ell(A)$ for a set A such that $\alpha(A) = 0.5$, under the multiplicative gamma process prior (3). $a_1 = 2.5$, $\phi = 2$. From left to right $H = 4, 8, 16$. The values of a_2 vary across the x-axis in each plot, the values of ν across the y-axis.

Indeed, the closer to one the $r^k_{j\ell}$'s are, the more similar $\widetilde{\mu}_j$ and $\widetilde{\mu}_\ell$ will be, since their weights will be similar and the atoms are shared. On the other hand, values of $r^k_{j\ell}$ closer to 0 or much larger than 1 indicate that $\widetilde{\mu}_j$ and $\widetilde{\mu}_\ell$ will be very different. A trivial upper bound for (7) is

$$r^k_{j\ell} \leq \sum_{h=1}^{H} \frac{\lambda_{jh}}{\lambda_{\ell h}}.$$

Multiplying and dividing by H in (7) and taking the logarithm yields

$$\log r^k_{j\ell} = \log\left(\frac{1}{H}\sum_{h=1}^{H} \lambda_{jh} m_{hk}\right) - \log\left(\frac{1}{H}\sum_{h=1}^{H} \lambda_{\ell h} m_{hk}\right).$$

By the strong law of large numbers, we have that $\log r^k_{j\ell} \to 0$ as $H \to \infty$ if, for instance, λ_{jh} and $\lambda_{\ell h}$ are independent and identically distributed across the values of h. This further strengthens the assumption that $H \ll g$ in order for the model to make sense. Moreover, it is clear that the variance of $r^k_{j\ell}$ increases with the variance of the λ_{jh}'s. On the contrary, observe that if $H = 1$ the $\widetilde{\mu}_j$'s are all different but, because of the normalization, $\widetilde{p}_j = \widetilde{p}_k = \mu^*_1/\mu^*_1(\Theta)$ for any $j, k = 1, \ldots, g$. In Section E of the Supplementary Material we report an a prior Monte Carlo simulation comparing $r_{j\ell}$ as a function of H under different priors for Λ, namely and i.i.d. prior with Ga(ψ) distributed λ_{jh}'s, the multiplicative gamma process in (3) and the variances of the cumulative shrinkage prior (Legramanti et al., 2020). It is clear that under the two latter shrinkage priors, the choice of H has a smaller impact on the prior. For the sake of computational efficiency, we will adopt the multiplicative gamma process prior in our simulations, when no additional group-specific covariates are present. Instead, when we consider the case of area-referenced groups, we consider H to be a hyperparameter and perform model selection based on predictive performance

3. Posterior Inference

Let α be a measure on Θ, ν^* a Lévy intensity on $\mathbb{R}_+$, and $\phi > 0$. We denote with CoRM(ϕ, ν^*, α) the law of a compound random measure with i.i.d. Ga(ϕ)-distributed scores with directing random measure with intensity $\nu^*(z)\mathrm{d}z\,\alpha(\mathrm{d}\theta)$. Our model can be compactly summarized as

$$
\begin{aligned}
y_{ji} \mid \theta_{ji} &\overset{\text{ind}}{\sim} k(\cdot \mid \theta_{ji}), \qquad i = 1, \dots, n_i \\
\theta_{ji} \mid \widetilde{\mu}_j &\overset{\text{iid}}{\sim} \widetilde{\mu}_j / \widetilde{\mu}_j(\Theta), \qquad i = 1, \dots, n_i \\
\widetilde{\mu}_j &:= \sum_{h=1}^{H} \lambda_{jh} \mu_h^*, \\
(\mu_1^*, \dots, \mu_H^*) &\sim \text{CoRM}(\phi, \nu^*, \alpha), \qquad \Lambda \sim \pi(\Lambda).
\end{aligned}
\tag{8}
$$

In this section, we describe a simple MCMC scheme based on a truncation of the random measures. While we find the truncation convenient for algorithmic purposes, this is not necessary to fit the model. Indeed, in Section C of the Supplementary Material, we also describe a slice sampling algorithm based on Griffin and Walker (2011) that does not require truncating the random measure. In particular, let $K > 0$ denote a fixed number of atoms, we set

$$
\mu_h^* = \sum_{k=1}^{K} m_{hk} J_k \delta_{\theta_k^*}
$$

where $J_k \overset{\text{iid}}{\sim} p_J$, with p_J being a probability distribution, and $\theta_k^* \overset{\text{iid}}{\sim} G_0 := \alpha/\alpha(\Theta)$. Campbell et al. (2019) provide a thorough review of truncation methods for completely random measures including the choice of p_J for different random measures. We use $p_J = \text{Beta}(\phi/K, \phi)$ so that $\sum_{k=1}^{K} J_k \delta_{\theta_k^*}$ converges to a Beta process as $K \to +\infty$. This combined with gamma-distributed m_{hk} imply that marginally μ_h^* follows a gamma process (see Griffin and Leisen, 2017). Although this simple truncation might result in an approximation error that is large a priori, as shown in Nguyen et al. (2020), posterior inference is usually robust and no significant difference is detected. The choice of fixing K also allows for (much) faster code since the number of parameters is now fixed, and our implementation can thus take advantage of modern parallelization and vectorization algorithms. This is in line with our ultimate goal of fitting very large datasets with our model.

3.1. MCMC Algorithm for the Truncated Model

Observe that in (8), $\theta_{ji} = \theta_k^*$ with positive probability. Therefore an alternative representation is achieved by introducing latent cluster indicator variables c_{ji} such that c_{ji} are independent categorical variables with support $\{1, \dots, K\}$ and

$$
P(c_{ji} = k \mid \{\lambda_{jh}\}, \{m_{hk}\}, \{J_k\}) \propto (\Lambda M)_{jk} J_k.
$$

Let $T_j := \sum_k (\Lambda M)_{jk} J_k$. Writing $p(\cdot \mid \cdot)$ for a generic conditional density, the joint distribution of data and parameters under (8) is then

$$
p(\{y_{j,i}\}, \{c_{j,i}\}, \{\lambda_{j,h}\}, \{m_{h,k}\}, \{J_\ell\}, \{\theta_\ell^*\}) = \\
\prod_{j=1}^{g} T_j^{-n_j} \prod_{i=1}^{n_j} f(y_{j,i} \mid \theta^*_{c_{j,i}}) (\Lambda M)_{j, c_{j,i}} J_{c_{j,i}} \times \prod_{h=1}^{K} \left[G_0(\theta_h^*) p_J(J_k) \prod_{k=1}^{K} \text{Ga}(m_{hk} \mid \phi) \right] \pi(\Lambda).
$$

To facilitate posterior inference, we introduce a set of auxiliary variables u_j, which are gamma distributed with shape parameter T_j and rate parameter n_j. Then

$$
\begin{aligned}
&p(\{y_{j,i}\}, \{c_{j,i}\}, \{\lambda_{j,h}\}, \{m_{h,k}\}, \{J_\ell\}, \{\theta^*_\ell\}, \{u_j\}) = \\
&\quad \prod_{j=1}^{g} \frac{1}{\Gamma(n_j)} u_j^{n_j-1} \prod_{i=1}^{n_j} f(y_{j,i} \,|\, \theta^*_{c_{j,i}})(\Lambda M)_{j,c_{j,i}} J_{c_{j,i}} \times \exp\left(- \sum_{j=1}^{g} u_j \sum_{\ell=1}^{K} (\Lambda M)_{j,\ell} J_\ell \right) \\
&\quad \prod_{h=1}^{K} \left[G_0(\theta^*_h) p_J(J_k) \prod_{k=1}^{K} \mathrm{Ga}(m_{hk} \,|\, \phi) \right] \pi(\Lambda)
\end{aligned}
$$

It is then possible to sample from the posterior distribution via a Gibbs sampler:

(a) Update the atoms from

$$
p(\theta^*_h \,|\, \cdots) \propto \prod_{j=1}^{g} \prod_{i:c_{j,i}=h} f(y_{j,i} \,|\, \theta^*_h) G_0(\theta^*_h)
$$

(b) Update the J's from

$$
p(J_\ell \,|\, \cdots) \propto J_\ell^{q_\ell} \exp\left(- \sum_{j=1}^{g} u_j (\Lambda M)_{j,\ell} J_\ell \right) p_J(J_\ell)
$$

where $q_\ell = \sum_{j=1}^{g} \sum_{i=1}^{n_j} I[c_{j,i} = h]$.

(c) Update the m's from

$$
p(M \,|\, \cdots) \propto \prod_{j=1}^{g} \prod_{\ell=1}^{K} (\Lambda M)_{j,\ell}^{q_\ell} \times \exp\left(- \sum_{j=1}^{g} u_j (\Lambda M)_{j,\ell} J_\ell \right) \times \prod_{h=1}^{H} \prod_{k=1}^{K} \mathrm{Ga}(m_{hk} \,|\, \phi)
$$

The update of M can be done in a single block via Hamiltonian Monte Carlo.

