On a class of sigma-stable Poisson-Kingman models and an effective marginalized sampler

Stat Comput (2015) 25:67–78
DOI 10.1007/s11222-014-9499-4

On a class of σ-stable Poisson–Kingman models and an effective marginalized sampler

S. Favaro · M. Lomeli · Y. W. Teh

Received: 15 March 2014 / Accepted: 29 July 2014 / Published online: 22 August 2014

Abstract We investigate the use of a large class of discrete random probability measures, which is referred to as the class $\mathcal{Q}$, in the context of Bayesian nonparametric mixture modeling. The class $\mathcal{Q}$ encompasses both the the two-parameter Poisson–Dirichlet process and the normalized generalized Gamma process, thus allowing us to comparatively study the inferential advantages of these two well-known nonparametric priors. Apart from a highly flexible parameterization, the distinguishing feature of the class $\mathcal{Q}$ is the availability of a tractable posterior distribution. This feature, in turn, leads to derive an efficient marginal MCMC algorithm for posterior sampling within the framework of mixture models. We demonstrate the efficacy of our modeling framework on both one-dimensional and multi-dimensional datasets.

Keywords Bayesian nonparametrics · Normalized generalized Gamma process · Marginalized MCMC sampler · Mixture model · σ-Stable Poisson–Kingman model · Two parameter Poisson–Dirichlet process

Electronic supplementary material The online version of this article (doi:10.1007/s11222-014-9499-4) contains supplementary material, which is available to authorized users.

S. Favaro
University of Torino and Collegio Carlo Alberto, Turin, Italy
e-mail: stefano.favaro@unito.it

M. Lomeli
Gatsby Computational Neuroscience Unit, University College London, London, UK
e-mail: mlomeli@gatsby.ucl.ac.uk

Y. W. Teh (✉)
Department of Statistics, University of Oxford, Oxford, UK
e-mail: y.w.teh@stats.ox.ac.uk

1 Introduction

A general approach for modeling continuous data in Bayesian nonparametrics was first proposed by Lo (1984) in terms of an infinite dimensional mixture model, and it is nowadays the subject of a rich and active literature. Let $P = \sum_{i\geq 1} P_i \delta_{\tilde{X}_i}$ be a random probability measure (RPM) such that $(P_i)_{i\geq 1}$ are non-negative r.v.s. that add up to one and $(\tilde{X}_i)_{i\geq 1}$ are r.v.s., independent of $(P_i)_{i\geq 1}$, and independent and identically distributed. Given a collection of continuous observations $Y_1, \ldots, Y_n$, the infinite dimensional mixture model is defined as

$$\begin{aligned} Y_i \mid X_i &\overset{\text{ind}}{\sim} G(\cdot \mid X_i) \\ X_i \mid P &\overset{\text{iid}}{\sim} P \qquad i = 1, \ldots, n \\ P &\sim \mathscr{P}, \end{aligned} \tag{1}$$

where $G(\cdot \mid X_i)$ is a continuous distribution parameterized by X_i and admitting a density function $g(\cdot \mid X_i)$ with respect to a dominating measure. The distribution $G(\cdot \mid X_i)$ is referred to as the kernel, whereas P is the mixing measure. By the a.s. discreteness of P, each pair of the X_i's takes on the same value with positive probability, with this value identifying a mixture component. In such a way, the r.v.s. X_i's allocate the Y_i's to a random number of mixture components, thus providing a model for the unknown number clusters within the data.

Lo (1984) assumed P to be the Dirichlet process (DP) by Ferguson (1973). Under this assumption (1) is termed DP mixture model. Several MCMC methods have been proposed for posterior sampling from the DP mixture model. On one hand, marginal MCMC methods remove the infinite dimensional aspect of the mixture model by exploiting the tractable marginalization with respect to the DP. See, e.g., Escobar

(1994), MacEachern (1994) and Escobar and West (1995) for early contributions, and Neal (2000) for an overview with noteworthy developments such as the celebrated Algorithm 8. On the other hand, conditional MCMC methods maintain the infinite dimensionality of the DP mixture model and find appropriate ways for sampling a sufficient but finite number of the atoms of the DP. See, e.g., Ishwaran and James (2001), Muliere and Tardella (1998), Walker (2007), Papaspiliopoulos and Roberts (2008), Papaspiliopoulos (2008) and Kalli et al. (2011).

It is apparent that one can replace the DP mixing measure with any discrete RPMs. Ishwaran and James (2001) proposed to replace the DP with the two parameter Poisson–Dirichet (PD) process, also known as Pitman–Yor process, introduced in Perman et al. (1992). See Pitman and Yor (1997) and Pitman (2006) for a detailed account on the two parameter PD process. Nieto-Barajas et al. (2004) proposed to replace the DP with the normalized random measures (NRMs) introduced in Regazzini et al. (2002). See Lijoi and Prünster (2010) for an up-to-date review of NRMs. As a notable example of NRM, Lijoi et al. (2007) focussed on the normalized generalized Gamma (GG) process. Marginal and conditional MCMC methods have been developed under mixing measures belonging to the classes of stick-breaking random probability measures and NRMs. See, e.g., Ishwaran and James (2001), Lijoi et al. (2007), Griffin and Walker (2009), Barrios et al. (2013), Favaro and Teh (2013) and Favaro and Walker (2013).

The two parameter PD process and the normalized GG process are noteworthy examples of σ-stable Poisson–Kingman models. These models form a large class of discrete RPMs, and correspond to the Gibbs-type RPMs with positive indices, introduced by Pitman (2003) and further investigated by Gnedin and Pitman (2005). Ishwaran and James (2001) and Lijoi et al. (2007) showed that the two parameter PD process and the normalized GG process are valid alternatives to the DP: while they preserve almost the same mathematical tractability as the DP, they have more elaborate clustering properties. Precisely, it is well known that the DP allocates observations to a specific mixture component with a probability depending solely on the number of times that the mixture component occurs. In contrast, under the two parameter PD process and the normalized GG process, the allocation probability depends heavily on the number of mixture components. Such a more flexible allocation mechanism, which is peculiar to σ-stable Poisson–Kingman models, turns out to be a key feature for making inference under the mixture model (1). See De Blasi et al. (2013) for an up-to-date review.

While the main advantages of replacing the DP with the two parameter PD process and the normalized GG process are well known from the seminal papers Ishwaran and James (2001) and Lijoi et al. (2007), there are no comprehensive studies which investigate the inferential advantages, if there are some, of replacing a two parameter PD process with a normalized GG process and vice versa. More generally, in the context of Bayesian nonparametric mixture modeling, there are no comprehensive studies which investigate the use of σ-stable Poisson–Kingman models different from the two parameter PD process and the normalized GG process. In this paper we shed some light on these aspects by suitably reparameterizing the two parameter PD process and the normalized GG process into a unique class of discrete RPMs. Such a class will be referred to as the class $\mathcal{Q}$.

The definition of the class $\mathcal{Q}$ arises from Proposition 21 in Pitman and Yor (1997), where a noteworthy relationship between the two parameter PD process and the normalized GG process is established. While maintaing the same mathematical tractability and clustering properties as the two parameter PD process and the normalized GG process, the class $\mathcal{Q}$ stands out for a highly flexible parameterization. Recently, the idea of exploiting Proposition 21 in Pitman and Yor (1997) in order to define a flexible class of tractable discrete RPMs has been independently proposed in James (2013). There, the class $\mathcal{Q}$ is referred to as the Poisson–Gamma (PG) class, and an explicit stick-breaking representation for RPMs in the PG class is derived and investigated. We point out that the use of the class $\mathcal{Q}$ also appeared, although in a context different from Bayesian nonparametrics, in James (2010). Differently from the contributions in James (2010) and (2013), in this paper we study the use of the class $\mathcal{Q}$ in the context of Bayesian nonparametric mixture modeling under the hierarchical framework (1).

Within the class of σ-stable Poisson–Kingman models, a distinguishing feature of the class $\mathcal{Q}$ is the availability of a tractable posterior distribution. The posterior distribution of the two parameter PD process was first derived in Pitman (1996a) by means of constructive arguments relying on the sampling properties of the process. See Lijoi and Prünster (2010) for an alternative proof. Recently, James et al. (2009) provided a posterior characterization for the entire class of NRMs and, as a special case, they obtained the posterior distribution of the normalized GG process. See Lijoi and Prünster (2010) for a posterior counterpart of Proposition 21 in Pitman and Yor (1997), namely a relation between the posterior distributions of two parameter PD process and the normalized GG process. In this paper we present a posterior characterization of the class $\mathcal{Q}$, thus providing a unified framework for the posterior distributions of the two parameter PD process and the normalized GG process.

