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

Raffaele Argiento · Ilaria Bianchini · Alessandra Guglielmi

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 Poisson processes 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 assume such a process as the mixing measure in a mixture model for density and cluster estimation, and build an efficient Gibbs sampler scheme to simulate from the posterior. Finally, we discuss applications and performance of the model to two popular datasets, as well as comparison with competitor algorithms, the slice sampler and a posteriori truncation.

Keywords Bayesian nonparametric mixture models · Normalized generalized gamma process · Blocked Gibbs sampler · Finite dimensional approximation · A priori truncation method

R. Argiento (✉)
CNR-IMATI, via Bassini 15, 20133 Milan, Italy
e-mail: raffaele@mi.imati.cnr.it

I. Bianchini · A. Guglielmi
Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milan, Italy
e-mail: ilaria.bianchini@polimi.it

A. Guglielmi
e-mail: alessandra.guglielmi@polimi.it

1 Introduction

The major goal of this work is the definition of a new class of nonparametric priors, which can be considered as an approx-imation 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 probabil-ity 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 parame-ters. NRMIs are a popular tool in a mixture context, where they are usually considered as mixing measures of para-metric densities for continuous data, and, as a consequence, also NRMI mixtures include infinite parameters. There are two main approaches to deal with this computational prob-lem, namely marginal and conditional Gibbs sampler algo-rithms for sampling from the posterior. The former inte-grate out the infinite dimensional parameter (i.e. the ran-dom probability), resorting to generalized Polya urn schemes (MacEachern 1998); see Neal (2000) for a review on the subject. Recently, Favaro and Teh (2013) developed algo-rithms 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

This is a post-peer-review, pre-copyedit version of an article published in Stat Comput, 2016, 26:641-661. The final authenticated version is available online at: http://dx.doi.org/10.1007%2Fs11222-015-9549-6

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 approxi-mated by truncation of the infinite sum defining the process. Truncation can be achieved a posteriori, when one approxi-mates 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 implemented. 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 (2014), 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 (2014), 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 Sethu-raman (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 Non-parametrics, we will consider this new discrete random prob-ability 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 is given, as well as for all the other scalar parameters defining the new process. As a main achievement of this paper, we design a blocked Gibbs sampler algorithm to simulate from the posterior. Moreover, we discuss guidelines for choosing the prior on ε.

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. Den-sity estimates are shown for the two applications, together with a thorough robustness analysis of the estimates with respect to prior choice, in particular in order to investigate the effect of the approximation parameter ε. In addition, for the Galaxy data, we compare our algorithm to a conditional method, the slice sampler by Griffin and Walker (2011), and to the simple adaptive truncation method in Argiento et al.(2010); in particular, the integrated autocorrelation times of the number of clusters and of the deviance of the estimated density show evidence in favor of the performances of our algorithm.

In Sect. 2 we introduce notation on homogeneous NRMIs, while in Sect. 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; besides, we provide an interpretation of parameter ε and suggest a family of marginal priors for ε itself. Section 5 (Galaxy data) and 6 (Yeast cell cycle data) discuss the two applications; in particular, Sect. 5.1 presents the comparison between our model and the other two from the literature. The article ends with wrap-up of the proposed model as well as with possible future developments in Sect. 7. Proofs are grouped in the Appendices.

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 ran-dom measure μ on Θ is completely random if for any finite sequence B_1, B_2,..., B_k of disjoint Borel sets in Θ, $\mu(B_1)$, $\mu(B_2)$, ... , $\mu(B_k)$ are independent. A purely atomic completely random measure is defined (see Kingman 1993, Sect. 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 Pois-

son 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 condition:

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

If $T := \mu(\Theta) = \sum_{j\geq 1} J_j$, 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:

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

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 $\rho(s) = (\kappa/\Gamma(1-\sigma))\, s^{-1-\sigma} e^{-\omega s}\mathbb{I}_{(0,+\infty)}(s)$, 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 $\omega > 0$ and $\sigma = 0$, the Dirichlet process (DP) is recovered.

One of the main arguments in favor 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. $N_\varepsilon \sim \mathscr{P}_0(\Lambda_\varepsilon)$ where

$$\Lambda_\varepsilon := \int_\varepsilon^{+\infty} \rho(x)dx = \frac{\kappa\omega^\sigma}{\Gamma(1-\sigma)}\Gamma(-\sigma, \omega\varepsilon), \tag{3}$$

and $\Gamma(a, x) = \int_x^{+\infty} t^{a-1}\mathrm{e}^{-t}dt$ 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

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

thanks to the well-known relationship between Poisson and Bernoulli processes; see, for instance, Kingman (1993), Sect. 2.4. Observe that ρ_ε is defined by restricting ρ to the interval $(\varepsilon, +\infty)$ first, and then normalizing by Λ_ε. 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 by ρ_ε, 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_j's, 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:

$$P_\varepsilon = \sum_{j=0}^{N_\varepsilon} P_j \delta_{\tau_j} = \frac{1}{T_\varepsilon}\sum_{j=0}^{N_\varepsilon} J_j \delta_{\tau_j}, \tag{4}$$

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 (4) 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. Let $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n)$ be a finite sample from a species sampling model P. We denote by $\boldsymbol{\theta}^* := (\theta_1^*, \ldots, \theta_k^*)$ the vector

of its unique distinct values. Then, $\boldsymbol{\theta}$ 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 $j = 1, \ldots, k$ and $\#C_i = n_i$ for $1 \leq i \leq k$. The marginal law of $\boldsymbol{\theta}$ has unique characterization in terms of $\boldsymbol{\theta}^*$ and the exchangeable partition $\boldsymbol{p}_n$ as follows:

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

where p is the exchangeable partition probability function (eppf) associated to P (see Pitman 1996); the eppf p induces a probability law on the set of the partitions of $\mathbb{N}_n$.

An analytic expression of the eppf of P_ε defined in (4) can be recovered resorting to (30) in Pitman (1996): if $P_\varepsilon \sim \varepsilon\text{−NGG}(\sigma, \kappa, \omega, P_0)$, then its eppf is

$$p_\varepsilon(n_1, \ldots, n_k) = \int_0^{+\infty} (u+\omega)^{k\sigma - n} \prod_{i=1}^{k} \Gamma(n_i - \sigma, (u+\omega)\varepsilon) \times \frac{\kappa^{k-1}}{\Gamma(1-\sigma)^{k-1}} \frac{\Lambda_{\varepsilon,u} + k}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} \mathrm{e}^{\Lambda_{\varepsilon,u} - \Lambda_\varepsilon} \frac{u^{n-1}}{\Gamma(n)} du, \quad (5)$$

for any $n_1, \ldots, n_k$ positive integers such that $\sum_{i=1}^k n_i = n$, where

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

Details of the proof of (5) are in the Appendices.

Now we state the main distributional result on P_ε, that is convergence in distribution of the sequence of ε-NGG processes to the NGG process; the proof of the following proposition requires, as a preliminary result, to show convergence of the sequence of the eppfs $p_\varepsilon(n_1, \ldots, n_k)$ to the eppf associated to the NGG process as $\varepsilon \to 0$; see Lemma 1 in the Appendices.

Proposition 1 *For any $\varepsilon > 0$, let P_ε be a ε-NGG$(\sigma, \kappa, \omega, P_0)$ process. Then we have*

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

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

As before, 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 (4), 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 (4) such that there exists a corresponding location for which $\tau_{l_i^*} = \theta_i^*$, $i = 1, \ldots, k$. The remaining values are called *non-allocated* jumps. We use the superscript (na) for random variables related to *non-allocated* jumps. Before stating the proposition, we introduce the random variable $U := \Gamma_n / T_\varepsilon$, where $\Gamma_n \sim gamma(n, 1)$, being Γ_n and T_ε independent. It will be clear that this variable is decisive when simulating posterior trajectories of ε−NGG processes.

Proposition 2 *If P_ε is an ε−NGG$(\sigma, \kappa, \omega, P_0)$ process, then the conditional distribution of P_ε, given $\boldsymbol{\theta}^*$ and $U = u$, satisfies the following distributional equation*

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

where

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

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

   being $\Lambda_{\varepsilon,u}$ as defined in (6), and denoting $\mathscr{P}_i(\lambda)$ the shifted Poisson distribution on $\{i, i+1, i+2, \ldots\}$ with mean $i + \lambda$, $i = 0, 1$;
2. *the* allocated *jumps $\{P_1^{(a)}, \ldots, P_k^{(a)}\}$ associated to the fixed points of discontinuity $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)$ of P_ε^* are obtained by normalization of $J_j^{(a)} \stackrel{ind}{\sim} gamma(n_j - \sigma, u + \omega)\mathbb{I}_{(\varepsilon, +\infty)}$, for $j = 1 \ldots, k$;*
3. *$P_{\varepsilon,u}^{(na)}(\cdot)$ 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 as 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 (4);*
5. *the posterior law of U given $\boldsymbol{\theta}^*$ has density*

$$f_{U|\boldsymbol{\theta}^*}(u|\boldsymbol{\theta}^*) \propto u^{n-1}(u+\omega)^{k\sigma - n}(\Lambda_{\varepsilon,u} + k)\mathrm{e}^{\Lambda_{\varepsilon,u}} \times \prod_{i=1}^{k} \Gamma(n_i - \sigma, (u+\omega)\varepsilon)\mathbb{I}_{(0,+\infty)}(u).$$

Observe that this proposition is merely a characterization of 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

described as the law of a random probability measure, which is a mixture between a ε-NGG process and a discrete probability measure with support given by the (observed) distinct values $\boldsymbol{\theta}^*$.

As a final remark in this section, we point out that our approximation P_ε (and the two propositions above) holds true for two particular cases of NGG processes too, namely the Dirichlet and the normalized σ-stable process, also known as the two parameters $(\sigma, 0)$ Poisson-Dirichlet process, when $0 < \sigma < 1$. In particular, the construction is as the one described here, on the ground of the following expressions:

$$\Lambda_\varepsilon = -\kappa\, Ei(-\omega\varepsilon),$$
$$\rho_\varepsilon(s) = -s^{-1}/Ei(-\omega\varepsilon)\mathrm{e}^{-\omega s}\mathbb{I}_{(\varepsilon,+\infty)}(s)$$

for the Dirichlet process, where $Ei(x) = \int_{-\infty}^{x} \mathrm{e}^t/t\, dt, x < 0$, is the exponential integral function, and

$$\Lambda_\varepsilon = \kappa\varepsilon^{-\sigma}/\left(\sigma\Gamma(1-\sigma)\right),$$
$$\rho_\varepsilon(s) = \sigma\varepsilon^{\sigma}s^{-1-\sigma}\mathbb{I}_{(\varepsilon.+\infty)}(s)$$

for the normalized σ-stable process.

