Bayesian Analysis (2019) **14**, Number 4, pp. 1303–1356

Latent Nested Nonparametric Priors (with Discussion)

Federico Camerlenghi[*†], David B. Dunson[‡], Antonio Lijoi[§], Igor Prünster[§] and Abel Rodríguez[¶]

Abstract. Discrete random structures are important tools in Bayesian nonparametrics and the resulting models have proven effective in density estimation, clustering, topic modeling and prediction, among others. In this paper, we consider nested processes and study the dependence structures they induce. Dependence ranges between homogeneity, corresponding to full exchangeability, and maximum heterogeneity, corresponding to (unconditional) independence across samples. The popular nested Dirichlet process is shown to degenerate to the fully exchangeable case when there are ties across samples at the observed or latent level. To overcome this drawback, inherent to nesting general discrete random measures, we introduce a novel class of latent nested processes. These are obtained by adding common and group-specific completely random measures and, then, normalizing to yield dependent random probability measures. We provide results on the partition distributions induced by latent nested processes, and develop a Markov Chain Monte Carlo sampler for Bayesian inferences. A test for distributional homogeneity across groups is obtained as a by-product. The results and their inferential implications are showcased on synthetic and real data.

AMS 2000 subject classifications: Primary 60G57, 62G05, 62F15.

Keywords: Bayesian nonparametrics, completely random measures, dependent nonparametric priors, heterogeneity, mixture models, nested processes.

1 Introduction

Data that are generated from different (though related) studies, populations or experiments are typically characterized by some degree of heterogeneity. A number of Bayesian nonparametric models have been proposed to accommodate such data structures, but analytic complexity has limited understanding of the implied dependence structure across samples. The spectrum of possible dependence ranges from homogeneity, corresponding to full exchangeability, to complete heterogeneity, corresponding to unconditional independence. It is clearly desirable to construct a prior that can cover

[*]Department of Economics, Management and Statistics, University of Milano - Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy, federico.camerlenghi@unimib.it

[†]Also affiliated to Collegio Carlo Alberto, Torino and BIDSA, Bocconi University, Milano, Italy

[‡]Department of Statistical Science, Duke University, Durham, NC 27708-0251 U.S.A., dunson@duke.edu

[§]Department of Decision Sciences and BIDSA, Bocconi University, via Röntgen 1, 20136 Milano, Italy, antonio.lijoi@unibocconi.it; igor@unibocconi.it

[¶]Department of Applied Mathematics and Statistics, University of California at Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, U.S.A., abel.rod@ucsc.edu

 https://doi.org/10.1214/19-BA1169

this full spectrum, leading to a posterior that can appropriately adapt to the true dependence structure in the available data.

This problem has been partly addressed in several papers. In Lijoi et al. (2014) a class of random probability measures is defined in such a way that proximity to full exchangeability or independence is expressed in terms of a $[0, 1]$-valued random variable. In the same spirit, a model decomposable into idiosyncratic and common components is devised in Müller et al. (2004). Alternatively, approaches based on Pólya tree priors are developed in Ma and Wong (2011); Holmes et al. (2015); Filippi and Holmes (2017), while a multi-resolution scanning method is proposed in Soriano and Ma (2017). In Bhattacharya and Dunson (2012) Dirichlet process mixtures are used to test homogeneity across groups of observations on a manifold. A popular class of dependent nonparametric priors that fits this framework is the *nested Dirichlet process* (nDP) of Rodríguez et al. (2008), which aims at clustering the probability distributions associated to d populations. For $d = 2$ this model is

$$\begin{aligned} (X_{i,1}, X_{j,2}) \mid (\tilde{p}_1, \tilde{p}_2) &\overset{\text{ind}}{\sim} \tilde{p}_1 \times \tilde{p}_2 \qquad (i, j) \in \mathbb{N} \times \mathbb{N} \\ (\tilde{p}_1, \tilde{p}_2) \mid \tilde{q} &\sim \tilde{q}^2, \qquad \tilde{q} = \sum_{i \geq 1} \omega_i \, \delta_{G_i}, \end{aligned} \tag{1}$$

where the random elements $\boldsymbol{X}_\ell := (X_{i,\ell})_{i \geq 1}$, for $\ell = 1, 2$, take values in a space $\mathbb{X}$, the sequences $(\omega_i)_{i \geq 1}$ and $(G_i)_{i \geq 1}$ are independent, with $\sum_{i \geq 1} \omega_i = 1$ almost surely, and the G_i's are i.i.d. random probability measures on $\mathbb{X}$ such that

$$G_i = \sum_{t \geq 1} w_{t,i} \delta_{\theta_{t,i}}, \qquad \theta_{t,i} \overset{\text{iid}}{\sim} Q_0 \tag{2}$$

for some non-atomic probability measure Q_0 on $\mathbb{X}$. In Rodríguez et al. (2008) it is assumed that $\tilde{q}$ and the G_i's are realizations of Dirichlet processes while in Rodríguez and Dunson (2014) it is assumed they are from a generalized Dirichlet process introduced by Hjort (2000). Due to discreteness of $\tilde{q}$, one has $\tilde{p}_1 = \tilde{p}_2$ with positive probability allowing for clustering at the level of the populations' distributions and implying $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have the same probability distribution.

The composition of random combinatorial structures, such as those in (1), lies at the heart of several other proposals of prior processes for modeling non-exchangeable data. A noteworthy example is the hierarchical Dirichlet process in Teh et al. (2006), which arises as a generalization of the latent Dirichlet allocation model Blei et al. (2003) and yields a partition distribution also known as the Chinese restaurant franchise. Generalizations beyond the Dirichlet process case together with an in-depth analysis of their distributional properties is provided in Camerlenghi et al. (2019a). Another approach sets a prior directly on the space of partitions, by possibly resorting to appropriate modifications of product partition models. See, e.g., Dahl et al. (2017); Müller et al. (2011); Page and Quintana (2016); Blei and Frazier (2011). In fact, the literature on priors over spaces of dependent probability distributions has rapidly grown in the last 15 years, spurred by the ideas of MacEachern (1999, 2000). The initial contributions in the area were mainly focused on providing dependent versions of the Dirichlet process (see, e.g., De Iorio et al. (2004); Gelfand et al. (2005); Griffin and Steel (2006);

De Iorio et al. (2009)). More recently, a number of proposals of more general classes of dependent priors have appeared, by either using a stick-breaking procedure or resorting to random measures-based constructions. Among them we mention Chung and Dunson (2009); Jara et al. (2010); Rodríguez et al. (2010); Rodríguez and Dunson (2011); Lijoi et al. (2014); Griffin et al. (2013); Griffin and Leisen (2017); Mena and Ruggiero (2016); Barrientos et al. (2017); Nguyen (2013, 2015). Our contribution, relying on a random measures-based approach, inserts itself into this active research area providing an effective alternative to the nDP.

The nDP has been widely used in a rich variety of applications, but it has an unappealing characteristic that provides motivation for this article. In particular, if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ share at least one value, then the posterior distribution of $(\tilde{p}_1, \tilde{p}_2)$ degenerates on $\{\tilde{p}_1 = \tilde{p}_2\}$, forcing homogeneity across the two samples. This occurs also in nDP mixture models in which the $X_{i,\ell}$ are latent, and is not specific to the Dirichlet process but is a consequence of nesting discrete random probabilities. For a more effective illustration, consider the case where one is examining measurements that are used to assess quality of hospitals in d different regions or territories. It is reasonable to assume that there is homogeneity (namely, exchangeability) among hospitals in the same region and heterogeneity across different regions. This is actually the setting that motivated the original formulation in Rodríguez et al. (2008), who are interested in clustering the d populations of hospitals based on the quality of care. However, one may also aim at identifying possible sub-populations of hospitals that are shared across the d regions, while still preserving some degree of heterogeneity. Unfortunately, the nDP cannot attain this and as soon as the model detects a shared sub-population between two different regions it leads to the conclusion that those two regions share the same probability distribution and are, thus, similar or homogeneous.

To overcome this major limitation, we propose a more flexible class of *latent nested processes*, which preserve heterogeneity *a posteriori*, even when distinct values are shared by different samples. Latent nested processes define $\tilde{p}_1$ and $\tilde{p}_2$ in (1) as resulting from normalization of an additive random measure model with common and idiosyncratic components, the latter with nested structure. Latent nested processes are shown to have appealing distributional properties. In particular, nesting corresponds, in terms of the induced partitions, to a convex combination of full exchangeability and unconditional independence, the two extreme cases. This naturally yields a methodology for testing equality of distributions.

2 Nested processes

2.1 Generalizing nested Dirichlet processes via normalized random measures

We first propose a class of nested processes that generalize nested Dirichlet processes by replacing the Dirichlet process components with a more flexible class of random measures. The idea is to define $\tilde{q}$ in (1) in terms of normalized completely random measures on the space $\mathsf{P}_{\mathbb{X}}$ of probability measures on $\mathbb{X}$. In order to provide a full

account of the construction, introduce a Poisson random measure $\tilde{N} = \sum_{i\geq 1} \delta_{(J_i,G_i)}$ on $\mathbb{R}^+ \times \mathsf{P}_{\mathbb{X}}$ characterized by a mean intensity measure ν such that for any $A \in \mathscr{B}(\mathbb{R}^+) \otimes \mathscr{B}(\mathsf{P}_{\mathbb{X}})$ for which $\nu(A) < \infty$ one has $\tilde{N}(A) \sim \text{Po}(\nu(A))$. It is further supposed that

$$\nu(\mathrm{d}s, \mathrm{d}p) = c\, \rho(s)\, \mathrm{d}s\, Q(\mathrm{d}p), \tag{3}$$

where Q is a probability distribution on $\mathsf{P}_{\mathbb{X}}$, ρ is some non-negative measurable function on $\mathbb{R}^+$ such that $\int_0^\infty \min\{1, s\}\, \rho(s)\, \mathrm{d}s < \infty$ and $c > 0$. Henceforth, we will also refer to ν as Lévy intensity. A *completely random measure* (CRM) $\tilde{\mu}$ without fixed points of discontinuity is, thus, defined as $\tilde{\mu} = \sum_{i\geq 1} J_i\, \delta_{G_i}$. It is well-known that ν characterizes $\tilde{\mu}$ through its Lévy-Khintchine representation

$$\begin{aligned} \mathbb{E}\left[\mathrm{e}^{-\lambda \tilde{\mu}(A)}\right] &= \exp\left\{ -\int_{\mathbb{R}^+ \times \mathsf{P}_{\mathbb{X}}} \left(1 - \mathrm{e}^{-\lambda s}\right) \nu(\mathrm{d}s, \mathrm{d}p) \right\} \\ &= \exp\left\{ -c\, Q(A) \int_0^\infty \left(1 - \mathrm{e}^{-\lambda s}\right) \rho(s)\, \mathrm{d}s \right\} =: \mathrm{e}^{-c\, Q(A)\, \psi(\lambda)} \end{aligned} \tag{4}$$

for any measurable $A \subset \mathsf{P}_{\mathbb{X}}$, we use the notation $\tilde{\mu} \sim \text{CRM}[\nu; \mathsf{P}_{\mathbb{X}}]$. The function ψ in (4) is also referred to as the *Laplace exponent* of $\tilde{\mu}$. For a more extensive treatment of CRMs, see Kingman (1993). If one additionally assumes that $\int_0^\infty \rho(s)\, \mathrm{d}s = \infty$, then $\tilde{\mu}(\mathsf{P}_{\mathbb{X}}) > 0$ almost surely and we can define $\tilde{q}$ in (1) as

$$\tilde{q} \overset{\mathrm{d}}{=} \frac{\tilde{\mu}}{\tilde{\mu}(\mathsf{P}_{\mathbb{X}})}. \tag{5}$$

This is known as a *normalized random measure with independent increments* (NRMI), introduced in Regazzini et al. (2003), and is denoted as $\tilde{q} \sim \text{NRMI}[\nu; \mathsf{P}_{\mathbb{X}}]$. The baseline measure, Q, of $\tilde{\mu}$ in (3) is, in turn, the probability distribution of $\tilde{q}_0 \sim \text{NRMI}[\nu_0; \mathbb{X}]$, with $\tilde{q}_0 \overset{\mathrm{d}}{=} \tilde{\mu}_0/\tilde{\mu}_0(\mathbb{X})$ and $\tilde{\mu}_0$ having Lévy measure

$$\nu_0(\mathrm{d}s, \mathrm{d}x) = c_0\, \rho_0(s)\, \mathrm{d}s\, Q_0(\mathrm{d}x) \tag{6}$$

for some non-negative function ρ_0 such that $\int_0^\infty \min\{1, s\}\, \rho_0(s)\, \mathrm{d}s < \infty$ and $\int_0^\infty \rho_0(s)\, \mathrm{d}s = \infty$. Moreover, Q_0 is a non-atomic probability measure on $\mathbb{X}$ and ψ_0 is the Laplace exponent of $\tilde{\mu}_0$. The resulting general class of nested processes is such that $(\tilde{p}_1, \tilde{p}_2)|\tilde{q} \sim \tilde{q}^2$ and is indicated by

$$(\tilde{p}_1, \tilde{p}_2) \sim \text{NP}(\nu_0, \nu).$$

The *nested Dirichlet process* (nDP) of Rodríguez et al. (2008) is recovered by specifying $\tilde{\mu}$ and $\tilde{\mu}_0$ to be gamma processes, namely $\rho(s) = \rho_0(s) = s^{-1}\, \mathrm{e}^{-s}$, so that both $\tilde{q}$ and $\tilde{q}_0$ are Dirichlet processes.

2.2 Clustering properties of nested processes

A key property of nested processes is their ability to cluster both population distributions and data from each population. In this subsection, we present results on: (i) the

prior probability that $\tilde{p}_1 = \tilde{p}_2$ and the resulting impact on ties at the observations' level; (ii) equations for mixed moments as convex combinations of fully exchangeable and unconditionally independent special cases; and (iii) a similar convexity result for the so called *partially exchangeable partition probability function* (pEPPF), describing the distribution of the random partition generated by the data. Before stating result (i) define

$$\tau_q(u) = \int_0^\infty s^q \, \mathrm{e}^{-us} \, \rho(s) \, \mathrm{d}s, \qquad \tau_q^{(0)}(u) = \int_0^\infty s^q \, \mathrm{e}^{-us} \, \rho_0(s) \, \mathrm{d}s,$$

for any $u > 0$, and agree that $\tau_0(u) \equiv \tau_0^{(0)}(u) \equiv 1$.

Proposition 1. *If* $(\tilde{p}_1, \tilde{p}_2) \sim \mathrm{NP}(\nu_0, \nu)$, *with* $\nu(\mathrm{d}s, \mathrm{d}p) = c\,\rho(s)\,\mathrm{d}s\; Q(\mathrm{d}p)$ *and* $\nu_0(\mathrm{d}s, \mathrm{d}x) = c_0\,\rho_0(s)\,\mathrm{d}s\,Q_0(\mathrm{d}x)$ *as before, then*

$$\pi_1 := \mathbb{P}(\tilde{p}_1 = \tilde{p}_2) = c \int_0^\infty u \, \mathrm{e}^{-c\psi(u)} \, \tau_2(u) \, \mathrm{d}u \tag{7}$$

and the probability that any two observations from the two samples coincide equals

$$\mathbb{P}(X_{j,1} = X_{k,2}) = \pi_1 \, c_0 \int_0^\infty u \, \mathrm{e}^{-c_0 \, \psi_0(u)} \, \tau_2^{(0)}(u) \, \mathrm{d}u > 0. \tag{8}$$

This result shows that the probability of $\tilde{p}_1$ and $\tilde{p}_2$ coinciding is positive, as desired, but also that this implies a positive probability of ties at the observations' level. Moreover, (7) only depends on ν and not ν_0, since the latter acts on the $\mathbb{X}$ space. In contrast, the probability that any two observations $X_{j,1}$ and $X_{k,2}$ from the two samples coincide given in (8) depends also on ν_0. If $(\tilde{p}_1, \tilde{p}_2)$ is an nDP, which corresponds to $\rho(s) = \rho_0(s) = \mathrm{e}^{-s}/s$, one obtains $\pi_1 = 1/(c+1)$ and $\mathbb{P}(X_{j,1} = X_{k,2}) = \pi_1/(c_0+1)$.

The following proposition [our result (ii)] provides a representation of mixed moments as a convex combination of full exchangeability and unconditional independence between samples.

Proposition 2. *If* $(\tilde{p}_1, \tilde{p}_2) \sim \mathrm{NP}(\nu_0, \nu)$ *and* $\pi_1 = \mathbb{P}(\tilde{p}_1 = \tilde{p}_2)$ *is as in* (7), *then*

$$\begin{aligned} \mathbb{E}\Big[\int_{\mathsf{P}_{\mathbb{X}}^2} f_1(p_1) f_2(p_2) \tilde{q}(\mathrm{d}p_1) \tilde{q}(\mathrm{d}p_2) \Big] &= \pi_1 \int_{\mathsf{P}_{\mathbb{X}}} f_1(p) f_2(p) Q(\mathrm{d}p) \\ &\quad + (1-\pi_1) \int_{\mathsf{P}_{\mathbb{X}}} f_1(p) Q(\mathrm{d}p) \int_{\mathsf{P}_{\mathbb{X}}} f_2(p) Q(\mathrm{d}p) \end{aligned} \tag{9}$$

for all measurable functions $f_1, f_2 : \mathsf{P}_{\mathbb{X}} \to \mathbb{R}^+$ *and the expected value is taken w.r.t.* $\tilde{q}$.

This convexity property is a key property of nested processes. The component with weight $1-\pi_1$ in (9) accounts for heterogeneity among data from different populations and it is important to retain this component also *a posteriori* in (1). Proposition 2 is instrumental to obtain our main result in Theorem 1 characterizing the partially exchangeable random partition induced by $\boldsymbol{X}_1^{(n_1)} = (X_{1,1}, \ldots, X_{n_1,1})$ and $\boldsymbol{X}_2^{(n_2)} = (X_{1,2}, \ldots, X_{n_2,2})$

in (1). To fix ideas consider a partition of the n_ℓ data of sample $\boldsymbol{X}_\ell^{(n_\ell)}$ into k_ℓ specific groups and k_0 groups shared with sample $\boldsymbol{X}_s^{(n_s)}$ $(s \neq \ell)$ with corresponding frequencies $\boldsymbol{n}_\ell = (n_{1,\ell}, \ldots, n_{k_\ell,\ell})$ and $\boldsymbol{q}_\ell = (q_{1,\ell}, \ldots, q_{k_0,\ell})$. In other terms, the two-sample data induce a partition of $[n_1 + n_2] = \{1, \ldots, n_1 + n_2\}$. For example, $\boldsymbol{X}_1^{(7)} = (0.5,\ 2,\ -1,\ 5,\ 5,\ 0.5,\ 0.5)$ and $\boldsymbol{X}_2^{(4)} = (5,\ -2,\ 0.5,\ 0.5)$ yield a partition of $n_1 + n_2 = 11$ objects into 5 groups of which $k_1 = 2$ and $k_2 = 1$ are specific to the first and the second sample, respectively, and $k_0 = 2$ are shared. Moreover, the frequencies are $\boldsymbol{n}_1 = (1, 1)$, $\boldsymbol{n}_2 = (1)$, $\boldsymbol{q}_1 = (3, 2)$ and $\boldsymbol{q}_2 = (2, 1)$. As already mentioned at the beginning of the present section, the partition of the data is characterized by a convenient probabilistic tool called *partially exchangeable partition probability function* (pEPPF), whose formal definition is as follows

$$\mathbb{E} \int_{\mathbb{X}^k} \prod_{j=1}^{k_1} \tilde{p}_1^{n_{j,1}}(\mathrm{d}x_{j,1}) \prod_{l=1}^{k_2} \tilde{p}_2^{n_{l,2}}(\mathrm{d}x_{l,2}) \prod_{r=1}^{k_0} \tilde{p}_1^{q_{r,1}}(\mathrm{d}x_r)\, \tilde{p}_2^{q_{r,2}}(\mathrm{d}x_r), \tag{10}$$

where $k = k_1 + k_2 + k_0$ and the expected value is taken w.r.t. the joint distribution of $(\tilde{p}_1, \tilde{p}_2)$. In the exchangeable framework the pEPPF reduces to the usual *exchangeable partition probability function* (EPPF), as introduced by Pitman (1995). See also Kingman (1978) who proved that the law of a random partition, satisfying certain consistency conditions and a symmetry property, can always be recovered as the random partition induced by an exchangeable sequence of observations.

