Sampling hierarchies of discrete random structures

Antonio Lijoi[1], Igor Prünster[1], and Tommaso Rigon[2]

[1]Department of Decision Sciences and BIDSA, Bocconi University, Milano, Italy
[2]Department of Statistical Science, Duke University, Durham, U.S.A.

Abstract

Hierarchical normalized discrete random measures identify a general class of priors that is suited to flexibly learn how the distribution of a response variable changes across groups of observations. A special case widely used in practice is the hierarchical Dirichlet process. Although current theory on hierarchies of nonparametric priors yields all relevant tools for drawing posterior inference, their implementation comes at a high computational cost. We fill this gap by proposing an approximation for a general class of hierarchical processes, which leads to an efficient conditional Gibbs sampling algorithm. The key idea consists of a deterministic truncation of the underlying random probability measures leading to a finite dimensional approximation of the original prior law. We provide both empirical and theoretical support for such a procedure.

Keywords: Bayesian nonparametrics; Hierarchical Dirichlet process; Normalized random measures; Pitman–Yor process.

1 Introduction

When investigating covariate–dependent observations $\{(X_{li})_{i\geq 1} : l \in \mathcal{L}\}$ in a Bayesian framework, the standard assumption of exchangeability is not appropriate since it amounts to considering the data as being homogeneous. The covariate $l \in \mathcal{L}$ is actually a source of heterogeneity that one has to take into account and a different symmetry condition among the data should be specified. Here we focus on the case where the covariate space is finite, i.e. $\mathcal{L} = \{1, \ldots, d\}$, and identifies data that are recorded under d different, though related, experimental conditions. In view of this, a natural dependence structure is implied by *partial exchangeability* according to which exchangeability holds true within each of the d separate groups of observations, but not across them. More formally, let $\mathbb{X}$ be the sample space and let $\mathscr{X}$ denote its Borel σ-algebra. For the sake of generality, the space $\mathbb{X}$ is assumed to be Polish, although in practice one typically has $\mathbb{X} \subseteq \mathbb{R}^p$. Moreover, $\mathsf{P}_{\mathbb{X}}$ stands for the space of probability measures on $\mathbb{X}$. The array of $\mathbb{X}$–valued random elements $\{(X_{li})_{i\geq 1} : l = 1, \ldots, d\}$ is partially exchangeable if and only if for any $i = 1, \ldots, n^{(l)}$ and any $l = 1, \ldots, d$

$$(X_{li} \mid \tilde{p}_l) \overset{\text{ind}}{\sim} \tilde{p}_l, \qquad (\tilde{p}_1, \ldots, \tilde{p}_d) \sim Q_d, \tag{1}$$

for some probability measure Q_d on the product space $\mathsf{P}_{\mathbb{X}}^d$. Hence, conditionally on the vector $(\tilde{p}_1, \dots, \tilde{p}_d)$, the X_{li}'s are independent and identically distributed within, but only independent across groups. The measure Q_d plays the role of prior distribution and in addition governs the dependence across groups. This setting also constitutes a crucial building block for the construction of more complex models in which latent quantities, rather than the raw data, are assumed to be partially exchangeable.

The definition and the investigation of Q_d (for $d \geq 1$) is an active field of research in Bayesian Nonparametrics (BNP). In the simple exchangeable case ($d = 1$), for example, some well–known limitations of the Dirichlet process (Ferguson, 1973) have fostered the research of novel *discrete* nonparametric priors, which are nowadays well established inferential tools. Among the several available alternatives, we recall the Pitman–Yor process (Pitman and Yor, 1997) and the normalized random measures with independent increments (NRMIs) (Regazzini et al., 2003). Both the Pitman–Yor process and the subclass of homogeneous NRMIs are, in turn, instances of *proper* species sampling models with infinitely many components (Pitman, 1996), namely random probability measures

$$\tilde{p}(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot),$$

where the random $\mathbb{X}$–valued locations $(\theta_h)_{h \geq 1}$ and the random weights $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$ are independent. Furthermore, the θ_h's are independent and identically distributed (iid) draws from a diffuse probability measure P, that is $P(\{x\}) = 0$ for any $x \in \mathbb{X}$.

As for the partially exchangeable case ($d \geq 2$), an early proposal for Q_d appeared in Cifarelli and Regazzini (1978), but the decisive boost to the literature came after the seminal paper of MacEachern (1999). Here we will rely on a hierarchical construction of Q_d and assume that the elements of the collection $\{\tilde{p}_1, \dots, \tilde{p}_d\}$ are conditionally iid, given another discrete random probability measure $\tilde{p}_0$, such that

$$\begin{aligned} \tilde{p}_l(\cdot) &= \sum_{h=1}^{\infty} \xi_{lh} \delta_{\theta_{lh}}(\cdot), \quad (\theta_{lh} \mid \tilde{p}_0) \overset{\text{iid}}{\sim} \tilde{p}_0, \\ \tilde{p}_0(\cdot) &= \sum_{h=1}^{\infty} \pi_{0h} \delta_{\phi_h}(\cdot), \quad \phi_h \overset{\text{iid}}{\sim} P_0, \end{aligned} \tag{2}$$

for $l = 1, \dots, d$ and $h \geq 1$, where P_0 is some diffuse probability measure on $\mathbb{X}$. Note that in view of this specification, one marginally has $\mathbb{E}(\tilde{p}_l \mid \tilde{p}_0) = \tilde{p}_0$ for each $l = 1, \dots, d$. Thus, dependence across groups in (1) is induced by considering an exchangeable collection $\{\tilde{p}_1, \dots, \tilde{p}_d\}$ of random probability measures. Such a model, when the $\tilde{p}_l$'s and $\tilde{p}_0$ are Dirichlet processes, has been proposed in Teh et al. (2006) and takes on the name of *hierarchical Dirichlet process* (HDP). The HDP has been successfully applied, e.g., to topic modeling (Teh et al., 2006), speaker diarization (Fox et al., 2011) and the analysis of fMRI data (Zhang et al., 2016). For a stimulating account on its use in several modeling and applied frameworks see Teh and Jordan (2010). An extension to the wider class of *normalized random measures* was proposed in Camerlenghi et al. (2019), which further provides a systematic investigation of the most relevant distributional properties for Bayesian inference. The achievement of these results heavily benefits from the nice probabilistic structure of the completely

random measures (CRMs) that are used to define the underlying random probability measures. See also Bassetti et al. (2019), Argiento et al. (2019) for further recent developments. It is worth recalling that other examples of CRM–based priors Q_d are available in the literature, the most recent examples being Lijoi et al. (2014a,b), Lijoi and Nipoti (2014) and Griffin and Leisen (2017).

As noted from equation (2), the random probability measures $\tilde{p}_l$ at the bottom of the hierarchy have a purely atomic base measure $\tilde{p}_0$ that stands at the top. Discreteness of $\tilde{p}_0$ is a significant hurdle as it entails challenging analytical difficulties that are effectively detailed in Camerlenghi et al. (2019). And while one still achieves a posterior characterization in this setting, its implementation might be computationally challenging in practice. Additionally, most of the current algorithms for posterior inference with hierarchical processes are of *marginal* type, that is they rely on the marginalization of the random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_d)$. While having some computational advantages, this rules out the possibility of obtaining complex posterior functionals of the vector $(\tilde{p}_1, \ldots, \tilde{p}_d)$, which are often of interest in several applied contexts such as, for example, credible intervals. To overcome this difficulty, we propose a simple and efficient *conditional* Gibbs sampler for a wide class of hierarchical discrete random probability measures that includes the HDP as a special case. The actual implementation of the algorithm is eased by an a priori approximation of the infinite dimensional process, based on a deterministic truncation of the random probability measure $\tilde{p}_0$. We provide theoretical support for such a truncation, borrowing ideas from the arguments of Muliere and Tardella (1998), Ishwaran and James (2001), Argiento et al. (2016) and Arbel et al. (2019), who described truncated approximations for discrete nonparametric priors, within the exchangeable setting.

It is finally worth noting that building upon model (2) and, then, truncating to the Hth term, one can obtain the building block of a mixture model for partially exchangeable data that is discussed in detail in Section 4. Most notably, it also has some connections with the Latent Dirichlet Allocation (LDA) of Blei et al. (2003), of which our proposal is a generalization. In fact, we work with a wider class of distributions compared to the Dirichlet distribution used in LDA. Additionally, while in LDA dependence among mixing distributions is induced through an approximate empirical Bayes procedure that determines the numerical value of certain hyperparameters of the model, here our full Bayesian analysis makes use of suitable prior laws for all the parameters and hyperparameters of the model. Finally, as for the choice of H, that is the number of latent topics in the terminology of topic modeling, in LDA it is selected so that it minimizes some out-of-sample goodness-of-fit metric. On the other hand, we choose H in order to achieve a satisfactory approximation of the infinite dimensional process; the actual number of latent topics is elegantly and effectively regulated by the prior. We stress that our model is not confined to topic modeling with categorical data: indeed, they may cope with observations taking values in general Polish spaces, thus allowing for a much broader applicability.

The paper is organized as follows: in Section 2 we review some background material on homogeneous normalized random measures with independent increments (NRMIs), and on the Pitman-Yor process. In Section 3, we propose a particular instance of hierarchical process and we discuss a finite dimensional approximation based on a deterministic truncation of $\tilde{p}_0$. In Section 4 the truncated process is employed to define an infinite mixture model for partially exchangeable data. The novel conditional Gibbs sampler to conduct posterior

inference is derived and described in detail. To assess the practical performance of both the novel algorithm and the aforementioned infinite mixture model, we conduct a simulation study in Section 5. Finally, as an illustration, we apply our algorithm on real data in Section 6.

2 Preliminaries and background

2.1 Normalized random measures

Throughout the manuscript we will make extensive use of the notion of homogeneous normalized completely random measures (NRMIs), and of the Pitman-Yor process (PY), which are hence briefly recalled here.

We start by providing some preliminary, and concise, background on completely random measures. To this end, we will denote with $\mathsf{M}_{\mathbb{X}}$ the space of boundedly finite measures on $(\mathbb{X}, \mathscr{X})$ and with $\mathscr{M}_{\mathbb{X}}$ the corresponding σ-algebra. See Appendix A.2 of Daley and Vere-Jones (2003) for details on this space and its properties.