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 parameterization is not unique) and change notation accordingly, i.e. ε-NGG(σ, κ, P_0). Details on the specific choices of P_0 are illustrated in Sects. 5 and 6. The model we assume is the following:

$$\begin{aligned}
&X_i|\theta_i \overset{\text{ind}}{\sim} k(\cdot;\theta_i),\ i = 1,\dots,n\\
&\theta_1,\dots,\theta_n|P_\varepsilon \overset{\text{iid}}{\sim} P_\varepsilon\\
&P_\varepsilon \sim \varepsilon - NGG(\sigma,\kappa,P_0)\ \text{process prior},\\
&\varepsilon,\sigma,\kappa \sim \pi(\varepsilon)\times\pi(\sigma)\times\pi(\kappa),
\end{aligned} \tag{7}$$

where $k(\cdot;\theta_i)$ is a parametric family of densities on $\mathbb{X}\subset\mathbb{R}^p$, for all $\theta\in\Theta$. In the rest of the paper, we assume the Gaussian kernel, where θ_i denotes the mean 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 any Borel set A and any $\varepsilon \geq 0$. Model (7) will be addressed here as *ε−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

$$f(x) = \int_\Theta k(x;\theta)P_\varepsilon(d\theta) = \sum_{j=0}^{N_\varepsilon} P_j\, k(x;\tau_j). \tag{8}$$

In general, computation of posterior inference for (7), when P_ε is substituted by a NGG process P, is not straightforward, since this model assumes an infinite number of parameters. As we mentioned in the Introduction, different approaches have been proposed in the literature. Here we exploit a prior truncation approach; as a matter of fact, from the algorithmic point of view, the finite dimensionality of the $\varepsilon - NGG$ process is a key point, since it allows expressing 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.

$$\begin{aligned}
&\mathscr{L}(\boldsymbol{X}|\boldsymbol{\theta})\mathscr{L}(\boldsymbol{\theta}|P_\varepsilon)\mathscr{L}(P_\varepsilon,\varepsilon,\sigma,\kappa)\\
&\quad = \mathscr{L}(\boldsymbol{X},\boldsymbol{\theta}|P_\varepsilon)\mathscr{L}(P_\varepsilon|\varepsilon,\sigma,\kappa)\mathscr{L}(\varepsilon,\sigma,\kappa).
\end{aligned} \tag{9}$$

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

$$\begin{aligned}
\mathscr{L}(\boldsymbol{X},\boldsymbol{\theta}|P_\varepsilon) &= \prod_{i=1}^{n} P_\varepsilon(\theta_i)k(X_i;\theta_i)\\
&= \left(\prod_{i\in C_1} k(X_i;\theta_1^*)\dots\prod_{i\in C_k} k(X_i;\theta_k^*)\right)\\
&\quad\times\left(\sum_{l_1^*,\dots,l_k^*} P_{l_1^*}^{n_1}\dots P_{l_k^*}^{n_k}\delta_{\tau_{l_1^*}}(\theta_1^*)\dots\delta_{\tau_{l_k^*}}(\theta_k^*)\right)\\
&= \frac{1}{T_\varepsilon^n}\sum_{l_1^*,\dots,l_k^*}\left(J_{l_1^*}^{n_1}\prod_{i\in C_1}k(X_i;\theta_1^*)\dots J_{l_k^*}^{n_k}\prod_{i\in C_k}k(X_i;\theta_k^*)\right),
\end{aligned} \tag{10}$$

while $\mathscr{L}(P_\varepsilon|\varepsilon,\sigma,\kappa)$ is the finite dimensional distribution of P_ε in Sect. 3, and the joint law $\mathscr{L}(\varepsilon,\sigma,\kappa) = \pi(\varepsilon)\pi(\sigma)\pi(\kappa)$ will be elicited in Sects. 5 and 6. In addition, we provide some guidelines on the choice of $\pi(\varepsilon)$ and discuss the prior mean of N_ε at the end of this section. We use the same notation as in the proof of Proposition 2. We augment the state space and apply Proposition 2, considering also the random variable U. Therefore, the sample space of the Gibbs sampler is the set of all values of the *parameter* $(\boldsymbol{\theta}, P_\varepsilon, \varepsilon, u, \sigma, \kappa)$. Consequently, the joint law of data and parameters can be written as follows:

$$\begin{aligned}
&\mathscr{L}(\boldsymbol{X},\boldsymbol{\theta},u,P_\varepsilon,\varepsilon,\sigma,\kappa)\\
&= \mathscr{L}(\boldsymbol{X}|\boldsymbol{\theta},u,P_\varepsilon,\varepsilon,\sigma,\kappa)\mathscr{L}(\boldsymbol{\theta},u,P_\varepsilon|\varepsilon,\sigma,\kappa)\mathscr{L}(\varepsilon,\sigma,\kappa)\\
&= \prod_{i=1}^{n} k(X_i;\theta_i)\mathscr{L}(\boldsymbol{\theta},u,P_\varepsilon|\varepsilon,\sigma,\kappa)\pi(\varepsilon)\pi(\sigma)\pi(\kappa)
\end{aligned}$$

$$= \frac{u^{n-1}}{\Gamma(n)} \prod_{j=0}^{N_\varepsilon} \left(\mathrm{e}^{-uJ_j} \rho_\varepsilon(J_j) P_0(\tau_j) \right) \frac{\Lambda_\varepsilon^{N_\varepsilon} \mathrm{e}^{-\Lambda_\varepsilon}}{N_\varepsilon!} \tag{11}$$

$$\times\ \pi(\varepsilon)\pi(\sigma)\pi(\kappa) \sum_{l_1^*,\ldots,l_k^*} \left(J_{l_1^*}^{n_1} \prod_{i \in C_1} k(X_i;\theta_1^*) \delta_{\tau_{l_1^*}}(\theta_1^*) \right.$$

$$\left. \cdots J_{l_k^*}^{n_k} \prod_{i \in C_k} k(X_i;\theta_k^*) \delta_{\tau_{l_k^*}}(\theta_k^*) \right),$$

where we used the hierarchical structure in (7). Note that $\mathscr{L}(\boldsymbol{\theta}, u, P_\varepsilon | \varepsilon, \sigma, \kappa)$ coincides with distribution in (16). Full details of the blocked Gibbs sampler can be found in Appendix 6; however, in the following steps, we sketch all the full-conditionals:

1. Sampling from $\mathscr{L}(u|\boldsymbol{X}, \boldsymbol{\theta}, P_\varepsilon, \varepsilon, \sigma, \kappa)$: since the joint law of data and parameter (see (11)) 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 $\mathscr{L}(\boldsymbol{\theta}|u, \boldsymbol{X}, P_\varepsilon, \varepsilon, \sigma, \kappa)$: by (11), each θ_i, for $i = 1, \ldots, n$, has discrete law with support $\{\tau_0, \ldots, \tau_{N_\varepsilon}\}$ and probabilities equal to $\mathbb{P}(\theta_i = \tau_j) \propto J_j k(X_i; \tau_j)$.
3. Sampling from $\mathscr{L}(P_\varepsilon, \varepsilon, \sigma, \kappa | u, \boldsymbol{\theta}, \boldsymbol{X})$: this step is not straightforward. In the Appendix 6 we show that it can be split into two consecutive substeps:

   3a. Sampling from $\mathscr{L}(\varepsilon, \sigma, \kappa | u, \boldsymbol{\theta}, \boldsymbol{X})$: a Gibbs sampler strategy will achieve it. For a detailed description of the full conditionals (i) $\mathscr{L}(\varepsilon|\sigma, \kappa, u, \boldsymbol{\theta}, \boldsymbol{X})$, (ii) $\mathscr{L}(\sigma|\varepsilon, \kappa, u, \boldsymbol{\theta}, \boldsymbol{X})$ and (iii) $\mathscr{L}(\kappa|\varepsilon, \sigma, u, \boldsymbol{\theta}, \boldsymbol{X})$, we refer to formulas (23), (24), (25) in the Appendices.

   3b. Sampling from $\mathscr{L}(P_\varepsilon|\varepsilon, \sigma, \kappa, u, \boldsymbol{\theta}, \boldsymbol{X})$: via characterization of the posterior in Proposition 2, since this distribution is equal to $\mathscr{L}(P_\varepsilon|\varepsilon, \sigma, \kappa, u, \boldsymbol{\theta})$. As a matter of fact, 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 Fig. 1. With regard to 3b. v, we do not directly apply Proposition 2, but add an acceleration step (see for instance Argiento et al. 2010) sampling from the distribution in Fig. 1. When sampling from non-standard distributions, we acknowledge that Accept–Reject or Metropolis–Hastings algorithms have been exploited.

We believe that a broader discussion on the prior $\mathscr{L}(\varepsilon, \sigma, \kappa)$ of the scalar parameters of the ε-NGG process is needed, or at least on ε, which is new with respect to the parameters of the NGG process. Here we have designed the algorithm when $\varepsilon, \sigma, \kappa$ are a priori independent, and possibly degenerate. However, it is easy to extend the algorithm in Fig. 1 to the case of a priori dependence. As a first remark, we stress that a large ε may easily lead to a prior for P_ε degenerate on δ_{τ_0}, $\tau_0 \sim P_0$, since less jumps in (2) will be considered in definition (4). Equivalently, from (3), it is easy to check that $\mathbb{E}(\Lambda_\varepsilon) \to 0$ for $\varepsilon \to +\infty$. On the other hand, the statistical goal here is the analysis of the prior and the algorithm for ε small, because this leads to an approximation of NGG process mixtures. Since T denotes the sum of all unnormalized jumps defining the NGG process (see (2)), we consider this random variable, or a summary of it (e.g. $\mathbb{E}(T)$) to select the jumps to include in (4). As a guideline, we suggest to assume ε random on a bounded interval $(0, \delta)$, with δ given by the minimum between a small number (0.1, say) and $\mathbb{E}(T)$. For instance, in experiment (C) in Sect. 5, we fix a prior for ε (when κ is fixed) that is a scaled beta on $(0, \delta)$. As an alternative, we recommend to fix a prior with full support $(0, +\infty)$, but in such a way that most mass is concentrated around 0. See Bianchini (2015) for further details.

In our opinion, parameter ε can be considered either as a "true" parameter as σ and κ, and the prior on it should be given on the ground of the prior information we have, or as a "tuning" parameter useful to approximate the exact NGG-process mixture, if it is credited to be "true" model. In the latter case, even if it is assumed random, ε has to be "small". On the other hand, if we do not believe in the infinite mixture model, but our prior belief supports a finite mixture model, a prior on ε favoring large values would be a better choice.

