Submitted to the Statistical Science

Modeling with normalized random measure mixture models

Ernesto Barrios[*], **Antonio Lijoi**[†], **Luis E. Nieto-Barajas**[*] **and Igor Prünster**[†]

ITAM, University of Pavia, ITAM and University of Torino

Abstract. The Dirichlet process mixture model and more general mixtures based on discrete random probability measures have been shown to be flexible and accurate models for density estimation and clustering. The goal of this paper is to illustrate the use of normalized random measures as mixing measures in nonparametric hierarchical mixture models and point out how possible computational issues can be successfully addressed. To this end, we first provide a concise and accessible introduction to normalized random measures with independent increments. Then, we explain in detail a particular way of sampling from the posterior using the Ferguson–Klass representation. We develop a thorough comparative analysis for location–scale mixtures that considers a set of alternatives for the mixture kernel and for the nonparametric component. Simulation results indicate that normalized random measure mixtures potentially represent a valid default choice for density estimation problems. As a byproduct of this study an R package to fit these models was produced and is available in the Comprehensive R Archive Network (CRAN).

Key words and phrases: Bayesian Nonparametrics, completely random measure, clustering, density estimation, Dirichlet process, increasing additive process, latent variables, mixture model, normalized generalized gamma process, normalized inverse Gaussian process, normalized random measure, normalized stable process.

Department of Statistics, ITAM, Mexico D.F. (e-mail: ebarrios@itam.mx; lnieto@itam.mx); Department of Economics and Management, University of Pavia, Italy, (e-mail: lijoi@unipv.it); Department of Economics and Statistics, University of Torino, Italy, (e-mail: igor@econ.unito.it).

[*] Supported by the National Council for Science and Technology of Mexico (CONACYT), Grant I130991.

[†] Also affiliated to Collegio Carlo Alberto, Moncalieri, Italy. Supported by the European Research Council (ERC) through StG “N-BNP” 306406.

1. INTRODUCTION

The Dirichlet process mixture model (DPM), introduced by Lo (1984), currently represents the most popular Bayesian nonparametric model. It is defined as

$$\tilde{f}(x) = \int k(x|\theta)\tilde{P}(\mathrm{d}\theta), \tag{1}$$

where k is a parametric kernel and $\tilde{P}$ is a random probability whose distribution is the Dirichlet process prior with (finite) parameter measure α, in symbols $\tilde{P} \sim \mathscr{D}_\alpha$. It is often useful to write $\alpha = aP_0$ where $P_0 = \mathrm{E}[\tilde{P}]$ is a probability measure and a is in $(0, +\infty)$. In other words, the DPM is a mixture of a kernel k with mixing distribution a Dirichlet process. See also Berry and Christensen (1979) for an early contribution to DPM.

Alternatively, the DPM can also be formulated as a hierarchical model (Ferguson, 1983). In this case, X_i, θ_i for $i = 1, \ldots, n$

$$\begin{aligned} X_i\,|\theta_i &\overset{\text{ind}}{\sim} k(\,\cdot\,|\,\theta_i), \\ \theta_i|\tilde{P} &\overset{\text{iid}}{\sim} \tilde{P}, \\ \tilde{P} &\sim \mathscr{D}_\alpha. \end{aligned} \tag{2}$$

The hierarchical representation of the DPM explicitly displays features of the model that are relevant for practical purposes. Indeed, Escobar and West (1995) developed an MCMC algorithm for simulating from the posterior distribution. This contribution paved the way for extensive uses of the DPM, and semiparametric variations of it, in many different applied contexts. See MacEachern and Müller (2000) and Müller and Quintana (2004) for reviews of recent progress, both computational and applied. The main idea behind Escobar and West's algorithm is represented by the marginalization of the infinite dimensional random component, namely the Dirichlet process $\tilde{P}$, which leads to work with generalized Pólya urn schemes. If the centering measure P_0 is further chosen to be the conjugate prior for kernel k, then one can devise a Gibbs sampler whose implementation is straightforward. In particular, the typical setup in applications involves a normal kernel: if the location (or location–scale) mixture of normals is combined with a conjugate normal (or normal–gamma) probability measure P_0, the full conditional distributions can be determined thus leading to a simple Gibbs sampler.

