Electronic Journal of Statistics
Vol. 10 (2016) 3516–3547
ISSN: 1935-7524
DOI: 10.1214/16-EJS1168

Posterior sampling from ε-approximation of normalized completely random measure mixtures

Raffaele Argiento

CNR-IMATI, Via Bassini 15, 20133 Milano, Italy
e-mail: raffaele@mi.imati.cnr.it

Ilaria Bianchini and Alessandra Guglielmi

Department of Mathematics, Politecnico di Milano,
Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: ilaria.bianchini@polimi.it; alessandra.guglielmi@polimi.it

Abstract: This paper adopts a Bayesian nonparametric mixture model where the mixing distribution belongs to the wide class of normalized homogeneous completely random measures. We propose a truncation method for the mixing distribution by discarding the weights of the unnormalized measure smaller than a threshold. We prove convergence in law of our approximation, provide some theoretical properties, and characterize its posterior distribution so that a blocked Gibbs sampler is devised.

The versatility of the approximation is illustrated by two different applications. In the first the normalized Bessel random measure, encompassing the Dirichlet process, is introduced; goodness of fit indexes show its good performances as mixing measure for density estimation. The second describes how to incorporate covariates in the support of the normalized measure, leading to a linear dependent model for regression and clustering.

Keywords and phrases: Bayesian nonparametric mixture models, normalized completely random measures, blocked Gibbs sampler, finite dimensional approximation.

Received September 2015.

Contents

1. Introduction

One of the livelier topic in Bayesian Nonparametrics concerns mixtures of parametric densities where the mixing measure is an almost surely discrete random probability measure. The basic model is what is known as the Dirichlet process mixture model, appeared first in [33], where the mixing measure is indeed the Dirichlet process. Dating back to [25] and [32], many alternative mixing measures have been proposed; the former paper replaced the Dirichlet process with stick-breaking random probability measures, while the latter focused on normalized completely random measures. These hierarchical mixtures play a pivotal role in modern Bayesian Nonparametrics, and their popularity is mainly due to the high flexibility in density estimation problems as well as in clustering, which is naturally embedded in the model.

In some statistical applications, the clustering induced by the Dirichlet process as mixing measure may be restrictive. In fact, it is well-know that the latter allocates observations to clusters with probabilities depending only on the cluster sizes, leading to the "the rich gets richer" behavior. Within some classes of more general processes, as, for instance, stick-breaking and normalized processes, the probability of allocating an observation to a specific cluster depends also on extra parameters, as well as on the number of groups and on the cluster size. We refer to [4] for a recent review of the state of the art on Bayesian nonparametric mixture models and clustering.

Since posterior inference for Bayesian nonparametric mixtures involves an infinite-dimensional parameter, this may lead to computational issues. However, there is a recent and lively literature focusing mainly on two different classes of MCMC algorithms, namely *marginal* and *conditional* Gibbs samplers. The former integrates out the infinite dimensional parameter (i.e. the random probability), resorting to generalized Polya urn schemes; see [17] or [34]. The latter includes the nonparametric mixing measure in the state space of the Gibbs sampler, updating it as a component of the algorithm; this class includes the slice sampler [see 23]. Among conditional algorithms there are truncation methods, where the infinite parameter (i.e. the mixing measure) is approximated by truncating the infinite sums defining the process, either a posteriori [5, 9] or a priori [3, 24].

In this work we introduce an almost surely finite dimensional class of random probability measures that approximates the wide family of homogeneous normalized completely random measures [41, 29]; we use this class as the building

block in mixture models and provide a simple but general truncation algorithm to perform posterior inference. Our approximation is based on the constructive definition of the weights of the completely random measure as the points of a Poisson process on $\mathbb{R}^+$. In particular, we consider only points larger than a threshold ε, controlling the degree of approximation. Conditionally on ε, our process is finite dimensional both a priori and a posteriori.

Here we illustrate two applications. In the first one, a new choice for the Lévy intensity ρ, characterizing the normalized completely random measure, is proposed: the Bessel intensity function that, up to our knowledge, has never been applied in a statistical framework, but known in finance [see 7, for instance]. We call this new process normalized Bessel random measure. In the second application, we set ρ to be the well-known generalized gamma intensity and consider a centering measure $P_{0\mathbf{x}}$ depending on a set of covariates $\mathbf{x}$, yielding a linear dependent normalized completely random measure. For a recent survey on dependent nonparametric processes in the Statistics and Machine Learning literature see [20].

In this paper, since the main objective is the approximation of the nonparametric process arising from the normalization of completely random measures, we fix ε to a small value. However, it is worth mentioning that it is possible to choose a prior for ε, but the computational cost might greatly increase for some intensity ρ.

The new achievements of this paper can be summarized as follows: (i) a generalization of the ε-approximation given in [3] for the NGG process to the whole family of normalized homogeneous completely random measures, (ii) a different technique providing the posterior distribution (and the exchangeable partition probability function) of this new random probability measure, making use of Palm's formula, and (iii) the introduction of the normalized Bessel random measure as mixing measure in Bayesian nonparametric mixtures.

In particular, after the introduction of the finite dimensional ε-approximation of a normalized completely random measure, we derive its posterior and show that the ε-approximation converges to its infinite dimensional counterpart (Section 3). Then we provide a Gibbs sampler for the ε-approximation hierarchical mixture model (Section 4). Section 5 illustrates some criteria to choose the approximation parameter ε. Section 6.1 is devoted to the introduction of the normalized Bessel random measure, and some of its properties; on the other hand, Section 6.2 discusses an application of the ε-Bessel mixture models to both simulated and real data. Section 7 defines the linear dependent $\varepsilon - NGG$'s, and considers linear dependent $\varepsilon - NGG$ mixtures to fit the AIS data set. To complete the set-up of the paper, Section 2 is devoted to a summary of basic notions about homogeneous normalized completely random measures, and Section 8 contains a conclusive discussion.

2. Preliminaries on normalized completely random measures

Let us briefly recall the definition of a homogeneous normalized completely random measure. Let $\Theta \subset \mathbb{R}^m$ for some positive integer m. A random measure

μ on Θ is completely random if for any finite sequence $B_1, B_2, \ldots, B_k$ of disjoint sets in $\mathcal{B}(\Theta)$, $\mu(B_1), \mu(B_2), \ldots, \mu(B_k)$ are independent. A purely atomic (with no fixed points) completely random measure (c.r.m.) is defined [see 30, Section 8.2] by

$$\mu(\cdot) = \int_{\mathbb{R}^+ \times \cdot} sN(ds, d\tau), \tag{2.1}$$

where N is a Poisson process on $\mathbb{R}^+ \times \Theta$ with mean intensity $\nu(ds, d\tau)$. A completely random measure is homogeneous if $\nu(ds, d\tau) = \rho(s)ds\kappa P_0(d\tau)$, where $\rho(s)$ is the density of a non-negative measure on $\mathbb{R}^+$, and κP_0 is a finite measure on Θ with total mass $\kappa > 0$. If

$$\int_0^{+\infty} \min\{1, s\}\rho(s)ds < +\infty, \tag{2.2}$$

then μ is characterized by the so-called Lévy-Khintchine representation: for any non-negative function f, the Laplace functional Ψ_μ of μ is given by

$$\Psi_\mu[f] := \mathbb{E}\left\{\mathrm{e}^{-\int_\Theta f(\tau)\mu(d\tau)}\right\} = \exp\left\{-\int_{\mathbb{R}^+ \times \Theta} (1 - \mathrm{e}^{-sf(\tau)})\nu(ds, d\tau)\right\}; \tag{2.3}$$

in this case $\nu(ds, d\tau)$ is called Lévy intensity measure. Furthermore, we assume that ρ satisfies the following regularity condition:

$$\int_0^{+\infty} \rho(s)ds = +\infty, \tag{2.4}$$

so that the total number of points of the process, $N(\mathbb{R}^+ \times \Theta)$, is Poisson distributed with mean $\int_{\mathbb{R}^+ \times \Theta} \nu(ds, d\tau) = \kappa \int_{\mathbb{R}^+} \rho(s)ds = +\infty$. This implies that any homogeneous completely random measure under (2.2) and (2.4) can be represented as $\mu(\cdot) = \sum_{j \geq 1} J_j \delta_{\tau_j}(\cdot)$. Since μ is homogeneous, the support points $\{\tau_j\}$ and the jumps $\{J_j\}$ of μ are independent, and the τ_j's are independent identically distributed (iid) random variables from P_0, while $\{J_j\}$ are the points of a Poisson process on $\mathbb{R}^+$ with mean intensity ρ. Moreover, if $T := \mu(\Theta) = \sum_{j \geq 1} J_i$, by (2.2) and (2.4), we have $\mathbb{P}(0 < T < +\infty) = 1$.

Therefore, a random probability measure (r.p.m.) P can be defined through normalization of μ:

$$P := \frac{\mu}{\mu(\Theta)} = \sum_{j=1}^{+\infty} \frac{J_j}{T}\ \delta_{\tau_j} = \sum_{j=1}^{+\infty} P_j \delta_{\tau_j}\ . \tag{2.5}$$

We refer to P in (2.5) as a (homogeneous) normalized completely random measure with parameter $(\rho, \kappa P_0)$. As an alternative notation, following [27], P is referred to as a homogeneous normalized measure with independent increments. The definition of normalized completely random measures appeared in [41] first. An alternative construction of normalized completely random measures can be given in terms of Poisson-Kingman models as in [39].

3. ε-approximation of normalized completely random measures

The goal of this section is the definition of a finite dimensional random probability measure that is an approximation of a general normalized completely random measure with Lévy intensity given by $\nu(ds, d\tau) = \rho(ds)\kappa P_0(d\tau)$, introduced above.

First of all, by the Restriction Theorem for Poisson processes, for any $\varepsilon > 0$, all the jumps $\{J_j\}$ of μ larger than a threshold ε are still a Poisson process, with mean intensity $\gamma_\varepsilon(s) := \kappa\rho(s)\mathbb{I}_{(\varepsilon,+\infty)}(s)$. Moreover, the total number of these points is Poisson distributed, i.e. $N_\varepsilon \sim \mathcal{P}_0(\Lambda_\varepsilon)$ where $\Lambda_\varepsilon := \kappa\int_\varepsilon^{+\infty}\rho(s)ds$. Since $\Lambda_\varepsilon < +\infty$ for any $\varepsilon > 0$ by (2.2), N_ε is almost surely finite. In addition, conditionally to N_ε, the points $\{J_1, \ldots, J_{N_\varepsilon}\}$ are iid from the density

$$\rho_\varepsilon(s) = \frac{\gamma_\varepsilon(s)}{\Lambda_\varepsilon} = \frac{\kappa\rho(s)}{\Lambda_\varepsilon}\mathbb{I}_{(\varepsilon,+\infty)}(s), \tag{3.1}$$

thanks to the relationship between Poisson and Bernoulli processes; see, for instance, [30], Section 2.4.

We denote by $\tilde{\mu}_\varepsilon$ the c.r.m. with Lévy intensity

$$\nu_\varepsilon(ds, d\tau) := \rho(ds)\mathbb{I}_{(\varepsilon,+\infty)}(s)ds\kappa P_0(d\tau). \tag{3.2}$$

This implies that $\tilde{\mu}_\varepsilon = \sum_{j=1}^{N_\varepsilon} J_j\delta_{\tau_j}$. However, it is not worth trying to normalize $\tilde{\mu}_\varepsilon$, since $\tilde{\mu}_\varepsilon(B) = 0$ for any B if $N_\varepsilon = 0$. We consider, instead, the c.r.m. μ_ε so defined:

$$\mu_\varepsilon(\cdot) \stackrel{d}{=} J_0\delta_{\tau_0}(\cdot) + \tilde{\mu}_\varepsilon(\cdot) \tag{3.3}$$

where (J_0, τ_0) is independent from $\{(J_j, \tau_j), j \geq 1\}$, J_0 and τ_0 are independent with density ρ_ε and P_0, respectively. Thus

$$\mu_\varepsilon(\cdot) = J_0\delta_{\tau_0}(\cdot) + \sum_{j=1}^{N_\varepsilon} J_j\delta_{\tau_j}(\cdot) = \sum_{j=0}^{N_\varepsilon} J_j\delta_{\tau_j}(\cdot).$$

Summing up, we define:

$$P_\varepsilon(\cdot) = \sum_{j=0}^{N_\varepsilon} P_j\delta_{\tau_j}(\cdot) = \sum_{j=0}^{N_\varepsilon} \frac{J_j}{T_\varepsilon}\delta_{\tau_j}(\cdot), \tag{3.4}$$

where $T_\varepsilon = \sum_{j=0}^{N_\varepsilon} J_j$, $\tau_j \overset{\text{iid}}{\sim} P_0$, $\{\tau_j\}$ and $\{J_j\}$ independent. We denote P_ε in (3.4) by $\varepsilon - NormCRM$ and write $P_\varepsilon \sim \varepsilon - NormCRM(\rho, \kappa P_0)$. When $\rho_\varepsilon(s) = 1/(\omega^\sigma\Gamma(-\sigma, \omega\varepsilon))s^{-\sigma-1}\mathrm{e}^{-\omega s}$, $s > \varepsilon$, P_ε is the $\varepsilon - NGG$ process introduced in [3], with parameter (σ, κ, P_0), $0 \leq \sigma \leq 1$, $\kappa \geq 0$.