The posterior characterization of the class $\mathcal{Q}$ leads to derive a marginal MCMC algorithm for posterior sampling from (1). This is the second distinguishing feature of the class $\mathcal{Q}$. Making use of our posterior analysis, we develop an efficient marginalized sampler that is uniformly applicable across the whole $\mathcal{Q}$ class. As compared to conditional samplers, which explicitly instantiate the RPM underlying the model using a variety of truncation techniques, marginal samplers integrate out the RPM, working directly with the induced random partition, and can thus mix more rapidly. Our marginal sampler is an extension, from NRMs to the class $\mathcal{Q}$, of the Algorithm 8 with Reuse developed by Favaro and Teh (2013). We demonstrate that while our marginal sampler is applicable to the entire class $\mathcal{Q}$, for two-parameter PD processes it is only slightly less efficient than the standard marginal sampler which exploit the simple analytic form of the distribution of the clustering structure.

The paper is structured as follows. In Sect. 2 we define the class $\mathcal{Q}$ and we present some posterior and marginal characterizations for priors belonging to the class $\mathcal{Q}$. In Sect. 3 we describe the Reuse algorithm for posterior sampling from (1) with a mixing measure in the class $\mathcal{Q}$, and we present some simulation studies in Sect. 4.

2 Preliminaries

We start by recalling the definition of completely random measure (CRM) introduced in Kingman (1967). Let $\mathbb{X}$ be a Polish space endowed with the Borel σ-field $\mathscr{X}$. A CRM μ is a random element on the space of boundedly finite measures on $\mathbb{X}$ and such that, for any $\{A_1,\ldots,A_n\}$ in $\mathscr{X}$ with $A_i \cap A_j = \emptyset$ for $i \neq j$, the r.v.s. $\mu(A_1),\ldots,\mu(A_n)$ are mutually independent. The distribution of μ is characterized by the Lévy Khintchine representation of the Laplace functional transform of μ, namely

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} f(x)\mu(dx)}\right] = \exp\left\{-\int_0^{+\infty}\int_{\mathbb{X}}\left(1-\mathrm{e}^{-sf(x)}\right)\nu(dx,ds)\right\},$$

for any $f:\mathbb{X}\to\mathbb{R}$ such that $\int_{\mathbb{X}}|f(x)|\mu(dx) < +\infty$ a.s. The Lévy intensity measure ν determines uniquely μ. Kingman (1967) showed that μ is discrete a.s. and, hence, it can be represented in terms of nonnegative random jumps $(J_i)_{i\geq 1}$ at $\mathbb{X}$-valued random locations $(\tilde{X}_i)_{i\geq 1}$, i.e.,

$$\mu(\cdot) = \sum_{i\geq 1} J_i \delta_{\tilde{X}_i}(\cdot).$$

In the present paper we consider Lévy intensity measures that can be factorized as follows: $\nu(dx,ds) = \rho(ds)\alpha_0(dx)$ where ρ is the Lévy measure driving the jump part of μ, and α_0 is the nonatomic probability measure driving the location part of μ. Such a factorization, intuitively, implies the independence between $(J_i)_{i\geq 1}$ and $(\tilde{X}_i)_{i\geq 1}$. Therefore, without loss of generality, the random locations $(\tilde{X}_i)_{i\geq 1}$ can be assumed to be r.v.s. independent and identically distributed according to α_0.

2.1 σ-PK models

The class of σ-PK models was introduced in Pitman (2003) as a generalization of the normalized σ-stable process by Kingman (1975). See Pitman (2006) for a detailed account. Specifically, for any $\sigma \in (0,1)$ let μ_σ be a σ-stable CRM, namely a CRM characterized by the Lévy intensity measure

$$\nu(dx,ds) = \rho_\sigma(ds)\alpha_0(dx) = \frac{\sigma}{\Gamma(1-\sigma)}s^{-\sigma-1}ds\alpha_0(dx).$$

Since $\int_0^\epsilon \rho_\sigma(s)ds = +\infty$, for any $\epsilon > 0$, $T_\sigma = \sum_{i\geq 1} J_i$ is finite a.s. In particular, the total mass T_σ is a positive σ-stable r.v. and its density function is denoted by f_σ. The normalized σ-stable process is defined as the a.s. discrete RPM

$$P_\sigma(\cdot) = \frac{\mu_\sigma(\cdot)}{T_\sigma} = \sum_{i\geq 1} P_i \delta_{\tilde{X}_i}(\cdot),$$

with $P_i = J_i/T_\sigma$ and where $(\tilde{X}_i)_{i\geq 1}$ are r.v.s., independent of $(P_i)_{i\geq 1}$, and independent and identically distributed according to α_0. See Regazzini et al. (2002) and James et al. (2009) for a generalization of P_σ by replacing μ_σ with any CRM. Such a generalization gives rise to the class of NRMs.

A σ-PK model is defined as a generalization of P_σ which is obtained by suitably deforming, or tilting, the distribution of the total mass T_σ. Precisely, let $(P_{(i)})_{i\geq 1}$ be the decreasing rearrangement of $(P_i)_{i\geq 1}$ and let $T_{\sigma,h}$ be a r.v. with density function $f_{T_{\sigma,h}}(t) = h(t)f_\sigma(t)$, for any nonnegative function h. Let $\mathrm{PK}(\rho_\sigma\,|\,t)$ be the conditional distribution of $(P_{(i)})_{i\geq 1}$ given $T_{\sigma,h} = t$. A σ-PK model with parameter h is defined as the a.s. discrete RPM

$$P_{\sigma,h}(\cdot) = \sum_{i\geq 1} P_{(i)}\delta_{\tilde{X}_i}(\cdot),$$

where $(\tilde{X}_i)_{i\geq 1}$ are r.v.s. independent of $(P_{(i)})_{i\geq 1}$ and independent and identically distributed as α_0, whereas $(P_{(i)})_{i\geq 1}$ is distributed as $\int_0^{+\infty}\mathrm{PK}(\rho_\sigma\,|\,t)f_{T_{\sigma,h}}(t)dt$. Accordingly, we can write $P_{\sigma,h}(\cdot) = \mu_{\sigma,h}(\cdot)/T_{\sigma,h}$ where $\mu_{\sigma,h}$ is an a.s. discrete random measure with distribution $\mathbb{P}_{\sigma,h}$ absolutely continuous with respect to the distribution $\mathbb{P}_\sigma$ of μ_σ, and such that $\mathrm{d}\mathbb{P}_{\sigma,h}(\mu)/\mathrm{d}\mathbb{P}_\sigma = h(\mu(\mathbb{X}))$, and $T_{\sigma,h}$ is the total mass of $\mu_{\sigma,h}$ with density function $f_{T_{\sigma,h}}$.

Clearly P_σ is the σ-PK model corresponding to the choice $h(t) = 1$. The two parameter PD process and the normalized GG process are other two noteworthy σ-PK models, and they both include P_σ as a special case. In the next two examples we briefly recall their definitions.

Example 1 For any $\sigma \in (0,1)$ and $\theta > -\sigma$ the two parameter PD process is a σ-PK model with parameter h of the form

$$h(t) = p(t;\sigma,\theta) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}t^{-\theta}.$$

We denote by $\mathcal{P}_{\sigma,\theta}$ the two parameter PD process. We refer to Perman et al. (1992), James (2002, 2013), Pitman and Yor (1997) and James et al. (2008) for details on $\mathcal{P}_{\sigma,\theta}$.

Example 2 For any $\sigma \in (0,1)$ and $\tau > 0$ the normalized GG process is a σ-PK model with parameter h of the form

$$h(t) = g(t;\sigma,\tau) = \mathrm{e}^{\tau^{\sigma}-\tau t}.$$

We denote by $\mathcal{G}_{\sigma,\tau}$ the normalized GG process. We refer to James (2002, 2013), Pitman (2003), Lijoi et al. (2005, 2007) and Favaro et al. (2012) for details on $\mathcal{G}_{\sigma,\tau}$.