Let us start by analyzing the two extreme cases. For the fully exchangeable case (in the sense of exchangeability holding true across both samples), one obtains the EPPF

$$\begin{aligned} \Phi_k^{(N)}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1 + \boldsymbol{q}_2) = \frac{c_0^k}{\Gamma(N)} \int_0^\infty u^{N-1} \mathrm{e}^{-c_0 \psi_0(u)} \\ \times \prod_{j=1}^{k_1} \tau_{n_{j,1}}^{(0)}(u) \prod_{i=1}^{k_2} \tau_{n_{i,2}}^{(0)}(u) \prod_{r=1}^{k_0} \tau_{q_{r,1}+q_{r,2}}^{(0)}(u)\, \mathrm{d}u \end{aligned} \tag{11}$$

having set $N = n_1 + n_2$, $k = k_0 + k_1 + k_2$. The marginal EPPFs for the individual sample $\ell = 1, 2$ are

$$\begin{aligned} \Phi_{\ell,k_0+k_\ell}^{(n_\ell)}(\boldsymbol{n}_\ell, \boldsymbol{q}_\ell) &= \Phi_{k_0+k_\ell}^{(n_\ell)}(\boldsymbol{n}_\ell, \boldsymbol{q}_\ell) \\ &= \frac{(c_0)^{k_0+k_\ell}}{\Gamma(n_\ell)} \int_0^\infty u^{n_\ell - 1}\, \mathrm{e}^{-c_0\, \psi_0(u)} \prod_{j=1}^{k_\ell} \tau_{n_{j,\ell}}^{(0)}(u) \prod_{r=1}^{k_0} \tau_{q_{r,\ell}}^{(0)}(u)\, \mathrm{d}u. \end{aligned} \tag{12}$$

Both (11) and (12) hold true with the constraints $\sum_{j=1}^{k_\ell} n_{j,\ell} + \sum_{r=1}^{k_0} q_{r,\ell} = n_\ell$ and $1 \leq k_\ell + k_0 \leq n_\ell$, for each $\ell = 1, 2$. Finally, the convention $\tau_0^{(0)} \equiv 1$ implies that whenever an argument of the function $\Phi_k^{(n)}$ is zero, then it reduces to $\Phi_{k-1}^{(n)}$. For example, $\Phi_3^{(6)}(0, 2, 4) = \Phi_2^{(6)}(2, 4)$. Both (11) and (12) solely depend on the Lévy intensity of the CRM and can be made explicit for specific choices. We are now ready to state our main result (iii).

Theorem 1. *The random partition induced by the samples* $\boldsymbol{X}_1^{(n_1)}$ *and* $\boldsymbol{X}_2^{(n_2)}$ *drawn from* $(\tilde{p}_1, \tilde{p}_2) \sim \mathrm{NP}(\nu_0, \nu)$*, according to* (1) *with* Q_0 *non-atomic, is characterized by the pEPPF*

$$\begin{aligned}\Pi_k^{(N)}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1, \boldsymbol{q}_2) &= \pi_1\, \Phi_k^{(N)}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1 + \boldsymbol{q}_2) \\ &\quad + (1-\pi_1)\, \Phi_{k_0+k_1}^{(n_1+|\boldsymbol{q}_1|)}(\boldsymbol{n}_1, \boldsymbol{q}_1) \Phi_{k_0+k_2}^{(n_2+|\boldsymbol{q}_2|)}(\boldsymbol{n}_2, \boldsymbol{q}_2) \mathbb{1}_{\{0\}}(k_0)\end{aligned} \tag{13}$$

having set $|\boldsymbol{a}| = \sum_{i=1}^p a_i$ *for any vector* $\boldsymbol{a} = (a_1, \ldots, a_p) \in \mathbb{R}^p$ *with* $p \geq 2$.

The two independent EPPFs in the second summand on the right-hand side of (13) are crucial for accounting for the heterogeneity across samples. However, the result shows that one shared value, i.e. $k_0 \geq 1$, forces the random partition to degenerate to the fully exchangeable case in (11). Hence, a single tie forces the two samples to be homogeneous, representing a serious limitation of all nested processes including the nDP special case. This result shows that degeneracy is a consequence of combining simple discrete random probabilities with nesting. In the following section, we develop a generalization that is able to preserve heterogeneity in presence of ties between the samples.

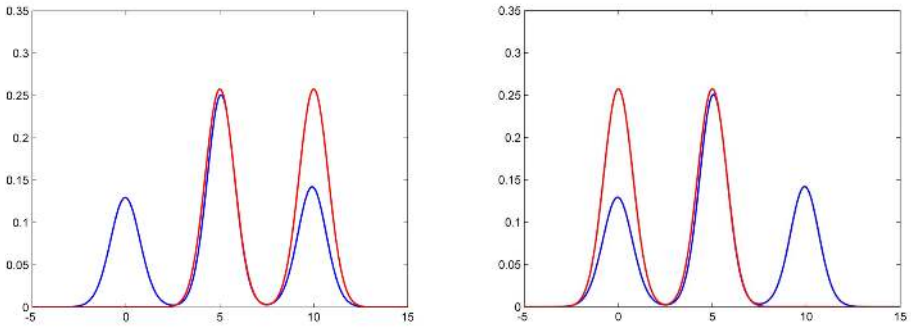
3 Latent nested processes

To address degeneracy of the pEPPF in (13), we look for a model that, while still able to cluster random probabilities, can also take into account heterogeneity of the data in presence of ties between $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$. The issue is relevant also in mixture models where $\tilde{p}_1$ and $\tilde{p}_2$ are used to model partially exchangeable latent variables such as, e.g., vectors of means and variances in normal mixture models. To see this, consider a simple density estimation problem, where two-sample data of sizes $n_1 = n_2 = 100$ are generated from

$$X_{i,1} \sim \frac{1}{2}\,\mathrm{N}(5,\, 0.6) + \frac{1}{2}\,\mathrm{N}(10,\, 0.6) \qquad X_{j,2} \sim \frac{1}{2}\,\mathrm{N}(5,\, 0.6) + \frac{1}{2}\,\mathrm{N}(0,\, 0.6).$$

This can be modeled by dependent normal mixtures with mean and variance specified in terms of a nested structure as in (1). The results, carried out by employing the algorithms detailed in Section 4, show two possible outcomes: either the model is able to estimate well the two bimodal marginal densities, while not identifying the presence of a common component, or it identifies the shared mixture component but does not yield a sensible estimate of the marginal densities, which both display three modes. The latter situation is displayed in Figure 1: once the shared component (5, 0.6) is detected, the two marginal distributions are considered identical as the whole dependence structure boils down to exchangeability across the two samples.

This critical issue can be tackled by a novel class of latent nested processes. Specifically, we introduce a model where the nesting structure is placed at the level of the underlying CRMs, which leads to greater flexibility while preserving tractability. In order to define the new process, let $\mathsf{M}_{\mathbb{X}}$ be the space of boundedly finite measures on $\mathbb{X}$



(a) (b)

Figure 1: Nested σ-stable mixture models: Estimated densities (blue) and true densities (red), for $\boldsymbol{X}_1^{(100)}$ in Panel (a) and for $\boldsymbol{X}_2^{(100)}$ in Panel (b).

and Q the probability measure on $\mathsf{M}_{\mathbb{X}}$ induced by $\tilde{\mu}_0 \sim \mathrm{CRM}[\nu_0; \mathbb{X}]$, where ν_0 is as in (6). Hence, for any measurable subset A of $\mathbb{X}$

$$\mathbb{E}\Big[\mathrm{e}^{-\lambda \tilde{\mu}_0(A)}\Big] = \int_{\mathsf{M}_{\mathbb{X}}} \mathrm{e}^{-\lambda\, m(A)}\, Q(\mathrm{d}m) = \exp\Big\{ - c_0\, Q_0(A) \int_0^{\infty} \Big(1 - \mathrm{e}^{-\lambda s}\Big)\, \rho_0(s)\, \mathrm{d}s \Big\}.$$

Definition 1. Let $\tilde{q} \sim \mathrm{NRMI}[\nu; \mathsf{M}_{\mathbb{X}}]$, with $\nu(\mathrm{d}s, \mathrm{d}m) = c\rho(s)\mathrm{d}s\, Q(\mathrm{d}m)$. Random probability measures $(\tilde{p}_1, \tilde{p}_2)$ are a *latent nested process* if

$$\tilde{p}_\ell = \frac{\mu_\ell + \mu_S}{\mu_\ell(\mathbb{X}) + \mu_S(\mathbb{X})} \qquad \ell = 1, 2, \tag{14}$$

where $(\mu_1, \mu_2, \mu_S) \,|\, \tilde{q} \sim \tilde{q}^2 \times \tilde{q}_S$ and $\tilde{q}_S$ is the law of a $\mathrm{CRM}[\nu_0^*; \mathbb{X}]$, where $\nu_0^* = \gamma\, \nu_0$, for some $\gamma > 0$. Henceforth, we will use the notation $(\tilde{p}_1, \tilde{p}_2) \sim \mathrm{LNP}(\gamma, \nu_0, \nu)$.

Furthermore, since

$$\tilde{p}_\ell = w_\ell \frac{\mu_\ell}{\mu_\ell(\mathbb{X})} + (1 - w_\ell) \frac{\mu_S}{\mu_S(\mathbb{X})}, \quad \text{where } w_\ell = \frac{\mu_\ell(\mathbb{X})}{\mu_S(\mathbb{X}) + \mu_\ell(\mathbb{X})}, \tag{15}$$

each $\tilde{p}_\ell$ is a mixture of two components: an idiosyncratic component $p_\ell := \mu_\ell/\mu_\ell(\mathbb{X})$ and a shared component $p_S := \mu_S/\mu_S(\mathbb{X})$. Here μ_S preserves heterogeneity across samples even when shared values are present. The parameter γ in the intensity ν_0^* tunes the effect of such a shared CRM. One recovers model (1) as $\gamma \to 0$. A generalization to nested CRMs of the results given in Propositions 1 and 2 is provided in the following proposition, whose proof is omitted.

Proposition 3. *If* $(\mu_1, \mu_2) \,|\, \tilde{q} \sim \tilde{q}^2$, *where* $\tilde{q} \sim \mathrm{NRMI}[\nu; \mathsf{M}_{\mathbb{X}}]$ *as in* Definition 1, *then*

$$\pi_1^* = \mathbb{P}(\mu_1 = \mu_2) = c \int_0^{\infty} u\, \mathrm{e}^{-c\psi(u)}\, \tau_2(u)\, \mathrm{d}u \tag{16}$$

and

$$\begin{aligned}&\mathbb{E}\Big[\int_{\mathsf{M}_{\mathbb{X}}^2} f_1(m_1)\, f_2(m_2)\, \tilde{q}^2(\mathrm{d}m_1, \mathrm{d}m_2)\Big] \\ &\qquad = \pi_1^* \int_{\mathsf{M}_{\mathbb{X}}} f_1(m)\, f_2(m)\, Q(\mathrm{d}m) + (1-\pi_1^*) \prod_{\ell=1}^{2} \int_{\mathsf{M}_{\mathbb{X}}} f_\ell(m)\, Q(\mathrm{d}m)\end{aligned} \tag{17}$$

for all measurable functions $f_1, f_2 : \mathsf{M}_{\mathbb{X}} \to \mathbb{R}^+$.

Proposition 4. *If* $(\tilde{p}_1, \tilde{p}_2) \sim \mathrm{LNP}(\gamma, \nu_0, \nu)$, *then* $\mathbb{P}(\tilde{p}_1 = \tilde{p}_2) = \mathbb{P}(\mu_1 = \mu_2)$.

Proposition 4, combined with $\{\tilde{p}_1 = \tilde{p}_1\} = \{\mu_1 = \mu_2\} \cup (\{\tilde{p}_1 = \tilde{p}_2\} \cap \{\mu_1 \neq \mu_2\})$, entails $\mathbb{P}[\{\tilde{p}_1 = \tilde{p}_2\} \cap \{\mu_1 \neq \mu_2\}] = 0$ namely

$$\mathbb{P}(\{\tilde{p}_1 = \tilde{p}_2\} \cap \{\mu_1 = \mu_2\}) + \mathbb{P}(\{\tilde{p}_1 \neq \tilde{p}_2\} \cap \{\mu_1 \neq \mu_2\}) = 1$$

and, then, the random variables $\mathbb{1}\{\tilde{p}_1 = \tilde{p}_2\}$ and $\mathbb{1}\{\mu_1 = \mu_2\}$ coincide almost surely. As a consequence the posterior distribution of $\mathbb{1}\{\mu_1 = \mu_2\}$ can be readily employed to test equality between the distributions of the two samples. Further details are given in Section 5.

For analytic purposes, it is convenient to introduce an augmented version of the latent nested process, which includes latent indicator variables. In particular, $(X_{i,1}, X_{j,2}) \mid (\tilde{p}_1, \tilde{p}_2) \sim \tilde{p}_1 \times \tilde{p}_2$, with $(\tilde{p}_1, \tilde{p}_2) \sim \mathrm{LNP}(\gamma, \nu_0, \nu)$ if and only if

$$\begin{aligned}(X_{i,1}, X_{j,2}) \mid (\zeta_{i,1}, \zeta_{j,2}, \mu_1, \mu_2, \mu_S) &\overset{\text{ind}}{\sim} p_{\zeta_{i,1}} \times p_{2\zeta_{j,2}} \\ (\zeta_{i,1}, \zeta_{j,2}) \mid (\mu_1, \mu_2, \mu_S) &\sim \mathrm{Bern}(w_1) \times \mathrm{Bern}(w_2) \\ (\mu_1, \mu_2, \mu_S) \mid (\tilde{q}, \tilde{q}_S) &\sim \tilde{q}^2 \times \tilde{q}_S.\end{aligned} \tag{18}$$

The latent variables $\zeta_{i,\ell}$ indicate which random probability measure is actually generating each observation $X_{i,\ell}$, for $i = 1, \ldots, n_\ell$. More specifically this random probability measure coincides with p_ℓ if the corresponding label $\zeta_{i,\ell} = 1$, otherwise, if $\zeta_{i,\ell} = 0$, this is $p_0 = p_S$. We will further write $\boldsymbol{\zeta}_\ell^* = (\zeta_{1,\ell}^*, \ldots, \zeta_{k_\ell,\ell}^*)$ to denote the latent variables that are associated to the k_ℓ distinct clusters, either shared or sample-specific, for $\ell = 0, 1, 2$. Moreover, $\bar{k}_\ell := |\boldsymbol{\zeta}_\ell^*|$ and define $\bar{k} := \bar{k}_0 + \bar{k}_1 + \bar{k}_2$. With $\odot$ denoting the component-wise multiplication of vectors, the frequencies corresponding to groups labeled $\zeta_{i,\ell} = 1$ will be denoted by $\bar{\boldsymbol{n}}_\ell := \boldsymbol{n}_\ell \odot \boldsymbol{\zeta}_\ell^*$ and $\bar{\boldsymbol{q}}_\ell := \boldsymbol{q}_\ell \odot \boldsymbol{\zeta}_0^*$, with $\bar{n}_\ell = |\bar{\boldsymbol{n}}_\ell|$ and $\bar{q}_\ell = |\bar{\boldsymbol{q}}_\ell|$, for $\ell = 1, 2$. Finally, if $\bar{\boldsymbol{q}} := \bar{\boldsymbol{q}}_1 + \bar{\boldsymbol{q}}_2$ and $\bar{n}_0 = |\bar{\boldsymbol{q}}|$, the overall number of observations having label 1 will be indicated by $\bar{n} = \bar{n}_0 + \bar{n}_1 + \bar{n}_2$. For instance, if $\boldsymbol{X}_1^{(7)} = (0.5, 2, -1, 5, 5, 0.5, 0.5)$, $\boldsymbol{X}_2^{(4)} = (5, -2, 0.5, 0.5)$, $\boldsymbol{\zeta}_1 = (0, 1, 0, 1, 1, 0, 0)$ and $\boldsymbol{\zeta}_2 = (1, 1, 0, 0)$, the labels attached to the 5 distinct observations are $\boldsymbol{\zeta}_1^* = (1, 0)$, $\boldsymbol{\zeta}_2^* = (1)$ and $\boldsymbol{\zeta}_0^* = (0, 1)$. From this, one has $\bar{k}_1 = \bar{k}_2 = \bar{k}_0 = 1$, $\bar{\boldsymbol{n}}_1 = (1, 0)$, $\bar{\boldsymbol{n}}_2 = 1$, $\bar{\boldsymbol{q}}_1 = (0, 2)$ and $\bar{\boldsymbol{q}}_2 = (0, 1)$.

Theorem 2. *The random partition induced by the samples* $\boldsymbol{X}_1^{(n_1)}$ *and* $\boldsymbol{X}_2^{(n_2)}$ *drawn from* $(\tilde{p}_1, \tilde{p}_2) \sim \mathrm{LNP}(\gamma, \nu_0, \nu)$*, as in* (18)*, is characterized by the pEPPF*

$$\begin{aligned}\Pi_k^{(N)}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1, \boldsymbol{q}_2) &= \pi_1^* \frac{c_0^k (1+\gamma)^k}{\Gamma(N)} \\ &\times \int_0^\infty s^{N-1} e^{-(1+\gamma)c_0\psi_0(s)} \prod_{\ell=1}^{2} \prod_{j=1}^{k_\ell} \tau_{n_{j,\ell}}^{(0)}(s) \prod_{j=1}^{k_0} \tau_{q_{j,1}+q_{j,2}}^{(0)}(s) \mathrm{d}s \\ &+ (1-\pi_1^*) \sum_{(*)} I_2(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1 + \boldsymbol{q}_2, \boldsymbol{\zeta}^*),\end{aligned} \tag{19}$$

where

$$\begin{aligned}I_2(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1 + \boldsymbol{q}_2, \boldsymbol{\zeta}^*) = \frac{c_0^k \gamma^{k-\bar{k}}}{\Gamma(n_1)\Gamma(n_2)} \int_0^\infty \int_0^\infty u^{n_1-1} v^{n_2-1} e^{-\gamma c_0 \psi_0(u+v) - c_0(\psi_0(u)+\psi_0(v))} \\ \times \prod_{j=1}^{k_1} \tau_{n_{j,1}}^{(0)}(u + (1-\zeta_{j,1}^*)v) \prod_{j=1}^{k_2} \tau_{n_{j,2}}^{(0)}((1-\zeta_{j,2}^*)u + v) \\ \times \prod_{j=1}^{k_0} \tau_{q_{j,1}+q_{j,2}}^{(0)}(u+v) \mathrm{d}u \mathrm{d}v\end{aligned}$$

and the sum in the second summand on the right hand side of (19) *runs over all the possible labels* $\boldsymbol{\zeta}^* \in \{0,1\}^{k_1+k_2}$*.*

The pEPPF (19) is a convex linear combination of an EPPF corresponding to full exchangeability across samples and one corresponding to unconditional independence. Heterogeneity across samples is preserved even in the presence of shared values. The above result is stated in full generality, and hence may seem somewhat complex. However, as the following examples show, when considering stable or gamma random measures, explicit expressions are obtained. When $\gamma \to 0$ the expression (19) reduces to (13), which means that the nested process is achieved as a special case.

Example 1. Based on Theorem 2 we can derive an explicit expression of the partition structure of *latent nested* σ*-stable processes*. Suppose $\rho(s) = \sigma\, s^{-1-\sigma}/\Gamma(1-\sigma)$ and $\rho_0(s) = \sigma_0\, s^{-1-\sigma_0}/\Gamma(1-\sigma_0)$, for some σ and σ_0 in $(0,1)$. In such a situation it is easy to see that $\pi_1^* = 1-\sigma$, $\tau_q^{(0)}(u) = \sigma_0(1-\sigma_0)_{q-1} u^{\sigma_0 - q}$ and $\psi_0(u) = u^{\sigma_0}$. Moreover let $c_0 = c = 1$, since the total mass of a stable process is redundant under normalization. If we further set

$$J_{\sigma_0,\gamma}(H_1, H_2; H) := \int_0^1 \frac{w^{H_1-1}(1-w)^{H_2-1}}{[\gamma + w^{\sigma_0} + (1-w)^{\sigma_0}]^H} \mathrm{d}w,$$

for any positive H_1, H_2 and H, and

$$\xi_a(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1 + \boldsymbol{q}_2) := \prod_{\ell=1}^{2} \prod_{j=1}^{k_\ell} (1-a)_{n_{j,\ell}-1} \prod_{j=1}^{k_0} (1-a)_{q_{j,1}+q_{j,2}-1},$$

for any $a \in [0,1)$, then the partially exchangeable partition probability function in (19) may be rewritten as

$$\Pi_k^{(N)}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{q}_1,\boldsymbol{q}_2) = \sigma_0^{k-1}\Gamma(k)\xi_{\sigma_0}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{q}_1+\boldsymbol{q}_2)\Big\{\frac{(1-\sigma)}{\Gamma(N)} + \frac{\sigma}{\Gamma(n_1)\Gamma(n_2)}\sum_{(*)}\gamma^{k-\bar{k}}\,J_{\sigma_0,\gamma}(n_1-\bar{n}_1+\bar{k}_1\sigma_0, n_2-\bar{n}_2+\bar{k}_2\sigma_0;k)\Big\}.$$

The sum with respect to $\boldsymbol{\zeta}^*$ can be evaluated and it turns out that

$$\Pi_k^{(n)}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{q}_1+\boldsymbol{q}_2) = \frac{\sigma_0^{k-1}\Gamma(k)}{\Gamma(n)}\xi_{\sigma_0}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{q}_1+\boldsymbol{q}_2)\Big[1-\sigma+\sigma\gamma^{k_0}\,\frac{B(k_1\sigma_0,\,k_2\sigma_0)}{B(n_1,n_2)} \times \int_0^1 \frac{\prod_{j=1}^{k_1}(1+\gamma w^{n_{j,1}-\sigma_0})\,\prod_{i=1}^{k_2}[1+\gamma(1-w)]^{n_{i,2}-\sigma_0}}{\Big[\gamma+w^{\sigma_0}+(1-w)^{\sigma_0}\Big]^k}\,\mathrm{Beta}(\mathrm{d}w;k_1\sigma_0,k_2\sigma_0)\Big],$$

where $\mathrm{Beta}(\,\cdot\,;a,b)$ stands for the beta distribution with parameters a and b, while $B(p,q)$ is the beta function with parameters p and q. As it is well-known, $\sigma_0^{k-1}\,\Gamma(k)\,\xi_{\sigma_0}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{q}_1+\boldsymbol{q}_2)/\Gamma(N)$ is the exchangeable partition probability function of a normalized σ_0-stable process. Details on the above derivation, as well as for the following example, can be found in the Supplementary Material (Camerlenghi et al., 2019b).

