A blocked Gibbs sampler for NGG-mixture models via a priori truncation

A Bayesian nonparametric model for density and cluster estimation: the ε-NGG process mixture model

Raffaele Argiento[1], Ilaria Bianchini[1] and Alessandra Guglielmi[2]
[1]CNR-IMATI and [2]Politecnico di Milano

Abstract

We define a new class of random probability measures, approximating the well-known normalized generalized gamma (NGG) process. Our new process is defined from the representation of NGG processes as discrete measures where the weights are obtained by normalization of the jumps of a Poisson process, and the support consists of independent identically distributed location points, however considering only jumps larger than a threshold ε. Therefore, the number of jumps of the new process, called ε-NGG process, is a.s. finite. A prior distribution for ε can be elicited. We will assume such a process as the mixing measure in a mixture model for density and cluster estimation. We also build an efficient Gibbs sampler scheme to simulate from the posterior. Finally, the performance of our model on two popular datasets will be illustrated.

Keywords: Bayesian nonparametric mixture models, normalized generalized gamma process, blocked Gibbs sampler, finite dimensional approximation, a priori truncation method.

1 Introduction

The first goal of this work is the definition of a new class of nonparametric priors, which can be considered as an approximation of the distribution of a homogeneous normalized random measure with independent increments, namely the normalized generalized gamma process. Any homogeneous normalized random measure with independent increments (NRMI) can be represented as a discrete random probability measure: the weights are obtained by normalization of the jumps (a countable set) of a Poisson process, while the support consists of a countable number of random points from some distribution. In this case, posterior inference is made difficult by the presence of infinite unknown parameters. NRMIs are a popular tool in a mixture context, where they are usually considered as mixing measures of parametric densities for continuous data, and therefore NRMI mixtures include infinite parameters. There

are two main approaches to deal with this computational problem, namely marginal and conditional Gibbs sampler algorithms for sampling from the posterior. The former integrate out the infinite dimensional parameter (i.e. the random probability), resorting to generalized Polya urn schemes (MacEachern, 1998); see Neal (2000) for a review on the subject. Recently, Favaro and Teh (2013) developed algorithms of both types for mixture models with NRMI mixing measures.

On the other hand, by a conditional algorithm we mean a Gibbs sampler imputing the nonparametric mixing measure and updating it as a component of the algorithm itself. The reference papers on conditional algorithms for Dirichlet process mixture models are Papaspiliopoulos and Roberts (2008) and Walker (2007). The former builds a retrospective algorithm, while the latter proposes a slice sampler algorithm. The slice sampler has been extended to NRMI mixtures in Griffin and Walker (2011). See also Favaro and Walker (2013).

Conditional algorithms are called *truncation* methods here if the infinite parameter (i.e. the mixing measure) is approximated by truncation of the infinite sum defining the process. Truncation can be achieved a posteriori, when one approximates the infinite parameter P given the data, as described in Gelfand and Kottas (2002) for the DPM model. On the other hand, truncation can be applied a priori to approximate the nonparametric mixing distribution with a finite dimensional random probability measure. In this case, a simpler mixture model has to be fitted. In the latter framework, pioneer papers for DPM models are Ishwaran and James (2001) and Ishwaran and Zarepour (2000, 2002). For instance, Ishwaran and James (2001) consider a (blocked) Gibbs sampler for a finite approximation of the stick-breaking prior in order to deal with a finite number of random variables, which are updated in "blocks". Barrios et al. (2013) propose an *a posteriori* truncation algorithm for NMRI mixtures using the Ferguson-Klass representation of completely random measures (Ferguson and Klass, 1972). Of course, when using truncation algorithms, the key-point is the choice of the truncation level; Argiento et al. (2010) propose a simple adaptive truncation method evaluating an upper bound in probability for the jumps excluded from the summation. Recently, an *a priori* truncation method has been introduced by Griffin (2013), who proposes an adaptive truncation algorithm for posterior inference with priors either of stick-breaking or NRMI type.

If we needed a motivation for conditional algorithms, with or without truncation, we should keep in mind that they are able to provide a full Bayesian analysis. On the other hand, as pointed out in Griffin (2013), there are two motivations for truncation: the study of the properties of the prior distribution, which is not our primary goal, and simpler calculation of posterior inference using these priors. Instead, with regard to theoretical results on approximation of Dirichlet processes based on the distributional equation for a DP given in

Sethuraman (1994), we refer here to Muliere and Tardella (1998) and Favaro et al. (2012).

In this work we introduce a new truncation prior by defining a random probability measure which depends (among the others) on a parameter ε, controlling the degree of approximation of the truncation method. In particular, our prior is a *truncated version* of a normalized generalized gamma (NGG) process (Lijoi et al., 2007), where this new random probability measure is built from the representation of the weights of a NGG process as normalized points of a Poisson process; however, in this representation, we consider only points larger than the threshold ε. We refer to this random probability measure as ε-NGG process. Conditionally on ε, our process is finite dimensional either a priori and a posteriori. To justify our proposal, we show that, for ε going to zero, the finite dimensional ε-NGG prior converges to its infinite dimensional counterpart. As often done in Bayesian Nonparametrics, we will consider this new discrete random probability as the mixing measure in a Gaussian mixture model, which is a very flexible tool for density and cluster estimation problems. A prior distribution for ε can be given, as well as for all the other potential parameters defining the new process. As a second goal of this paper, we design a blocked Gibbs sampler algorithm to simulate from the posterior.

For illustration purposes, we fitted our mixture model to two popular datasets: the Galaxy data, and the Yeast cell cycle data, which is an interesting multivariate dataset consisting of gene expression profiles measured at 9 different times. Density estimates are shown for the two applications, together with a thorough robustness analysis of the estimates with respect to prior choice, in particular to investigate the effect of the approximation parameter ε.

In Section 2 we introduce notation on homogeneous NRMIs, while in Section 3 we define the new ε-NGG process, show convergence in distribution to a NGG process and describe its posterior, given a sample from it. Section 4 introduces ε-NGG mixtures and describes the MCMC algorithm for computing its posterior. Section 5 (Galaxy data) and 6 (Yeast cell cycle data) discuss the two applications. The article ends up with wrap-up of the proposed model as well as with possible future developments in Section 7.

2 Homogeneous normalized random measures

In this section we sketch the basic ingredients to construct homogeneous NRMIs in order to smooth the introduction of our new prior. Further details can be found in James et al. (2009) and Regazzini et al. (2003) and the references therein.

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 completely random measure is defined (see Kingman, 1993,

Section 8.2) by $\mu(\cdot) = \sum_{j\geq 1} J_j \delta_{\tau_j}(\cdot)$, where the $\{(J_j, \tau_j)\}_{j\geq 1}$ are the points of a Poisson process on $\mathbb{R}^+ \times \Theta$. We denote by $\nu(ds, d\tau)$ the intensity of the mean measure of such a Poisson process. A completely random measure is homogeneous if $\nu(ds, d\tau) = \rho(s)ds P_0(d\tau)$, where $\rho(s)$ is the density of a non-negative measure on $\mathbb{R}^+$, while P_0 is a probability measure on Θ. If μ is homogeneous, the support points, that is $\{\tau_j\}$, and the jumps of μ, $\{J_j\}$, 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. Furthermore, we assume that ρ satisfies the following regularity conditions:

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

If $T := \mu(\Theta) = \sum_{j>1} J_j$, the former condition in (1) guarantees that $P(T < +\infty) = 1$, while the latter yields $P(T = 0) = 0$. Therefore, a random probability measure (r.p.m.) P can be defined through normalization of μ:

$$P := \frac{\mu}{\mu(\Theta)} = \sum_{j=1}^{\infty} \frac{J_j}{T}\ \delta_{\tau_j} = \sum_{j=1}^{\infty} P_i \delta_{\tau_j}\ . \tag{2}$$

Following James et al. (2009) we refer to P in (2) as a *homogeneous normalized random measure with independent increments* (HNRMI). The definition of HNRMIs appeared in Regazzini et al. (2003) first. An alternative construction of HNRMI can be given in terms of Poisson-Kingman models as in Pitman (2003).

In particular, in this paper we are going to propose a new r.p.m. on the ground of a HNRMI, namely the normalized generalized gamma process, introduced in Lijoi et al. (2007). We use the same notation as in Argiento et al. (2010). By a NGG$(\sigma, \kappa, \omega, P_0)$ process P we denote the HNRMI as in (2) where the mean intensity of the Poisson process defining the jumps is $\rho(s) = (\kappa/\Gamma(1-\sigma))\, s^{-1-\sigma} \mathrm{e}^{-s\omega} \mathbb{I}_{(0,+\infty)}(s)$, and $0 \leq \sigma \leq 1$, $\kappa, \omega \geq 0$. This parametrization is not unique, as the scaling property in Pitman (2003) shows, since $(\sigma, \kappa, \omega, P_0)$ and $(\sigma, s^\sigma \kappa, \omega/s, P_0)$, for any $s > 0$, give the same distribution for P. When $\omega = 1$ and $\sigma = 0$, the Dirichlet process (DP) is recovered.

One of the main reasons in favour of NGG process, instead of DP, is its higher flexibility in clustering. For instance, when considering a sample of size n from a NGG process, the distribution of the number K_n of distinct values in the sample has a further degree of freedom, σ, which tunes its variance, contrary to the DP case, where the distribution of K_n can be highly peaked. The parameter σ also drives a richer reinforcement mechanism in the predictive distributions of the sample. Moreover, NGG processes are of Gibbs-type, a class of r.p.m.'s which stands out for their mathematical tractability (see Lijoi et al., 2008).

Recent works that include NGG processes as an ingredient in their models are Caron (2012) and Caron and Fox (2014), both on statistical networks: the former for bipartite

random graphs, while the latter for sparse and exchangeable random graphs. Griffin et al. (2013) and Lijoi et al. (2014) propose a vector of dependent NGG processes for comparing distributions. See also Chen et al. (2012) for an application of such multivariate priors in a dynamic topic modeling context.

3 ε-NGG processes

The goal of this section is the definition of a finite dimensional random probability measure that is an approximation of the NGG process with parameters $(\sigma, \kappa, \omega, P_0)$, introduced above. The idea is the following: it is straightforward to show that, for any $\varepsilon > 0$, all the jumps $\{J_j\}$ of μ larger then a threshold ε are still a Poisson process, with mean intensity $\tilde{\rho}_\varepsilon(s) := \rho(s)\mathbb{I}_{(\varepsilon,+\infty)}(s)$. Moreover, the total number of these points is Poisson distributed, i.e. $N_\varepsilon \sim \mathcal{P}_0(\Lambda_\varepsilon)$ where

$$\Lambda_\varepsilon := \int_\varepsilon^{+\infty} \rho(x)dx = \frac{\kappa\omega^\sigma}{\Gamma(1-\sigma)}\Gamma(-\sigma, \omega\varepsilon),$$

where $\Gamma(a,x) = \int_x^{+\infty} t^{a-1}\mathrm{e}^{-t}dt$ is the incomplete gamma function. Since $\Lambda_\varepsilon < +\infty$ for any $\varepsilon > 0$, 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

$$\rho_\varepsilon(s) = \frac{1}{\omega^\sigma\Gamma(-\sigma, \omega\varepsilon)} s^{-\sigma-1}\mathrm{e}^{-\omega s}\mathbb{I}_{(\varepsilon,\infty)}(s).$$

This is the well-known relationship between Poisson and Bernoulli processes; see, for instance, Kingman (1993), Section 2.4. However, in this case, while $\mathbb{P}(\sum_{j=1}^{N_\varepsilon} J_j < \infty) = 1$, the condition on the right of (1) is not satisfied, so that $\mathbb{P}(\sum_{j=1}^{N_\varepsilon} J_j = 0) > 0$, or, in other terms, $\mathbb{P}(N_\varepsilon = 0) > 0$ for any $\varepsilon > 0$. To overcome this problem, we add one more point J_0, independent on the previous J_js, but identically distributed, so that we consider $N_\varepsilon+1$ iid points $\{J_0, J_1, \ldots, J_{N_\varepsilon}\}$. We are ready to define an $\varepsilon-$NGG process as:

$$P_\varepsilon = \sum_{j=0}^{N_\varepsilon} P_j\delta_{\tau_j} = \frac{1}{T_\varepsilon}\sum_{j=0}^{N_\varepsilon} J_j\delta_{\tau_j}, \tag{3}$$

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) by $\varepsilon-NGG(\sigma, \kappa, \omega, P_0)$ *process.*

Observe that P_ε is a proper species sampling model (Pitman, 1996) with a random number $N_\varepsilon + 1$ of different species. In particular we observe that if $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n)$ is a finite sample from a species sampling model P, its marginal law has unique characterization in term of its unique distinct values $\boldsymbol{\theta}^* := (\theta_1^*, \ldots, \theta_k^*)$ and its exchangeable partition $\boldsymbol{p}_n$ as follows:

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

where p is the exchangeable partition probability function (eppf) associated to P and n_i is the number of elements of $\boldsymbol{\theta}$ equal to θ_i^* for $1 \leq i \leq k$. See Pitman (1996).

Coming back to P_ε defined in (3), by formula (30) in Pitman (1996), we have that the eppf corresponding to P_ε is such that

$$p_\varepsilon(n_1, \ldots, n_k) = \sum_{j_1, \ldots, j_k} \mathbb{E}\left(\prod_{i=1}^{k} P_{j_i}^{n_i}\right), \tag{4}$$

where $(j_1, \ldots, j_k)$ ranges over all permutations of k positive integers. It is useful to introduce the following notation: the random vector $(\theta_1, \ldots, \theta_n)$ 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 $j = 1, \ldots, k$. In particular $\#C_i = n_i$ for $1 \leq i \leq k$, and the eppf p can be viewed as a probability law on the set of the partitions of $\mathbb{N}_n$.

The following proposition provides an expression for the eppf of the $\varepsilon-$NGG process.