It is also worth describing prior mean and variance of N_ε, the number of jumps (minus 1) in the P_ε definition:

$$\mathbb{E}(N_\varepsilon) = \mathbb{E}(\mathbb{E}(N_\varepsilon|\varepsilon, \sigma, \kappa)) = \mathbb{E}(\Lambda_\varepsilon) = \mathbb{E}\left(\kappa \frac{\Gamma(-\sigma, \varepsilon)}{\Gamma(1-\sigma)} \right),$$

$$\begin{aligned} \mathrm{Var}(N_\varepsilon) &= \mathrm{Var}(\mathbb{E}(N_\varepsilon|\varepsilon, \sigma, \kappa)) + \mathbb{E}(\mathrm{Var}(N_\varepsilon|\varepsilon, \sigma, \kappa)) \\ &= \mathrm{Var}(\Lambda_\varepsilon) + \mathbb{E}(\Lambda_\varepsilon) \\ &= \mathrm{Var}\left(\kappa \frac{\Gamma(-\sigma, \varepsilon)}{\Gamma(1-\sigma)} \right) + \mathbb{E}\left(\kappa \frac{\Gamma(-\sigma, \varepsilon)}{\Gamma(1-\sigma)} \right). \end{aligned}$$

As far as Λ_ε (i.e. the conditional mean of N_ε) is concerned, the effect of κ is linear, while the influence of (ε, σ) is given by $\Gamma(-\sigma, \varepsilon)/\Gamma(1-\sigma)$, and it is antithetic. For any fixed $\sigma \in [0, 1)$, this function converges to $+\infty$ when $\varepsilon \to 0$ and to 0 when $\varepsilon \to +\infty$, and it decreases with ε. On the other hand, for any $\varepsilon > 0$, a qualitative study of the function $\sigma \mapsto \Gamma(-\sigma, \varepsilon)/\Gamma(1-\sigma)$ shows that it has a maximum located at a point that is closer to 1 as ε gets smaller. For this reason, when we want to approximate a NGG process with a large σ, we need to assume a very small ε; of course, this behavior is not surprising, since, when σ is large, NGG

Repeat for g in 1…**G**:

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

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

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

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

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 $\dfrac{\Lambda_{\varepsilon u}}{\Lambda_{\varepsilon u}+k}\mathscr{P}_1(\Lambda_{\varepsilon u}) + \dfrac{k}{\Lambda_{\varepsilon u}+k}\mathscr{P}_0(\Lambda_{\varepsilon u})$, then set $\boldsymbol{N}_\varepsilon^{(g)} + 1 = N_{na}^{(g)} + k$.

3.b.*ii* **Non-allocated jumps**: for $j = 1,\ldots,N_{na}$ sample independently from
$$\mathscr{L}(J_j) \propto e^{-uJ_j}\rho_\varepsilon(J_j).$$

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

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

3.b.*v* **Allocated points of support**: for $i = 1,\ldots,k$ sample independently from
$$\mathscr{L}(\tau_i^*) \propto \{\textstyle\prod_{j\in C_i} k(X_j;\tau_i)\}P_0(\tau_i).$$

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

mixtures and the parametric hierarchical mixture

$$X_i|\theta_i \overset{\text{ind}}{\sim} k(\cdot,\theta_i),\ \theta_i \overset{\text{iid}}{\sim} P_0$$

are close under any distance metrizing weak convergence.

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 (7) when $k(\cdot;\theta)$ is the Gaussian density on $\mathbb{R}$ and $\theta = (\mu,\sigma^2)$ stands for its mean and variance. Moreover, $P_0(d\mu, d\sigma^2) = \mathscr{N}(d\mu; m_0, \sigma^2/\kappa_0) \times inv-gamma(d\sigma^2; a, b)$; here $\mathscr{N}(m_0, \sigma^2/\kappa_0)$ 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 (2014a). Here we shed light on five 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, (D) when both σ and κ are random and ε is fixed, and, after that, (E) when $\mathbb{E}(N_\varepsilon)$ is set equal to 50.

We have implemented our Gibbs sampler in C++; the code is available on request from the second author. 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 of the sum defining the process P_ε (see (4)) and, consequently, less components in

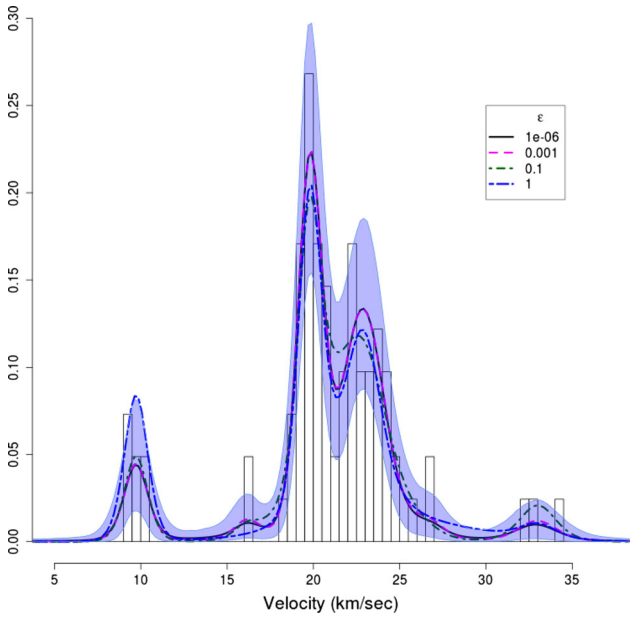

Fig. 2 Density estimates for different values of ε, while $\sigma = 0.4$ and $\kappa = 0.45$, case (A). The shaded region denotes 90% CI around the density estimates for $\varepsilon = 10^{-6}$

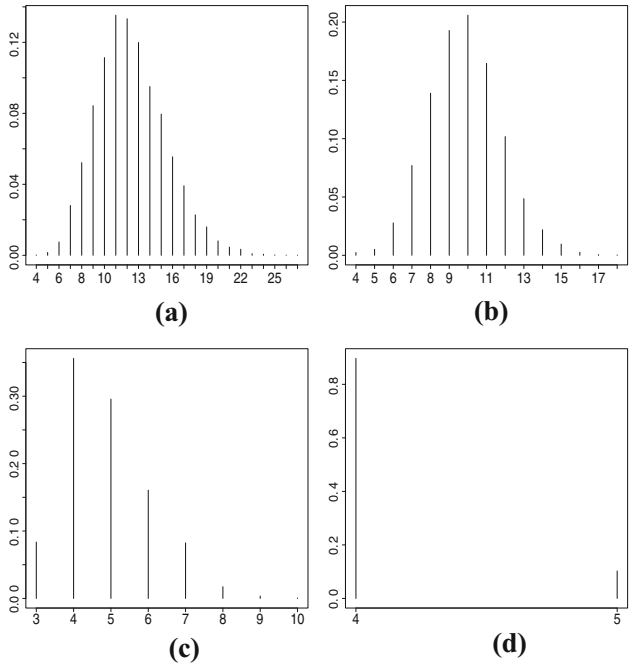

Fig. 3 Posterior distributions of the number K_n of groups under the $\varepsilon-$NGG mixture with hyperparameter set (A). **a** $\varepsilon = 10^{-6}$, **b** $\varepsilon = 10^{-3}$, **c** $\varepsilon = 10^{-1}$, **d** $\varepsilon = 1$

the mixture (8) are considered. In particular, the prior mean of N_ε is equal to 188.63, 10.86, 0.90, $5.6 \cdot 10^{-2}$ under the four different values of ε; the last two values are very small, but $\varepsilon = 10^{-1}$, or 1, yields a model which is pretty far from the NGG-mixture. Therefore the posterior estimate of the number K_n of groups, i.e. the number of unique values among $(\theta_1, \ldots, \theta_n)$ in (7), will be concentrated on smaller integer values as ε increases (see Fig. 3). It is worth underlining that, as another consequence of the smaller number of components in the mixture (8) 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 $\{0.001, 0.1, \ldots, 0.8\}$. 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.

For set (C) of hyperparameters, we have considered $\sigma \in \{0.001, 0.1, 0.2, \ldots, 0.8\}$, $\kappa = 0.45$ and ε random, uniformly distributed on the interval $(0, \delta)$, with $\delta = \min(0.1, \mathbb{E}(T) = \kappa) = 0.1$ (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 (4). 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 of σ for three different prior on ε. The increasing trend in $\mathbb{E}(K_n|data)$ is milder when ε is beta (red stars) or uniform distributed (blue diamonds) than when ε is equal to 10^{-6} (green dots). It is worth remarking that the prior distribution of N_ε makes an impact on the run-time, of course; for instance, if the prior mean of N_ε is large, the run-time will be large as well. For this reason we have com-

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

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

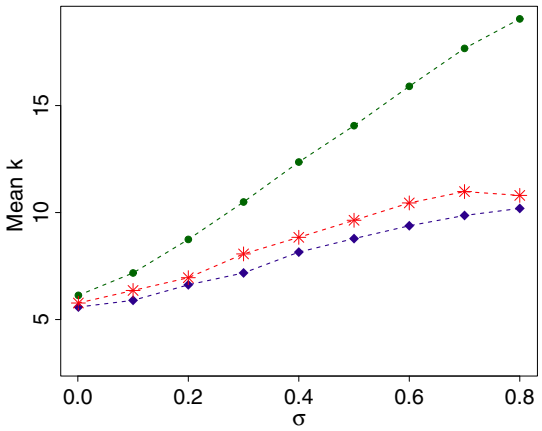

Fig. 4 Posterior mean of K_n as a function of σ, under different priors for ε in experiment (C): degenerate on 10^{-6} (green dots), uniform (blue diamonds) and scaled beta (red stars)

Table 2 Prior mean and variance of N_ε for cases (B) and (C)

σ	(B)	(C) Uniform		(C) beta	
	$\mathbb{E}(N_\varepsilon) = \mathrm{Var}(N_\varepsilon)$	$\mathbb{E}(N_\varepsilon)$	$\mathrm{Var}(N_\varepsilon)$	$\mathbb{E}(N_\varepsilon)$	$\mathrm{Var}(N_\varepsilon)$
0.001	5.99	1.24	1.44	1.72	2.2
0.1	12.26	1.42	1.84	2.12	3.54
0.2	28.38	1.61	2.55	2.68	7.64
0.3	71.41	1.85	4.22	3.51	27.98
0.4	188.63	2.12	9.09	4.82	$+\infty$
0.5	506.9	2.42	$+\infty$	7.12	$+\infty$
0.6	1345.3	2.75	$+\infty$	11.85	$+\infty$
0.7	3405.1	3.1	$+\infty$	$+\infty$	$+\infty$
0.8	7730.4	3.37	$+\infty$	$+\infty$	$+\infty$

puted prior means, variances and medians of N_ε (see Table 2) under (B) and (C); the medians are not reported here, but, for example, when the prior mean is infinite ($\sigma = 0.7, 0.8$), both medians are equal to 3. Run-times of experiments (C) are smaller than (B), and, in fact, prior means (or medians) are 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 Sect. 3, that is harder to be numerically evaluated.

Looking at the posterior distribution of ε in Fig. 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.

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 $\pi(\sigma) \times \pi(\kappa) = beta(a_1, b_1) \times gamma(c_1, d_1)$, 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 Fig. 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.