Increasing Lévy processes are completely random measures for $\Theta = \mathbb{R}$ (or $\mathbb{R}^+$). Therefore, it is worth mentioning some literature on ε-approximation of such processes in the financial context. In particular, the book by Asmussen and Glynn [6, Chapter XII] provides a justification for the approximation of infinite

activity Lévy processes by compound Poisson processes: any Lévy jump process J on $\mathbb{R}$ can be represented as the sum of two independent Lévy processes

$$J(s) = J_1(s) + J_2(s), \ s \in \mathbb{R},$$

where the Lévy measures of J_1 and J_2 are restrictions of the "whole" Lévy measure on $(-\varepsilon, \varepsilon)$ and $(-\infty, -\varepsilon] \cup [\varepsilon, +\infty)$, respectively. When considering the homogeneous completely random measure μ under (2.2) and (2.4) as here, this theory yields that μ is the sum of two independent homogeneous completely random measures $\mu^{(0,\varepsilon]}$ and $\tilde{\mu}_\varepsilon$, corresponding to mean intensities $\rho(s)\mathbb{I}_{(0,\varepsilon]}(s)$ and ρ_ε as in (3.1), respectively. Note that $\tilde{\mu}_\varepsilon$ is the c.r.m. in the right hand-side of (3.3). The basic idea of the ε-approximation is that, if ε is small enough, $\mu^{(0,\varepsilon]}$ can be neglected and μ can be approximated by $\tilde{\mu}_\varepsilon$; see [6, Chapter XII] and [43].

The approach to ε-approximation taken here is similar, though not identical, since we first add the random mass J_0 in the random point τ_0 to $\tilde{\mu}_\varepsilon$ to define the c.r.m. μ_ε as in (3.3). The r.p.m. P_ε in (3.4) is then defined by normalization of μ_ε. We will show in Proposition 3.3 that P_ε converges in distribution to P as ε goes to 0, but the basic idea of the approximation is that the point mass we add to $\tilde{\mu}_\varepsilon$ is negligible; see Section 5.

Several other methods have been proposed in order to approximate a normalized measure; first of all, we mention the inverse Lévy measure method, referred to as Ferguson-Klass representation [19] in this context, representing the Poisson process of the jumps of a subordinator as a series of trasformed (via the survival function of the Lévy intensity) points of a unit rate Poisson process. Of course, to get implementable simulation algorithms, the series expansion has to be truncated at a fixed and large integer N, or whenever the new jump to be added to the series is smaller that a threshold ε. In the latter case, the truncation rule would yield only jumps of size greater than ε, obtaining an algorithm that is similar to that proposed here; see [6, Chapter XII]. On the other hand, [2] proposes a truncation rule of the series representation at a fixed integer N quantifying the error through a moment-matching criterion, i.e. evaluating a measure of discrepancy between actual moments of the whole series and moments of the truncated sum based on the simulation output. More series representations of the jump process can be considered, with corresponding truncation rules; see [11] and [42]. Alternatively, [43] proposed a novel class of r.p.m.'s, that is dense in the class of homogeneous normalized completely random measures. These authors first approximate any c.r.m. μ with $\tilde{\mu}_\varepsilon$ which, as we have already mentioned, has finite Lévy measure. Then, resorting to the "denseness" of the novel class, they approximate $\tilde{\mu}_\varepsilon$ with an element of this class, with Lévy intensity given by the weighted sum of a finite number of intensities of finite activity processes, plus the intensity of the gamma process.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be a sample from P_ε, a $\varepsilon - NormCRM(\rho, \kappa P_0)$ as defined in (3.4), and let $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$ be the (observed) distinct values in $\boldsymbol{\theta}$. We denote by *allocated* jumps of the process the values $P_{l_1^*}, P_{l_2^*}, \ldots, P_{l_k^*}$ in (3.4) such that there exists a corresponding location for which $\tau_{l_i^*} = \theta_i^*$, $i = 1, \ldots, k$.

The remaining values are *non-allocated* jumps. We use the superscript (na) for random variables related to *non-allocated* jumps. The first result is a characterization of the posterior law of the random measure μ_ε, not yet normalized; however, we need introducing two more ingredients first. We consider an auxiliary random variable U such that $U|\mu_\varepsilon \sim Gamma(n, T_\varepsilon)$, so that the marginal density of U is

$$\begin{aligned} f_U(u;n) = \frac{u^{n-1}}{\Gamma(n)}\mathbb{E}(T_\varepsilon^n \mathrm{e}^{-T_\varepsilon u}) &= \frac{u^{n-1}}{\Gamma(n)}(-1)^n \frac{d}{du^n}\mathbb{E}(\mathrm{e}^{-uT_\varepsilon}) \\ &= \frac{u^{n-1}}{\Gamma(n)}(-1)^n \frac{d}{du^n}\frac{\Lambda_{\varepsilon,u}\mathrm{e}^{\Lambda_{\varepsilon,u}}}{\Lambda_\varepsilon \mathrm{e}^{\Lambda_\varepsilon}}, \end{aligned} \tag{3.5}$$

and the last equality follows easily from the definition of T_ε and (2.3), using notation defined in (3.8). We also formulate the following lemma, whose proof is straightforward.

Lemma 3.1. *Let $\tilde{\mu}_\varepsilon$ be a finite c.r.m. with Lévy intensity ν_ε as in (3.2), and let μ_ε be defined as in (3.3). Consider a c.r.m. $\mu^\star$ such that*

$$\mu^\star(\cdot) \stackrel{d}{=} X\mu_\varepsilon(\cdot) + (1-X)\tilde{\mu}_\varepsilon(\cdot), \tag{3.6}$$

where $X \sim Bernoulli(p)$, $p = a/(a+b)$, $a, b > 0$, and X is independent on $\tilde{\mu}_\varepsilon$ and (J_0, τ_0). The Laplace functional of $\mu^\star$ is:

$$\Psi[f] = \frac{aA[f]+b}{a+b}\exp\left\{-\int_{\mathbb{R}^+\times\Theta}\left(1-\mathrm{e}^{-f(\tau)s}\right)\nu_\varepsilon(ds,d\tau)\right\}, \tag{3.7}$$

for any positive f, where

$$\begin{aligned} A[f] := \mathbb{E}\left(\mathrm{e}^{-f(\tau_0)J_0}\right) &= \int_{\mathbb{R}^+\times\Theta}\mathrm{e}^{-f(\tau)s}\rho_\varepsilon(s)dsP_0(d\tau) \\ &= \frac{1}{\Lambda_\varepsilon}\int_{\mathbb{R}^+\times\Theta}\mathrm{e}^{-sf(\tau)}\nu_\varepsilon(ds,d\tau) \end{aligned}$$

is the Laplace functional of the random measure $J_0\delta_{\tau_0}$.

The posterior distribution of μ_ε has the following characterization.

Theorem 3.1. *If P_ε is an $\varepsilon-NormCRM(\rho, \kappa P_0)$, then the conditional distribution of P_ε, given $\boldsymbol{\theta}^*$ and $U = u$, is obtained by normalization of the following random measure*

$$\mu_\varepsilon^*(\cdot) \stackrel{d}{=} \mu_{\varepsilon,u}^{(na)}(\cdot) + \mu_{\varepsilon,u}^{(a)}(\cdot) = \mu_{\varepsilon,u}^{(na)}(\cdot) + \sum_{j=1}^{k} J_j^{(a)}\delta_{\theta_k^*}(\cdot)$$

where

1. *the law of the process of non-allocated jumps* $\mu_{\varepsilon,u}^{(na)}(\cdot)$ *is distributed as the c.r.m.* $\mu^\star$ *defined in* (3.6)*, corresponding to Lévy intensity in* (3.7) *given by* $\mathrm{e}^{-us}\nu_\varepsilon(ds, d\tau)$ *and probability* p *of success* $p = \Lambda_{\varepsilon,u}/(\Lambda_{\varepsilon,u} + k)$*, where*

$$\Lambda_{\varepsilon,u} := \kappa \int_\varepsilon^{+\infty} \mathrm{e}^{-us}\rho(s)ds, \quad u \geq 0; \tag{3.8}$$

2. *the process of allocated jumps* $\mu_{\varepsilon,u}^{(a)}(\cdot)$ *has fixed points of discontinuity* $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$ *with weights* $J_j^{(a)} \overset{ind}{\sim} s^{n_i}\mathrm{e}^{-us}\rho(s)\mathbb{I}_{(\varepsilon,+\infty)}(s)ds$*,* $j = 1, \ldots, k$*;*
3. $\mu_{\varepsilon,u}^{(na)}(\cdot)$ *and* $\mu_{\varepsilon,u}^{(a)}(\cdot)$ *are independent, conditionally to* $\boldsymbol{l}^* = (l_1^*, \ldots, l_k^*)$*, the vector of locations of the* allocated *jumps;*
4. *the posterior law of* U *given* $\boldsymbol{\theta}^*$ *has density on the positive real numbers given by*

$$f_{U|\boldsymbol{\theta}^*}(u|\boldsymbol{\theta}^*) \propto u^{n-1}\mathrm{e}^{\Lambda_{\varepsilon,u}-\Lambda_\varepsilon}\frac{\Lambda_{\varepsilon,u}+k}{\Lambda_\varepsilon}\prod_{i=1}^k\int_\varepsilon^{+\infty}\kappa s^{n_i}\mathrm{e}^{-us}\rho(s)ds, \quad u > 0.$$

The proof of the above proposition, as well as of all the others in this section, is in Appendix B. An immediate consequence of Theorem 3.1 is the next proposition.

Corollary 3.1. *The conditional distribution of* P_ε*, given* $\boldsymbol{\theta}^*$ *and* $U = u$*, verifies the distributional equation*

$$P_\varepsilon^*(\cdot) \overset{d}{=} wP_{\varepsilon,u}^{(na)}(\cdot) + (1-w)\sum_{j=1}^k P_j^{(a)}\delta_{\theta_k^*}(\cdot)$$

where $P_{\varepsilon,u}^{(na)}(\cdot)$ *is the null measure if* $\mu_{\varepsilon,u}^{(na)}(\Theta) = 0$*,* $w = \mu_{\varepsilon,u}^{(na)}(\Theta)/(\mu_{\varepsilon,u}^{(na)}(\Theta) + \sum_{j=1}^k J_j^{(a)})$*, and the jumps* $\{P_1^{(a)}, \ldots, P_k^{(a)}\}$ associated *to the fixed points of discontinuity* $\theta_1^*, \ldots, \theta_k^*$ *are defined as* $P_j^{(a)} = J_j^{(a)}/\sum_{j=1}^k J_j^{(a)}$*,* $j = 1 \ldots, k$*.*

Theorem 3.1 and Corollary 3.1 conceive the "finite dimensional" counterpart of Proposition 1 in [27].

Both the infinite and finite dimensional processes defined in (2.5) and (3.4), respectively, belong to the wide class of species sampling models, deeply investigated in [38], and we use some of the results there to derive ours. Let $(\theta_1, \ldots, \theta_n)$ be a sample from (2.5) or (3.4) (or, more generally, from a species sampling model); since it is a sample from a discrete probability, it induces a random partition $\boldsymbol{p}_n := \{C_1, \ldots, C_k\}$ on the set $\mathbb{N}_n := \{1, \ldots, n\}$ where $C_j = \{i : \theta_i = \theta_j^*\}$ for $j = 1, \ldots, k$. If $\#C_i = n_i$ for $1 \leq i \leq k$, the marginal law of $(\theta_1, \ldots, \theta_n)$ has unique characterization:

$$\mathcal{L}(\boldsymbol{p}_n, \theta_1^*, \ldots, \theta_k^*) = p(n_1, \ldots, n_k)\prod_{j=1}^k \mathcal{L}(\theta_j^*),$$

where p is the exchangeable partition probability function (eppf) associated to the random probability. The eppf p is a probability law on the set of the partitions of $\mathbb{N}_n$. The following proposition provides an expression for the eppf of a general $\varepsilon - NormCRM$.

Proposition 3.1. *Let $(n_1, \dots, n_k)$ be a vector of positive integers such that $\sum_{i=1}^k n_i = n$. Then, the eppf associated with a $P_\varepsilon \sim \varepsilon$-NormCRM$(\rho, \kappa P_0)$ is*

$$p_\varepsilon(n_1, \dots, n_k) = \int_0^{+\infty} \left[\frac{u^{n-1}}{\Gamma(n)} \frac{(k + \Lambda_{\varepsilon,u})}{\Lambda_\varepsilon} e^{(\Lambda_{\varepsilon,u} - \Lambda_\varepsilon)} \prod_{i=1}^k \int_\varepsilon^{+\infty} \kappa s^{n_i} \mathrm{e}^{-us} \rho(s) ds \right] du \tag{3.9}$$

where $\Lambda_{\varepsilon,u}$ has been defined in (3.8).

A result concerning the eppf of a generic normalized (homogeneous) completely random measure can be obtained from [39], formulas (36)-(37):

$$p(n_1, \dots, n_k) = \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)} \mathrm{e}^{\kappa \int_0^{+\infty} (\mathrm{e}^{-us} - 1)\rho(s) ds} \left(\prod_{i=1}^k \int_0^{+\infty} \kappa s^{n_i} \mathrm{e}^{-us} \rho(s) ds \right) du. \tag{3.10}$$

It follows that the eppf of (3.4) converges pointwise to that of the corresponding (homogeneous) normalized completely random measure (2.5) when ε tends to 0.