2.2 Sampling properties of σ-PK models

Pitman (2003), and later on Gnedin and Pitman (2005), provided a comprehensive study of the sampling properties of σ-PK models. Let $\mathscr{P}_{\sigma,h}$ be the distribution of $P_{\sigma,h}$, and let $\mathbf{X} = (X_1,\ldots,X_n)$ be a sample from $P_{\sigma,h}$, namely

$$\begin{aligned} X_i \mid P_{\sigma,h} &\overset{\text{iid}}{\sim} P_{\sigma,h} \qquad i = 1,\ldots,n \\ P_{\sigma,h} &\sim \mathscr{P}_{\sigma,h}, \end{aligned}$$

By the discreteness of $P_{\sigma,h}$, $\mathbf{X}$ induces a random partition Π_n of $[n] = \{1,\ldots,n\}$. Specifically, Π_n is defined in such a way that indices i and j belong to the same block c of Π_n if and only if $X_i = X_j$. According to Pitman (1995), Π_n is exchangeable, namely the distribution of Π_n depends only on the number $|\Pi_n|$ of blocks and their frequencies $\{|c| : c \in \Pi_n\}$. This distribution is known as the exchangeable partition probability function (EPPF).

Pitman (2003) characterized the EPPF induced by the σ-PK model in terms of a suitable product form, a feature which is crucial for guaranteeing mathematical tractability. In particular, a sample $\mathbf{X}$ from $P_{\sigma,h}$ induces an exchangeable random partition Π_n admitting the EPPF

$$\Pr[\Pi_n = \pi] = V_{n,|\pi|}\prod_{c\in\pi}(1-\sigma)_{|c|-1}, \tag{2}$$

where $(a)_{(n)} = \prod_{i=0}^{n-1}(a+i)$ with the proviso $(a)_{(0)} = 1$, and

$$\begin{aligned} V_{n,|\pi|} = &\frac{\sigma^{|\pi|}}{\Gamma(n-|\pi|\sigma)} \\ &\times \int_0^{+\infty}\int_0^t s^{n-|\pi|\sigma-1}t^{-n}h(t)f_\sigma(t-s)dsdt. \end{aligned} \tag{3}$$

The EPPF (2), with $V_{n,|\pi|}$ in (3), has been characterized by Gnedin and Pitman (2005) in the class of the Gibbs-type EPPFs. See also the monograph by Pitman (2006) for a comprehensive account on Gibbs-type EPPFs, and De Blasi et al. (2013) for the use of Gibbs-type EPPFs in Bayesian nonparametrics.

According to (2) and (3), σ and h play a crucial role in determining the probabilistic structure of the random partition induced by the sample $\mathbf{X}$ from $P_{\sigma,h}$. By marginalizing (2) with respect to the frequencies one recovers the distribution of the number $|\Pi_n|$ of distinct observations in $\mathbf{X}$. Specifically,

$$\Pr[|\Pi_n| = |\pi|] = V_{n,|\pi|}\frac{\mathscr{C}(n,|\pi|;\sigma)}{\sigma^{|\pi|}}, \tag{4}$$

where $\mathscr{C}$ is the generalized factorial. See Charalambides (2005) for details. From (2) and (4), the conditional distribution of Π_n given $|\Pi_n| = |\pi|$ depends only upon σ. In other words the frequencies of Π_n are independent of h given $|\Pi_n| = |\pi|$. Hence, h determines the distribution of Π_n only via $|\Pi_n|$.

The important role of σ and h also appears in the study of the large n asymptotic behaviour of $|\Pi_n|$. This asymptotic behaviour has been characterized in Proposition 13 by Pitman (2003) by means of a positive and almost surely finite r.v. $S_{\sigma,h}$. Specifically, as $n \to +\infty$, one has

$$\frac{|\Pi_n|}{n^{\sigma}} \xrightarrow{\text{a.s.}} S_{\sigma,h}.$$

The r.v. $S_{\sigma,h}$, which is referred to as the σ-diversity of $P_{\sigma,h}$, is related to the total mass $T_{\sigma,h}$ by the following identity

$$S_{\sigma,h} \overset{\text{d}}{=} (T_{\sigma,h})^{-\sigma}. \tag{5}$$

The distribution of $T_{\sigma,h}$ is governed by σ and h through $f_{T_{\sigma,h}}$ and, hence, the asymptotic behaviour of $|\Pi_n|$ is ultimately governed by σ and h. The specification of σ and h thus encodes a knowledge on the asymptotic number of blocks in the partition induced by a sample from $P_{\sigma,h}$.

3 The class $\mathcal{Q}$

The definition of the class $\mathcal{Q}$ aries from Proposition 21 in Pitman and Yor (1997), where a noteworthy relationship between $\mathcal{P}_{\sigma,\theta}$ and $\mathcal{G}_{\sigma,\tau}$ is established. Such a result holds for any $\sigma \in (0,1)$ and $\theta > 0$ and it relies on a suitable randomization of the parameter τ in $\mathcal{G}_{\sigma,\tau}$. In particular, let

$$F_{\sigma,\theta}(d\tau) = \frac{\sigma}{\Gamma(\theta/\sigma)}\tau^{\theta-1}e^{-\tau^{\sigma}}d\tau, \tag{6}$$

for any $\sigma \in (0,1)$ and $\theta > 0$. Then, the marginal distribution of $\mathcal{G}_{\sigma,\tau}$ when τ is randomized with respect to $F_{\sigma,\theta}$ coincides with the distribution of $\mathcal{P}_{\sigma,\theta}$. This link can be shown in terms of the parameters p and g defining the distributions of $\mathcal{P}_{\sigma,\theta}$ and $\mathcal{G}_{\sigma,\tau}$, respectively. Indeed

$$\begin{tikzcd}
 & \sigma - PK \arrow[d, "h=q(\sigma,F)"] & \\
 & \mathcal{Q} \arrow[dl, "F=F_{\sigma,\theta}"'] \arrow[dr, "F=\delta_\tau"] & \\
PY \arrow[d, "\sigma\to0"] \arrow[dr, "\theta=0"] & & NGG \arrow[dll, "\sigma\to0"] \arrow[dl, "\tau=0"] \arrow[d, "\sigma=0.5"] \\
DP & NS & NIG
\end{tikzcd}$$

Fig. 1 Relationships between the class $\mathcal{Q}$ and the Dirichlet process (DP), the normalized σ-stable process (NS), the normalized inverse Gaussian process (NIG), the normalized GG process (NGG), the two parameter PD process (PY), the class of σ-stable PK models (σ-PK)

$$p(t;\sigma,\theta)=\int_0^{+\infty}g(t;\sigma,\tau)F_{\sigma,\theta}(d\tau).$$

It is apparent that one can replace the distribution (6) with any other distribution F over the positive real line, or over a subset of it. This procedure leads to the definition of the class $\mathcal{Q}$, namely a flexible class of priors indexed by σ and F, and including as special cases both the $\mathcal{P}_{\sigma,\theta}$ and $\mathcal{G}_{\sigma,\tau}$.

Definition 3 A prior in the class $\mathcal{Q}$ is a σ-PK model with parameter

$$h(t)=q(t;\sigma,F)=\int_D g(t;\sigma,\tau)F(d\tau), \tag{7}$$

where F is a distribution over any subset D of the positive real line.

We denote by $\mathcal{Q}_{\sigma,F}$ a prior in the class $\mathcal{Q}$. According to Definition 3, $\mathcal{Q}_{\sigma,F}(\cdot)=\mu_{\sigma,q}(\cdot)/T_{\sigma,q}$ where $\mu_{\sigma,q}$ is an a.s. discrete random measure with distribution $\mathbb{P}_{\sigma,q}$ absolutely continuous with respect to $\mathbb{P}_\sigma$, and such that $\mathrm{d}\mathbb{P}_{\sigma,q}(\mu)/\mathrm{d}\mathbb{P}_\sigma=q(\mu(\mathbb{X});\sigma,F)$, and $T_{\sigma,q}$ is the total mass of $\mu_{\sigma,q}$ with density function $f_{T_{\sigma,q}}(t)=q(t;\sigma,F)f_\sigma(t)$. If F is the generalized Gamma distribution in (6), then $\mathcal{Q}_{\sigma,F}$ coincides with $\mathcal{P}_{\sigma,\theta}$. Note that such a choice of F does not include the case $-\sigma<\theta<0$. If $F=\delta_\tau$, for any $\tau>0$, then $\mathcal{Q}_{\sigma,F}$ coincides with $\mathcal{G}_{\sigma,\tau}$. The normalized σ-stable process by Kingman (1975) corresponds to $\mathcal{G}_{\sigma,0}$ and $\mathcal{P}_{\sigma,0}$, whereas the normalized inverse Gaussian process by Lijoi et al. (2005) corresponds to $\mathcal{G}_{1/2,\tau}$. Figure 1 shows the relationships between elements in $\mathcal{Q}$.