(d) Update the λ's from

$$
p(\Lambda \,|\, \cdots) \propto \prod_{j=1}^{g} \prod_{\ell=1}^{K} (\Lambda M)_{j,\ell}^{q_\ell} \times \exp\left(- \sum_{j=1}^{g} u_j (\Lambda M)_{j,\ell} J_\ell \right) \pi(\Lambda)
$$

Again, we can update Λ using a single step of Hamiltonian Monte Carlo.

(e) Update the cluster indicators from a categorical distribution over $\{1, \ldots, K\}$ with weights

$$
P(c_{j,i} = h \,|\, \cdots) \propto f(y_{j,i} \,|\, \theta^*_h)(\Lambda M)_{j,h} J_h
$$

(f) update the u's from $p(u_j \,|\, \cdots) \propto u_j^{n_j} e^{-T_j u_j}$, where we recognize the kernel of a $\mathrm{Gamma}(n_j, T_j)$ random variable

Finally, when the prior for Λ is the multiplicative gamma process (3) we propose to gain computational efficiency by selecting H through an adaptive Gibbs sampling scheme as in Bhattacharya and Dunson (2011). In particular, when adaptation occurs, we look at the "empty columns" of Λ. We define a column h of Λ to be empty if

$$
\sum_{j=1}^{g} \frac{\lambda_{jh}}{\sum_{k=1}^{H} \lambda_{jk}} < \varepsilon \bar{\lambda}
$$

where $\bar{\lambda} = H^{-1} \sum_{h=1}^{H} \sum_{j=1}^{g} \frac{\lambda_{jh}}{\sum_{k=1}^{H} \lambda_{jk}}$. In our experience $\varepsilon = 0.05$ provides satisfactory results. If there are no empty columns, we add a column sampled from the prior to Λ and a row sampled from the prior to M. Instead, if empty columns are found, we drop them from Λ and the corresponding rows from M.

Bhattacharya and Dunson (2011) propose to adapt Λ at each iteration ℓ with a probability p_ℓ that decreases exponentially fast. This choice is possible also within our algorithm but, in our experience, it significantly impacts run-time. This is due to the choice of using HMC to sample Λ and M and, in particular, to the use of the `tensorflow-probability` Python package, in combination with `LAX` compilation. For technical reasons, every time the size of Λ and M change, big chunks of the code must be recompiled, so that it's not efficient to adapt every few iterations. Instead, we propose to have a fixed adaptation window of $1,000$ iterations, where the adaptation occurs every 50 iterations. In our experience, this simple modification reduces the overall runtime by at least one order of magnitude.

4. Resolving the non-identifiability via post-processing

As already mentioned in the introduction, our model is not identifiable due to the multiplicative relation between Λ and $(\mu_1^*, \ldots, \mu_h^*)$. This is not surprising, as the same holds for common latent factor models (Geweke and Singleton, 1980), where the likelihood is invariant to the action of orthogonal matrices. In that context, a common practice to recover identifiability is to constrain the matrix Λ to be lower triangular with positive entries on the diagonal (Geweke and Zhou, 1996). More recently, it has been proposed to ignore the identifiability issue and obtain a point-estimate of the posterior distribution either by post-processing the MCMC chains (see Papastamoulis and Ntzoufras, 2022; Poworoznek et al., 2021, and the references therein) or by choosing the maximum a posteriori (Schiavon et al., 2022). In particular, Poworoznek et al. (2021) propose to orthogonalize each posterior sample of Λ and then solve the sign ambiguity and label switching via a greedy matching algorithm.

The non-identifiability in our model is more severe than the one of common latent factor models. In fact, for any Q s.t. Q^{-1} is well defined, the likelihood is invariant when considering $\Lambda' = \Lambda Q^{-1}$ and $M' = QM$. Nonetheless, the constraints that $\Lambda' \geq 0$ (element-wise) and $M' \geq 0$ greatly reduce the number of matrices Q that can cause non-identifiability. In particular, we don't need to worry about sign ambiguity.

4.1. The Objective Function

Consider equation (2). Factorizations of the kind $\Gamma = \Lambda M$ where all the three matrices have nonnegative entries are common in blind source separation (BSS) problems, where the goal is to estimate "source components" M and "mixing proportions" Λ such that the observed signal Γ is approximately ΛM. Two well-established approaches to BSS are nonnegative matrix factorization (NMF, Sra and Dhillon, 2005) and independent component analysis (ICA, Hyvärinen, 2013). The main difference between the two consists in the loss function optimized. In NMF it is usually the norm of the approximation error, while, in ICA, the mutual information between the source components is minimized alongside the approximation error. This takes into account the goal of separating the components. Since in our analogy the sample size of the latent factor model is just one (i.e., in our model there is one single vector $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_p$ instead of multiple realizations), it is not possible to use the same criteria of ICA to define what we mean by "separated components". Hence, we propose to optimize with respect to the following *interpretability* criterion:

$$L(Q; M, J, \theta) = \sum_{i<j} \left(\int_{\mathbb{Y}} \left[\int_{\Theta} f(y \mid \theta) \mu_i'(\mathrm{d}\theta) \right] \left[\int_{\Theta} f(y \mid \theta) \mu_j'(\mathrm{d}\theta) \right] \mathrm{d}y \right)^2 . \qquad (9)$$

where

$$\mu'_j = \sum_{k=1}^{K} (QM)_{jk} J_k \delta_{\theta^*_k}$$

Low values of $L(Q; M, J, \theta)$ in (9) are attained when the transformed random measures μ'_h, mixed with the mixture kernel f, result in well separated densities. Indeed, note that, defining $g_i(y) := \int_\Theta f(y \mid \theta)\mu'_i(\mathrm{d}\theta)$ it is clear that (9) can be interpreted as the sum of the squared inner products (in the L_2 sense) between g_i and g_j. Since the g_i's are positive functions, low values of the inner products can be obtained only if μ'_i and μ'_j give mass to different portions of the domains. The L_2 distance is not commonly used to measure the discrepancy of densities. A more familiar option would be to consider $\int \sqrt{g_i(y)}\sqrt{g_j(y)}\mathrm{d}y$, that is $1 - d_{\mathcal{H}}(g_i, g_j)$ where $d_{\mathcal{H}}$ denotes the Hellinger distance. However, this choice of loss function leads to a more complex optimization problem, that cannot be solved with our approach. Indeed, as discussed later in Section 4.3, the positivity of the g_i's might not be preserved by the intermediate steps of the algorithm. Therefore, we need a loss function that continues to make sense for negative g_i's.

4.2. The Optimization Space

Consider now the space over which one should minimize (9). First of all, we must require the existence of Q^{-1} to interpet $\Lambda' = \Lambda Q^{-1}$. Moreover, for the model to make sense we need to ensure the positivity of the coefficients involved, i.e. $\Lambda' = \Lambda Q^{-1} \geq 0$ and $M' = QM \geq 0$. Finally, we observe that (i) given an "optimal" Q such that $L(Q; M, J, \theta) = 0$, $L(\gamma Q; M, J, \theta) = 0$ for any $\gamma > 0$, and (ii) $L(Q; M, J, \theta)$ attains lower values when the entries in Q are small. Despite the preference for small Q in the optimization problem, the resulting model is invariant to such rescalings since it involves the normalization of the underlying random measures. Hence, to overcome both issues we propose to add a further constraint in the optimization problem, namely $\det Q = 1$, which prevents having several optimal solutions differing by a constant and does not allow for matrices with entries too close to 0.

In conclusion, we propose to optimize (9) over the special linear group $SL(H) = \{Q \in \mathbb{R}^{H \times H} : \det Q = 1\}$, with the additional positivity constraints, i.e. our optimization problem becomes

$$\min_{Q \in SL(H)} \sum_{h,k=1}^{H} L(Q; M, J, \theta) \quad \text{s.t.} \quad \Lambda Q^{-1} \geq 0, \ QM \geq 0. \tag{10}$$

The special linear group is not a linear space, therefore common gradient-based optimization techniques cannot be used to solve (10). However, we can take advantage of the differential structure of $SL(H)$. In fact, it is a Lie group (hence, a smooth differentiable manifold) with associated Lie algebra $\mathfrak{sl}(H) = \{A \in \mathbb{R}^{H \times H} : \mathrm{tr} A = 0\}$. See Section A.2 of the Supplementary Material for some basic details regarding Riemannian manifolds and Lie groups.