Definition 1. *A measurable function $\tilde{\mu}$ defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathsf{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$ is a completely random measure (CRM) if for any pairwise disjoint sets $A_1, \ldots, A_M$ in $\mathscr{X}$ and for any $M \geq 2$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_M)$ are mutually independent.*

Henceforth we mostly focus on CRMs of the form

$$\tilde{\mu}(\cdot) = \sum_{h=1}^{\infty} J_h \delta_{\theta_h}(\cdot), \tag{3}$$

for some sequence of $\mathbb{X}$–valued random elements $(\theta_h)_{h\geq 1}$ and positive jumps $(J_h)_{h\geq 1}$. Under representation (3), a CRM $\tilde{\mu}$ is characterized by the so-called Lévy-Khintchine representation, which states that

$$\mathbb{E}\left(\exp\left\{-\int_{\mathbb{X}} f(x)\tilde{\mu}(\mathrm{d}x)\right\}\right) = \exp\left\{-\int_{\mathbb{R}^+\times\mathbb{X}} \left(1-\mathrm{e}^{-sf(x)}\right)\nu(\mathrm{d}s, \mathrm{d}x)\right\}, \tag{4}$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\int_{\mathbb{X}} |f(x)|\tilde{\mu}(\mathrm{d}x) < \infty$ almost surely, with ν being a measure on $\mathbb{R}^+ \times \mathbb{X}$ such that for any set A in $\mathscr{X}$ it holds $\int_A \int_{\mathbb{R}^+} \min\{s, 1\}\nu(\mathrm{d}s, \mathrm{d}x) < \infty$. By virtue of (4), the measure ν characterizes $\tilde{\mu}$ and it is referred to as the Lévy intensity of $\tilde{\mu}$. We additionally assume *homogeneity*, implying that the Lévy intensity ν can be factorized as the product of two measures, that is

$$\nu(\mathrm{d}s, \mathrm{d}x) = \rho(\mathrm{d}s)cP(\mathrm{d}x), \tag{5}$$

with c a positive constant and P a probability measure over $(\mathbb{X}, \mathscr{X})$. This amounts to assuming that the sequences $(J_h)_{h\geq 1}$ and $(\theta_h)_{h\geq 1}$ in (3) are independent. Note also that the points $(J_h, \theta_h)_{h\geq 1}$ can be regarded as a sample from a Poisson process with intensity ν.

Definition 2. *Let $\tilde{\mu}$ be a CRM with Lévy intensity ν as in (5) and such that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ almost surely. Then, a random probability measure $\tilde{p}$ is named*

*homogeneous normalized random measure with independent increments (*NRMI*) if*

$$\tilde{p}(\cdot) = \frac{\tilde{\mu}(\cdot)}{\tilde{\mu}(\mathbb{X})} = \sum_{h=1}^{\infty} \frac{J_h}{\sum_{h'=1}^{\infty} J_{h'}} \delta_{\theta_h}(\cdot) \sim \text{NRMI}(c, \rho, P),$$

where $\tilde{\mu}(\mathbb{X}) = \sum_{h'=1}^{\infty} J_{h'}$ *and* $\theta_h \overset{\text{iid}}{\sim} P$.

The homogeneity assumption (5) greatly simplifies the development of theoretical properties and has several practical advantages. Indeed, most of the NRMI currently employed in Bayesian nonparametrics are homogeneous (Lijoi and Prünster, 2010), and nonetheless they represent a wide class of random probability measure. We shall stress the fact that this definition of homogeneous NRMI does not make any further assumptions on the baseline probability measure P. In particular, P is allowed to have atoms.

While Definition 2 encompasses several nonparametric priors used in the Bayesian literature such as, e.g., the Dirichlet process (Ferguson, 1973), the normalized stable process (Kingman, 1975) and the normalized inverse–Gaussian process (Lijoi et al., 2005), an important instance of discrete random probability measure that does not belong to this class of NRMIs is the Pitman–Yor process. The PY process can be defined in terms of a simple stick-breaking construction, which will henceforth be relevant for us.

Definition 3. *Let* $(v_{0h})_{h\geq 1}$, *with* $v_{0h} \overset{\text{ind}}{\sim} \text{BETA}(1-\sigma_0,\, c_0 + h\sigma_0)$, *and*

$$\tilde{p}_0(\cdot) = \sum_{h=1}^{\infty} \pi_{0h} \delta_{\phi_h}(\cdot), \qquad \pi_{0h} = v_{0h} \prod_{h'<h} (1 - v_{0h'}), \qquad \phi_h \overset{\text{iid}}{\sim} P_0, h \geq 1,$$

where P_0 *is a diffuse probability measure on* $(\mathbb{X}, \mathscr{X})$ *and we agree on* $\prod_{h'<1}(1-v_{0h'}) \equiv 1$. *Then* $\tilde{p}_0$ *is a Pitman–Yor process whenever the parameters are such that* $\sigma_0 \in [0,1)$ *and* $c_0 > -\sigma_0$ *or* $\sigma_0 < 0$ *and* $c_0 = m|\sigma_0|$ *for some integer* m. *We will use the notation* $\tilde{p}_0 \sim \text{PY}(\sigma_0, c_0, P_0)$.

In the sequel we will only consider a subset of the collection of admissible parameters (σ_0, c_0), namely that for which $\sigma_0 \in [0,1)$. Clearly, setting $\sigma_0 = 0$ one obtain the stick-breaking construction of Sethuraman (1994) for the Dirichlet process, whereas for $c_0 = 0$ one is able to recover the stick-breaking construction of the σ_0-stable process given in Perman (1990). See also Perman et al. (1992). The distribution of the weights $\boldsymbol{\pi}_0 = (\pi_{01}, \pi_{02}, \dots)$ will be denoted with

$$\boldsymbol{\pi}_0 \sim \text{GEM}(\sigma_0, c_0),$$

after Griffiths, Engen, and McCloskey, and is also referred to as the *two–parameter Poisson–Dirichlet process*.

2.2 NRMI with finitely supported base measure

The hierarchical specification of discrete random probability measures given in (2) entails that each $\tilde{p}_l$ has, conditionally on $\tilde{p}_0$, an atomic base measure. In our case the $\tilde{p}_l$'s are homogeneous NRMIs and this motivates our interest in discussing specific features of NRMIs whose base measure P is purely atomic. Accordingly, henceforth we will suppose that, for some $H \geq 1$, there exists $\{x_1, \dots, x_H\} \subset \mathbb{X}$ such that $P(\{x_h\}) > 0$ for any $h \in \{1, \dots, H\}$ and

$\sum_{h=1}^{H} P(\{x_h\}) = 1$. This corresponds to normalizing a CRM with fixed points of discontinuity.

Let us first consider a finite collection $\{\zeta_1, \ldots, \zeta_H\}$ of independent and infinitely divisible positive random variables such that for any $\lambda > 0$, $c_h > 0$ and $h = 1, \ldots, H$ one has $\mathbb{E}\{\exp(-\lambda\zeta_h)\} = \exp\{-c_h\,\psi(\lambda)\}$. The function ψ is the so-called *Laplace exponent* corresponding to jump measure ρ, which is defined as $\psi(\lambda) = \int_{\mathbb{R}^+} \left(1 - \mathrm{e}^{-\lambda s}\right) \rho(\,\mathrm{d}s)$ for any $\lambda > 0$. We also assume that $\zeta_h = 0$ almost surely if $c_h = 0$.

Definition 4. *If $\bar{\zeta} = \sum_{h=1}^{H} \zeta_h$ and $\pi_h = \zeta_h/\bar{\zeta}$, then we say that $(\pi_1, \ldots, \pi_{H-1})$ identifies a normalized infinitely divisible distribution and will use the notation*

$$(\pi_1, \ldots, \pi_{H-1}) \sim \text{NID}(c_1, \ldots, c_H; \rho).$$

These distributions have been discussed at length in Favaro et al. (2011) and Lijoi et al. (2019). If $\tilde{p} \sim \text{NRMI}(c, \rho, P)$, for any finite and measurable partition $\{B_1, \ldots, B_M\}$ of $\mathbb{X}$ the vector $(\tilde{p}(B_1), \ldots, \tilde{p}(B_{M-1}))$ clearly identifies a probability distribution on the simplex $\mathcal{S}_{M-1} = \{(\omega_1, \ldots, \omega_{M-1}) : \omega_m \geq 0;\ \sum_{m=1}^{M-1} \omega_m \leq 1\}$. Combining standard properties of CRMs and NRMIs with Definition 4 one can show that

$$(\tilde{p}(B_1), \ldots, \tilde{p}(B_{M-1})) \sim \text{NID}(cP(B_1), \ldots, cP(B_M); \rho),$$

with the proviso that $\tilde{p}(B_m) = 0$, almost surely, if $P(B_m) = 0$. If we set $c_h = cP(\{x_h\})$ for each $h = 1, \ldots, H$, and note that $P(\mathbb{X} \setminus \{x_1, \ldots, x_H\}) = 0$, the random probability measure $\tilde{p} \sim \text{NRMI}(c, \rho, P)$ is fully characterized by the random vector

$$(\tilde{p}(\{x_1\}), \ldots, \tilde{p}(\{x_{H-1}\})) \sim \text{NID}(c_1, \ldots, c_H; \rho),$$

and the support of $\tilde{p}$ is the finite set $\{x_1, \ldots, x_H\}$, almost surely. This motivates the shorter notation $\tilde{p} \sim \text{NRMI}(c_1, \ldots, c_H; \rho)$ we use in this setting. Note that the posterior distribution of a NID random vector, if data are generated under a multinomial sampling, can be obtained in closed form. This is detailed in Lijoi et al. (2019) and it will be of great practical importance in the implementation of conditional algorithms for hierarchical processes.

While NIDs have been defined on a finite–dimensional simplex, they can be easily extended to an infinite dimensional setting. This is illustrated in the following.

Definition 5. *Let $\boldsymbol{c} = (c_1, c_2, \ldots)$ be an infinite collection of non–negative numbers such that $0 < \sum_{h=1}^{\infty} c_h < \infty$. An infinite random vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ such that $\sum_{h=1}^{\infty} \pi_h = 1$, almost surely, is a normalized infinitely divisible process (NIDP) with parameters ρ and $\boldsymbol{c}$ if, for any $M \geq 2$ and finite partition $\mathcal{H}_1, \ldots, \mathcal{H}_M$ of $\mathbb{N}$, one has*

$$\Big(\sum_{j \in \mathcal{H}_1} \pi_j, \ldots, \sum_{j \in \mathcal{H}_{M-1}} \pi_j\Big) \sim \text{NID}\Big(\sum_{j \in \mathcal{H}_1} c_j, \ldots, \sum_{j \in \mathcal{H}_M} c_j; \rho\Big),$$

and it will be denoted $\boldsymbol{\pi} \sim \text{NIDP}(\boldsymbol{c}, \rho)$.

If we take $\tilde{p} \sim \text{NRMI}(c, \rho, P)$ with $cP = \sum_{h=1}^{\infty} c_h \delta_{x_h}$ and let, for any $h \geq 1$

$$\pi_h = \tilde{p}(\{x_h\}) = \sum_{h=1}^{\infty} \frac{J_h}{\sum_{h'=1}^{\infty} J_{h'}} \delta_{\theta_h}(\{x_h\}) = \sum_{\{j:\, \theta_j = x_h\}} \frac{J_j}{\sum_{h' \geq 1} J_{h'}},$$

then $\boldsymbol{\pi} \sim \text{NIDP}(\boldsymbol{c}, \rho)$ with $c_h = cP(\{x_h\})$ for each h. Because of their connection with NRMIs with countable baseline measure, NIDP processes will play a central role also in the description of general hierarchical processes.

We move on presenting some examples of homogeneous NRMIs, the associated NID distributions and their densities, that will play a relevant role in the sequel.

Example 1 (Gamma process). *Let $\rho(\,ds) = s^{-1}\mathrm{e}^{-s}\,ds$, thus implying that the* CRM *$\tilde{\mu}$ is a gamma process since $\tilde{\mu}(B) \overset{ind}{\sim} \text{GAMMA}(cP(B), 1)$ for any $B \in \mathscr{X}$. Hence, $\tilde{p} \sim \text{NRMI}(c, \rho, P)$ is a Dirichlet process and for any measurable partition $\{B_1 \ldots, B_M\}$ of $\mathbb{X}$*

$$(\tilde{p}(B_1), \ldots, \tilde{p}(B_{M-1})) \sim \text{DIRICHLET}\,(c_1, \ldots, c_M)$$

with $c_m = cP(B_m)$ for $m = 1, \ldots, M$. If $c_m > 0$ for each $m = 1, \ldots, M$, its density function is

$$f(\boldsymbol{\omega}) = \frac{\Gamma\,(c_1 + \cdots + c_M)}{\Gamma(c_1) \times \cdots \times \Gamma(c_M)}\ \omega_1^{c_1 - 1} \cdots \omega_{M-1}^{c_{M-1} - 1}(1 - |\boldsymbol{\omega}|)^{c_M - 1} I_{\mathcal{S}_{M-1}}(\boldsymbol{\omega}),$$

where $|\boldsymbol{\omega}| = \sum_{m=1}^{M-1} \omega_m$.