Example 2. Let $\rho(s) = \rho_0(s) = \mathrm{e}^{-s}/s$. Recall that $\tau_q^{(0)}(u) = \Gamma(q)/(u+1)^q$ and $\psi_0(u) = \log(1+u)$, furthermore $\pi_1^* = 1/(1+c)$ by standard calculations. From Theorem 2 we obtain the partition structure of the *latent nested Dirichlet process*

$$\Pi_k^{(N)}(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{q}_1,\boldsymbol{q}_2) = \xi_0(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{q}_1+\boldsymbol{q}_2)c_0^k\Big\{\frac{1}{1+c}\,\frac{(1+\gamma)^k}{(c_0(1+\gamma))_N} + \frac{c}{1+c}\sum_{(*)}\frac{\gamma^{k-\bar{k}}}{(\alpha)_{n_2}(\beta)_{n_1}}\,{}_3F_2(c_0+\bar{n}_2,\alpha,n_1;\alpha+n_2,\beta+n_1;1)\Big\},$$

where $\alpha = (\gamma+1)c_0+n_1-\bar{n}_1$, $\beta = c_0(2+\gamma)$ and ${}_3F_2$ is the generalized hypergeometric function. In the same spirit as in the previous example, the first element in the linear convex combination above $c_0^k(1+\gamma)^k\,\xi_0(\boldsymbol{n}_1,\boldsymbol{n}_2,\boldsymbol{q}_1+\boldsymbol{q}_2)/(c_0(1+\gamma))_N$ is nothing but the Ewens' sampling formula, i.e. the exchangeable partition probability function associated to the Dirichlet process whose base measure has total mass $c_0(1+\gamma)$.

4 Markov Chain Monte Carlo algorithm

We develop a class of MCMC algorithms for posterior computation in latent nested process models relying on the pEPPFs in Theorem 2, as they tended to be more effective. Moreover, the sampler is presented in the context of density estimation, where

$$(X_{i,1},X_{j,2}) \mid (\boldsymbol{\theta}_1^{(n_1)},\boldsymbol{\theta}_2^{(n_2)}) \overset{\text{ind}}{\sim} h(\,\cdot\,;\theta_{i,1}) \times h(\,\cdot\,;\theta_{j,2}) \qquad (i,j) \in \mathbb{N}\times\mathbb{N}$$

and the vectors $\boldsymbol{\theta}_\ell^{(n_\ell)} = (\theta_{1,\ell},\ldots,\theta_{n_\ell,\ell})$, for $\ell = 1,2$ and with each $\theta_{i,\ell}$ taking values in $\Theta \subset \mathbb{R}^b$, are partially exchangeable and governed by a pair of $(\tilde{p}_1,\tilde{p}_2)$ as in (18). The

discreteness of $\tilde{p}_1$ and $\tilde{p}_2$ entails ties among the latent variables $\boldsymbol{\theta}_1^{(n_1)}$ and $\boldsymbol{\theta}_2^{(n_2)}$ that give rise to $k = k_1 + k_2 + k_0$ distinct clusters identified by

- the k_1 distinct values specific to $\boldsymbol{\theta}_1^{(n_1)}$, i.e. not shared with $\boldsymbol{\theta}_2^{(n_2)}$. These are denoted as $\boldsymbol{\theta}_1^* := (\theta_{1,1}^*, \ldots, \theta_{k_1,1}^*)$, with corresponding frequencies $\boldsymbol{n}_1$ and labels $\boldsymbol{\zeta}_1^*$;
- the k_2 distinct values specific to $\boldsymbol{\theta}_2^{(n_2)}$, i.e. not shared with $\boldsymbol{\theta}_1^{(n_1)}$. These are denoted as $\boldsymbol{\theta}_2^* := (\theta_{1,2}^*, \ldots, \theta_{k_2,2}^*)$, with corresponding frequencies $\boldsymbol{n}_2$ and labels $\boldsymbol{\zeta}_2^*$;
- the k_0 distinct values shared by $\boldsymbol{\theta}_1^{(n_1)}$ and $\boldsymbol{\theta}_2^{(n_2)}$. These are denoted as $\boldsymbol{\theta}_0^* := (\theta_{1,0}^*, \ldots, \theta_{k_0,0}^*)$, with $\boldsymbol{q}_\ell$ being their frequencies in $\boldsymbol{\theta}_\ell^{(n_\ell)}$and shared labels $\boldsymbol{\zeta}_0^*$.

As a straightforward consequence of Theorem 2, one can determine the joint distribution of the data $\boldsymbol{X}$, the corresponding latent variables $\boldsymbol{\theta}$ and labels $\boldsymbol{\zeta}$ as follows

$$f(\boldsymbol{x} \mid \boldsymbol{\theta})\, \Pi_k^{(N)}(\boldsymbol{n}_1, \boldsymbol{n}_2, \boldsymbol{q}_1, \boldsymbol{q}_2) \prod_{\ell=0}^{2} \prod_{j=1}^{k_\ell} Q_0(\mathrm{d}\theta_{j,\ell}^*), \tag{20}$$

where $\Pi_k^{(N)}$ is as in (19) and, for $C_{j,\ell} := \{i :\ \theta_{i,\ell} = \theta_{j,\ell}^*\}$ and $C_{r,\ell,0} := \{i :\ \theta_{i,\ell} = \theta_{r,0}^*\}$,

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_{\ell=1}^{2} \prod_{j=1}^{k_\ell} \prod_{i \in C_{j,\ell}} h(x_{i,\ell}; \theta_{j,\ell}^*) \prod_{r=1}^{k_0} \prod_{i \in C_{r,\ell,0}} h(x_{i,\ell}; \theta_{r,0}^*).$$

We do now specialize (20) to the case of latent nested σ-stable processes described in Example 1. The Gibbs sampler is described just for sampling $\boldsymbol{\theta}_1^{(n_1)}$, since the structure is replicated for $\boldsymbol{\theta}_2^{(n_2)}$. To simplify the notation, v^{-j} denotes the random variable v after the removal of $\theta_{j,1}$. Moreover, with $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \sigma, \sigma_0, \boldsymbol{\phi})$, we let $\boldsymbol{T}_{-\theta_{j,1}}$ stand for $\boldsymbol{T}$ after deleting $\theta_{j,1}$, $I = \mathbb{1}\{\tilde{p}_1 = \tilde{p}_2\}$ and $Q_j^*(\mathrm{d}\theta) = h(x_{j,1}; \theta)\, Q_0(\mathrm{d}\theta) / \int_\Theta h(x_{j,1}; \theta)\, Q_0(\mathrm{d}\theta)$. Here $\boldsymbol{\phi}$ denotes a vector of hyperparameters entering the definition of the base measure Q_0. The updating structure of the Gibbs sampler is as follows

(1) Sample $\theta_{j,1}$ from

$$\begin{aligned}
\mathbb{P}(\theta_{j,1} \in \mathrm{d}\theta \,|\, \boldsymbol{T}_{-\theta_{j,1}}, I = 1) &= w_0 Q_{j,1}^*(\mathrm{d}\theta) + \sum_{\{i:\, \zeta_{i,0}^{*,-j} = \zeta_{j,1}\}} w_{i,0} \delta_{\{\theta_{i,0}^{*,-j}\}}(\mathrm{d}\theta) \\
&\quad + \sum_{\{i:\, \zeta_{i,1}^{*,-j} = \zeta_{j,1}\}} w_{i,1} \delta_{\{\theta_{i,1}^{*,-j}\}}(\mathrm{d}\theta) + \sum_{\{i:\, \zeta_{i,2}^{*,-j} = \zeta_{j,1}\}} w_{i,2} \delta_{\{\theta_{i,2}^{*,-j}\}}(\mathrm{d}\theta), \\
\mathbb{P}(\theta_{j,1} \in \mathrm{d}\theta \,|\, \boldsymbol{T}_{-\theta_{j,1}}, I = 0) &= w_0' Q_{j,1}^*(\mathrm{d}\theta) + \sum_{\{i:\, \zeta_{i,1}^{*,-j} = \zeta_{j,1}\}} w_{i,1}' \delta_{\{\theta_{i,1}^{*,-j}\}}(\mathrm{d}\theta) \\
&\quad + \mathbb{1}_{\{0\}}(\zeta_{j,1}) \Big[\sum_{\{i:\, \zeta_{i,2}^{*,-j} = 0\}} w_{i,2}' \delta_{\{\theta_{i,2}^{*,-j}\}}(\mathrm{d}\theta) + \sum_{r=1}^{k_0} w_{r,0}' \delta_{\{\theta_{r,0}^{*,-j}\}}(\mathrm{d}\theta) \Big],
\end{aligned}$$

where

$$w_0 \propto \frac{\gamma^{1-\zeta_{j,1}}\sigma_0\, k^{-r}}{1+\gamma}\int_\Theta h(x_{j,1};\theta)Q_0(\mathrm{d}\theta), \quad w_{i,\ell} \propto (n_{i,\ell}^{-j}-\sigma_0)h(x_{j,1};\theta_{i,\ell}^{*,-j}) \quad \ell=1,2,$$

$$w_{i,0} \propto (q_{i,1}^{-j}+q_{i,2}^{-j}-\sigma_0)h(x_{j,1};\theta_{i,0}^{*,-j})$$

and, with $a_1 = n_1 - (\bar{n}_1^{-j}+\zeta_{j,1}) + \bar{k}_1^{-j}\sigma_0$ and $a_2 = n_2 - \bar{n}_2 + \bar{k}_2\sigma_0$, one further has

$$\begin{aligned}
w_0' &\propto \gamma^{1-\zeta_{j,1}}\sigma_0 k^{-j} J_{\sigma_0}(a_1+\zeta_{j,1}\sigma_0, a_2; k^{-j}+1)\int_\Theta h(x_{j,1};\theta)Q_0(\mathrm{d}\theta),\\
w_{i,\ell}' &\propto J_{\sigma_0}(a_1,a_2;k^{-j})\,(n_{i,\ell}^{-j}-\sigma_0)\,h(x_{j,\ell};\theta_{j,\ell}^{*,-j}) \qquad \ell=1,2,\\
w_{i,0}' &\propto J_{\sigma_0}(a_1,a_2;k^{-j})\,(q_{i,1}^{-j}+q_{i,2}^{-j}-\sigma_0)\,h(x_{j,1};\theta_{i,0}^{*,-j}).
\end{aligned}$$

(2) Sample $\zeta_{j,1}^*$ from

$$\begin{aligned}
&\mathbb{P}(\zeta_{j,1}^*=x \mid \boldsymbol{T}_{-\zeta_{j,1}^*}, I=1) = \frac{\gamma^{1-x}}{1+\gamma},\\
&\mathbb{P}(\zeta_{j,1}^*=x \mid \boldsymbol{T}_{-\zeta_{j,1}^*}, I=0) \propto \gamma^{k-k_x-\bar{k}_0-\bar{k}_2} J_{\sigma_0}(n_1-n_x+k_x\sigma_0, n_2-\bar{n}_2+\bar{k}_2\sigma_0; k),
\end{aligned}$$

where $x\in\{0,1\}$, $k_x := x + |\boldsymbol{\zeta}_1^{*,-j}|$ and $n_x = n_{j,1}x + |\boldsymbol{\zeta}_1^{*,-j}\odot \boldsymbol{n}_1^{-j}|$, where $\boldsymbol{a}\odot\boldsymbol{b}$ denotes the component-wise product between two vectors $\boldsymbol{a}$, $\boldsymbol{b}$. Moreover, it should be stressed that, conditional on $I=0$, the labels $\zeta_{r,0}^*$ are degenerate at $x=0$ for each $r=1,\ldots,k_0$.

(3) Update I from

$$\mathbb{P}(I=1\mid\boldsymbol{T}) = 1-\mathbb{P}(I=0\,|\,\boldsymbol{T}) = \frac{(1-\sigma)B(n_1,n_2)}{(1-\sigma)B(n_1,n_2)+\sigma J_{\sigma_0}(\bar{a}_1,\bar{a}_2;k)(1+\gamma)^k},$$

where $\bar{a}_1 = n_1-\bar{n}_1+\bar{k}_1\sigma_0$ and $\bar{a}_2 = n_2-\bar{n}_2+\bar{k}_2\sigma_0$. This sampling distribution holds true whenever $\boldsymbol{\theta}_1^{(n_1)}$ and $\boldsymbol{\theta}_2^{(n_2)}$ do not share any value $\theta_{j,0}^*$ with label $\zeta_{j,0}^*=1$. If this situation occurs, then $\mathbb{P}(I=1\mid\boldsymbol{T})=1$.

(4) Update σ and σ_0 from

$$\begin{aligned}
f(\sigma_0\,|\,\boldsymbol{T}_{-\sigma_0}, I) &\propto J_{\sigma_0}^{1-I}(\bar{a}_1,\bar{a}_2;k)\,\sigma_0^{k-1}\kappa_0(\sigma_0)\prod_{\ell=1}^{2}\prod_{j=1}^{k_\ell}(1-\sigma_0)_{n_{j,\ell}-1}\prod_{r=1}^{k_0}(1-\sigma_0)_{q_{r,1}+q_{r,2}-1},\\
f(\sigma\,|\,\boldsymbol{T}_{-\sigma}, I) &\propto \kappa(\sigma)\left[(1-\sigma)\mathbb{1}_{\{1\}}(I)+\sigma\mathbb{1}_{\{0\}}(I)\right],
\end{aligned}$$

where κ and κ_0 are the priors for σ and σ_0, respectively.

(5) Update γ from

$$f(\gamma \,|\, \boldsymbol{T}_{-\gamma}, I) \;\propto\; \gamma^{k-\bar{k}}\, g(\gamma) \left[\frac{1-\sigma}{(1+\gamma)^k}\, \mathbb{1}_{\{1\}}(I) + \sigma\, J_{\sigma_0}(\bar{a}_1, \bar{a}_2; k)\, \mathbb{1}_{\{0\}}(I) \right],$$

where g is the prior distribution for γ.

Finally, the updating of the hyperparameters depends on the specification of Q_0 that is adopted. They will be displayed in the next section, under the assumption that Q_0 is a normal/inverse-Gamma.

The evaluation of the integral $J_{\sigma_0}(h_1, h_2; h)$ is essential for the implementation of the MCMC procedure. This can be accomplished through numerical methods based on quadrature. However, computational issues arise when h_1 and h_2 are both less than 1 and the integrand defining J_{σ_0} is no longer bounded, although still integrable. For this reason we propose a plain Monte Carlo approximation of J_{σ_0} based on observing that

$$J_{\sigma_0}(h_1, h_2; h) = B(h_1, h_2)\, \mathbb{E}\Big\{ \frac{1}{[\gamma + W^{\sigma_0} + (1-W)^{\sigma_0}]^h} \Big\},$$

with $W \sim \text{Beta}(h_1, h_2)$. Then generating an i.i.d. sample $\{W_i\}_{i=1}^L$ of length L, with $W_i \sim W$, we get the following approximation

$$J_{\sigma_0}(h_1, h_2; h) \approx B(h_1, h_2) \frac{1}{L} \sum_i^L \frac{1}{[\gamma + W_i^{\sigma_0} + (1-W_i)^{\sigma_0}]^h}.$$

5 Illustrations

The algorithm introduced in Section 4 is employed here to estimate dependent random densities. Before implementation, we need first to complete the model specification of our latent nested model (14). Let $\Theta = \mathbb{R} \times \mathbb{R}^+$ and $h(\cdot; (M, V))$ be Gaussian with mean M and variance V. Moreover, as customary, Q_0 is assumed to be a normal/inverse-Gamma distribution

$$Q_0(\mathrm{d}M, \mathrm{d}V) = Q_{0,1}(\mathrm{d}V) Q_{0,2}(\mathrm{d}M|V)$$

with $Q_{0,1}$ an inverse-Gamma probability distribution with parameters (s_0, S_0) and $Q_{0,2}$ a Gaussian with mean m and variance τV. Furthermore, the hyperpriors are

$$\tau^{-1} \sim \text{Gam}(w/2, W/2), \qquad m \sim \text{N}(a, A),$$

for some real parameters $w > 0$, $W > 0$, $A > 0$ and $a \in \mathbb{R}$. In the simulation studies we have set $(w, W) = (1, 100)$, $(a, A) = ((n_1 \bar{X} + n_2 \bar{Y})/(n_1 + n_2), 2)$. The parameters τ and m are updated on the basis of their full conditional distributions, which can be easily derived, and correspond to

$$\mathscr{L}(\tau | \boldsymbol{T}_{-\tau}, I) \sim \text{IG}\Big(\frac{w}{2} + \frac{k}{2}, \frac{W}{2} + \sum_{\ell=0}^{2} \sum_{i=1}^{k_\ell} \frac{(M^*_{i,\ell} - m)^2}{2V^*_{i,\ell}} \Big),$$

$$\mathscr{L}(m|\boldsymbol{T}_{-m}, I) \sim \mathrm{N}\Big(\frac{R}{D}, \frac{1}{D}\Big),$$

where

$$R = \frac{a}{A} + \sum_{\ell=0}^{2}\sum_{i=1}^{k_\ell} \frac{M^*_{i,\ell}}{\tau V^*_{i,\ell}}, \quad D = \frac{1}{A} + \sum_{\ell=0}^{2}\sum_{i=1}^{k_\ell} \frac{1}{\tau V^*_{i,\ell}}.$$

The model specification is completed by choosing uniform prior distributions for σ_0 and σ. In order to overcome the possible slow mixing of the Pólya urn sampler, we include the acceleration step of MacEachern (1994) and West et al. (1994), which consists in resampling the distinct values $(\theta^*_{i,\ell})_{i=1}^{k_\ell}$, for $\ell = 0, 1, 2$, at the end of every iteration. The numerical outcomes displayed in the sequel are based on $50,000$ iterations after $50,000$ burn-in sweeps.

Throughout we assume the data $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ to be independently generated by two densities f_1 and f_2. These will be estimated jointly through the MCMC procedure and the borrowing of strength phenomenon should then allow improved performance. An interesting byproduct of our analysis is the possibility to examine the clustering structure of each distribution, namely the number of components of each mixture. Since the expression of the pEPPF (19) consists of two terms, in order to carry out posterior inference we have defined the random variable $I = \mathbb{1}_{\{\mu_1=\mu_2\}}$. This random variable allows to test whether the two samples come from the same distribution or not, since $I = \mathbb{1}_{\{\tilde{p}_1=\tilde{p}_2\}}$ almost surely (see also Proposition 4). Indeed, if interest lies in testing

$$H_0 : \tilde{p}_1 = \tilde{p}_2 \quad \text{versus} \quad H_1 : \tilde{p}_1 \neq \tilde{p}_2,$$

based on the MCMC output, it is straightforward to compute an approximation of the Bayes factor

$$\mathrm{BF} = \frac{\mathbb{P}(\tilde{p}_1 = \tilde{p}_2|\boldsymbol{X})}{\mathbb{P}(\tilde{p}_1 \neq \tilde{p}_2|\boldsymbol{X})} \frac{\mathbb{P}(\tilde{p}_1 \neq \tilde{p}_2)}{\mathbb{P}(\tilde{p}_1 = \tilde{p}_2)} = \frac{\mathbb{P}(I = 1|\boldsymbol{X})}{\mathbb{P}(I = 0|\boldsymbol{X})} \frac{\mathbb{P}(I = 0)}{\mathbb{P}(I = 1)}$$

leading to acceptance of the null hypothesis if BF is sufficiently large. In the following we first consider simulated datasets generated from normal mixtures and then we analyze the popular Iris dataset.

5.1 Synthetic examples

We consider three different simulated scenarios, where $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ are independent and identically distributed draws from densities that are both two component mixtures of normals. In both cases $(s_0, S_0) = (1, 1)$ and the sample size is $n = n_1 = n_2 = 100$.