Intuitively, the interpretation of F is directly related to the interpretation of the parameter τ in the normalized GG process $\mathcal{G}_{\sigma,\tau}$. In this respect, Lijoi et al. (2007) showed that τ and σ tune the distribution of $|\Pi_n|$ in a sample from $\mathcal{G}_{\sigma,\tau}$. See also Lijoi et al. (2005) for details. The parameter $\tau>0$ controls the location of the distribution of $|\Pi_n|$: the bigger τ the larger the expected number of distinct observations tends to be. The parameter $\sigma\in(0,1)$ controls the flatness of the distribution of $|\Pi_n|$: the bigger σ the flatter is the distribution of $|\Pi_n|$. Accordingly, the more general parameterization of $\mathcal{Q}_{\sigma,F}$ determines a more flexible control in the clustering behaviour induced by $\mathcal{Q}_{\sigma,F}$. Note however, from our discussion in Sect. 2 on the random partition induced by a sample from a σ-PK model, that the only effect of F on the clustering behaviour is through the number of cluster $|\Pi_n|$, with the clustering behaviour conditioned on $|\Pi_n|=|\pi|$ only depending on the parameter σ.

3.1 Posterior analysis

Let $\mathbf{X}$ be a sample of size n from $\mathcal{Q}_{\sigma,F}$, and recall that one can always represent $\mathbf{X}$ in terms of a random partition Π_n of $[n]$, consisting of $|\Pi_n|$ distinct observations, where each $c\in\Pi_n$ corresponds to a distinct observed value $\tilde{X}_c$ with frequency $|c|$. Given $\mathbf{X}$, hereafter we present a comprehensive posterior analysis of $\mathcal{Q}_{\sigma,F}$. We start by providing a posterior characterization of $\mu_{\sigma,q}$ in terms of auxiliary r.v.s. This, then, will lead to the distribution of $\mathcal{Q}_{\sigma,F}\,|\,\mathbf{X}$. The posterior characterization of $\mu_{\sigma,q}$ follows by combining the definition of $\mu_{\sigma,q}$ with the posterior characterization of $\mu_{\sigma,g}$ derived from Theorem 1 of James et al. (2009). Indeed, given a r.v. $\mathcal{T}$ with distribution F, $\mu_{\sigma,q}\,|\,\mathcal{T}$ is an a.s. discrete random measure with distribution $\mathbb{P}_{\sigma,g}$ absolutely continuous with respect to $\mathbb{P}_\sigma$, and such that $\mathrm{d}\mathbb{P}_{\sigma,g}(\mu)/\mathrm{d}\mathbb{P}_\sigma=g(\mu(\mathbb{X});\sigma,\mathcal{T})$. See Example 3.24 in Lijoi and Prünster (2010) for details. To make the paper self-contained, in the online Appendix we present a proof of the next proposition which relies on the absolute continuity of $\mathbb{P}_{\sigma,q}$ with respect to $\mathbb{P}_\sigma$. This approach was first exploited in Lijoi and Prünster (2010) to provide an alternative proof of the posterior distribution of the two parameter PD process.

Proposition 4 *Let* $\mathbf{X}$ *be a sample from* $\mathcal{Q}_{\sigma,F}$ *and let* $(\mathcal{T},U)$ *be a r.v. such that*

$$\Pr[\mathcal{T}\in d\tau, U\in du\,|\,\mathbf{X}]$$
$$=\frac{u^{n-1}e^{\tau^\sigma-(u+\tau)^\sigma}(u+\tau)^{\sigma|\pi|-n}du\,F(d\tau)}{\int_D\int_0^{+\infty}u^{n-1}e^{\tau^\sigma-(u+\tau)^\sigma}(u+\tau)^{\sigma|\pi|-n}du\,F(d\tau)}.$$

Then,

$$\mu_{\sigma,q}\,|\,(\mathcal{T},U,\mathbf{X})\overset{d}{=}\sum_{c\in\pi}J_{c,\mathcal{T},U}\delta_{\tilde{X}_c}+\mu_{\sigma,g^*} \tag{8}$$

where

(i) *μ_{σ,g^*} is an a.s. discrete random measure with distribution $\mathbb{P}_{\sigma,g^*}$ absolutely continuous with respect to $\mathbb{P}_\sigma$, and such that $\mathrm{d}\mathbb{P}_{\sigma,g^*}(\mu)/\mathrm{d}\mathbb{P}_\sigma = g(\mu(\mathbb{X}); \sigma, \mathcal{T} + U)$*

(ii) *the $J_{c,\mathcal{T},U}$'s are independent r.v.s., independent of μ_{σ,g^*}, and distributed according to a Gamma distribution with parameter $(|c| - \sigma, \mathcal{T} + U)$, for any $c \in \pi$.*

Note that the continuous part of (8) coincides with μ_{σ,g^*}, which depends on $\mathcal{T}$ and U only through $V = U + \mathcal{T}$. Accordingly, the continuous part of $\mathcal{Q}_{\sigma,F} \,|\, \mathbf{X}$ will be also in the class $\mathcal{Q}$ with the distribution of $V \,|\, \mathbf{X}$ playing the role of the mixing parameter F. See the online Appendix for an explicit expression of the distribution of $V \,|\, \mathbf{X}$. In the next two examples we apply Proposition 4 under the assumptions that F coincides with (6) and F coincides with δ_τ, for any $\tau > 0$, respectively. The former assumption leads to the posterior characterization of the two parameter PD process in Lijoi and Prünster (2010), whereas the latter leads to the posterior characterization of the normalized GG process in James et al. (2009).

Example 5 For any $\sigma \in (0, 1)$ and $\theta > 0$, let $\mathbf{X}$ be a sample from $\mathcal{P}_{\sigma,\theta}$. Then,

$$\mu_{\sigma,p} \,|\, (V, \mathbf{X}) \overset{\mathrm{d}}{=} \sum_{c\in\pi} J_{c,V} \delta_{\tilde{X}_c} + \mu_{\sigma,g^*},$$

where

$$\Pr[V \in dv \,|\, \mathbf{X}] = \frac{\sigma}{\Gamma(\theta/\sigma + |\pi|)} v^{\theta+|\pi|\sigma-1} \mathrm{e}^{-v^\sigma} dv.$$

The $J_{c,V}$'s are independent Gamma r.v.s. with parameter $(|c| - \sigma, V)$.

Example 6 For any $\sigma \in (0, 1)$ and $\tau > 0$, let $\mathbf{X}$ be a sample from $\mathcal{G}_{\sigma,\tau}$. Then,

$$\mu_{\sigma,g} \,|\, (V, \mathbf{X}) \overset{\mathrm{d}}{=} \sum_{c\in\pi} J_{c,V} \delta_{\tilde{X}_c} + \mu_{\sigma,g^*},$$

where

$$\Pr[V \in dv \,|\, \mathbf{X}] = \frac{\sigma v^{|\pi|\sigma-n}(v-\tau)^{n-1}\mathrm{e}^{-v^\sigma}}{\sum_{i=0}^{n-1}\binom{n-1}{i}(-\tau)^i \Gamma(|\pi| - i/\sigma, \tau^\sigma)} dv.$$

The $J_{c,V}$'s are independent Gamma r.v.s. with parameter $(|c| - \sigma, V)$.

We conclude by stating the aforementioned posterior characterization of $\mathcal{Q}_{\sigma,F}$. Of course an application of the next proposition under the assumptions that F coincides with (6) and F coincides with δ_τ, for any $\tau > 0$, leads to the posterior characterizations originally provided by Pitman (1996a) and James et al. (2009), respectively.

Proposition 7 *Let $\mathbf{X}$ be a sample from $\mathcal{Q}_{\sigma,F}$. Then, the random probability measure $\mathcal{Q}_{\sigma,F} \,|\, \mathbf{X}$ is equal in distribution to*

$$\sum_{c\in\pi} W_c \delta_{\tilde{X}_c} + W_{|\pi|+1} \mathcal{Q}_{\sigma,F^*},$$

where $\mathcal{Q}_{\sigma,F^*}$ is a prior in the class $\mathcal{Q}$ with F^* being the distribution of $V \,|\, \mathbf{X}$, and $(W_1, \ldots, W_{|\pi|}, W_{|\pi|+1})$ is a r.v. on the $|\pi|$-th dimensional simplex with density function

$$\begin{aligned} &g_{(W_1,\ldots,W_{|\pi|})}(w_1, \ldots, w_j) \\ &\quad = \frac{\Gamma(n)}{\prod_{c\in\pi}\Gamma(|c|-\sigma)} \prod_{c\in\pi} w_c^{|c|-\sigma-1} \left(1 - \sum_{c\in\pi} w_c\right)^{|\pi|\sigma-1} \\ &\quad\times \frac{\int_D \mathbb{E}[(T_\sigma)^{-|\pi|\sigma} e^{\tau^\sigma - \tau \frac{T_\sigma}{1-\sum_{c\in\pi} w_c}}]F(d\tau)}{\int_D \int_\tau^{+\infty} v^{|\pi|\sigma-n}(v-\tau)^{n-1} e^{\tau^\sigma - v^\sigma} dv F(d\tau)}. \end{aligned}$$

The r.v. $(W_1, \ldots, W_{|\pi|}, W_{|\pi|+1})$ is independent of $\mathcal{Q}_{\sigma,F^*}$ if and only if F is the generalized Gamma distribution (6).