Given the importance of the DPM model, much attention has been devoted to the development of alternative and more efficient algorithms. According to the terminology of Papaspiliopoulos and Roberts (2008) these can be divided into two classes: marginal and conditional methods. Marginal methods, such as the Escobar and West algorithm, integrate out the Dirichlet process in (2) and resort to the predictive distributions, within a Gibbs sampler, to obtain posterior samples. In this framework an important advance is due to MacEachern and Müller (1998): they solve the issue of providing algorithms, which effectively tackle the case where the kernel k and P_0 are not a conjugate pair. On the other hand, conditional methods work directly on (2) and clearly have to face the problem of sampling the trajectories of an infinite–dimensional random element such as the

Dirichlet process. The first contributions along this line are given in Muliere and Tardella (1998) and Ishwaran and James (2001) who use truncation arguments. Exact simulations can be achieved by the retrospective sampling technique introduced in Papaspiliopoulos and Roberts (2008) and slice sampling schemes as in Walker (2007).

In this paper we focus on mixture models more general than the DPM, namely mixtures with mixing measure given by normalized random measures with independent increments (NRMI), namely a class of random probability measures introduced in Regazzini, Lijoi and Prünster (2003). Several applications of specific members of this class, or closely related distributions, are now present in the literature and deal with species sampling problems, mixture models, clustering, reliability and models for dependence. See Lijoi and Prünster (2010) for references. Here we describe in detail a conditional algorithm which allows one to draw posterior simulations from mixtures based on a general NRMI. As we shall point out, it works equally well regardless of k and P_0 forming a conjugate pair or not and readily yields credible intervals. Our description is a straightforward implementation of the posterior characterization of NRMI provided in James, Lijoi and Prünster (2009) combined with the representation of an increasing additive process given in Ferguson and Klass (1972). The R package `BNPdensity`, available in the Comprehensive R Archive Network (CRAN) implements this algorithm. For contributions containing thorough and insightful comparisons of algorithms for Bayesian nonparametric mixture models, both marginal and conditional, the reader is referred to Papaspiliopoulos and Roberts (2008) and Favaro and Teh (2012).

The `BNPdensity` package is used to carry out a comparative study that involves a variety of datasets both real and simulated. For the real datasets we show the impact of choosing different kernels and compare the performance of location-scale nonparametric mixtures. We also examine different mixing measures and show some advantages and disadvantages fitting the data and the number of induced clusters. Model performance is assessed by referring to conditional predictive ordinates and to suitable numerical summaries of these values. For the simulated examples, we rely on the relative mean integrated squared error to measure the performance of NRMI mixtures with respect to competing methods such as kernel density estimators, Bayesian wavelets and finite mixtures of normals. The outcome clearly shows that NRMI mixtures, and in particular mixtures of stable NRMIs, potentially represent a valid default choice for density estimation problems.

The outline of the paper is as follows. We provide in Section 2 an informal review of normalized random measures and highlight their uses for Bayesian nonparametric inference. Particular emphasis is given to the posterior representation since it plays a key role in the elaboration of the sampling scheme that we use; in Section 3 a conditional algorithm for simulating from the posterior of NRMI mixtures is described in great detail; Section 4 contains a comprehensive data analysis highlighting the potential of NRMI mixtures.

2. DIRICHLET PROCESS AND NRMIS

A deeper understanding of NRMI mixture models defined in (1) is eased by an accessible introduction to the notions of completely random measures and

NRMIs. This section aims at providing a concise review of the most relevant distributional properties of completely random measures and NRMIs in view of their application to Bayesian inference. These are also important for addressing the computational issues we shall focus on in later sections.