First consider a scenario where $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ are drawn from the same density

$$X_{i,1} \sim X_{j,2} \sim \frac{1}{2}\,\mathrm{N}(0, 1) + \frac{1}{2}\,\mathrm{N}(5, 1).$$

The posterior distributions for the number of mixture components, respectively denoted by K_1 and K_2 for the two samples, and for the number of shared components, denoted

by K_{12}, are reported in Table 1. The maximum a posteriori estimate is highlighted in bold. The model is able to detect the correct number of components for each distribution as well as the correct number of components shared across the two mixtures. The density estimates, not reported here, are close to the true data generating densities. The Bayes factor to test equality between the distributions of $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ has been approximated through the MCMC output and coincides with BF = 5.85, providing evidence in favor of the null hypothesis.

scen.	# *comp.*	0	1	2	3	4	5	6	≥ 7
I	K_1	0	0	**0.638**	0.232	0.079	0.029	0.012	0.008
	K_2	0	0	**0.635**	0.235	0.083	0.029	0.011	0.007
	K_{12}	0	0	**0.754**	0.187	0.045	0.012	0.002	0.001
II	K_1	0	0	**0.679**	0.232	0.065	0.018	0.004	0.002
	K_2	0	0	**0.778**	0.185	0.032	0.004	0.001	0
	K_{12}	0	**0.965**	0.034	0.001	0	0	0	0
III	K_1	0	0	**0.328**	0.322	0.188	0.089	0.041	0.032
	K_2	0	0	**0.409**	0.305	0.152	0.073	0.034	0.027
	K_{12}	0	0.183	**0.645**	0.138	0.027	0.006	0.001	0

Table 1: Simulation study: Posterior distributions of the number of components in the first sample (K_1), in the second sample (K_2) and shared by the two samples (K_{12}) corresponding to the three scenarios. The posterior probabilities corresponding to the MAP estimates are displayed in bold.

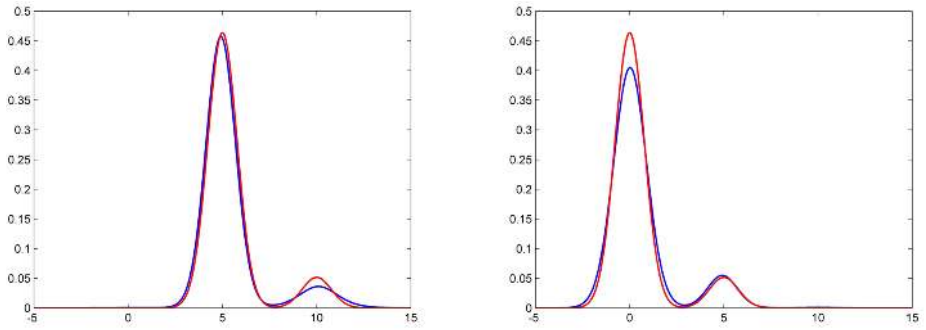
Scenario II corresponds to samples $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ generated, respectively, from

$$X_{i,1} \sim 0.9\,\mathrm{N}(5, 0.6) + 0.1\,\mathrm{N}(10, 0.6) \quad X_{j,2} \sim 0.1\,\mathrm{N}(5, 0.6) + 0.9\,\mathrm{N}(0, 0.6).$$

Both densities have two components but only one in common, i.e. the normal distribution with mean 5. Moreover, the weight assigned to $\mathrm{N}(5, 0.6)$ differs in the two cases. The density estimates are displayed in Figure 2. The spike corresponding to the common component (concentrated around 5) is estimated more accurately than the idiosyncratic components (around 0 and 10, respectively) of the two samples nicely showcasing the borrowing of information across samples. Moreover, the posterior distributions of the number of components are reported in Table 1. The model correctly detects that each mixture has two components with one of them shared and the corresponding distributions are highly concentrated around the correct values. Finally the Bayes factor BF to test equality between the two distributions equals 0.00022 and the null hypothesis of distributional homogeneity is rejected.

Scenario III consists in generating the data from mixtures with the same components but differing in their weights. Specifically, $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ are drawn from, respectively,

$$X_{i,1} \sim 0.8\,\mathrm{N}(5, 1) + 0.2\,\mathrm{N}(0, 1) \quad X_{j,2} \sim 0.2\,\mathrm{N}(5, 1) + 0.8\,\mathrm{N}(0, 1),$$



(a) (b)

Figure 2: Simulated scenario II (mixtures of normal distributions with a common component): the estimated densities (blue) and true densities (red) generating $\boldsymbol{X}_1^{(100)}$ in Panel (a) and $\boldsymbol{X}_2^{(100)}$ in Panel (b).

The posterior distribution of the number of components is again reported in Table 1 and again the correct number is identified, although in this case the distributions exhibit a higher variability. The Bayes factor BF to test equality between the two distributions is 0.54, providing weak evidence in favor of the alternative hypothesis that the distributions differ.

5.2 Iris dataset

Finally, we examine the well known Iris dataset, which contains several measurements concerning three different species of Iris flower: setosa, versicolor, virginica. More specifically, we focus on petal width of those species. The sample $\boldsymbol{X}$ has size $n_1 = 90$, containing 50 observations of setosa and 40 of versicolor. The second sample $\boldsymbol{Y}$ is of size $n_2 = 60$ with 10 observations of versicolor and 50 of virginica.

Since the data are scattered across the whole interval $[0, 30]$, we need to allow for large variances and this is obtained by setting $(s_0, S_0) = (1, 4)$. The model neatly identifies that the two densities have two components each and that one of them is shared as showcased by the posterior probabilities reported in Table 2. As for the Bayes factor, we obtain BF ≈ 0 leading to the unsurprising conclusion that the two samples come from two different distributions. The corresponding estimated densities are reported in Figure 3.

We have also monitored the convergence of the algorithm that has been implemented. Though we here provide only details for the Iris dataset, we have conducted similar analyses also for each of the illustrations with synthetic datasets in Section 5.1. Notably, all the examples with simulated data have experienced even better performances than those we are going to display henceforth. Figure 4 depicts the partial autocorrelation function for the sampled parameters σ and σ_0. The partial autocorrelation function

# *comp.*	0	1	2	3	4	5	6	≥ 7
K_1	0	0	**0.466**	0.307	0.141	0.055	0.020	0.011
K_2	0	0.001	**0.661**	0.248	0.068	0.017	0.004	0.001
K_{12}	0	**0.901**	0.093	0.006	0	0	0	0

Table 2: Real data: Posterior distributions of the number of components in the first sample (K_1), in the second sample (K_2) and shared by the two samples (K_{12}). The posterior probabilities corresponding to the MAP estimates are displayed in bold.

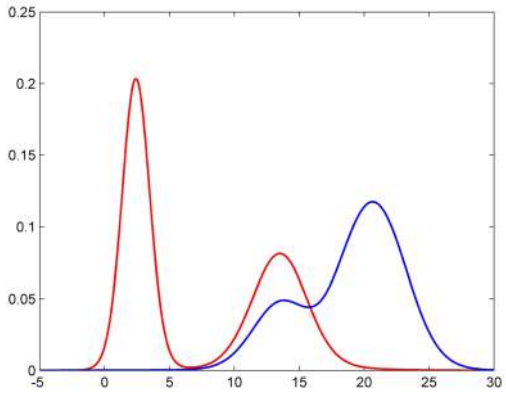


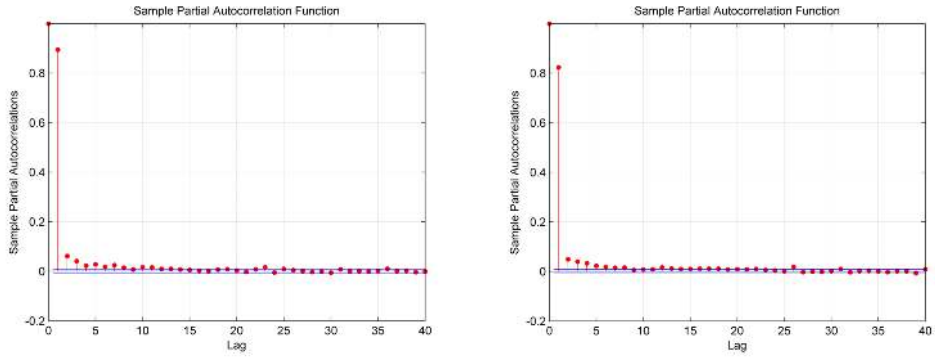
Figure 3: Iris dataset: the estimated densities for the first sample $\mathbf{X}$ (observations of setosa and versicolor) are shown in red, while the estimated densities for the second sample $\mathbf{Y}$ (observations of versicolor and virginica) are shown in blue.

apparently has an exponential decay and after the first lag exhibits almost negligible peaks.

We have additionally monitored the two estimated densities near the peaks, which identify the mixtures' components. More precisely, Figure 5(a) displays the trace plots of the density referring to the first sample at the points 3 and 13, whereas Figure 5(b) shows the trace plots of the estimated density function of the second sample at the points 13 and 21.

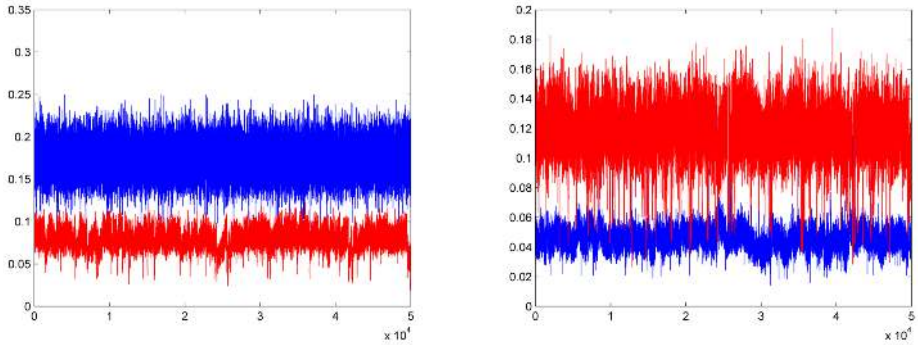
6 Concluding remarks

We have introduced and investigated a novel class of nonparametric priors featuring a latent nested structure. Our proposal allows flexible modeling of heterogeneous data and deals with problems of testing distributional homogeneity in two-sample problems. Even if our treatment has been confined to the case $d = 2$, we stress that the results may be formally extended to $d > 2$ random probability measures. However, their implementation would be more challenging since the marginalization with respect to $(\tilde{p}_1, \ldots, \tilde{p}_d)$ leads to considering all possible partitions of the d random probability mea-



(a) (b)

Figure 4: Iris dataset: plots of the partial autocorrelation functions for the parameters σ (a) and σ_0 (b).



(a) (b)

Figure 5: Iris dataset. Panel (a): trace plots of the estimated density, say $f_1(x)$, generating $\boldsymbol{X}$ at points $x = 3$ and $x = 13$; panel (b): trace plots of the estimated density, say $f_2(x)$, generating $\boldsymbol{Y}$ at the points $x = 13$ and $x = 21$.

sures. While sticking to the same model and framework which has been shown to be effective both from a theoretical and practical point of view in the case $d = 2$, a more computationally oriented approach would be desirable in this case. There are two possible paths. The first, along the lines of the original proposal of the nDP in Rodríguez et al. (2008), consists in using tractable stick-breaking representations of the underlying random probabilities, whenever available to devise an efficient algorithm. The second, which needs an additional significant analytical step, requires the derivation of a posterior characterization of $(\tilde{p}_1, \ldots, \tilde{p}_d)$ that allows sampling of the trajectories of latent nested processes and build up algorithms for which marginalization is not needed. Both will be the object of our future research.

Supplementary Material

Supplementary material to Latent nested nonparametric priors
(DOI: 10.1214/19-BA1169SUPP; .pdf).

References

Barrientos, A. F., Jara, A., and Quintana, F. A. (2017). "Fully nonparametric regression for bounded data using dependent Bernstein polynomials." *Journal of the American Statistical Association*, to appear. MR3671772. doi: https://doi.org/10.1080/01621459.2016.1180987. 1304

Bhattacharya, A. and Dunson, D. (2012). "Nonparametric Bayes classification and hypothesis testing on manifolds." *Journal of Multivariate Analysis*, 111: 1–19. MR2944402. doi: https://doi.org/10.1016/j.jmva.2012.02.020. 1304

Blei, D. M. and Frazier, P. I. (2011). "Distance dependent Chinese restaurant process." *Journal of Machine Learning Research*, 12: 2383–2410. MR2834504. 1304

Blei, D. M., NG, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3: 993–1022. 1304

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019a). "Distribution theory for hierarchical processes." *Annals of Statistics*, 47(1): 67–92. MR3909927. doi: https://doi.org/10.1214/17-AOS1678. 1304

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodríguez, A. (2019b). "Supplementary material to Latent nested nonparametric priors." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1169SUPP. 1313

Chung, Y. and Dunson, D. B. (2009). "Nonparametric Bayes conditional distribution modeling with variable selection." *Journal of the American Statistical Association*, 104(488): 1646–1660. MR2750582. doi: https://doi.org/10.1198/jasa.2009.tm08302. 1304

Dahl, D. B., Day, R., and Tsai, J. W. (2017). "Random partition distribution indexed by pairwise information." *Journal of the American Statistical Association* to appear. MR3671765. doi: https://doi.org/10.1080/01621459.2016.1165103. 1304

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). "Bayesian nonparametric nonproportional hazards survival modeling." *Biometrics*, 65(3): 762–771. MR2649849. doi: https://doi.org/10.1111/j.1541-0420.2008.01166.x. 1304

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). "An ANOVA model for dependent random measures." *Journal of the American Statistical Association*, 99(465): 205–215. MR2054299. doi: https://doi.org/10.1198/016214504000000205. 1304

Filippi, S. and Holmes, C. C. (2017). "A Bayesian nonparametric approach for quantifying dependence between random variables." *Bayesian Analysis*, 12(4): 919–938. MR3724973. doi: https://doi.org/10.1214/16-BA1027. 1304

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association*, 100(471): 1021–1035. MR2201028. doi: https://doi.org/10.1198/016214504000002078. 1304

Griffin, J. E., Kolossiatis, M., and Steel, M. F. J. (2013). "Comparing distributions by using dependent normalized random-measure mixtures." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 75(3): 499–529. MR3065477. doi: https://doi.org/10.1111/rssb.12002. 1304

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*, 79(2): 525–545. MR3611758. doi: https://doi.org/10.1111/rssb.12176. 1304

Griffin, J. E. and Steel, M. F. J. (2006). "Order-based dependent Dirichlet processes." *Journal of the American Statistical Association*, 101(473): 179–194. MR2268037. doi: https://doi.org/10.1198/016214505000000727. 1304

Hjort, N. L. (2000). "Bayesian analysis for a generalized Dirichlet process prior." Technical report, University of Oslo. 1304

Holmes, C., Caron, F., Griffin, J. E., and Stephens, D. A. (2015). "Two-sample Bayesian nonparametric hypothesis testing." *Bayesian Analysis*, 10(2): 297–320. MR3420884. doi: https://doi.org/10.1214/14-BA914. 1304

Jara, A., Lesaffre, E., De Iorio, M., and Quintana, F. (2010). "Bayesian semiparametric inference for multivariate doubly-interval-censored data." *Annals of Applied Statistics*, 4(4): 2126–2149. MR2829950. doi: https://doi.org/10.1214/10-AOAS368. 1304

Kingman, J. F. C. (1978). "The representation of partition structures." *Journal of the London Mathematical Society (2)*, 18(2): 374–380. MR0509954. doi: https://doi.org/10.1112/jlms/s2-18.2.374. 1308

Kingman, J. F. C. (1993). *Poisson processes*. Oxford University Press. MR1207584. 1306

Lijoi, A., Nipoti, B., and Prünster, I. (2014). "Bayesian inference with dependent normalized completely random measures." *Bernoulli*, 20(3): 1260–1291. MR3217444. doi: https://doi.org/10.3150/13-BEJ521. 1304

Ma, L. and Wong, W. H. (2011). "Coupling optional Pólya trees and the two sample problem." *Journal of the American Statistical Association*, 106(496): 1553–1565. MR2896856. doi: https://doi.org/10.1198/jasa.2011.tm10003. 1304

MacEachern, S. N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics. Simulation and Computation*, 23(3): 727–741. MR1293996. doi: https://doi.org/10.1080/03610919408813196. 1316

MacEachern, S. N. (1999). "Dependent nonparametric processes." In *ASA proceedings of the section on Bayesian statistical science*, 50–55. 1304

MacEachern, S. N. (2000). "Dependent Dirichlet processes." *Tech. Report, Department of Statistics, The Ohio State University*. 1304

Mena, R. H. and Ruggiero, M. (2016). "Dynamic density estimation with diffusive Dirichlet mixtures." *Bernoulli*, 22(2): 901–926. MR3449803. doi: https://doi.org/10.3150/14-BEJ681. 1304

Müller, P., Quintana, F., and Rosner, G. (2004). "A method for combining inference across related nonparametric Bayesian models." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 66(3): 735–749. MR2088779. doi: https://doi.org/10.1111/j.1467-9868.2004.05564.x. 1304

Müller, P., Quintana, F., and Rosner, G. L. (2011). "A product partition model with regression on covariates." *Journal of Computational and Graphical Statistics*, 20(1): 260–278. MR2816548. doi: https://doi.org/10.1198/jcgs.2011.09066. 1304

Nguyen, X. (2013). "Convergence of latent mixing measures in finite and infinite mixture models." *Annals of Statistics*, 41(1): 370–400. MR3059422. doi: https://doi.org/10.1214/12-AOS1065. 1304

Nguyen, X. (2015). "Posterior contraction of the population polytope in finite admixture models." *Bernoulli*, 21(1): 618–646. MR3322333. doi: https://doi.org/10.3150/13-BEJ582. 1304

Page, G. L. and Quintana, F. A. (2016). "Spatial product partition models." *Bayesian Analysis*, 11(1): 265–298. MR3465813. doi: https://doi.org/10.1214/15-BA971. 1304

Pitman, J. (1995). "Exchangeable and partially exchangeable random partitions." *Probab. Theory Related Fields*, 102(2): 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 1308

Regazzini, E., Lijoi, A., and Prünster, I. (2003). "Distributional results for means of random measures with independent increments." *Annals of Statistics*, 31: 560–585. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 1306

Rodríguez, A. and Dunson, D. B. (2011). "Nonparametric Bayesian models through probit stick-breaking processes." *Bayesian Analysis*, 6(1): 145–177. MR2781811. doi: https://doi.org/10.1214/11-BA605. 1304

Rodríguez, A. and Dunson, D. B. (2014). "Functional clustering in nested designs: modeling variability in reproductive epidemiology studies." *Annals of Applied Statistics*, 8(3): 1416–1442. MR3271338. doi: https://doi.org/10.1214/14-AOAS751. 1304

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103(483): 1131–1144. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 1304, 1305, 1306, 1321

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2010). "Latent stick-breaking processes." *Journal of the American Statistical Association*, 105(490): 647–659. MR2724849. doi: https://doi.org/10.1198/jasa.2010.tm08241. 1304

Soriano, J. and Ma, L. (2017). "Probabilistic multi-resolution scanning for two-sample differences." *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(2): 547–572. MR3611759. doi: https://doi.org/10.1111/rssb.12180. 1304

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 1304

West, M., Müller, P., and Escobar, M. D. (1994). "Hierarchical priors and mixture models, with application in regression and density estimation." In *Aspects of uncertainty*, 363–386. Wiley, Chichester. MR1309702. 1316

Acknowledgments

A. Lijoi and I. Prünster are partially supported by MIUR, PRIN Project 2015SNS29B.

Invited Discussion

Mario Beraha[†*] and Alessandra Guglielmi[*]

We thank the authors (also denoted by Camerlenghi *et al.* hereafter) for a very interesting paper, which addresses the problem of testing homogeneity between two populations/groups. They start from pointing out a drawback of the Nested Dirichlet Process (NDP) by Rodriguez et al. (2008), i.e. its degeneracy to the exchangeable case: when the NDP is a prior for two population distributions (or for the corresponding mixing measures in mixture models), it forces homogeneity across the two samples in case of ties across samples at the observed or latent level. In fact, as pointed out by Camerlenghi *et al.*, the NDP does not accommodate for shared atoms across populations. This limitation, which is clear from the definition of NDP in Rodriguez et al. (2008), has a strong impact on the inference: as showed in this paper, if two populations share at least one common latent variable in the mixture model, the posterior distribution would either identify the two random measures associated to the populations as completely different (i.e. it would not recover the shared components) or it would identify them as identical. The need for a more flexible framework is elegantly addressed by the authors who propose a novel class of Latent Nested Nonparametric priors, where a shared random measure is added to the draws from a nested random measure, hence accommodating for shared atoms. There are two key ideas in their model: (i) nesting discrete random probability measures as in the case of the nested Dirichlet process by Rodriguez et al. (2008), and (ii) contaminating the population distributions with a common component as in Müller et al. (2004) (or as in Lijoi et al., 2014). The latter yields dependency among the random probability measures of the populations and avoids the degeneracy issue pointed out by the authors, while the former accounts for testing homogeneity in two-sample problems.