4.3. A Riemannian Augmented Lagrangian Method

We are now in place to state the algorithm. For notational convenience, define the functions $c^1_{jh}(Q) = -(\Lambda Q^{-1})_{jh}$ and $c^2_{hk} = -(QM)_{hk}$. Denote with c_j the collection of all such functions. The positivity constraints are equivalent to $c_j \leq 0$ for all j's. Following the augmented Lagrangian method (Birgin and Martínez, 2014), we can deal with the constraints $\Lambda Q^{-1} \geq 0$ and $QM \geq 0$ by introducing auxiliary parameters ρ, γ_j and define the

Algorithm 1. Augmented Lagrangian Multiplier Method

```
input Starting point $Q$, initial values $\rho$, $\gamma_j$, target threshold $\varepsilon^*$, initial threshold $\varepsilon$.
repeat
    $Q = Q'$
    solve $Q' = \arg\min_Q \mathcal{L}_\rho(Q, \gamma)$ for fixed $\rho, \gamma$ with theshold $\varepsilon$ using Algorithm 2
    $\gamma_j = \gamma_j + \rho c_j(Q')$
    $\rho = 0.9\rho \; \varepsilon = \max\{\varepsilon^*, 0.9\varepsilon\}$
until $\varepsilon \leq \varepsilon^*$; $\|Q - Q'\| \leq \varepsilon$
end
```

Algorithm 2. Lie RATTLE Optimization

```
input Starting point $Q, P$, momentum $\tau$, stepsize $s$, threshold $\varepsilon$.
repeat
    $P = \tau\left(P - s\Pi_{\mathfrak{sl}(H)}(\partial_Q \mathcal{L}_\rho(Q, \gamma), Q)\right)$
    $Q = Q \exp_m(\chi P)$, $\chi = \cosh(-\log \tau)$
    $P = \tau\left(P - s\Pi_{\mathfrak{sl}(H)}(\partial_Q \mathcal{L}_\rho(Q, \gamma), Q)\right)$
until $\|Q - Q'\| \leq \varepsilon$
end
```

augmented loss function

$$\mathcal{L}_\rho(Q, \gamma) = L(Q; M, J, \theta) + \frac{\rho}{2} \sum_j \max\left\{0, \frac{\gamma_j}{\rho} c_j(Q)\right\} \tag{11}$$

Then, we can solve (10) by alternating between minimizing (11) for fixed values of ρ, γ_j and updating ρ, γ_j as in Algorithm 1. As in the usual augmented Lagrangian method, the constraints might be violated in the intermediate steps. Intuitively, the fact that the penalty term γ_j is increased at every iteration if the constraint is violated should force the solution of the problem inside the feasible region. See Birgin and Martínez (2014) for convergence results of the augmented Lagrangian method.

It is now left to discuss how to solve (11) for fixed ρ and γ_j. We propose to tackle this problem with the Riemannian dissipative RATTLE algorithm in França et al. (2021), reported for the special case of optimization over $SL(H)$ in Algorithm 2. In particular, $\Pi_{\mathfrak{sl}(H)}$ is the projection over the Lie algebra $\mathfrak{sl}(H)$ while $\exp_m$ denotes the matrix exponential, which is a map $\mathfrak{sl}(H) \to SL(H)$. Informally, Algorithm 2 resembles an accelerated gradient method, where a momentum term is introduced to speed up the convergence. We further have

$$\partial_Q \mathcal{L}_\rho(Q, \gamma)_{ij} = \frac{\partial_Q \mathcal{L}_\rho(Q, \gamma)}{\partial Q_{ji}}$$

(note the index flip $ij \to ji$, in other words $\partial_Q f(Q) = \nabla_Q f(Q)^\top$ where ∇ stands for the usual Euclidean gradient). Moreover, the following proposition gives a computationally convenient way of evaluating $\Pi_{\mathfrak{sl}(H)}$.

Proposition 5. *Let X an $H \times H$ real valued matrix. Then*

$$\Pi_{\mathfrak{sl}(H)}(X) = (X - diag(X))^T + \sum_{\ell=1}^{H-1} X_\ell^*$$

where $diag(X)$ is the diagonal matrix with entries equal to the diagonal of X and X_ℓ^* is a diagonal matrix whose only nonzero entries are the (ℓ, ℓ)-th and the $(\ell + 1, \ell + 1)$-th ones, which equal to $X_{i,i} - X_{i+1,i+1}$ and $-X_{i,i} - X_{i+1,i+1}$ respectively.

The parameters involved in the optimization problem are: the stepsize s and momentum factor τ in Algorithm 2 as well as the initial values ρ, γ_j and the target and thresholds ε^*, ε in Algorithm 1. We suggest as defaults $s = 10^{-6}$, $\tau = 0.9$, $\rho = \gamma_j = 10$, $\varepsilon^* = 10^{-6}$, $\varepsilon = 10^{-2}$. Finally, to set the starting point Q we we solve the unconstrained optimization problem (equivalent to setting $\gamma_j = 0$ in (11)) using Algorithm 2 and use that solution as starting point for the constrained optimization. The initial momentum term P in Algorithm 2 is always the zero matrix.

Convergence of the augmented Lagrangian method is difficult to study, particularly so in the case of optimization on Riemannian manifolds. Given that the objective function is nonconvex, we con only establish that our algorithm converges to a local minimizer of (9). See Liu and Boumal (2020) for further discussions

4.4. The Label-Switching Problem

Observe that another source of non-identifiability comes from the labeling of $\mu_1^*, \dots, \mu_H^*$. Namely, the likelihood and the loss function (9) are invariant under permutation of the indices$\{1, \dots, H\}$, provided that the columns of Λ are permuted as well. This prevents the possibility of computing reliable posterior summaries of the μ_h^*'s and Λ from the MCMC chains.

We propose to post-process the output of our sampling algorithm to get rid of this problem. In particular, as in Poworoznek et al. (2021), we propose to align the latent measures at each iteration to a given template. Let $\hat{\mu}_1, \dots, \hat{\mu}_H$ denote the template. For instance,

$$\hat{\mu}_h = \sum_{k=1}^{K} (Q^{(\ell)} M^{(\ell)})_{jk} J_k^{(\ell)} \delta_{\theta_k^{(\ell)}}$$

where we denote with the superscript ℓ the index of the MCMC sample. We choose ℓ to approximate the maximum a posteriori. $Q^{(\ell)}$ denotes the associated optimal transformation matrix obtained as outlined above. Let $d(\hat{\mu}_h, \mu_j')$ denote a dissimilarity between two measures. Two specific choices are discussed later. We align each $(\mu_1^{\prime(j)}, \dots, \mu_H^{\prime(j)}) := Q^{(j)}(\mu_1^{*(j)}, \dots, \mu_H^{*(j)})$ to $\hat{\mu}_1, \dots, \hat{\mu}_H$ by learning an optimal permutation σ of $\{1, \dots, H\}$, associated to a permutation matrix P_σ that minimizes $\sum_h d(\hat{\mu}_h, \mu_{\sigma(h)}^{(j)\prime})$ by solving

$$\inf_{P \in \mathrm{Perm}_H} \sum_{h,k=1}^{H} d(\hat{\mu}_h, \mu_k^{(j)\prime}) P_{hk}$$

where Perm_H denotes the space of $H \times H$ permutation matrices. Naively, this would require $H!$ computations. Instead, we solve the relaxed optimization problem by looking for the P stochastic matrix (i.e., rows and columns sum to one) that minimizes the objective above. That is, we solve for the Wasserstein distance between the empirical measures ν_1 and ν_2 defined as

$$\nu_1 = \frac{1}{H} \sum_{h=1}^{H} \delta_{\hat{\mu}_h}, \qquad \nu_2 = \frac{1}{H} \sum_{k=1}^{H} \delta_{\mu_k^{(j)\prime}}$$

where ν_i is a probability measure on the space of positive measures over Θ. Birkhoff's theorem ensures that the solution to the relaxed optimization problem is a permutation matrix.

As far as the dissimilarity $d(\hat{\mu}_h, \mu_j')$ is concerned, in our examples we considered

$$d(\hat{\mu}_h, \mu_j') = \left\| \hat{\mu}_h(\Theta)^{-1} \int_\Theta f(y \mid \theta) \hat{\mu}_h(\mathrm{d}\theta) - \mu_j'(\Theta)^{-1} \int_\Theta f(y \mid \theta) \mu_j'(\mathrm{d}\theta) \right\|$$

where $\| \cdot \|$ stands for the L_2 norm. This distance requires the numerical evaluation of a mixture density on a fixed grid, to compute the associated L_2 distance. This is easy when the dimension of the data space is small, typically when data are uni or bi-dimensional. See Section D of the Supplementary Material for a more efficient alternative in higher dimensions.

5. Simulation Study

We present two simulations to assess the performance of our model. In all the examples, we consider Gaussian mixture models, i.e., $\theta^*_h = (\mu_h, \sigma^2_h)$ and $f(\cdot \mid \theta) = \mathcal{N}(\cdot \mid \mu, \sigma^2)$. The scores m_{hk} in the CoRM are gamma distributed and each μ^*_h is marginally a gamma process (before the truncation) with total mass equal to 1 and base measure equal to the Normal-inverse-Gamma distribution, i.e. $G_0(\mu, \sigma^2) = \mathcal{N}(\mu \mid \mu_0, \sigma^2/\lambda) IG(\sigma^2 \mid a, b)$. We set μ_0 equal to the empirical mean of the observations, $\lambda = 0.01$, $a = b = 2$. We truncate the CoRM to $K = 20$ jumps to perform posterior inference. Specific choices of the prior for Λ are discussed case-by-case.