3.2 Sampling properties

Apart from the posterior distribution of $\mathcal{Q}_{\sigma,F}$, we are also interested in distributional properties of a sample $\mathbf{X}$ from $\mathcal{Q}_{\sigma,F}$. Since $\mathcal{Q}_{\sigma,F}$ is a σ-stable PK model, $\mathbf{X}$ induces an exchangeable random partition of $[n]$ which is distributed according to an EPPF of the form (2) with $V_{n,|\pi|}$ obtained by combining (3) with (7). In the next proposition we present a characterization of this EPPF in terms of auxiliary r.v.s. This characterization follows by combining the definition of $\mathcal{Q}_{\sigma,F}$ with Proposition 3 in James et al. (2009). See also Lijoi and Prünster (2010) for details.

Proposition 8 *Let $\mathbf{X}$ be a sample from $\mathcal{Q}_{\sigma,F}$. Then, one has*

$$\begin{aligned} &\Pr[\Pi_n = \pi, \{\tilde{X}_c \in dx_c : c \in \pi\}, \mathcal{T} \in d\tau, U \in du] \\ &\qquad = \frac{u^{n-1}}{\Gamma(n)} e^{-\psi_{\sigma,\tau}(u)} du F(d\tau) \prod_{c\in\pi} \kappa_{\sigma,u+\tau}(|c|)\alpha_0(d\tilde{x}_c), \end{aligned}$$

where $\psi_{\sigma,\tau}(u) = (u+\tau)^\sigma - \tau^\sigma$ and $\kappa_{\sigma,u+\tau}(m) = \sigma(1-\sigma)_{m-1}/(u+\tau)^{m-\sigma}$ are the Laplace exponent of $\mu_{\sigma,g}$ and the m-th moment of $\rho(ds) = e^{-(u+\tau)s}\rho_\sigma(ds)$, respectively.

Proposition 8 provides an augmented version, with respect to the auxiliary r.v. $(\mathcal{T}, U)$, of the EPPF induced by $\mathcal{Q}_{\sigma,F}$. In the next section Proposition 8 will be applied in order to implement the Algorithm 8 with Reuse for posterior sampling from Bayesian nonparametric mixture models with a mixing measure in the class $\mathcal{Q}$.

The class $\mathcal{Q}$ generalizes the two-parameter PD process and the normalized GG process in order to incorporate a large and tractable family of σ-PK models. Each member of $\mathcal{Q}$ is characterized by the distribution F over the tilting parameter τ of the normalized GG process. When viewed as a prior over exchangeable random partitions, each member of the class $\mathcal{Q}$ gives rise to a different distribution over the number $|\Pi_n| = |\pi|$ of clusters in the induced random partition $|\Pi_n|$, while all other aspects of $|\Pi_n|$ depend only on the parameter σ.

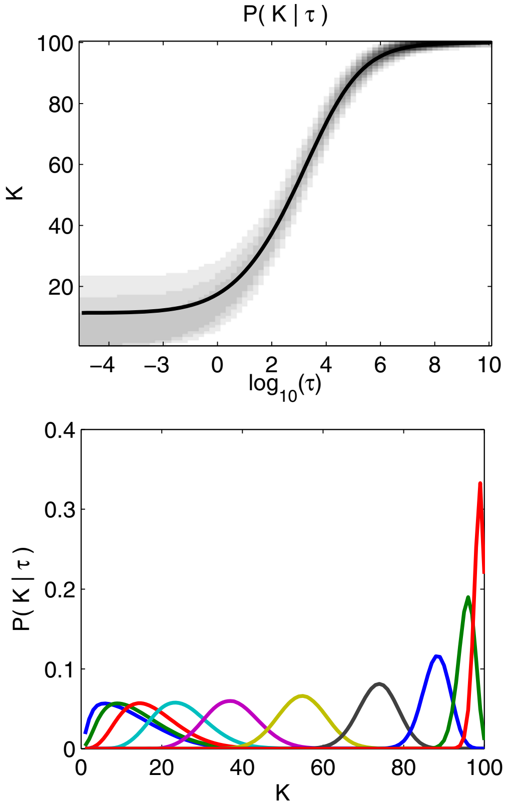

Fig. 2 Distribution over the number of clusters $|\Pi_n| = |\pi|$ under the normalized generalized Gamma process as a function of τ, with $\sigma = 0.5$ and $n = 100$. *Top* visualizing the probability mass function and mean of $|\Pi_n|$ as a function of $\log_{10}(\tau)$. *Bottom* the probability mass function of $|\Pi_n|$ for $\tau = 10^{-2}, 10^{-1}, \ldots, 10^{7}$

In Fig. 2 we visualize the distribution over $|\Pi_n|$ as it varies with τ, for $\sigma = 0.5$ and $n = 100$. We see that there is a monotonically increasing relationship between the two variables. Different distributions F induce different priors over $|\Pi_n|$ simply via the following convolution form

$$\Pr[\Pi_n = |\pi| \,|\, \sigma, F] \\ = \int_0^{+\infty} \Pr[\Pi_n = |\pi| \,|\, \sigma, \tau] F(d\tau)$$

A number of prior distributions over the random partition $|\Pi_n|$ achievable in the model are shown in Fig. 3, where we see that a variety of effects can be achieved with different distributions F, including multi-modality and larger spread as compared to the two parameter PD process.

4 Marginalized samplers

In this section, we develop a marginal MCMC algorithm for posterior simulation from a Bayesian nonparametric mixture model with a mixing measure in the class $\mathcal{Q}$. We have n observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$, each observation Y_i being associated with a latent r.v X_i. The latent r.v.s. are modeled as independent and identically distributed draws from a RPM $\mathcal{Q}_{\sigma,F}$. Formally, we can write

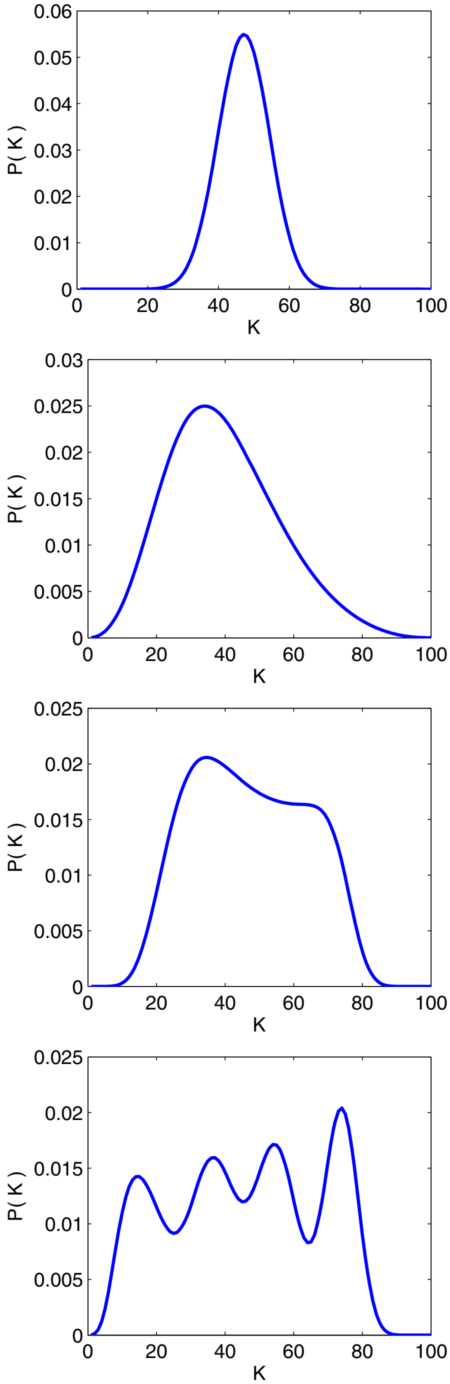

Fig. 3 Distribution over the number of clusters $|\Pi_n| = |\pi|$ under a prior in the class $\mathcal{Q}$ with $\sigma = 0.5$, $n = 100$, and different choices of the distribution F. From *top* to *bottom* a generalized Gamma distrbution leading to the two-parameter Poisson–Dirichlet process with $\theta = 10$, a Normal distribution over $\log(\tau)$ with parameters $\mu = 2\log(10)$ and $\sigma^2 = \log(10)$, a Uniform distribution over $\log(\tau)$ with range $[\log(10), \log(10^4)]$, and a discrete distribution with probability 1/4 at points $\tau = 1, 10^2, 10^3, 10^4$

$$\begin{aligned}
Y_i \mid X_i &\overset{\text{ind}}{\sim} G(\cdot \mid X_i) \\
X_i \mid \mathcal{Q}_{\sigma,F} &\overset{\text{iid}}{\sim} \mathcal{Q}_{\sigma,F} \qquad i = 1, \ldots, n \\
\mathcal{Q}_{\sigma,F} &= \frac{\mu_{\sigma,q}}{T_{\sigma,q}} \\
\mu_{\sigma,q} &\sim \mathbb{P}_{\sigma,q}, \qquad (9)
\end{aligned}$$

where $G(\cdot \mid X_i)$ denotes a continuous distribution admitting a density function $g(\cdot \mid X_i)$ with respect to a dominating measure. The marginal characterization in Proposition 8 turns out to be fundamental for deriving a marginal MCMC algorithm, the so-called Algorithm 8 with Reuse, for posterior sampling from the mixture model (9).