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

$$\begin{aligned} p_\varepsilon(n_1, \ldots, n_k) = \int_0^{+\infty} & \frac{1}{\Gamma(n)} u^{n-1} (u+\omega)^{k\sigma - n} \prod_{i=1}^{k} \Gamma(n_i - \sigma, (u+\omega)\varepsilon) \\ & \times \frac{\kappa^{k-1}}{\Gamma(1-\sigma)^{k-1}} \frac{\Lambda_{\varepsilon,u} + k}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} \exp\left\{\Lambda_{\varepsilon,u} - \Lambda_\varepsilon\right\} du, \end{aligned} \tag{5}$$

where

$$\Lambda_{\varepsilon,u} := \int_\varepsilon^\infty \rho_{\varepsilon,u}(x) dx = \frac{\kappa(u+\omega)^\sigma}{\Gamma(1-\sigma)} \Gamma(-\sigma, (u+\omega)\varepsilon) \tag{6}$$

with

$$\rho_{\varepsilon,u}(x) = \frac{\kappa}{\Gamma(1-\sigma)} x^{-1-\sigma} \mathrm{e}^{-(\omega+u)x} \mathbb{I}_{(0,\infty)}(x). \tag{7}$$

Proof. First observe that, since N_ε has a Poisson distribution with parameter Λ_ε, we have

$$p_\varepsilon(n_1, \ldots, n_k) = \sum_{N_\varepsilon = 0}^{+\infty} p_\varepsilon(n_1, \ldots, n_k | N_\varepsilon) \frac{\Lambda_\varepsilon^{N_\varepsilon}}{N_\varepsilon!} \mathrm{e}^{-\Lambda_\varepsilon}. \tag{8}$$

Then, (4) yields

$$p_\varepsilon(n_1, \ldots, n_k | N_\varepsilon) = \mathbb{I}_{\{1, \ldots, N_\varepsilon + 1\}}(k) \sum_{j_1, \ldots, j_k} \mathbb{E}\left(\prod_{i=1}^{k} P_{j_i}^{n_i}\right),$$

where the vector $(j_1,\dots,j_k)$ ranges over all permutations of k elements in $\{0,\dots,N_\varepsilon\}$. Then, using the gamma function identity,

$$\frac{1}{T_\varepsilon^n} = \int_0^{+\infty} \frac{1}{\Gamma(n)} u^{n-1} e^{-uT_\varepsilon} du, \tag{9}$$

we have:

$$\begin{aligned}
p_\varepsilon(n_1,..,n_k|N_\varepsilon) &= \mathbb{I}_{\{1,\dots,N_\varepsilon+1\}}(k) \sum_{j_1,\dots,j_k} \int \prod_{i=1}^k \frac{J_{j_i}^{n_i}}{T_\varepsilon^{n_i}} \mathcal{L}(dJ_0,\dots,dJ_{N_\varepsilon}) \\
&= \mathbb{I}_{\{1,\dots,N_\varepsilon+1\}}(k) \sum_{j_1,\dots,j_k} \int \int_0^{+\infty} \frac{1}{\Gamma(n)} u^{n-1} \mathrm{e}^{-uT_\varepsilon} du \prod_{i=1}^k J_{j_i}^{n_i} \mathcal{L}(dJ_0,\dots,dJ_{N_\varepsilon}) \\
&= \mathbb{I}_{\{1,\dots,N_\varepsilon+1\}}(k) \sum_{j_1,\dots,j_k} \int_0^{+\infty} du \left(\frac{1}{\Gamma(n)} u^{n-1} \prod_{i=1}^k \int_0^{+\infty} J_{j_i}^{n_i} \mathrm{e}^{-J_{j_i}u} \rho_\varepsilon(J_{j_i}) dJ_{j_i} \right. \\
&\qquad \left. \times \prod_{j \notin \{j_1,\dots,j_k\}} \int_0^{+\infty} \mathrm{e}^{-J_j u} \rho_\varepsilon(J_j) dJ_j \right) \\
&= \mathbb{I}_{\{1,\dots,N_\varepsilon+1\}}(k) \sum_{j_1,\dots,j_k} \int_0^{+\infty} du \left(\frac{1}{\Gamma(n)} u^{n-1} \prod_{i=1}^k \int_\varepsilon^{+\infty} \frac{J_{j_i}}{\omega^\sigma \Gamma(-\sigma,\omega\varepsilon)} J_{j_i}^{-\sigma-1} \mathrm{e}^{-(\omega+u)J_{j_i}} dJ_{j_i} \right. \\
&\qquad \left. \times \prod_{j \notin \{j_1,\dots,j_k\}} \int_0^{+\infty} \frac{1}{\omega^\sigma \Gamma(-\sigma,\omega\varepsilon)} J_j^{-\sigma-1} \mathrm{e}^{-(\omega+u)J_j} dJ_j \right) \\
&= \mathbb{I}_{\{1,\dots,N_\varepsilon+1\}}(k) \sum_{j_1,\dots,j_k} \int_0^{+\infty} du \left(\frac{1}{\Gamma(n)} u^{n-1} \prod_{i=1}^k \frac{(u+\omega)^{\sigma-n_i} \Gamma(n_i-\sigma,(u+\omega)\varepsilon)}{\omega^\sigma \Gamma(-\sigma,\omega\varepsilon)} \right. \\
&\qquad \left. \times \left(\frac{(u+\omega)^\sigma \Gamma(-\sigma,(u+\omega)\varepsilon)}{\omega^\sigma \Gamma(-\sigma,\omega\varepsilon)} \right)^{N_\varepsilon+1-k} \right).
\end{aligned}$$

If we switch the finite sum and the integral, since the integrand function does not depend on the position of the clusters j_i's, $i=1,\dots,k$, but only on the sizes n_i, and there are $(N_\varepsilon+1)(N_\varepsilon)\dots(N_\varepsilon+1-k) = \dfrac{(N_\varepsilon+1)!}{(N_\varepsilon+1-k)!}$ sequences of k distinct elements from $\{0,\dots,N_\varepsilon\}$, we get:

$$\begin{aligned}
p_\varepsilon(n_1,\dots,n_k|N_\varepsilon) = \mathbb{I}_{\{1,\dots,N_\varepsilon+1\}}(k) \int_0^{+\infty} du \left(\frac{1}{\Gamma(n)} u^{n-1} \frac{(N_\varepsilon+1)!}{(N_\varepsilon+1-k)!} \right. \\
\left. \times \prod_{i=1}^k \frac{(u+\omega)^{\sigma-n_i} \Gamma(n_i-\sigma,(u+\omega)\varepsilon)}{\omega^\sigma \Gamma(-\sigma,\omega\varepsilon)} \left(\frac{(u+\omega)^\sigma \Gamma(-\sigma;(u+\omega)\varepsilon)}{\omega^\sigma \Gamma(-\sigma,\omega\varepsilon)} \right)^{N_\varepsilon+1-k} \right).
\end{aligned}$$

Observe that, because of the indicator function in the above formula, summation in (8) has to be taken for N_ε from $k-1$ to $+\infty$. Then, by the change of variable $N_{na} = N_\varepsilon + 1 - k$ in

the summation ($N_\varepsilon + 1 - k$ is the number of *non-allocated* jumps), simple calculations give

$$p_\varepsilon(n_1,\dots,n_k) = \sum_{N_{na}=0}^{+\infty} \int_0^{+\infty} du \left(\frac{1}{\Gamma(n)} u^{n-1} (u+\omega)^{k\sigma-n} \prod_{i=1}^{k} \Gamma(n_i - \sigma, (u+\omega)\varepsilon) \right.$$
$$\left. \times \frac{1}{\omega^\sigma \Gamma(-\sigma,\omega\varepsilon)} \frac{\kappa^{k-1}}{\Gamma(1-\sigma)^{k-1}} \frac{N_{na}+k}{N_{na}!} \left(\frac{\kappa(u+\omega)^\sigma}{\Gamma(1-\sigma)} \Gamma(-\sigma,(u+\omega)\varepsilon) \right)^{N_{na}} \mathrm{e}^{-\Lambda_\varepsilon} \right).$$

By Fubini's theorem, we can switch integration and summation, and introduce $\Lambda_{\varepsilon,u}$ as defined in (7), so that

$$p_\varepsilon(n_1,\dots,n_k) = \int_0^{+\infty} du \left(\frac{u^{n-1}}{\Gamma(n)} (u+\omega)^{k\sigma-n} \prod_{i=1}^{k} \Gamma(n_i - \sigma, (u+\omega)\varepsilon) \frac{1}{\omega^\sigma \Gamma(-\sigma,\omega\varepsilon)} \frac{\kappa^{k-1}}{\Gamma(1-\sigma)^{k-1}} \right.$$
$$\left. \times \sum_{N_{na}=0}^{+\infty} \frac{N_{na}+k}{N_{na}!} (\Lambda_{\varepsilon,u}^{N_{na}}) \mathrm{e}^{-\Lambda_\varepsilon} \right),$$

that is (5), since

$$\sum_{N_{na}=0}^{+\infty} \frac{N_{na}+k}{N_{na}!} \Lambda_{\varepsilon,u}^{N_{na}} = \mathrm{e}^{\Lambda_{\varepsilon,u}} \left(\Lambda_{\varepsilon,u} + k \right).$$

□

Lemma 1. *Let (a_n) and (b_n) be two sequences of real numbers, such that*

$$\lim_{n\to+\infty} (a_n + b_n) = l, \quad \liminf_{n\to+\infty} a_n = a_0, \quad \liminf_{n\to+\infty} b_n = b_0,$$

where l, a_0, b_0 are finite, and $a_0 + b_0 = l$. Then

$$\lim_{n\to\infty} a_n = a_0, \quad \lim_{n\to\infty} b_n = b_0.$$

Proof. By definition of lim inf and lim sup we have:

$$\liminf a_n + \liminf b_n \le \liminf(a_n + b_n) \le \liminf a_n + \limsup b_n \le \limsup(a_n + b_n)$$
$$\le \limsup a_n + \limsup b_n.$$

From the hypothesis we have

$$a_0 + b_0 = l = \liminf(a_n + b_n) \le a_0 + \limsup b_n \le \limsup(a_n + b_n) = l = a_0 + b_0,$$

so that $\limsup b_n = b_0$, but by hypothesis $b_0 = \liminf b_n$, and consequently

$$\lim_{n\to+\infty} b_n = b_0.$$

We prove similarly that $\lim_{n\to\infty} a_n = a_0$. □

Of course, this lemma can be generalized to any finite number of sequences. Now we are ready to show that the eppf of an $\varepsilon-$NGG process converges pointwise to that of an NGG process when $\varepsilon \to 0$.

Proposition 2. *Let $p_\varepsilon(\cdot)$ be the eppf of a $\varepsilon-NGG(\sigma,\kappa,\omega,P_0)$ process. Then for each $n_1,\ldots,n_k \in \mathbb{N}$ with $k>0$ and $\sum_{i=1}^k n_i = n$*

$$\lim_{\varepsilon\to 0} p_\varepsilon(n_1,\ldots,n_k) = p_0(n_1,\ldots,n_k), \tag{10}$$

where $p_0(\cdot)$ is the eppf of a $NGG(\sigma,\kappa,\omega,P_0)$ process.

Proof. By Proposition 1

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

where f_ε is the integrand in equation (5). Moreover the eppf of a NGG$(\sigma,\kappa,\omega,P_0)$ process can be written as

$$p_0(n_1,\ldots,n_k) = \int_0^{+\infty} f_0(u;n_1,\ldots,n_k)du$$

where

$$f_0(u;n_1,\ldots,n_k) = \frac{u^{n-1}}{\Gamma(n)}(u+\omega)^{k\sigma-n}\prod_{i=1}^k \Gamma(n_i-\sigma)\left(\frac{\kappa}{\Gamma(1-\sigma)}\right)^{k-1} \times \frac{\kappa}{\Gamma(1-\sigma)}\exp\left\{-\kappa\frac{(\omega+u)^\sigma-\omega^\sigma}{\sigma}\right\};$$

see, for instance, Lijoi et al. (2007). We first show that

$$\lim_{\varepsilon\to 0} f_\varepsilon(u;n_1,\ldots,n_k) = f_0(u;n_1,\ldots,n_k) \quad \text{for any } u>0.$$

This is straightforward by the following remarks:

1. $\lim_{\varepsilon\to 0}\Gamma(n_i-\sigma,(u+\omega)\varepsilon) = \Gamma(n_i-\sigma)$, for any $i=1,2,\ldots,k$, by the Dominated Convergence Theorem, since $n_i-\sigma \geq 1-\sigma > 0$;

2. since $\lim_{\varepsilon\to 0}\Gamma(-\sigma,\omega\varepsilon) = +\infty$ and

$$\Gamma(1-\sigma,x) = -\sigma\Gamma(-\sigma,x) + x^{-\sigma}\mathrm{e}^{-x}$$

(Gradshteyn and Ryzhik, 2000), we have:

$$\lim_{\varepsilon\to 0}\frac{\Lambda_{\varepsilon,u}+k}{\omega^\sigma\Gamma(-\sigma,\omega\varepsilon)} = \frac{\kappa}{\Gamma(1-\sigma)}, \quad \lim_{\varepsilon\to 0}(\Lambda_{\varepsilon,u}-\Lambda_\varepsilon) = -\kappa\frac{(\omega+u)^\sigma-\omega^\sigma}{\sigma}.$$