Proposition 3.2. *Let $p_\varepsilon(\cdot)$ be the eppf of a $\varepsilon-$NormCRM$(\rho, \kappa P_0)$. Then for any sequence $n_1, \dots, n_k$ of positive integers with $k > 0$ and $\sum_{i=1}^k n_i = n$,*

$$\lim_{\varepsilon \to 0} p_\varepsilon(n_1, \dots, n_k) = p_0(n_1, \dots, n_k), \tag{3.11}$$

where $p_0(\cdot)$ is the eppf of the NormCRM$(\rho, \kappa P_0)$ as in (3.10).

Convergence of the sequence of eppfs yields convergence of the sequences of $\varepsilon - NormCRMs$, generalizing a result obtained for $\varepsilon - NGG$ processes.

Proposition 3.3. *Let P_ε be a $\varepsilon - NormCRM(\rho, \kappa P_0)$, for any $\varepsilon > 0$. Then*

$$P_\varepsilon \xrightarrow{d} P \text{ as } \varepsilon \to 0,$$

where P is a NormCRM$(\rho, \kappa P_0)$. Moreover, as ε tends to $+\infty$, $P_\varepsilon \xrightarrow{d} \delta_{\tau_0}$, where $\tau_0 \sim P_0$.

The proof of the above proposition is along the same lines as the proof of Proposition 1 in [3], and therefore it is omitted here.

Furthermore, the m-th moment of P_ε, $m = 1, 2, \dots$, is equal to:

$$\mathbb{E}\left[(P_\varepsilon(B))^m\right] = \mathbb{E}\left[(P_0(B))^{K_m}\right] \tag{3.12}$$

where $B \in \mathcal{B}(\Theta)$ and K_m is the number of distinct values in a sample of size m from P_ε. In particular, when $m = 2$, K_m assumes values in $\{1, 2\}$, and the

probability that $K_2 = 1$ is the probability that, in a sample of size 2 from P_ε, the sample values coincide, i.e. $p_\varepsilon(2)$. Therefore $\mathbb{E}(P_\varepsilon(B)^2) = P_0(B)p_\varepsilon(2) + (P_0(B))^2(1 - p_\varepsilon(2))$, and consequently

$$\mathrm{Var}(P_\varepsilon(B)) = p_\varepsilon(2)P_0(B)\left(1 - P_0(B)\right). \tag{3.13}$$

Analogously, the covariance structure of P_ε is as follows:

$$\mathrm{Cov}(P_\varepsilon(B_1), P_\varepsilon(B_2)) = p_\varepsilon(2)\left(P_0(B_1 \cap B_2) - P_0(B_1)P_0(B_2)\right) \tag{3.14}$$

for any B_1, $B_2 \in \mathcal{B}(\Theta)$. Proofs of (3.12) and (3.14) are given in Appendix B.

4. $\varepsilon - NormCRM$ process mixtures

We consider mixtures of parametric kernels as the distribution of data, where the mixing measure is the $\varepsilon - NormCRM(\rho, \kappa P_0)$. The model we assume is the following:

$$\begin{aligned} Y_i|\theta_i &\overset{\text{ind}}{\sim} f(\cdot;\theta_i),\ i = 1,\dots,n \\ \theta_i|P_\varepsilon &\overset{\text{iid}}{\sim} P_\varepsilon,\ i = 1,\dots,n \\ P_\varepsilon &\sim \varepsilon - NormCRM(\rho, \kappa P_0), \\ \varepsilon &\sim \pi(\varepsilon), \end{aligned} \tag{4.1}$$

where $f(\cdot;\theta_i)$ is a parametric family of densities on $\mathbb{Y} \subset \mathbb{R}^p$, for all $\theta \in \Theta \subset \mathbb{R}^m$. Remember that P_0 is a non-atomic probability measure on Θ, such that $\mathbb{E}(P_\varepsilon(A)) = P_0(A)$ for all $A \in \mathcal{B}(\Theta)$ and $\varepsilon \geq 0$. Model (4.1) will be addressed here as $\varepsilon - NormCRM$ *hierarchical mixture* model.

The design of a Gibbs scheme to sample from the posterior distribution of model (4.1) is straightforward, once we have augmented the state space with the variable u, by using the posterior characterization in Theorem 3.1. The Gibbs sampler generalizes that one provided in [3] for $\varepsilon - NGG$ mixtures, but it is designed for any Lévy intensity ρ under (2.2) and (2.4). Description of the full-conditionals is below, and further details can be found in Appendix A.

1. **Sampling from** $\boldsymbol{\mathcal{L}}(u|\boldsymbol{Y}, \boldsymbol{\theta}, P_\varepsilon, \varepsilon)$: it is clear that, conditionally to P_ε, u is independent from the other variables and distributed according to gamma with parameters (n, T_ε).
2. **Sampling from** $\boldsymbol{\mathcal{L}}(\boldsymbol{\theta}|u, \boldsymbol{Y}, P_\varepsilon, \varepsilon)$: each θ_i, for $i = 1,\dots,n$, has discrete law with support $\{\tau_0, \tau_1, \dots, \tau_{N_\varepsilon}\}$, and probabilities $\mathbb{P}(\theta_i = \tau_j) \propto J_j f(Y_i;\tau_j)$.
3. **Sampling from** $\boldsymbol{\mathcal{L}}(P_\varepsilon, \varepsilon|u, \boldsymbol{\theta}, \boldsymbol{Y})$: this step is not straightforward and can be split into two consecutive substeps:
   - 3.a **Sampling from** $\boldsymbol{\mathcal{L}}(\varepsilon|u, \boldsymbol{\theta}, \boldsymbol{Y})$: see Appendix A.
   - 3.b **Sampling from** $\boldsymbol{\mathcal{L}}(P_\varepsilon|\varepsilon, u, \boldsymbol{\theta}, \boldsymbol{Y})$: via characterization of the posterior in Theorem 3.1, since this distribution is equal to $\mathcal{L}(P_\varepsilon|\varepsilon, u, \boldsymbol{\theta})$.

To put into practice, we have to sample (i) the number N_{na} of *non-allocated* jumps, (ii) the vector of the unnormalized *non-allocated* jumps $\boldsymbol{J}^{(na)}$, (iii) the vector of the unnormalized *allocated* jumps $\boldsymbol{J}^{(a)}$, the support of the *allocated* (iv) and *non-allocated* (v) jumps. See Appendix A for a wider description.

We highlight that, when sampling from non-standard distributions, Accept-Reject or Metropolis-Hastings algorithms have been exploited.

5. Some ideas on the choice of ε

We believe that a brief discussion on the choice of the approximation parameter ε is worth doing. We could also consider it random, as we did in [3], where the ε-NGG mixture model was proposed. In our general view, this parameter can be considered either as a true parameter, and then it should be fixed on the ground of the prior information we have, or as a tuning parameter to approximate the "exact" model (normalized completely random measure mixtures). If we prefer the latter alternative as we did here, ε has to be small. However, since the result on ε-approximation (Theorem 3.3) concerns the prior distribution in (4.1), the only suggestions we can give refer to a priori criteria. Here we suggest to set ε such that the sum of the masses $\mu((0,\varepsilon])$ and J_0 we perturb μ with, obtaining μ_ε, is small. In particular, since the interest is in normalized random measures, "small" is fixed with respect to the expectation $\mathbb{E}(T)$ of the total mass of μ, i.e. we choose ε such that

$$r(\varepsilon) := \frac{\mathbb{E}(\mu(0,\varepsilon]) + \mathbb{E}(J_0)}{\mathbb{E}(T)} \leq \nu, \tag{5.1}$$

where ν is typically a small value. Rather, alternative criteria are available; for instance, as in [3], we could choose ε to achieve a prefixed value for $\mathbb{E}(N_\varepsilon)$ or $\mathrm{Var}(N_\varepsilon)$. As far as (5.1) is concerned, observe that

$$\mathbb{E}(\mu(0,\varepsilon]) = \kappa \int_0^\varepsilon s\rho(s)ds, \qquad \mathrm{Var}(\mu(0,\varepsilon]) = \kappa \int_0^\varepsilon s^2\rho(s)ds;$$

from (2.2), it follows that

$$\mathbb{E}(\mu(0,\varepsilon]) \to 0 \quad \mathrm{Var}(\mu(0,\varepsilon]) \to 0 \text{ as } \varepsilon \to 0,$$

i.e. the r.v. $\mu(0,\varepsilon]$ converges to 0 in $\mathcal{L}_2$ and this implies convergence in probability. Besides, we have that

$$\varepsilon \leq \mathbb{E}(J_0) = \frac{\kappa \int_\varepsilon^{+\infty} s\rho(s)ds}{\Lambda_\varepsilon} \leq \frac{\mathbb{E}(T)}{\mathbb{E}(N_\varepsilon)}.$$

Consequently, when $\varepsilon \to 0$, $\mathbb{E}(N_\varepsilon) \to +\infty$ and thus $\mathbb{E}(J_0)$ converges to 0.

As an interesting example, we evaluate the ratio $r(\varepsilon)$ when $\rho(s) = 1/\Gamma(1-\sigma)s^{-1-\sigma}\mathrm{e}^{-\omega\varepsilon}$ for $0 \leq \sigma < 1$, $\kappa > 0$ and $\omega = 1$, that means when μ is the

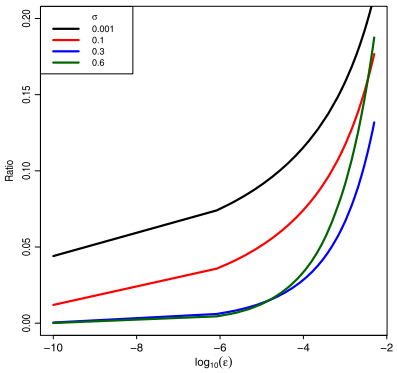


FIG 1. *Values of $r(\varepsilon)$ when ρ is the Lévy intensity of the generalized gamma c.r.m., with $\kappa = 1$ and different values of σ, as a function of $\log_{10}(\varepsilon)$.*

generalized gamma process, i.e. the unnormalized c.r.m. defining NGG processes by normalization. By 8.354.2 in [22], we have that

$$\begin{aligned}\mathbb{E}\left(\mu(0,\varepsilon]\right) &= \frac{\kappa}{\Gamma(1-\sigma)}\left(\Gamma(1-\sigma) - \Gamma(1-\sigma;\varepsilon)\right) \\ &= \frac{\kappa}{\Gamma(1-\sigma)}\left(\sum_{n=0}^{+\infty}\frac{(-1)^n\varepsilon^{1-\sigma+n}}{n!(1-\sigma+n)}\right) \overset{\varepsilon\to 0}{\sim} \frac{\kappa\varepsilon^{1-\sigma}}{\Gamma(2-\sigma)},\end{aligned}$$

and $\mathbb{E}(J_0) = \Gamma(1-\sigma,\varepsilon)/\Gamma(-\sigma,\varepsilon)$. We also mention that $\mathrm{Var}(\mu(0,\varepsilon]) \sim (\kappa\varepsilon^{2-\sigma})/\Gamma(2-\sigma)$ as ε tends to 0. Figure 1 shows $r(\varepsilon)$ when μ is the generalized gamma process with $\kappa = 1$ and different values of σ, as a function of ε. Note that a smaller threshold ε is needed in order to obtain the same value of ν when the parameter σ decreases to 0.

Similar calculations can be derived when μ is the Bessel random measure introduced in the next section.

6. Normalized Bessel random measure mixtures: density estimation

In this section we introduce a new normalized process, called normalized Bessel random measure. Section 6.1 describes theoretical results: in particular, we show that this family encompasses the well-known Dirichlet process. Then we fit the mixture model to synthetic and real datasets in Section 6.2. Results are illustrated through a density estimation problem.