As a comment on the computational perspective, we note that their Markov Chain Monte Carlo (MCMC) method relies on the analytical expression of the Partially Exchangeable Partition Probability Function (pEPPF), which the authors obtain in the special case of $I = 2$ populations. However, the sampling scheme poses significant computational issues even in the case of $I = 2$, needing to rely on Monte Carlo integration to approximate some intractable integrals.

In this comment, we address the problem of extending their mixture model class for testing homogeneity of I populations, with $I > 2$, according to the first *path* the authors mention in their concluding remarks. In particular, we assume the mixture model for I populations/groups, when the mixing random probability measures $(\widetilde{p}_1, \ldots, \widetilde{p}_I)$ have a prior distribution that is the Latent Nested Dirichlet process (LNDP) measure. This prior is more manageable than their general proposal, thanks to the stick-breaking representation of all the random probability measures involved, which can be easily truncated to give an approximation and is straightforward to compute. Here, we apply

[*]Politecnico di Milano, Milano, Italy, mario.beraha@polimi.it, alessandra.guglielmi@polimi.it
[†]Also affiliated with Università degli Studi di Bologna

the Latent Nested Dirichlet Process mixtures to simulated datasets from this paper, while the authors adopt a different latent nested nonparametric prior for $I = 2$ populations. By using the truncation approximation of stick breaking random probabilities, we do not need to resort to the pEPPF anymore and we are able to extend the analysis to cases with more than two populations.

However, our experience shows that this vanilla-truncation MCMC scheme does not scale well with I: the computational burden becomes demanding even for moderate values of I, which are common when testing homogeneity for different groups, for example while comparing a treatment in a small group of hospitals. If one assumes the LNDP as a prior for the mixing random probability measures $(\widetilde{p}_1, \ldots, \widetilde{p}_I)$, we have showed that we really need to derive either the posterior characterization of the LNDP, as suggested by the authors, or significantly more efficient truncation-based schemes.

1 Latent Nested Dirichlet Process Mixture Models

In this section, we make explicit the details of the definition of the Latent Nested Process that was introduced by the authors, and then consider the Latent Nested Dirichlet Process as the mixing distributions for I different populations. We also apply this model to synthetic data.

In what follows, we use the same acronyms as the authors, specifically NRMI and CRM for normalized random measure with independent increments and completely random measure respectively.

Consider the (Euclidean) space Θ and let $\mathbb{M}_\Theta$ be the space of all bounded measures on Θ. Let $\tilde{q}$ be a random probability measure, $\tilde{q} \sim \text{NRMI}[\nu, \mathbb{M}_\Theta]$ with intensity $\nu(ds, dm) = c\rho(s)dsQ(dm)$; here $c > 0$, ρ is a function defined on $\mathbb{R}^+$ under conditions

$$\int_0^{+\infty} \min\{1, s\}\rho(s)ds < +\infty, \quad \int_0^{+\infty} \rho(s)ds = +\infty,$$

and Q is a probability measure on $\mathbb{M}_\Theta$. We skip the details on the σ-algebras attached to the spaces we consider. We know that $\tilde{q} = \sum_{j=1}^{\infty} \tilde{\omega}_j \delta_{\tilde{\eta}_j}$, where $\{(\tilde{\omega}_j, \tilde{\eta}_j)\}$ are the points of a Poisson process with mean intensity $\nu(ds, dm)$. In particular, $\tilde{\eta}_j \overset{\text{iid}}{\sim} Q$, i.e. each $\tilde{\eta}_j$ is itself a CRM on Θ with Lévy intensity $\nu_0(ds, d\theta) = c_0\rho_0(s)dsQ_0(d\theta)$, which implies $\tilde{\eta}_j = \sum_{k=1}^{\infty} J_k^j \delta_{\theta_k^j}$, where, for each j, $\{(J_k^j, \theta_k^j), k \geq 1\}$ are the points of a Poisson process with mean intensity $\nu_0(ds, d\theta)$. Here $c_0 > 0$, ρ_0 is a function on $\mathbb{R}^+$ under the same conditions as $\rho(s)$ and Q_0 is a probability measure on Θ. Finally, let q_S be the law of μ_S, a CRM on Θ, with Lévy intensity $\nu_0^* = \gamma\nu_0$, where $\gamma > 0$.

Similarly to the authors, we define a Latent Nested Process as a collection of random probability measures $\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_I$ on Θ such that

$$\tilde{p}_i = \frac{\mu_i + \mu_s}{\mu_i(\mathbb{X}) + \mu_s(\mathbb{X})} = w_i \frac{\mu_i}{\mu_i(\mathbb{X})} + (1 - w_i)\frac{\mu_S}{\mu_S(\mathbb{X})}, \ i = 1, \ldots, I,$$

where

$$\mu_1, \mu_2, \ldots, \mu_I, \mu_S | \tilde{q}, q_S \sim \tilde{q} \times \tilde{q} \ldots \times \tilde{q} \times q_S.$$

In particular, if we set $\rho(s) = \rho_0(s) = s^{-1}\mathrm{e}^{-s}$, $s > 0$, we obtain the Latent Nested Dirichlet Process; since the μ_i's and μ_S are independent gamma processes in this case, the μ_i's also being iid, and

$$p_i = \frac{\mu_i}{\mu_i(\mathbb{X})}, i = 1, \ldots, I, \quad p_S = \frac{\mu_S}{\mu_S(\mathbb{X})},$$

i.e. p_i and p_S are draws from two independent Dirichlet processes, we have

$$p_i \mid G \overset{\text{iid}}{\sim} G = \sum_{l=1}^{\infty} \pi_l \delta_{G_l^*}, \ i = 1, \ldots, I, \tag{1.1}$$

$$p_S = \sum_{h=1}^{\infty} w_h^S \delta_{\theta_h^S}, \tag{1.2}$$

where G is a Nested Dirichlet process, i.e. a DP whose atoms are DPs. We use notation $(\tilde{p}_1, \ldots, \tilde{p}_I) \sim LNDP(\gamma, \nu_0, \nu)$ for

$$\widetilde{p}_i = w_i p_i + (1 - w_i) p_S, \quad i = 1, \ldots, I.$$

Note that each $\widetilde{p}_i$ is a mixture of two components: an idiosyncratic component p_i and a shared component p_S, where the latter preserves heterogeneity across populations even when shared values are present. As pointed out by the authors, the random indicator functions of the two events $\widetilde{p}_i = \widetilde{p}_{i'}$ and $p_i = p_{i'}$ coincide a.s., if $i \neq i'$. This latter event has positive prior probability for any couple of distinct indexes i, i' in $\{1, \ldots, I\}$. Summing up, this prior induces a prior distribution for the parameter $\boldsymbol{\rho}$, the partition of population indexes $\{1, 2, \ldots, I\}$: two populations are clustered together if they share the same mixing measure.

Now, suppose that we have data from I different populations (e.g. measurements on patients in different hospitals). Let y_{ji}, $j = 1, \ldots, n_i$, be observations for different subjects in population i, for $i = 1, \ldots, I$. We assume that, for any $i = 1, \ldots, I$,

$$\begin{aligned} y_{ji} \mid \widetilde{p}_i &\overset{\text{iid}}{\sim} \int_{\Theta} f(y_{ji} \mid \theta) \widetilde{p}_i(d\theta), \quad j = 1, \ldots, n_i \\ (\widetilde{p}_1, \ldots \widetilde{p}_I) &\sim LNDP(\gamma, \nu_0, \nu). \end{aligned} \tag{1.3}$$

For computing posterior inference, instead of considering model (1.3), we consider a truncation approximation of the stick-breaking representation of the LNDP, similarly as in Rodriguez et al. (2008). In particular, instead of (1.1)-(1.2), we consider the p_i's iid from a L-H truncation of a nested Dirichlet process, i.e.,

$$p_i | G \overset{\text{iid}}{\sim} \sum_{l=1}^{L} \pi_l \delta_{G_l^*}, \quad \pi_l = \nu_l \prod_{s=1}^{l-1} (1 - \nu_s), \quad \nu_l \overset{\text{iid}}{\sim} Beta(1, c) \quad l = 1, \ldots, L-1, \quad \nu_L = 1$$

$$G_l^* = \sum_{h=1}^{H} w_{lh} \delta_{\theta_{lh}^*}, \quad w_{lh} = u_{lh} \prod_{s=1}^{h-1} (1 - u_{ls}), \quad u_{lh} \overset{\text{iid}}{\sim} Beta(1, c_0) \quad h = 1, \ldots, H-1, \quad u_{lH} = 1$$

$$\theta_{lh}^* \overset{\text{iid}}{\sim} Q_0 \text{ for all } l, h$$

and p_S itself is an H–truncated Dirichlet Process of parameters γc_0 and Q_0. Since w_i is defined from the total masses of independent gamma processes, then

$$w_i = \frac{\mu_i(\Theta)}{\mu_i(\Theta) + \mu_S(\Theta)} \sim Beta(c_0, \gamma c_0), \quad i = 1, \dots, I.$$

This truncation approximation could be exploited to design blocked Gibbs sampling schemes as in Ishwaran and James (2001), or more general truncation schemes (see the references in Argiento et al., 2016); in the next section we use this truncation approximation in order to write a JAGS code to fit the data from the examples.

2 Simulation Study

We have fitted the truncated Latent Nested Dirichlet Process mixture model to simulated data via JAGS, using $L = 30$ and $H = 50$. The parametric kernel $f(y|\theta)$ in (1.3) is the unidimensional Gaussian density with mean θ and variance σ^2, i.e. $\theta = (\mu, \sigma)$. For every simulated dataset, we have considered the base measure $Q_0(\mu, \sigma) = \mathcal{N}\left(0, \lambda\sigma^2\right) \times \mathcal{U}(\sigma \mid 0, 2)$, with $\lambda = 10$. Moreover we set $c = c_0 = 1$ and let $\gamma \sim \mathcal{U}(0.25, 5)$. Chains were run for 10,000 iterations after 15,000 iterations of adaptation and 5,000 iterations of burn-in, thinning every 10 iterations for a final sample size equal to 1,000.

First, we considered two of the simulated scenarios examined in the paper, specifically scenarios I and II, and we simulated $n_1 = n_2 = 100$ observations from each group. Scenario I corresponds to full exchangeability across two groups of data, i.e.

$$y_{j1}, y_{j2} \overset{\text{iid}}{\sim} 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(5, 1),$$

while scenario II corresponds to partial exchangeability with a shared component between the populations

$$y_{j1} \overset{\text{iid}}{\sim} 0.9\mathcal{N}(5, 0.6) + 0.1\mathcal{N}(10, 0.6) \quad y_{j2} \overset{\text{iid}}{\sim} 0.1\mathcal{N}(5, 0.6) + 0.9\mathcal{N}(0, 0.6).$$

Both scenarios were tested in the paper under the same Gaussian kernel we consider, with a latent nested σ-stable mixture model instead of the LNDP as a prior for the mixing distributions. We have considered another simulated dataset from $I = 3$ populations, with $n_1 = n_2 = n_3 = 100$, that is

$$y_{j1} \overset{\text{iid}}{\sim} 0.2\mathcal{N}(5, 0.6) + 0.8\mathcal{N}(0, 0.6) \quad y_{j2} \overset{\text{iid}}{\sim} 0.2\mathcal{N}(5, 0.6) + 0.8\mathcal{N}(0, 0.6) \quad y_{j3} \overset{\text{iid}}{\sim} \mathcal{N}(-3, 0.6),$$

which corresponds to full exchangeability across populations $1, 2$ but not across $1, 2, 3$.

As pointed out by the authors, Bayes factors for homogeneity tests across populations are available as a by-product of their model. Homogeneity tests with hypotheses

$$H_0 : \tilde{p}_i = \tilde{p}_j \quad vs \quad H_1 : \tilde{p}_i \neq \tilde{p}_j \tag{2.1}$$

are performed by the authors in case $(i, j) = (1, 2)$, by introducing the auxiliary variable $\mathbb{I}_{\{\tilde{p}_1 = \tilde{p}_2\}}$ in their MCMC state space, so that draws from its posterior are straightforwardly available. In our formulation of the LNDP mixture model instead, we resort to

the cluster allocation variables of the nested process, $s_j = l$ iff $p_j = G_l^*$ for $j = 1, \ldots, I$, to perform the same tests.

In case of $I > 2$ populations, it is also possible to perform global tests on the cluster structure arising among the populations. In our new (third) scenario, we are interested in testing the presence of one single group against the presence of three groups (for example), i.e.

$$H_0 : \tilde{p}_1 = \tilde{p}_2 = \tilde{p}_3 \quad vs \quad H_1 : \tilde{p}_1 \neq \tilde{p}_2 \neq \tilde{p}_3.$$

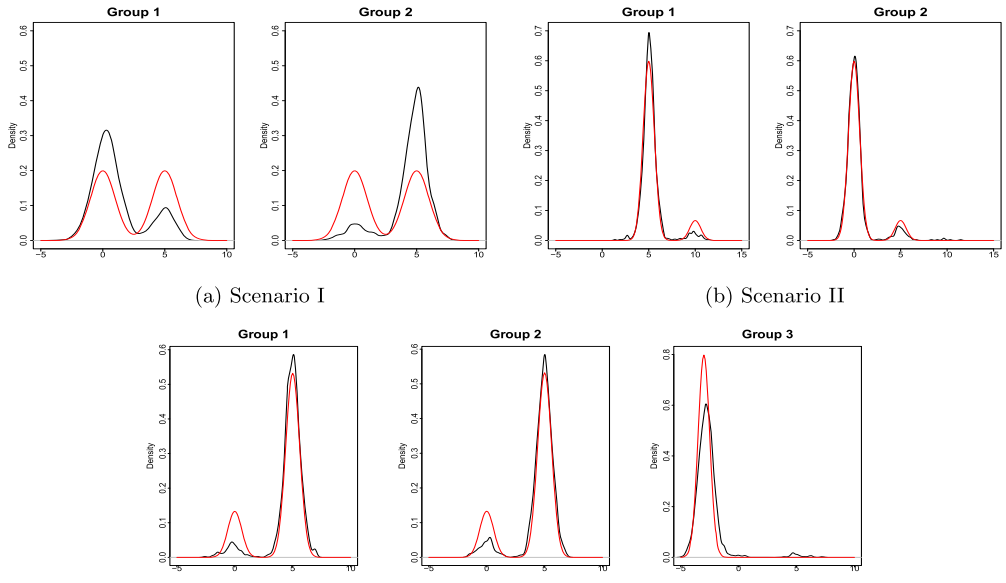
This type of tests are straightforward to obtain, since they are based on the EPPF of the nested process. Indeed, a priori, $P(\tilde{p}_1 = \tilde{p}_2 = \tilde{p}_3) = P(\boldsymbol{\rho} = \{1, 2, 3\})$ while $P(\tilde{p}_1 \neq \tilde{p}_2 \neq \tilde{p}_3) = P(\boldsymbol{\rho} = \{1\}, \{2\}, \{3\})$, where $\boldsymbol{\rho}$ is the partition of $\{1, 2, 3\}$ arising from the nested process; posterior odds are obtained once again monitoring the values of the allocation variables s_j's. The Bayes factor for this specific test equals 0.18, providing evidence in favour of H_1.

Scenario	(i, j)	BF_{01}
I	(1, 2)	1.00
II	(1, 2)	0.08
3 populations	(1, 2)	1.27
	(1, 3)	0.07
	(2, 3)	0.09

Table 1: Bayes factors for hypotheses (2.1) under the three simulated scenarios.

Table 1 reports the Bayes factors for tests (2.1) computed via our MCMC, while Figure 1 displays the predictive densities in each population. As far as the Bayes factors are concerned, we have computed those corresponding to hypotheses (2.1) with $(i, j) = (1, 2)$ for scenarios I and II, while for the new scenario we consider all the possible pairwise tests, i.e. $(i, j) = (1, 2),\ (1, 3),\ (2, 3)$. The Bayes factors in Table 1 correctly indicate strong evidence in favour of the alternative hypothesis for the second and third test of the 3-populations scenario, as well as for scenario II, while for the other tests there is no clear evidence in either direction. The BF_{01} for scenario II is much larger than the corresponding Bayes factor computed by the authors, obtained under the latent nested σ-stable mixture model; similarly, our BF_{01} for scenario I is equal to 1, while the authors obtain a larger value, giving evidence in favour of the true hypothesis. Of course, the mixing of the chain produced by JAGS, especially for scenario I with equal mixture weights, is generally worse than any specifically-designed MCMC scheme, as the one described by the authors. However, the density estimates (in black) for scenario II in Figure 1(b) are accurate, unlike those in Figure 1(a) where we clearly see that the JAGS code is not able to recover the weights in the true density in each group, while recovering the locations. Predictive densities in Figure 1(c) are close to the true population distributions in all the groups, even though we experienced the same difficulties in recovering the true weights of all the mixtures because of the large number of allocation parameters in the JAGS model, which makes sampling much less efficient.

To conclude our experiments, we have also designed a scenario with 4 populations simulating $n_i = 100$ observations from each true population distribution, which is a



(a) Scenario I

(b) Scenario II

(c) New scenario

Figure 1: Density estimates for scenario I (a), II (b) and the new scenario with $I = 3$ populations (c). In every panel, the black line denotes the predictive density in the population, while the red line is the density which generated the data.

mixture of two Gaussian components. The Bayes factors for hypotheses (2.1), computed via our JAGS MCMC, are in agreement with the true underlying clustering, that is $\{1, 2\}, \{3, 4\}$. However, even with as little as 100 observations per group, the MCMC simulation took more than 8 hours to run. To make a comparison, in our experience, the runtime of our JAGS code for $I = 3$ populations was about 2.5 times longer than for $I = 2$ populations, and that for $I = 4$ groups was approximately 4 times larger than for $I = 2$.

Despite the construction of ad-hoc Gibbs sampling schemes, possibly based on the truncated stick breaking representation, which could greatly improve the performances we reported, we believe that this model, generalized as we have presented here to the case of $I > 2$ populations and using a truncation approximation for the LNDP, contains inherent computational difficulties which are not easy to deal with. Assuming a larger value for I, even though a moderate value as in case of, e.g., comparing a patient treatment in a few dozens of hospitals, will still be challenging using the model we have considered here, taking into action the suggestion Camerlenghi *et al.* made in their concluding remarks.

References

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): 641–661. MR3489862. doi: https://doi.org/10.1007/s11222-015-9549-6. 1329

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 1329

Lijoi, A., Nipoti, B., Prünster, I., et al. (2014). "Bayesian inference with dependent normalized completely random measures." *Bernoulli*, 20(3): 1260–1291. MR3217444. doi: https://doi.org/10.3150/13-BEJ521. 1326

Müller, P., Quintana, F., and Rosner, G. (2004). "A method for combining inference across related nonparametric Bayesian models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3): 735–749. MR2088779. doi: https://doi.org/10.1111/j.1467-9868.2004.05564.x. 1326

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103(483): 1131–1154. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 1326, 1328

Invited Discussion

Vera Liu* and Peter Müller†

Camerlenghi et al. introduce two generalizations of the Nested Dirichilet Process (NDP) (Rodríguez et al., 2008). The first generalization (nested non-parametric process, NP) replaces the DP prior in the NDP construction by a normalized CRM (completley random measure). The change is methodologically minor, but practically very important. It allows substantially more flexibility with respect to the implied clusters. For any application that focuses on the implied partitions, the NP would seem a more appropriate choice than the NDP.

The major extension is introduced in the second generalization, the latent non-parametric process (LNP). The LNP model allows two random distributions (sub-populations), say $\tilde{p}_1$ and $\tilde{p}_2$, to share some related clusters and at the same time allows some clusters that are distinct and specific to each subpopulation. In contrast, the NDP only allows two cases: $\tilde{p}_1$ and $\tilde{p}_2$ are either identical or share no common atoms, i.e., implied clusters of experimental units would either be all in common across the two sub-populations, or all distinct. The NP model inherits the same limitation. The LNP maintains a positive probability for these two extreme cases, but also allows intermediate configurations with some shared and some subpopulation-distinct clusters.

Using the notation from the paper, the random probability measure $\tilde{p}_\ell$ for sub-population ℓ is defined as $\tilde{p}_\ell = w_\ell \frac{\mu_\ell}{\mu_\ell(X)} + (1 - w_\ell)\frac{\mu_s}{\mu_s(X)}$ with weight $w_\ell = \frac{\mu_\ell(X)}{\mu_\ell(X)+\mu_s(X)}$. Since μ_s includes all shared atoms that are in common across the two sub-populations, all shared atoms must have the same relative weights across different sub-populations. One example comes from the authors' simulation study, where the two Gaussian mixtures $X_{i,1} \sim 0.8N(5,1) + 0.2N(0,1)$ and $X_{i,2} \sim 0.2N(5,1) + 0.8N(0,1)$ have (only) shared components, but with different relative weights. In the posterior distribution each sample picks up one cluster that is not practically different from the shared clusters, just to accommodate the discrepancy in the weights. The model is not set up to accommodate varying relative weights. An extension or variation of LNP might be useful in some applications where more flexible weight assignment is desired.