Finally, to better understand the effect of $\mathbb{E}(N_\varepsilon)$ on our algorithm, we fixed four different sets of hyperparameters for case (E), with $\mathbb{E}(N_\varepsilon) = 50$ for $\kappa = 5$: (a) $\varepsilon = 0.025$ and $\sigma = 0.025$, (b) ε uniform on $(0, 0.1)$ and $\sigma = 0.025$, (c) $\varepsilon = 10^{-4}$ and $\sigma = 0.0035$, (d) $\varepsilon = 10^{-4}$ and $\sigma \sim beta(0.11, 5.42)$. This experiment aims at showing the influence of antithetic parameters ε and σ on the posterior of N_ε, the prior mean of N_ε being fixed. Here we report only prior and posterior histograms of N_ε. From Fig. 7, it is clear that, in these cases, prior and posterior of N_ε do not substantially differ since the prior mean of N_ε is sufficiently large to explain the estimated number of groups under the mixture. On the other hand, if ε or σ are random (panels (b) and (d)), N_ε tends to favor smaller values a posteriori. This confirms the larger flexibility of the model when the parameters are random. We underline once more that even if both σ and ε influence N_ε, they are quite different in their meaning and in their statistical usage: σ is a parameter of the "exact" NGG-mixture itself (refer to the literature for its interpretation), while ε is the degree of approximation of our P_ε, that we assume random only to gain more flexibility and lower computational costs.

Finally, we mention that a gamma prior can be given to ε. In particular, Bianchini (2015) analyzes the same dataset with different gamma priors for ε, all of them concentrating most mass around 0.

5.1 Comparison with other methods

As we mentioned in the Introduction, there are many other computational methods that can be used to fit NGG-mixture models. Here we compare our algorithm to an exact method, i.e. the two slice sampler algorithms in Griffin and Walker (2011), and the simple adaptive truncation method in Argiento et al. (2010). The choice of the former was recommended by the referees, and we agree that a comparison

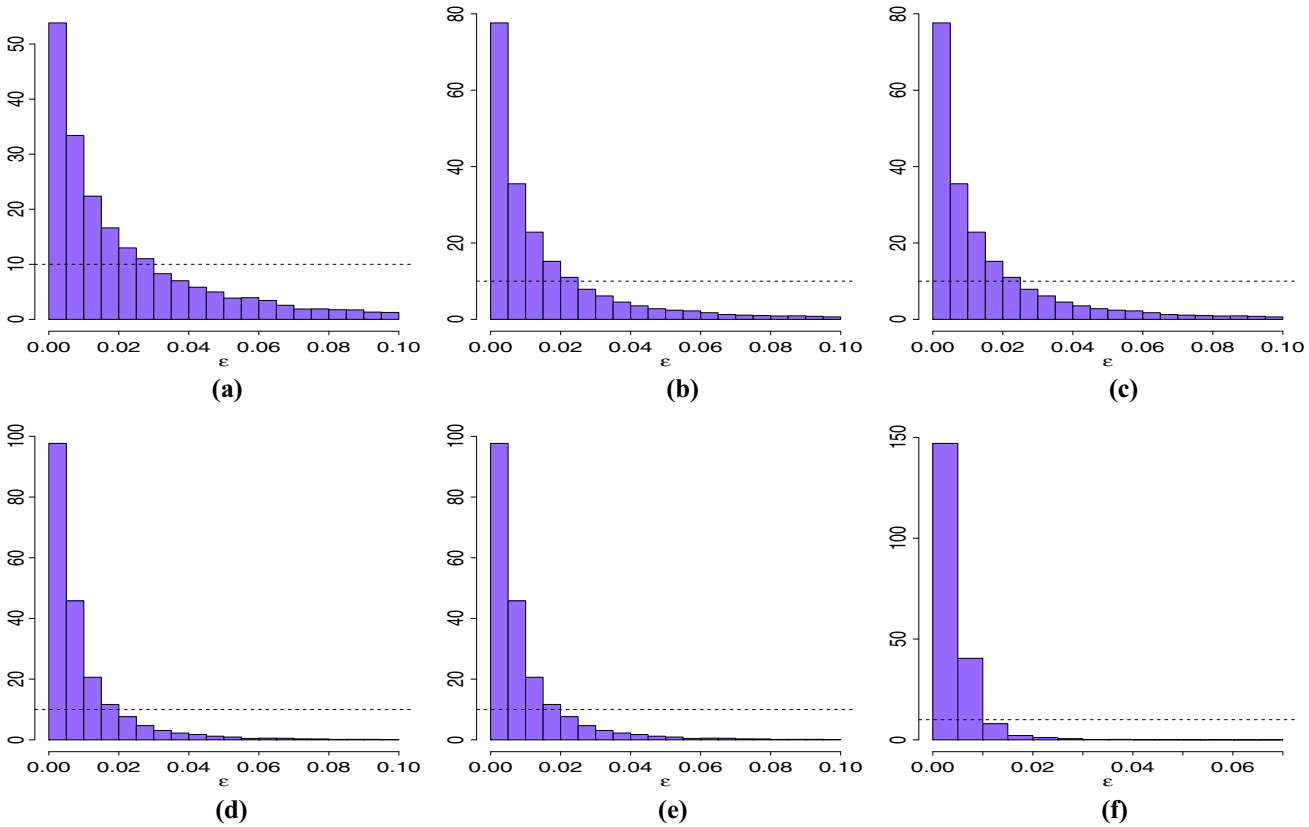

Fig. 5 Posterior distribution of ε for experiment (C), together with $\mathscr{U}(0,\delta)$ prior (dashed). **a** $\sigma = 0.001$, **b** $\sigma = 0.1$, **c** $\sigma = 0.2$, **d** $\sigma = 0.5$, **e** $\sigma = 0.7$, **f** $\sigma = 0.9$

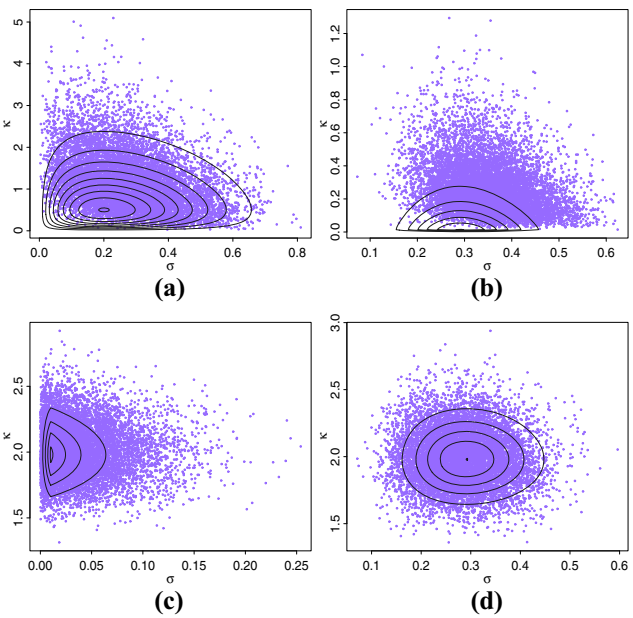

Fig. 6 Scatterplots of posterior values of (σ,κ) with contour levels of the prior for case (D)

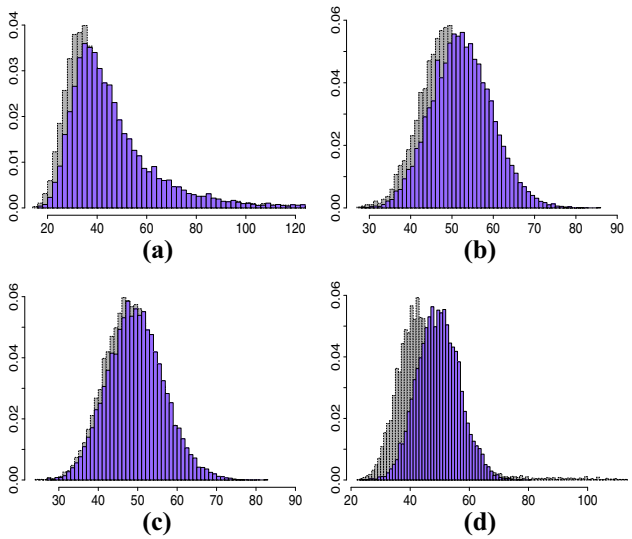

Fig. 7 Histograms of the prior and posterior (*light grey*) values of N_ε in the test cases (E): **a** $\varepsilon = 0.025, \sigma = 0.025$; **b** ε uniform on $(0, 0.1)$, $\sigma = 0.025$; **c** $\varepsilon = 10^{-4}, \sigma = 0.0035$; **d** $\varepsilon = 10^{-4}, \sigma \sim beta(0.11, 5.42)$

between ours and an exact method (as the slice sampler) is needed. On the other hand, the latter method is the main competitor, since it is a conditional algorithm using a truncation strategy from the posterior point of view, while here the truncation is a priori. We used the Matlab code for the two slice samplers in Griffin and Walker (2011) available online

Table 3 Estimates of integrated autocorrelation times for the deviance, $\hat{\tau}_D$, and number of groups, $\hat{\tau}_C$, and run-times, using different methods

	σ	σ and κ fixed						σ or κ random		
		0.001	0.1	0.2	0.3	0.4	0.5	(i)	(ii)	(iii)
Slice 1	$\hat{\tau}_D$	118.9	70.4	31.7	12.14	9.5	8.4	161.7	43.3	280.0
	$\hat{\tau}_C$	188.4	158.9	146.4	129.0	124.7	80.4	263.9	142.0	363.7
Slice 2	$\hat{\tau}_D$	84.9	17.2	5.0	1.3	1.0	1.0	33.4	101.1	114.6
	$\hat{\tau}_C$	201.6	112.8	219.2	198.5	90.8	52.1	145.6	47.8	161.2
A posteriori truncation Argiento et al. (2010)	$\hat{\tau}_D$	22.1	15.8	5.2	3.0	2.6	2.2	–	–	–
	$\hat{\tau}_C$	27.6	18.3	12.7	11.3	9.2	9.2	–	–	–
	Run-time	42 min 23 s	1 h 54 min	2 h 35 min	5 h 6 min	14 h 30 min	∼48 h			
$\varepsilon = 10^{-6}$	$\hat{\tau}_D$	25.3	8.5	2.1	2.0	2.5	1.6	8.5	3.0	22.3
	$\hat{\tau}_C$	30.1	25.2	28.0	35.0	30.2	18.0	196.3	12.6	47.9
	Run-time	1 min 28 s	56 s	1 min 30 s	3 min	8 min	17 min			
$\varepsilon \sim Unif(0, \delta)$	$\hat{\tau}_D$	59.6	48.4	13.7	12.0	2.2	7.0	15.3	6.3	19.7
	$\hat{\tau}_C$	103.0	71.6	56.6	63.4	28.8	59.6	51.7	37.0	75.6
	Run-time	48s	31s	35s	38s	43s	48s			

at the journal Supplemental page. For the second method, we used the code described in Argiento et al. (2010), available on request from the authors.

