NORMALIZED LATENT MEASURE FACTOR MODELS

Mario Beraha [1] and Jim E. Griffin[2]

[1] Department of Economics and Statistics, University of Torino,
(e-mail: mario.beraha@unito.it)

[2] Department of Statistical Sciences, University College London
(e-mail: j.griffin@ucl.ac.uk)

ABSTRACT: 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. 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

In this short paper, we review the methodology for modeling and comparing probability distributions discussed in Beraha and Griffin (2022). 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. We consider here specifically the case where data are naturally divided into groups or subpopulations, and data are partially exchangeable. Let $(y_1,\dots,y_g)$ denote a sample of observations divided into g groups where $y_j = (y_{j1},\dots,y_{jn_j})$. By de Finetti's theorem, partial exchangeability is tantamount to assuming that there is a vector of random probability measures $(p_1,\dots,p_g) \sim Q$ such that

$$\begin{aligned} y_{j1},\dots,y_{jn_j} \mid p_j &\overset{\text{iid}}{\sim} p_j, \qquad j = 1,\dots,g \\ p_1,\dots,p_g &\sim Q \end{aligned} \tag{1}$$

and independence holds across groups. In particular, we focus here on mixture models of the kind

$$p_j(y) = \int_\Theta f(y \mid \theta)\widetilde{p}_j(\mathrm{d}\theta)$$

where the $\widetilde{p}_j$'s are almost surely discrete random probability measures.

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).

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. In particular, 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. 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.

2 The Model

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

$$\widetilde{\mu}_j(\cdot) = \sum_{h=1}^{H} \lambda_{jh}\mu_h^*(\cdot). \tag{2}$$

Note that (2) generalizes the construction in Griffin et al. (2013) and Lijoi et al. (2014).

A 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 $\nu^*(\mathrm{d}z)\alpha(\mathrm{d}x)$. In this case we can write

$$\widetilde{\mu}_j(\cdot) = \sum_{k \geq 1} (\Lambda M)_{jk} J_k \delta_{\theta_k^*}(\cdot), \tag{3}$$

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. Suitable prior distributions for all the parameters will be specified in later sections.

Equations (2) and (3) share analogies with 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 and call Λ the loadings matrix and the μ_h^*'s the latent measures.

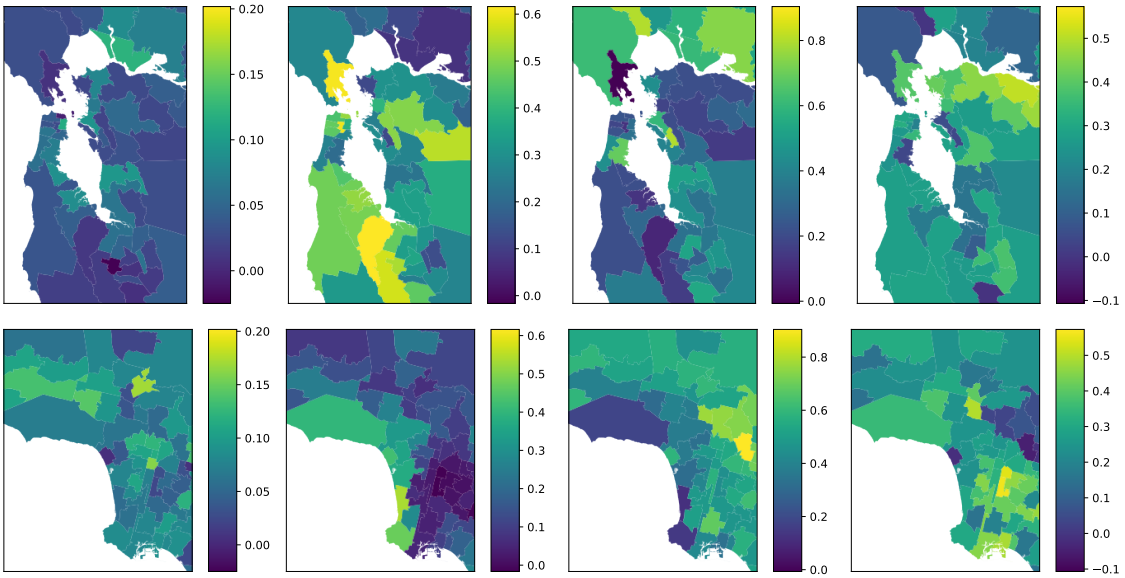

Figure 1: Spatial distribution of the scores in the Californian income dataset. Top row: San Francisco area. Bottom row: Los Angeles ares. Columns represent overall average, high, median, and low income prevalence, respectively.

3 Some details about posterior inference and post-processing

We perform posterior inference by proposing an ad-hoc Markov chain Monte Carlo algorithm, which combines Gibbs updates (when the full conditionals belong to known parametric families) with Hamiltonian Monte Carlo steps (when they do not). For computational convenience, we truncate the support of the CoRM to K atoms, but a slice sampler could be alternatively employed. Software is implemented using the `JAX` Python package.

Our model is not identifiable due to the multiplicative relationship between Λ and $(\mu_1^*, \dots, \mu_h^*)$. 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. 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 $M' = QM$. Nevertheless, 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 do not need to worry about sign ambiguity.

As in Poworoznek et al. (2021), we propose to find an optimal Q via an ad-hoc post processing that aims to maximally separate the latent measure μ_h^*'s, according to some notion of distance between measures. We formalize the post-processing into a constrained optimization problem over the special linear group, that is the set of matrices with determinant equal to one. The special linear group is not a linear space, but can take advantage of its differential structure, and tackle the problem via a Riemannian augmented Lagrangian method that leverages recent advances in Riemannian optimization (França et al., 2021).

4 Analysis of Californian Income Data

We consider the 2021 American Community Survey census data publicly available at `https://www.census.gov/programs-surveys`. 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 PUMA,

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.

We assume independent log-Gaussian Markov random field priors for each column of Λ, beta priors for the J's and gamma priors for the m_{hk}'s and fix $H = 4$. Although not shown here, the four latent measures can be interpreted as representing *average incomes* (i.e. the distribution is equal to the whole population distribution), *high incomes*, *median incomes* (i.e., the distribution is concentrated on median values) and *low incomes*. Figure 1 shows the values of Λ for the four latent measures associated with the PUMAs in the San Francisco and Los Angeles areas. In particular, 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 some areas in Los Angeles. In particular, the PUMA around the port and the one corresponding to the "south LA" neighborhoods going from University Park to Green Meadows. This is in agreement with the 2008 *Concentrated Poverty in Los Angeles* report, which estimates that the percentage of households in poverty is typically greater than 40% in those areas.

References

Argiento, R., A. Cremaschi, and M. Vannucci (2019). Hierarchical normalized completely random measures to cluster grouped data. *J. Am. Stat. Assoc. 115*(529), 318–333.

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.

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

Beraha, M. and J. E. Griffin (2022). Normalized latent measure factor models. *arXiv preprint arXiv:2205.15654*.

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

França, G., A. Barp, M. Girolami, and M. I. Jordan (2021). Optimization on manifolds: A symplectic approach. *arXiv preprint arXiv:2107.11231*.

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 nonparametrics. *J. R. Statist. Soc. B 79*(2), 525–545.

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

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.

Poworoznek, E., F. Ferrari, and D. Dunson (2021). Efficiently resolving rotational ambiguity in Bayesian matrix sampling with matching. *arXiv preprint arXiv:2107.13783*.

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.

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.