Now let $\mathcal{C} = \{C_1, \ldots, C_k\}$ be a partition of $\{1, \ldots, n\}$ with group sizes $(n_1, \ldots, n_k)$, and let Π_n be the set of all the possible partitions of $\{1, \ldots, n\}$, of any size $k = 1, \ldots, n$. Of course, by definition of eppf,

$$\sum_{\mathcal{C} \in \Pi_n} p(n_1, \ldots, n_k) = 1$$

and, in particular this holds for either p_ε and p_0. Moreover, by Fatou's Lemma we have

$$\begin{aligned} p_0(n_1, \ldots, n_k) = \int_0^{+\infty} \lim_{\varepsilon \to 0} f_\varepsilon(u; n_1, \ldots, n_k) du &= \int_0^{+\infty} \liminf_{\varepsilon \to 0} f_\varepsilon(u; n_1, \ldots, n_k) du \\ &\leq \liminf_{\varepsilon \to 0} \int_0^{+\infty} f_\varepsilon(u; n_1, \ldots, n_k) du = \liminf_{\varepsilon \to 0} p_\varepsilon(n_1, \ldots, n_k). \end{aligned}$$

Suppose now that for a particular sequence $\mathcal{C} \in \Pi_n$, we had $p_0(n_1, \ldots, n_k) < \liminf_{\varepsilon \to 0} p_\varepsilon(n_1, \ldots, n_k)$. In this case

$$1 = \sum_{\mathcal{C} \in \Pi_n} p_0(n_1, .., n_k) < \sum_{\mathcal{C} \in \Pi_n} \liminf_{\varepsilon \to 0} p_\varepsilon(n_1, .., n_k) \leq \liminf_{\varepsilon \to 0} \sum_{\mathcal{C} \in \Pi_n} p_\varepsilon(n_1, \ldots, n_k) = 1,$$

that is a contradiction. Therefore we can conclude that

$$p_0(n_1, \ldots, n_k) = \liminf_{\varepsilon \to 0} p_\varepsilon(n_1, \ldots, n_k), \text{ for all } n_1, \ldots, n_k, \text{ all } k.$$

Summing up, we have proved so far that:

$$\lim_{\varepsilon \to 0} \sum_{\mathcal{C} \in \Pi_n} p_\varepsilon(n_1, \ldots, n_k) = 1,$$

$$\liminf_{\varepsilon \to 0} (n_1, \ldots, n_k) = p_0(n_1, \ldots, n_k) \text{ for all } \mathcal{C} = (C_1, \ldots, C_k) \in \Pi_n, \sum_{\mathcal{C} \in \Pi_n} p_0(n_1, \ldots, n_k) = 1.$$

By Lemma 1, equation (10) follows. □

Convergence of the sequence of eppfs yield convergence of the sequences of ε-NGG processes. The main distributional result on P_ε is the following:

Proposition 3. *Let P_ε be a ε-NGG$(\sigma, \kappa, \omega, P_0)$ process, for any $\varepsilon > 0$. Then*

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

where P is a NGG$(\sigma, \kappa, \omega, P_0)$ process. Moreover, as $\varepsilon \to +\infty$, $P_\varepsilon \xrightarrow{d} \delta_{\tau_0}$, where $\tau_0 \sim P_0$.

Proof. As mentioned before, P_ε is a proper species sampling model, so that p_ε defines a probability law on the sets of all partitions of $\mathbb{N}_n := \{1, \ldots, n\}$, once that we have set a positive integer n. Therefore, we introduce $(N_1^\varepsilon, \ldots, N_k^\varepsilon)$, the sizes of the blocks (in order of appearance, of the random partition $C_{\varepsilon,n}$ defined by p_ε, for any $\varepsilon \geq 0$. The probability

distributions of $\{(N_1^\varepsilon, \ldots, N_k^\varepsilon), \varepsilon \geq 0\}$ are proportional to the values of p_ε (for any $\varepsilon \geq 0$) in (2.6) in Pitman (2006). Hence, by Proposition 2, for any $k = 1, \ldots, n$ and any n,

$$(N_1^\varepsilon, \ldots, N_k^\varepsilon) \xrightarrow{d} (N_1^0, \ldots, N_k^0) \text{ as } \varepsilon \to 0.$$

Here $(N_1^0, \ldots, N_k^0)$ denote the sizes of the blocks (in order of appearance, of the random partition $C_{\varepsilon,n}$ defined by p_0, the eppf of a NGG$(\sigma, \kappa, \omega, P_0)$ process. By formula (2.30) in Pitman (2006), we have

$$\begin{array}{ccc} \left(\frac{N_j^\varepsilon}{n}\right) & \xrightarrow[n\to+\infty]{d} & (\tilde{P}_j^\varepsilon) \\ {\scriptstyle \varepsilon\to 0}\downarrow{\scriptstyle d} & & \\ \left(\frac{N_j^0}{n}\right) & \xrightarrow[n\to+\infty]{d} & (\tilde{P}_j) \end{array}$$

where P_j^ε and $\tilde{P}_j$ are the j-th weights of a ε-NGG and a NGG process (with parameters $(\sigma, \kappa, \omega, P_0)$), respectively. Note that the sequences depending on n have only a finite number of positive weights.

Recall that the weak convergence of a sequence of random probability measures is equivalent to the pointwise convergence of the Laplace transforms (see Kallenberg, 1983, Theorem 4.2). Let $f(\cdot)$ be a continuous and bounded function on Θ. If we can invert the order of the limit operations below, then we have:

$$\begin{aligned} \lim_{\varepsilon\to 0} \mathbb{E}\left(\mathrm{e}^{-\int_\Theta f d\mu^\varepsilon}\right) &= \lim_{\varepsilon\to 0}\lim_{n\to\infty} \mathbb{E}\left(\mathrm{e}^{-\int_\Theta f d\mu_n^\varepsilon}\right) = \lim_{n\to\infty}\lim_{\varepsilon\to 0} \mathbb{E}\left(\mathrm{e}^{-\int_\Theta f d\mu_n^\varepsilon}\right) \\ &= \lim_{n\to\infty} \mathbb{E}\left(\mathrm{e}^{-\int f d\mu_n^0}\right) = \mathbb{E}\left(\mathrm{e}^{-\int f d\mu^0}\right). \end{aligned} \tag{11}$$

Here we have denoted by

$$\mu_n^\varepsilon := \sum_j \frac{N_j^\varepsilon}{n}\delta_{\tau_j} \quad \text{and} \quad \mu^\varepsilon := \sum_j \tilde{P}_j^\varepsilon \delta_{\tau_j} \quad \text{for any } \varepsilon \geq 0;$$

thus (11) proves the stated convergence, conditioning on $\{\tau_0, \tau_1, \tau_2, \ldots\}$, which are iid from P_0. To justify the interchange of the two limits above, we must prove that the sequence $\left\{\mathbb{E}\left(\mathrm{e}^{-\int f d\mu_n^\varepsilon}\right), n \geq 1\right\}$ converges uniformly. To this end, it is sufficient to show that difference between two next terms in the sequence does not depend on ε; in fact, for any $M > 0$, since

$$|\mathrm{e}^{-x} - \mathrm{e}^{-y}| \leq \mathrm{e}^M |x - y| \text{ for any } x, y \in [-M, M],$$

we have

$$\begin{aligned} \left|\mathbb{E}\left(\mathrm{e}^{-\int f d\mu_{n+1}^\varepsilon}\right) - \mathbb{E}\left(\mathrm{e}^{-\int f d\mu_n^\varepsilon}\right)\right| &\leq \mathbb{E}\left(\left|\mathrm{e}^{-\int f d\mu_{n+1}^\varepsilon} - \mathrm{e}^{-\int f d\mu_n^\varepsilon}\right|\right) \\ &\leq \mathrm{e}^M \mathbb{E}\left(\left|\int f d\mu_{n+1}^\varepsilon - \int f d\mu_n^\varepsilon\right|\right), \end{aligned}$$

where $M \geq \sup f$. Let now $C_{\varepsilon,n+1}$ be a random partition on $\{1,\ldots,n+1\}$ such that its restriction to $\{1,\ldots,n\}$ corresponds to $C_{\varepsilon,n}$. We can distinguish two cases

1. $C_{\varepsilon,n+1}$ has the same number of clusters of $C_{\varepsilon,n}$, one of that, say the one with index j^*, with n_{j^*+1} elements;

2. $C_{\varepsilon,n+1}$ has one more cluster of numerosity one than $C_{\varepsilon,n}$.

In both cases, it is not difficult to prove that

$$\mathbb{E}\left(\left|\int_\Theta f d\mu^\varepsilon_{n+1} - \int_\Theta f d\mu^\varepsilon_n\right|\right) \leq \frac{2M\mathrm{e}^M}{n+1}.$$

Finally, it is straightforward to show that the stated convergence follows from the convergence in distribution conditioning on $\{\tau_0,\tau_1,\tau_2,\ldots\}$, with an argument on Laplace transforms as before. Convergence as $\varepsilon \to +\infty$ is straightforward. □

Let $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)$ be a sample from P_ε, a ε-NGG$(\sigma,\kappa,\omega,P_0)$ process as defined in (3), and let $\boldsymbol{\theta}^* = (\theta^*_1,\ldots,\theta^*_k)$ be the (observed) distinct values in $\boldsymbol{\theta}$. The following proposition gives a "finite dimensional" version of the characterization of the posterior law of a NGG process in James et al. (2009). We will denote by *allocated* jumps of the process the values $P_{l^*_1}, P_{l^*_2},\ldots,P_{l^*_k}$ in (3) 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.

Proposition 4. *If P_ε is an $\varepsilon-NGG(\sigma,\kappa,\omega,P_0)$ process, then the conditional distribution of P_ε, given $\boldsymbol{\theta}^*$ and a latent scalar variable $U = u$, coincides with that of the random measure*

$$P^*_\varepsilon(\cdot) \stackrel{d}{=} wP^{(na)}_{\varepsilon,u}(\cdot) + (1-w)\sum_{j=1}^k P^{(a)}_j \delta_{\theta^*_k}(\cdot)$$

where

1. *$P^{(na)}_{\varepsilon,u}(\cdot)$, the process of* non-allocated *jumps, is distributed according to an $\varepsilon-NGG(\sigma,\kappa,\omega+u,P_0)$ process, given that exactly N_{na} jumps of the process were obtained, where the posterior law of N_{na} is*

$$\frac{\Lambda_{\varepsilon,u}}{k+\Lambda_{\varepsilon,u}}\mathcal{P}_1(\Lambda_{\varepsilon,u}) + \frac{k}{k+\Lambda_{\varepsilon,u}}\mathcal{P}_0(\Lambda_{\varepsilon,u}),$$

   being $\Lambda_{\varepsilon,u}$ as defined in (6)*, and denoting $\mathcal{P}_i(\lambda)$ the shifted Poisson distribution on $\{i,i+1,i+2,\ldots\}$ with mean $i+\lambda$, $i = 0,1$.*

2. *The jumps $\{P^{(a)}_1,\ldots,P^{(a)}_k\}$* assigned *to the fixed points of discontinuity $\boldsymbol{\theta}^* = (\theta^*_1,\ldots,\theta^*_k)$ of P^*_ε are obtained by normalization of $J^{(a)}_j \stackrel{\text{ind}}{\sim} gamma(n_j-\sigma, u+\omega)\mathbb{I}_{(\varepsilon,+\infty)}$, for $j = 1\ldots,k$.*

3. *$P_{\varepsilon,u}(\cdot)$ and $\{J_1^{(a)}, \cdots, J_k^{(a)}\}$ are independent, conditionally to $\boldsymbol{l}^* = (l_1^*, \ldots, l_k^*)$, the vector of locations of the* allocated *jumps.*

4. *when $N_{na} = 0$, w is defined to be equal to 0, while if N_{na} is different from 0, then $w = T_{\varepsilon,u}/(T_{\varepsilon,u} + \sum_{j=1}^k J_j^{(a)})$, where $T_{\varepsilon,u}$ is the total sum of the jumps in representation of $P_{\varepsilon,u}^{(na)}(\cdot)$ as in* (3).

5. *the posterior law of U given $\boldsymbol{\theta}^*$ has density on the positive real given by*

$$f_{U|\boldsymbol{\theta}^*}(u|\boldsymbol{\theta}^*) \propto u^{n-1}(u+\omega)^{k\sigma-n}(\Lambda_{\varepsilon,u}+k)\mathrm{e}^{\Lambda_{\varepsilon,u}}\prod_{i=1}^k \Gamma(n_i-\sigma,(u+\omega)\varepsilon).$$

Observe that this proposition is just way of describing (characterizing) the posterior of an ε-NGG process. As in the infinite dimensional case, the posterior distribution of an ε-NGG$(\sigma,\kappa,\omega,P_0)$ process, conditionally on U and $\boldsymbol{\theta}$, can be expressed as the law of a random probability measure, which is a mixture between an ε-NGG process and a discrete probability measure with support given by the (observed) distinct values $\boldsymbol{\theta}^*$.