In order to compare the algorithms, we computed the integrated autocorrelation time τ for two variables: the number of clusters and the deviance D of the estimated density, defined as

$$D = -2\sum_{i=1}^{n} log\left(\sum_{j=1}^{k} \frac{n_j}{n} k(x_i; \theta_j^*)\right).$$

The integrated autocorrelation time controls the accuracy of Monte Carlo estimates computed using the MCMC chain and provides a measure of the efficiency of the method. The same indexes have been also used in Papaspiliopoulos and Roberts (2008) and Griffin and Walker (2011) to assess the performance of their methods. An estimator for τ is

$$\hat{\tau} = \frac{1}{2} + \sum_{l=1}^{C-1} \hat{\rho}_l,$$

where C is a threshold value, chosen as the minimum lag l such that $|\hat{\rho}_l| < 2/\sqrt{G}$ (see Kalli et al. 2011, for further details) and $\hat{\rho}_l$ is the estimated autocorrelation at lag l. Obviously, a small value of τ implies good mixing and hence an efficient method.

For the aim of comparison, we ran 100,000 iterations with an extra burn-in of 10,000 iterations and computed the value of integrated autocorrelation time of the two variables mentioned above for the three algorithms and for each set of hyperparameters. For the algorithm in Argiento et al. (2010), we fixed the truncation level M in the sum defining the NGG process so that

$$\mathbb{P}\left(\sum_{j=M+1}^{+\infty} \tilde{J}_j \leq 0.1\ \mathbb{E}(T)\right) = 0.99,$$

where $\{\tilde{J}_j\}$ is the decreasing sequence of the jumps. We selected hyperparameters as in experiment (B), while σ ranges in $\{0.001, 0.1, \ldots, 0.5\}$. Moreover, we consider the cases where ε is uniform on $(0, \delta)$, like in (C), and (i) $\sigma \sim beta(1, 19), \kappa = 0.45$, (ii) $\sigma = 0.1, \kappa \sim gamma(2, 2)$, (iii) $\sigma \sim beta(1.1, 30)$, $\kappa \sim gamma(1.1, 8)$.

It is clear from Table 3 that, when κ and σ are fixed, both truncation methods are generally more efficient than the slice samplers and provide very similar values of the integrated autocorrelation times; in addition, note that our algorithm achieves very good performances when ε is fixed. Specifically, the values of $\hat{\tau}_D$ are the smallest under our a priori truncation algorithm, while the best $\hat{\tau}_C$ s are obtained under the truncation method in Argiento et al. (2010), at the cost of much longer run-times (see Table 3). Observe that run-times of the two truncation methods might be compared, since the algorithm in Argiento et al. (2010) was coded in C language and the computations were carried out on the same machine; it is apparent that the algorithm proposed in this work is much faster. In cases (i), (ii) and (iii), where σ and κ are assumed random, algorithm in Argiento et al. (2010) cannot be applied, while Matlab codes for the slice samplers allow these cases too. For all these experiments (but $\hat{\tau}_C$ under Slice 2-(i)), the values of the integrated autocorrelation times show that our algorithm outperforms the two slice samplers. The better performance of the truncation algorithms is not surprising; as noted by Griffin and Walker (2011), "the slice sampler introduces auxiliary variables to help simulation which will slow convergence through over-conditioning".

This comment also applies when ε is random: we gain in run-time during the simulation, but the efficiency decreases, due to the larger autocorrelation of the variables. Summing up, in this particular example, our method with $\varepsilon = 10^{-6}$ outperforms the competitors.

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 $\left[X_{ij},\ 1 \leq i \leq n,\ 1 \leq j \leq p\right]$, where the rows $(X_1, \ldots, X_n)$ contain the expression patterns of genes and are our data points. Each entry 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. (2014). We assume the Gaussian kernel $k(\cdot; \theta_i) = \mathscr{N}_p(\cdot; \theta_i)$ 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 $P_0(d\mu, d\Sigma) = \mathscr{N}_p(d\mu | m_0, \Sigma/s_0) \times \prod_{k=1}^{p} inv - gamma(d\sigma^2_k | a, b)$.

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 $\text{Var}(\mu) = \mathbb{I}_p$ 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 (F)), then ε uniformly distributed over the interval $(0, 0.01)$, $\sigma \in \{0.01, 0.1, 0.2, \ldots, 0.5\}$ and $\kappa = 0.7$ (case (G)). Finally, we set $\varepsilon = 10^{-4}$, $\sigma \sim beta(2, 15)$ and $\kappa \sim gamma(2, 0.1)$ (case (H)).

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 (F) 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 is that, if ε is fairly large, the jumps J_j's are approximately independently sampled from a degenerate distribution on ε, and therefore, they assume the same value; consequently, the full-conditional of

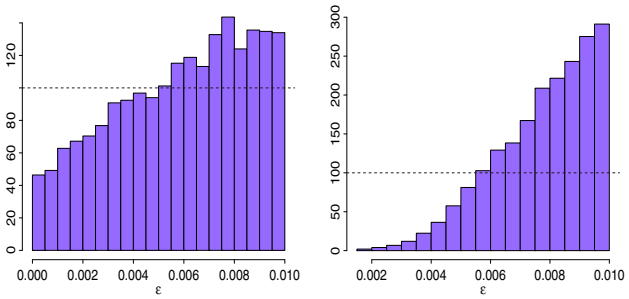

Fig. 8 Posterior distribution of ε for experiment (G): $\sigma = 0.001$ (*left*) and $\sigma = 0.5$ (*right*). The prior is $\mathscr{U}(0, 0.01)$ (*dashed*)

$\boldsymbol{\theta}$, as in Step 2. of the algorithm (see Fig. 1), depends only on the parametric kernel, evaluated at data points, yielding that $N_\varepsilon + 1$ and K_n coincide. In this case, the prior means (and variances) of N_ε are 9.32, 7.69, 6.06, 4.44, respectively, for the four different values of ε.

For experiment (G), the choice of hyperparameters implies that the prior mean of N_ε increases from 3.61 to 14.44 as σ increases, while the prior variance grows from 4.16 to 649.64. Figure 8 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, density estimates seem to fit the data well. Figure 9 shows all the unidimensional marginal predictive densities for case (G). We have not observed substantial differences among the predictive density estimates in all the experiments we run.

For experiment (H), we set a vague prior for κ, and a more informative prior on σ to speed up and improve the mixing; in this case, both prior mean and variance of N_ε are very large. The posterior of (σ, κ) is displayed in Fig. 10, 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 (7), we adopt a standard approach in the Bayesian framework. First of all, we recall that (7) induces a prior for the random partition $\boldsymbol{p}_n = \{C_1, \ldots, C_k\}$ of the data labels (see notation in Sect. 3), and, as a consequence, cluster estimates are based on its posterior. As such an estimate we consider the partition $\widehat{\boldsymbol{p}}_n$ minimizing the so-called Binder loss function with equal misclassification costs, using the same approach as in Argiento et al. (2014). To compare different cluster estimates, we evaluate the posterior expectation $\mathbb{E}(H(\boldsymbol{p}_n)|data)$ when 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 (2014a). In Fig. 11 we report one of the best cluster estimate, which was obtained when hyperparameters are those of case (H).

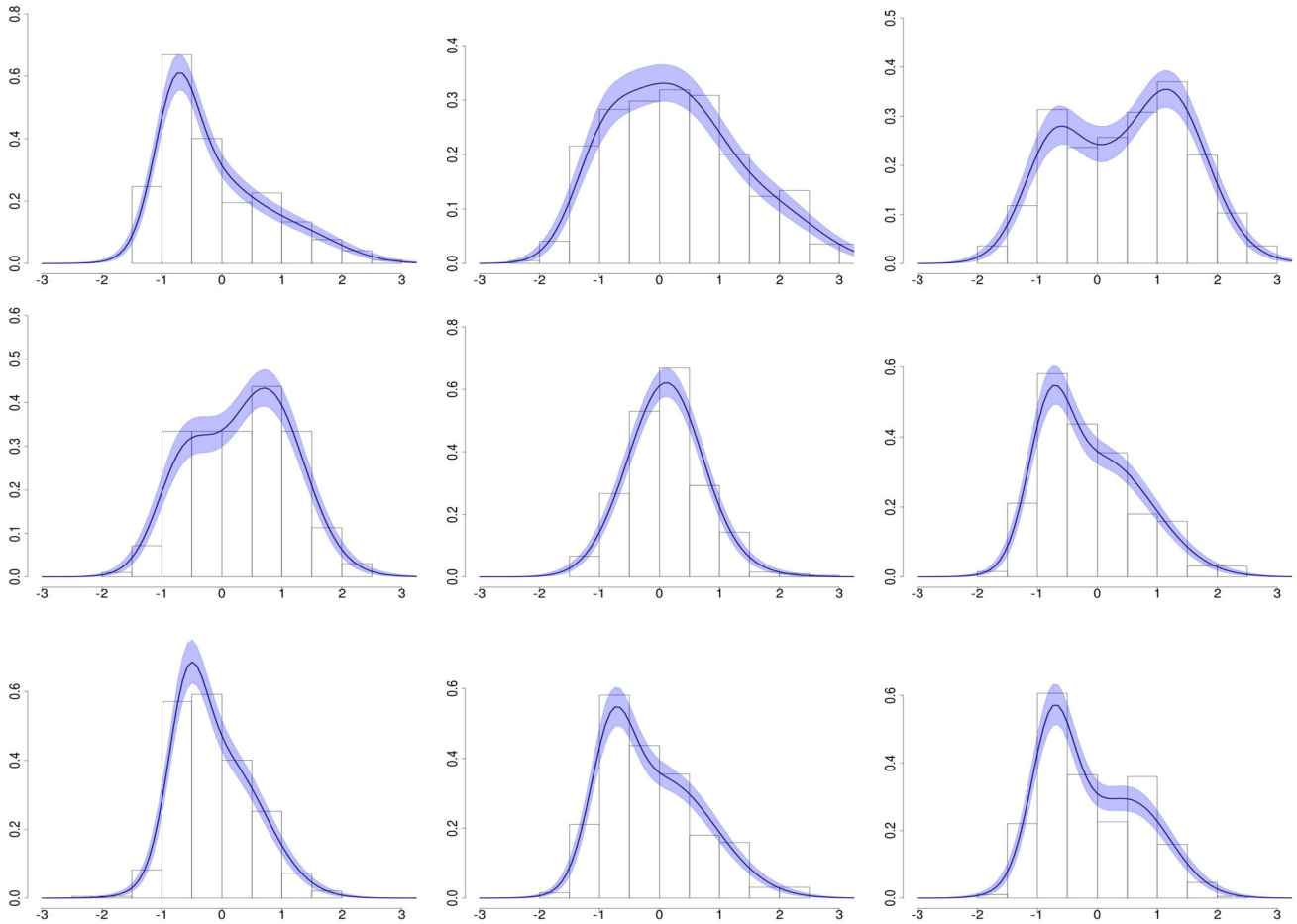

Fig. 9 Marginal density estimates for experiment (G) when $\sigma = 0.001$, $\kappa = 0.7$, $\varepsilon \sim \mathscr{U}(0, 0.01)$. The *shaded regions* denote 90% CI's around the density estimates