5.1. *Interpretation of the posterior distribution*

Before giving details on the numerical illustration, we discuss how to obtain interpretable summaries of the posterior distribution, after post-processing. This also allows us to set some notation used in the next sections.

Interpreting the unnormalized *latent factor densities* $\int_\Theta f(\cdot \mid \theta) \mu^*_h(\mathrm{d}\theta)$ is difficult because of the lack of a common scale to which the densities should be referred. In fact, note that these are not probability densities. Let p_j be the j-th group-specific density. We can write

$$p_j = \int_\Theta f(\cdot \mid \theta) \bar{\tilde{p}}(\mathrm{d}\theta) + \sum_{h=1}^{H} s_{jh} \int_\Theta f(\cdot \mid \theta) \epsilon_h(\mathrm{d}\theta)$$

where $\bar{\tilde{p}}(\mathrm{d}\theta)$ is the average of $\tilde{p}_1, \ldots, \tilde{p}_g$, $p'_h = \mu'_h / \mu'_h(\Theta)$, $\epsilon_h = p'_h - \bar{\tilde{p}}(\mathrm{d}\theta)$ and the scores s_{jh}'s are defined as

$$s_{jh} = \frac{\lambda'_{jh} \mu'_h(\Theta)}{\sum_{k=1}^{H} \lambda'_{jk} \mu'_k(\Theta)} \tag{12}$$

Note that ϵ_h is a signed measure. Instead of comparing the latent factor densities, we find it considering the *residual factor densities* $\int_\Theta f(\cdot \mid \theta) \epsilon_h(\mathrm{d}\theta)$ leads to easier interpretations. Note that, by definition, the residual factor densities might be negative.

Moreover, we can associated to each μ'_h an *importance score* I_h defined as $I_h = \sum_{j=1}^{g} s_{jh}$ The rationale comes from writing $\mu'_h = \mu'_h(\Theta) p'_h$ so that

$$p_j = \int_\Theta f(\cdot \mid \theta) \sum_{h=1}^{H} \frac{\lambda'_{jh} \mu'_h(\Theta)}{\sum_{k=1}^{H} \lambda_{jk} \mu'_k(\Theta)} p'_h(\mathrm{d}\theta) = \sum_{h=1}^{H} s_{jh} \int_\Theta f(\cdot \mid \theta) p'_h(\mathrm{d}\theta)$$

that is, we express each $\tilde{p}_j$ as a convex combination of probability measures and with weight s_{jh}.

As far as posterior summaries of the μ^*_h's and Λ are concerned, we first perform the post-processing of the MCMC chains described in Section 4. Then we define the posterior point estimates μ'_h and Λ' as

$$\mu'_h = \frac{1}{M} \sum_{\ell=1}^{M} \sum_{k \geq 1} \left(P^{(\ell)} Q^{(\ell)} M^{(\ell)} \right)_{hk} J^{(\ell)}_k \delta_{\theta^{*(\ell)}_k}, \qquad \Lambda' = \frac{1}{M} \sum_{\ell=1}^{M} \left(\Lambda^{(\ell)} (Q^{(\ell)})^{-1} \right) (P^{(\ell)})^\top \tag{13}$$

where the superscript ℓ, $\ell = 1, \dots, M$ is used to denote the iteration of the MCMC algorithm, $Q^{(\ell)}$ is the matrix found with Algorithm 1, and $P^{(\ell)}$ is the permutation matrix found as in Section 4.4. Essentially, the posterior point estimates are obtained by first collapsing the MCMC draws from the multimodal posterior around one specific mode, and then averaging these transformed draws.

5.2. Only Group Information

We consider here a simulated example with $g = 100$ groups of data, where each $n_j = 25$. We consider the situation where we tend to observe only small differences across populations by considering the following data generation process

$$y_{j,i} \overset{\text{iid}}{\sim} w_{j1} \mathcal{N}(-2, 2) + w_{j2} \mathcal{N}(0, 2) + w_{j3} \mathcal{N}(2, 2), \qquad i = 1, \dots, n_j$$

and for each group we simulate $\boldsymbol{w}_j = (w_{j1}, w_{j2}, w_{j3}) \overset{\text{iid}}{\sim} \text{Dirichlet}(1, 1, 1)$. In most of the groups, the data generating density is unimodal and they differ mainly because of different levels of skewness.

As prior for Λ, we assume the multiplicative gamma process (3) setting $H = 20$. We run the MCMC chains for a total of $11,000$ of which the first $1,000$ are used for the adaptation and the following $5,000$ are discarded as burn-in. The adaptation phase quickly finds between 3 and 5 latent measures, 4 being the final value. We post-process the chains as in Section 4.

Figure 3 shows the inferred latent factors densities before and after the post-processing. It is clear that solving the label switching is essential. Although not particularly evident from the plot, the matrices $Q^{(j)}$ found by the optimization algorithm were significantly different from the identity, hence showing the usefulness of the post-processing. Our approach identifies the main common traits in the data. Factors 1 and 3 peak around -2 and 2 respectively, while the second and fourth factors are both more concentrated around the origin, with the second one presenting a light skewness and heavier right tail. The residual factor densities can be used to infer the same description of the latent measures. With the exception of finding one extra latent factor, our model infers exactly the sources of variability in the data generating process, which are the three Gaussian densities centered in -2, 0, and $+2$ respectively.

We also compare the fit of our model to other possible competitors. As two standard baselines, we consider (i) pooling all the data together, i.e., disregarding the group information and (ii) fitting a simple Dirichlet process (DP) mixture to the whole dataset, and fitting a separate DP mixture to each of the groups independently. Then, we consider (iii) the GM-dependent DP in Lijoi et al. (2014) implemented in the R package `BNPMix` (Corradin et al., 2021), and (iv) the CoRM model with Gamma marginals in Griffin and Leisen (2017) and we use the code from Camerlenghi et al. (2022) to fit the model. As last competitor, we consider (v) a dependent stick-breaking process defined as

$$\widetilde{p}_{j,i} := \sum_{k=1}^{K} w_k(\boldsymbol{x}_{j,i}) \delta_{\theta_k^*}$$

where all the atoms are shared (as in our model) and the weights follow a logit stick-breaking process prior (Rigon and Durante, 2021) where we associate to each observation a covariate $\boldsymbol{x}_{j,i} \in \mathbb{R}^{100}$ such that $x_{j,i,h} = 1$ if and only if observation $y_{j,i}$ belongs to group j. To fit this model, we used the `C++` library `BayesMix` (Beraha et al., 2022) and fix $K = 20$.

We report the boxplots of the Kullback-Leibler divergences between the true data generating densities and the estimated ones (across the 100 groups of data) for all the models under analysis in Figure 4. Our model achieves the best performance with the GM-DP

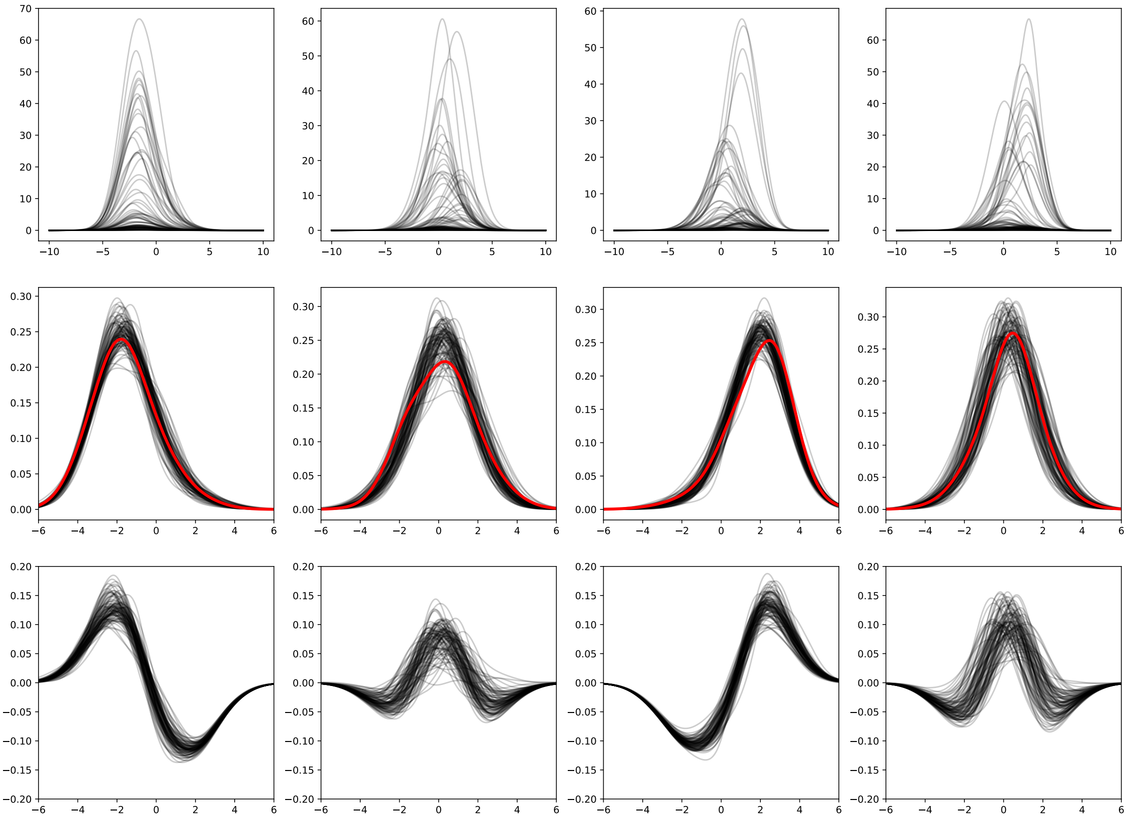

Fig. 3: Posterior summaries for the simulation in Section 5.2. Top row: draws from the posterior distribution of the latent factor densities. Middle row: draws after post-processing and normalization, the red density denotes the template. Bottom row: posterior draws of the residual factor densities.