Example 2 (Inverse Gaussian process). *Let the intensity function $\rho(\,ds) = (\sqrt{2\pi})^{-1}\, s^{-3/2}\, \mathrm{e}^{-s/2}\,ds$, which identifies an inverse–Gaussian* CRM *$\tilde{\mu}$. Then $\tilde{\mu}(B) \sim \text{IG}(cP(B), 1)$ for any $B \in \mathscr{X}$ and we shall, then, use the notation*

$$(\tilde{p}(B_1), \ldots, \tilde{p}(B_{M-1})) \sim \text{N-IG}\,(c_1, \ldots, c_M)\,,$$

with $c_m = cP(B_m)$ for $m = 1, \ldots, M$. If $c_m > 0$ for each m, its density function can be obtained in closed form (Lijoi et al., 2005) and coincides with

$$\begin{aligned} f(\boldsymbol{\omega}) = & \frac{\mathrm{e}^{\sum_{m=1}^{M} c_m} \prod_{m=1}^{M} c_m}{2^{M/2-1}\Gamma(1/2)^M} \frac{\mathcal{K}_{-M/2}\left(\sqrt{\mathcal{A}_M(\boldsymbol{\omega})}\right)}{\mathcal{A}_M(\boldsymbol{\omega})^{M/4}} \times \\ & \times \left\{\omega_1 \cdots \omega_{M-1}(1 - |\boldsymbol{\omega}|)\right\}^{-3/2} I_{\mathcal{S}_{M-1}}(\boldsymbol{\omega}), \end{aligned}$$

where $\mathcal{A}_M(\boldsymbol{\omega}) = \sum_{m=1}^{M-1}(c_m^2/\omega_m) + c_M^2/(1 - |\boldsymbol{\omega}|)$ and $\mathcal{K}_q(\cdot)$ denotes the modified Bessel function of the third type.

Example 3 (1/2 stable process). *Let the intensity function $\rho(\,ds) = (\sqrt{2\pi})^{-1}\, s^{-3/2}\,ds$, so that the* CRM *$\tilde{\mu}$ is a σ-stable process with parameter $\sigma = 1/2$ since $\tilde{\mu}(B) \sim \text{STABLE}(cP(B), 1/2)$ for any measurable B in $\mathbb{X}$. If $\tilde{p} \sim \text{NRMI}(c, \rho, P)$ we will write*

$$(\tilde{p}(B_1), \ldots, \tilde{p}(B_{M-1})) \sim \text{N-STABLE}\,(c_1, \ldots, c_M)\,,$$

with $c_m = cP(B_m)$ for $m = 1, \ldots, M$. If $c_m > 0$ for each m, its density function is

$$f(\boldsymbol{\omega}) = \frac{\Gamma(M/2) \prod_{m=1}^{M} c_m}{\Gamma(1/2)^M\, \mathcal{A}_M(\boldsymbol{\omega})^{M/2}} \left\{\omega_1 \cdots \omega_{M-1}(1 - |\boldsymbol{\omega}|)\right\}^{-3/2} I_{\mathcal{S}_{M-1}}(\boldsymbol{\omega}),$$

where, as before $\mathcal{A}_M(\boldsymbol{\omega}) = \sum_{m=1}^{M-1}(c_m^2/\omega_m) + c_M^2/(1 - |\boldsymbol{\omega}|)$. See Carlton (2002). A well–known property of the normalized stable process is that it does not depend on the total mass c and this is clearly reflected by the expression of the density function above.

3 Hierarchical NRMI-PY process

In order to define a prior Q_d that governs a d-dimensional partially exchangeable array $\{(X_{li})_{i\geq 1} : l = 1, \dots, d\}$, according to (1), we rely on (2) and resort to a special instance of hierarchical discrete random probabilities. More specifically, we will deal with the following setting

$$\begin{aligned} (X_{li} \mid \tilde{p}_l) &\overset{\text{ind}}{\sim} \tilde{p}_l, \quad i = 1, \dots, n^{(l)}, \quad l = 1, \dots, d, \\ (\tilde{p}_l \mid \tilde{p}_0) &\overset{\text{iid}}{\sim} \text{NRMI}(c, \rho, \tilde{p}_0), \quad l = 1, \dots, d, \\ \tilde{p}_0 &\sim \text{PY}(\sigma_0, c_0, P_0), \end{aligned} \tag{6}$$

where P_0 is a *diffuse* probability measure defined on $\mathbb{X}$. We will identify this model as a hierarchical NRMI-PY process. Notice that both the HDP (Teh et al., 2006) and the hierarchical stable process (Camerlenghi et al., 2019) can be recovered as particular cases.

A key feature of hierarchical species sampling models (2), and consequently also of the NRMI-PY process (6), is that with positive probability they induce ties among the X_{li}'s, because of the almost sure discreteness of both $(\tilde{p}_l \mid \tilde{p}_0)$ and $\tilde{p}_0$. Ties might occur both within and across groups, because the $(\tilde{p}_l \mid \tilde{p}_0)$ share the same discrete baseline measure, for $l = 1, \dots, d$. Thus, investigating the a priori clustering mechanism is of greater importance to highlight possible limitations induced by specific choices of $(\tilde{p}_l \mid \tilde{p}_0)$ and $\tilde{p}_0$. Indeed, compared to the HDP, specification (6) allows for a more flexible modeling of the clustering mechanism while still preserving analytical tractability: one can resort to the general theory set forth in Camerlenghi et al. (2019) in order to derive the partially exchangeable partition function, the full posterior characterization, and a closed form expression for the distribution of the number of clusters. See also Bassetti et al. (2019) for further developments in this direction. In addition, formulation (6) is also a suitable choice for computational reasons, as we will discuss in Section 4. Indeed, the stick–breaking construction of the PY process $\tilde{p}_0$ leads to a simple simulation strategy, both a priori and a posteriori, whereas NRMIs are a good candidate for each $(\tilde{p}_l \mid \tilde{p}_0)$ whenever it is relatively simple to study their finite-dimensional distribution, as discussed in Section 2.

An alternative representation of the model in (6) highlights a direct connection with a hierarchical collection of random weights following NIDP and GEM distributions, respectively. This approach provides a deeper understanding of the model and, in addition, has relevant computational advantages. Let us first recall that, in view of Definition 2, one has

$$\tilde{p}_l(\cdot) = \sum_{h=1}^{\infty} \frac{J_{lh}}{\sum_{h'=1}^{\infty} J_{lh'}} \delta_{\theta_{lh}}(\cdot), \qquad l = 1, \dots, d, \tag{7}$$

where $(\theta_{lh} \mid \tilde{p}_0) \overset{\text{iid}}{\sim} \tilde{p}_0$, for $h = 1, 2, \dots$, and $l = 1, \dots, d$. Moreover, the sequences of random jumps J_{lh} are independent from the locations θ_{lh} and also conditionally independent across groups, given $\tilde{p}_0$. As for the PY process $\tilde{p}_0$, we will refer to the notation used in Definition 3. From the above construction, each random probability measures $\tilde{p}_l$ places positive probability on the locations that are sampled from $\tilde{p}_0$. Because of the almost sure discreteness of $\tilde{p}_0$, one can equivalently rewrite (7) as follows

$$\tilde{p}_l(\cdot) = \sum_{h=1}^{\infty} \pi_{lh} \delta_{\phi_h}(\cdot), \qquad l = 1, \dots, d, \tag{8}$$

in which, conditionally on $\tilde{p}_0$, the locations ϕ_h are fixed whereas the "modified weights" $\boldsymbol{\pi}_l = (\pi_{l1}, \pi_{l2}, \dots)$ are

$$\pi_{lh} = \sum_{\{j:\, \theta_{lj} = \phi_h\}} \left(\frac{J_{lj}}{\sum_{h'}^{\infty} J_{lh'}} \right), \qquad h = 1, 2, \dots,$$

for any $l = 1, \dots, d$. Remarkably, the conditional law of the perturbed weights $\boldsymbol{\pi}_l$, given $\tilde{p}_0$, can be derived and it follows a NIDP process. This can be easily seen from additivity of NRMIs, since for any finite and measurable partition $\{B_1, \dots, B_M\}$ of $\mathbb{X}$,

$$(\tilde{p}_l(B_1), \dots, \tilde{p}_l(B_{M-1}) \mid \tilde{p}_0) = \Big(\sum_{j \in \mathcal{H}_1} \pi_{lj}, \dots, \sum_{j \in \mathcal{H}_{M-1}} \pi_{lj} \mid \tilde{p}_0 \Big),$$

where $\mathcal{H}_m = \{h \geq 1 : \ \phi_h \in B_m\}$, for $m = 1, \dots, M$, form a partition of $\mathbb{N}$. Then, we have that

$$\Big(\sum_{j \in \mathcal{H}_1} \pi_{lj}, \dots, \sum_{j \in \mathcal{H}_{M-1}} \pi_{lj} \mid \tilde{p}_0 \Big) \sim \text{NID}\Big(c \sum_{j \in \mathcal{H}_1} \pi_{0j}, \dots, c \sum_{j \in \mathcal{H}_M} \pi_{0j}; \rho \Big),$$

since $\tilde{p}_0(B_m) = \sum_{j \in \mathcal{H}_m} \pi_{0j}$, for any $m = 1, \dots, M$. This implies, by definition of a NIDP, that $(\boldsymbol{\pi}_l \mid \boldsymbol{\pi}_0) \overset{\text{iid}}{\sim} \text{NIDP}(c\boldsymbol{\pi}_0, \rho)$, for any $l = 1, \dots, d$. Now let us introduce a collection of assignment variables $G_{li} \in \{1, 2, \dots\}$, denoting the cluster membership of each observation, namely $X_{li} = \phi_{G_{li}}$. Then, we express model (6) in the following equivalent form

$$\begin{aligned} (G_{li} \mid \boldsymbol{\pi}_l) &\overset{\text{iid}}{\sim} \text{CATEGORICAL}(\boldsymbol{\pi}_l), \\ (\boldsymbol{\pi}_l \mid \boldsymbol{\pi}_0) &\overset{\text{iid}}{\sim} \text{NIDP}(c\boldsymbol{\pi}_0, \rho), \quad \boldsymbol{\pi}_0 \sim \text{GEM}(\sigma_0, c_0), \\ \phi_h &\overset{\text{iid}}{\sim} P_0, \qquad h \geq 1, \end{aligned} \tag{9}$$

for $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, d$. Specification (9) in the particular case of the HDP is already available from Teh et al. (2006) and it is extended here to the NRMI-PY process.