4.1 Algorithm 8 with Reuse

Since $\mathcal{Q}_{\sigma,F}$ is discrete a.s., the r.v.s. $\mathbf{X}$ may take on repeated values. Let Π_n the random partition of $[n]$ induced by $\mathbf{X}$, with each cluster $c \in \Pi_n$ corresponding to a unique value $\tilde{X}_c$. We will make use of Proposition 8 to marginalize out the random measure $\mu_{\sigma,q}$, leaving $\mathcal{T}$, U, Π_n and $\{\tilde{X}_c : c \in \Pi_n\}$ as the r.v.s. whose joint posterior distribution is to be simulated. Such a posterior distribution is

$$\begin{aligned}
&\Pr[\mathcal{T} \in d\tau, U \in du, \Pi_n = \pi, \{\tilde{X}_c \in dx_c : c \in \Pi_n\} \mid \mathbf{Y}] \\
&\quad \propto u^{n-1}(u+\tau)^{\sigma|\pi|-n}\sigma^{|\pi|}e^{-(u+\tau)^\sigma + \tau^\sigma}\, du\, F(d\tau) \\
&\quad \times \prod_{c\in\pi}(1-\sigma)_{|c|-1}\alpha_0(dx_c)\prod_{i\in c} G(Y_i \mid \tilde{X}_c).
\end{aligned}$$

and a Gibbs sampler can be derive from it. In particular, the conditional distributions for $\mathcal{T}$, U and $\{\tilde{X}_c : c \in \Pi_n\}$ are

$$\begin{aligned}
&\Pr[\mathcal{T} \in d\tau \mid \text{rest}] \propto (u+\tau)^{\sigma|\pi|-n}e^{-(u+\tau)^\sigma+\tau^\sigma}F(d\tau), \\
&\Pr[U \in du \mid \text{rest}] \propto u^{n-1}(u+\tau)^{\sigma|\pi|-n}e^{-(u+\tau)^\sigma}du,
\end{aligned}$$

and

$$\Pr[\tilde{X}_c \in dx_c \mid \text{rest}] \propto \alpha_0(dx_c)\prod_{i\in c} G(Y_i \mid \tilde{X}_c),$$

respectively. These conditional distributions are not in forms from which values can be easily simulated. Instead a variety of MCMC simulation techniques can be employed, including Metropolis–Hastings, slice sampling, or Hamiltonian Monte Carlo. In our simulations we used slice sampling updates to the logarithms of the auxiliary variables for numerical stability, using the Stepping Out procedure, with steps of size 2 and a maximum of 20 steps (Neal 2003). Slice sampling is a simple and efficient update which does not require gradient information and is robust against the need to specify appropriate length scales to update the variables. We used conditional Gibbs updates for the cluster variables $\tilde{X}_c$ as this is suitable for the mixture component hierarchy we considered.

Finally, we can use the Algorithm 8 with Reuse of Favaro and Teh (2013) to update the partition Π_n given the other r.v.s. The Reuse algorithm uses $C \in \mathbb{N}$ auxiliary r.v.s. $(X_1^e, \ldots, X_C^e)$ which play the role of parameters associated with empty clusters, and are independent and identically distributed according to α_0. Each update of the Reuse algorithm updates the cluster assignment of an observation, say $i \in [n]$, according to the following scheme.

(1) Remove i from the cluster it belongs to, say $c \in \Pi$.
(2) If c becomes empty as a result, pick $k \in [C]$ uniformly at random and replace X_k^e with the value of the parameter $\tilde{X}_c$ associated with cluster c, and remove c from Π_n.
(3) Assign i to the clusters with the following probabilities:

$$\Pr[\text{assign } i \text{ to cluster } c' \mid \text{rest}] \\
\propto \begin{cases} (|c'| - \sigma)G(Y_i \mid \tilde{X}_{c'}) & \text{for } c' \in \Pi_n, \\ \dfrac{\sigma(U+\mathcal{T})^\sigma}{C}G(Y_i \mid X_{c'}^e) & \text{for } c' \in [C]. \end{cases}$$

The first terms on the right hand side is proportional to the conditional probability of being assigned to the corresponding cluster (with the C empty clusters sharing the probability of creating a new cluster), while the second terms are the likelihoods associated with observation Y_i given the cluster parameters.
(4) If an empty cluster $c' \in [C]$ was chosen, then we assign i to a new cluster in Π_n with parameter $X_{c'}^e$, and replace $X_{c'}^e$ with a new independent draw from α_0.

At regular intervals, e.g., after the cluster assignments of all observations have been updated, the parameters of the empty clusters are refreshed by drawing them as independent and identically distributed according to α_0. The main steps of the Algorithm 8 with Reuse are summarized in the following Algorithm 1 and Algorithm 2. Each iteration of the algorithm takes $O(n|\Pi_n|C)$ time, where C is the unit cost of a cluster likelihood computation or of updating one cluster with respect to one observation. This computational complexity has the same scaling as existing marginal samplers for other Bayesian nonparametric mixture models.

Algorithm 1 MS($\mathcal{T}$, U, $\{\tilde{X}_c\}_{c\in\Pi_n}$, Π_n, $\{Y_i\}_{i\in[n]}$)

```
for $t = 1 \to iter$ do
    Update $\mathcal{T}$: Slice sample $\Pr[\mathcal{T} \in \mathrm{d}\tau \mid \text{rest}]$
    Update $U$ : Slice sample $\Pr[U \in \mathrm{d}u \mid \text{rest}]$
    for $c \in \Pi_n$ do
        Update $\tilde{X}_c$: Slice sample $\Pr[\tilde{X}_c \in \mathrm{d}x_c \mid \text{rest}]$
    end for
    Update $\Pi_n$: ReUse($\Pi_n$, $C$, $\{\tilde{X}_c\}_{c\in\Pi_n}$, rest)
end for
```

Algorithm 2 ReUse($\Pi_n, C, \{\tilde{X}_c\}_{c \in \Pi_n}$, rest)

```
Draw $\{X_j^e\}_{j=1}^C \overset{\text{i.i.d.}}{\sim} \alpha_0$
**for** $i = 1 \to n$ **do**
    Let $c \in \Pi_n$ be such that $i \in c$
    $c \leftarrow c \setminus \{i\}$
    **if** $c = \emptyset$ **then**
        $k \sim \text{DiscreteUniform}(\frac{1}{C})$
        $X_k^e \leftarrow \tilde{X}_c$
        $\Pi_n \leftarrow \Pi_n \setminus \{c\}$
    **end if**
    Set $c'$ according to Pr[assign $i$ to cluster $c'$ | rest]
    **if** $c' \in [C]$ **then**
        $\Pi_n \leftarrow \Pi_n \cup \{\{i\}\}$
        $\tilde{X}_{\{i\}} \leftarrow X_{c'}^e$
        $X_{c'}^e \sim \alpha_0$
    **else**
        $c' \leftarrow c' \cup \{i\}$
    **end if**
**end for**
```

In the special case of the two-parameter PD process, the generalized gamma distribution (6) leads to a significant simplification of the conditional joint distribution of $\mathcal{T}$ and U given the rest. In particular, by using the reparameterization $V = U + \mathcal{T}$ and $Z = U/V$, we obtain

$$\Pr[V \in dv, Z \in dz \mid \text{rest}] \\ \propto z^{n-1}(1-z)^{\theta-1} v^{\theta+\sigma|\pi|-1} e^{-v^{\sigma}}.$$

Hence the r.v.s. V and Z are conditionally independent and distributed according to a Beta distribution and a generalized Gamma distribution, respectively. Therefore, it is possible to marginalize out the r.v.s. V and Z, resulting in an expression for the well-known EPPF of the two-parameter PD process. See Pitman (1995) for details.