being a close runner-up. As expected, pooling together all the data and fitting each group independently produces the worst performance. We did not thoroughly investigate the runtimes, since each method is implemented in a different programming language and with different programming techniques. We limit ourselves to report that the dependent Dirichlet process is the slowest to fit, while the single DP with pooled data and the independently fitted DPs are the fastest, also due to the heavy optimizations of the `BayesMix` library. Our method is slightly slower than the GM-DP (but this is implemented in `C++` while ours in pure `Python`) and an order of magnitude faster than the CoRM.

5.3. Area-Referenced Data

We consider data over a regular lattice on $0, 1, \dots, q \times 0, 1, \dots, q \subset \mathbb{Z}^2$. We consider $q = 4, 8, 16$ so that the number of groups is $g = 16, 64, 256$ respectively. Following the simulation study in Beraha et al. (2021), we generate data at each location from a three-component Gaussian mixture with means $-5, 0, 5$ respectively and variances equal to one. Let x_j, y_j denote the x and y coordinate of location j on the lattice. The location-specific weights are

$$(w_{j1}, w_{j2}, w_{j3}) = \left(e^{\widetilde{w}_{j1}}, e^{\widetilde{w}_{j2}}, 1\right) / \left(1 + e^{\widetilde{w}_{j1}} + e^{\widetilde{w}_{j2}}\right)$$

where

$$\widetilde{w}_{j1} = 3(x_j - \bar{x}) + 3(y_j - \bar{y}), \quad \widetilde{w}_{j2} = -3(x_j - \bar{x}) - 3(y_j - \bar{y})$$

and $(\bar{x}, \bar{y})$ denote the center of the lattice. For each location, 25 observations are simulated.

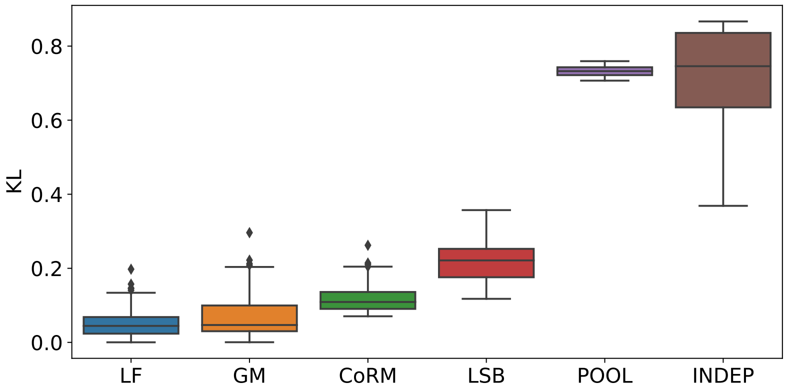

Fig. 4: Boxplots of KL divergences between the true data generating densities and the estimated ones in the 100 groups of data of the simulation in Section 5.2. From left to right, ordered by the median error, our latent factor model (LF) the GM-dependent DP (GM), the compound random measure (CoRM), the dependent Dirichlet process based on the logit stick-breaking prior (DDP), the estimate obtained pooling together all the groups (POOL), and the one obtained fitting independently each group (INDEP).

We compare our model with prior (4) for $H = 1, 2, 3, 5, 10$ with the spatially dependent mixture model (SPMIX, Beraha et al., 2021) and the Hierarchical Dirichlet Process (HDP, Teh et al., 2006). Although the latter does not take into account the spatial dependence, it is shown in Beraha et al. (2021) that the HDP performs well when the number of groups g is small.

We truncate the CoRM to $K = 20$ jumps and set the number of components in SPMIX to 20 as well. Prior distributions can be assumed for τ and ρ in (4). However, since the likelihood is invariant with respect to rescalings of Λ, we found that having a prior on τ led to non-convergent MCMC chains for Λ. In particular, after a few thousand iterations, the values of the entries in Λ were in the order of 10^{100}. Hence, we suggest fixing τ to a sufficiently large value. In our simulations, we always set $\tau \equiv 2.5$. Assuming a prior for ρ does not have such an impact on posterior inference. However, it would require re-computing the determinant of Σ^{-1} at every MCMC iteration, which requires $O(g^3)$ operations. Hence, we fix ρ to 0.95 to encourage strong spatial dependence in our examples. Another possibility would be to fix a grid of values in $(0, 1)$ and assume a discrete prior for ρ over it, which allows the computation of all the required matrix determinants beforehand.

All the MCMC chains are run for $10,000$ iterations, discarding the first $5,000$ as burn-in. It is clear from Figure 5 (top row) that our model outperforms the competitors when $g = 16, 64$ and performs slightly better than the spatial mixture model when $g = 256$. In all the settings, the best performance is associated with $H = 3$ latent measures. Posterior samples of the latent factor densities are reported in Figure 5 (bottom row) for the setting with $g = 64$ and $H = 3$. In this case, the latent densities are already well separated so that there is no need to post-process the MCMC chains using the algorithm described in Section 4. The three latent densities give mass to one of the three modes in the data each.

6. Real Data Illustrations

In this section, we illustrate our methodology on two real datasets. In both cases, data are univariate and we let $f(\cdot \mid \theta)$ be the Gaussian density with parameters $\theta = (\mu, \sigma^2)$. The base measure G_0 is the Normal-inverse-Gamma distribution, whose parameters are set as in Section 5. Moreover, we always truncate to $K = 20$ points the support of the random

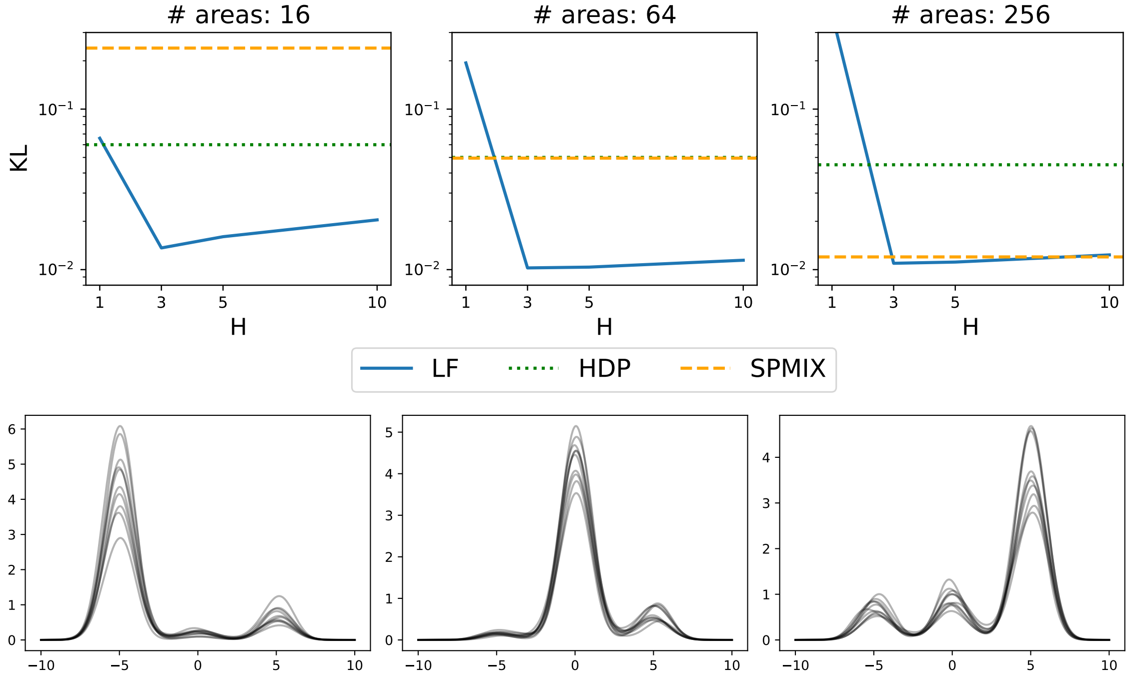

Fig. 5: Top row: Average Kullback–Leibler divergence between the true data generating density and the Bayesian estimate, as a function of the number of latent measures H, for our model (LF) the Hierarchical Dirichlet process (HDP) and the spatially dependent mixture models (SPMIX). From left to right $g = 16, 64, 256$. Bottom row: Posterior samples for the latent factor densities when $g = 64$ and $H = 3$

measures.