3.1 Deterministic truncation of the infinite process

Posterior inference for the NRMI-PY hierarchical processes of equation (6) is complicated by the infinite amount of parameters involved in the prior specification. A possible strategy for circumventing the problem is the marginalization with respect to the random probability measures $\tilde{p}_1, \dots, \tilde{p}_d, \tilde{p}_0$ to obtain generalized Pólya urn schemes that are building blocks of Gibbs samplers of the type proposed in Camerlenghi et al. (2019). This approach is very effective when one wants to approximate Bayesian point estimators under squared error loss or, more generally, evaluate linear functionals of the underlying posterior distribution. On the contrary, it is not ideal if one is interested in non–linear functionals such as those needed for determining credible intervals that are relevant for uncertainty quantification. In order to address the issue, we first introduce a deterministic truncation of the stick-breaking construction of the PY process. This obviously has a cascade effect also on the conditional distributions of the $\tilde{p}_l$'s, given such a truncated version of $\tilde{p}_0$, since they boil down

to finite–dimensional random elements, without the need of further approximations. More precisely, we approximate model (6) with the following truncated specification

$$\begin{aligned}(X_{li} \mid \tilde{p}_l^H) &\overset{\text{ind}}{\sim} \tilde{p}_l^H, \quad i = 1, \dots, n^{(l)}, \quad l = 1, \dots, d, \\ (\tilde{p}_l^H \mid \tilde{p}_0^H) &\overset{\text{iid}}{\sim} \text{NRMI}\Big(c\tilde{p}_0^H; \rho\Big), \quad l = 1, \dots, d, \\ \tilde{p}_0^H &\sim \text{PY}_H(\sigma_0, c_0, P_0),\end{aligned} \tag{10}$$

where $\tilde{p}_0^H \sim \text{PY}_H(\sigma_0, c_0, P_0)$ denotes a truncated PY process with H components. In other terms $\tilde{p}_0^H$ is obtained as in Definition 3, provided that $\nu_{0H} = 1$, so that $\sum_{h=1}^H \pi_{0h} = 1$ almost surely. Clearly, the truncated measure $\text{PY}_H(\sigma_0, c_0, P_0)$ converges weakly, almost surely, to a proper Pitman-Yor process as $H \to \infty$, and hence implying also the weak convergence, almost surely, of the bottom level NRMIs.

Remark 1. *The approximation strategy employed in (10) does not use peculiar properties of the* PY *process. Hence, this approach may be extended to other discrete random probability measure $\tilde{p}_0$ if an approximation $\tilde{p}_0^H$ were available. For example, one may alternatively assume that $\tilde{p}_0 \sim \text{NRMI}(c_0, \rho_0, P_0)$ and let $\tilde{p}_0^H$ be its* NID*-multinomial approximation (Lijoi et al., 2019). An advantage of the truncated* NRMI-PY *process relies on the fact that the density function of a truncated* GEM *distribution is available in closed form for general values of (σ_0, c_0), and this enables posterior inference via Gibbs sampling. However, this strategy might be adapted to general approximations $\tilde{p}_0^H$ as long as the density function of the weights $(\pi_{01}, \dots, \pi_{0H})$ associated to $\tilde{p}_0^H$ is available.*

An assessment of the effect of such a deterministic truncation can be obtained by determining an upper bound of the total variation distance between $\tilde{p}_l$ of the hierarchical process (6) and its finite-dimensional approximation $\tilde{p}_l^H$ in (10), for each $l = 1, \dots, d$. This can provide some guidance on the value at which H can be fixed. It is apparent that such an upper bound turns out to be random and we will rely on its expected value in order to gain some intuitive insight on the accuracy of the proposed truncation. To this end, we need to introduce $\tau_2(u) = \int_{\mathbb{R}^+} s^2 \mathrm{e}^{-us} \rho(\mathrm{d}s)$ and let $(x)_r = x(x+1)\cdots(x+r-1)$ denote the Pochammer symbol. Moreover, we recall that ψ is the Laplace exponent associated to ρ, i.e. $\psi(u) = \int_{\mathbb{R}^+} (1 - \mathrm{e}^{-us})\, \rho(\mathrm{d}s)$ for any $u > 0$.

Theorem 1. *Let $(\tilde{p}_1, \dots, \tilde{p}_d)$ be a hierarchical* NRMI-PY *process as in (6) and $(\tilde{p}_1^H, \dots, \tilde{p}_d^H)$ be the truncated version defined in (10). Then, for any $l = 1, \dots, d$,*

$$d_{\text{TV}}\left(\tilde{p}_l, \tilde{p}_l^H\right) = \sup_{A \in \mathscr{X}} \left|\tilde{p}_l(A) - \tilde{p}_l^H(A)\right| \leq R_{lH} = \sum_{h > H} \pi_{lh},$$

almost surely implying that

$$\mathbb{E}\left(d_{\text{TV}}\left(\tilde{p}_l, \tilde{p}_l^H\right)\right) \leq \mathbb{E}(R_{lH}) = \prod_{h=1}^{H} \frac{c_0 + \sigma_0 h}{c_0 + \sigma_0(h-1) + 1}.$$

In addition, set $\mathcal{R}_1(H) = \prod_{h=1}^H \frac{c_0+\sigma_0 h}{c_0+\sigma_0(h-1)+1}$ and $\mathcal{R}_2(H) = \prod_{h=1}^H \frac{(c_0+\sigma_0 h)_2}{(c_0+\sigma_0(h-1)+1)_2}$, then

$$Var(R_{lH}) = \mathcal{I}(c, \rho)\mathcal{R}_1(H) + (1 - \mathcal{I}(c, \rho))\mathcal{R}_2(H) - \mathcal{R}_1(H)^2,$$

where $\mathcal{I}(c, \rho) = c \int_{\mathbb{R}^+} u \mathrm{e}^{-c\psi(u)} \tau_2(u)\, du$.

Theorem 1 is similar in spirit to the one of Ishwaran and James (2001); Ishwaran and Zarepour (2002) in the exchangeable case, being based on *a priori* considerations. However, while we work directly with the underlying random probability measures, their results are based on the discrepancy between the marginal densities of a mixture model.

The upper bound R_{lH} has a simple interpretation: it consists on the part of $\boldsymbol{\pi}_l$ neglected by the truncation, and hence it is sometimes called *truncation error* in the exchangeable setting (Arbel et al., 2019). As a natural and intuitive consequence of Theorem 1, we get that $d_{\text{TV}}\left(\tilde{p}_l, \tilde{p}_l^H\right) \xrightarrow{\text{a.s.}} 0$ as $H \to \infty$. More importantly, the first two moments of R_{lH} can be used to determine a suitable truncation level H; for example, one might select the value of H such that the expected value of R_{lH} is below a certain threshold. Some further insight on R_{lH} may be gained by using the fact that

$$(R_{lH} \mid \boldsymbol{\pi}_0) \sim \text{NID}\Big(c\Big(1 - \sum_{h=1}^{H} \pi_{0h}\Big), c\sum_{h=1}^{H} \pi_{0h}; \rho\Big),$$

so that one can simulate its realizations, conditionally on $\boldsymbol{\pi}_0$. When $\tilde{p}_0$ is a Dirichlet process the expected value of the random variable R_{lH} goes to zero exponentially fast, meaning that H has not to be very large in practice. This is illustrated in the following example.

Example 4 (Truncated HDP)**.** *If $\rho(\,ds) = s^{-1}\mathrm{e}^{-s}$ and $\sigma_0 = 0$, then $\tilde{p}_0^H$ in (10) is a truncated Dirichlet process and the $\tilde{p}_l^H$ are, conditionally on $\tilde{p}_0^H$, iid draws from a Dirichlet distribution. Specializing Theorem 1 we get*

$$\mathbb{E}\left(d_{\text{TV}}\left(\tilde{p}_l, \tilde{p}_l^H\right)\right) \leq \left(\frac{c_0}{c_0+1}\right)^H.$$

Therefore, on average, the distance $d_{\text{TV}}\left(\tilde{p}_l, \tilde{p}_l^H\right)$ goes to zero exponentially fast as a function of H. Moreover,

$$Var(R_{lH}) = \frac{1}{c+1}\Big[c\left(\frac{c_0}{c_0+2}\right)^H - (c+1)\left(\frac{c_0}{c_0+1}\right)^{2H} + \left(\frac{c_0}{c_0+1}\right)^H\Big],$$

which is, again, exponentially decreasing as a function of H, implying that the upper bound R_{lH} is quite concentrated on its expected value for reasonably large values of H.

As apparent from Theorem 1, the parameters (c_0, σ_0) of the (truncated) PY process $\tilde{p}_0^H$ directly impact the quality of the approximation. Indeed, the expectation $\mathbb{E}(R_{lH})$ increases as a function of both c_0 and σ_0. However, if $\sigma_0 > 0$ the decay is not exponential anymore, implying that to achieve reasonable approximations we need a larger H, especially for values of σ_0 close to one. This is consistent with the discussions in Ishwaran and James (2001) and Arbel et al. (2019) in the exchangeable case. We will discuss the choice of H through examples in the simulation study of Section 5 and in the illustration of Section 6.

Another natural aspect that is worth pointing out is the dependence between $\tilde{p}_l^H$ and $\tilde{p}_{l'}^H$, for any $l \neq l'$, and how this differ from the one associated to the original hierarchical process specification in (2). To this end one can, for instance, evaluate the correlation between $\tilde{p}_l^H(A)$ and $\tilde{p}_{l'}^H(A)$ for any $A \in \mathscr{X}$ and truncation level H.

Theorem 2. *Let $(\tilde{p}_1, \ldots, \tilde{p}_d)$ be a hierarchical approximate* NRMI-PY *process as in (10). Then, for any $A \in \mathscr{X}$ such that $0 < P_0(A) < 1$ and any $l \neq l'$*

$$\mathrm{Corr}\left(\tilde{p}_l^H(A), \tilde{p}_{l'}^H(A)\right) = \frac{\mathcal{I}_0(\sigma_0, c_0, H)}{\mathcal{I}(c,\rho) + \mathcal{I}_0(\sigma_0, c_0, H)(1 - \mathcal{I}(c,\rho))}, \tag{11}$$

where $\mathcal{I}(c,\rho)$ is as in Theorem 1 and

$$\mathcal{I}_0(\sigma_0, c_0, H) = \sum_{h=1}^{H-1} \left[\frac{(1-\sigma_0)_2}{(1+c_0+(h-1)\sigma_0)_2} \prod_{l=1}^{h-1} \frac{(c_0+l\sigma_0)_2}{(1+c_0+(l-1)\sigma_0)_2} \right] + \prod_{h=1}^{H-1} \frac{(c_0+h\sigma_0)_2}{(1+c_0+(h-1)\sigma_0)_2}.$$

Moreover, taking the limit

$$\lim_{H\to\infty} \mathcal{I}_0(\sigma_0, c_0, H) = \frac{1-\sigma_0}{1+c_0},$$

which entails that $\mathrm{Corr}(\tilde{p}_l^H(A), \tilde{p}_{l'}^H(A))$ converges to the actual $\mathrm{Corr}(\tilde{p}_l(A), \tilde{p}_{l'}(A))$, implied by the model (6), as $H \to \infty$. It is apparent that the correlation coefficient is always positive and, unsurprisingly, does not depend on the specific set A as a consequence of homogeneity of the underlying random probability measures at the different levels of the hierarchy. As such, it is generally interpreted as an overall measure of dependence between the random probability measures.