Figure 1 highlights how different models accommodate different levels of flexibility in modeling heterogeneity across subpopulations. In the figure, "heterogeneity of clusters" refers to the mix of shared versus subpopulation-specific clusters; and "varying shared weights" refers to allowing the relative weights of shared clusters to vary across subpopulations. The figure places NDP (and NP), LNP, and the hierarchical DP (HDP) (Teh et al., 2006) with respect to these model features.

The restriction to common relative weights of the shared components becomes a practical limitation especially in the extension to $d > 2$ subpopulations. In that case

*Department of Statistics & Data Science, The University of Texas at Austin, TX, veraliu@utexas.edu

†Department of Statistics & Data Science, The University of Texas at Austin, TX

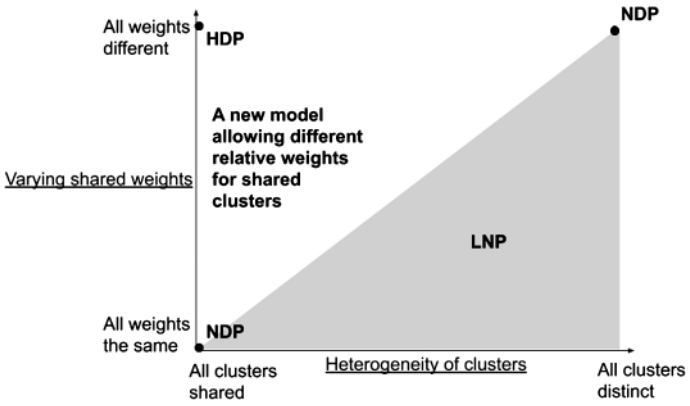


Figure 1: The figure places NDP, LNP and hierarchical DP (HDP) in a diagram to summarize flexibility with respect to (i) allowing a mix of shared and subpopulation-specific clusters (horizontal axis, "heterogeneity of clusters"); and (ii) allowing the relative weights of shared clusters to vary across subpopulations (vertical axis, "varying shared weights").

also the housekeeping in an implementation of posterior simulation becomes very burdensome, as one needs to distinguish unique values that are in common to all, or any subset of the d subpopulations. The authors mention this challenge in the conclusion. We would like to add some observations related to this issue.

Example. *Consider breast cancer gene expression data. Earlier studies have demonstrated that three subtypes of breast cancer – basal-like, HER2-enriched and luminal A – include some shared, overlapping features in addition to some subpopulation-specific features. This could be formalized as cluster locations in a mixture of normal model for gene expression (after suitable transformation), as, for example, in Xu et al. (2015). Let $y_{\ell i}$ denote the (transformed) data on a selected set of biomarkers for the i-th patient in subpopulation ℓ. If we were to use a model as in equation (1) (in the paper), we would use $d = 3$ random probability measures $\tilde{p}_\ell$, $\ell = 1, 2, 3$, and add an additional convolution with a normal kernel to define*

$$y_{\ell i} \sim G_\ell(y_{\ell i}) = \int N(y_{\ell i} \mid \theta)\, d\tilde{p}(\theta)$$

where θ is the pair of a multivariate normal mean and covariance matrix.

In this example, the assumption of common relative weights for shared components would be an unreasonable restriction (and this is no critique of the paper, as the authors never suggested any such applications). We feel the issue is not so much the restriction in the model, but rather in how the problem is set up. To start, we propose to first introduce a matrix of indicators to keep track of shared versus subpopulation-specific features (in anticipation of the upcoming argument we stop using the term "cluster"). For example, with the example in section 2.2 of the paper, $X_1 = (0.5, 2, -1, 5, 5, 0.5, 0.5)$

and $X_2 = (5, -2, 0.5, 0.5)$, instead of introducing k_1, k_2, k_0, we would set up a binary $(d \times k)$ matrix with $d = 2$ rows and $k = k_1 + k_2 + k_0$ columns,

$$Z = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

with parameters corresponding to each column, $\theta = (0.5, 2, -1, 5, -2)$. Finally, each subpopulation has a weight vector $w_\ell = (w_{\ell 1}, \ldots, w_{\ell k})$, which allows then to state a straightforward sampling model. The advantage is that now the parameters are $\omega = (Z, \theta, w)$, with k, k_0, k_1, k_2 being simple summaries of Z. The representation trivially generalizes to $d > 2$, without an explosion of notation and without tedious housekeeping needs.

The model is completed with a prior on Z, θ, w. Each column of Z defines a random subset (in this case, of the subpopulations), with $Z_{\ell c} = 1$ when feature c includes subpopulation ℓ. A prior on a family of random subsets ("features") is known as feature allocation. See, for example, Broderick et al. (2013) for a good review.

In summary, we argue that when the inference goal is to identify common and distinct features across d subpopulation, then perhaps it is more appropriate to set it up as a feature allocation problem. The downside, of course, is to lose the mathematical tractability of the specification with hierarchies over random probability measures. This includes, in particular, the characterization of the pEPPF (partially exchangable partition probability function) and related results in the paper. We appreciate the contribution of the paper in introducing an interesting new class of models for families of discrete random probability measures and related random partitions, and the elegant results on pEPPF's.

References

Broderick, T., Pitman, J., and Jordan, M. I. (2013). "Feature Allocations, Probability Functions, and Paintboxes." *Bayesian Analysis*, 8(4): 801–836. MR3150470. doi: https://doi.org/10.1214/13-BA823. 1335

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The Nested Dirichlet Process." *Journal of the American Statistical Association*, 103(483): 1131–1154. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 1333

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 1333

Xu, Y., Mueller, P., and Telesca, D. (2015). "Bayesian Inference for Latent Biologic Structure with Determinantal Point Processes (DPP)." *Biometrics*, 72. MR3545688. doi: https://doi.org/10.1111/biom.12482. 1334

Invited Discussion*

Fernando A. Quintana[†,‡,‖] and Alejandro Jara[§,¶,**]

1 The Proposal

We congratulate Federico Camerlenghi, David Dunson, Antonio Lijoi, Igor Prünster and Abel Rodríguez, from now on referred to as CDLPR, for an an interesting paper. CDLPR are indeed to be commended by a very fine work. By focusing on partitions they uncovered a critical degeneracy issue underlying the nested Dirichlet process (NDP) and other discrete nested processes. The key tool to understand the problem is the partially exchangeable partition probability function (pEPPF), that describes the probability model on partitions induced by models such as the NDP. CDLPR explicitly find the pEPPF in the case of $d = 2$ samples, showing that it is expressed as a mixture of patterns that may be described as partially exchangeable (both samples are marginally exchangeable, arising from different discrete random probability measures) and fully exchangeable (both samples arise from a common discrete random measure). Specifically, the pEPPF is a convex combination of these two forms. However, if it so happens that one atom is shared by these two random measures, then the structure degenerates to the fully exchangeable case. This is certainly a limitation of nested processes. As discussed in the manuscript, the problem is not restricted to the NDP, but will manifest itself in the case of *any* nested discrete random measure. CDLPR illustrate further this point with real data examples and synthetic data simulations. Solving the degeneracy problem motivated the (very clever) introduction of their latent nested process (LNP) approach. By allowing idiosyncratic and shared components, CDLPR overcome the degeneracy while retaining modeling flexibility. Essentially, the shared component can explicitly provide atoms in the mixture that are common to both samples, while the idiosyncratic components can adjust to local behavior without being forced to combine all of the mass in a single random measure. Their specific construction involves three random measures with which they create the LNP by normalizing the sum of idiosyncratic (μ_ℓ) and shared measures (μ_S) thus yielding $\tilde{p}_\ell$, as given in (14), where $\ell = 1, 2$. The structure is emphasized by noting that each $\tilde{p}_\ell$ can be expressed as a convex combination of normalized versions of μ_ℓ and μ_S. Indispensable to the proposal is the fact that flexibility does not come at the expense of practical tractability, which is of course vital for the practical success of models based on LNPs.

We discuss next some possible extensions to the model constructed by CDLPR.

*Supported by Millennium Science Initiative of the Ministry of Economy, Development, and Tourism, grant "Millenium Nucleus Center for the Discovery of Structures in Complex Data".

[†]Department of Statistics, Pontificia Universidad Católica de Chile, quintana@mat.uc.cl, url: http://www.mat.uc.cl/~quintana

[‡]Millennium Nucleus Center for the Discovery of Structures in Complex Data

[§]Department of Statistics, Pontificia Universidad Católica de Chile, atjara@uc.cl, url: http://www.mat.uc.cl/~ajara

[¶]Millennium Nucleus Center for the Discovery of Structures in Complex Data

[‖]Supported by Fondecyt grant 1180034.

[**]Supported by Fondecyt grant 1180640.

2 Model Extensions

2.1 $d > 2$

CDLPR consider the case of more than two samples in the final Discussion. An obvious problem of this setting is that analytic formulas get more complicated. Instead of dealing with atoms that are either shared by the two samples or specific to each sample, a general formula would require keeping track of all possible combinations of atoms shared by two samples, three samples, etc. and all atoms that are not shared by any other sample. This amounts to enlarging (19) to a summation over all possible partitions of $[d] = \{1, \ldots, d\}$, which is certainly substantially less appealing than the current formula. Nevertheless, in a simple case like $d = 3$, one may still be able to carry out these calculations analytically. But more generally, the extended computational approaches described in the discussion should be most welcome.

2.2 Dependence

Assume we have covariates $Z_{i,1}$ and $Z_{j,2}$ available for each of the elements in the $d = 2$ samples. We make the (implicit) assumption of matching dimensions in the dimensions of $Z_{i,1}$ and $Z_{j,2}$, just as is assumed in the case of $X_{i,1}$ and $X_{j,2}$. It is natural then to consider modifying the model to account for this extra information. An obvious alternative is to include a regression in each of the populations likelihood factors, probably with a suitable link function to account for cases where the sample mean is restricted to a subset of Euclidean spaces. Using the notation in §4 of the manuscript, this could be achieved by setting $h(x_{i,\ell}; \theta^*_{j,\ell}, \beta_i) = h(x_{i,\ell}; \theta^*_{j,\ell} + \beta_i^T z_{i,\ell})$, i.e. using the $\theta^*_{j,\ell}$ terms as random intercepts, with the convention that no constant variables are present in the design vectors. Of course, many more alternatives are possible.

An alternative way to consider much more general forms of dependence, which does not require big changes to the current analytical derivations would rely on the conditional approach introduced by Müller et al. (1996) to model jointly responses and covariates, $(\boldsymbol{X}, \boldsymbol{Z})$, say, and then examine the induced conditional distribution $\boldsymbol{X} \mid \boldsymbol{Z}$. Specifically, one may reinterpret $X_{i,1}$ and $X_{j,2}$ in formula (18) as $(X_{i,1}, Z_{i,1})$ and $(X_{j,1}, Z_{j,1})$ and proceed with the desired inference from there, with obvious adjustment to the dimensions of random variables.

Of more historical interest from the Bayesian nonparametric viewpoint is to consider dependence that affects the distributions of responses in a more general way. In other words, the idea would be to create a dependent version of the latent nested nonparametric priors in the sense of a collection of random probability measures indexed by predictors, in the spirit of the pioneering work of MacEachern (1999; 2000). Several alternatives are in turn available for such extension. In a general context, this may require sophisticated model choices, but to illustrate the point, consider a very simple version of the LNP, where each of the three defining measures μ_S, μ_1 and μ_2 are given by Gamma processes, which are special cases of completely random measures (CRMs). Extend now the definition of these Gamma processes to $\mathbb{R} \times \mathcal{Z}$, where $\mathcal{Z}$ is the covariate (or predictor) space, so that we have three CRMs indexed by predictors, $\{\mu_{S,z}\}_{z \in \mathcal{Z}}$,

and $\{\mu_{\ell,z}\}_{z\in\mathcal{Z}}$, for $\ell = 1, 2$. From this, a covariate-dependent version of a special case of LNP follows, and for each predictor value $z \in \mathcal{Z}$, the resulting normalized construction

$$\tilde{p}_{\ell,z} = \frac{\mu_{\ell,z} + \mu_{S,z}}{\mu_{\ell,z}(\mathbb{X}) + \mu_{S,z}(\mathbb{X})}, \qquad \ell = 1, 2,$$

follows the LNP model. A similar approach to construct Pólya Trees indexed by predictors was presented in Trippa et al. (2011).

3 Final Words

CDLPR present their LNP in the context of CRMs which include several well known and popular models as special cases. As in many cases of Bayesian Nonparametric priors, the real practical value of models lies in all potential applications that use such priors as building blocks for more complicated structures. In this sense, we look forward to many such applications and/or developments of the ideas discussed in the manuscript. Until then, we congratulate CDLPR again.

References

MacEachern, S. N. (1999). "Dependent Nonparametric Processes." In *ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association.*. 1337

MacEachern, S. N. (2000). "Dependent Dirichlet processes." Technical report, Department of Statistics, The Ohio State University. 1337

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian Curve Fitting Using Multivariate Normal Mixtures." *Biometrika*, 83(1): 67–79. MR1399156. doi: https://doi.org/10.1093/biomet/83.1.67. 1337

Trippa, L., Müller, P., and Johnson, W. (2011). "The multivariate beta process and an extension of the Polya tree model." *Biometrika*, 98(1): 17–34. MR2804207. doi: https://doi.org/10.1093/biomet/asq072. 1338

Acknowledgments

The authors acknowledge the support of Millennium Science Initiative of the Ministry of Economy, Development, and Tourism, grant "Millenium Nucleus Center for the Discovery of Structures in Complex Data".

Contributed Discussion

Li Ma*

I congratulate the authors on this excellent paper that provides deep insights into nested discrete processes, especially in revealing the distributional properties of the random partitions induced by these processes. Moreover, the authors skillfully embed the classical idea from Müller et al. (2004) into nested completely random measures to form a new class of latent nested nonparametric processes (LNNPs), thereby resolving a key difficulty in applying the nested Dirichlet process (DP) (Rodríguez et al., 2008) and its variant (Rodriguez and Dunson, 2014) when tied observations are present across data sets that are otherwise different. I shall discuss two challenges that a practitioner of the LNNP might face and how the LNNP might be further generalized to address them. Because nested discrete processes involve two layers of clustering structures—one at the sample level and the other at the observation level, to avoid confusion, I shall distinguish the two types by referring to them as sample cluster (SC) and observation cluster (OC) respectively.

To see the two challenges, note that by constructing each sampling distribution as a weighted average between a shared measure and a SC-specific idiosyncratic measure, the LNNP assumes that (i) samples that belong to the same SC must share exactly the same sampling distribution without any within-SC sample-to-sample variability; (ii) samples from different SCs must share the same relative weights over the OCs induced by the shared measure. Both of these assumptions can be too restrictive in applications. Next I shall consider each of them in turn.

Incorporating within-SC variation. Sample-to-sample variation is prevalent in applications. Even otherwise similar data sets collected under the same controllable conditions will inevitably display variability to various extents from each other. In the current context, not incorporating such variation within-SC can lead to many small or even singleton SCs when the number of observations in each sample grows. This was previously pointed out by MacEachern (2008) in his illuminating discussion on the nested DP.

It appears that the flexible nested process framework proposed in this paper is capable of incorporating such within-SC variation with some extension. In particular, for each SC, we can introduce a dispersion parameter—which can be scalar, multivariate, or even infinite-dimensional—that characterizes how samples within the SC differ. Specifically, let Ω_d denote the (Borel measurable) space of all possible values of the dispersion parameter. Let the Poisson random measure be defined as $\tilde{N} = \sum_{i\geq 1} \delta_{(J_i,G_i,w_i)}$ on $\mathbb{R}^+ \times \mathrm{P}_{\mathbb{X}} \times \Omega_d$ with a mean intensity function $\nu(\mathrm{d}s, \mathrm{d}p, \mathrm{d}v) = C\rho(s)\mathrm{d}s\, Q(\mathrm{d}p \times \mathrm{d}v)$, where Q is now a probability measure on the product space $\mathrm{P}_{\mathbb{X}} \times \Omega_d$.

Accordingly, define a completely random measure $\tilde{\mu} = \sum_{i\geq 1} J_i \delta_{(G_i,w_i)}$, and its normalized version

$$\tilde{q} := \frac{\tilde{\mu}}{\tilde{\mu}(\mathrm{P}_{\mathbb{X}} \times \Omega_d)}.$$

*Department of Statistical Science, Duke University, Durham, NC 27708, USA, li.ma@duke.edu

Given $\tilde{q}$, we can generate K sampling distributions $\tilde{p}_1, \tilde{p}_2, \ldots \tilde{p}_K$ hierarchically from their respective SCs with the corresponding within-SC variation as follows. First, we generate the SC centers (e.g., the mean distributions) and the within-SC dispersions

$$(\tilde{p}_{01}, \tilde{v}_1), \ldots, (\tilde{p}_{0K}, \tilde{v}_K) \,|\, \tilde{q} \sim \tilde{q}^K.$$

Given the SC centers and within-SC dispersions, we generate the sampling distributions

$$\tilde{p}_i \,|\, \tilde{p}_{0i}, \tilde{v}_i \overset{\text{ind}}{\sim} F(\tilde{p}_{0i}, \tilde{v}_i) \qquad \text{for } i = 1, 2, \ldots, K,$$

where $F(p, v)$ represents a "location-scale" family in the space of probability distributions with location (or center) p and scale (or dispersion) v.

While popular random measures such as the DP and the Pitman-Yor process can serve as this "location-scale" family, they are limited in that their dispersion parameter is either scalar or otherwise low-dimensional, and thus they cannot characterize flexible within-SC variation. A particularly flexible "location-scale" family is the Pólya tree (PT), which has an infinite-dimensional dispersion parameter with $\Omega_d = [0, \infty)^\infty$ as pointed out in Berger and Guglielmi (2001). Following this route, Christensen and Ma (In press) demonstrated a special case of the above nested model with $\rho(s) = s^{-1}e^{-s}$ and Q the product of an "adaptive PT" and a "stochastically increasing shrinkage" hyperprior on the dispersion parameter, both introduced in Ma (2017).

Allowing different relative weights on shared OCs. In many applications, the samples belonging to different SCs share OCs, and this is indeed one of the key motivations for the authors to introduce the LNNP as nested processes alone do not allow this feature. By introducing a shared component and building each sampling distribution as a weighted average of a shared and an idiosyncratic component in the style of Müller et al. (2004), the LNNP assumes that the idiosyncratic components do not share any OCs whereas the shared component must endow all the shared OCs with exactly the same relative weights among them. These constraints can be overly restrictive in practice. For example, Soriano and Ma (2019) considered the application of flow cytometry, where the observations are blood cells and each OC corresponds to a cell subtype (e.g., T-cells, B-cells, etc.). Different subtypes of patients (the SCs) will share some of the same cell subtypes, or one should hope so as they are all humans! In other words, the actual SCs might differ only in the weights of some OCs, not their identities.

The strategy that Soriano and Ma (2019) employed to address this issue is to let all the samples share a common set of OCs, and introduce shared and idiosyncratic components only in generating the weights of these OCs. This way, while the shared component still corresponds to the OCs with the same relative weights across SCs as in the LNNP, the idiosyncratic components now allow SCs to have distinct weights on common OCs. (Note that this strategy still allows SCs to have unique OCs in that an SC without a certain OC will just have a very small weight on that OC.) As Surya Tokdar pointed out through personal communications, this strategy is essentially a (limiting) version of a Müller et al. (2004) style mixture of a shared DP and idiosyncratic DPs generated from a hierarchical DP (Teh et al., 2006). Unlike the LNNP, this "LHDP" does not allow inference on the partition at the sample level. It is of interest to investigate how to allow different weights on shared OCs across SCs in the LNNP in such a way that maintains its ability to carry out the partitioning or clustering on the samples.