6.1. The Invalsi Dataset

We consider the *Invalsi* dataset† that collects the evaluation of a unified math test undertaken by all Italian high-school students. Grades vary from 1 to 10 with 6 being the passing grade. We pre-process the data by adding a small Gaussian noise with zero mean and standard deviation equal to 0.25. The dataset contains the scores of 39377 students, subdivided into 1048 schools. The number of students per school varies from 4 to 131, with 37 students per school on average with a standard deviation of 12 approximately. Another approach to model this kind of data would be to assume another mixture kernel (other than the normal one) which corresponds to the probability mass function of a random variable taking values $\{1, \ldots, 10\}$. For instance, we could adopt the method in Canale and Dunson (2011), whereby a Gaussian mixture is assumed for latent variables (one for each observation). However, our main interest here is not density modelling but rather explaining the difference in distribution across different schools, and we believe that posterior inference on the latent measure factors would remain qualitatively unchanged using this more complex approach.

We assume the multiplicative gamma process prior for Λ as in (3) with $H = 20$. The initial adaptation phase identifies 5 latent factors. Draws from the latent factor densities are displayed in Figure 6. It is clear that some label switching is happening between the fourth and fifth factors. After the post-processing, for ease of visualization, we discretized the estimated normalized latent factor densities to the original grades $i = 1, \ldots, 10$ by

†available for research purposes at https://invalsi-serviziostatistico.cineca.it

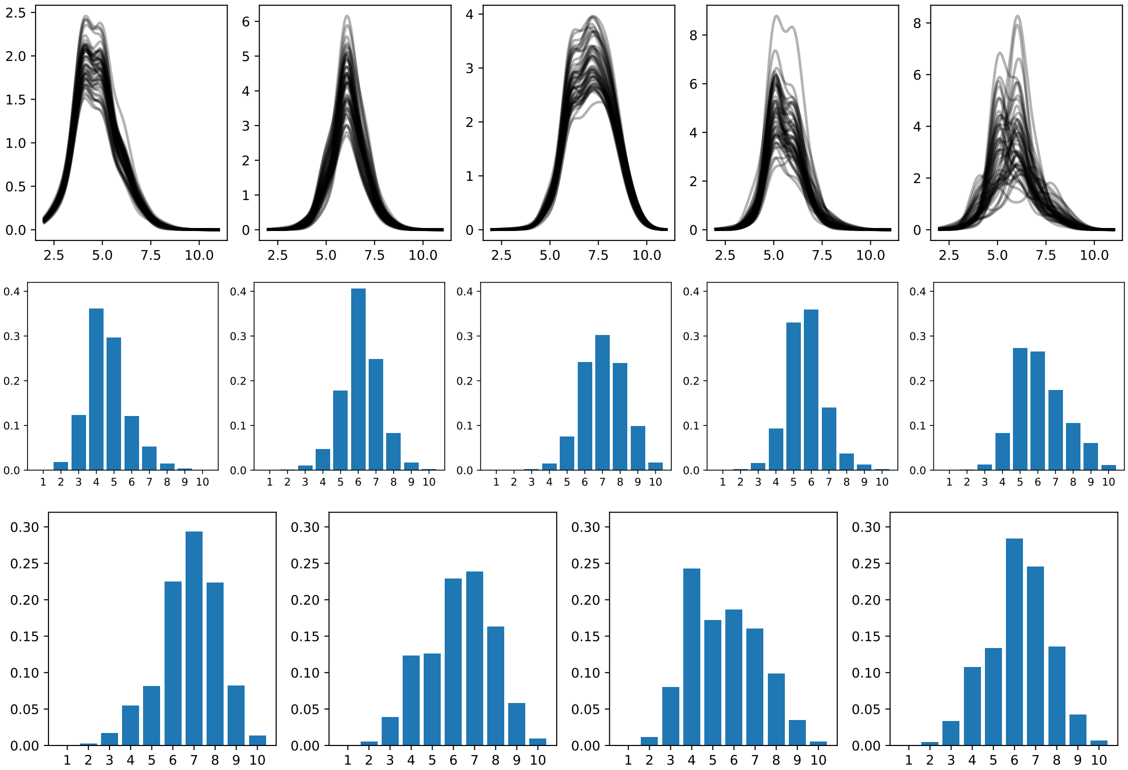

Fig. 6: Summary of posterior inference on the Invalsi dataset. Top row: draws from the posterior distribution of the latent factor densities. Middle row: estimates of the discretised normalised latent factor densities after post-processing. Bottom row: average density in each cluster discretised on the intervals $[i - 0.5, i + 0.5)$, $i = 1, \ldots, 10$.

evaluating $\int_{i-0.5}^{i+0.5} f(y \mid \theta)\mu'_h(\mathrm{d}\theta)/\mu'_h(\Theta)$. The estimated factors are displayed in the first two rows of Figure 6. They represent a wide range of behaviors: the first one is concentrated on negative grades below the passing threshold, the second one is centered on the passing grade, and the third one on grades way above the passing grade. The fourth and the fifth represent more complex distributions: the former one covering the range of "just below the passing grade and just above it", the latter one instead represents a distribution peaked at 5 with a heavy right tail.

The importance scores I_h are approximately $331, 184, 351, 165, 16$. Hence, we can interpret that the two most relevant common traits are the ones represented by μ'_1 (that combines a sharp peak in 4, with a heavy right tail), and by μ'_3, which gives mass to grades above the passing threshold.

Finally, we look at the scores λ_{jh}'s after the post-processing. We can understand the similarities between schools by clustering the scores for each school from the corresponding row of the matrix Λ'. Using a hierarchical clustering algorithm yields four clusters (the dendrogram is shown in Figure E.4 in the Supplementary Material). We then compute the average value $\hat{\lambda}_\ell = (\hat{\lambda}_{\ell 1}, \ldots, \hat{\lambda}_{\ell H})$ for each of the four clusters, to which a probability measure $\widetilde{p}_\ell \propto \sum_{h=1}^{H} \hat{\lambda}_{\ell h}\mu'_h$ and report the associated mixture density in the bottom row Figure 6.We define a cluster-specific mean distributions $\widetilde{p}_\ell \propto \sum_{h=1}^{H} \hat{\lambda}_{\ell h}\mu'_h$ by taking the average value $\hat{\lambda}_\ell = (\hat{\lambda}_{\ell 1}, \ldots, \hat{\lambda}_{\ell H})$ for each of the four cluster. the associated mixture densities are shown in the bottom row Figure 6. The clusters are easily interpretable and the mean distributions $\widetilde{p}_1, \ldots \widetilde{p}_4$ are substantially different.

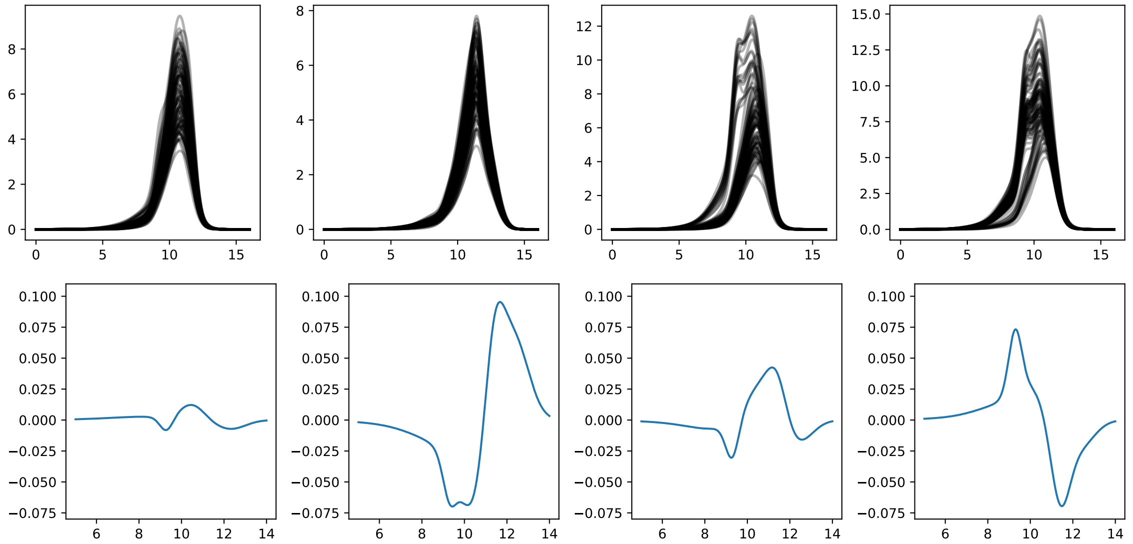

Fig. 7: Summary of posterior inference on the Californian income dataset. Top row: draws from the posterior distribution of the latent factor densities. Bottom row: average of the residual factor densities after post-processing.