Remark 2. *Note that the parameters c_0 and σ_0 do not play the same role as in the infinite dimensional case. Indeed, one can show that $\lim_{c_0\to\infty} \mathcal{I}_0(\sigma_0, c_0, H) = 1$, which clearly entails that $\lim_{c_0\to\infty} Corr(\tilde{p}_l^H(A), \tilde{p}_{l'}^H(A)) = 1$. On the other hand, it is clear that when $H = \infty$ one has the opposite limiting behaviour, namely that $\lim_{c_0\to\infty} Corr(\tilde{p}_l(A), \tilde{p}_{l'}(A)) = 0$, for any $l \neq l'$. Similar considerations can be made when considering $\sigma_0 \to 1$. The truncation effect that explains this different limiting dependence structure is quite intuitive: when either σ_0 or c_0 increase more mass is placed on the Hth atom of the stick-breaking construction, so that $\tilde{p}_0^H$ eventually converges to a point mass at ϕ_H. To sum up, if we let σ_0 (or c_0) be fixed and consider the correlation as a function of c_0 (or of σ_0) it first decreases until it reaches a minimum and, then, increases.*

Example 5 (Truncated HDP, cont'd.)**.** *In the* HDP *case, the above correlation can be significantly simplified. Indeed, a straightforward application of Theorem 2 yields*

$$Corr(\tilde{p}_l^H(A), \tilde{p}_{l'}^H(A)) = \frac{(1+c)\left(1 + c_0\left(\frac{c_0}{c_0+2}\right)^{H-1}\right)}{1 + c_0 + c\left(1 + c_0\left(\frac{c_0}{c_0+2}\right)^{H-1}\right)}.$$

In the infinite case $H \to \infty$ the correlation reduces to $(1+c)/(1+c_0+c)$, as already obtained in Camerlenghi et al. (2019). Thus, the truncation of $\tilde{p}_0$ induces a perturbation of the correlation of the HDP *through a factor which is exponentially decreasing in H.*

4 Hierarchical NRMI-PY mixture model

In several applied contexts the discreteness of the hierarchical NRMI-PY prior is not a realistic assumption. Nonetheless, we can adapt formulation (6) by adding a further level in the hierarchy, giving rise to a mixture model for partially exchangeable observations. Within the exchangeable framework, this idea was firstly suggested by Lo (1984), and discussed in practice for instance in Escobar and West (1995) in the Dirichlet process case, and e.g. by Barrios et al. (2013) for general homogeneous NRMIs.

Let Y_{li} for $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, d$ be a sample of observations taking values in a complete and separable metric space $\mathbb{Y}$ and let $K : \mathbb{Y} \times \mathbb{X} \to \mathbb{R}^+$ a transition kernel such that $y \mapsto K(y; x)$ is a density function on $\mathbb{Y}$, for any $x \in \mathbb{X}$, with respect to some dominating σ–finite measure. Exploiting representation (9), for any truncation level H the approximate hierarchical NRMI-PY mixture model is

$$\begin{aligned} (Y_{li} \mid G_{li}, \boldsymbol{\phi}) &\overset{\text{ind}}{\sim} K(y; \phi_{G_{li}}), \\ (G_{li} \mid \boldsymbol{\pi}_l) &\overset{\text{iid}}{\sim} \text{CATEGORICAL}(\pi_{l1}, \dots, \pi_{lH}), \\ (\boldsymbol{\pi}_l \mid \boldsymbol{\pi}_0) &\overset{\text{iid}}{\sim} \text{NID}(c\pi_{01}, \dots, c\pi_{0H}; \rho), \\ \boldsymbol{\pi}_0 \sim \text{GEM}_H(\sigma_0, c_0), &\quad \phi_h \overset{\text{iid}}{\sim} P_0, \end{aligned} \tag{12}$$

for $h = 1, \dots, H$, $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, d$, with $\boldsymbol{\phi} = (\phi_1, \dots, \phi_H)$ and $\boldsymbol{\pi}_l = (\pi_{l1}, \dots, \pi_{lH-1})$, and where $\text{GEM}_H(\sigma_0, c_0)$ denotes the truncated sequence of probabilities $\boldsymbol{\pi}_0 = (\pi_{01}, \dots, \pi_{0H-1})$, associated to the aforementioned truncated PY process. Also, we set $\pi_{lH} = 1 - |\boldsymbol{\pi}_l|$ for $l = 0, 1, \dots, d$. Marginalizing over the cluster indicators G_{li}, we obtain a finite mixture representation

$$(Y_{li} \mid \boldsymbol{\pi}_l, \boldsymbol{\phi}) \overset{\text{ind}}{\sim} f_l(y \mid \boldsymbol{\pi}_l, \boldsymbol{\phi}) = \sum_{h=1}^{H} \pi_{lh} K(y; \phi_h), \tag{13}$$

for $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, d$. As apparent from equations (12)-(13), under this hierarchical constructions the distributions $f_l(y \mid \boldsymbol{\pi}_l, \boldsymbol{\phi})$ for $l = 1, \dots, d$ share the same mixture components $K(y; \phi_h)$. However, they have different mixing weights $\boldsymbol{\pi}_l$, accounting for heterogeneity across groups. We remark that the conditional density $f_l(y \mid \boldsymbol{\pi}_l, \boldsymbol{\phi})$ is often of direct inferential interest and one may want to obtain its posterior distribution rather than just confining herself to a point estimate. In this case, one cannot rely on marginal algorithms that integrate out the random weights π_{lh} and a different (conditional) sampler must be adopted.

4.1 Blocked Gibbs sampler

In this Section we propose a simple Markov Chain Monte Carlo (MCMC) scheme that makes use of the approximate specification in equation (12) and enables posterior inference. The algorithms originally proposed for the HDP in Teh et al. (2006) are of marginal type, thus being characterized by their pros and cons: very effective for point estimation, but unreliable when it comes to uncertainty quantification. In the supplementary material of Fox et al. (2011) a conditional algorithm for the HDP is discussed, and it is based on a finite-dimensional approximation of $\tilde{p}_0$; however, its applicability is limited to the

HDP case. A general marginal algorithm for hierarchical NRMI processes and hierarchical PY processes was proposed by Camerlenghi et al. (2019). In this very same paper, the authors discuss also a conditional algorithm based on a representation of CRMs that can be traced back to Ferguson and Klass (1972). Its actual implementation must still rely on some truncation of the underlying infinite dimensional process that can be achieved through a specific approach as the one suggested, e.g., in Arbel and Prünster (2017). Since the representation in Ferguson and Klass (1972) displays jumps arranged in decreasing order, any truncation rule will retain the most relevant jumps. On the other hand, any computational procedure based on this construction will require the inversion of an underlying Lévy measure attainable and this may cause some computational issues.

The blocked Gibbs sampler we propose does not rely on the augmented scheme proposed in Camerlenghi et al. (2019) nor does it makes use of the (suitably truncated) Ferguson & Klass representation, while still being a conditional algorithm. Furthermore, the effect of the approximations can be explicitly assessed a priori thanks to Theorem 1. The main relevant constraint implied by our proposal is the availability of the density function $f(\boldsymbol{\pi}_l \mid \boldsymbol{\pi}_0)$ in closed form since it needs to be evaluated. Nonetheless there are some noteworthy examples of NRMIs that comply with this requirement, namely the Dirichlet process, the normalized inverse–Gaussian process, and the 1/2-stable process. See Section 2.

We now review the steps of the blocked Gibbs sampler, outlined in Algorithm 1, highlighting practical difficulties and suggesting possible solutions. Each step represents a full conditional distribution for a block of random variables, and we will denote with a "$-$" the conditioning to all the other variables.
Step [*1*]. Observations are randomly and independently allocated to different clusters. Since we have truncated the sequence of weights $\boldsymbol{\pi}_0$ up to the Hth term, the number of mixture component is finite. In turns, this implies that the normalizing constant can be obtained as a simple summation of the involved quantities.
Step [*2*]. The mixing probabilities $\boldsymbol{\pi}_l$ are sampled independently for $l = 1, \ldots, d$. In the Dirichlet case, the density $f(\boldsymbol{\pi}_l \mid -)$ is still a Dirichlet with updated parameters, thanks to conjugacy. For general NID distributions, the law of $(\boldsymbol{\pi}_l \mid -)$ has been recently obtained in closed form by Lijoi et al. (2019) when $\boldsymbol{\pi}_0 = (c_0/H, \ldots, c_0/H)$; the extension to general baseline probabilities $\boldsymbol{\pi}_0$ is a straightforward modification of their results. This in particular enables the exact sampling from the full-conditional $f(\boldsymbol{\pi}_l \mid -)$, without the need of Metropolis steps, for all the NID distributions discussed in this paper. See Lijoi et al. (2019) for the details.
Step [*3*]. The baseline mixing weights $\boldsymbol{\pi}_0$ are sampled. Notice that the vector $\boldsymbol{\pi}_0$ is a particular instance of a generalized Dirichlet distribution (Connor and Mosimman, 1969), and its density is

$$f(\boldsymbol{\omega}) = \frac{(1 - |\boldsymbol{\omega}|)^{c_0 + \sigma_0(H-1)-1}}{\prod_{h=1}^{H-1} \mathrm{Beta}(1 - \sigma_0, c_0 + h\sigma_0)} \prod_{h=1}^{H-1} \left[\omega_h^{-\sigma_0} \left(\sum_{j=h}^{H} \omega_j \right)^{-1} \right] I_{\mathcal{S}_{H-1}}(\boldsymbol{\omega}).$$

where $B(a, b)$ is the beta function evaluated at $a, b > 0$. Unfortunately, the full conditional $f(\boldsymbol{\pi}_0 \mid -)$ has no closed form—even in the Dirichlet case—and therefore we must resort to a Metropolis-Hastings step. Having tried

Algorithm 1: Steps of the Gibbs sampler

begin

Step [1] Assign each unit $i = 1, \dots, n^{(l)}$ and $l = 1, \dots, d$, to a mixture component;

for l *from* 1 *to* d **do**

for i *from* 1 *to* $n^{(l)}$ **do**

Sample $G_i \in (1, \dots, H)$ independently from the categorical variable with probabilities

$$\mathbb{P}(G_{li} = h \mid -) = \frac{\pi_{lh} K(Y_{li}; \phi_h)}{\sum_{h'=1}^{H} \pi_{lh'} K(Y_{li}; \phi_{h'})},$$

for every $h = 1, \dots, H$.

Step [2] Update the mixing parameters $\boldsymbol{\pi}_l$, for any $l = 1, \dots, d$;

for l *from* 1 *to* d **do**

Sample $\boldsymbol{\pi}_l$ independently from the full conditional having density proportional to

$$f(\boldsymbol{\pi}_l \mid -) \propto f(\boldsymbol{\pi}_l \mid \boldsymbol{\pi}_0) \prod_{h=1}^{H} \pi_{lh}^{n_{lh}},$$

where $n_{lh} = \sum_{i=1}^{n^{(l)}} I(G_{li} = h)$, and where $I(\cdot)$ denotes the indicator function.