Proof. The conditional distribution of $\boldsymbol{\theta}$ is:

$$\begin{aligned}
\mathcal{L}(\theta_1,..,\theta_n|P_\varepsilon) &= \prod_{i=1}^n P_\varepsilon(\theta_i) = \prod_{i=1}^n \sum_{j=0}^{N_\varepsilon} \left(P_j\delta_{\tau_j}(\theta_i)\right) \\
&= \sum_{l_1=0}^{N_\varepsilon} P_{l_1}\delta_{\tau_{l_1}}(\theta_1)\sum_{l_2=0}^{N_\varepsilon} P_{l_2}\delta_{\tau_{l_2}}(\theta_2)\cdots\sum_{l_n=0}^{N_\varepsilon} P_{l_n}\delta_{\tau_{l_n}}(\theta_n) \\
&= \mathbb{I}_{\{1,\ldots,N_\varepsilon+1\}}(k)\frac{1}{(T_\varepsilon)^n}\sum_{l_1^*,\ldots,l_k^*} J_{l_1^*}^{n_1}\ldots J_{l_k^*}^{n_k}\delta_{\tau_{l_1^*}}(\theta_1^*)\ldots\delta_{\tau_{l_k^*}}(\theta_k^*)
\end{aligned}$$

where $(\theta_1^*,\theta_2^*,\ldots,\theta_k^*)$ is the vector of the unique values in the sample. We will omit the indicator $\mathbb{I}_{\{1,\ldots,N_\varepsilon+1\}}(k)$ till we need it. We introduce the latent variable U in the statement of this proposition as already done in the proof of Proposition 1, i.e. $U = \Gamma_n/T_\varepsilon$, where $\Gamma_n \sim gamma(n,1)$, Γ_n and T_ε being independent, so that (9) holds. Therefore,

$$\mathcal{L}(\boldsymbol{\theta},u|P_\varepsilon) = \frac{1}{\Gamma(n)}u^{n-1}e^{-uT_\varepsilon}\sum_{l_1^*,..,l_k^*}\left(J_{l_1^*}^{n_1}\delta_{\tau_{l_1^*}^*}(\theta_1^*)\ldots J_{l_k^*}^{n_k}\delta_{\tau_{l_k^*}^*}(\theta_k^*)\right)$$

Therefore, by Bayes' theorem, we have:

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\theta}, u, P_\varepsilon) = \mathcal{L}(\boldsymbol{\theta}, u|P_\varepsilon)\mathcal{L}(P_\varepsilon) \\
&= \frac{1}{\Gamma(n)} u^{n-1} e^{-uT_\varepsilon} \sum_{l_1^*,\ldots,l_k^*} \big(J_{l_1^*}^{n_1}\delta_{\tau_{l_1^*}^*}(\theta_1^*)\ldots J_{l_k^*}^{n_k}\delta_{\tau_{l_k^*}^*}(\theta_k^*)\big)\mathcal{L}(P_\varepsilon) \\
&= \frac{1}{\Gamma(n)} u^{n-1} \prod_{j=0}^{N_\varepsilon}(e^{-uJ_j}) \sum_{l_1^*,\ldots,l_k^*} \big(J_{l_1^*}^{n_1}\delta_{\tau_{l_1^*}^*}(\theta_1^*)\ldots J_{l_k^*}^{n_k}\delta_{\tau_{l_k^*}^*}(\theta_k^*)\big) \\
&\qquad\qquad \times \prod_{j=0}^{N_\varepsilon}\big(\rho_\varepsilon(J_j)P_0(\tau_j)\big)\mathcal{P}_0(N_\varepsilon;\Lambda_\varepsilon) \\
&= \frac{1}{\Gamma(n)} u^{n-1} \prod_{j=0}^{N_\varepsilon}\big(e^{-uJ_j}\rho_\varepsilon(J_j)P_0(\tau_j)\big) \sum_{l_1^*,\ldots,l_k^*} \big(J_{l_1^*}^{n_1}\delta_{\tau_{l_1^*}^*}(\theta_1^*)\ldots J_{l_k^*}^{n_k}\delta_{\tau_{l_k^*}^*}(\theta_k^*)\big)\mathcal{P}_0(N_\varepsilon;\Lambda_\varepsilon)
\end{aligned}
\tag{12}
$$

where, in this proof, $\mathcal{P}_0(N_\varepsilon;\Lambda_\varepsilon)$ is the density of the Poisson distribution with parameter Λ_ε, evaluated in N_ε and $P_0(\tau)$ is the density of P_0 evaluated in τ.

The conditional distribution of P_ε, given $U = u$ and $\boldsymbol{\theta}$, is as follows:

$$
\mathcal{L}(P_\varepsilon|u,\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\tau},\boldsymbol{J},N_\varepsilon|u,\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\tau},\boldsymbol{J}|N_\varepsilon,u,\boldsymbol{\theta})\mathcal{L}(N_\varepsilon|u,\boldsymbol{\theta}). \tag{13}
$$

The second factor in the right handside is proportional to

$$
\begin{aligned}
\mathcal{L}(N_\varepsilon,u,\boldsymbol{\theta}) &= \int dJ_0\ldots dJ_{N_\varepsilon}d\tau_0\ldots d\tau_{N_\varepsilon}\mathcal{L}(\boldsymbol{\tau},\boldsymbol{J},N_\varepsilon,u,\boldsymbol{\theta}) \\
&= \sum_{l_1^*,\ldots,l_k^*}\Bigg\{\Bigg[\prod_{i=1}^{k}\int J_{l_i^*}^{n_i}\delta_{\tau_{l_i^*}}(\theta_i^*)e^{-uJ_{l_i^*}}\rho_\varepsilon(J_{l_i^*})P_0(\tau_{l_i^*})dJ_{l_i^*}d\tau_{l_i^*}\Bigg] \\
&\qquad\times\Bigg[\prod_{j\neq\{l_1^*,..,l_k^*\}}\int e^{-uJ_j}\rho_\varepsilon(J_j)P_0(\tau_j)dJ_jd\tau_j\Bigg]\Bigg\}\frac{1}{\Gamma(n)}u^{n-1}\mathcal{P}_0(N_\varepsilon;\Lambda_\varepsilon).
\end{aligned}
$$

Observe that, for any $j \neq \{l_1^*,..,l_k^*\}$,

$$
\begin{aligned}
\int e^{-uJ_j}\rho_\varepsilon(J_j)P_0(\tau_j)dJ_jd\tau_j &= \int_0^{+\infty} e^{-uJ_j}\rho_\varepsilon(J_j)dJ_j \\
&= \frac{1}{\omega^\sigma\Gamma(-\sigma,\omega\varepsilon)}\int_0^{+\infty} x^{-\sigma-1}\mathrm{e}^{(u+\omega)x}\mathbb{I}_{(\varepsilon,+\infty)}(x)dx \\
&= \frac{(\omega+u)^\sigma}{\omega^\sigma\Gamma(-\sigma,\omega\varepsilon)}\int_{(\omega+u)\varepsilon}^{+\infty} e^{-y}y^{-\sigma-1}dy \\
&= \frac{(\omega+u)^\sigma\Gamma(-\sigma,(\omega+u)\varepsilon)}{\omega^\sigma\Gamma(-\sigma,\omega\varepsilon)}.
\end{aligned}
\tag{14}
$$

The integrand function in the second line of the formula above is the kernel of the mean intensity of a ε-NGG$(\sigma,\kappa,\omega+u,P_0)$ process. On the other hand, for $i = 1,\ldots,k$:

$$
\begin{aligned}
&\int J_{l_i^*}^{n_i}\delta_{\tau_{l_i^*}}(\theta_i^*)\mathrm{e}^{-uJ_{l_i^*}}\rho_\varepsilon(J_{l_i^*})P_0(\tau_{l_i^*})dJ_{l_i^*}d\tau_{l_i^*} \\
&\qquad = \left(\int J_{l_i^*}\mathrm{e}^{-uJ_{l_i^*}}\rho_\varepsilon(J_{l_i^*})dJ_{l_i^*}\right)\left(\int \delta_{\tau_{l_i^*}}(\theta_i^*)P_0(\theta_i^*)d\theta_i^*\right) \\
&\qquad = \frac{P_0(\theta_i^*)}{\omega^\sigma\Gamma(-\sigma,\omega\varepsilon)}\int_0^{+\infty} x^{n_i}\mathrm{e}^{-ux}x^{-1-\sigma}\mathrm{e}^{-\omega x}\mathbb{I}_{(\varepsilon,+\infty)}(x)dx \qquad (15)\\
&\qquad = \frac{(\omega+u)^{\sigma-n_i}}{\omega^\sigma}\frac{\Gamma(n_i-\sigma,(u+\omega)\varepsilon)}{\Gamma(-\sigma,\omega\varepsilon)}P_0(\theta_i^*).
\end{aligned}
$$

The integrand function in (15) is the kernel of a gamma density with parameters $(n_i-\sigma, u+\omega)$, restricted to $(\varepsilon,+\infty)$. Summing up, we have

$$
\begin{aligned}
\mathcal{L}(N_\varepsilon|u,\boldsymbol{\theta}) \propto \mathcal{L}(N_\varepsilon,u,\boldsymbol{\theta}) &= \frac{1}{\Gamma(n)}u^{n-1}\sum_{l_1^*,..,l_k^*}\left\{\left(\frac{(\omega+u)^{k\sigma-n}\prod_{i=1}^k\Gamma(n_i-\sigma,(\omega+u)\varepsilon)P_0(\theta_i^*)}{\omega^{\sigma k}\Gamma(-\sigma,\omega\varepsilon)^k}\right)\right. \\
&\quad \left.\times\left(\frac{(\omega+u)^{\sigma(N_\varepsilon+1-k)}\Gamma(-\sigma,(u+\omega)\varepsilon)^{N_\varepsilon+1-k}}{\omega^{\sigma(N_\varepsilon+1-k)}\Gamma(-\sigma,\omega\varepsilon)^{N_\varepsilon+1-k}}\right)\right\}\mathcal{P}_0(N_\varepsilon;\Lambda_\varepsilon) \\
&= \frac{u^{n-1}}{\Gamma(n)}\mathcal{P}_0(N_\varepsilon;\Lambda_\varepsilon)\frac{(N_\varepsilon+1)!}{(N_\varepsilon+1-k)!}\prod_{i=1}^k\Big(P_0(\theta_i^*)\Gamma(n_i-\sigma,\varepsilon(\omega+u))\Big) \qquad (16)\\
&\quad \times\frac{(\omega+u)^{\sigma k-n}}{\omega^{\sigma k}\Gamma(-\sigma,\omega\varepsilon)^k}\frac{(\omega+u)^{\sigma N_{na}}\Gamma(-\sigma,\varepsilon(\omega+u))^{N_{na}}}{\omega^{\sigma N_{na}}\Gamma(-\sigma,\omega\varepsilon)^{N_{na}}}\mathbb{I}_{\{(N_\varepsilon+1)\geq k\}}.
\end{aligned}
$$

As in the proof of Proposition 1, $N_{na}=N_\varepsilon+1-k$ is the number of *non-allocated* jumps. Therefore, since k is given, the conditional distribution $\mathcal{L}(N_\varepsilon|u,\boldsymbol{\theta})$ is identified by $\mathcal{L}(N_{na}|u,\boldsymbol{\theta})$; we have

$$
\mathcal{L}(N_{na}|u,\boldsymbol{\theta}) \propto \mathbb{I}_{(N_{na}\geq 0)}\frac{(\omega+u)^{\sigma k-n}}{\omega^\sigma\Gamma(-\sigma,\omega\varepsilon)}\frac{(N_{na}+k)}{N_{na}!}\left(\frac{\kappa(u+\omega)^\sigma}{\Gamma(1-\sigma)}\Gamma(-\sigma,(u+\omega)\varepsilon)\right)^{N_{na}}.
$$

Let $\Lambda_{\varepsilon,u}$ be as in (6); it easily follows that

$$
\begin{aligned}
\mathcal{L}(N_{na}|\varepsilon,u,\boldsymbol{\theta}) &\propto \frac{N_{na}+k}{N_{na}!}\mathrm{e}^{-\Lambda_{\varepsilon,u}}\Lambda_{\varepsilon,u}^{N_{na}} = \left(\frac{N_{na}}{N_{na}!}+\frac{k}{N_{na}!}\right)\mathrm{e}^{-\Lambda_{\varepsilon,u}}\Lambda_{\varepsilon,u}^{N_{na}} \\
&= \frac{\Lambda_{\varepsilon,u}}{(N_{na}-1)!}\Lambda_{\varepsilon,u}^{(N_{na}-1)}\mathrm{e}^{-\Lambda_{\varepsilon,u}}+\frac{k}{N_{na}!}\Lambda_{\varepsilon,u}^{N_{na}}\mathrm{e}^{-\Lambda_{\varepsilon,u}} \qquad (17)\\
&= \frac{\Lambda_{\varepsilon,u}}{\Lambda_{\varepsilon,u}+k}\mathcal{P}_1(N_{na};\Lambda_{\varepsilon,u})+\frac{k}{\Lambda_{\varepsilon,u}+k}\mathcal{P}_0(N_{na};\Lambda_{\varepsilon,u}).
\end{aligned}
$$

The first factor in the right handside of (13) can be computed by the following comment. Denote by $\boldsymbol{l}^*=(l_1^*,\ldots,l_k^*)$ the vector of locations of the *allocated* jumps. From (12) it is clear that, since

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{J},\boldsymbol{\tau},\boldsymbol{l}^*|N_{na},u,\boldsymbol{\theta}) &= J_{l_1^*}^{n_1}\delta_{\tau_{l_1^*}^*}(\theta_1^*)\ldots J_{l_k^*}^{n_k}\delta_{\tau_{l_k^*}^*}(\theta_k^*)\prod_{j=0}^{N_{na}+k-1}\rho_\varepsilon(J_j)P_0(\tau_j)\mathrm{e}^{-uJ_j} \\
&= \left(\prod_{i=1}^k J_{l_i^*}^{n_i}\delta_{\tau_{l_i^*}^*}(\theta_i^*)\mathrm{e}^{-uJ_{l_i^*}}\rho_\varepsilon(J_{l_i^*})P_0(J_{l_i^*})\right)\times\left(\prod_{j\neq\{l_1^*,..,l_k^*\}} e^{-uJ_j}\rho_\varepsilon(J_j)P_0(\tau_j)\right). \qquad (18)
\end{aligned}
$$

The first factor in the last expression refers to the unnormalized *allocated* process: the support is $\boldsymbol{\theta}^*$, while the jumps follow independent restricted gamma densities, as clearly observed after (15). This shows point 2. of the Proposition.

By the remark made after (14), we have that $\mathcal{L}(\boldsymbol{J}, \boldsymbol{\tau}, \boldsymbol{l}^* | N_{na}, u, \boldsymbol{\theta})$, describing the law of the *non-allocated* process, is a ε-NGG process with N_{na} jumps, and the conditional distribution of N_{na} is described in (17). This shows point 1. of the Proposition.

Point 3 follows straightforwardly from (18). Normalization of the jumps (*allocated* and *non-allocated*) gives 4.