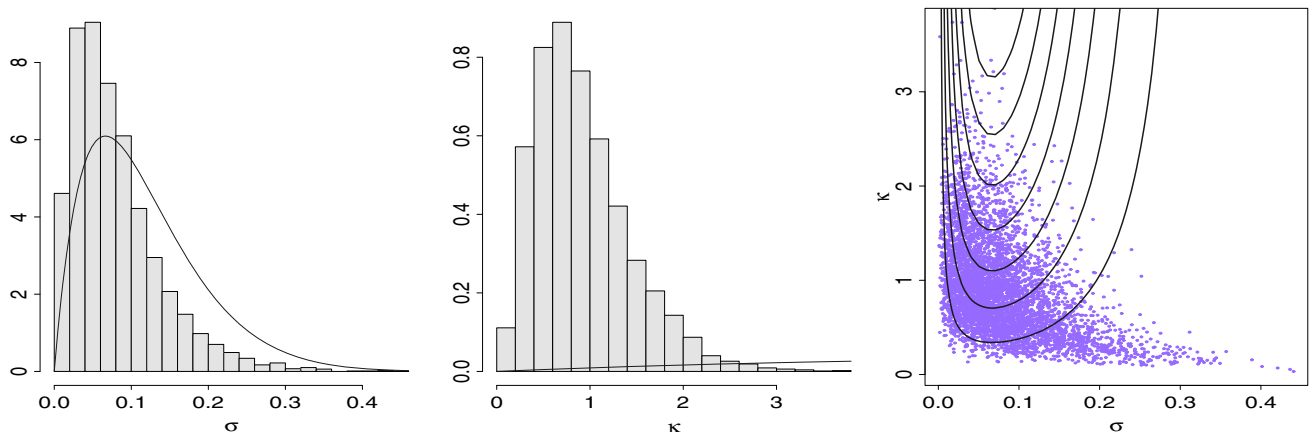

Fig. 10 Posteriors of σ (*left*), κ (*center*), and (σ, κ) (*right*). The priors are superimposed as *gray lines*

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 adequate 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 the 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.

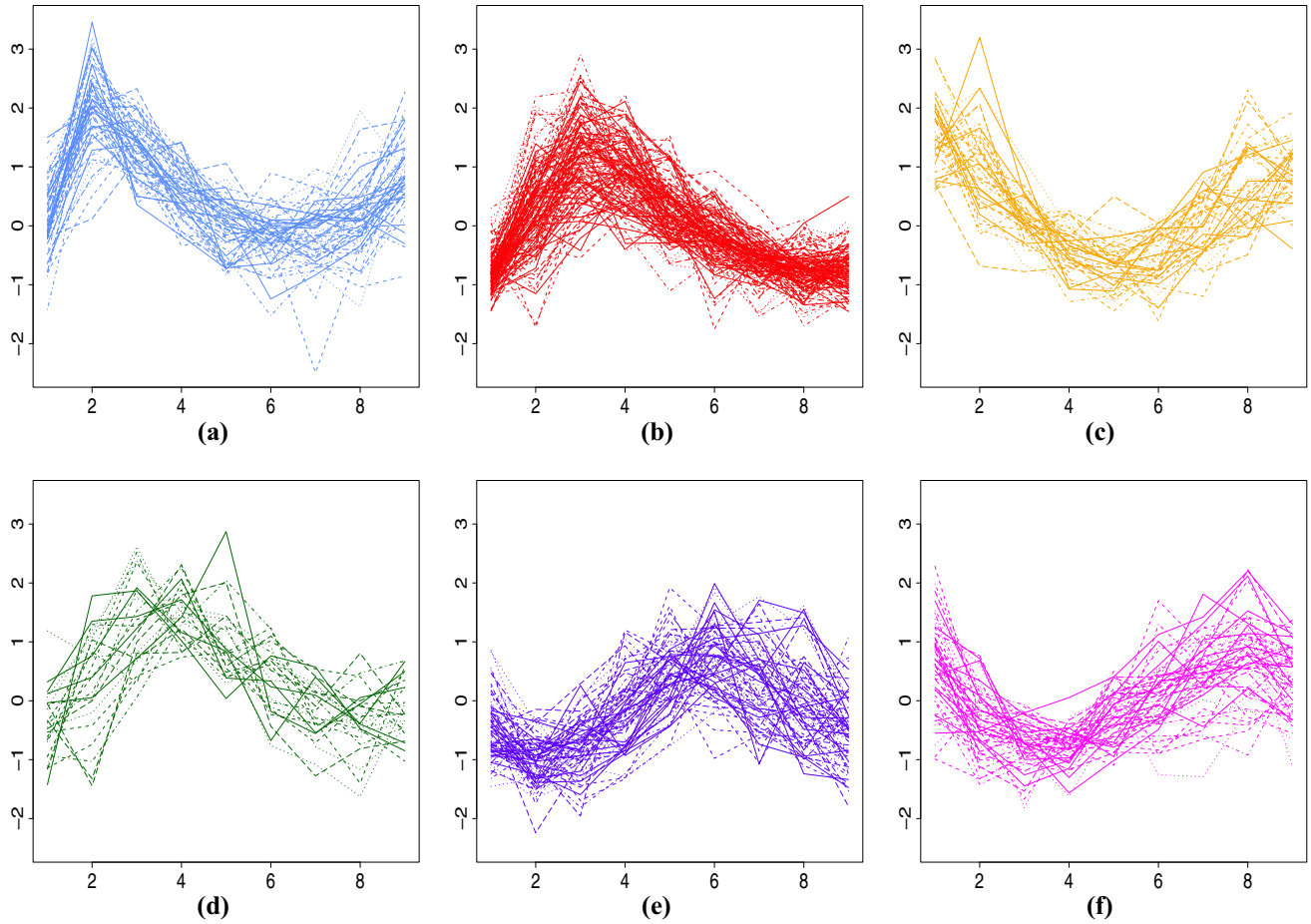

Fig. 11 Bayesian cluster estimates for experiment (H)

As a last remark in this section, we would like to point out that all the cluster estimates, here and in Bianchini (2014a), 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.

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 assumed 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 those in the 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 made clear; moreover, the robustness of the model with respect to the choice of the hyperparameters has been checked. As a conclusion, even if both parameters

σ and ε influence the number of groups in this new mixture model, they are quite different in the meaning: σ is a parameter of the "exact" NGG-mixture itself, while ε is the degree of approximation of P_ε, that we assume random only to gain more flexibility and lower computational costs. 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 mixing distribution (4) grows very fast with σ, at least as far as σ reaches σ_0, the argmax of $\Gamma(-\sigma, \varepsilon)/\Gamma(1-\sigma)$ (σ_0 is close to 1), slowing down the run-time of the algorithm. On the other hand, when σ is larger than σ_0, our P_ε may fail to approximate the "exact" NGG process, unless ε becomes smaller. However, if σ is close to 1, the NGG-mixture model itself is very close to a parametric mixture model.

In this paper, we have detailed the ε-approximation of a class of homogeneous normalized random measures with independent increments, the NGG processes. This construction is based on the specific features of ρ, the mean intensity of the process defining the jumps of the "exact" random probability measure (i.e., ρ has an asymptote in 0 and has finite mass between 1 and $+\infty$). Consequently, in general, the ε-version of any HNRMI can be defined, considering $\Lambda_\varepsilon = \int_\varepsilon^{+\infty} \rho(x)dx$ and $\rho_\varepsilon(s) = (1/\Lambda_\varepsilon)\rho(s)\mathbb{I}_{(\varepsilon,+\infty)}(s)$. However, the proof of Propositions 1 and 2 in this more general setting and follow-up analysis are the subject of future research.

Acknowledgments The authors would like to thank the referees and the Associate Editor for their valuable help in improving this manuscript. Ilaria Bianchini was partially funded by CNR-IMATI *Flagship project - Factory of the future - Fab at Hospital*.

8 Appendices

8.1 Appendix 1: Proof of (5)

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

$$p_\varepsilon(n_1, \ldots, n_k) = \sum_{N_\varepsilon=0}^{+\infty} p_\varepsilon(n_1, \ldots, n_k|N_\varepsilon)\frac{\Lambda_\varepsilon^{N_\varepsilon}}{N_\varepsilon!}\mathrm{e}^{-\Lambda_\varepsilon}. \quad (12)$$

Then, formula (30) in Pitman (1996) 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 $(j_1, \ldots, j_k)$ ranges over all permutations of k elements in $\{0, \ldots, N_\varepsilon\}$. Using the gamma function identity,

$$\frac{1}{T_\varepsilon^n} = \int_0^{+\infty} \frac{1}{\Gamma(n)} u^{n-1} e^{-uT_\varepsilon} du, \quad (13)$$

we have:

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

Substituting the expression of ρ_ε in Sect. 3:

$$\begin{aligned}
&p_\varepsilon(n_1, \ldots, n_k|N_\varepsilon) \\
&= \sum_{j_1,\ldots,j_k} \int_0^{+\infty} du \Big(\frac{1}{\Gamma(n)} u^{n-1} \\
&\quad \times \prod_{i=1}^k \int_\varepsilon^{+\infty} \frac{J_{j_i}}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} J_{j_i}^{-\sigma-1} \mathrm{e}^{-(\omega+u)J_j} dJ_{j_i} \\
&\quad \times \prod_{j\notin\{j_1,\ldots,j_k\}} \int_0^{+\infty} \frac{1}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} J_j^{-\sigma-1} \mathrm{e}^{-(\omega+u)J_j} dJ_j \Big) \\
&= \sum_{j_1,\ldots,j_k} \int_0^{+\infty} \Big(\frac{1}{\Gamma(n)} u^{n-1} \\
&\quad \times \prod_{i=1}^k \frac{(u+\omega)^{\sigma-n_i} \Gamma(n_i-\sigma; (u+\omega)\varepsilon)}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} \\
&\quad \times \left(\frac{(u+\omega)^\sigma \Gamma(-\sigma; (u+\omega)\varepsilon)}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} \right)^{N_\varepsilon+1-k} \Big) du.
\end{aligned}$$

We have crossed out the indicator function in all previous lines. 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, $i = 1, \ldots, k$, but only on the sizes n_i, and there are $(N_\varepsilon+1)(N_\varepsilon)\ldots(N_\varepsilon+1-k) = (N_\varepsilon+1)!/(N_\varepsilon+1-k)!$ sequences of k distinct elements from $\{0, \ldots, N_\varepsilon\}$, we get:

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

Observe that, because of the indicator function in the above formula, summation in (12) 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

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

By Fubini's theorem, we can switch integration and summation, and introduce $\Lambda_{\varepsilon,u}$ as defined in (6), so that

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

that is (5), since

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

8.2 Appendix 2: Lemma 1

We report this simple lemma for a thorough understanding of Lemma 2.

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 hypotheses 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.