References

Berger, J. O. and Guglielmi, A. (2001). "Bayesian and Conditional Frequentist Testing of a Parametric Model versus Nonparametric Alternatives." *Journal of the American Statistical Association*, 96(453): 174–184. URL http://www.jstor.org/stable/2670357. MR1952730. doi: https://doi.org/10.1198/016214501750333045. 1340

Christensen, J. and Ma, L. (In press). "A Bayesian hierarchical model for related densities using Pólya trees." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, doi: https://doi.org/10.1111/rssb.12346. 1340

Ma, L. (2017). "Adaptive Shrinkage in Pólya Tree Type Models." *Bayesian Analysis*, 12(3): 779–805. MR3655876. doi: https://doi.org/10.1214/16-BA1021. 1340

MacEachern, S. N. (2008). "Discussion of "The nested Dirichlet process" by A. E. Gelfand, D. B. Dunson and A. Rodriguez." *Journal of the American Statistical Association*, 103: 1149–1151. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 1339

Müller, P., Quintana, F., and Rosner, G. (2004). "A method for combining inference across related nonparametric Bayesian models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3): 735–749. MR2088779. doi: https://doi.org/10.1111/j.1467-9868.2004.05564.x. 1339, 1340

Rodriguez, A. and Dunson, D. B. (2014). "Functional clustering in nested designs: Modeling variability in reproductive epidemiology studies." *The Annals of Applied Statistics*, 8(3): 1416–1442. MR3271338. doi: https://doi.org/10.1214/14-AOAS751. 1339

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The Nested Dirichlet Process." *Journal of the American Statistical Association*, 103(483): 1131–1154. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 1339

Soriano, J. and Ma, L. (2019). "Mixture Modeling on Related Samples by ψ-Stick Breaking and Kernel Perturbation." *Bayesian Analysis*, 14(1): 161–180. MR3910042. doi: https://doi.org/10.1214/18-BA1106. 1340

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). MR2279480. doi: https://doi.org/10.1198/016214506000000302. 1340

Contributed Discussion*

Christian P. Robert†

While this paper is a fairly theoretical piece of work, it manages in completely and beautifully validating a Bayesian approach for the non-parametric clustering of separate populations with "common" clusters. More formally, it constructs a new family of models that allows for a partial or complete equality between two probability measures, but does not result in the artifact that forces full identity when the associated samples do share some common observations. Indeed, more traditional structures prohibit one or the other, from the Dirichlet process (DP) preventing two probability measure realisations from being equal or partly equal to some hierarchical DP (HDP) already allowing for common atoms across measure realisations, but failing to authorise a complete identity between two realised distributions, to nested DP offering one extra level of randomness, but with an infinity of DP realisations that bars common atomic support from happening, besides completely identical support (and hence distribution).

The current paper thus posits two realisations of random measures to be decomposed as a sum of (i) a common random measure and (ii) one among two separate almost independent random measures: equation (14) is the core representation in the paper that allows for partial or total equality. An extension to a setting larger than facing two samples seems however complicated if only because of the number of common measures one has to introduce, from the integrally common measure to measures that are only shared by a subset of the samples. Except in the simplified framework when a single and universally common measure is adopted (with enough justification). The randomness of the model is handled via different completely random measures that involve the unusual recourse to four degrees of hierarchy in the Bayesian model.

Since the mixture example is central to the paper, the case of one or rather of two two-component Normal mixtures with a common component (but with different mixture weights) is handled by the advertised approach, although it seems to me that it could be covered by a simpler HDP. Having exactly the same term (i.e., with the very same weight) is not covered, but this may be of lesser appeal in real life applications. Note that alternative, easily constructed, parametric constructs are already available in this specific case, involving a limited prior input and a lighter computational burden, although the Gibbs sampler behind the model proves extremely simple for the approach advocated therein. (One may still wonder at the robustness of the sampler once a case of identical distributions occurs.)

Due to the combinatoric explosion associated with a higher number of observed samples, despite obvious practical situations, one may wonder at any feasible (and possibly

*I am most grateful to the Bayesian reading group in CEREMADE, Paris Dauphine, for conducting a discussion on this paper. This work is partly supported by a Senior Institut Universitaire de France Fellowship.

†CEREMADE, Université Paris Dauphine, PSL, France, Department of Statistics, University of Warwick, UK, xian@ceremade.dauphine.fr

sequential) extension, one that would further keep a coherence under marginalisation (in the number of samples). And also whether or not multiple testing could be coherently envisioned in this setting, for instance when handling all hospitals in the UK. Another consistency question covers the Bayes factor used to assess whether the two distributions behind the samples are or not identical. (One may further question the relevance of the question, hopefully applied to more relevant dataset than the Iris data.)

Contributed Discussion

Fabrizio Leisen* and Alan Riva Palacio†

We would like to congratulate the authors of Camerlenghi et al. (2018) for their insightful paper on nested processes. We strongly believe that this paper not only provides a new perspective on the nested Dirichlet process of Rodriguez et al. (2008) but also on the modelling of heterogeneous data with Bayesian nonparametric priors. We welcome contributions to Bayesian nonparametrics which exhibit a theoretical rigour as the one displayed in this paper.

1 Extending the LNP with CoRMs

In this discussion we would like to highlight an extension of the process displayed in (14) which would provide a further modelling flexibility. The authors' proposal relies on normalizing the vector of measures $(\mu_1 + \mu_S, \mu_2 + \mu_S)$. A first observation is that the measure μ_S influences the dependence among the two components although, the fact that is shared by both measures, limits the dependence modelling. We propose to consider a vector $(\mu_1 + \mu_S^1, \mu_2 + \mu_S^2)$ such that (μ_S^1, μ_S^2) is a *Compound Random Measure* (CoRM), see Griffin and Leisen (2017). Before commenting the benefits of this approach, we provide a description of CoRMs in terms of discrete measures. Consider the CRM:

$$\tilde{\mu}^* = \sum_{i \geq 1} J_i \delta_{X_i}$$

with *directing Lévy measure* ν^* and $X_k \overset{i.i.d.}{\sim} G_0$, for some non-atomic measure G_0. Compound Random Measures could be described in an intuitive way as a perturbation of the above CRM:

$$\tilde{\mu}_1 = \sum_{i \geq 1} m_{1,i} J_i \delta_{X_i} \qquad \cdots \qquad \tilde{\mu}_d = \sum_{i \geq 1} m_{d,i} J_i \delta_{X_i},$$

where $(m_{1i}, \ldots, m_{di}) \overset{i.i.d.}{\sim} h$ (*score distribution*) and $X_k \overset{i.i.d.}{\sim} G_0$, for some non-atomic measure G_0. Griffin and Leisen (2018) employed CoRMs in a density regression context whereas Riva Palacio and Leisen (2018) used CoRMs to develop new modelling tools in survival analysis.

The vector (μ_S^1, μ_S^2) is chosen as a CoRM with $d = 2$. The score distribution h is a parametric distribution which allows to model the dependence across measures. As shown in Riva Palacio and Leisen (2019) in the context of subordinators, this distribution influences the correlation between components. Such modelling might be appealing in a

*School of Mathematics, Statistics and Actuarial Sciences, University of Kent, fabrizio.leisen@gmail.com

†Departamento de Probabilidad y Estadística, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, alan_riva@ciencias.unam.mx

scenario where a component is shared but with different weights. If the Lévy intensity of (μ_S^1, μ_S^2) is supported on the diagonal line then the above pEPPF collapses in the one defined by Camerlenghi et al. (2018); if, on the other hand, it is supported on the axis $\{(x,y) \in (\mathbb{R}^+)^2 : x = 0\} \cup \{(x,y) \in (\mathbb{R}^+)^2 : y = 0\}$ then the pEPPF reduces to the EPPF framework. We can effectively use a CoRM to modulate between such behaviours. If we consider a multivariate Log-normal distribution for the score distribution in a CoRM we can allocate the mass of the respective Lévy measure in a 2-dimensional space and modulate between the behaviours discussed above by suitably choosing the mean vector and covariance-variance matrix of the underlying Gaussian distribution. Let $\boldsymbol{v} \in \mathbb{R}^2$ be a mean vector and $\boldsymbol{\Sigma}$ be a variance-covariance matrix defining a bivariate Gaussian distribution, we denote LogNormal$(\boldsymbol{v}, \boldsymbol{\Sigma})$ for the associated LogNormal distribution. The CoRM given by an arbitrary directing Lévy measure $\nu^\star$ and score distribution LogNormal$(\boldsymbol{v}, \boldsymbol{\Sigma})$ can be seen to be well defined by using the integrability condition result in Riva Palacio and Leisen (2019).

References

Camerlenghi, F., Dunson, D. B., Lijoi, A., Pruenster, I., and Rodríguez, A. (2018). "Latent Nested Nonparametric Priors (with discussion)." *Bayesian Analysis*. Advance publication. URL https://doi.org/10.1214/19-BA1169. 1344, 1345

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. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12176. MR3611758. doi: https://doi.org/10.1111/rssb.12176. 1344

Griffin, J. E. and Leisen, F. (2018). "Modelling and Computation Using NCoRM Mixtures for Density Regression." *Bayesian Analysis*, 13(3): 897–916. MR3807871. doi: https://doi.org/10.1214/17-BA1072. 1344

Riva Palacio, A. and Leisen, F. (2018). "Bayesian nonparametric estimation of survival functions with multiple-samples information." *Electronic Journal of Statistics*, 12(1): 1330–1357. MR3797716. doi: https://doi.org/10.1214/18-EJS1420. 1344

Riva Palacio, A. and Leisen, F. (2019). "Compound vectors of subordinators and their associated positive Lévy copulas." *Arxiv preprint*. URL https://arxiv.org/abs/1909.12112. 1344, 1345

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The Nested Dirichlet Process." *Journal of the American Statistical Association*, 103(483): 1131–1154. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 1344

Contributed Discussion

Emanuele Aliverti[*], Sally Paganin[†], Tommaso Rigon[‡], and Massimiliano Russo[§,¶]

We congratulate the authors on an interesting paper, which provides a concrete contribution in Bayesian nonparametric methods. The proposed latent nested process (LNP) of Camerlenghi *et al.* is a notable generalization of the nested Dirichlet process (NDP) of Rodríguez et al. (2008). In the first place, Camerlenghi *et al.* extend the NDP to a broader class of nested processes (NP), leveraging on homogeneous random measures with independent increments (Regazzini et al., 2003). They elegantly frame this novel class of priors within the theory of completely random measures.

The rigorous theoretical study of the involved clustering mechanism allows Camerlenghi *et al.* to identify a potential pitfall of general NPs. Specifically, two random discrete distributions $\tilde{p}_\ell$ and $\tilde{p}_{\ell'}$, associated to different groups (populations) and distributed according to a NP, are either identical (i.e. $\tilde{p}_\ell = \tilde{p}_{\ell'}$ a.s.), or they do not have common atoms. This behavior implies that NPs can borrow information across groups only in an extreme fashion, that is, by assuming full homogeneity across populations. In contrast, the LNP generalization accommodates smooth transitions between the full homogeneity and the independence cases, while still accounting for clustering across different populations.

We will focus on the latent nested Dirichlet process special case, which has been considered by Camerlenghi *et al.* in their Example 2. First recall that the NDP of Rodríguez et al. (2008), in presence of $d \geq 2$ populations, can be alternatively defined through a Blackwell and MacQueen (1973) urn-scheme. Let δ_x denote a point mass at x. If $(p_1, \ldots, p_d)$ is a collection of random probability measures on a complete and separable metric space $\mathbb{X}$ following an NDP, then for any $c > 0$

$$p_{\ell+1} \mid p_1, \ldots, p_\ell \sim \frac{c}{c+\ell} Q + \frac{1}{c+\ell} \sum_{i=1}^{\ell} \delta_{p_i}, \qquad \ell = 1, \ldots, d-1, \tag{1}$$

where Q is the probability distribution of a Dirichlet process $\tilde{q}_0 \sim \text{DP}(c_0 Q_0)$, with precision parameter $c_0 > 0$ and with Q_0 being a non-atomic probability measure on $\mathbb{X}$. In other words, each p_ℓ is either a sample from a $\text{DP}(c_0 Q_0)$ or is set equal to one of the previously observed random measures.

The latent nested Dirichlet process is built upon (1). More precisely, the vector of random probability measures $(\tilde{p}_1, \ldots, \tilde{p}_d)$ characterizing such a process is obtained as a

[*]Department of Statistical Sciences, Università degli studi di Padova, Padova, Italy, aliverti@stat.unipd.it

[†]Department of Environmental Science, Policy & Management, University of California Berkeley, Berkeley, USA, sally.paganin@berkeley.edu

[‡]Department of Statistical Science, Duke University, Durham, USA, tommaso.rigon@duke.edu

[§]Harvard–MIT Center for Regulatory Science, Harvard Medical School, Boston, USA, massimiliano_russo@hms.harvard.edu

[¶]Department of Data Science Dana-Farber Cancer Institute, Boston, USA

convex combination of two random probability measures, namely

$$\tilde{p}_\ell = w_\ell p_\ell + (1 - w_\ell) p_S, \qquad \ell = 1, \dots, d, \tag{2}$$

where $p_S \sim \text{DP}(\gamma c_0 Q_0)$, with $\gamma > 0$, is independent on $p_1, \dots, p_d$ whereas $w_\ell \overset{\text{iid}}{\sim} \text{Beta}(c_0, \gamma c_0)$, independently on the random probability measures $p_1, \dots, p_d$ and p_S.

As formalized by Proposition 4 in Camerlenghi *et al.*, some random probability measures among $\tilde{p}_1, \dots, \tilde{p}_d$ will be identical with positive probability. Broadly speaking, this occurs if ties are present in the underlying urn-scheme of (1). In the Dirichlet case, the *a priori* probability of homogeneity among two distributions is

$$\pi_1^* := \mathbb{P}(p_\ell = p_{\ell'}) = \mathbb{P}(\tilde{p}_\ell = \tilde{p}_{\ell'}) = \frac{1}{c+1}, \qquad \ell \neq \ell'.$$

Thus, Camerlenghi *et al.* suggest to evaluate the posterior probability $\mathbb{P}(p_\ell = p_{\ell'} \mid \boldsymbol{X})$, to test the null hypothesis $H_0 : \tilde{p}_\ell = \tilde{p}_{\ell'}$ against the alternative $H_1 : \tilde{p}_\ell \neq \tilde{p}_{\ell'}$. Such an approach is appealing as it naturally follows from the model construction.

Although this testing procedure is theoretically well-justified, there might be few practical difficulties that are worth emphasizing. Consider the example in Scenario II of Section 5.1 in Camerlenghi *et al.*, in which there are two mixtures of two normal distributions with a common component. The two distributions can be made equal either allowing the weight of the idiosyncratic component to be zero, or having arbitrary weights and letting the distribution-specific components to have the same parameters. The former case can easily be encountered. In fact, when the parameter γ is large enough one has that $w_\ell \approx 0$, in turn implying that $\tilde{p}_\ell \approx \tilde{p}_S$. This statement is formalized in the following lemma, whose proof is omitted.

Lemma 1. *Let $(\tilde{p}_1, \dots, \tilde{p}_d)$ be a latent nested Dirichlet process of (1)–(2). Then $\tilde{p}_1 = \cdots = \tilde{p}_d$ almost surely, as $\gamma \to \infty$.*

Lemma 1 holds for general LNPs and it has relevant consequences. Strictly speaking, it implies that homogeneity among populations is recovered as limiting case when $\gamma \to \infty$, regardless of the ties occurring in the Pólya-sequence of (1). Besides, (2) suggests that homogeneity between two groups (i.e. $\tilde{p}_\ell = \tilde{p}_{\ell'}$) is attained exactly whenever $p_\ell = p_{\ell'}$ but also approximately if $w_\ell \approx 0$. This could affect the rationale underlying the testing procedure, because an LNP model may struggle in discriminating between the case of two identical latent distributions ($p_\ell = p_{\ell'}$) and that of two similar, yet different, random probability measures ($w_\ell \approx 0$). Note that this issue is specific to the LNP, since nested processes correspond to the case $w_\ell = 1$.

As a consequence, if an LNP is employed for testing purposes, the probability of homogeneity $\mathbb{P}(\tilde{p}_\ell = \tilde{p}_{\ell'} \mid \boldsymbol{X})$ might be deflated, possibly leading to biased decisions. Hence, we recommend to select the parameter γ with great care. In contrast, if the LNP were used for density estimation, these considerations would not be a concern.

References

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1(2): 353–355. MR0362614. 1346

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. MR1983542. doi: https://doi.org/10.1214/aos/1051027881. 1346

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103(483): 1131–1144. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 1346

Rejoinder

Federico Camerlenghi[*,†], David B. Dunson[‡], Antonio Lijoi[§], Igor Prünster[§] and Abel Rodríguez[¶]

We are extremely grateful to all the discussants for their insightful comments and stimulating ideas. While the last 15 years have witnessed a vast literature on BNP modeling for heterogeneous data, there are still many aspects to be investigated and, maybe even more importantly, to gain a deep understanding of the inferential implications of the available modeling choices. And, the discussion of our paper is clear evidence of this.

The main goal we pursue in our paper is to propose and investigate a model that is able to account for clustering both probability distributions and observations (or latent features) across multiple samples. Indeed, we have shown that the Nested Dirichlet Process (NDP), which was originally proposed with this purpose, fails to accomplish the task since as soon as clusters of data (or latent variables) are shared across samples, the samples themselves are clustered as well. This feature is formalized in terms of degeneracy of the posterior distribution of $(\tilde{p}_1, \tilde{p}_2)$, induced by a NDP, on the diagonal $p_1 = p_2$, when conditioning on samples sharing a common cluster. Such a limitation has been overlooked in the literature. In addition to singling out this limitation, by introducing a new class of models, Latent Nested Priors (LNPs), we also highlight that it can be overcome so to preserve the appealing nested structure. As mentioned in the paper, our proposal is based on an additive composition of shared and idiosyncratic random measures and the nested structure is specified at the level of the idiosyncratic components. Its main goal was to preserve heterogeneity even when clusters are shared across samples and this was achieved. Obviously our construction has also limitations, such as those mentioned in the discussions, and we hope this paper together with its discussion will spur important improvements. In particular, either efficient computational-based approaches or alternative model formulations are required to handle the multiple population scenario when the number of groups equals $d > 2$ (see Section 1). At the same time, greater flexibility may be achieved by allowing the weights of the shared component in the model to be different across different populations. These two points are the main focus of our Rejoinder, which also investigates other interesting properties of LNP's relevant for prior specification as well as further extensions to allow for more complex covariate-dependence.

[*]Department of Economics, Management and Statistics, University of Milano - Bicocca, Piazza dell'Ateneo Nuovo 1, 20126 Milano, Italy, federico.camerlenghi@unimib.it

[†]Also affiliated to Collegio Carlo Alberto, Torino and BIDSA, Bocconi University, Milano, Italy

[‡]Department of Statistical Science, Duke University, Durham, NC 27708-0251, U.S.A., dunson@duke.edu

[§]Department of Decision Sciences and BIDSA, Bocconi University, via Röntgen 1, 20136 Milano, Italy, antonio.lijoi@unibocconi.it; igor@unibocconi.it

[¶]Department of Applied Mathematics and Statistics, University of California at Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, U.S.A., abel.rod@ucsc.edu

1 Beyond the two samples case

In the paper we focused on the case of $d = 2$ groups of observations: both theoretical results and marginal algorithms are obtained for this special case. As stressed in the Concluding Remarks (Section 6 paper), one may still formally display results analogous to Propositions 2–4, Theorems 1–2 for $d > 2$, although the combinatorial hurdles that are involved make them of little practical use. A possible solution is to resort to a computationally oriented approach as effectively remarked in the discussions by **M. Beraha & A. Guglielmi**, **F. A. Quintana & A. Jara** and **C.P. Robert**.

The contribution by **Beraha & Guglielmi** is centered around the scenario with multiple-populations, namely with $d > 2$: they focus on a suitable extension of the mixture model to test homogeneity across the samples and to allow for posterior inference. They specify a Latent Nested Dirichlet Process (LNDP) for the vector $(\tilde{p}_1, \ldots, \tilde{p}_d)$ so that they can rely on the stick-breaking construction of the Dirichlet process and on its conjugacy properties. Their simulations are run in JAGS, but unfortunately the computational cost is still demanding. They also highlight that the results obtained using the LNDP along with the truncation scheme are generally worse than the ones we obtain using the latent nested σ-stable process implemented via a marginal MCMC sampler. For example the Bayes factor (BF) derived using the LNDP in Scenario I is equal to 1 (see Table 1 in **Beraha & Guglielmi**), while the BF obtained with stable LNP in Scenario I equals 5.85. We think this behavior is due to two main factors: the limited flexibility induced by the Dirichlet Process and the algorithm, which is not exact since it is based on the truncation of a series representation. An *ad hoc* MCMC scheme designed for LNDP could probably improve the mixing of MCMC algorithm, however one may look for more efficient solutions to overcome both computational issues and the poor model flexibility of the LNDP. An option that is definitely worth investigating relies on the possible characterization of the posterior distribution of the vector of LNP $(\tilde{p}_1, \ldots, \tilde{p}_d)$. One may, then, use suitable conditional methods such as the importance conditional Gibbs sampler introduced in Canale et al. (2019) for the Dirichlet and the Pitman–Yor process.