4.2 Simulation studies

In Sect. 4.1 we described a novel marginalized sampler for posterior sampling from Bayesian nonparametric mixture models reposing on prior distributions in the class $\mathcal{Q}$ of mixing measures. In particular, under the assumption of the two parameter PD process, a variant based on the reparameterization $V = U + \mathcal{T}$ and $Z = U/V$ leads to a simpler marginalized sampler. Marginalizing out V and Z finally leads to the EPPF induced by a sample from a two parameter PD process, on which we can base an even simpler marginalized sampler.

We explored the relative efficiencies of the three resulting marginalized samplers on the galaxy dataset[1]. This dataset consists of $n = 82$ velocities of galaxies and it is an obligatory exercise when working with infinite mixture models. See Roeder (1990) for details. Specifically we used a Gaussian component models parameterized by cluster-specific means and variances with a non-conjugate base distribution. We refer to Favaro and Teh (2013) for additional details. All algorithms were implemented in Java and used the same code base. We have found that the run times of all three algorithms are comparable.

Table 1 Comparison of sampler efficiencies on the galaxy dataset

Algorithm	Effective sample size	
	$\|\Pi_n\| = \|\pi\|$	$\log \tau$
Standard $\mathcal{Q}$	4,772	NA
V and Z updates	4,767	NA
Marginalized 2PPD	4,588	NA
Standard $\mathcal{Q}$	2,835	1,986
V and Z updates	2,636	3,534
Marginalized 2PPD	3,572	8,107

Each of 10 runs produces 10,000 samples, at intervals of 10 iterations, after an initial burn-in period of 10,000 iterations. First three lines are with fixed hyperparameters, while second set of three lines are with updates to hyperparameters σ and θ

We considered two scenarios. In the first scenario we fixed the hyperparameters at $\sigma = 1/3$ and $\theta = 1$, whereas in the second scenario we allowed σ and θ to be sampled as well, with prior $\sigma \sim \text{Beta}(2, 4)$ and $\theta \sim \text{Gamma}(1, 1)$. These priors are chosen to be broad and not very informative, with a preference for smaller values of σ. The fixed values are chosen as the corresponding prior means. The reported results are qualitatively not sensitive to the choices made here. Table 1 shows the effective sample sizes (ESSs) obtained by the three algorithms. For each algorithm we collected 10,000 samples with a thinning factor of 10 and an initial burn-in phase of 10,000 iterations. ESSs were computed using the R Coda package. When the hyperparameters are fixed we see that all algorithms achieved comparable and good ESSs. However when the hyperparameters are allowed to vary, the ESSs of the samplers with additional auxiliary variables are lowered. This is to be expected, since the additional auxiliary variables induce additional dependencies which slows down convergence. However, the ESSs of the data augmentation schemes are still good, while being applicable to the much larger $\mathcal{Q}$ class of priors. Although in these experiments we have found little difference between the samplers using the $\mathcal{T}, U$ and the V, Z representations, we expect the V, Z sampler to perform better on average and on more complex models.

In order to demonstrate that the algorithm running on a non-standard $\mathcal{Q}$ class prior, we applied the model with a log-normal F distribution over τ with log-scale 0 and shape 1, to the galaxy dataset, and to a dataset of vegetable oil spectral profiles[2]. There are 120 observations in the oil dataset,

[1] This dataset is included in the MASS package in the R statistical computing environment.

[2] This dataset can be obtained from the University of Copenhagen, Department of Food Science repository of public datasets for multivariate analysis: http://www.models.life.ku.dk/oliveoil.

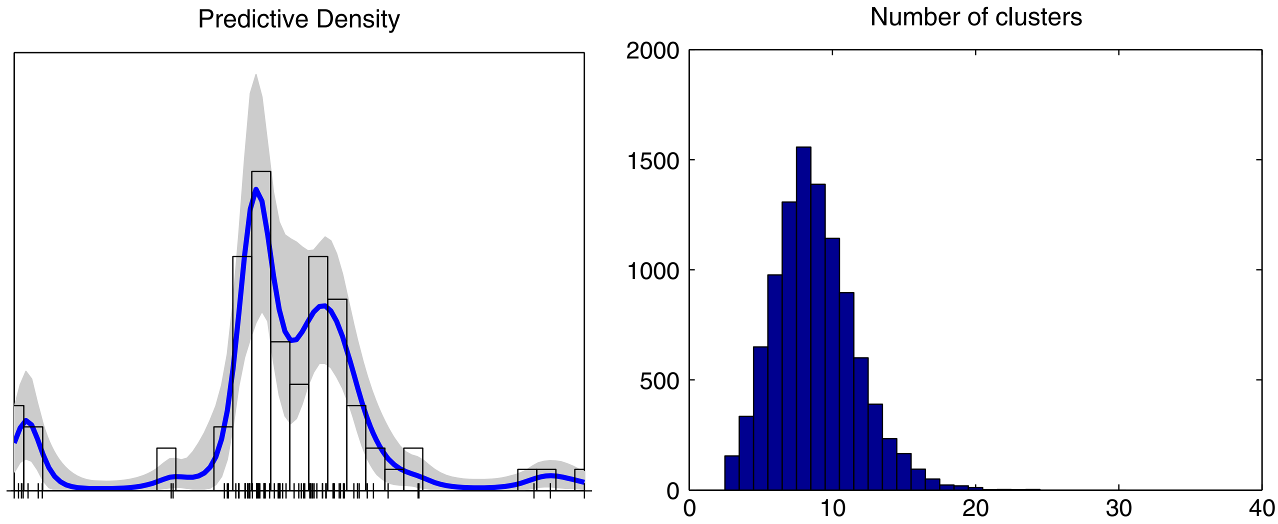

Fig. 4 Galaxy data. *Left* posterior distribution over the density. The *thick curve* is the mean density while the *shaded area* gives the 95 % credible intervals. *Right* posterior distribution over the number of clusters

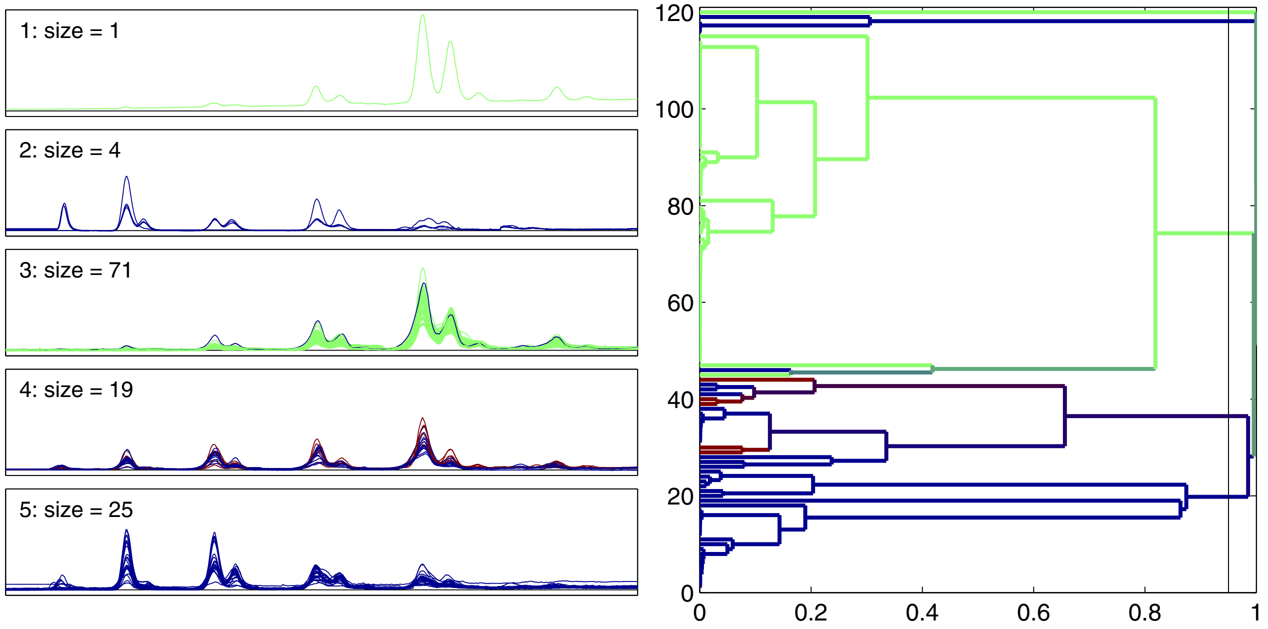

Fig. 5 Vegetable oils data. *Right* dendrogram summarizing the posterior distribution over partitions of the data. *Colours* denote the proportion of oils in each subset belonging to each class (*blue* vegetable oil, *green* olive oil, *red* blends). *Left* five subsets of oils obtained by thresholding the dendrogram at 0.95

consisting of olive oils, other vegetable oils, and blends (See De la Mata-Espinosa et al. (2011) for details). The spectral profiles are 4001 dimensional, which we pre-process using PCA to reduce the dimensionality to 6. This retained 95 % of the variance.