6.1. Definition

Let us consider a normalized completely random measure corresponding to mean intensity

$$\rho(s;\omega) = \frac{1}{s}\mathrm{e}^{-\omega s}I_0(s), \quad s > 0,$$

where $\omega \geq 1$ and

$$I_\nu(s) = \sum_{m=0}^{+\infty} \frac{(s/2)^{2m+\nu}}{m!\Gamma(\nu+m+1)}$$

is the modified Bessel function of order $\nu > 0$ [see 15, Sect 7.2.2]. It is straightforward to see that, for $s > 0$,

$$\rho(s;\omega) = \frac{1}{s}\mathrm{e}^{-\omega s} + \sum_{m=1}^{+\infty} \frac{1}{2^{2m}(m!)^2} s^{2m-1}\mathrm{e}^{-\omega s}, \tag{6.1}$$

so that ρ is the sum of the Lévy intensity of the gamma process with rate parameter ω and of the Lévy intensities

$$\rho_m(s;\omega) = \frac{1}{2^{2m}(m!)^2} s^{2m-1}\mathrm{e}^{-\omega s}, \quad s > 0, \qquad m = 1, 2, \ldots \tag{6.2}$$

corresponding to finite activity Poisson processes. It is simple to check that (2.2) and (2.4) hold. Hence, following (2.5) in Section 2, we introduce the *normalized Bessel random measure* P, with parameters (ω, κ), where $\omega \geq 1$ and $\kappa > 0$. Thanks to (6.1) and the Superposition Property of Poisson processes the total mass T in (2.5) can be written as

$$T \stackrel{d}{=} T_G + \sum_{m=1}^{+\infty} T_m, \tag{6.3}$$

where T_G, T_1, $T_2, \ldots$ are independent random variables, T_G being the total mass of the gamma process and T_m the total mass of a completely random measure corresponding to the intensity $\nu_m(ds, d\tau) = \rho_m(s)ds\kappa P_0(d\tau)$. In particular, $T_G \sim gamma(\kappa, \omega)$, while $T_m = \sum_{j=1}^{N_m} J_j^{(m)}$, where $N_m \sim Poi(\kappa\Gamma(2m)/((2\omega)^{2m}(m!)^2))$, and $\{J_j^{(m)}\}$ are the points of a Poisson process on $\mathbb{R}^+$ with intensity $\kappa\rho_m$. By this notation we mean that T_m is equal to 0 when $N_m = 0$, while, conditionally to $N_m > 0$, $J_j^{(m)} \stackrel{\text{iid}}{\sim} gamma(2m, \omega)$. We can write down the density function of T, via (2.3):

$$\begin{aligned}
\psi(\lambda) &:= -\log\left(\mathbb{E}(\mathrm{e}^{-\lambda T})\right) = \kappa \int_0^{+\infty} (1 - \mathrm{e}^{-\lambda s})\rho(s;\omega)ds \\
&= \kappa\left(\log\left(\frac{\omega+\lambda}{\omega}\right) + \sum_{m=1}^{+\infty} \frac{\Gamma(2m)}{2^{2m}(m!)^2\omega^m} - \sum_{m=1}^{+\infty} \frac{\Gamma(2m)}{2^{2m}(m!)^2(\omega+\lambda)^m}\right) \\
&= \kappa \log\left(\frac{\omega+\lambda+\sqrt{(\omega+\lambda)^2-1}}{\omega+\sqrt{\omega^2-1}}\right).
\end{aligned}$$

The same expression is obtained when $T \sim f_T(t) = \kappa(\omega+\sqrt{\omega^2-1})^\kappa \dfrac{\mathrm{e}^{-\omega t}}{t} I_\kappa(t)$, $t > 0$ [see 22, formula (17.13.112)]. Observe that, when $\omega = 1$, f_T is called

Bessel function density [18]. By (3.10), the eppf of the normalized Bessel random measure is:

$$p_B(n_1,\dots,n_k;\omega,\kappa) = \kappa^k \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)} \left(\frac{\omega+\sqrt{\omega^2-1}}{\omega+u+\sqrt{(\omega+u)^2-1}} \right)^{\kappa} \frac{1}{(u+\omega)^n} \times \prod_{j=1}^{k} \Gamma(n_j) \;\; {}_2F_1\left(\frac{n_j}{2}, \frac{n_j+1}{2}; 1; \frac{1}{(u+\omega)^2} \right) du, \tag{6.4}$$

where

$${}_2F_1(\alpha_1,\alpha_2;\gamma;z) := \sum_{m=0}^{\infty} \frac{(\alpha_1)_m (\alpha_2)_m}{(\gamma)_m} \frac{1}{m!} (z)^m, \qquad \text{with } (\alpha)_m := \frac{\Gamma(\alpha+m)}{\Gamma(\alpha)}$$

is the hypergeometric series [see 22, formula (9.100)].

The following proposition shows that the eppf of the normalized Bessel random measure converges to the eppf of the Dirichlet process as the parameter ω increases. The proof is given in Appendix B.

Proposition 6.1. *Let $(n_1,\dots,n_k)$ be a vector of positive integers such that $\sum_{i=1}^{k} n_i = n$, where $k = 1,\dots,n$. Then, the eppf (6.4), associated with the normalized Bessel random measure P with parameter (ω,κ), $\omega \geq 1$, $\kappa > 0$, and mean measure P_0, is such that*

$$\lim_{\omega\to+\infty} p_B(n_1,\dots,n_k;\omega,\kappa) = p_D(n_1,\dots,n_k;\kappa),$$

where $p_D(n_1,\dots,n_k;\kappa)$ is the eppf of the Dirichlet process with measure parameter κP_0.

The prior distribution of K_n, the number of distinct values in a sample of size n from the normalized Bessel random measure, could be derived from its eppf in (6.4). However, this is not an easy task from a computational point of view, so that we prefer to use a Monte Carlo strategy to simulate from the prior of the K_n. The simulation strategy is also useful to understanding the meaning of the parameters of the normalized Bessel random measure: κ has the usual interpretation of the mass parameter, since, when fixing ω, $\mathbb{E}(K_n)$ increases with κ. On the other hand, the effect of ω is quite peculiar: decreasing ω (thus drifting apart from the Dirichlet process), with κ fixed, the prior distribution of K_n shifts towards smaller values. However, when $\mathbb{E}(K_n)$ is kept fixed, the distribution has heavier tails if ω is small (see Figures 2 and 4 (a)).

The Lévy intensity (6.1) of the normalized Bessel completely random measure has a similar expression as the intensity corresponding to an element of the class $\mathcal{C}$ in [43]. Both intensities are linear combinations of intensities of the gamma process and of the type $s^{i-1}\mathrm{e}^{-\omega s}\mathbb{I}_{(0,+\infty)}(s)$, corresponding to finite activity Poisson processes. Here, the intensity of the Bessel r.p.m. corresponds to an infinite mixture with fixed weights, where the indexes i are even integers (see (6.2)), while [43] assume a linear combination of a finite number of components, through a vector of parameters.

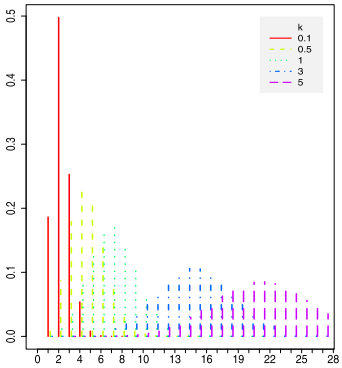


FIG 2. *Prior distribution of K_n under a sample from ε-NB process with $\varepsilon = 10^{-6}$, $\omega = 1.05$ and several values for κ, as reported in the legend.*

6.2. Application

In this section let us consider the hierarchical mixture model (4.1), where the mixing measure is P_ε, the ε-approximation of the normalized Bessel random measure, as introduced above (here ε-NB$(\omega,\ \kappa P_0)$ mixture model). Of course, when ε is small, this model approximates the corresponding mixture when the mixing measure is P; to the best of our knowledge, this normalized Bessel completely random measure has never been considered in the Bayesian nonparametric literature. By decomposition (6.3), we argue that this model is suitable when the unknown density shows many different components, where a few of them are very spiky (they should correspond to Lévy intensities (6.2)), while there is a folk of flatter components which are explained by the intensity $(1/s)\mathrm{e}^{-\omega s}$ of the Gamma process. For this reason, we consider a simulated dataset which is a sample from a mixture of 5 Gaussian distributions with means and standard deviations equal to $\{(15, 1.1), (50, 1), (20, 4), (30, 5), (40, 5)\}$, and weights proportional to $\{10, 9, 4, 5, 5\}$. The histogram of the simulated data, for $n = 1000$, is reported in Figure 3.

We report posterior estimates for different sets of hyperparameters of the ε-NB mixture model when $f(\cdot;\theta)$ is the Gaussian density on $\mathbb{R}$ and $\theta = (\mu, \sigma^2)$ stands for its mean and variance. Moreover, $P_0(d\mu, d\sigma^2) = \mathcal{N}(d\mu; \bar{y}_n, \sigma^2/\kappa_0) \times inv - gamma(d\sigma^2; a, b)$; here $\mathcal{N}(\bar{y}_n, \sigma^2/\kappa_0)$ is the Gaussian distribution with mean $\bar{y}_n$(the empirical mean) and variance σ^2/κ_0, and $inv - gamma(d\sigma^2; a, b)$ is the inverse-gamma distribution with mean $b/(a-1)$ (if $a > 1$). We set $\kappa_0 = 0.01$, $a = 2$ and $b = 1$ as proposed first in [16]. We shed light on three sets of hyperparameters in order to understand sensitivity of the estimates under different conditions of variability; indeed, each set has a different value of $p_\varepsilon(2)$, which tunes the a-priori variance of P_ε, as reported in (3.13). We tested three different values for $p_\varepsilon(2)$: $p_\varepsilon(2) = 0.9$ in set (A), $p_\varepsilon(2) = 0.5$ in set (B) and $p_\varepsilon(2) = 0.1$ in set (C). Moreover, in each scenario we let the parameter $1/\omega$ range in $\{0.01, 0.25, 0.5, 0.75, 0.95\}$; note that the extreme case of $\omega = 100$ (or

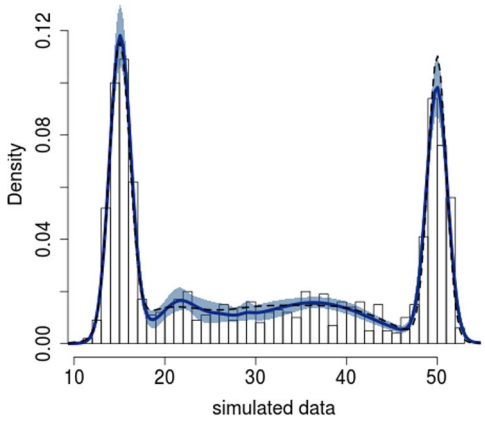


FIG 3. *Density estimate for case A5: posterior mean (line), 90% pointwise credibility intervals (shadowed area), true density (dashed) and the histogram of simulated data.*

equivalently $1/\omega = 0.01$) corresponds to an approximation of the DPM model. The mass parameter κ is then fixed to achieve the desired level of $p_\varepsilon(2)$. As far as the choice of ε concerns, we set it equal to 10^{-6}: this provides pretty good approximation a priori (see Section 5); moreover, posterior inference proved to be fairly robust with respect to ε. At the end, we got 15 tests, listed in Table 1. As mentioned before, it is possible to choose a prior for ε, even if, for the ρ in (6.1), the computational cost would greatly increase due to the evaluation of functions ${}_2F_1$ in (6.4).

We have implemented our Gibbs sampler in C++. All the tests in Sections 6 and 7 were made on a laptop with Intel Core i7 2670QM processor, with 6GB of RAM. Every run produced a final sample size of 5000 iterations, after a thinning of 10 and an initial burn-in of 5000 iterations. Every time the convergence was checked by standard R package CODA tools.

Here, we focus on density estimation: all the tests provide similar estimates, quite faithful to the true density. Figure 3 shows density estimate and pointwise 90% credibility intervals for case A5; the true density is superimposed as dashed line. Figure 4 (a) and (b) display prior and posterior distributions, respectively, of the number K_n of groups, i.e. the number of unique values among $(\theta_1, \ldots, \theta_n)$ in (4.1) under two sets of hyperparameters, A1, representing an approximation of the DPM model, and A5, where the parameter ω is nearly 1. From Figure 4 it is clear that A5 is more flexible than A1: for case A5, a priori the variance of K_n is larger, and, on the other hand, the posterior probability mass in 5 (the true value) is larger.

In order to compare different priors, we consider five different predictive goodness-of-fit indexes: (i) the sum of squared errors (SSE) , i.e. the sum of the squared differences between the y_i and the predictive mean $\mathbb{E}(Y_i|data)$ (yes, we are using data twice!); (ii) the sum of standardized absolute errors (SSAE), given by the sum of the standardized error $|y_i - \mathbb{E}(Y_i|data)|/\sqrt{\mathrm{Var}(Y_i|data)}$;

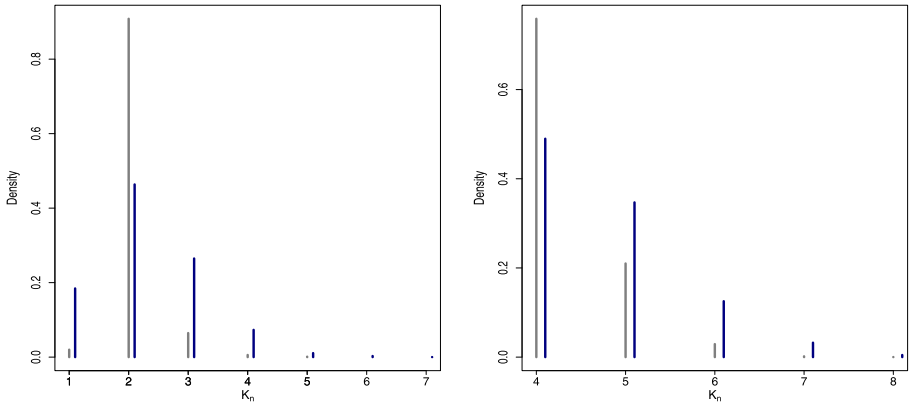
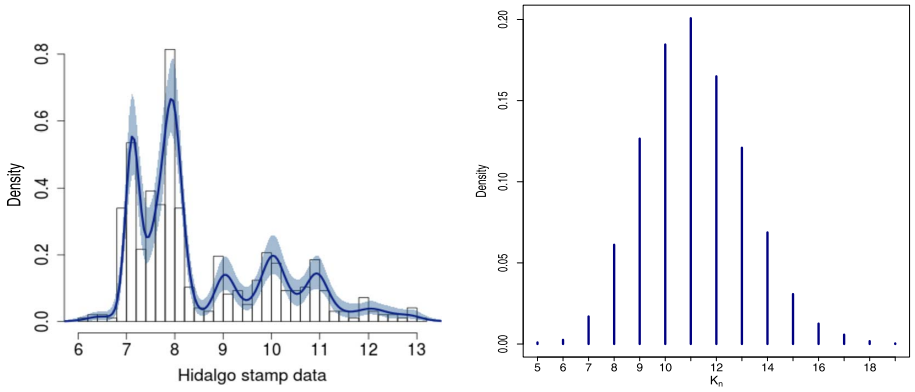


FIG 4. *Prior (a) and posterior (b) distributions of the number K_n of groups for test A1 (gray) and A5 (blue).*

(*iii*) log-pseudo marginal likelihood (LPML), quite standard in the Bayesian literature, defined as the sum of $\log(CPO_i)$, where CPO_i is the conditional predictive ordinate of y_i, the value of the predictive distribution evaluated at y_i, conditioning on the training sample given by all data except y_i. The last two indexes, (*iv*) $WAIC_1$ and (*v*) $WAIC_2$, as denoted here, were proposed in [44] and deeply analyzed in [21]: they are generalizations of the AIC, adding two types of penalization, both accounting for the "effective number of parameters". The bias correction in $WAIC_1$ is similar to the bias correction in the definition of the DIC, while $WAIC_2$ is the sum of the posterior variances of the conditional density of the data. See [21] for their precise definition. Table 1 shows the values of the five indexes for each test: the optimal (according to each index) tests are

TABLE 1
Predictive goodness-of-fit indexes for the simulated dataset.