With regard to 5., we need to integrate out N_ε in $\mathcal{L}(N_\varepsilon, u, \boldsymbol{\theta})$ displayed in (16). We have already made these computations in the proof of Proposition 1, and thus $f_{U|\boldsymbol{\theta}^*}(u|\boldsymbol{\theta}^*)$ is proportional to the integrand in (5). □

4 ε-NGG process mixtures

Often, in Bayesian nonparametric problems, it happens that discrete random probabilities, as our ε-NGG process, appear as mixing measures in a mixture context. Indeed, we are going to consider a mixture of Gaussian kernels as the distribution of the i-th observation, where the mixing measure is the ε-NGG$(\sigma, \kappa, \omega, P_0)$ process. In the rest of paper we set $\omega = 1$ (since the original parametrization is not unique) and fix P_0 (see Sections 5 and 6 for details), and change notation accordingly, i.e. ε-NGG(σ, κ, P_0). The model we assume is the following:

$$
\begin{aligned}
X_i|\theta_i &\overset{\text{ind}}{\sim} k(\cdot;\theta_i),\ i = 1, \ldots, n \\
\theta_1, \ldots, \theta_n|P_\varepsilon &\overset{\text{iid}}{\sim} P_\varepsilon \\
P_\varepsilon &\sim \varepsilon - NGG(\sigma, \kappa, P_0) \text{ process prior,} \\
\varepsilon, \sigma, \kappa &\sim \pi(\varepsilon) \times \pi(\sigma) \times \pi(\kappa),
\end{aligned} \tag{19}
$$

where $k(\cdot;\theta_i)$ is a parametric family of densities on $\mathbb{X} \subset \mathbb{R}^p$, for all $\theta \in \Theta \subset \mathbb{R}^m$. In the rest of the paper, we assume the Gaussian kernel, where θ_i denotes the means and the covariance matrix. Remember that P_0 is a non-atomic probability measure on Θ; it is straightforward to see that $\mathbb{E}(P_\varepsilon(A)) = P_0(A)$ for all $A \in \mathcal{B}(\theta)$ and all $\varepsilon \geq 0$. Model (19) will be addressed here as $\varepsilon-NGG$ *hierarchical mixture* model. It is well known that this model is equivalent to assume that the X_i's, conditionally on P_ε, are independently distributed according to the random density

$$
f(x) = \int_\Theta k(x;\theta) P_\varepsilon(d\theta) = \sum_{j=0}^{N_\varepsilon} P_j\ k(x;\tau_j). \tag{20}
$$

In general, computation of posterior inference for (19), when P_ε is substituted by a NGG process P, is not straightforward, since this model assumes an infinite number of parame-

ters. As we mentioned in the Introduction, different approaches have been proposed in the literature. Here we exploit a prior truncation approach; in fact, from the algorithmic point of view, the finite dimensionality of the $\varepsilon - NGG$ process is a key point since it allows us to express our r.p.m. in terms of a finite number of random variables. In particular, we are able to build a blocked Gibbs sampler to update blocks of parameters, which are drawn from multivariate distributions. The *parameter* is $(P_\varepsilon, \varepsilon, \sigma, \kappa, \boldsymbol{\theta})$, and the posterior is proportional to the product of the conditional distribution of the data, given the parameter, times the prior, i.e.

$$\mathcal{L}(\boldsymbol{X}|\boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta}|P_\varepsilon)\mathcal{L}(P_\varepsilon, \varepsilon, \sigma, \kappa) = \mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}|P_\varepsilon)\mathcal{L}(P_\varepsilon|\varepsilon)\mathcal{L}(\varepsilon, \sigma, \kappa). \tag{21}$$

The conditional law $\mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}|P_\varepsilon)$ can be expressed as follows:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{X}, \boldsymbol{\theta}|P_\varepsilon) &= \prod_{i=1}^{n} \big(P_\varepsilon(\theta_i)k(X_i;\theta_i)\big) \\
&= \Big(\prod_{i\in C_1} k(X_i;\theta_1^*)\dots\prod_{i\in C_k} k(X_i;\theta_k^*)\Big)\Big(\sum_{l_1^*,\dots,l_k^*} P_{l_1^*}^{n_1}..P_{l_k^*}^{n_k}\delta_{\tau_{l_1^*}}(\theta_1^*)..\delta_{\tau_{l_k^*}}(\theta_k^*)\Big) \\
&= \frac{1}{T_\varepsilon^n}\sum_{l_1^*,\dots,l_k^*}\Big(J_{l_1^*}^{n_1}\prod_{i\in C_1}k(X_i;\theta_1^*)J_{l_2^*}^{n_2}\prod_{i\in C_2}k(X_i;\theta_2^*)..J_{l_k^*}^{n_k}\prod_{i\in C_k}k(X_i;\theta_k^*)\Big),
\end{aligned} \tag{22}$$

while $\mathcal{L}(P_\varepsilon|\varepsilon)$ is the finite dimentional distribution of P_ε in Section 3, and the joint law $\mathcal{L}(\varepsilon, \sigma, \kappa) = \pi(\varepsilon)\pi(\sigma)\pi(\kappa)$ will be elicited in Sections 5 and 6. We use the same notation as in the proof of Proposition 4. We augment the stace space and apply Proposition 4, considering also the random variable U. Therefore, the sample space of the Gibbs sampler is the set of all values of the *parameter* $(\boldsymbol{\theta}, P_\varepsilon, \varepsilon, u, \sigma, \kappa)$. Details of the blocked Gibbs sampler can be found in the Appendix; however, in the following steps, we describe all the full-conditionals:

1. Sampling from $\mathcal{L}(u|\boldsymbol{X}, \boldsymbol{\theta}, P_\varepsilon, \varepsilon, \sigma, \kappa)$: since the joint law of data and parameters (see (23) in the Appendix) depends on u only through its prior density, this conditional distribution is equal to the prior of U, that is the gamma distribution with parameters (n, T_ε).

2. Sampling from $\mathcal{L}(\boldsymbol{\theta}|u, \boldsymbol{X}, P_\varepsilon, \varepsilon, \sigma, \kappa)$: by (23), each θ_i, for $i = 1, \dots, n$, has discrete law with support $\{\tau_0, \tau_1, \dots, \tau_{N_\varepsilon}\}$, and probabilities $\mathbb{P}(\theta_i = \tau_j) \propto J_j k(X_i; \tau_j)$.

3. Sampling from $\mathcal{L}(P_\varepsilon, \varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}, \boldsymbol{X})$: this step is not straigthforward. In the Appendix we show that it can be split into two consecutive substeps:

   3.a Sampling from $\mathcal{L}(\varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}, \boldsymbol{X})$: a Gibbs sampler strategy will achieve it. For a detailed description of the full conditionals (i) $\mathcal{L}(\varepsilon|\sigma, \kappa, u, \boldsymbol{\theta}, \boldsymbol{X})$, (ii) $\mathcal{L}(\sigma|\varepsilon, \kappa, u, \boldsymbol{\theta}, \boldsymbol{X})$ and (iii) $\mathcal{L}(\kappa|\varepsilon, \sigma, u, \boldsymbol{\theta}, \boldsymbol{X})$, we refer to the Appendix (see formulas (24), (25), (26)).

3.b Sampling from $\mathcal{L}(P_\varepsilon|\varepsilon,\sigma,\kappa,u,\boldsymbol{\theta},\boldsymbol{X})$: via characterization of the posterior in Proposition 4, since this distribution is equal to $\mathcal{L}(P_\varepsilon|\varepsilon,\sigma,\kappa,u,\boldsymbol{\theta})$. To put in 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.

Summing up, our algorithm is outlined in Figure 1. With regard to 3.b.v, we do not directly apply Proposition 4, but add an acceleration step (see for instance Argiento et al., 2010) for sampling from the distribution in Figure 1. As a final remark in this section, when sampling from non-standard distributions, Accept-Reject or Metropolis-Hastings algorithms have been exploited.

Repeat for g in 1...**G**:

1. Sample $\mathbf{u}^{(g)}$ from a Gamma(n, T_ε).

2. For i=1,..,n sample $\boldsymbol{\theta}_{\boldsymbol{i}}^{(g)}$ from a discrete distribution s.t. $\mathbb{P}(\theta_i=\tau_j)\propto J_j k(X_i;\tau_j),\ \ j=0,..,N_\varepsilon$.

3.a.*i* Sample $\boldsymbol{\varepsilon}^{(g)}$ from $\mathcal{L}(\varepsilon)\propto\prod_{i=1}^k\Gamma(n_i-\sigma,\varepsilon(u+\omega))e^{\Lambda_{\varepsilon u}-\Lambda_\varepsilon}\dfrac{\Lambda_{\varepsilon u}+k}{\Gamma(-\sigma,\omega\varepsilon)}\pi(\varepsilon)$.

3.a.*ii* Sample $\boldsymbol{\sigma}^{(g)}$ from $\mathcal{L}(\sigma)\propto\dfrac{(u+\omega)^{k\sigma}}{\omega^\sigma}\dfrac{\Lambda_{\varepsilon u}+k}{\Gamma(-\sigma,\omega\varepsilon)}e^{\Lambda_{\varepsilon u}-\Lambda_\varepsilon}\prod_{i=1}^k\Gamma(n_i-\sigma,\varepsilon(u+\omega))\Gamma(1-\sigma)^{1-k}\pi(\sigma)$.

3.a.*iii* If the prior for κ is a gamma(α,β), sample $\boldsymbol{\kappa}^{(g)}$ from a mixture of gamma densities: $p_1 gamma(\alpha+k,R+\beta)+(1-p_1)gamma(\alpha+k-1,R+\beta)$.

3.b.*i* Sample $N_{na}^{(g)}$ from $\dfrac{\Lambda_{\varepsilon u}}{\Lambda_{\varepsilon u}+k}\mathcal{P}_1(\Lambda_{\varepsilon u})+\dfrac{k}{\Lambda_{\varepsilon u}+k}\mathcal{P}_0(\Lambda_{\varepsilon u})$, then set $\boldsymbol{N_\varepsilon^{(g)}}+1=N_{na}^{(g)}+k$.

3.b.*ii* **Non-allocated jumps**: sample independently from $\mathcal{L}(J_j)\propto e^{-uJ_j}\rho_\varepsilon(J_j)$.

3.b.*iii* **Allocated jumps**: sample independently from $\mathcal{L}(J_{l_i^*})\propto\text{gamma}(n_i-\sigma,u+\omega)\mathbb{1}_{(\varepsilon,\infty)}$.

3.b.*iv* **Non-allocated points of support**: sample independently from P_0.

3.b.*v* **Allocated points of support**: sample independently from $\mathcal{L}(\tau_i^*)\propto\{\prod_{j\in C_i}k(X_j;\tau_i)\}P_0(\tau_i)$.

Figure 1: Blocked Gibbs sampler scheme; the conditioning arguments of all full conditionals have been cut out to simplify notation.

5 Galaxy data

This super-popular dataset contains $n = 82$ measured velocities of different galaxies from six well-separated conic sections of space. Values are expressed in Km/s, scaled by a factor of 10^{-3}. We report posterior estimates for different sets of hyperparameters of the $\varepsilon-$NGG mixture model (19) when $k(\cdot;\theta)$ is the Gaussian density on $\mathbb{R}$ and $\theta = (\mu, \sigma^2)$ stands for its mean and variance, and $P_0(d\mu, d\sigma^2) = \mathcal{N}(d\mu; m_0, \sigma^2/\kappa_0) \times inv - gamma(d\sigma^2; a, b)$; here $\mathcal{N}(m_0, \sigma^2/\kappa_0)$ is the Gaussian distribution with m_0 mean and σ^2/κ_0 variance, and $inv - gamma(a, b)$ is the inverse-gamma distribution with mean $b/(a-1)$ (if $a > 1$). We set $m_0 = \bar{x}_n = 20.8315$, $\kappa_0 = 0.01$, $a = 2$, $b = 1$ as proposed first in Escobar and West (1995).

We did an extensive robustness analysis with respect to ε, σ, κ; see Bianchini (2014). Here we shed light on four sets of hyperparameters only, to understand sensitivity of the estimates (A) when ε varies, but it is not random, (B) when σ varies (but it is not random), then (C) when ε is assumed random and σ and κ are fixed, and finally (D) when both σ and κ are random and ε is fixed.

We have implemented our Gibbs sampler in C++. Tests were made on a laptop with Intel Core i7 2670QM processor, with 6 GB of RAM. Every run produced a final sample size of 10,000 iterations, after a thinning of 10 and an initial burn-in of 10,000 iterations. Every time the convergence was checked by standard R package CODA tools.

With reference to (A), we set $\sigma = 0.4$ and $\kappa = 0.45$, and $\varepsilon = 10^{-6}, 10^{-3}, 10^{-1}, 1$. Figure 2 shows the predictive density estimates under different values of ε: all the estimates are similar and they fit well the data. Observe that, when ε increases, more jumps J_j's are cut out from the sum defining the process P_ε (see (3)) and, consequently, less components in the mixture (20) are considered. Therefore the posterior estimate of the number K_n of components will be concentrated on smaller integer values as ε increases (see Figure 3).

It is worth underlining that, as another consequence of the smaller number of components in the mixture (20) when ε increases, we have observed a huge gain in run-time: for instance, with our machine, the run-time ranges from approximately 7 minutes ($\varepsilon = 10^{-6}$) to less than 1 minute ($\varepsilon = 1$).

The second set (B) of hyperparameters is specified by $\varepsilon = 10^{-6}$ and $\kappa = 0.45$, while σ ranges in $\{0.001, 0.1, 0.2, \ldots, 0.8\}$. The posterior density estimates are similar to those obtained before, and for this reason they are not reported here. On the other hand, we are interested to understand the effect of σ on the posterior distribution of K_n, as shown in Table 1. Note that we are also including the Dirichlet process mixture model here (for $\sigma = 0.001 \simeq 0$ and ε small). As expected, the posterior mean of K_n, as well as its variance, increases with σ.