It would be interesting to compare the results in **Beraha & Guglielmi** with those that one would obtain from a Latent Nested Pitman–Yor processes, which is more flexible a prior specification compared to the Dirichlet process and still tractable from an analytical standpoint. More precisely let us denote by Q_σ, the law of a σ-stable CRM on the space of boundedly finite measure $\mathsf{M}_{\mathbb{X}}$ on $\mathbb{X}$. For any $\theta > 0$, define $Q_{\sigma,\theta}$ on $\mathsf{M}_{\mathbb{X}}$ as absolutely continuous w.r.t. Q_σ with Radon–Nikodym derivative

$$\frac{\mathrm{d}Q_{\sigma,\theta}}{\mathrm{d}Q_\sigma}(m) = \frac{\sigma\,\Gamma(\theta)}{\Gamma(\theta/\sigma)}\, m^{-\theta}(\mathbb{X}).$$

Note that if one considers a random measure $\tilde{\mu}_{\sigma,\theta}$ with distribution $Q_{\sigma,\theta}$, then under normalization one has the Pitman–Yor process. Hence latent nested Pitman–Yor processes may be defined as follows: consider a vector $(\tilde{p}_1, \tilde{p}_2)$ as in Equation (14) of the paper with $\mu_S \sim Q_{\sigma,\theta}$ and $\mu_1, \mu_2 | \tilde{q} \overset{\text{iid}}{\sim} \tilde{q}$, where $\tilde{q}$ is a NRMI arising from a CRM with Lévy intensity $c\rho(s)\mathrm{d}s\, Q_{\sigma,\theta}(\mathrm{d}p)$. One may think to develop efficient and fast algorithms for this class of models, exploiting the quasi-conjugacy property of the Pitman–Yor process and its stick breaking representation.

2 Improving the model flexibility

Most of the discussions focus on suitable extension of the LNP to allow for increased modeling flexibility. The main critical issue that has been raised is related to the same common component, which is crucial to preserve heterogeneity across different groups when clusters are shared. Indeed, we agree this is not consistent with more realistic situations whereby shared clusters of observations have different relative weights and we are happy to see this generated interesting ideas by **F. Leisen & A. R. Palacio**, **V. Liu & P. Müller** and **L. Ma**. From the mixture representation of each random probability $\tilde{p}_\ell$

$$\tilde{p}_\ell = w_\ell \frac{\mu_\ell}{\mu_\ell(\mathbb{X})} + (1 - w_\ell)\frac{\mu_S}{\mu_S(\mathbb{X})}, \quad \text{where } w_\ell = \frac{\mu_\ell(\mathbb{X})}{\mu_S(\mathbb{X}) + \mu_\ell(\mathbb{X})}.$$

It is apparent that the component of $\tilde{p}_\ell$ with the common atoms is $(1-w_\ell)\frac{\mu_S}{\mu_S(\mathbb{X})}$. Hence, the relative weights of the shared atoms for $\tilde{p}_1$ and $\tilde{p}_2$ differ only in the multiplicative factors $(1 - w_\ell)$, with $\ell = 1, 2$. This may be way too restrictive in some applications as nicely illustrated, e.g., in **Liu & Müller** through breast cancer gene expression data. Hence, unsurprisingly, it is essential to explore further extensions of LNPs in order to accommodate for different relative weights of the common atoms. Some of them have been already pointed out by the discussants. For example, **Leisen & Palacio** suggest the use of vectors (μ_S^1, μ_S^2) of Compound Random Measures (CoRMs) to replace the additive structure that makes the model analytically intractable when the number of samples d is larger than 2, as recalled in Section 1. The weights of μ_S^ℓ are different but dependent and one may still achieve clustering of samples by resorting to a suitable specification of the Lévy intensity for (μ_S^1, μ_S^2). Another solution is suggested by **Liu & Müller** whose proposal relies on a feature allocation model for identifying shared and idiosyncratic mixture components. Both a are definitely promising lines of investigation that is worth pursuing.

Obviously, starting from the hierarchical representation where

$$\tilde{p}_1, \tilde{p}_2 \overset{\text{iid}}{\sim} \tilde{q}', \qquad \tilde{q}' = \sum_{j \geq 1} \tilde{\pi}_j \delta_{\tilde{G}_j}$$

one may rely on appropriate specifications of the G_j's that yield different relative weights for the shared atoms. For example, one may proceed along the lines of the discussion by **Ma** and define the $\tilde{G}_j$'s as in Soriano and Ma (2019), i.e. all the $\tilde{G}_j$'s share the same atoms while their weights have shared and idiosyncratic components. This looks very much in the spirit of our paper. Another point touched by **Ma** is related to the limited flexibility of the dispersion parameter for LNPs based on CRMs. This is an interesting point, as the model typically has very few parameters to be tuned. A flexible model with an infinite-dimensional parameter which certainly merits further investigation could be based on nesting of Pólya trees (see also Christensen and Ma, 2019).

3 Correlation and related issues

E. Aliverti *et al.* and **Quintana & Jara** discuss some issues related to the dependence structure induced by LNPs. **Quintana & Jara** introduce the issue of possible extensions to accommodate for covariate-dependence, which we did not consider. Indeed, the most natural option would be a semiparametric covariate-dependent model where one may even introduce the dependence within the Lévy intensities of the underlying CRMs. While this is a nice and useful extension of LNPs that appears more appropriate for complex data structures, its implementation may follow from results in our paper.

Aliverti *et al.* unveil another interesting property of LNPs, which we actually overlooked. Indeed, as $\gamma \to +\infty$, they nicely show that all the random probability measures coincide almost surely and, hence, degeneracy to exchangeability may still arise though as a limiting case. This can also be deduced by the expression of the correlation in the two examples of interest (latent nested Dirichlet and stable processes). In order to do this we provide a general expression for the correlation $\text{Corr}(\tilde{p}_1(A), \tilde{p}_2(A))$ on the same set A.

Proposition 1. *If $(\tilde{p}_1, \tilde{p}_2) \sim \text{LNP}(\gamma, \nu_0, \nu)$, then for any set A belonging to the Borel σ-field on $\mathbb{X}$, the correlation between $\tilde{p}_1(A)$ and $\tilde{p}_2(A)$ is given by*

$$\text{Corr}(\tilde{p}_1(A), \tilde{p}_2(A)) = \pi_1^* + (1-\pi_1^*)\frac{\gamma \mathscr{I}_2(\gamma, \nu_0)}{(1+\gamma)\mathscr{I}_1(\gamma, \nu_0)}, \tag{1}$$

where

$$\mathscr{I}_1(\gamma, \nu_0) = \int_0^\infty u e^{-c_0(\gamma+1)\psi_0(u)} \tau_2^{(0)}(u) \mathrm{d}u,$$
$$\mathscr{I}_2(\gamma, \nu_0) = \int_0^\infty \int_0^\infty e^{-c_0\psi_0(u)-c_0\psi_0(v)-c_0\gamma\psi_0(u+v)} \tau_2^{(0)}(u+v) \mathrm{d}u\mathrm{d}v.$$

Proof. By definition we have that the covariance is

$$\text{Corr}(\tilde{p}_1(A), \tilde{p}_2(A)) = \frac{\mathbb{E}[\tilde{p}_1(A)\tilde{p}_2(A)] - \mathbb{E}[\tilde{p}_1(A)]\mathbb{E}[\tilde{p}_2(A)]}{\sqrt{\text{Var}(\tilde{p}_1(A))\text{Var}(\tilde{p}_2(A))}}. \tag{2}$$

We start by evaluating $\mathbb{E}[\tilde{p}_1(A)\tilde{p}_2(A)]$, using Proposition 3 and conditioning w.r.t. μ_S, we get

$$\mathbb{E}[\tilde{p}_1(A)\tilde{p}_2(A)] = \mathbb{E}\Big[\prod_{\ell=1}^{2} \frac{\mu_\ell(A) + \mu_S(A)}{\mu_\ell(\mathbb{X}) + \mu_S(\mathbb{X})}\Big]$$
$$= \pi_1^* \mathbb{E}_{\mu_S} \int_{\mathsf{M}_\mathbb{X}} \Big(\frac{m(A) + \mu_S(A)}{m(\mathbb{X}) + \mu_S(\mathbb{X})}\Big)^2 Q(\mathrm{d}m) + (1-\pi_1^*)\mathbb{E}_{\mu_S}\Big(\int_{\mathsf{M}_\mathbb{X}} \frac{m(A) + \mu_S(A)}{m(\mathbb{X}) + \mu_S(\mathbb{X})} Q(\mathrm{d}m)\Big)^2,$$

where $\mathbb{E}_{\mu_S}$ denotes the expected value with respect to the random variable μ_S, which is distributed as a CRM with Lévy intensity $\gamma\nu_0$. Denoting by $\tilde{\mu}_S^*, \tilde{\mu}_1^*$ and $\tilde{\mu}_2^*$ three

independent CRMs with respective Lévy intensities given by $\gamma\nu_0, \nu_0$ and ν_0 we get

$$\begin{aligned}&\mathbb{E}[\tilde{p}_1(A)\tilde{p}_2(A)]\\&\quad=\pi_1^*\mathbb{E}\left[\frac{\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\right]^2+(1-\pi_1^*)\mathbb{E}\left[\frac{\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\cdot\frac{\tilde{\mu}_2^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_2^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\right].\end{aligned}\tag{3}$$

We now evaluate the two expected value in (3), as for the first one we get

$$\begin{aligned}\mathbb{E}\left[\frac{\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\right]^2&=\int_0^\infty u\mathbb{E}[e^{-u(\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X}))}(\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A))^2]\mathrm{d}u\\&=\int_0^\infty u\mathbb{E}[e^{-u(\tilde{\mu}_1^*(A^c)+\tilde{\mu}_S^*(A^c))}]\frac{\mathrm{d}^2}{\mathrm{d}u^2}\mathbb{E}[e^{-u(\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A))}]\mathrm{d}u,\end{aligned}$$

by observing that $\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)$ has Laplace transform given by

$$\mathbb{E}[e^{-u(\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A))}]=\exp\left\{-c_0(\gamma+1)\psi_0(u)Q_0(A)\right\},$$

straightforward calculations show that

$$\begin{aligned}&\mathbb{E}\left[\frac{\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\right]^2\\&\quad=\int_0^\infty ue^{-c_0(\gamma+1)\psi_0(u)}[(c_0(\gamma+1)Q_0(A)\tau_1^{(0)}(u))^2+c_0Q_0(A)(\gamma+1)\tau_2^{(0)}(u)]\mathrm{d}u.\end{aligned}\tag{4}$$

If one considers the set $A=\mathbb{X}$, the previous equation boils down to the identity

$$c_0^2(\gamma+1)^2\int_0^\infty ue^{-c_0(\gamma+1)\psi_0(u)}(\tau_1^{(0)}(u))^2\mathrm{d}u=1-c_0(\gamma+1)\int_0^\infty ue^{-c_0(\gamma+1)\psi_0(u)}\tau_2^{(0)}(u)\mathrm{d}u$$

using this expression in (4) we finally obtain

$$\begin{aligned}&\mathbb{E}\left[\frac{\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\right]^2\\&\quad=Q_0(A)^2+Q_0(A)(1-Q_0(A))c_0(\gamma+1)\int_0^\infty ue^{-c_0(\gamma+1)\psi_0(u)}\tau_2^{(0)}(u)\mathrm{d}u.\end{aligned}\tag{5}$$

We now focus on the second expected value in (3)

$$\begin{aligned}&\mathbb{E}\left[\frac{\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\cdot\frac{\tilde{\mu}_2^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_2^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\right]\\&\quad=\int_0^\infty\int_0^\infty\mathbb{E}[e^{-u(\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X}))-v(\tilde{\mu}_2^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X}))}(\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A))(\tilde{\mu}_2^*(A)+\tilde{\mu}_S^*(A))]\mathrm{d}u\mathrm{d}v\end{aligned}$$

resorting to the first order derivatives of the Laplace functional of the three completely

random measures, some elementary calculations show that

$$\begin{aligned}
&\mathbb{E}\left[\frac{\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\cdot\frac{\tilde{\mu}_2^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_2^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\right]\\
&\quad=\int_0^\infty\int_0^\infty e^{-c_0\psi_0(u)-c_0\psi_0(v)-c_0\gamma\psi_0(u+v)}\\
&\qquad\times\Big\{c_0^2Q_0(A)^2[\tau_1^{(0)}(u)\tau_1^{(0)}(v)+\gamma\tau_1^{(0)}(u+v)(\tau_1^{(0)}(u)+\tau_1^{(0)}(v)+\gamma\tau_1^{(0)}(u+v))]\\
&\qquad\qquad+c_0Q_0(A)\gamma\tau_2^{(0)}(u+v)\Big\}\,\mathrm{d}u\mathrm{d}v.
\end{aligned}$$

Choosing $A=\mathbb{X}$ in the previous expression, one can obtain a suitable identity and get

$$\begin{aligned}
&\mathbb{E}\left[\frac{\tilde{\mu}_1^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_1^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\cdot\frac{\tilde{\mu}_2^*(A)+\tilde{\mu}_S^*(A)}{\tilde{\mu}_2^*(\mathbb{X})+\tilde{\mu}_S^*(\mathbb{X})}\right]=Q_0(A)^2\\
&\quad+Q_0(A)(1-Q_0(A))c_0\gamma\int_0^\infty\int_0^\infty e^{-c_0\psi_0(u)-c_0\psi_0(v)-c_0\gamma\psi_0(u+v)}\tau_2^{(0)}(u+v)\mathrm{d}u\mathrm{d}v.
\end{aligned}\tag{6}$$

Substituting the expressions (5)–(6) in (3) we obtain

$$\begin{aligned}
&\mathbb{E}[\tilde{p}_1(A)\tilde{p}_2(A)]=\\
&\quad Q_0(A)^2+Q_0(A)(1-Q_0(A))c_0\{\pi_1^*(\gamma+1)\mathscr{I}_1(\gamma,\nu_0)+(1-\pi_1^*)\gamma\mathscr{I}_2(\gamma,\nu_0)\}.
\end{aligned}\tag{7}$$

Along similar lines one may determine $\mathrm{Var}(\tilde{p}_1(A))$ showing that

$$\mathrm{Var}(\tilde{p}_1(A))=Q_0(A)(1-Q_0(A))c_0(\gamma+1)\mathscr{I}_1(\gamma,\nu_0),\tag{8}$$

and the same formula holds true for $\mathrm{Var}(\tilde{p}_2(A))$. Substituting expressions (7)–(8) in (2) and since $\mathbb{E}[\tilde{p}_1(A)]=\mathbb{E}[\tilde{p}_2(A)]=Q_0(A)$, the conclusion follows. □

As for many other proposals of dependent priors, (1) does not depend on the particular set A and it is, thus, interpreted as a measure of the overall dependence across the two random probability measures. One can now determine explicit expressions of the correlation in some examples of interest and we show that it goes to 1 as $\gamma\to+\infty$: this is in line with Lemma 1 of **Aliverti *et al.***.

Example 1. Consider the latent nested stable process introduced in Example 1 of the paper. One can evaluate the integrals $\mathscr{I}_1(\gamma,\nu_0)$ and $\mathscr{I}_2(\gamma,\nu_0)$ to obtain

$$\mathscr{I}_1(\gamma,\nu_0)=\frac{(1-\sigma_0)_1}{c_0(\gamma+1)},$$

by a change of variable and some algebra we also get that

$$\mathscr{I}_2(\gamma,\nu_0)=\frac{(1-\sigma_0)_1}{c_0}\int_0^1\frac{1}{\gamma+w^{\sigma_0}+(1-w)^{\sigma_0}}\mathrm{d}w.$$

Using the previous expression in (1) and recalling that $\pi_1^*=1-\sigma$ we get

$$\mathrm{Corr}(\tilde{p}_1(A),\tilde{p}_2(A))=(1-\sigma)+\sigma\gamma\int_0^1\frac{1}{\gamma+w^{\sigma_0}+(1-w)^{\sigma_0}}\mathrm{d}w.\tag{9}$$

We point out that as $\gamma\to+\infty$ the correlation goes to 1. □

Example 2. Let us consider the latent nested Dirichlet process of Example 2 in the paper, which corresponds to the choice $\rho_0(s) = \rho(s) = e^{-s}/s$. Recalling that $\tau_q^{(0)}(u) = \Gamma(q)/(u+1)^q$ and $\psi_0(u) = \log(1+u)$ we can easily evaluate the integral $\mathscr{I}_1(\gamma, \nu_0)$, in particular we obtain

$$\mathscr{I}_1(\gamma, \nu_0) = \frac{1}{c_0(\gamma+1)(c_0(\gamma+1)+1)}.$$

As for the second integral in the expression of the correlation (1), one obtains

$$\mathscr{I}_2(\gamma, \nu_0) = \int_0^\infty \int_0^\infty \frac{1}{(1+u)^{c_0}(1+v)^{c_0}(1+u+v)^{c_0\gamma+2}} \mathrm{d}u\mathrm{d}v,$$

where we have exploited identity 3.197.1 in Gradshteyn and Ryzhik (2007) we get

$$\mathscr{I}_2(\gamma, \nu_0) = B(1, c_0\gamma+2) \int_0^\infty (1+u)^{-c_0(\gamma+2)-1} {}_2F_1(c_0, c_0(\gamma+1)+1; c_0(\gamma+1)+2; -u)\mathrm{d}u$$

a change of variable $t = u/(u+1)$ and identity 7.512.5 of Gradshteyn and Ryzhik (2007) imply the following expression for the integral under consideration

$$\mathscr{I}_2(\gamma, \nu_0) = \frac{{}_3F_2(c_0, c_0(\gamma+1)+1, 1; c_0(\gamma+1)+2, c_0(\gamma+2)+1; 1)}{(c_0\gamma+2)c_0(\gamma+2)}.$$

Thanks to the previous considerations and remembering that $\pi_1^* = 1/(1+c)$, the correlation becomes

$$\begin{aligned}\mathrm{Corr}(\tilde{p}_1(A), \tilde{p}_2(A)) = \frac{1}{1+c} + \frac{c}{1+c}\frac{\gamma(c_0(\gamma+1)+1)}{(c_0\gamma+2)(\gamma+2)} \\ \times {}_3F_2(c_0, c_0(\gamma+1)+1, 1; c_0(\gamma+1)+2, c_0(\gamma+2)+1; 1).\end{aligned} \tag{10}$$

In order to study the behavior of the correlation as $\gamma \to +\infty$, we have to determine the limit as $\gamma \to +\infty$ of the hypergeometric function, to this end note that

$${}_3F_2(c_0, \lambda, 1; \lambda+1, c_0+\lambda; 1) = \frac{1}{\Gamma(c_0)} \sum_{k\geq 0} \frac{\lambda}{\lambda+k} \frac{\Gamma(\lambda+c_0)}{\Gamma(\lambda+c_0+k)} \Gamma(c_0+k),$$

where we have put $\lambda = \lambda(\gamma) = c_0(\gamma+1)+1 \to +\infty$. We observe that for any $\lambda \geq 2$

$$\begin{aligned}\frac{\lambda}{\lambda+k}\frac{\Gamma(\lambda+c_0)}{\Gamma(\lambda+c_0+k)}\Gamma(c_0+k) &\leq \frac{\Gamma(\lambda+c_0)}{\Gamma(\lambda+c_0+k)}\Gamma(c_0+k) \\ &\leq \frac{\Gamma(2+c_0)}{\Gamma(2+c_0+k)}\Gamma(c_0+k) = \frac{c_0(c_0+1)}{(c_0+k)(c_0+k+1)},\end{aligned}$$

where we have exploited the fact that the ratio of gamma functions is decreasing in λ. Thanks to the previous inequality we can conclude that the general term of the summation over k is uniformly bounded by the term of a convergent series in k, therefore the dominated convergence theorem may be applied and then

$$\lim_{\gamma\to\infty} {}_3F_2(c_0, \lambda, 1; \lambda+1, c_0+\lambda; 1) = \frac{1}{\Gamma(c_0)} \sum_{k\geq 0} \lim_{\gamma\to\infty} \frac{\lambda}{\lambda+k} \frac{\Gamma(\lambda+c_0)}{\Gamma(\lambda+c_0+k)} \Gamma(c_0+k)$$

$$= \frac{1}{\Gamma(c_0)} \sum_{k \geq 0} \lim_{\gamma \to \infty} \lambda^{-k} \Gamma(c_0 + k) = 1.$$

This implies again that the correlation in (10) goes to 1 as γ tends to infinity. □

Due to our interest in testing homogeneity across populations, another important aspect that should be investigated is related to consistency issues of the Bayes Factor, as underlined by **Robert**. This should also be complemented by the study of consistency for the random dependent densities in a partially exchangeable setting. Both problems will be object of future investigation.

We conclude expressing again our gratitude to all discussants for their enlightening contributions, which have helped outlining several relevant aspects about LNPs and the general problem of pursuing clustering of distributions and clustering of observations across multiple samples by means of a nested structure and beyond.

References

Canale, A., Corradin, R., and Nipoti, B. (2019). "Importance conditional sampling for Bayesian nonparametric mixtures." *arXiv preprint arXiv:1906.08147*. 1350

Christensen, J. and Ma, L. (2019). "A Bayesian hierarchical model for related densities using Pólya trees." *Journal of the Royal Statistical Society, Series B*. To appear. 1351

Gradshteyn, I. S. and Ryzhik, I. M. (2007). *Tables of integrals, sums, series, and products*. Academic Press, 7th edition. MR0669666. 1355

Soriano, J. and Ma, L. (2019). "Mixture modeling on related samples by ψ-stick breaking and kernel perturbation." *Bayesian Anal.*, 14(1): 161–180. MR3910042. doi: https://doi.org/10.1214/18-BA1106. 1351