8.3 Appendix 3: Lemma 2

Lemma 2 *Let p_ε 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 $k\ge0$ and $\sum_{i=1}^k n_i=n$, we have*

$$\lim_{\varepsilon\to0}p_\varepsilon(n_1,\ldots,n_k)=p_0(n_1,\ldots,n_k),\tag{14}$$

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

Proof Formula (5) can be restated as

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

where f_ε denotes the integrand in the right hand side of (5) itself. In addition, 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

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

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 The-orem, 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:

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

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

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

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 $\mathscr{C} \in \Pi_n$, we have $p_0(n_1, \ldots, n_k) < \liminf_{\varepsilon\to 0} p_\varepsilon(n_1, \ldots, n_k)$. In this case

$$\begin{aligned} 1 &= \sum_{\mathscr{C}\in\Pi_n} p_0(n_1, \ldots, n_k) < \sum_{\mathscr{C}\in\Pi_n} \liminf_{\varepsilon\to 0} p_\varepsilon(n_1, \ldots, n_k) \\ &\leq \liminf_{\varepsilon\to 0} \sum_{\mathscr{C}\in\Pi_n} p_\varepsilon(n_1, \ldots, n_k) = 1, \end{aligned}$$

that is a contradiction. Therefore we conclude that

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

for all $n_1, \ldots, n_k$, all k. Summing up, we have proved so far that:

- $\lim_{\varepsilon\to 0} \sum_{\mathscr{C}\in\Pi_n} p_\varepsilon(n_1, \ldots, n_k) = 1$;
- $\liminf_{\varepsilon\to 0} p_\varepsilon(n_1, \ldots, n_k) = p_0(n_1, \ldots, n_k)$ for all $\mathscr{C} = (C_1, \ldots, C_k) \in \Pi_n$;
- $\sum_{\mathscr{C}\in\Pi_n} p_0(n_1, \ldots, n_k) = 1$.

By Lemma 1, equation (14) follows.

8.4 Appendix 4: Proof of Proposition 1

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 Lemma 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 $(N_1^0, \ldots, N_k^0)$ denotes the sizes of the blocks (in order of appearance), of the random partition $C_{0,n}$ 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:

$$\begin{aligned} &\lim_{\varepsilon\to 0} \mathbb{E}\left(\mathrm{e}^{-\int_\Theta f d\mu^\varepsilon}\right) \\ &= \lim_{\varepsilon\to 0} \lim_{n\to+\infty} \mathbb{E}\left(\mathrm{e}^{-\int_\Theta f d\mu_n^\varepsilon}\right) \\ &= \lim_{n\to+\infty} \lim_{\varepsilon\to 0} \mathbb{E}\left(\mathrm{e}^{-\int_\Theta f d\mu_n^\varepsilon}\right) \\ &= \lim_{n\to+\infty} \mathbb{E}\left(\mathrm{e}^{-\int f d\mu_n^0}\right) = \mathbb{E}\left(\mathrm{e}^{-\int f d\mu^0}\right). \end{aligned} \tag{15}$$

Here we have introduced notation

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

thus (15) 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 the 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

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

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 distinguish two cases:

1. $C_{\varepsilon,n+1}$ has the same number of clusters of $C_{\varepsilon,n}$; one of these clusters (the j-th for instance) has one more element and, as a consequence, has size equal to $n_j + 1$;
2. $C_{\varepsilon,n+1}$ has one more cluster than $C_{\varepsilon,n}$; this cluster contains only one element.

In both cases, it is not difficult to prove that

$$\mathbb{E}\left(\left|\int_\Theta f d\mu_{n+1}^\varepsilon - \int f d\mu_n^\varepsilon\right|\right) \leq \frac{2M}{n+1},$$

so that we are able to interchange the two limits. Finally, it is straightforward to show that the stated convergence follow from the convergence in distribution conditioning on $\{\tau_0, \tau_1, \tau_2, \ldots\}$, with an argument on Laplace transforms as before. This ends the first part of the Proposition, i.e. convergence for $\varepsilon \to 0$.

Convergence as $\varepsilon \to +\infty$ is straightforward as well. In fact, when ε increases to $+\infty$, there are no jumps to consider in (4) but the extra J_0, so that P_ε degenerates on δ_{τ_0}.

8.5 Appendix 5: Proof of Proposition 2

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

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

where $(\theta_1^*, \theta_2^*, \ldots, \theta_k^*)$ is the vector of the unique values in the sample. We omit the indicator $\mathbb{I}_{\{1,\ldots,N_\varepsilon+1\}}(k)$ till we need it. Introducing the auxiliary variable U, by (13) we have:

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

Hence we have:

$$\begin{aligned}
\mathscr{L}(\boldsymbol{\theta}, u, P_\varepsilon) &= \mathscr{L}(\boldsymbol{\theta}, u | P_\varepsilon)\mathscr{L}(P_\varepsilon) \\
&= \frac{u^{n-1}}{\Gamma(n)} e^{-uT_\varepsilon} \sum_{l_1^*,\ldots,l_k^*} \left(J_{l_1^*}^{n_1} \delta_{\tau_{l_1^*}^*}(\theta_1^*) \ldots J_{l_k^*}^{n_k} \delta_{\tau_{l_k^*}^*}(\theta_k^*)\right)\mathscr{L}(P_\varepsilon) \\
&= \frac{u^{n-1}}{\Gamma(n)} \prod_{j=0}^{N_\varepsilon} (e^{-uJ_j}) \sum_{l_1^*,\ldots,l_k^*} \left(J_{l_1^*}^{n_1} \delta_{\tau_{l_1^*}^*}(\theta_1^*) \ldots J_{l_k^*}^{n_k} \delta_{\tau_{l_k^*}^*}(\theta_k^*)\right) \\
&\quad \times \prod_{j=0}^{N_\varepsilon} \left(\rho_\varepsilon(J_j) P_0(\tau_j)\right) \mathscr{P}_0(N_\varepsilon; \Lambda_\varepsilon) \qquad (16) \\
&= \frac{1}{\Gamma(n)} u^{n-1} \prod_{j=0}^{N_\varepsilon} \left(e^{-uJ_j} \rho_\varepsilon(J_j) P_0(\tau_j)\right) \\
&\quad \times \sum_{l_1^*,\ldots,l_k^*} \left(J_{l_1^*}^{n_1} \delta_{\tau_{l_1^*}^*}(\theta_1^*) \ldots J_{l_k^*}^{n_k} \delta_{\tau_{l_k^*}^*}(\theta_k^*)\right) \mathscr{P}_0(N_\varepsilon; \Lambda_\varepsilon)
\end{aligned}$$

where, in this proof, $\mathscr{P}_0(N_\varepsilon; \Lambda_\varepsilon)$ 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:

$$\begin{aligned}
\mathscr{L}(P_\varepsilon | u, \boldsymbol{\theta}) &= \mathscr{L}(\boldsymbol{\tau}, \boldsymbol{J}, N_\varepsilon | u, \boldsymbol{\theta}) \\
&= \mathscr{L}(\boldsymbol{\tau}, \boldsymbol{J} | N_\varepsilon, u, \boldsymbol{\theta})\mathscr{L}(N_\varepsilon | u, \boldsymbol{\theta}).
\end{aligned} \qquad (17)$$

The second factor in the right handside is proportional to

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

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

$$
\begin{aligned}
&\int e^{-uJ_j} \rho_\varepsilon(J_j) P_0(\tau_j) dJ_j d\tau_j \\
&= \int_0^{+\infty} e^{-uJ_j} \rho_\varepsilon(J_j) dJ_j \\
&= \frac{1}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} \int_0^{+\infty} x^{-\sigma-1} \mathrm{e}^{(u+\omega)x} \mathbb{I}_{(\varepsilon,+\infty)}(x) dx \\
&= \frac{(\omega+u)^\sigma}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} \int_{(\omega+u)\varepsilon}^{+\infty} e^{-y} y^{-\sigma-1} dy \qquad (18) \\
&= \frac{(\omega+u)^\sigma \Gamma(-\sigma, (\omega+u)\varepsilon)}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)}.
\end{aligned}
$$

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 $i = 1, \dots, k$:

$$
\begin{aligned}
&\int J_{l_i^*}^{n_i} \delta_{\tau_{l_i^*}}(\theta_i^*) \mathrm{e}^{-uJ_{l_i^*}} \rho_\varepsilon(J_{l_i^*}) P_0(\tau_{l_i^*}) dJ_{l_i^*} d\tau_{l_i^*} \\
&= \left(\int J_{l_i^*} \mathrm{e}^{-uJ_{l_i^*}} \rho_\varepsilon(J_{l_i^*}) dJ_{l_i^*} \right) \left(\int \delta_{\tau_{l_i^*}}(\theta_i^*) P_0(\theta_i^*) d\theta_i^* \right) \\
&= \frac{P_0(\theta_i^*)}{\omega^\sigma \Gamma(-\sigma, \omega\varepsilon)} \int_0^{+\infty} x^{n_i} \mathrm{e}^{-ux} x^{-1-\sigma} \mathrm{e}^{-\omega x} \mathbb{I}_{(\varepsilon,+\infty)} dx \qquad (19) \\
&= \frac{(\omega+u)^{\sigma-n_i}}{\omega^\sigma} \frac{\Gamma(n_i - \sigma, (u+\omega)\varepsilon)}{\Gamma(-\sigma, \omega\varepsilon)} P_0(\theta_i^*).
\end{aligned}
$$

The integrand function in (19) is the kernel of a gamma density with parameters $(n_i - \sigma, u+\omega)$, restricted to $(\varepsilon, +\infty)$. Summing up, we have

$$
\begin{aligned}
&\mathscr{L}(N_\varepsilon | u, \boldsymbol{\theta}) \propto \mathscr{L}(N_\varepsilon, u, \boldsymbol{\theta}) = \frac{u^{n-1}}{\Gamma(n)} \mathscr{P}_0(N_\varepsilon; \Lambda_\varepsilon) \\
&\times \sum_{l_1^*,\dots,l_k^*} \left\{ \left(\frac{(\omega+u)^{k\sigma-n} \prod_{i=1}^k \Gamma(n_i - \sigma, (\omega+u)\varepsilon) P_0(\theta_i^*)}{\omega^{\sigma k} \Gamma(-\sigma, \omega\varepsilon)^k} \right) \right. \\
&\left. \times \left(\frac{(\omega+u)^{\sigma(N_\varepsilon+1-k)} \Gamma(-\sigma, (u+\omega)\varepsilon)^{N_\varepsilon+1-k}}{\omega^{\sigma(N_\varepsilon+1-k)} \Gamma(-\sigma, \omega\varepsilon)^{N_\varepsilon+1-k}} \right) \right\} \\
&= \frac{u^{n-1}}{\Gamma(n)} \mathscr{P}_0(N_\varepsilon; \Lambda_\varepsilon) \frac{(N_\varepsilon+1)!}{(N_\varepsilon+1-k)!} \frac{(\omega+u)^{\sigma k-n}}{\omega^{\sigma k} \Gamma(-\sigma, \omega\varepsilon)^k} \qquad (20) \\
&\times \frac{(\omega+u)^{\sigma N_{na}} \Gamma(-\sigma, \varepsilon(\omega+u))^{N_{na}}}{\omega^{\sigma N_{na}} \Gamma(-\sigma, \omega\varepsilon)^{N_{na}}} \\
&\times \prod_{i=1}^k \left(P_0(\theta_i^*) \Gamma(n_i - \sigma, \varepsilon(\omega+u)) \right) \mathbb{I}_{\{(N_\varepsilon+1) \geq k\}}.
\end{aligned}
$$