Test	ω	κ	SSE	SSAE	WAIC1	WAIC2	LPML
A1	100	0.06	6346.59	811.16	-3312.44	-3312.55	-3312.55
A2	4	0.09	5812.86	810.43	-3312.33	-3312.42	-3312.43
A3	2	0.1	6089.19	810.99	-3312.38	-3312.47	-3312.48
A4	1.33	0.11	6498.23	811.29	-3312.54	-3312.62	-3312.63
A5	1.05	0.11	**5725.18**	**810.39**	**-3312.27**	**-3312.36**	**-3312.36**
B1	100	0.43	5184.25	809.61	-3311.95	-3312	-3312.01
B2	4	0.67	5125.41	809.7	-3312.19	-3312.25	-3312.26
B3	2	0.81	4610.39	809.42	-3311.92	-3311.98	-3312
B4	1.33	0.93	4246.43	**809.07**	**-3311.75**	**-3311.83**	**-3311.84**
B5	1.05	1	**4571.09**	809.08	-3311.96	-3312.05	-3312.06
C1	100	1.56	3707.5	809.36	**-3311.73**	**-3311.86**	**-3311.88**
C2	4	2.67	2194.1	808.8	-3312.02	-3312.23	-3312.26
C3	2	3.64	1223.86	809.28	-3312.62	-3312.96	-3312.99
C4	1.33	5.29	748.85	808.7	-3313.05	-3313.51	-3313.54
C5	1.05	8.95	**685**	**807.96**	-3312.9	-3313.36	-3313.38



(a) (b)

FIG 5. *Posterior inference for the Hidalgo stamp data for Test 4: histogram of the data, density estimate and 90% pointwise credibility intervals (a); posterior distribution of K_n (b).*

highlighted in bold for the experiments (A), (B) and (C). It is apparent that the different tests provide similar values of the indexes, but SSE, indicating that, from a predictive viewpoint, there are no significant differences among the priors. However, especially when the value of κ is small, i.e. in all tests A and B, a model with a smaller ω tends to outperform the Dirichlet process case (approximately, when $\omega = 100$). On the other hand, the SSE index shows quite different values among the tests: it is well-known that this is a index favoring complex models and leading to better results when data are over-fitted. Therefore, tests with an higher value of κ are always preferable according to this criterion.

We fitted our model also to a real dataset, the Hidalgo stamps data of [45] consisting of $n = 485$ measurements of stamp thickness in millimeters (here multiplied by 10^3). The stamps have been printed between 1872 and 1874 on different paper types, see data histogram in Figure 5. This dataset has been analyzed by different authors in the context of mixture models: see, for instance, [37].

We report posterior inference for the set of hyperparameters which is most in agreement with our prior belief: the mean distribution is $P_0(d\mu, d\sigma^2) = \mathcal{N}(d\mu; \bar{y}_n, \sigma^2/\kappa_0) \times inv - gamma(d\sigma^2; a, b)$ as before, and $\kappa_0 = 0.005$, $a = 2$ and $b = 0.1$. The approximation parameter ε of the ε-NB$(\omega, \kappa P_0)$ random measure is fixed to 10^{-6}; on the other hand, in order to set parameters ω and κ, we argue as follows: ω ranges in $\{1.05, 5, 10, 1000\}$ and we choose the mass parameter κ such that the prior mean of the number of clusters, i.e. $\mathbb{E}(K_n)$, is the desired one. As noted in Section 6.1, a closed form of the prior distribution of K_n is not available, so we resort to Monte Carlo simulation to estimate it. Table 2 shows the four couples of (ω, κ) yielding $\mathbb{E}(K_n) = 7$: indeed, according to [26] and [36] and references therein, there are at least 7 different groups (but the true number is unknown), corresponding to the number of types of paper used. For an in-depth discussion about the appropriate number of groups in Hidalgo stamps data, we refer the reader to [10]. Table 2 also reports prior stan-

TABLE 2
Predictive goodness-of-fit indexes for the Hidalgo stamps data.

Test	ω	κ	$\mathbb{E}(K_n)$	$sd(K_n)$	SSE	SSAE	WAIC1	WAIC2	LPML
1	1000	0.98	7	2.04	15.17	384.1	-713.12	-713.96	-714.12
2	10	0.91	7	2.13	12.85	383.51	-713.22	-714.04	-714.25
3	5	0.92	7	2.18	13.52	383.68	-713.52	-714.3	-714.4
4	1.05	1.02	7	2.32	**11.12**	**383.38**	**-712.84**	**-713.66**	**-714.05**

dard deviations of K_n: even if the a-priori differences are small, the posteriors appear to be quite different among the 4 tests. All the posterior distributions on K_n support the conjecture of at least seven distinct modes in the data; in particular, Figure 5 (b) displays the posterior distribution of K_n for Test 4. A modest amount of mass is given to less than 7 groups, and the mode is in 11. Even Test 1, corresponding to the Dirichlet process case, does not give mass to less than 7 groups, where 9 is the mode. Density estimates seem pretty good; an example is given in Figure 5 (a), with 90% credibility band for Test 4.

As in the simulated data example, some predictive goodness-of-fit indexes are reported in Table 2: the optimal value for each index is indicated in bold. The SSE is significantly lower when ω is small, thus suggesting a greater flexibility of the model with small values of ω. The other indexes assume the optimal value in Test 4 as well, even if those values are similar along the tests.

Our ε-approximation method turned out to be accurate and fast when compared with competitors (the slice sampler and an a-posteriori truncation method) when the mixing r.p.m is the NGG process and the kernel is Gaussian; see [3], Section 5.

7. Linear dependent *NGG* mixtures: application to sports data

Let us consider a regression problem, where the response Y is univariate and continuous, for ease of notation. We model the relationship (in distributional terms) between the vector of covariates $\boldsymbol{x} = (x_1, \ldots, x_p)$ and the response Y through a mixture density, where the mixing measure is a collection $\{P_{\boldsymbol{x}}, \mathbf{x} \in \mathcal{X}\}$ of $\varepsilon - NormCRMs$, being $\mathcal{X}$ the space of all possible covariates. We follow the same approach as in [35] and [14] for the dependent Dirichlet process. We define the *dependent* $\varepsilon - NormCRM$ *process* $\{P_{\boldsymbol{x}}, \boldsymbol{x} \in \mathcal{X}\}$, conditionally to $\boldsymbol{x}$, as:

$$P_{\boldsymbol{x}} \stackrel{d}{=} \sum_{j=0}^{N_\varepsilon} P_j \delta_{\gamma_j(\boldsymbol{x})}. \tag{7.1}$$

The weights P_j are the normalized jumps as in (3.4), while the locations $\gamma_j(\boldsymbol{x})$, $j = 1, 2, \ldots$, are independent stochastic processes with index set $\mathcal{X}$ and $P_{0\boldsymbol{x}}$ marginal distributions. Model (7.1) is such that, marginally, $P_{\boldsymbol{x}}$ follows a $\varepsilon - NormCRM$ process, with parameter $(\rho, \kappa P_{0\boldsymbol{x}})$, where ρ is the intensity of a Poisson process on $\mathbb{R}^+$, $\kappa > 0$, and $P_{0\boldsymbol{x}}$ is a probability on $\mathbb{R}$. Observe that, since N_ε and P_j do not depend on $\boldsymbol{x}$, (7.1) is a generalization of the single weights

dependent Dirichlet process [see 8, for this terminology]. We also assume the functions $\boldsymbol{x} \mapsto \gamma_j(\boldsymbol{x})$ to be continuous.

The dependent $\varepsilon - NormCRM$ process in (7.1) takes into account the vector of covariates $\boldsymbol{x}$ only through $\gamma_j(\boldsymbol{x})$. In particular, when the kernel of the mixture (4.1) belongs to the exponential family, for each j, $\gamma_j(\boldsymbol{x}) = \gamma(\boldsymbol{x}; \boldsymbol{\tau}_j)$ can be assumed as the link function of a generalized linear model, so that (4.1) specializes to

$$\begin{aligned} Y_i|\boldsymbol{\theta}_i, \boldsymbol{x}_i &\overset{\text{ind}}{\sim} f(\boldsymbol{y}; \boldsymbol{\gamma}(\boldsymbol{x}_i, \boldsymbol{\theta}_i)) \quad i = 1, \ldots, n \\ \boldsymbol{\theta}_i|P_\varepsilon &\overset{\text{iid}}{\sim} P_\varepsilon \quad i = 1, \ldots, n \qquad \text{where } P_\varepsilon \sim \varepsilon - NormCRM(\rho, \kappa P_0). \end{aligned} \tag{7.2}$$

This last formulation is convenient because it facilitates parameters interpretation as well as numerical posterior computation.

We analyze the Australian Institute of Sport (AIS) data set [12], which consists of 11 physical measurements on 202 athletes (100 females and 102 males). Here the response is the lean body mass (lbm), while three covariates are considered, the red cell count (rcc), the height in cm (Ht) and the weight in Kg (Wt). The data set is contained in the R package `DPpackage` [28]. The actual model (7.2) we consider here is when $f(\cdot; \mu, \eta^2)$ is the Gaussian distribution with μ mean and η^2 variance; moreover, $\mu = \gamma(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^t\boldsymbol{\theta}$, and the mixing measure P_ε is the ε-NGG(κ, σ, P_0), as introduced in [3]. We have considered two cases, when mixing the variance η^2 with respect to the NGG process or when the variance η^2 is given a parametric density; in both cases, by linearity of the mean $\boldsymbol{x}^t\boldsymbol{\theta}$, the model (here called linear dependent NGG mixture) can be interpreted as a NGG process mixture model, and inference can be achieved via an algorithm similar to that in Section 4. We set $\varepsilon = 10^{-6}$, which provides a moderate value for the ratio $r(\varepsilon)$ in (5.1), and $\sigma \in \{0.001, 0.125, 0.25\}$, κ such that $\mathbb{E}(K_n) \simeq 5$ or 10. When the variance η^2 is included in the location points of the $\varepsilon - NGG$ process, then P_0 is $\mathcal{N}_4(\boldsymbol{b}_0, \Sigma_0) \times inv - gamma(\nu_0/2, \nu_0\eta_0^2/2)$; on the other hand, when η^2 is given a parametric density, then $\eta^2 \sim inv - gamma(\nu_0/2, \nu_0\eta_0^2/2)$. We fixed hyperparameters in agreement with the least squares estimate: $\boldsymbol{b}_0 = (-50, 5, 0, 0)$, $\Sigma_0 = diag(100, 10, 10, 10)$, $\nu_0 = 4$, $\eta_0^2 = 1$. For all the experiments, we computed the posterior of the number of groups, the predictive densities at different values of the covariate vectors and the cluster estimate via posterior maximization of Binder's loss function [see 31]. Moreover, we compared the different prior settings computing predictive goodness-of-fit tools, specifically log pseudo-marginal likelihood (LPML) and the sum of squared errors (SSE), as introduced in Section 6.2. The minimum value of SSE, among our experiments, was achieved when η^2 is included in the location of the $\varepsilon - NGG$ process, $\sigma = 0.001$ and $\kappa = 0.8$ so that $\mathbb{E}(K_n) \simeq 5$. On the other hand, the optimal LPML was achieved when $\sigma = 0.125$, $\kappa = 0.4$, and $\mathbb{E}(K_n) \simeq 5$. Posterior of K_n and cluster estimate under this last hyperparameter setting are in Figure 6 ((a) and (b), respectively); in particular the cluster estimate is displayed in the scatterplot of the Wt vs lbm. In spite of the vague prior, the posterior of K_n is almost degenerate on 2, giving evidence to the existence of two linear relationships between lbm and Wt.