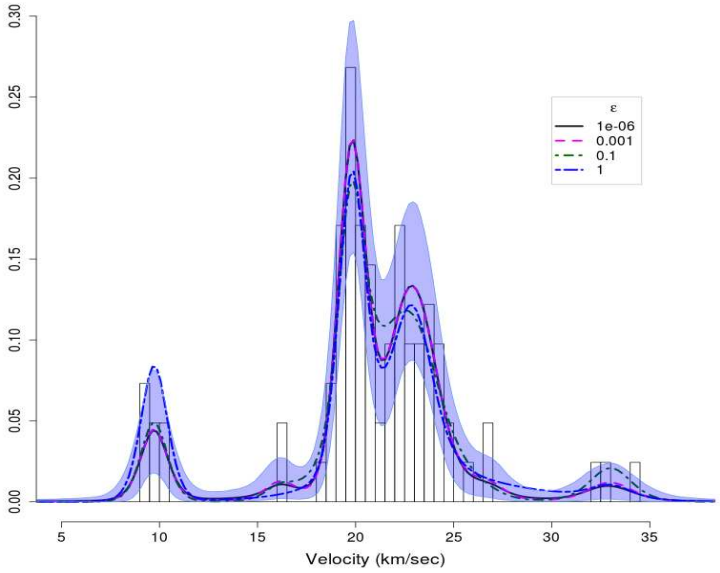
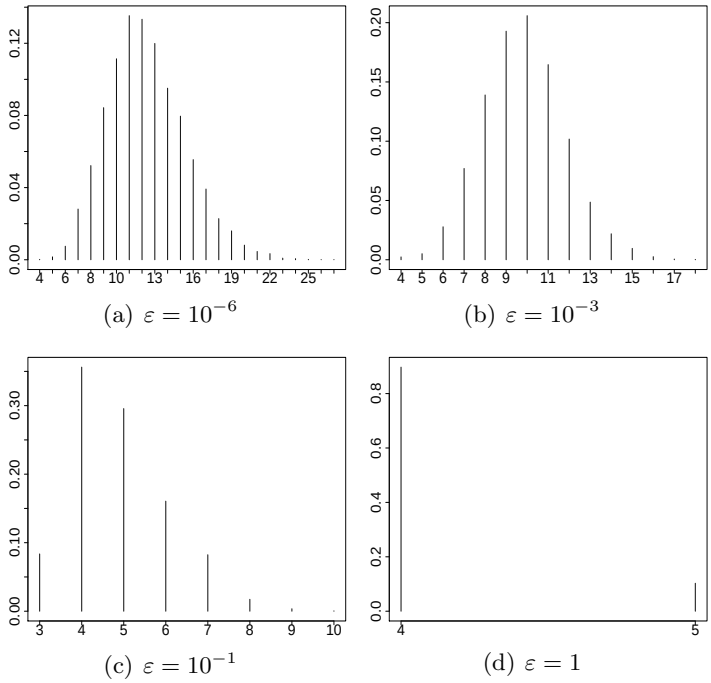


Figure 2: Density estimates for different values of ε, while $\sigma = 0.4$ and $\kappa = 0.45$, case (A). The shaded region denotes 90% CI around the density estimates for $\varepsilon = 10^{-6}$.



(a) $\varepsilon = 10^{-6}$ (b) $\varepsilon = 10^{-3}$

(c) $\varepsilon = 10^{-1}$ (d) $\varepsilon = 1$

Figure 3: Posterior distributions of the number K_n of components in the $\varepsilon-$NGG mixture with hyperparameter set (A).

σ	Prior mean	Posterior mean	Posterior variance
0.001	3	6.13	1.73
0.1	4.06	7.18	2.39
0.2	5.6	8.74	4.25
0.3	7.8	10.49	6.39
0.4	10.9	12.36	9.30
0.5	15.3	14.06	11.49
0.6	21.5	15.90	14.61
0.7	30.2	17.67	17.66
0.8	42.3	19.05	20.16

Table 1: Posterior (and prior) summaries of K_n under case (B).

For set (C) of hyperparameters, we have considered $\sigma \in \{0.001, 0.1, 0.2, \ldots, 0.9\}$, $\kappa = 0.45$ and ε random, uniformly distributed on the interval $(0, \delta)$, with $\delta = \min(0.1, E(T_\varepsilon))$ (non-informative prior) or with a scaled beta distribution on the same interval with mean equal to 0.25δ and variance $0.05\delta^2$ (a more informative prior). When ε is random, the model is expected to be more flexible, since it would "adjust" for the number of jumps of the process P_ε that must be considered. Furthermore, on one hand, if ε increases, the process will be significantly different from the NGG process (indeed, $P_\varepsilon \xrightarrow{d} \delta_{\tau_0}$), since, in this case, many small jumps will not be included in (3). As in the previous cases, density estimates are pretty good and we do not include them here. Figure 4 shows the posterior mean of K_n as a function

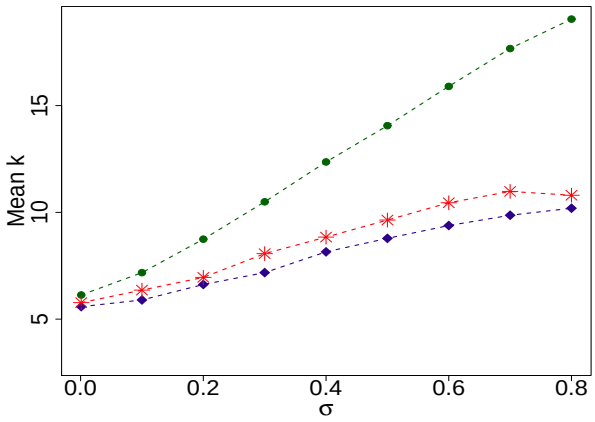


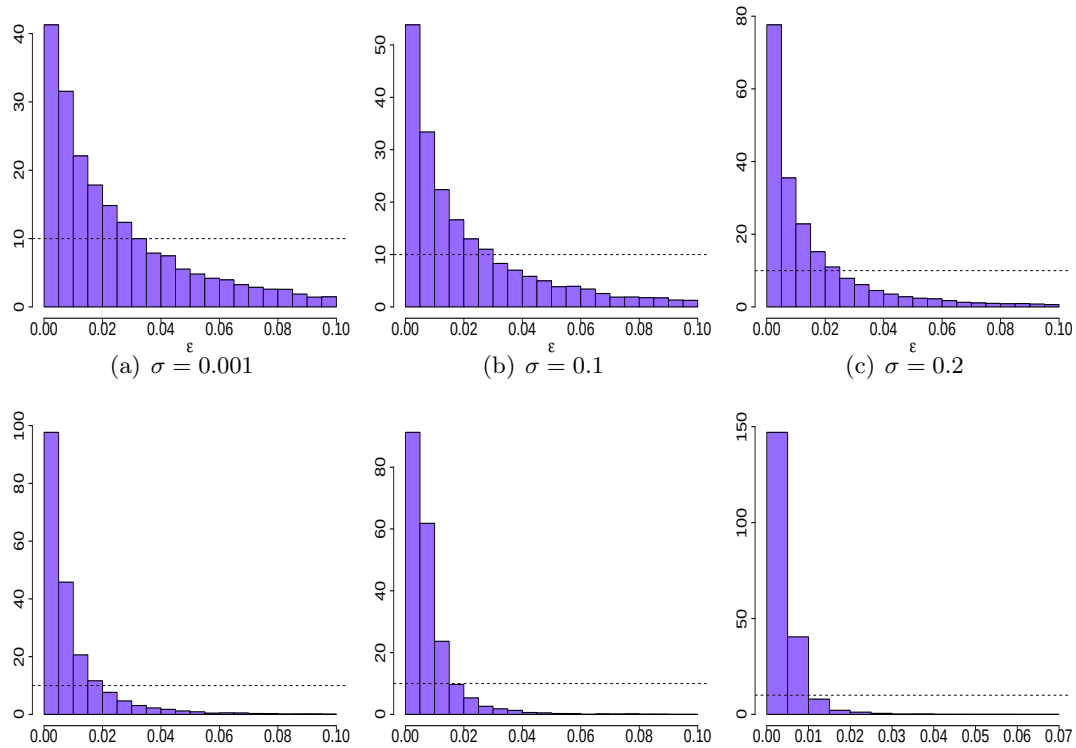
Figure 4: Posterior mean of K_n as a function of σ, under different priors for ε in experiment (C): degenerate on 10^{-6} (green dots), uniform (blue diamonds) and scaled beta (red stars).

of σ for three different priors on ε. The linear increase in $\mathbb{E}(K_n|data)$ is smaller when ε is

beta (red stars) or uniform distributed (blue diamonds) than when ε is equal to 10^{-6} (green dots). In this case, even run-times are shorter than in (B), since the posterior number of *allocated* jumps is usually smaller.

As far as robustness with respect to σ is concerned, we should acknowledge that, as σ increases, more computational problems come up, because of the incomplete gamma function, appearing in the expression of ρ_ε given in Section 3, that is harder to be numerically evaluated.

Looking at the posterior distribution of ε in Figure 5, data suggest that small values of ε are the "best" fit, when the prior of ε is uniform. In particular, increasing σ, and consequently increasing the prior expected number of components in the $\varepsilon-$NGG mixture, we get that the posterior of ε is concentrated on smaller values, which implies larger values for K_n a posteriori.



(a) $\sigma = 0.001$ (b) $\sigma = 0.1$ (c) $\sigma = 0.2$

(d) $\sigma = 0.5$ (e) $\sigma = 0.7$ (f) $\sigma = 0.9$

Figure 5: Posterior distribution of ε for experiment (C), together with $\mathcal{U}(0, \delta)$ prior (dashed).

Finally, we have considered case (D), when both σ and κ are random, and ε is small ($\varepsilon = 10^{-4}$). In particular, we set four different priors $\pi(\sigma) \times \pi(\kappa) = Beta(a_1, b_1)gamma(c_1, d_1)$, with $(a_1, b_1, c_1, d_1) \in \{(2, 5, 2, 2), (10, 23, 1.1, 8), (1.1, 30, 1.1, 8), (10, 23, 100, 50)\}$; the prior information on (σ, κ), and consequently on K_n, is quite different among these four cases: diffuse

prior marginals first, then two conflicting prior marginal beliefs, and last prior marginal beliefs in agreement. For all priors we have got density estimates similar to those reported in Figure 2, while the posterior distribution of K_n is in accordance to the prior information. In particular, σ influences the posterior variance of N_{na}, the number of *non-allocated* jumps: in fact, if a priori σ is concentrated on large values, then the tail of the posterior distribution of N_{na} is heavy. Figure 6 shows the scatterplots of posterior values of (σ, κ); contour plots of the priors are superimposed. Note that, in panels (b) and (c), the posterior is in strong disagreement with the prior, since the prior on (σ, κ) has been assigned too restrictive in these two cases.

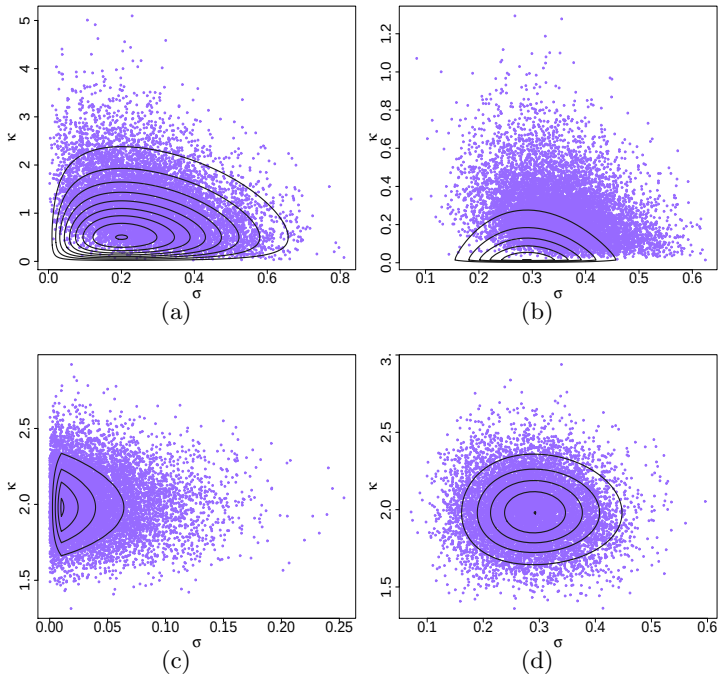


Figure 6: Scatterplots of posterior values of (σ, κ) with contour levels of the prior, case (D).

6 Yeast cell cycle data

We fitted our model to a multivariate dataset used in the literature for clustering gene expression profiles, usually called Yeast cell cycle data (see Cho et al., 1998). A gene expression data set from a microarray experiment can be represented by a real-valued matrix $[X_{ij},\ 1 \leq i \leq n,\ 1 \leq j \leq p]$, where the rows $(X_1, \ldots, X_n)$ contain the expression patterns of genes and are our data points. Each cell X_{ij} is the measured expression level of gene i at time j. The Yeast cell cycle dataset contains $n = 389$ gene expression profiles, observed at 17 different time values, one every 10 minutes from time zero. We consider

only a part of the data, and filter them: the final dataset ($n = 389$, $p = 9$) is the same as in Argiento et al. (2013). We assume the Gaussian kernel $k(\cdot;\theta_i) = \mathcal{N}_p(\cdot;\theta_i)$ where $\theta_i = (\mu_i, \Sigma_i)$ and Σ_i, the covariance matrix, is assumed diagonal with entries $(\sigma^2_{1,i}, \ldots, \sigma^2_{p,i})$. Here $P_0(d\mu, d\Sigma) = \mathcal{N}_p\left(d\mu | m_0, \frac{1}{s_0}\Sigma\right) \times \prod_{k=1}^{p} inv-gamma(d\sigma^2_k | a, b)$.