As in the proof of formula (5), $N_{na} = N_\varepsilon + 1 - k$ is the number of *non-allocated* jumps. Therefore, since k is given, the conditional distribution $\mathscr{L}(N_\varepsilon |\, u, \boldsymbol{\theta})$ is identified by $\mathscr{L}(N_{na} | u, \boldsymbol{\theta})$; we have

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

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

$$
\begin{aligned}
&\mathscr{L}(N_{na} | \varepsilon, u, \boldsymbol{\theta}) \propto \frac{N_{na}+k}{N_{na}!} \mathrm{e}^{-\Lambda_{\varepsilon,u}} \Lambda_{\varepsilon,u}^{N_{na}} \\
&= \frac{\Lambda_{\varepsilon,u}}{(N_{na}-1)!} \Lambda_{\varepsilon,u}^{(N_{na}-1)} \mathrm{e}^{-\Lambda_{\varepsilon,u}} + \frac{k}{N_{na}!} \Lambda_{\varepsilon,u}^{N_{na}} \mathrm{e}^{-\Lambda_{\varepsilon,u}} \qquad (21) \\
&= \frac{\Lambda_{\varepsilon,u}}{\Lambda_{\varepsilon,u}+k} \mathscr{P}_1(N_{na}; \Lambda_{\varepsilon,u}) + \frac{k}{\Lambda_{\varepsilon,u}+k} \mathscr{P}_0(N_{na}; \Lambda_{\varepsilon,u}).
\end{aligned}
$$

On the other hand, the first factor in the right handside of (17) can be computed by the following comment. Denote by $\boldsymbol{l}^* = (l_1^*, \dots, l_k^*)$ is the vector of locations of the *allocated* jumps. From (16), it is clear that

$$
\begin{aligned}
&\mathscr{L}(\boldsymbol{J}, \boldsymbol{\tau}, \boldsymbol{l}^* | N_{na}, u, \boldsymbol{\theta}) \\
&= J_{l_1^*}^{n_1} \delta_{\tau_{l_1^*}^*}(\theta_1^*) \dots J_{l_k^*}^{n_k} \delta_{\tau_{l_k^*}^*}(\theta_k^*) \\
&\quad \times \prod_{j=0}^{N_{na}+k-1} \rho_\varepsilon(J_j) P_0(\tau_j) \mathrm{e}^{-uJ_j} \\
&= \left(\prod_{i=1}^k J_{l_i^*}^{n_i} \delta_{\tau_{l_i^*}^*}(\theta_i^*) \mathrm{e}^{-uJ_{l_i^*}} \rho_\varepsilon(J_{l_i^*}) P_0(J_{l_i^*}) \right) \\
&\quad \times \left(\prod_{j \neq \{l_1^*,\dots,l_k^*\}} e^{-uJ_j} \rho_\varepsilon(J_j) P_0(\tau_j) \right). \qquad (22)
\end{aligned}
$$

The first factor in (22) refers to the unnormalized *allocated* process: the support is $\boldsymbol{\theta}^*$, while the jumps follows independent restricted gamma densities, as clearly observed after (19). This shows point 2. of the Proposition.

On the other hand, the second factor in (22) shows that the *non-allocated* jumps are indeed the jumps of ε-NGG$(\sigma, \kappa, \omega+u, P_0)$ process, given that exactly N_{na} jumps of the process were obtained; moreover, the conditional distribution of N_{na} is described in (21). This shows point 1. of the Proposition.

Point 3 follows straightforward from (22).
Normalization of the jumps (*allocated* and *non-allocated*) gives 4.
With regard to 5., we need to integrate out N_ε in $\mathscr{L}(N_\varepsilon, u, \boldsymbol{\theta})$ displayed in (20). We have already made these computations in the proof of formula (5), and thus $f_{U|\boldsymbol{\theta}^*}(u|\boldsymbol{\theta}^*)$ is proportional to the integrand in (5).

8.6 Appendix 6: Details of the blocked Gibbs sampler

We explicitly derive every step of the Gibbs sampler in Fig. 1, starting from the joint distribution of data and parameters in (11).

1. The first step is straightforward, since

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

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

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

   and this proves Step 2.
3. As far as $\mathscr{L}(P_\varepsilon, \varepsilon, \sigma, \kappa|u, \boldsymbol{\theta}, \boldsymbol{X})$ is concerned, we have

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

   so that Step 3. can be split into two consecutive substeps. First we simulate from $\mathscr{L}(\varepsilon, \sigma, \kappa|u, \boldsymbol{\theta})$ as follows: we integrate out N_ε (or equivalently N_{na}) from (20) and obtain

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

   In practical terms, Step 3a can be obtained in three substeps:

$$\mathscr{L}(\varepsilon|u, \boldsymbol{\theta}, \boldsymbol{X}) \propto \prod_{i=1}^{k} \Gamma(n_i - \sigma, \varepsilon(u+\omega)) \mathrm{e}^{(\Lambda_{\varepsilon,u} - \Lambda_\varepsilon)}$$
$$\times \frac{\Lambda_{\varepsilon,u} + k}{\Gamma(-\sigma, \omega\varepsilon)} \pi(\varepsilon), \tag{23}$$

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

$$\mathscr{L}(\kappa|u, \boldsymbol{\theta}, \boldsymbol{X}) = p_1 gamma(\alpha + k, R + \beta)$$
$$+ (1 - p_1) gamma(\alpha + k - 1, R + \beta), \tag{25}$$

   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 is equal to

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

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

References

Argiento, R., Guglielmi, A., Pievatolo, A.: Bayesian density estimation and model selection using nonparametric hierarchical mixtures. Comput. Stat. Data Anal. **54**, 816–832 (2010)

Argiento, R., Cremaschi, A., Guglielmi, A.: A "density-based" algorithm for cluster analysis using species sampling Gaussian mixture models. J. Comput. Graph. Stat. **23**, 1126–1142 (2014)

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

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

Bianchini, I.: A new finite approximation for the NGG mixture model: an application to density estimation. In: The Contribution of Young Researchers to Bayesian Statistics: Proceedings of BAYSM2014. Springer, Berlin (2015)

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

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

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

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., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell. **2**, 65–73 (1998)
Escobar, M., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**, 577–588 (1995)
Favaro, S., Teh, Y.: MCMC for normalized random measure mixture models. Stat. Sci. **28**(3), 335–359 (2013)
Favaro, S., Walker, S.G.: Slice sampling σ-stable Poisson-Kingman mixture models. J. Comput. Graph. Stat. **22**(4), 830–847 (2013)
Favaro, S., Guglielmi, A., Walker, S.: A class of measure-valued Markov chains and Bayesian nonparametrics. Bernoulli **18**(3), 1002–1030 (2012)
Ferguson, T.S., Klass, M.: A representation of independent increment processes without Gaussian components. Ann. Math. Stat. **43**, 1634–1643 (1972)
Gelfand, A.E., Kottas, A.: A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. J. Comput. Graph. Stat. **11**, 289–305 (2002)
Gradshteyn, I., Ryzhik, L.: Table of Integrals, Series, and Products, 6th edn. Academic Press, San Diego (2000)
Griffin, J., Walker, S.G.: Posterior simulation of normalized random measure mixtures. J. Comput. Graph. Stat. **20**, 241–259 (2011)
Griffin, J.E.: An adaptive truncation method for inference in Bayesian nonparametric models. Stat. Comput. doi:10.1007/s11222-014-9519-4 (2014)
Griffin, J.E., Kolossiatis, M., Steel, M.F.: Comparing distributions by using dependent normalized random-measure mixtures. J. R. Stat. Soc. B **75**(3), 499–529 (2013)
Ishwaran, H., James, L.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**, 161–173 (2001)
Ishwaran, H., Zarepour, M.: Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. Biometrika **87**, 371–390 (2000)
Ishwaran, H., Zarepour, M.: Exact and approximate sum representations for the Dirichlet process. Can. J. Stat. **30**, 269–283 (2002)
James, L., Lijoi, A., Prünster, I.: Posterior analysis for normalized random measures with independent increments. Scand. J. Stat. **36**, 76–97 (2009)
Kallenberg, O.: Random Measures, 4th edn. Akademie, Berlin (1983)
Kalli, M., Griffin, J.E., Walker, S.G.: Slice sampling mixture models. Stat. Comput. **21**(1), 93–105 (2011)
Kingman, J.F.C.: Poisson Processes, vol. 3. Oxford University Press, Oxford (1993)
Lijoi, A., Mena, R.H., Prünster, I.: Controlling the reinforcement in Bayesian nonparametric mixture models. J. R. Stat. Soc. B **69**, 715–740 (2007)
Lijoi, A., Prunster, I., Walker, S.G.: Investigating nonparametric priors with Gibbs structure. Stat. Sin. **18**, 1653–1668 (2008)
Lijoi, A., Nipoti, B., Prunster, I.: Bayesian inference with dependent normalized completely random measures. Bernoulli **20**, 1260–1291 (2014)
MacEachern, S.N.: Computational methods for mixture of Dirichlet process models. In: Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statistics, vol. 133, pp. 23–43. Springer, New York (1998)
Muliere, P., Tardella, L.: Approximating distributions of random functionals of Ferguson-Dirichlet priors. Can. J. Stat. **26**(2), 283–297 (1998)
Neal, R.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**, 249–265 (2000)
Papaspiliopoulos, O., Roberts, G.O.: Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika **95**, 169–186 (2008)
Pitman, J.: Some developments of the Blackwell-Macqueen urn scheme. In: Ferguson TS, Shapley LS, Macqueen JB (eds) Statistics, Probability and Game Theory: Papers in Honor of David Blackwell. IMS Lecture Notes-Monograph Series, vol. 30, pp. 245–267. Institute of Mathematical Statistics, Hayward (1996)
Pitman, J.: Poisson-Kingman partitions. In: Science and Statistics: A Festschrift for Terry Speed. IMS Lecture Notes-Monograph Series, vol. 40, pp. 1–34. Institute of Mathematical Statistics, Hayward (2003)
Pitman, J.: Combinatorial Stochastic Processes. Lecture Notes in Mathematics, vol. 1875, pp. 1–34. Springer, New York (2006)
Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of random measures with independent increments. Ann. Stat. **31**, 560–585 (2003)
Sethuraman, J.: A constructive definition of Dirichlet priors. Stat. Sinica **4**(2), 639–650 (1994)
Walker, S.G.: Sampling the Dirichlet mixture model with slices. Commun. Stat. Simulat. **36**, 45–54 (2007)