Step [3] Sample the baseline mixing parameter $\boldsymbol{\pi}_0$ from the full conditional having density proportional to

$$f(\boldsymbol{\pi}_0 \mid -) \propto f(\boldsymbol{\pi}_0) \prod_{l=1}^{d} f(\boldsymbol{\pi}_l \mid \boldsymbol{\pi}_0).$$

Step [4] Update the kernel parameters ϕ_h, for any $h = 1, \dots, H$;

for h *from* 1 *to* H **do**

Sample the kernel parameters ϕ_h independently from the full conditional having density proportional to

$$f(\phi_h \mid -) \propto f(\phi_h) \prod_{(l,i) \in \mathcal{G}_h} K(Y_{li}; \phi_h),$$

where $\mathcal{G}_h = \{i = 1, \dots, n^{(l)}, l = 1, \dots, d : G_{li} = h\}$.

several different proposal distributions, we obtained very good performance by working in the unconstrained space $\boldsymbol{v}_0 = (v_{01}, \dots, v_{0H-1}) \in \mathbb{R}^{H-1}$—with $v_{0h} = \log(\pi_{0h}/\pi_{0H})$, for any $h = 1, \dots, H-1$—and then by applying a componentwise Gaussian random walk. The variances on the Gaussian proposal were adaptively and automatically selected as in Roberts and Rosenthal (2009).

Step [4]. The atoms ϕ_h are sampled independently for $h = 1, \dots, H$, proceeding as in the exchangeable setting and considering only within-cluster observations. The complexity of this sampling step depends both on the chosen kernel

K and on the prior distribution P_0. However, if the kernel belongs to an exponential family, then one might adopt a conjugate prior distribution (Diaconis and Ylvisaker, 1979), and hence simplify the computations.

As a final remark, we notice that the deterministic truncation allows for the implementation of other well-established MCMC techniques, essentially because it shifts the original nonparametric formulation to a finite-dimensional problem, whose likelihood and prior distribution can be readily evaluated. As such, automatic tools like STAN (Carpenter et al., 2017) might be used for posterior inference.

5 Simulation study

To assess the empirical performance of model (12) and the associated Gibbs sampling algorithm, we conduct a simple simulation study. The target of this analysis is the comparison between the HDP and more general hierarchical processes in terms of inference on the clustering structure of the data.

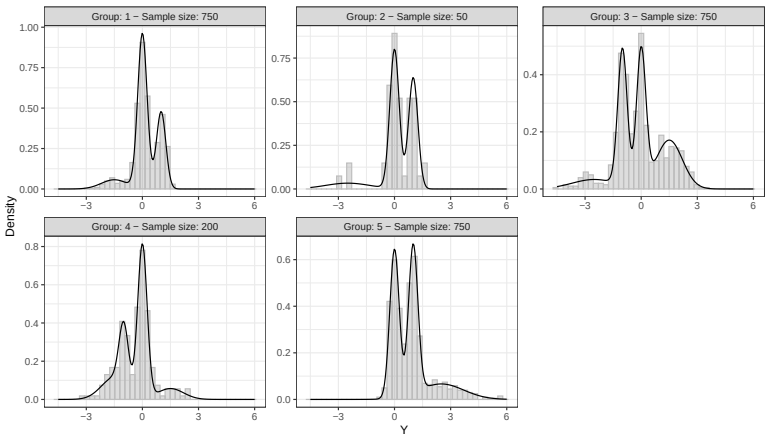

Figure 1: Graphical representation of the dataset used in the simulation study. Solid lines represent the true densities from which the data were simulated.

We consider $n = 2500$ observations divided in $d = 5$ different groups, each having a different sample size, precisely $(n^{(1)}, \dots, n^{(5)}) = (750, 50, 750, 200, 750)$. Within group, the simulated data are iid draws from a group-specific finite mixture of Gaussian distributions, whereas across groups they are independently sampled. The Gaussian mixtures densities were chosen so that different groups share some mixture components. In particular, there are a total of 7 latent Gaussian mixture components having means $(-2.5, -1.5, -1, 0, 1, 1.5, 2.5)$ and standard deviations $(1.2, 0.7, 0.25, 0.25, 0.25, 0.7, 1.2)$, which are split over the $d = 5$ groups, as reported in Table 1. For instance, the mixture component with 0 mean and standard deviation 0.25 is shared by all the groups. The mixing proportions are not uniform within groups nor equal across groups: this means, for example, that some mixture components are specific of two groups but they are not shared by the other three. The simulated dataset is illustrated in Figure 1.

In the hierarchical mixture (12), we employ a Gaussian kernel $K(y; \phi_h) = \mathrm{N}\left(y; \mu_h, \tau_h^{-2}\right)$, and we choose a conditionally conjugate prior distribution for

		Mixture component						
		1	2	3	4	5	6	7
Group	1	0.0	0.1	0.0	0.6	0.3	0.0	0.0
	2	0.1	0.0	0.0	0.5	0.4	0.0	0.0
	3	0.1	0.0	0.3	0.3	0.0	0.3	0.0
	4	0.0	0.2	0.2	0.5	0.0	0.1	0.0
	5	0.0	0.0	0.0	0.4	0.4	0.0	0.2

Table 1: True mixing proportions of the simulated data for each group $l = 1, \ldots, 5$, and for each of the 7 mixture components.

Model	c	c_0	σ_0	Correlation	Expected # clusters	$\mathbb{E}(R_{lH})$	H
HDP	18	13	0	0.59	≈ 41	$< 10^{-6}$	250
HDP-PY	7	5	0.5	0.43	≈ 40	0.042	250
HST-PY	-	7	0.5	0.12	≈ 39	0.057	250
HIG-PY	2.5	2	0.5	0.50	≈ 40	0.020	250

Table 2: Hyperparameter settings for each hierarchical mixture model. The correlation coefficient is evaluated using Theorem 2. The expected number of clusters is obtained via Monte Carlo simulations, averaging over 100′000 values from the truncated prior. The expected value of the upper bound R_{lH}, defined as in Theorem 1, is also reported.

the random locations $\phi_h = (\mu_h, \tau_h^{-2})$, so that their baseline measure is

$$P_0\left(\mathrm{d}\mu, \mathrm{d}\tau^2\right) = P_{0,1}(\mathrm{d}\mu) P_{0,2}\left(\mathrm{d}\tau^2\right),$$

where $P_{0,1}$ is a Gaussian distribution with mean 0 and standard deviation 10, whereas $P_{0,2}$ is a Gamma distribution with parameters $(1, 1)$. To simplify our treatment, we decided not to place any hyperprior distribution on the parameters in P_0, although this further hierarchical layer could be easily handled with a straightforward modification of the blocked Gibbs sampler in Algorithm 1.

We fitted four different hierarchical mixture models to the same simulated dataset, for different choices of the jump measure $\rho(\mathrm{d}s)$ and of the hyperparameters c, c_0 and σ_0, whose value are presented in Table 2. These models include: i) a hierarchical Dirichlet Process (HDP); ii) a hierarchical Dirichlet and Pitman-Yor process (HDP-PY); iii) a hierarchical 1/2-stable and Pitman-Yor process (HST-PY); iv) a hierarchical normalized inverse Gaussian and Pitman-Yor process (HIG-PY). Notice that in the 1/2-stable case the total mass parameter is irrelevant and therefore it was omitted. We fixed a common truncation level $H = 250$, which we found to be sufficiently large to guarantee a good approximation of the infinite hierarchical mixture model. Indeed, in Table 2 we also report the expected value of upper bound R_{lH}, defined as in Theorem 1, which in the worst case scenario is approximately equal to 0.06.

The hyperparameters c, c_0 and σ_0 were selected so that peculiar characteristics of each model can be appreciated—especially compared to the HDP. In particular, the a priori expected number of clusters, obtained via Monte Carlo after averaging over 100′000 draws from the truncated prior in (10), is centered approximately around 40, as reported in Table 2 and depicted in Figure 2. That is, we set on purpose the a priori expected number of clusters to be much

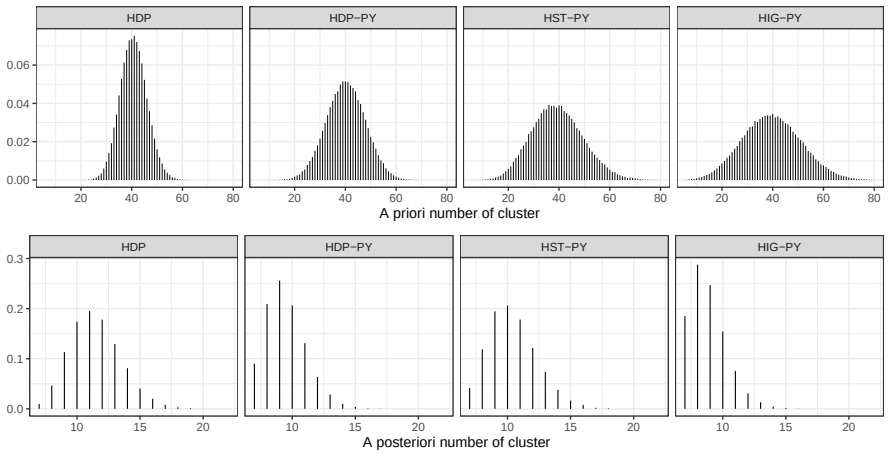

Figure 2: Top figures: a priori distribution of the number of clusters, based on 100′000 simulations from the truncated prior. Bottom figures: a posteriori distribution of the number of clusters, based on 20′000 MCMC draws. Both top and bottom figures refers to the models in Table 2.

higher than the true number of mixture components. An extensive description of the underlying clustering behaviors is beyond the aim of this paper, and one can refer e.g. to Lijoi et al. (2007); De Blasi et al. (2015) in the exchangeable case and to Camerlenghi et al. (2019) in the partially exchangeable setting with hierarchical processes. To our purposes, it suffices to notice that the a priori distribution of the number of clusters is much "flatter"—i.e. less informative—in general hierarchical mixture models compared to the one of the HDP, as empirically evidenced in Figure 2. This is due to the stable parameter σ_0 in the Pitman-Yor specification, but also to the specific choice of jump measure ρ. For example, the normalized inverse Gaussian distribution might be regarded as less informative compared to the Dirichlet (Lijoi et al., 2005), essentially leading to a flatter cluster configuration. Thus, we aim at showing that hierarchical models beyond the HDP might be more robust in identifying a suitable number of components, especially in severely misspecified prior settings. This behavior was already noticed in Lijoi et al. (2007) for exchangeable data, and extend to the case of truncated hierarchical processes.

We run the chain for 200′000 iterations—after a burn-in period of 100′000 draws—and we thin the chain every 10 iterations, thus comprising a total of 20′000 posterior samples. The traceplots show good mixing and no evidence against convergence. As expected, the posterior distribution of the number of clusters—depicted in the bottom row of Figure 2—differs across models: in the HDP the values having highest probabilities are located between 10 and 12, whereas in all the other cases the posterior distribution is shifted towards 7, the correct number of mixture components. This is particularly evident in the HIG-PY case, whose a priori distribution was indeed the less informative.