We made a thorough robustness analysis, with respect to the choice of P_0 and $(\varepsilon, \sigma, \kappa)$-prior. We were able to compute the log-pseudo marginal likelihood (LPML) for every set of hyperparameters; however, here we report posterior inference for the set of hyperparameters which is most in agreement with the prior information given by the reference partition of Cho et al. (1998): $\boldsymbol{m}_0 = \mathbf{0}$, $s_0 = 1$, $a = 3$, $b = 2$, so that $\mathrm{Var}(\mu) = \mathbb{I}_p$ and $\mathbb{E}(\Sigma) = \mathbb{I}_p$. To understand the effect of $\varepsilon, \sigma, \kappa$, first we set $\sigma = 0.001$ and $\kappa = 0.7$, so that $\mathbb{E}(K_n) = 5$ as in the reference partition, and let ε vary in $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$ (case (E)), then $\varepsilon \sim \mathcal{U}(0, 0.01)$ and $\sigma \in \{0.01, 01, 0.2, \ldots, 0.5\}$ and $\kappa = 0.7$ (case (F)). Finally, we set $\varepsilon = 10^{-4}$, $\sigma \sim Beta(2, 15)$ and $\kappa \sim gamma(2, 0.1)$ (case (G)).

The posterior inference was computed via MCMC chains as before, with a final sample size of 5,000, after a thinning of 20 and a burn-in of 5,000. As far as case (E) is concerned, we do not report the inference, but make only one comment: a priori, we have to assume ε on rather small values, otherwise the model would get stuck into a parametric one (remember that for $\varepsilon \to +\infty$ our model is parametric). From a computational point of view, what happens in pratice is that, if ε is fairly large, the jumps J_j's are approximately independent sampled from a degenerate distribution on ε, and therefore, they assume the same value; consequently, the full-conditional of $\boldsymbol{\theta}$, as in Step 2. of the algorithm (see Figure 1), depends only on the parametric kernel, evaluated at data points, yielding that $N_\varepsilon + 1$ and K_n coincide.

For experiment (F), Figure 7 illustrates the posterior of ε with $\sigma = 0.001$ (left) and $\sigma = 0.5$ (right). It is clear that ε assumes pretty "large" values: data do not fancy the nonparametric model ($\varepsilon = 0$). In all the experiments, the density estimates seem to fit well the data. Figure 8 shows the marginal predictive densities for case (F). We have not observed substantial differences in the rest of the cases.

For experiment (G), we set a vague prior for κ, and a more informative prior on σ to speed up and improve the mixing; the posterior of (σ, κ) is displayed in Figure 9, showing a noteworthy update of the prior to the posterior.

The reference partition into five groups in Cho et al. (1998) was obtained by visual inspection. In order to provide cluster estimates with our model (19), we adopt a standard approach in the Bayesian framework. First of all, remind that (19) induces a prior for the random partition $\boldsymbol{p}_n = \{C_1, \ldots, C_k\}$ of the data labels (see notation in Section 3), so that the cluster estimates are based on its posterior. As such an estimate we consider $\widehat{\boldsymbol{p}}_n$ minimizing the so-called Binder loss function with equal misclassification costs, using the

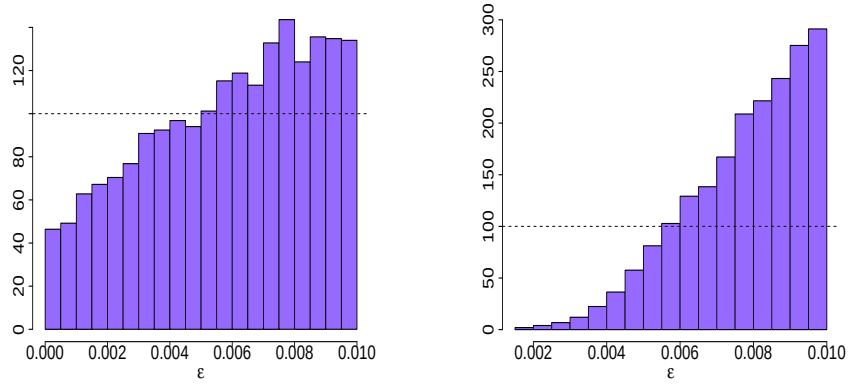


Figure 7: Posterior distribution of ε for case (F): $\sigma = 0.001$ (left) and $\sigma = 0.5$ (right). The prior is $\mathcal{U}(0, 0.01)$ (dashed).

same approach as in Argiento et al. (2013). To compare different cluster estimates, we evaluate the posterior expectation $\mathbb{E}(H(\pi_n)|data)$ where the function H is a standard tool as the silhouette coefficient or the adjusted Rand index. We compared cluster estimates for more sets of hyperparameters than those reported here; see Bianchini (2014). In Figure 10 we report one of the best cluster estimate, which was obtained when hyperparameters are those of case (G). The Silhouette coefficient in any group can be computed, obtaining

(a)	(b)	(c)	(d)	(e)	(f)
0.22	0.23	0.22	0.04	0.18	0.14

Compared to other experiments we did, these figures indicate a good clustering. Note that there is only one group (d), with a coefficient near to 0: indeed, it has a large empirical variance with respect to the other clusters. On the other hand, while the first two clusters are very similar to first two in the reference partition in Cho et al. (1998), in the rest of the groups we seem to tide up their partition. The posterior mean of the overall Silhouette coefficient is 0.2.

As a final remark, we would like to point out that all the cluster estimates, here and in Bianchini (2014), were robust with respect to the choice of the prior of $(\varepsilon, \sigma, \kappa)$, while, on the contrary, they are very sensible with respect to P_0.

7 Conclusions

We have proposed a new model for density and cluster estimates in the Bayesian nonparametric framework. In particular, a finite dimensional process, the ε-NGG process, has been defined, which converges in distribution to the well-known NGG process, when ε tends to 0. Here, the ε-NGG process is the mixing measure in a mixture model.

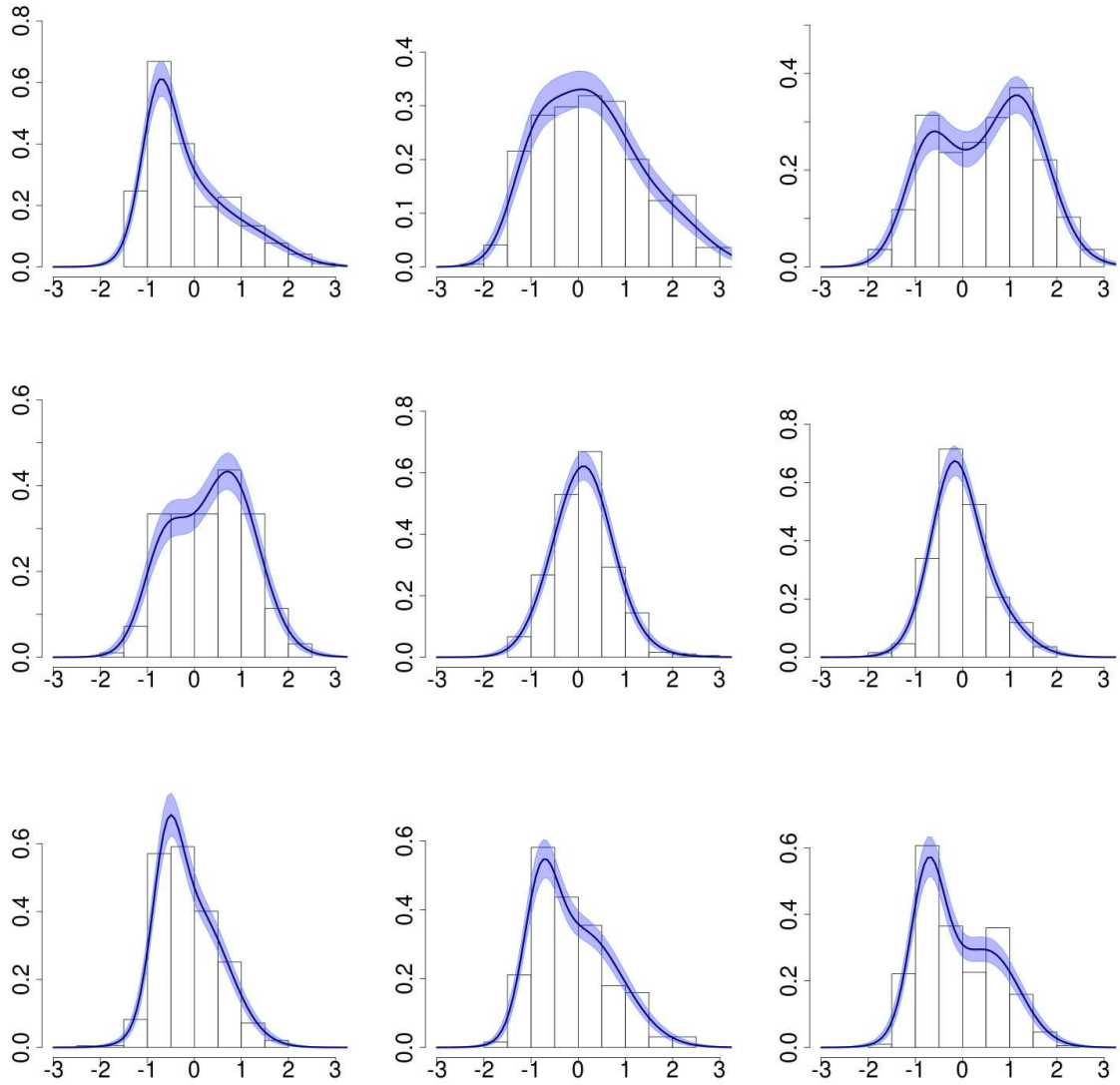


Figure 8: Marginal density estimates for experiment (F) when $\sigma = 0.001$, $\kappa = 0.7$, $\varepsilon \sim \mathcal{U}(0, 0.01)$. The shaded regions denote 90% CI's around the density estimates.

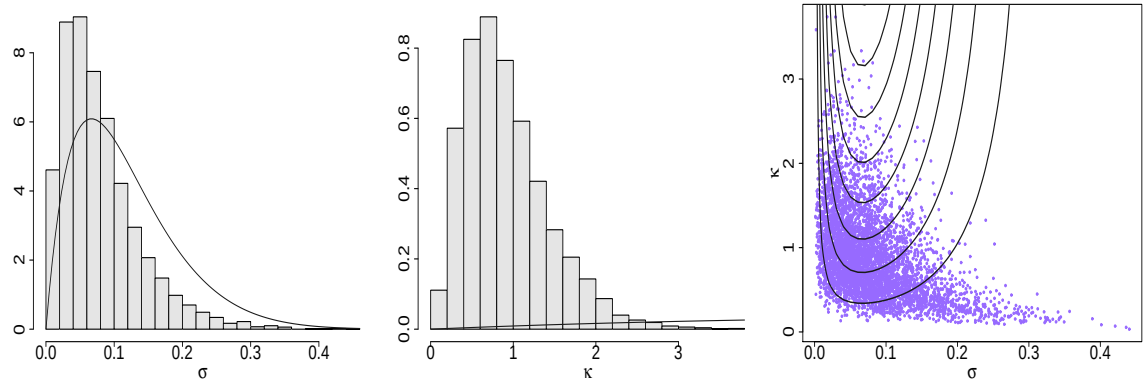


Figure 9: Posteriors of σ (left), κ (center), and (σ, κ) (right). The priors are superimposed as gray lines.

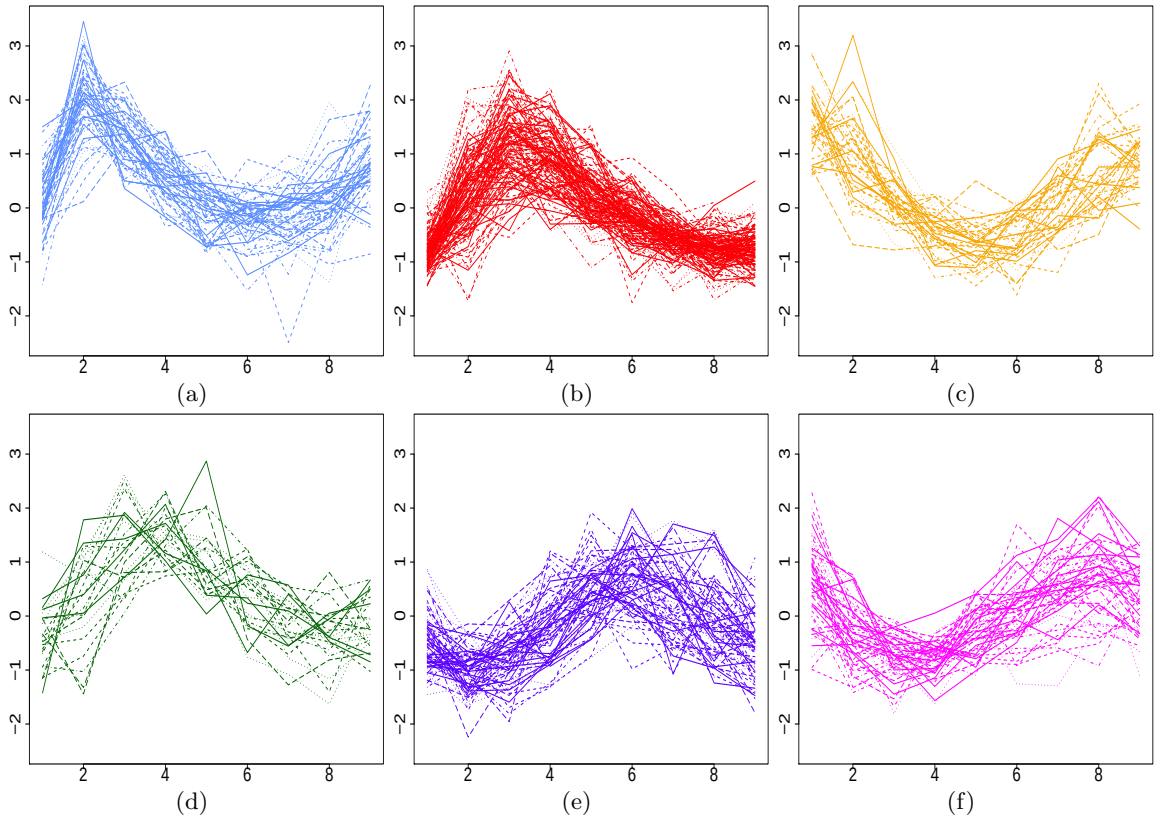


Figure 10: Bayesian cluster estimates for experiment (G).

