Bayesian Analysis (2019) **TBA**, Number TBA, pp. 1–23

Latent Nested Nonparametric Priors

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

[*]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 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 $\tilde{N}(A) \sim \mathrm{Po}(\nu(A))$. 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 $\tilde{\mu} \sim \mathrm{CRM}[\nu; \mathsf{P}_{\mathbb{X}}]$. 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 $\tilde{q} \sim \mathrm{NRMI}[\nu; \mathsf{P}_{\mathbb{X}}]$. The baseline measure, Q, of $\tilde{\mu}$ in (3) is, in turn, the probability distribution of $\tilde{q}_0 \sim \mathrm{NRMI}[\nu_0; \mathbb{X}]$, 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

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

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.

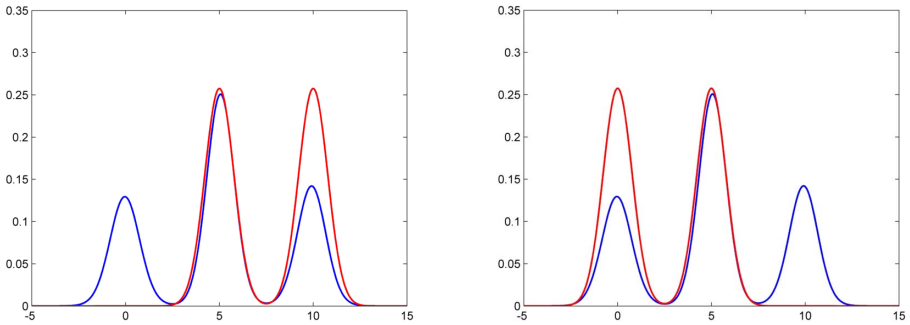
3 Latent nested processes

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 $\mathbb{M}_{\mathbb{X}}$ 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

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

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

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

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.

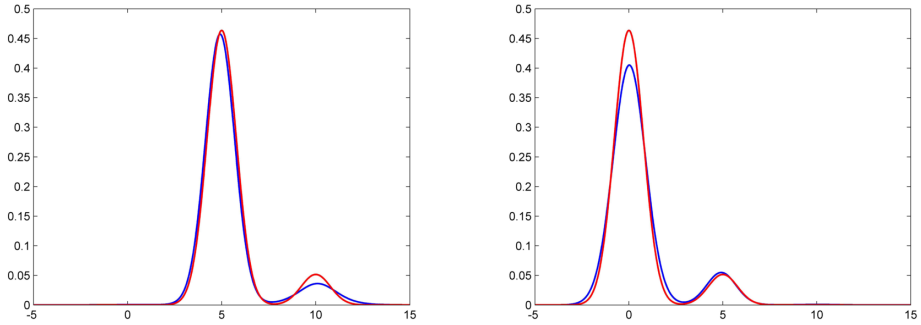
Scenario II corresponds to samples $\boldsymbol{X}_1^{(n_1)}$ and $\boldsymbol{X}_2^{(n_2)}$ generated, respectively, from

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

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.

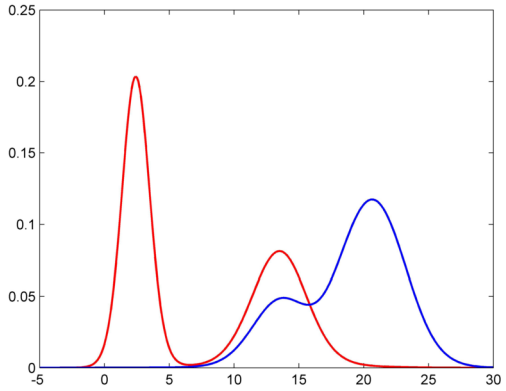


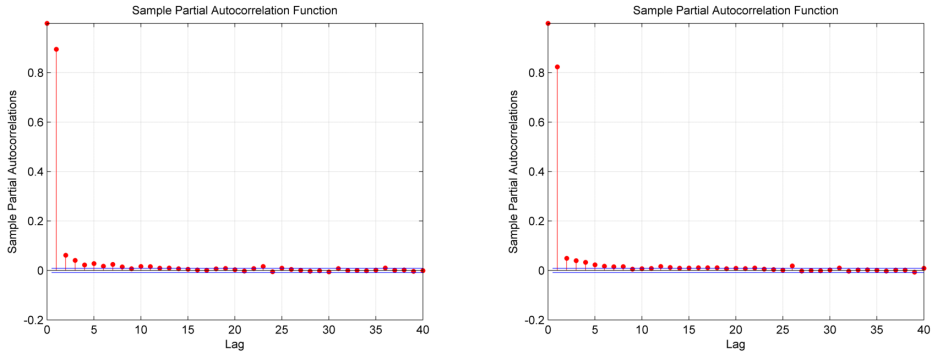
Figure 3: Iris dataset: the estimated densities for the first sample $\mathbf{X}$ (observations of setosa and versicolor) are shown in red, while the estimated densities for the second sample $\mathbf{Y}$ (observations of versicolor and virginica) are shown in blue.

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.

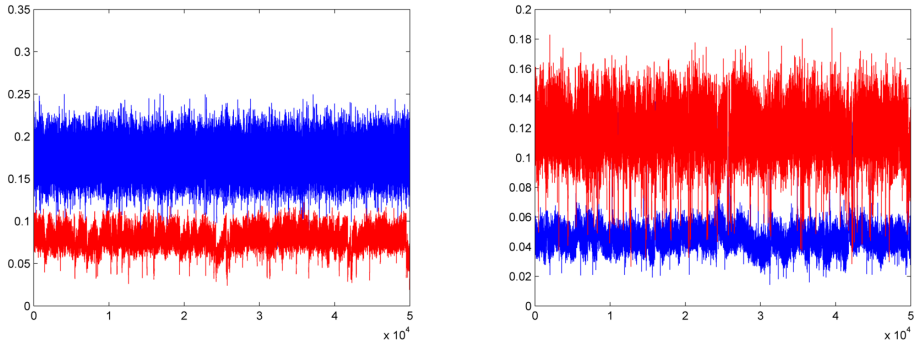
6 Concluding remarks

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. 3

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. 2

Blei, D. M. and Frazier, P. I. (2011). "Distance dependent Chinese restaurant process." *Journal of Machine Learning Research*, 12: 2383–2410. MR2834504. 2

Blei, D. M., NG, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3: 993–1022. 2

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. 2

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. 11

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. 3

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. 2

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. 2

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. 2

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. 2

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. 2

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. 3

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. 3

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. 2

Hjort, N. L. (2000). "Bayesian analysis for a generalized Dirichlet process prior." Technical report, University of Oslo. 2

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. 2

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. 3

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. 6

Kingman, J. F. C. (1993). *Poisson processes*. Oxford University Press. MR1207584. 4

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. 2, 3

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. 2

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. 15

MacEachern, S. N. (1999). "Dependent nonparametric processes." In *ASA proceedings of the section on Bayesian statistical science*, 50–55. 2

MacEachern, S. N. (2000). "Dependent Dirichlet processes." *Tech. Report, Department of Statistics, The Ohio State University*. 2

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. 3

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. 2

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. 2

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. 3

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. 3

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. 2

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. 6

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. 4

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. 3

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. 2

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. 2, 3, 4, 19

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. 3

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. 2

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. 2

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. 15

Acknowledgments

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