6 Illustration

To further corroborate the practical relevancy of the proposed conditional algorithm, in this section we discuss an application of the NRMI-PY process to latent

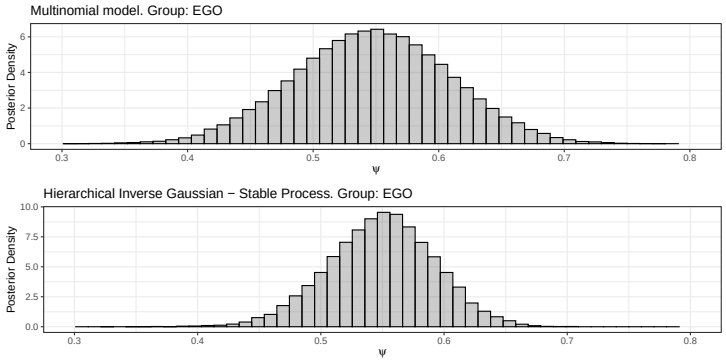

Figure 3: Posterior distribution of ψ_1, group: EGO. Top: posterior distribution of ψ_1 under the alternative multinomial formulation. Bottom: posterior distribution of ψ_1 under the HIG-ST mixture model of equation (14).

class analysis, in presence of qualitative covariates (Lazarsfeld and Henry, 1968; Goodman, 1974; Hagenaars and McCutcheon, 2002). As an illustration, we analyze the dataset presented in Stouffer and Toby (1951) and reported in Appendix A.3. This has been the object of several investigations (e.g. Goodman, 1974, 1975; Clogg and Goodman, 1986; Hagenaars and McCutcheon, 2002) through latent class analysis, and from a frequentist perspective. The data are based on a short questionnaire completed by $n = 648$ undergraduate students at Harvard and Radcliffe, in 1950. Four ethical dilemmas, denoted as A,B,C and D, were posed to these students: a response coded as 1 represents a preference towards particularistic values, and conversely 0 indicates a preference towards universalistic values. The questions were presented in slightly different forms to $d = 3$ independent and equally sized groups of students, meaning $n^{(1)} = n^{(2)} = n^{(3)} = 216$. The first group received each dilemma so that it refers to themselves (EGO), the second group so that it refers to a stranger (SMITH), and the third group so that it refers to a friend (FRIEND).

Clearly, some degree of agreement of the responses among different groups is expected, since the ethical dilemmas are the same. Nonetheless, the three groups should not be treated as identical, because the way in which each dilemma is posed might influence the response. Hence, within a Bayesian framework, the partial exchangeability assumption seem fairly natural in this setting, and it provides practical advantages. In particular, it allows to borrow information across groups and therefore to take stronger inferential conclusion compared to single-group analyses. Relying on the notation of Section 4, we assume that our observations are drawn from a collection of partially exchangeable binary random vectors $Y_{li} = (Y_{li1}, \ldots, Y_{li4}) \in \{0,1\}^4$, for $i = 1, \ldots, 216$ and $l = 1, 2, 3$, where the components of each Y_{li} refer to items A,B,C and D, respectively.

Latent class models are essentially mixture models in which, given a latent class (cluster) indicator G_{li}, the qualitative random variables $(Y_{li1}, \ldots, Y_{li4})$ are mutually independent. However, as noted in Dunson and Xing (2009), in this setting it is not straightforward to obtain a well-justified estimate for the number of mixture components. In addition, they proved that any probability mass function $\mathbb{P}(Y_{li} = y_{li})$ can be represented in terms of a latent class mixture model, when the number of mixture components is large enough. This leads us

to assuming a mixture model with infinitely many components that we truncate up to the Hth term, thus obtaining a flexible and theoretically justified model for contingency tables. Hence, we can extend the approach of Dunson and Xing (2009) to the partially exchangeable setting, whereby d groups of contingency tables are observed and a product of multinomial kernel in the NRMI-PY mixture model of equations (12)-(13) is specified. More precisely

$$\begin{aligned} Y_{li} \overset{\text{ind}}{\sim} \mathbb{P}(Y_{li1} = y_{li1}, \ldots, Y_{li4} = y_{li4} \mid \boldsymbol{\pi}_l, \boldsymbol{\phi}) = \\ = \sum_{h=1}^{H} \pi_{lh} \left(\prod_{k=1}^{4} \phi_{hk}^{y_{lik}} (1 - \phi_{hk})^{1 - y_{lik}} \right), \end{aligned} \tag{14}$$

independently for $i = 1, \ldots, 216$, and $l = 1, 2, 3$, where $\boldsymbol{\pi}_l = (\pi_{l1}, \ldots, \pi_{lH})$ has the same hierarchical prior distribution as in equation (12) and where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_H)$ is such that $\phi_h = (\phi_{h1}, \ldots, \phi_{h4})$ for any $h = 1, \ldots, H$. As for the baseline measure P_0, we selected a uniform prior over the space $(0, 1)^4$, which is conditionally conjugate and hence facilitates posterior computations. A possible alternative specification for P_0 consists of independent beta distributions, for $k = 1, \ldots, 4$, which would still preserve conjugacy while allowing for the inclusion of more specific prior information in the model.

As for the prior setting of $\boldsymbol{\pi}_l$, we specified a hierarchical normalized inverse–Gaussian and stable process (NIG-ST), with hyperparameter settings $c = 1/2$, $c_0 = 0$ and $\sigma_0 = 3/10$ and with a truncation level $H = 150$. We achieve a good approximation of the infinite dimensional process, since $\mathbb{E}(R_{lH}) < 10^{-4}$. Moreover, this specification induces high correlation a priori (to be meant in terms of the statement of Theorem 2) among the random probability measures $\tilde{p}_l$ (≈ 0.86): this is consistent with our prior belief that the same ethical dilemma should lead to very similar responses, regardless the way it was presented. The a priori expected number of clusters, evaluated via Monte Carlo, is approximately 3.9; however, the a priori distribution of the number of clusters is quite dispersed, consistently with the findings of previous analyses, which indeed do not provide a univocal recommendation about the number of latent components (Stouffer and Toby, 1951; Goodman, 1974, 1975; Clogg and Goodman, 1986). Posterior inference was conducted via MCMC, using the blocked Gibbs sampler described in Section 4. We run the chain for 200′000 iterations—after a burn-in period of 50′000 draws—and we thin the chain every 10 iterations, thus comprising a total of 20′000 posterior samples. The traceplots show good mixing and no evidence against convergence.

In Clogg and Goodman (1986) it is suggested that these dilemmas can be ordered (D $\to$ C $\to$ B $\to$ A), according to a Guttman scale. This means, for instance, that a negative answer to C should imply, on average, also a negative response to dilemmas B and A. While such an assumption greatly simplifies the analysis, it seems clear from the subsequent results that it can only provide a reasonable approximation of the phenomenon. Indeed, we aim at studying for instance the conditional probability of B $= 1$ given that C $= 0$, D $= 1$ for each group of respondents, which should be close to zero under the Guttman scale assumption. As it will turn out, these probabilities not only are away from zero, but they are also significantly greater than $1/2$. More formally, we are interested in the posterior distribution of

$$\psi_l = \mathbb{P}(Y_{li2} = 1 \mid Y_{li3} = 0, Y_{li4} = 1, \boldsymbol{\pi}_l, \boldsymbol{\phi}),$$

for any group $l = 1, 2, 3$, and given the data. Once more, we remark that the posterior distribution of each ψ_l can be obtained only through conditional algorithms, which therefore represent the only possible choice to conduct inference in this specific application.

In Figure 3 we compare the posterior distribution of ψ_1 (EGO group) obtained using the aforementioned HIG-ST model in equation (14), with the posterior distribution of ψ_1 obtained under a much simpler multinomial model. More precisely, under the alternative model we treat the $2^4 = 16$ possible combination of responses as mutually exclusive categories. Among groups, we assume full heterogeneity—i.e. independence—whereas within group observations are conditionally iid draws from a multinomial distribution having 16 possible outcomes, and with a uniform prior. In both cases, the posterior distribution of ψ_1 is far from zero, suggesting that the Guttman scaling adopted in Clogg and Goodman (1986) should be interpreted with care. However, as apparent from Figure 3, our HIG-ST model is able to substantially reduce the posterior uncertainty compared to the benchmark multinomial model. Essentially, this is due to two reasons: i) the latent class representation of equation (14), albeit flexible, allows for a parsimonious characterization of the distribution function $\mathbb{P}(Y_{li} = y_{li})$ compared to the alternative multinomial formulation (Dunson and Xing, 2009); ii) in our hierarchical formulation we flexibly borrow information across the three groups, and this translates in a lower variability of the the posterior distribution. The posterior distributions of ψ_2, ψ_3 for other groups (SMITH, FRIEND), lead to similar conclusions.

Acknowledgements

Most of the paper was completed while T. Rigon was a Ph.D. student at the Bocconi University, Milano. A. Lijoi and I. Prünster were partially supported by MIUR, PRIN Project 2015SNS29B. T. Rigon was partially supported by grant R01ES027498 of the National Institute of Environmental Health Sciences of the United States National Institutes of Health.

References

Arbel, J., De Blasi, P., and Prünster, I. (2019). Stochastic approximations to the Pitman-Yor process. *Bayesian Analysis*, 14(4):1201–1219.

Arbel, J. and Prünster, I. (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*, 27(1):3–17.

Argiento, R., Bianchini, I., and Guglielmi, A. (2016). A blocked gibbs sampler for ngg-mixture models via a priori truncation. *Statistics and Computing*, 26(3):641–666.

Argiento, R., Cremaschi, A., and Vannucci, M. (2019). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association*, Forthcoming.

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statistical Science*, 28(3):313–334.

Bassetti, F., Casarin, R., and Rossini, L. (2019). Hierarchical species sampling models. *Bayesian Analysis*, Forthcoming.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). Distribution theory for hierarchical processes. *The Annals of Statistics*, 49(1):67–92.

Carlton, M. A. (2002). A family of densities derived from the three-parameter Dirichlet process. *Journal of Applied Probability*, 39:764—774.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddel, A. (2017). Stan: a probabilistic programming language. *Journal of Statistical Software*, 76(1):1—32.

Cifarelli, D. and Regazzini, E. (1978). Problemi statistici non parametrici in condizioni di scambiabilità parziale. Technical report, Quaderni Istituto Matematica Finanziaria, Università di Torino Serie III, 12.

Clogg, C. C. and Goodman, L. A. (1986). On scaling models applied to data from several groups. *Psychometrika*, 51(1):123–135.

Connor, R. J. and Mosimman, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194– 206.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition. Elementary theory and methods.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):212–229.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate prior for exponential families. *The Annals of Statistics*, 7(2):269–292.

Dunson, D. B. and Xing, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Favaro, S., Hadjicharalambous, G., and Prünster, I. (2011). On a class of distributions on the simplex. *Journal of Statistical Planning and Inference*, 141(9):2987–3004.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Ferguson, T. S. and Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Annals of Mathematical Statistics*, 43(5):1634—1643.

Fox, E. B., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A):1020–1056.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unindentifiable models. *Biometrika*, 61(2):215–231.

Goodman, L. A. (1975). A new model for scaling response patterns: an application of quasi independence concept. *Journal of the American Statistical Association*, 70(352):755–768.

Griffin, J. E. and Leisen, F. (2017). Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 79(2):525–545.

Hagenaars, J. A. and McCutcheon, A. L. (2002). *Applied Latent Class Analysis*. Cambridge University Press.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canadian Journal of Statistics*, 30(2):269–283.

James, L. F., Lijoi, A., and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, 33(1):105–120.

Kingman, J. F. C. (1975). Random discrete distribution. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 37:1–22.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 69(4):715–740.

Lijoi, A. and Nipoti, B. (2014). A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association*, 109(506):802–814.

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

Lijoi, A., Nipoti, B., and Prünster, I. (2014b). Dependent mixture models: clustering and borrowing information. *Computational Statistics and Data Analysis*, 71:417–433.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In Hjort, N. L., Holmes, C. C., Muller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*. Cambridge University Press.

Lijoi, A., Prünster, I., and Rigon, T. (2019). Finite-dimensional discrete random structures and Bayesian clustering. Technical report, Collegio Carlo Alberto.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357.