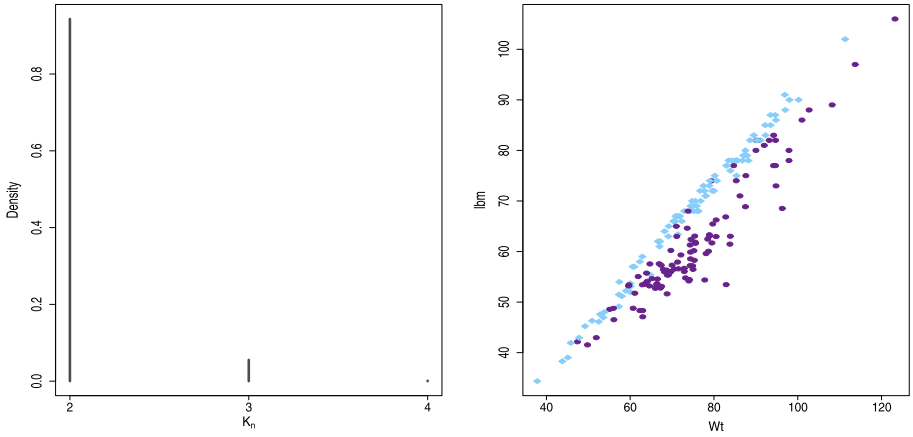


FIG 6. *Posterior distribution of the number K_n of groups (a) and cluster estimate (b) under the linear dependent $\varepsilon - NGG$ mixture.*

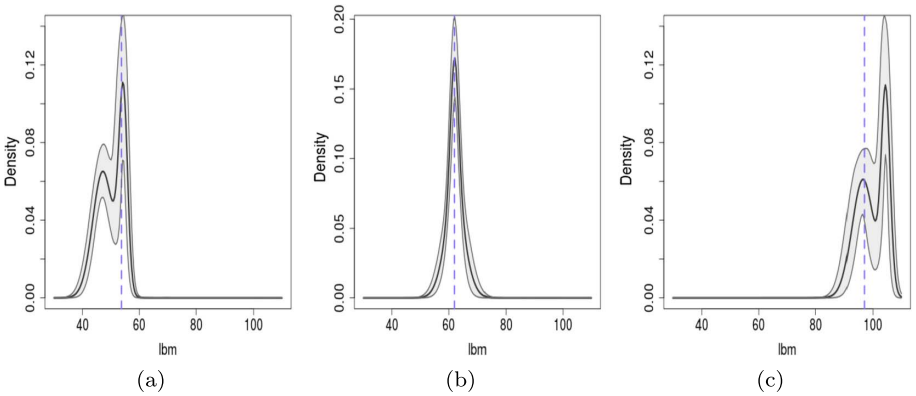


FIG 7. *Predictive distributions of lbm for three different athletes: Wt=60, rcc=3.9, Ht=176 (a), Wt=67.1, rcc=5.34, Ht=178.6 (b), Wt=113.7, rcc=5.17, Ht=209.4 (c). The shaded area is the predictive 95% pointwise credible interval, while the dashed vertical line denotes the observed value of the response.*

Finally, Figure 7 displays predictive densities and 95% credibility bands for 3 athletes, a female (Wt=60, rcc=3.9, Ht=176 and lbm=53.71), and two males (Wt=67.1,113.7, rcc=5.34,5.17, Ht=178.6, 209.4 and lbm=62,97) respectively, under the same hyperparameter setting of Figure 6; the dashed lines are observed values of the response. Depending on the covariate values, the distribution shows one or two peaks: this reflects the dependence of the grouping of the data on the value of $\mathbf{x}$. This figure highlights the versatility of nonparametric priors in a linear regression setting with respect to the customary parametric priors: indeed, the model is able to capture in detail the behavior of the data, even when several clusters are present.

8. Discussion

We have proposed a new model for density and cluster estimation in the Bayesian nonparametric framework. In particular, a finite dimensional process, the $\varepsilon - NormCRM$, has been defined, which converges in distribution to the corresponding normalized completely random measure, when ε tends to 0. Here, the $\varepsilon - NormCRM$ is the mixing measure in a mixture model. In this paper we have fixed ε very small, but we could choose a prior for ε and include this parameter into the Gibbs sampler scheme. Among the achievements of the work, we have generalized all the theoretical results obtained in the special case of NGG in [3], including the expression of the eppf for an $\varepsilon - NormCRM$ process, its convergence to the corresponding eppf of the nonparametric underlying process and the posterior characterization of P_ε. Moreover, we have provided a general Gibbs Sampler scheme to sample from the posterior of the mixture model. To show the performance of our algorithm and the flexibility of the model, we have illustrated two examples via normalized completely random measure mixtures: in the first application, we have introduced a new normalized completely random measure, named normalized Bessel random measure; we have studied its theoretical properties and used it as the mixing measure in a model to fit simulated and real datasets. The second example we have dealt with is a linear dependent $\varepsilon - NGG$ mixture, where the dependence lies on the support points of the mixing random probability, to fit a well known dataset. Current and future research is devoted on the use of our approximation on more complex dependence structures.

Appendix A: Details on full-conditionals for the Gibbs sampler

Here, we provide some details about Step 3 of the Gibbs Sampler in Section 4. As far as Step 3a is concerned, the full-conditional $\mathcal{L}(\varepsilon|u,\boldsymbol{\theta},\boldsymbol{Y})$ is obtained integrating out N_ε (or equivalently N_{na}) from the law $\mathcal{L}(N_\varepsilon,u,\boldsymbol{\theta},\boldsymbol{Y})$, as follows:

$$\begin{aligned}\mathcal{L}(\varepsilon|u,\boldsymbol{\theta},\boldsymbol{Y}) &\propto \sum_{N_{na}=0}^{+\infty}\mathcal{L}(N_{na},\varepsilon,u,\boldsymbol{\theta},\boldsymbol{Y})\\ &= \sum_{N_{na}=0}^{+\infty}\pi(\varepsilon)e^{-\Lambda_\varepsilon}\frac{\Lambda_{\varepsilon,u}^{N_{na}}}{\Lambda_\varepsilon}\frac{(N_{na}+k)}{N_{na}!}\prod_{i=1}^{k}\int_\varepsilon^{+\infty}\kappa s^{n_i}\mathrm{e}^{-us}\rho(s)ds\\ &= \left(\prod_{i=1}^{k}\int_\varepsilon^{+\infty}\kappa s^{n_i}\mathrm{e}^{-us}\rho(s)ds\right)e^{\Lambda_{\varepsilon,u}-\Lambda_\varepsilon}\frac{\Lambda_{\varepsilon,u}+k}{\Lambda_\varepsilon}\pi(\varepsilon) = f_\varepsilon(u;n_1,\ldots,n_k)\pi(\varepsilon),\end{aligned}$$

where we used the identity $\sum_{N_{na}=0}^{+\infty}\Lambda_{\varepsilon,u}^{N_{na}}(N_{na}+k)/(N_{na}!) = e^{\Lambda_{\varepsilon,u}}(\Lambda_{\varepsilon,u}+k)$. Moreover, $f_\varepsilon(u;n_1,\ldots,n_k)$ is defined in (B.7). This step depends explicitly on the expression of $\rho(s)$. Step 3.b consists in sampling from $\mathcal{L}(P_\varepsilon|\varepsilon,u,\boldsymbol{\theta})$ as reported in Corollary 3.1. In order to sample a draw from the posterior distribution of the (unnormalized) measure, we follow Theorem 3.1. The component $\mu_{\varepsilon,u}^{(a)}$ is

obtained generating independently from $\mathcal{L}(J_{l_i^*}) \propto J_{l_i^*}^{n_i} e^{-uJ_{l_i^*}} \rho(J_{l_i^*}) \mathbb{1}_{(\varepsilon,\infty)}(J_{l_i^*})$, $i = 1, \ldots, k$. On the other hand, $\mu_{\varepsilon,u}^{(na)}$ satisfies the distributional identity described at point 1 of the proposition, and therefore we simulate it as follows:

1. Draw x from the Bernoulli distribution with parameter $p = \Lambda_{\varepsilon,u}/(\Lambda_{\varepsilon,u}+k)$.
2. Draw $N^{(na)}$ from $\mathcal{P}_x(\Lambda_\varepsilon)$, where $\mathcal{P}_x(\Lambda_\varepsilon)$ denotes the shifted Poisson distribution, with support on $\{x, x+1, x+2, \ldots\}$ and mean $\lambda + x$.
3. If $N^{(na)} = 0$, let $\mu_{\varepsilon,u}^{(na)}$ be the null measure. Otherwise, draw an iid sample $\{(J_j, \tau_j),\ j = 1, \ldots, N^{(na)}\}$, from $\rho_\varepsilon(s) ds P_0(d\tau)$, and set $\mu_{\varepsilon,u}^{(na)} = \sum_{j=1}^{N^{(na)}} J_j \delta_{\tau_j}$.

Appendix B: Proofs of the theorems

B.1. Proof of Theorem 3.1

Conditionally to the unnormalized measure μ_ε (see (3.3)), the law of $\boldsymbol{\theta}$ is given by

$$\mathbb{P}(\theta_1 \in d\theta_1, \ldots, \theta_n \in d\theta_n | \mu_\varepsilon) = \frac{1}{T_\varepsilon^n} \prod_{j=1}^{k} \mu_\varepsilon(d\theta_j^*)^{n_j}.$$

By considering the variable U in (3.5), we express the joint conditional distribution of $\boldsymbol{\theta}$ and U as

$$\mathbb{P}(\theta_1 \in d\theta_1, \ldots, \theta_n \in d\theta_n, U \in du | \mu_\varepsilon) = \frac{u^{n-1}}{\Gamma(n)} \mathrm{e}^{-T_\varepsilon u} du \prod_{j=1}^{k} \mu_\varepsilon(d\theta_j^*)^{n_j}. \tag{B.1}$$

The posterior distribution of μ_ε can be characterized by its Laplace functional; we have

$$\begin{aligned} &\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\tau)\mu_\varepsilon(d\tau)} | \theta_1 \in d\theta_1^*, \ldots, \theta_n \in d\theta_n, U \in du\right) \\ &\quad = \frac{\mathbb{E}\left\{\mathrm{e}^{-\int_\Theta f(\tau)\mu_\varepsilon(d\tau)} \mathbb{P}(\theta_1 \in d\theta_1^*, \ldots, \theta_n \in d\theta_n, U \in du | \mu_\varepsilon)\right\}}{\mathbb{E}\left\{\mathbb{P}(\theta_1 \in d\theta_1^*, \ldots, \theta_n \in d\theta_n, U \in du | \mu_\varepsilon)\right\}}. \end{aligned} \tag{B.2}$$

Let us focus on the numerator in (B.2); by (B.1) we obtain:

$$\begin{aligned} &\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\tau)\mu_\varepsilon(d\tau)} \mathbb{P}(\theta_1 \in d\theta_1^*, \ldots, \theta_n \in d\theta_n, U \in du | \mu_\varepsilon)\right) \\ &= \frac{u^{n-1} du}{\Gamma(n)} \mathbb{E}\left(\mathrm{e}^{-J_0(f(\tau_0)+u)} \mathrm{e}^{-\int_\Theta (f(\tau)+u)\tilde{\mu}_\varepsilon(d\tau)} \prod_{j=1}^{k} (\tilde{\mu}_\varepsilon(d\theta_j^*) + J_0 \delta_{\tau_0}(d\theta^*))^{n_j}\right). \end{aligned} \tag{B.3}$$

Moreover, if P_0 is an absolutely continuous probability, then, for each $j = 1, \ldots, k$,

$$(\tilde{\mu}_\varepsilon(d\theta_j^*) + J_0 \delta_{\tau_0}(d\theta^*))^{n_j} = \tilde{\mu}_\varepsilon(d\theta_j^*)^{n_j} + J_0^{n_j} \delta_{\tau_0}(d\theta_j^*),$$

so that

$$\prod_{j=1}^{k}(\tilde{\mu}_\varepsilon(d\theta_j^*)^{n_j}+J_0^{n_j}\delta_{\tau_0}(d\theta^*))=\prod_{j=1}^{k}\tilde{\mu}_\varepsilon(d\theta^*)^{n_j}+\sum_{l=1}^{k}\delta_{\tau_0}(d\theta_l^*)J_0^{n_l}\prod_{j\neq l}\tilde{\mu}_\varepsilon(d\theta^*)^{n_j}.$$

Therefore, the expected value on the right hand side of (B.3) is:

$$\begin{aligned}&\mathbb{E}\left(\mathrm{e}^{-J_0(f(\tau_0)+u)}\right)\mathbb{E}\left\{\mathrm{e}^{-\int_\Theta f(\tau)+u\tilde{\mu}_\varepsilon(d\tau)}\prod_{j=1}^{k}\tilde{\mu}_\varepsilon(d\theta_j^*)^{n_j}\right\}\\&+\sum_{l=1}^{k}\mathbb{E}(\mathrm{e}^{-J_0(f(\tau_0)+u)}J_0^{n_l}\delta_{\tau_0}(d\theta_l^*))\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\tau)+u\tilde{\mu}_\varepsilon(d\tau)}\prod_{j\neq l}\tilde{\mu}_\varepsilon(d\theta_j^*)^{n_j}\right).\end{aligned}$$