6.2. Californian Income Data

We consider the 2021 American Community Survey census data publicly available at https://www.census.gov/programs-surveys/acs/data/experimental-data/2020-1-year-pums.html. Specifically, we consider the PINCP variable that represents the personal income of the survey responders and restrict to the citizens of the state of California. For privacy reasons, data are grouped into geographical units denoted as PUMAs, roughly corresponding to $100,000$ inhabitants. There are 265 PUMAs in California. We consider $y_{j,i}$ to be the logarithm of the income of the i-th person in the j-th PUMA. The total number of responders is $43,380$, with the median number of observations per PUMA being 164.

As shown in Figure E.5 in the Supplementary Material, the distributions of the income in different PUMAs are quite varied with clear spatial dependence. This is also confirmed by the analysis of Moran's I index for the average log-incomes, which is approximately 0.55. A permutation test confirmed that the spatial correlation is not-negligible. We assume independent log Gaussian Markov random fields priors for each column of Λ as in (4), where we fix $\tau = 2.5$ and $\rho = 0.95$. We choose H by evaluating the predictive goodness of fit for $H = 1, \ldots, 10$ using the widely applicable information criterion (WAIC, Watanabe, 2013). The best performance is associated with $H = 4$, therefore we comment on the posterior inference obtained under this model.

Figures 7 and 8 summarize the posterior findings. The draws from the latent factor densities (top row) show some evidence of label-switching in the third and fourth factors. Post-processing the chains with our algorithm estimates the four latent factors in Figure E.6 in the Supplementary Material. However, it is easier to interpret the residual factor densities displayed in the bottom row of Figure 7. The second and the fourth factors are associated with the largest variations. In particular, the second one gives mass to higher incomes while the fourth one gives mass to lower incomes. The first one is more representative of the average population since the variations are small. The third factor instead corresponds to average incomes and gives less mass (compared to the average population) to both low and high incomes. To visualize the spatial effect of the latent factors, we plot the scores s_{jh} for each factor. Note that the third latent factor is predominant in several areas, where s_{j3} is larger than 0.8. Instead, s_{j2} is small in all of California except for a

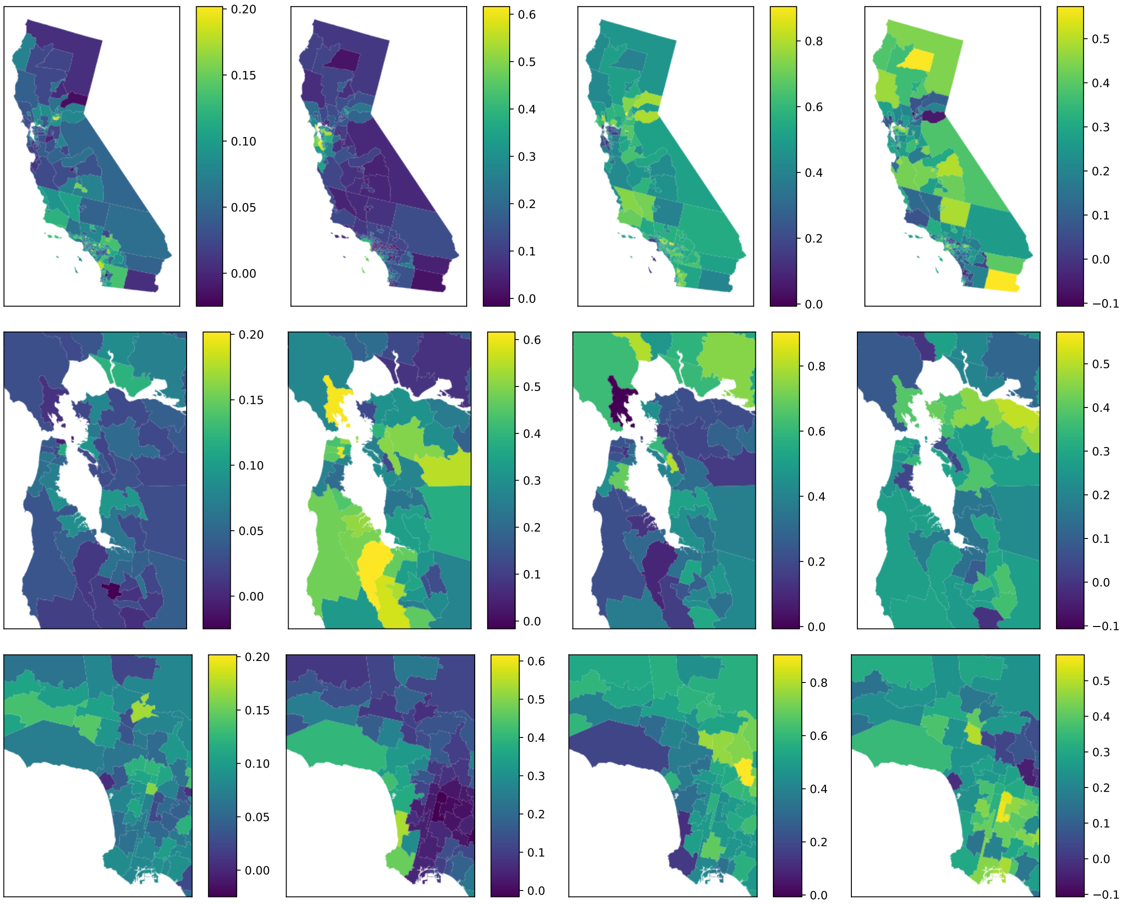

Fig. 8: Spatial distribution of the scores in the Californian income dataset. Top row: the scores s_{jh} for $h = 1, \dots, 4$ from left to right. Middle row: zoom on the San Francisco area. Bottom row: zoom on the Los Angeles area

few PUMAs in San Francisco, Long Beach, and San Diego, where the highest incomes are observed. In particular, zooming on San Francisco (middle row of Figure 8), we note that the second factor is highly represented in Palo Alto, home to several tech tycoons, and San Rafael, home to entertainers. Finally, note that the fourth factor (associated with the lowest incomes) has a high weight in the two PUMAs neighboring Mexico as well in some areas in Los Angeles. Notably, the PUMA around the port and the one corresponding to the "south LA" neighborhoods going from University Park to Green Meadows. This is in accordance with the 2008 *Concentrated Poverty in Los Angeles* report (Flaming and Matsunaga, 2008), which estimates that the percentage of households in poverty is typically above 40% in those areas.

7. Discussion

Modeling a collection of random probability measures is an old problem that has received considerable attention in the Bayesian nonparametric literature, see, e.g. Quintana et al. (2022) for a recent review. In this article, we have considered specifically the case when data are naturally divided into groups or subpopulations, and data are partially exchangeable. Taking a nonparametric Bayesian approach, we assumed that observations in each group can be suitably modelled by a mixture density, and proposed *normalized latent measure factor models* as a prior for the collection of mixing measures in each group. Similar to the

Gaussian latent factor model, our model assumes that each group-specific directing measure is a linear combination of a set of latent random measures. We can interpret the latent random measures as the latent common traits shared by the subpopulations. Moreover, the prior for the linear combination weights can include additional group-specific information such as geographical location. As a result of our construction, the group-specific random measures are not completely random. This precludes the study of important theoretical properties of our model, such as the posterior distribution of these random measures, at least with the classical tools employed in BNP. Moreover, expressions for marginal covariances between different measures are complex and not available analytically, which makes prior elicitation more demanding: we do not provide default values for the hyperparameters and suggest that prior elicitation should be carried out on a case-by-case basis, possibly through a priori Monte Carlo simulation as done in this paper.

To account for the non-identifiability of our model, we developed an ad-hoc post-processing algorithm leading to a constrained optimization algorithm over the special linear group, that is the group of matrices whose determinant is equal to one. To solve the optimization problem, we leveraged recent work on optimization on manifolds, proposing a Riemannian augmented Lagrangian method. Through simulations and illustrations on two real datasets, we validate our approach and show its usefulness, focusing in particular on the interpretation of the latent measures and the associated weights. We remark here that the post-processing is only necessary to estimate the latent random measures, and is superfluous if one's interest is predictive performance or estimating the group-specific densities. The model opens up many directions for future research which we discuss below and aim to investigate thoroughly in the future.