Figure 4 shows the posterior distribution over densities obtained by the model and over the number of clusters in the model applied to the galaxy dataset. The posterior densities are a good fit to the dataset, and are consistent with previous analysis. Note that the posterior number of clusters in the partition is relatively large compared to the clusters exhibited by the dataset. This is due to the fact that such nonparametric priors always produce a number of small clusters in posterior samples. Such clusters are spurious in nature and often can be suppressed easily using posterior summarization techniques as in our analysis of the vegetable oils dataset.

Figure 5 shows a summary of the posterior clustering structure obtained on the vegetable oils dataset. The dendrogram is obtained by a complete linkage algorithm, where the distance between two observations X_i and X_j is the posterior probability that i and j are not in the same cluster. The dataset consists three classes of oils, which are represented by three colours on the dendrogram: blue for vegetable oil, green for olive oil, and red for blends. Each segment on the dendrogram corresponds to a subset of observations, and is coloured depending on the proportion of each class of observations in the subset. We see that the model has successfully separated the olive oils from the vegetable oils, while the blended oils were not successfully separated from the vegetable oils.

We can further visualize the partitioning structure by thresholding the dendrogram at a level of 0.95. This is given by the vertical black line in the dendrogram plot. Note that there are five subsets at this level, with all pairs of observations in each subset being placed in the same cluster in Π_n with posterior probability greater than $1 - 0.95 = 0.05$. We also plotted the profiles of the observations in the five subsets in the Fig. 5, with the plots coloured depending on the class of each oil.

We see that the clustering obtained is reasonably sensible. The first subset consists of only one spectral profile, which is an outlier as it has an upward trend and may indicate an error in the processing which produced the data. The third subset consists mostly of olive oils, plus a vegetable oil with similar spectral profile as the olive oils. The second, fourth and fifth subset are mostly vegetable oils, though the spectral profiles of the three subsets are indeed quite distinct from each other, perhaps corresponding to different types of vegetable oils. The second subset has a bump on the left which is not present in other subsets. The fourth subset has a larger fifth pair of bumps, while the fifth subset has a larger second and third pair of bumps. The spectral profiles of the vegetable oils and blends in the fourth subset are reasonably similar to each other.

Acknowledgments The authors are grateful to an Associate Editor and two Referees for their valuable remarks and suggestions that have led to a substantial improvement of the paper. We would also like to thank Lancelot F. James for helpful suggestions. Stefano Favaro is supported by the European Research Council (ERC) through StG N-BNP 306406. Yee Whye Teh's research leading to these results is supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) ERC Grant Agreement No. 617411. Maria Lomeli is supported by the Gatsby Charitable Foundation.

References

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

Charalambides, C.A.: Combinatorial Methods in Discrete Distributions. Wiley-Interscience, Hoboken (2005)

De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Pruñster, I., Ruggiero, M.: Are Gibbs-type priors the most natural generalization of the Dirichlet process? IEEE Trans. Pattern Anal. Mach. Intell. (in press) (2013)

De la Mata-Espinosa, P., Bosque-Sendra, J.M., Bro, R., Cuadros-Rodriguez, L.: Discriminating olive and non-olive oils using HPLC-CAD and chemometrics. Anal Bioanal Chem **399**, 2083–2092 (2011)

Escobar, M.D.: Estimating normal means with a Dirichlet process prior. J. Am. Stat. Assoc. **89**, 268–277 (1994)

Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**, 577–588 (1995)

Favaro, S., Lijoi, A., Prünster, I.: On the stick-breaking representation of normalized inverse Gaussian priors. Biometrika **99**, 663–674 (2012)

Favaro, S., Teh, Y.W.: MCMC for normalized random measure mixture models. Stat. Sci. **28**, 335–359 (2013)

Favaro, S., Walker, S.G.: Slice sampling σ-stable Poisson–Kingman mixture models. J. Comput. Gr. Stat. **22**, 830–847 (2013)

Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. **1**, 209–230 (1973)

Gnedin, A., Pitman, J.: Exchangeable Gibbs partitions and Stirling triangles. Zap. Nauchn. Sem. S. Peterburg. Otdel. Mat. Inst. Steklov. (POMI) 325, 83–102 (2005)

Griffin, J.E., Walker, S.G.: Posterior simulation of normalized random measure mixtures. J. Comput. Gr. Stat. **20**, 241–259 (2009)

Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. **96**, 161–173 (2001)

James, L.F.: Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics (preprint) (2002). arXiv:math/0205093

James, L.F.: Coag–Frag duality for a class of stable Poisson–Kingman mixtures (preprint) (2010). arXiv:math/1008.2420

James, L.F.: Stick-breaking PG(α, ζ)-generalized Gamma processes (preprint) (2013). arXiv:math/1308.6570

James, L.F., Lijoi, A., Prünster, I.: Distributions of linear functionals of two parameter Poisson–Dirichlet random measures. Ann. Appl. Probab. **18**, 521–551 (2008)

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

Kalli, M., Griffin, J.E., Walker, S.G.: Slice sampling mixture models. Stat. Comput. **21**, 93–105 (2011)

Kingman, J.F.C.: Completely random measures. Pac. J. Math. **21**, 59–78 (1967)

Kingman, J.F.C.: Random discrete distributions. J. R. Stat. Soc. B **37**, 1–22 (1975)

Lijoi, A., Mena, R.H., Prünster, I.: Hierarchical mixture modelling with normalized inverse-Gaussian priors. J. Am. Stat. Assoc. **100**, 1278–1291 (2005)

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., Prünster, I.: Models beyond the Dirichlet process. In: Hjort, N.L., Holmes, C.C., Müller, P., Walker, S.G. (eds.) Bayesian Nonparametrics, pp. 80–136. Cambridge University Press, Cambridge (2010)

Lo, A.I.: On a class of Bayesian nonparametric estimates: I. Density estimates. Ann. Stat. **12**, 351–357 (1984)

MacEachern, S.N.: Estimating normal means with a conjugate style Dirichlet process prior. Commun. Stat. Simul. Comput. **23**, 727–741 (1994)

Muliere, P., Tardella, L.: Approximating distributions of random functionals of Ferguson–Dirichlet priors. Can. J. Stat. **26**, 283–298 (1998)

Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Gr. Stat. **9**, 249–265 (2000)

Neal, R.M.: Slice sampling. Ann. Stat. **31**, 705–767 (2003)

Nieto-Barajas, L.E., Prünster, I., Walker, S.G.: Normalized random measures driven by increasing additive processes. Ann. Stat. **32**, 2343–2360 (2004)

Papaspiliopoulos, O.: A note on posterior sampling from Dirichlet mixture models. Unpublished manuscript (2008)

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

Perman, M., Pitman, J., Yor, M.: Size-biased sampling of Poisson point processes and excursions. Probab. Theory Relat. Fields **92**, 21–39 (1992)

Pitman, J.: Exchangeable and partially exchangeable random partitions. Probab. Theory Relat. Fields **102**, 145–158 (1995)

Pitman, J.: Some developments of the Blackwell–MacQueen urn scheme. In: Ferguson, T.S., et al. (eds.) Statistics, Probability and Game Theory: Papers in Honor of David Blackwell. Lecture Notes Monograph Series, vol. 30, pp. 245–267. IMS, Beachwood (1996a)

Pitman, J., Yor, M.: The two parameter Poisson–Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**, 855–900 (1997)

Pitman, J.: Poisson–Kingman partitions. In: Goldstein, D.R. (ed.) Science and Statistics: A Festschrift for Terry Speed. Lecture Notes Monograph Series, pp. 1–34. IMS, Beachwood (2003)

Pitman, J.: Combinatorial stochastic processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875. Springer, New York (2006)

Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. Ann. Stat. **31**, 560–585 (2002)

Roeder, K.: Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. J. Am. Stat. Assoc. **36**, 45–54 (1990)

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