Representation (2.1) can be extended to $\tilde{\mu}_\varepsilon(d\theta_j^*)^{n_j}=\int_{\mathbb{R}^+\times\Theta}s^{n_j}\delta_\tau(d\theta_j^*)N(ds,d\tau)$ where N is a Poisson process with mean intensity $\nu_\varepsilon(ds,d\tau)$. If we apply Palm's formula [see 13, Proposition 13.1.IV] to $\tilde{\mu}_\varepsilon(d\theta_k^*)^{n_k}$, we have that

$$\begin{aligned}&\mathbb{E}\left\{\mathrm{e}^{-\int_\Theta(f(\tau)+u)\tilde{\mu}_\varepsilon(d\tau)}\prod_{j=1}^{k}\tilde{\mu}_\varepsilon(d\theta_j^*)^{n_j}\right\}\\&=\mathbb{E}\left\{\mathrm{e}^{-\int_\Theta(f(\tau)+u)\tilde{\mu}_\varepsilon(d\tau)}\prod_{j=1}^{k-1}\tilde{\mu}_\varepsilon(d\theta_j^*)^{n_j}\int_{\mathbb{R}^+\times\Theta}s_k^{n_k}\delta_{\tau_k}(d\theta_k^*)N(ds_k,d\tau_k)\right\}\\&=\mathbb{E}\left\{\mathrm{e}^{-\int_\Theta(f(\tau)+u)(\tilde{\mu}_\varepsilon)(d\tau)}\prod_{j=1}^{k-1}\tilde{\mu}_\varepsilon(d\theta_j^*)^{n_j}\right\}P_0(d\theta_k^*)\int_\varepsilon^\infty\mathrm{e}^{-(f(\theta_k^*)+u)s_k}s_k^{n_k}\kappa\rho(s_k)ds_k\end{aligned}$$

(by iterating again Palm's formula $k-1$ times)

$$\begin{aligned}&=\mathbb{E}\left\{\mathrm{e}^{-\int_\Theta(f(\tau)+u)(\tilde{\mu}_\varepsilon)(d\tau)}\right\}\prod_{j=1}^{k}\left(P_0(d\theta_j^*)\int_\varepsilon^\infty\mathrm{e}^{-(f(\theta_j^*)+u)s_j}s_j^{n_j}\kappa\rho(s_j)ds_j\right)\end{aligned}$$

(by(2.3))

$$\begin{aligned}&=\exp\left\{-\int_{\mathbb{R}^+\times\Theta}\left(1-\mathrm{e}^{-s(f(\tau)+u)}\right)\nu_\varepsilon(ds,d\tau)\right\}\\&\quad\times\prod_{j=1}^{k}P_0(d\theta_j^*)\int_\varepsilon^\infty\mathrm{e}^{-(f(\theta_j^*)+u)s_j}s_j^{n_j}\kappa\rho(s_j)ds_j.\end{aligned}$$

In other words, the numerator of (B.2) is equal to

$$\begin{aligned}&\frac{u^{n-1}}{\Gamma(n)}\frac{\int_{\mathbb{R}^+\times\Theta}\mathrm{e}^{-s(f(\tau)+u)}\nu_\varepsilon(ds,d\tau)+k}{\Lambda_\varepsilon}\mathrm{e}^{\left\{-\int_{\mathbb{R}^+\times\Theta}\left(1-\mathrm{e}^{-s(f(\tau)+u)}\right)\nu_\varepsilon(ds,d\tau)\right\}}\\&\qquad\times\prod_{j=1}^{k}P_0(d\theta_j^*)\int_\varepsilon^\infty\mathrm{e}^{-(f(\theta_j^*)+u)s}s^{n_j}\kappa\rho(s)ds.\end{aligned}\tag{B.4}$$

Observe that, if we plug the function $f \equiv 0$ in (B.4), we obtain the denominator of the ratio (B.2), that is

$$\begin{aligned}&\mathbb{P}(\theta_1 \in d\theta_1, \ldots, \theta_n \in d\theta_n, U \in du)\\ &\quad = \frac{u^{n-1}}{\Gamma(n)} \frac{\Lambda_{\varepsilon,u} + k}{\Lambda_\varepsilon} \mathrm{e}^{(\Lambda_{\varepsilon,u} - \Lambda_\varepsilon)} \prod_{j=1}^{k} P_0(d\theta_j^*) k_\varepsilon(u, n_j),\end{aligned} \tag{B.5}$$

where for $n > 0$, $k_\varepsilon(u,n) = \int_\varepsilon^\infty \mathrm{e}^{-us} s^n \kappa\rho(s) ds = (-1)^n \frac{d}{du^n} \psi_\varepsilon(u)$, and $\psi_\varepsilon(u) := -\log\left(\mathbb{E}(\mathrm{e}^{-u\tilde{T}_\varepsilon})\right) = \Lambda_\varepsilon - \Lambda_{\varepsilon,u}$.

We are ready to compute the posterior Laplace functional of μ_ε: by substituting (B.4) and (B.5) in the numerator and denominator of (B.2), we have

$$\begin{aligned}&\mathbb{E}\left(\mathrm{e}^{-\int_\Theta f(\tau)\mu_\varepsilon(d\tau)} | \theta_1 \in d\theta_1, \ldots, \theta_n \in d\theta_n, U \in du\right)\\ &= \left\{ \frac{\int_{\mathbb{R}^+\Theta} \mathrm{e}^{-sf(\tau)} \mathrm{e}^{-su} \nu_\varepsilon(ds, d\tau) + k}{\Lambda_{\varepsilon,u} + k} \ \mathrm{e}^{\left(-\int_{\mathbb{R}\times\Theta} (1 - \mathrm{e}^{-sf(\tau)}) \mathrm{e}^{-su} \nu_\varepsilon(du, d\tau)\right)} \right\}\\ &\quad \times \left(\prod_{j=1}^{k} \int_0^\infty \mathrm{e}^{-sf(\theta_j^*)} \frac{e^{-su} s^{n_j} \rho(s) \mathbb{I}_{(\varepsilon,\infty)}(s)}{k_\varepsilon(u, n_j)} ds \right).\end{aligned} \tag{B.6}$$

This expression gives that the posterior Laplace functional of μ_ε, conditionally to $U \in du$, factorizes in two terms. This proves the independence property in point 3. We denote the unnormalized process of non-allocated jumps by $\mu_{u,\varepsilon}^{(na)}$. Its conditional Laplace transform is given by the first factor (between $\{\}$) in the right hand side of (B.6). In order to obtain point 1. of the theorem, characterization (3.7) gives that the law of $\mu_{u,\epsilon}^{(na)}$ coincides with the law of a process $\mu^\star$ as given in (3.6), with (exponential tilted) Lévy intensity $\mathrm{e}^{-su}\nu_\epsilon(ds, d\tau)$ and probability of success of the Bernoulli mixing random variable $p = \frac{\Lambda_{\epsilon,u}}{k + \Lambda_{\epsilon,u}}$. As far as point 2. is concerned, the Laplace functional (B.6) gives that the process of the allocated jumps has fixed atoms at the observed unique values $\theta_1^*, \ldots, \theta_k^*$, i.e. it can be represented as

$$\mu_\epsilon^{(a)}(\cdot) = \sum_{j=1}^{k} J_j^{(a)} \delta_{\theta_j^*}(\cdot).$$

In this case, the weigths of the allocated masses $J_j^{(a)}$ are independent and distributed according to

$$P(J_j^{(a)} \in ds | \theta_1 \in d\theta_1, \ldots, \theta_n \in d\theta_n, U \in du) = \frac{e^{-su} s^{n_j} \rho(s) \mathbb{I}_{(\epsilon,\infty)}(s)}{k_\epsilon(u, n_j)} ds,$$

for any $j = 1, \ldots, k$. Finally, point 4. follows easily from (B.5).

B.2. Proof of Proposition 3.1

This proposition follows from (B.5). In fact, we first observe that $\mathbb{P}(\theta_1 \in d\theta_1, \ldots, \theta_n \in d\theta_n, U \in du) = \mathbb{P}(\boldsymbol{p}_n, \theta_1^* \in d\theta_1^*, \ldots, \theta_k^* \in d\theta_k^*, U \in du)$, and then integrate out $\theta_1^*, \ldots, \theta_k^*$ and U from (B.5) to obtain (3.9).

B.3. Proof of Proposition 3.2

By Proposition 3.1, $p_\varepsilon(n_1, \ldots, n_k) = \int_0^{+\infty} f_\varepsilon(u; n_1, \ldots, n_k)du$, where

$$f_\varepsilon(u; n_1, \ldots, n_k) = \frac{u^{n-1}}{\Gamma(n)} \frac{(k + \Lambda_{\varepsilon,u})}{\Lambda_\varepsilon} e^{(\Lambda_{\varepsilon,u} - \Lambda_\varepsilon)} \prod_{i=1}^{k} \int_\varepsilon^{+\infty} \kappa s^{n_i} \mathrm{e}^{-us} \rho(s)ds, \quad \text{(B.7)}$$

with $u > 0$. On the other hand, the eppf of a $NormCRM(\rho, \kappa P_0)$ can be written as $p_0(n_1, \ldots, n_k) = \int_0^{+\infty} f_0(u; n_1, \ldots, n_k)du$, where

$$f_0(u; n_1, \ldots, n_k) = \frac{u^{n-1}}{\Gamma(n)} \exp\left\{\kappa \int_0^{+\infty} (\mathrm{e}^{-us} - 1)\rho(s)ds\right\} \prod_{i=1}^{k} \int_0^{+\infty} \kappa s^{n_i} \mathrm{e}^{-us} \rho(s),$$

with $u > 0$. We first show that

$$\lim_{\varepsilon \to 0} f_\varepsilon(u; n_1, \ldots, n_k) = f(u; n_1, \ldots, n_k) \quad \text{for any } u > 0. \quad \text{(B.8)}$$

In particular, we have that

$$\lim_{\varepsilon \to 0} \int_\varepsilon^{+\infty} s^{n_i} \mathrm{e}^{-us} \rho(s)ds = \int_0^{+\infty} s^{n_i} \mathrm{e}^{-us} \rho(s)ds$$

and

$$\lim_{\varepsilon \to 0} \mathrm{e}^{\Lambda_{\varepsilon,u} - \Lambda_\varepsilon} = \exp\left\{\kappa \int_0^{+\infty} (\mathrm{e}^{-us} - 1)\rho(s)ds\right\},$$

being this limit finite for any $u > 0$. Using standard integrability criteria, it is straightforward to check that, for any $u > 0$, $\lim_{\varepsilon \to 0} \Lambda_{\varepsilon,u} = \lim_{\varepsilon \to 0} \Lambda_\varepsilon = +\infty$ and they are equivalent infinite, i.e.

$$\lim_{\varepsilon \to 0} \frac{k + \Lambda_{\varepsilon,u}}{\Lambda_\varepsilon} = \lim_{\varepsilon \to 0} \frac{\Lambda_{\varepsilon,u}}{\Lambda_\varepsilon} = 1.$$

We can therefore conclude that (B.8) holds true. The rest of the proof follows as in the second part of the proof of Lemma 2 in [3], where we prove that (i) $\lim_{\varepsilon \to 0} \sum_{\mathcal{C} \in \Pi_n} p_\varepsilon(n_1, \ldots, n_k) = 1$; (ii) $\liminf_{\varepsilon \to 0} p_\varepsilon(n_1, \ldots, n_k) = p_0(n_1, \ldots, n_k)$ for all $\mathcal{C} = (C_1, \ldots, C_k) \in \Pi_n$, the set of all partitions of $\{1, 2, \ldots, n\}$; (iii) $\sum_{\mathcal{C} \in \Pi_n} p_0(n_1, \ldots, n_k) = 1$. By Lemma 1 in [3], equation (3.11) follows.

B.4. Proof of formula 3.12

First of all, observe that

$$\begin{aligned}\left(x_1+\cdots+x_{N_\varepsilon^*}\right)^m &= \sum_{\substack{m_1+\cdots+m_{N_\varepsilon^*}=m\\ m_1,\ldots,m_{N_\varepsilon^*}\geq 0}} \binom{m}{m_1,\ldots,m_{N_\varepsilon^*}} \prod_{j=1}^{N_\varepsilon^*} x_j^{m_j} \\ &= \sum_{k=1}^{m} \mathbb{I}_{\{1,\ldots,N_\varepsilon^*\}}(k)\frac{1}{k!} \sum_{\substack{n_1+\cdots+n_k=m\\ n_j=1,2,\ldots}} \binom{m}{n_1,\ldots,n_k} \left(\sum_{j_1,\ldots,j_k}\prod_{i=1}^{k} x_{j_i}^{n_i}\right) \qquad \text{(B.9)}\end{aligned}$$

where $N_\varepsilon^* = N_\varepsilon + 1$, $x_j^0 = 1$ for all $x_j \geq 0$, and the last summation is over all positive integers, being (B.9) the multinomial theorem. The second equality follows straightforward from different identifications of the set of all partitions of m [see 40, Section 1.2]. Therefore, for any $B \in \mathcal{B}(\Theta)$, $m = 1, 2, \ldots$, we have (here, instead of P_0 and τ_0 as in (3.4), there are $P_{N_\varepsilon^*}$ and $\tau_{N_\varepsilon^*}$):

$$\begin{aligned}&\mathbb{E}(P_\varepsilon(B)^m) \\ &= \mathbb{E}\left(\mathbb{E}\left((\sum_{j=1}^{N_\varepsilon^*} P_j\delta_{\tau_j}(B))^m|N_\varepsilon\right)\right) \\ &= \mathbb{E}\left(\mathbb{E}\left(\sum_{\substack{m_1+\cdots+m_{N_\varepsilon^*}=m\\ m_1,\ldots,m_{N_\varepsilon^*}\geq 0}} \binom{m}{m_1,\ldots,m_{N_\varepsilon^*}} \prod_{j=1}^{N_\varepsilon^*}(P_j\delta_{\tau_j}(B))^{m_j}|N_\varepsilon\right)\right) \\ &= \mathbb{E}\left(\sum_{k=1}^{m}\mathbb{I}_{\{1,\ldots,N_\varepsilon^*\}}(k)\frac{1}{k!}\sum_{\substack{n_1+\cdots n_k=m\\ n_j=1,2,\ldots}}\binom{m}{n_1,\ldots,n_k}\right. \\ &\qquad \left.\times \sum_{j_1,\ldots,j_k}\mathbb{E}(\prod_{i=1}^{k}P_{j_i}^{n_i}|N_\varepsilon)\prod_{i=1}^{k}\mathbb{E}(\delta_{\tau_j}(B)|N_\varepsilon)\right) \\ &= \mathbb{E}\left(\sum_{k=1}^{m}\mathbb{I}_{\{1,\ldots,N_\varepsilon^*\}}(k)\frac{1}{k!}\sum_{\substack{n_1+\cdots+n_k=m\\ n_j=1,2,\ldots}}\binom{m}{n_1,\ldots,n_k}p_\varepsilon(n_1,\ldots,n_k)(P_0(B))^k\right).\end{aligned}$$