Our factor model approach can be extended to a wide range of dependence structures between the groups. For example, including observation-specific covariates in the model or time-dependent data. We can also build models which allow for the discovery of latent structure in the groups by further modeling the loadings matrix Λ. For instance, Rodríguez et al. (2008), Camerlenghi et al. (2019), Beraha et al. (2021), and Denti et al. (2021) build models which cluster groups according to the homogeneity of their distributions. We could achieve this by assuming that each of the group-specific directing measures is equal to one of the latent measures, *i.e.* only one of $\lambda_{j1}, \ldots, \lambda_{jH}$ are non-zero, which would be similar to exploratory factor analysis (Conti et al., 2014). Alternatively, we can achieve a "soft clustering" of the group-specific distributions *i.e.* cluster together similar distributions as opposed to homogeneous ones by assuming a mixture model for the rows of the matrix Λ. More generally, Λ could be expressed in terms of further low-rank matrix to find similarities between the group-specific factor loadings.

The post-processing identification scheme leads to estimated latent measures which are maximally separated according to the interpretability criterion. This allows us to interpret the factor loadings as an H-dimensional summary of the group-specific distribution where each element of the summary measures different parts of the distribution. In a similar way to scores from dimension reduction techniques, such as Principal Components Analysis, or embeddings in machine learning, these estimates can then be used as inputs into other statistical analyses. We effectively use this idea in the analysis of the Invalsi data-set where the estimated group-specific factor loadings are clustered to find groups of schools with similar distributions. This approach could have much wider applications. For example, the analysis of the Californian income data leads to estimated factor loadings for each PUMA which could be used in a regression model in place of other summaries such as median income, or the percentage of incomes below/above a threshold. These estimated factor loadings should provide more information than a single measure and be a more efficient representation than a large number of measures (for example, using a large number of thresholds). It would be particularly interesting to investigate this approach beyond univariate data, such as continuous or discrete multivariate observations where it's difficult to find efficient low-dimensional summaries of distributions.

Acknowledgements

Mario Beraha received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Mario Beraha thanks Matteo Pegoraro for helpful discussions on the post-processing algorithm.

References

Álvarez-Esteban, P. C., E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán (2018). Wide consensus aggregation in the Wasserstein space. Application to location-scatter families. *Bernoulli 24*(4A), 3147 – 3179.

Argiento, R., A. Cremaschi, and M. Vannucci (2019). Hierarchical normalized completely random measures to cluster grouped data. *J. Am. Stat. Assoc. 0*(0), 1–26.

Arminger, G. and B. O. Muthén (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika 63*(3), 271–300.

Baccelli, F., B. Błaszczyszyn, and M. Karray (2020). Random measures, point processes, and stochastic geometry. *HAL preprint available at https://hal.inria.fr/hal-02460214/*.

Bassetti, F., R. Casarin, and L. Rossini (2020). Hierarchical species sampling models. *Bayesian Anal. 15*(3), 809–838.

Beraha, M., A. Guglielmi, and F. A. Quintana (2021). The semi-hierarchical Dirichlet Process and its application to clustering homogeneous distributions. *Bayesian Anal. 16*(4), 1187–1219.

Beraha, M., B. Guindani, M. Gianella, and A. Guglielmi (2022). Bayesmix: Bayesian mixture models in C++. *arXiv preprint arXiv:2205.08144*.

Beraha, M., M. Pegoraro, R. Peli, and A. Guglielmi (2021). Spatially dependent mixture models via the logistic multivariate CAR prior. *Spat. Stat. 46*, 100548.

Bhattacharya, A. and D. B. Dunson (2011). Sparse Bayesian infinite factor models. *Biometrika 98*, 291–306.

Birgin, E. G. and J. M. Martínez (2014). *Practical augmented Lagrangian methods for constrained optimization.* SIAM.

Camerlenghi, F., R. Corradin, and A. Ongaro (2022). Conditional methods for compound random measures. *Manuscript in Preparation*.

Camerlenghi, F., D. B. Dunson, A. Lijoi, I. Prünster, and A. Rodríguez (2019, 12). Latent nested nonparametric priors (with discussion). *Bayesian Anal. 14*(4), 1303–1356.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *Ann. Stat. 47*(1), 67–92.

Campbell, T., J. H. Huggins, J. P. How, and T. Broderick (2019). Truncated random measures. *Bernoulli 25*, 1256–1288.

Canale, A. and D. B. Dunson (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association 106*(496), 1528–1539. PMID: 22523437.

Conti, G., S. Frühwirth-Schnatter, J. J. Heckman, and R. Piatek (2014). Bayesian exploratory factor analysis. *J. Econom. 183*, 31–57.

Corradin, R., A. Canale, and B. Nipoti (2021). BNPmix: An R package for Bayesian nonparametric modeling via Pitman-Yor mixtures. *J. Stat. Softw. 100*, 1–33.

Denti, F., F. Camerlenghi, M. Guindani, and A. Mira (2021). A common atoms model for the Bayesian nonparametric analysis of nested data. *J. Am. Stat. Assoc. 0*(0), 1–12.

Elliott, L. T., M. D. Iorio, S. Favaro, K. Adhikari, and Y. W. Teh (2019). Modeling Population Structure Under Hierarchical Dirichlet Processes. *Bayesian Anal. 14*(2), 313 – 339.

Flaming, D. and M. Matsunaga (2008). Concentrated poverty in Los Angeles (February 9, 2008). *Economic Roundtable Research Report, February 2008, Available at SSRN: https://ssrn.com/abstract=2772191*.

França, G., A. Barp, M. Girolami, and M. I. Jordan (2021). Optimization on manifolds: A symplectic approach.

Geweke, J. and G. Zhou (1996). Measuring the Pricing Error of the Arbitrage Pricing Theory. *Rev. Financ. Stud. 9*(2), 557–587.

Geweke, J. F. and K. J. Singleton (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *J. Am. Stat. Assoc. 75*(369), 133–137.

Griffin, J. E., M. Kolossiatis, and M. F. J. Steel (2013). Comparing distributions by using dependent normalized random-measure mixtures. *J. R. Statist. Soc. B 75*(3), 499–529.

Griffin, J. E. and F. Leisen (2017). Compound random measures and their use in Bayesian non-parametrics. *J. R. Statist. Soc. B 79*(2), 525–545.

Griffin, J. E. and S. G. Walker (2011). Posterior simulation of normalized random measure mixtures. *J. Comput. Graph. Stat. 20*(1), 241–259.

Hyvärinen, A. (2013). Independent component analysis: recent advances. *Philos. Trans. Royal Soc. A 371*(1984), 20110534.

Kingman, J. F. C. (1967). Completely random measures. *Pac. J. Math. 21*(1), 59 – 78.

Kingman, J. F. C. (1993). *Poisson processes*, Volume 3 of *Oxford Studies in Probability*. New York: The Clarendon Press Oxford University Press. Oxford Science Publications.

Legramanti, S., D. Durante, and D. B. Dunson (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika 107*, 745–752.

Lijoi, A., B. Nipoti, and I. Prünster (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli 20*(3), 1260–1291.

Lijoi, A., I. Prünster, and G. Rebaudo (2022). Flexible clustering via hidden hierarchical Dirichlet priors. *Scand. J. Stat 50*(1), 213–234.

Liu, C. and N. Boumal (2020). Simple algorithms for optimization on Riemannian manifolds with constraints. *Appl. Math. Optim. 82*(3), 949–981.

Montagna, S., S. T. Tokdar, B. Neelon, and D. B. Dunson (2012). Bayesian latent factor regression for functional and longitudinal data. *Biometrics 68*(4), 1064–1073.

Müller, P., F. Quintana, and G. Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *J. R. Stat. Soc. B 66*(3), 735–749.

Nguyen, T. D., J. H. Huggins, L. Masoero, L. Mackey, and T. Broderick (2020). Independent versus truncated finite approximations for Bayesian nonparametric inference. In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop.*

Papastamoulis, P. and I. Ntzoufras (2022, feb). On the identifiability of Bayesian factor analytic models. *Stat. Comput. 32*(2).

Pegoraro, M. and M. Beraha (2022). Projected Statistical Methods for Distributional Data on the Real Line with the Wasserstein Metric. *J. Mach. Learn. Res. 23*(37), 1–59.

Poworoznek, E., F. Ferrari, and D. Dunson (2021). Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching.

Quintana, F. A., P. Müller, A. Jara, and S. N. MacEachern (2022). The dependent Dirichlet process and related models. *Stat. Sci. 37*(1), 24–41.

Regazzini, E., A. Lijoi, and I. Prünster (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Stat. 31*(2), 560 – 585.

Rigon, T. and D. Durante (2021). Tractable Bayesian density regression via logit stick-breaking priors. *J. Stat. Plan. Inference 211*, 131–142.

Rodríguez, A., D. B. Dunson, and A. E. Gelfand (2008). The nested Dirichlet process. *J. Am. Stat. Assoc. 103*(483), 1131–1154.

Schiavon, L., A. Canale, and D. B. Dunson (2022, 01). Generalized infinite factorization models. *Biometrika 109*(3), 817–835. asab056.

Sra, S. and I. Dhillon (2005). Generalized nonnegative matrix approximations with Bregman divergences. *Adv. Neural Inf. Process. Syst. 18.*

Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei (2006). Hierarchical Dirichlet Processes. *J. Am. Stat. Assoc. 101*(476), 1566–1581.

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