MacEachern, S. N. (1999). Dependent nonparametric processes. In *ASA proceedings of the section on Bayesian statistical science*, Alexandria, VA: American Statistical Association, pages 50–55.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canadian Journal of Statistics*, 26(2):283–297.

Perman, M. (1990). *Random discrete distributions derived from subordinators*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–University of California, Berkeley.

Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory Related Fields*, 92(1):21–39.

Pitman, J. (1996). Some Developments of the Blackwell-Macqueen Urn Scheme. *Statistics, Probability and Game Theory*, 30:245–267.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.

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

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

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639 – 650.

Stouffer, S. A. and Toby, J. (1951). Role conflict and personality. *American Journal of Sociology*, 56(5):395–406.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In Hjort, N. L., Holmes, C. C., Muller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*. Cambridge University Press.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1–41.

Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., and Vannucci, M. (2016). A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data. *The Annals of Applied Statistics*, 10(2):638–666.

A Appendix

A.1 Proof of Theorem 1

Recall that $(\tilde{p}_1, \dots, \tilde{p}_d)$ comes from a hierarchical NRMI-PY process as in (6). Moreover, let $(\tilde{p}_1^H, \dots, \tilde{p}_d^H)$ be the hierarchical approximate process NRMI-PY defined in (10), with truncation level H. Then for any $A \in \mathscr{X}$, and exploiting representation (8), we have that almost surely

$$\begin{aligned}
\left|\tilde{p}_l(A) - \tilde{p}_l^H(A)\right| &= \left| \sum_{h=1}^{\infty} \pi_{lh}\delta_{\phi_h}(A) - \left(\sum_{h=1}^{H-1} \pi_{lh}\delta_{\phi_h}(A) + \left(1 - \sum_{h=1}^{H-1} \pi_{lh}\right) \delta_{\phi_H}(A)\right)\right| \\
&= \left|\pi_{lH}\delta_{\phi_H}(A) + \sum_{h>H} \pi_{lh}\delta_{\phi_h}(A) - \left(1 - \sum_{h=1}^{H-1} \pi_{lh}\right) \delta_{\phi_H}(A)\right| \\
&= \left| \delta_{\phi_H}(A) \sum_{h>H} \pi_{lh} - \sum_{h>H} \pi_{lh}\delta_{\phi_h}(A)\right| \leq \sum_{h>H} \pi_{lh} = R_{lH}.
\end{aligned}$$

Note that $\sum_{h>H} \pi_{lh} \geq \sum_{h>H} \pi_{lh}\delta_{\phi_h}(A)$. Hence, if $\delta_{\phi_H}(A) = 0$, then the last inequality easily follows, and the same holds true if $\delta_{\phi_H}(A) = 1$ almost surely. Hence,

$$d_{\text{TV}}\left(\tilde{p}_l, \tilde{p}_l^H\right) = \sup_{A \in \mathscr{X}} \left|\tilde{p}_l(A) - \tilde{p}_l^H(A)\right| \leq R_{lH} = \sum_{h>H} \pi_{lh},$$

almost surely. Moreover, notice that

$$\left(\sum_{h>H} \pi_{lh} \mid \boldsymbol{\pi}_0\right) \sim \text{NID}\left(c\left(1 - \sum_{h=1}^{H} \pi_{0h}\right), c\sum_{h=1}^{H} \pi_{0h}; \rho\right),$$

from which it follows that the expected value is equal to

$$\begin{aligned}
\mathbb{E}\left(\sum_{h>H} \pi_{lh}\right) = \mathbb{E}\left(\mathbb{E}\left(\sum_{h>H} \pi_{lh} \mid \boldsymbol{\pi}_0\right)\right) = \mathbb{E}\left(\sum_{h>H} \pi_{0h}\right) = \\
\prod_{h=1}^{H} \mathbb{E}\left(1 - \nu_{0h}\right) = \prod_{h=1}^{H} \frac{c_0 + \sigma_0 h}{c_0 + \sigma_0(h-1) + 1}.
\end{aligned}$$

Now recall that $\mathcal{I}(c, \rho) = c\int_{\mathbb{R}^+} u\mathrm{e}^{-c\psi(u)}\tau_2(u)\,\mathrm{d}u$ with $\tau_2(u) = \int_{\mathbb{R}^+} s^2\mathrm{e}^{-us}\rho(\,\mathrm{d}s)$ and let $R_{0H} = \sum_{h>H} \pi_{0h}$, then

$$\begin{aligned}
\text{Var}\left(R_{lH}\right) &= \mathbb{E}\left(\text{Var}\left(\sum_{h>H} \pi_{lh} \mid \boldsymbol{\pi}_0\right)\right) + \text{Var}\left(\mathbb{E}\left(\sum_{h>H} \pi_{lh} \mid \boldsymbol{\pi}_0\right)\right) = \\
&= \mathcal{I}(c, \rho)\mathbb{E}\left(R_{0H} - R_{0H}^2\right) + \mathbb{E}(R_{0H}^2) - \mathbb{E}\left(R_{0H}\right)^2,
\end{aligned}$$

where $\mathbb{E}(R_{0H})$ can be computed as before and

$$\mathbb{E}\left(R_{0H}^2\right) = \prod_{h=1}^{H} \mathbb{E}\left(\left(1 - \nu_{0h}\right)^2\right) = \prod_{h=1}^{H} \frac{(c_0 + \sigma_0 h)_2}{(c_0 + \sigma_0(h-1) + 1)_2},$$

recalling that $(x)_r = x(x+1)\cdots(x+r-1)$ denotes the Pochhammer symbol.

A.2 Proof of Theorem 2

First of all, notice that the expected value of the truncated Pitman-Yor process $\tilde{p}_0^H \sim \text{PY}_H(\sigma_0, c_0, P_0)$, for any $A \in \mathscr{X}$ and any $H = 1, 2, \ldots$, is equal to the baseline measure

$$\mathbb{E}(\tilde{p}_0^H(A)) = \sum_{h=1}^{H} \mathbb{E}(\pi_{0h})\mathbb{E}(\delta_{\phi_h}(A)) = P_0(A).$$

Moreover, one can show that

$$\text{Var}(\tilde{p}_0^H(A)) = P_0(A)(1 - P_0(A)) \sum_{h=1}^{H} \mathbb{E}(\pi_{0h}^2),$$

for any $H = 1, 2, \ldots$, and $A \in \mathscr{X}$. Define $\mathcal{I}_0(\sigma_0, c_0, H) = \sum_{h=1}^{H} \mathbb{E}(\pi_{0h}^2)$ and recall that $\mathcal{I}(c, \rho) = c \int_{\mathbb{R}^+} u e^{-c\psi(u)} \tau_2(u)\, \mathrm{d}u$ with $\tau_2(u) = \int_{\mathbb{R}^+} s^2 e^{-us} \rho(\mathrm{d}s)$. From Proposition 1 of James et al. (2006), one has that $\text{Var}(\tilde{p}_l^H(A) \mid \tilde{p}_0^H) = P_0(A)(1 - P_0(A))\mathcal{I}(c, \rho)$ for any $A \in \mathscr{X}$. Hence, for any $l = 1, \ldots, d$,

$$\begin{aligned}
\text{Var}(\tilde{p}_l^H(A)) &= \mathbb{E}(\text{Var}(\tilde{p}_l^H(A) \mid \tilde{p}_0^H)) + \text{Var}(\tilde{p}_0^H(A)) \\
&= \mathcal{I}(c, \rho)\mathbb{E}(\tilde{p}_0^H(A)(1 - \tilde{p}_0^H(A))) + \\
&\quad + P_0(A)(1 - P_0(A))\mathcal{I}_0(\sigma_0, c_0, H) \\
&= P_0(A)(1 - P_0(A))(\mathcal{I}(c, \rho) - \mathcal{I}(c, \rho)\mathcal{I}_0(\sigma_0, c_0, H) + \\
&\quad + \mathcal{I}_0(\sigma_0, c_0, H)).
\end{aligned}$$

Moreover, following Camerlenghi et al. (2019, Appendix A.1), for any $l \neq l'$

$$\begin{aligned}
\text{Cov}(\tilde{p}_l^H(A), \tilde{p}_{l'}^H(A)) &= \text{Var}(\tilde{p}_0^H(A)) = \\
&= P_0(A)(1 - P_0(A))\mathcal{I}_0(\sigma_0, c_0, H),
\end{aligned}$$

from which it follows that

$$\text{Cor}(\tilde{p}_l^H(A), \tilde{p}_{l'}^H(A)) = \frac{\mathcal{I}_0(\sigma_0, c_0, H)}{\mathcal{I}(c, \rho) + \mathcal{I}_0(\sigma_0, c_0, H)(1 - \mathcal{I}(c, \rho))}.$$

It remains to find the explicit formulation of $\mathcal{I}_0(\sigma_0, c_0, H)$, being equal to

$$\begin{aligned}
\mathcal{I}_0(\sigma_0, c_0, H) &= \sum_{h=1}^{H} \mathbb{E}(\pi_{0h}^2) = \sum_{h=1}^{H} \mathbb{E}\left(\nu_{0h}^2 \prod_{l=1}^{h-1} (1 - \nu_{0l})^2\right) = \\
&= \sum_{h=1}^{H-1} \left[\frac{(1 - \sigma_0)_2}{(1 + c_0 + (h-1)\sigma_0)_2} \left(\prod_{l=1}^{h-1} \frac{(c_0 + l\sigma_0)_2}{(1 + c_0 + (l-1)\sigma_0)_2}\right)\right] + \\
&\quad + \left(\prod_{l=1}^{H-1} \frac{(c_0 + l\sigma_0)_2}{(1 + c_0 + (l-1)\sigma_0)_2}\right).
\end{aligned}$$

Notice that all the above results hold also for the infinite case, having replaced $\mathcal{I}_0(\sigma_0, c_0, H)$ with its limit $\mathcal{I}_0(\sigma_0, c_0)$, so that

$$\begin{aligned}
\lim_{H \to +\infty} \mathcal{I}_0(\sigma_0, c_0, H) = \mathcal{I}_0(\sigma_0, c_0) &= \mathbb{E}\left(\sum_{h=1}^{\infty} \pi_{0h}^2\right) = \\
&= \sum_{h=1}^{\infty} \mathbb{E}\left(\pi_{0h}^2\right) = \frac{1 - \sigma_0}{1 + c_0},
\end{aligned}$$

where the last equality follows for instance from Ishwaran and James (2001, Appendix A.2).

A.3 Dataset

We report in Table 3 the dataset used in the illustrative analysis of Section 6 and originally presented in Stouffer and Toby (1951).

A	B	C	D	EGO	SMITH	FRIEND
0	0	0	0	42	37	35
0	0	0	1	23	31	17
0	0	1	0	6	6	9
0	0	1	1	25	15	26
0	1	0	0	6	5	3
0	1	0	1	24	29	27
0	1	1	0	7	6	3
0	1	1	1	38	25	32
1	0	0	0	1	2	3
1	0	0	1	4	4	5
1	0	1	0	1	3	2
1	0	1	1	6	4	5
1	1	0	0	2	3	0
1	1	0	1	9	23	20
1	1	1	0	2	3	3
1	1	1	1	20	20	26
Total				216	216	216

Table 3: The Stouffer and Toby (1951) dataset. We report the frequencies for each possible combination of the $2^4 = 16$ responses, divided over the the three groups EGO, SMITH and FRIEND.