We identify this last expression as $\mathbb{E}\left(\sum_{k=1}^{m} P_0(B)^k\mathbb{P}(K_m = k|N_\varepsilon)\right)$, where K_m is the number of distinct values in a sample of size m from P_ε. Hence, we have proved that

$$\mathbb{E}(P_\varepsilon(B)^m) = \mathbb{E}\left(\mathbb{E}(P_0(B)^{K_m}|N_\varepsilon)\right) = \mathbb{E}\left(P_0(B)^{K_m}\right).$$

B.5. Proof of formula 3.14

Suppose that $B_1, B_2 \in \mathcal{B}(\Theta)$ are disjoint. Therefore

$$\begin{aligned}\mathbb{E}(P_\varepsilon(B_1)P_\varepsilon(B_2)) &= \mathbb{E}\left(\mathbb{E}\left(\sum_{j=1}^{N_\varepsilon^*} P_j\delta_{\tau_j}(B_1)\sum_{l=1}^{N_\varepsilon^*} P_l\delta_{\tau_l}(B_2)|N_\varepsilon\right)\right)\\ &= \mathbb{E}\left(\sum_{\substack{l\neq j\\ j,l=1,\dots,N_\varepsilon^*}} \mathbb{E}(P_jP_l|N_\varepsilon)\mathbb{E}(\delta_{\tau_j}(B_1))\mathbb{E}(\delta_{\tau_l}(B_2)))\right)\\ &= \mathbb{E}\left(P_0(B_1)P_0(B_2)\sum_{\substack{l\neq j\\ j,l=1,\dots,N_\varepsilon^*}} \mathbb{E}(P_jP_l|N_\varepsilon)\right) = P_0(B_1)P_0(B_2)p_\varepsilon(1,1).\end{aligned}$$

The general case when B_1 and B_2 are not disjoint follows easily:

$$\begin{aligned}\mathbb{E}(P_\varepsilon(B_1)P_\varepsilon(B_2)) &= \mathbb{E}\left((P_\varepsilon(B_1\cap B_2))^2\right) + \mathbb{E}\left(P_\varepsilon(B_1\setminus B_2)P_\varepsilon(B_1\cap B_2)\right)\\ &\quad + \mathbb{E}\left(P_\varepsilon(B_2\setminus B_1)P_\varepsilon(B_1\cap B_2)\right) + \mathbb{E}(P_\varepsilon(B_1\setminus B_2)P_\varepsilon(B_2\setminus B_1)),\end{aligned}$$

where now the sets are disjoint. Applying the result above we first find that

$$\mathbb{E}(P_\varepsilon(B_1)P_\varepsilon(B_2)) = p_\varepsilon(2)P_0(B_1\cap B_2) + (1-p_\varepsilon(2))P_0(B_1)P_0(B_2),$$

and consequently formula 3.14 holds true.

B.6. Proof of Proposition 6.1

The eppf of the Dirichlet process appeared first in [1] [see 38]; anyhow, it is straightforward to derive it from (3.10):

$$\begin{aligned}p_D(n_1,\dots,n_k;\kappa) &= \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)}\mathrm{e}^{-\kappa\log\frac{u+\omega}{\omega}}\prod_{j=1}^k \kappa\frac{\Gamma(n_j)}{(u+\omega)^{n_j}}du\\ &= \kappa^k\int_0^{+\infty}\frac{u^{n-1}}{\Gamma(n)}\left(\frac{\omega}{\omega+u}\right)^\kappa\frac{1}{(u+\omega)^n}\prod_{j=1}^k\Gamma(n_j)du = \frac{\Gamma(\kappa)}{\Gamma(\kappa+n)}\kappa^k\prod_{j=1}^k\Gamma(n_j)\end{aligned}$$

where the last equality follows from formula (3.194.3) in [22]. By definition of the hypergeometric function, we have

$$1 \le {}_2F_1\left(\frac{n_j}{2},\frac{n_j+1}{2};1;\frac{1}{(u+\omega)^2}\right) \le {}_2F_1\left(\frac{n_j}{2},\frac{n_j+1}{2};1;\frac{1}{\omega^2}\right).$$

Moreover

$$\frac{\omega+\sqrt{\omega^2-1}}{(u+\omega)+\sqrt{((u+\omega)^2-1)}} = \frac{\omega}{u+\omega}\frac{1+\sqrt{1-1/\omega^2}}{1+\sqrt{1-1/(u+\omega)^2}}$$

and

$$\frac{1+\sqrt{1-1/\omega^2}}{2} \leq \frac{1+\sqrt{1-1/\omega^2}}{1+\sqrt{1-1/(u+\omega)^2}} \leq 1,$$

so that

$$\left(\frac{1+\sqrt{1-1/\omega^2}}{2}\right)^{\kappa} p_D(n_1,\ldots,n_k;\kappa) \leq p_B(n_1,\ldots,n_k;\omega,\kappa)$$
$$\leq \prod_{j=1}^{k} {}_2F_1\left(\frac{n_j}{2}, \frac{n_j+1}{2}; 1; \frac{1}{\omega^2}\right) p_D(n_1,\ldots,n_k;\kappa).$$

The left hand-side of these inequalities obviously converges to $p_D(n_1,\ldots,n_k;\kappa)$ as ω goes to $+\infty$. On the other hand,

$${}_2F_1\left(\frac{n_j}{2}, \frac{n_j+1}{2}; 1; \frac{1}{\omega^2}\right) \to 1 \text{ as } \omega \to +\infty,$$

thanks to the uniform convergence of the hypergeometric series ${}_2F_1(\frac{n_j}{2}, \frac{n_j+1}{2}; 1; z)$ on a disk of radius smaller that 1. We conclude that, for any $n_1,\ldots,n_k$ such that $n_1+\cdots+n_k=n$, $k=1,\ldots,n$, and any $\kappa>0$,

$$\lim_{\omega\to+\infty} p_B(n_1,\ldots,n_k;\omega,\kappa) = p_D(n_1,\ldots,n_k;\kappa).$$

References

[1] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*, 1152–1174. MR0365969

[2] Arbel, J. and Prünster, I. (2016). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*, doi: 10.1007/s11222-016-9676-8.

[3] Argiento, R., Bianchini, I. and Guglielmi, A. (2016). A blocked Gibbs sampler for NGG-mixture models via a priori truncation. *Statistics and Computing 26*, 641–661. MR3489862

[4] Argiento, R., Guglielmi, A., Hsiao, C., Ruggeri, F. and Wang, C. (2015). Modelling the association between clusters of SNPs and disease responses. In R. Mitra and P. Mueller (Eds.), *Nonparametric Bayesian Methods in Biostatistics and Bioinformatics*. Springer. MR3382181

[5] Argiento, R., Guglielmi, A. and Pievatolo, A. (2010). Bayesian density estimation and model selection using nonparametric hierarchical mixtures. *Computational Statistics and Data Analysis 54*, 816–832. MR2580918

[6] Asmussen, S. and Glynn, P.W. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, New York, 2007. MR2331321

[7] Barndorff-Nielsen, O.E. (2000). *Probability densities and Lévy densities*. University of Aarhus. Centre for Mathematical Physics and Stochastics.

[8] Barrientos, A.F., Jara, A. and Quintana, F.A. (2012). On the support of MacEacherns dependent Dirichlet processes and extensions. *Bayesian Analysis 7*, 277–310. MR2934952

[9] Barrios, E., Lijoi, A., Nieto-Barajas, L.E. and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statistical Science 28*, 313–334. MR3135535

[10] Basford, K., McLachlan, G. and York, M. (1997). Modelling the distribution of stamp paper thickness via finite normal mixtures: The 1872 Hidalgo stamp issue of Mexico revisited. *Journal of Applied Statistics 24*, 169–180.

[11] Bondesson, L. (1982). On simulation from infinitely divisible distributions. *Advances in Applied Probability 14*, 855–869. MR0677560

[12] Cook, R.D. and Weisberg, S. (1994). *An introduction to regression graphics*. John Wiley & Son. MR1285353

[13] Daley, D.J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: vol. II: general theory and structure*. Springer. MR2371524

[14] De Iorio, M., Johnson, W.O., Müller, P. and Rosner G.L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics 65*, 762–771. MR2649849

[15] Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G. and Bateman, H. (1953). *Higher transcendental functions*, Volume 2. McGraw-Hill New York.

[16] Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*, 577–588. MR1340510

[17] Favaro, S. and Teh, Y. (2013). MCMC for normalized random measure mixture models. *Statistical Science 28*, 335–359. MR3135536

[18] Feller, W. (1971). *An introduction to probability theory and its Applications, vol. II* (Second Edition ed.). John Wiley, New York. MR0270403

[19] Ferguson, T.S. and Klass, M.J. (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics 43*, 1634–1643. MR0373022

[20] Foti, N. and Williamson, S. (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Transactions on pattern Analysis and Machine Intelligence 37*, 359–371.

[21] Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing 24*, 997–1016. MR3253850

[22] Gradshteyn, I. and Ryzhik, L. (2007). *Table of integrals, series, and products - Seventh Edition* (Sixth ed.). San Diego (USA): Academic Press. MR2360010

[23] Griffin, J. and Walker, S.G. (2011). Posterior simulation of normalized random measure mixtures. *Journal of Computational and Graphical Statistics 20*, 241–259. MR2816547

[24] Griffin, J.E. (2013). An adaptive truncation method for inference in Bayesian nonparametric models. *arXiv:1308.2045*. MR3439383

[25] Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc. 96*, 161–173. MR1952729

[26] Ishwaran, H. and James, L.F. (2002). Approximate Dirichlet process computing in finite normal mixtures. *Journal of Computational and Graphical Statistics 11*, 508–532. MR1938445

[27] James, L., Lijoi, A. and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics 36*, 76–97. MR2508332

[28] Jara, A., Hanson, T.E., Quintana, F.A., Müller, P. and Rosner, G.L. (2011). DPpackage: Bayesian semi-and nonparametric modeling in R. *Journal of Statistical Software 40*, 1.

[29] Kingman, J.F.C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society 37*, 1–22. MR0368264

[30] Kingman, J.F.C. (1993). *Poisson processes*, Volume 3. Oxford university press. MR1207584

[31] Lau, J.W. and Green, P.J. (2007). Bayesian model based clustering procedures. *Journal of Computational and Graphical Statistics 16*, 526–558. MR2351079

[32] Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-gaussian priors. *Journal of the American Statistical Association 100*, 1278–1291. MR2236441

[33] Lo, A.J. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics 12*, 351–357. MR0733519

[34] Lomelí, M., Favaro, S. and Teh, Y.W. (2016). A marginal sampler for σ-stable Poisson-Kingman mixture models. *Journal of Computational and Graphical Statistics, Latest articles*.

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

[36] McAuliffe, J.D., Blei, D.M. and Jordan, M.I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Statistics and Computing 16*, 5–14. MR2224185

[37] Nieto-Barajas, L.E. (2013). Lévy-driven processes in bayesian nonparametric inference. *Boletín de la Sociedad Matemática Mexicana (3) 19*. MR3183997

[38] Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. In T. S. Ferguson, L. S. Shapley, and M. J. B. (Eds.), *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, Volume 30 of *IMS Lecture Notes-Monograph Series*, pp. 245–267. Hayward (USA): Institute of Mathematical Statistics. MR1481784

[39] Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: a Festschrift for Terry Speed*, Volume 40 of *IMS Lecture Notes-Monograph Series*, pp. 1–34. Hayward (USA): Institute of Mathematical Statistics. MR2004330

[40] Pitman, J. (2006). *Combinatorial Stochastic Processes – Ecole D'Eté de Probabilités de Saint-Flour XXXII*. New York: Springer. MR2245368

[41] Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of random measures with independent increments. *The Annals of Statistics 31*, 560–585. MR1983542

[42] Rosinski, J. Series representations of lévy processes from the perspective of point processes. In *Lévy processes*, pages 401–415. Springer, 2001. MR1833707

[43] Trippa, L. and Favaro, S. (2012). A class of normalized random measures with an exact predictive sampling scheme. *Scandinavian Journal of Statistics*, 39, 444–460. MR2971631

[44] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research 11*, 3571–3594. MR2756194

[45] Wilson, I. (1983). Add a new dimension to your philately. *The American Philatelist 97*, 342–349.