An interesting achievement is that, as ε varies, a large range of models can be obtained: from a nonparametric NGG mixture model, when ε decreases to 0, to a parametric model, when ε assumes large values. Hence, on one hand, the model can be used as an approximation of a NGG mixture model on which many theoretical results are available in the literature. On the other hand, our process can be viewed as a model different from the NGG process, with

a new prior: since it is finite dimensional, the inference will be quite simple. Furthermore, the precision parameter ε can be considered as a random variable, once we have elicited a prior for it: in this case, the data and the prior, via the posterior, drive the degree of approximation. Of course, under this model, the posterior distribution must be computed via simulation methods: a Gibbs sampler algorithm has been built to reach this goal. All the updating steps are as easy to implement as in the popular DPM model, but the new model is more flexible. In addition, thanks to the finite approximation, there is no need to integrate out the mixing component (i.e. the infinite dimensional parameter) itself, thus pursuing a full nonparametric Bayesian inference, in order to get posterior estimates of linear and non linear functionals of the population distribution.

We have illustrated our proposal through a density estimation problem: thanks to an extensive robustness analysis, the role and the influence of the parameters ε, σ and κ of our prior on the mixing of the chain and on posterior estimates have been clarified; moreover, the robustness of the model with respect to the choice of the hyperparameters has been checked. In addition to density estimation, a clustering problem has also been tackled in the multivariate case; the cluster estimates are pretty satisfactory.

As far as the drawbacks of the model are concerned, the first issue consists in the choice of the mean distribution P_0. As in each Bayesian nonparametric mixture model, especially when the dimension of data is large, P_0 strongly affects the estimates and the mixing of the MCMC chains. A second problem concerns the parameter σ: when it assumes values close to 1, on one hand the computation becomes difficult because of the presence of the incomplete gamma functions in the algorithm, which are very unstable in this case, while, on the other, correlation between U and ε heavily increases. Moreover, the number of components in the mixture grows very fast with σ, slowing down the run-time of the algorithm.

APPENDIX: DETAILS ON THE BLOCKED GIBBS SAMPLER

First of all, the joint law of data and parameters can be written as follows:

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{X},\boldsymbol{\theta},u,P_\varepsilon,\varepsilon,\sigma,\kappa) = \mathcal{L}(\boldsymbol{X}|\boldsymbol{\theta},u,P_\varepsilon,\varepsilon,\sigma,\kappa)\mathcal{L}(\boldsymbol{\theta},u,P_\varepsilon|\varepsilon,\sigma,\kappa)\mathcal{L}(\varepsilon,\sigma,\kappa)\\
&=\prod_{i=1}^{n} k(X_i;\theta_i)\mathcal{L}(\boldsymbol{\theta},u,P_\varepsilon|\varepsilon,\sigma,\kappa)\pi(\varepsilon)\pi(\sigma)\pi(\kappa)\\
&=\frac{u^{n-1}}{\Gamma(n)}\prod_{j=0}^{N_\varepsilon}\left(e^{-uJ_j}\rho_\varepsilon(J_j)P_0(\tau_j)\right)\sum_{l_1^*,..,l_k^*}\Big(J_{l_1^*}^{n_1}\prod_{i\in C_1}k(X_i;\theta_1^*)\delta_{\tau_{l_1^*}}(\theta_1^*)..\\
&\quad ..J_{l_k^*}^{n_k}\prod_{i\in C_k}k(X_i;\theta_k^*)\delta_{\tau_{l_k^*}}(\theta_k^*)\Big)\frac{\Lambda_\varepsilon^{N_\varepsilon}e^{-\Lambda_\varepsilon}}{N_\varepsilon!}\pi(\varepsilon)\pi(\sigma)\pi(\kappa),
\end{aligned}
\tag{23}
$$

where we used the hierarchical structure in (19). Note that $\mathcal{L}(\boldsymbol{\theta}, u, P_\varepsilon|\varepsilon, \sigma, \kappa)$ has been computed in (12). Now we derive every step of the Gibbs sampler in Figure 1.

1. The first step is straightforward, since

$$\mathcal{L}(u|\boldsymbol{X}, \boldsymbol{\theta}, P_\varepsilon, \varepsilon, \sigma, \kappa) \propto \mathcal{L}(u, \boldsymbol{X}, \boldsymbol{\theta}, P_\varepsilon, \varepsilon, \sigma, \kappa).$$

2. Thanks to the hierarchical structure of the model, the following relation holds true:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}, P_\varepsilon, \varepsilon, \sigma, \kappa, u) &\propto \prod_{i=1}^{n} k(X_i; \theta_i) \sum_{j=0}^{N_\varepsilon} J_j \delta_{\tau_j}(\theta_i) \\ &= \prod_{i=1}^{n} \sum_{j=0}^{N_\varepsilon} J_j k(X_i; \theta_i) \delta_{\tau_j}(\theta_i) = \prod_{i=1}^{n} J_i k(X_i; \tau_i),\end{aligned}$$

   therefore the second step is recovered.

3. As far as $\mathcal{L}(P_\varepsilon, \varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}, \boldsymbol{X})$ is concerned, we have

$$\mathcal{L}(P_\varepsilon, \varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}, \boldsymbol{X}) = \mathcal{L}(P_\varepsilon, \varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}) = \mathcal{L}(P_\varepsilon|\varepsilon, \sigma, \kappa, u, \boldsymbol{\theta}) \mathcal{L}(\varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}),$$

   so that step 3. can be split into two consecutive substeps. First we simulate from $\mathcal{L}(\varepsilon, \sigma, \kappa|u, \boldsymbol{\theta})$ as follows: we integrate out N_ε (or equivalently N_{na}) from (16) and obtain

$$\begin{aligned}\mathcal{L}(\varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}, \boldsymbol{X}) &\propto \sum_{N_{na}=0}^{+\infty} \mathcal{L}(N_{na}, \varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}, \boldsymbol{X}) \\ &= \frac{u^{n-1}}{\Gamma(n)} \left(\frac{\kappa}{\Gamma(1-\sigma)}\right)^{k-1} \prod_{i=1}^{k} \Big[\Gamma(n_i - \sigma, \varepsilon(u+\omega))\Big] \pi(\varepsilon)\pi(\sigma)\pi(\kappa) \\ &\quad \times \frac{(\omega+u)^{\sigma k - n}}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} e^{\Lambda_{\varepsilon,u} - \Lambda_\varepsilon} \left(\Lambda_{\varepsilon,u} + k\right).\end{aligned}$$

   In practice step 3.a can be obtained in three substeps:

$$\mathcal{L}(\varepsilon|u, \boldsymbol{\theta}, \boldsymbol{X}) \propto \prod_{i=1}^{k} \Gamma(n_i - \sigma, \varepsilon(u+\omega)) e^{(\Lambda_{\varepsilon,u} - \Lambda_\varepsilon)} \frac{\Lambda_{\varepsilon,u} + k}{\Gamma(-\sigma, \omega\varepsilon)} \pi(\varepsilon), \tag{24}$$

$$\begin{aligned}\mathcal{L}(\sigma|u, \boldsymbol{\theta}, \boldsymbol{X}) &\propto \frac{(u+\omega)^{k\sigma}}{\omega^\sigma} \frac{\Lambda_{\varepsilon,u} + k}{\Gamma(-\sigma, \omega\varepsilon)} \prod_{i=1}^{k} \Gamma(n_i - \sigma, \varepsilon(u+\omega)) \\ &\quad \times e^{(\Lambda_{\varepsilon,u} - \Lambda_\varepsilon)} \Gamma(1-\sigma)^{1-k} \pi(\sigma),\end{aligned} \tag{25}$$

$$\mathcal{L}(\kappa|u, \boldsymbol{\theta}, \boldsymbol{X}) = p_1 gamma(\alpha + k, R + \beta) + (1 - p_1) gamma(\alpha + k - 1, R + \beta), \tag{26}$$

where

$$R = \frac{\omega^{\sigma}\Gamma(-\sigma, \varepsilon\omega)}{\Gamma(1-\sigma)} - \frac{(\omega+u)^{\sigma}\Gamma(-\sigma, \varepsilon(\omega+u))}{\Gamma(1-\sigma)}$$

and

$$p_1 = \frac{(\alpha+k-1)(u+\omega)^{\sigma}\Gamma(-\sigma, \varepsilon(\omega+u))}{(\alpha+k-1)(u+\omega)^{\sigma}\Gamma(-\sigma, \varepsilon(\omega+u)) + k(R+\beta)\Gamma(1-\sigma)}.$$

Here we assume that $\pi(\kappa)$ is $gamma(\alpha, \beta)$. Step 3.b consists in sampling from $\mathcal{L}(P_{\varepsilon}|\varepsilon, \sigma, \kappa, u, \boldsymbol{\theta})$ and has already been described in Section 4.

References

Argiento, R., Cremaschi, A., and Guglielmi, A. (2013). "A Density-Based Algorithm for Cluster Analysis Using Species Sampling Gaussian Mixture Models." *Journal of Computational and Graphical Statistics*, Latest Articles.

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.

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

Bianchini, I. (2014). "A Bayesian nonparametric model for density and cluster estimation: the ε-NGG mixture model." Tesi di laurea magistrale, Ingegneria Matematica, Politecnico di Milano.

Caron, F. (2012). "Bayesian nonparametric models for bipartite graphs." In *NIPS*, 2060–2068.

Caron, F. and Fox, E. B. (2014). "Bayesian nonparametric models of sparse and exchangeable random graphs." *arXiv preprint arXiv:1401.1137*.

Chen, C., Ding, N., and Buntine, W. (2012). "Dependent hierarchical normalized random measures for dynamic topic modeling." *arXiv preprint arXiv:1206.4671*.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). "A genome-wide transcriptional analysis of the mitotic cell cycle." *Molecular Cell*, 2, 65–73.

Escobar, M. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *J. Amer. Statist. Assoc.*, 90, 577–588.

Favaro, S., Guglielmi, A., and Walker, S. (2012). "A class of measure-valued Markov Chains and Bayesian Nonparametrics." *Bernoulli*, 18(3), 1002–1030.

Favaro, S. and Teh, Y. (2013). "MCMC for Normalized Random Measure Mixture Models." *Statistical Science*, 28(3), 335–359.

Favaro, S. and Walker, S. G. (2013). "Slice sampling σ-stable Poisson-Kingman mixture models." *Journal of Computational and Graphical Statistics*, 22(4), 830–847.

Ferguson, T. S. and Klass, M. (1972). "A representation of independent increment processes without Gaussian components." *Ann. Math. Statist.*, 43, 1634–1643.

Gelfand, A. E. and Kottas, A. (2002). "A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models." *J. Comput. Graph. Statist.*, 11, 289–305.

Gradshteyn, I. and Ryzhik, L. (2000). *Table of integrals, series, and products - Sixth Edition*. San Diego (USA): Academic Press, sixth edition.

Griffin, J. and Walker, S. G. (2011). "Posterior Simulation of Normalized Random Measure Mixtures." *Journal of Computational and Graphical Statistics*, 20, 241–259.

Griffin, J. E. (2013). "An adaptive truncation method for inference in Bayesian nonparametric models." *arXiv preprint arXiv:1308.2045*.

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

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

Ishwaran, H. and Zarepour, M. (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models." *Biometrika*, 87, 371–390.

— (2002). "Exact and approximate sum representations for the Dirichlet process." *Canadian Journal of Statistics*, 30, 269–283.

James, L., Lijoi, A., and Prünster, I. (2009). "Posterior analysis for normalized random measures with independent increments." *Scand. J. Statist.*, 36, 76–97.

Kallenberg, O. (1983). *Random Measures*. Berlin: Akademie-Verlag, fourth edition.

Kingman, J. F. C. (1993). *Poisson processes*, volume 3. Oxford university press.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Controlling the reinforcement in Bayesian nonparametric mixture models." *Journal of the Royal Statistical Society B*, 69, 715–740.

Lijoi, A., Nipoti, B., and Prunster, I. (2014). "Bayesian inference with dependent normalized completely random measures." *Bernoulli*, Forthcoming papers.

Lijoi, A., Prunster, I., and Walker, S. G. (2008). "Investigating nonparametric priors with Gibbs structure." *Statistica Sinica*, 18, 1653–1668.

MacEachern, S. N. (1998). "Computational methods for mixture of Dirichlet process models." In *Practical nonparametric and semiparametric Bayesian statistics*, volume 133 of *Lecture Notes in Statist.*, 23–43. New York: Springer.

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

Neal, R. (2000). "Markov Chain sampling Methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9, 249–265.

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95, 169–186.

Pitman, J. (1996). "Some Developments of the Blackwell-Macqueen urn Scheme." In Ferguson, T. S., Shapley, L. S., and B., M. J. (eds.), *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, volume 30 of *IMS Lecture Notes-Monograph Series*, 245–267. Hayward (USA): Institute of Mathematical Statistics.

— (2003). "Poisson-Kingman Partitions." In *Science and Statistics: a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes-Monograph Series*, 1–34. Hayward (USA): Institute of Mathematical Statistics.

— (2006). *Combinatorial Stochastic Processes*. LNM n. 1875. New York: Springer.

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.

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

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Commun. Stat. Simulat.*, 36, 45–54.