2.1 Exchangeability and discrete nonparametric priors

In order to best describe the nonparametric priors we are going to deal with, we first recall the notion of exchangeability, its implication in terms of Bayesian inference and some useful notation. Let $(Y_n)_{n\geq 1}$ be an (ideally) infinite sequence of observations, defined on some probability space $(\Omega, \mathscr{F}, P)$, with each Y_i taking values in $\mathbb{Y}$ (a complete and separable metric space endowed with its Borel σ–algebra). While in a frequentist setting one typically assumes that the Y_i's are independent and identically distributed (iid) with some fixed and unknown distribution, in a Bayesian approach the independence is typically replaced by a weaker assumption of conditional independence, given a random probability distribution on $\mathbb{Y}$, which corresponds to assuming *exchangeable* data. Formally, this corresponds to an invariance condition according to which, for any $n \geq 1$ and any permutation π of the indices $1, \ldots, n$, the probability distribution of $(Y_1, \ldots, Y_n)$ coincides with the distribution of $(Y_{\pi(1)}, \ldots, Y_{\pi(n)})$. Then, the celebrated de Finetti representation theorem states that the sequence $(Y_n)_{n\geq 1}$ is exchangeable if and only if its distribution can be represented as a mixture of sequences of iid random variables. In other terms, $(Y_n)_{n\geq 1}$ is exchangeable if and only if there exists a probability distribution Q on the space of probability measures on $\mathbb{Y}$, say $\mathscr{P}_{\mathbb{Y}}$, such that

$$
\begin{aligned}
Y_i \,|\, \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P}, \qquad i = 1, \ldots, n \\
\tilde{P} &\sim Q,
\end{aligned}
\tag{3}
$$

for any $n \geq 1$. Hence, $\tilde{P}$ is a random probability measure on $\mathbb{Y}$, namely a random element on $(\Omega, \mathscr{F}, P)$ taking values in $\mathscr{P}_{\mathbb{Y}}$ (endowed with the topology of weak convergence). The probability distribution Q of $\tilde{P}$ is also termed *de Finetti measure* and represents the prior distribution in a Bayesian setup. Whenever Q degenerates on a finite dimensional subspace of $\mathscr{P}_{\mathbb{Y}}$, the inferential problem is usually called *parametric*. On the other hand, when the support of Q is infinite–dimensional then this is typically referred to as a *nonparametric* inferential problem. It is generally agreed that having a large topological support is a desirable property for a nonparametric prior (see, e.g., Ferguson, 1974).

In the context of nonparametric mixture models, which identify the main focus of the paper, a key role is played by *discrete nonparametric priors* Q, i.e. priors which select discrete distributions with probability 1. Clearly, any random probability measure $\tilde{P}$ associated to a discrete prior Q can be represented as

$$
\tilde{P} = \sum_{j\geq 1} \tilde{p}_j \, \delta_{Z_j}, \tag{4}
$$

where $(\tilde{p}_j)_{j\geq 1}$ is a sequence of non–negative random variables such that $\sum_{j\geq 1} \tilde{p}_j = 1$, almost surely, $(Z_j)_{j\geq 1}$ is a sequence of random variables taking values in $\mathbb{Y}$ and δ_Z is the Dirac measure.

As far as the observables Y_i's are concerned, the discrete nature of $\tilde{P}$ in (4) implies that any sample $Y_1, \dots, Y_n$ in (3) will feature ties with positive probability and, therefore, display $r \le n$ distinct observations $Y_1^*, \dots, Y_r^*$ with respective frequencies $n_1, \dots, n_r$ such that $\sum_{i=1}^r n_i = n$. Such a grouping implied by the discrete nature of the nonparametric prior lies at the heart of Bayesian nonparametric procedures for clustering purposes. Henceforth, R_n will denote the random variable identifying the number of distinct values appearing in the sample $Y_1, \dots, Y_n$.

The simplest and most familiar illustration one can think of is the Dirichlet process prior introduced by Ferguson (1973), which represents the cornerstone of Bayesian Nonparametrics. Its original definition was given in terms of a consistent family of finite–dimensional distributions that coincide with multivariate Dirichlet distributions. To make this explicit, introduce the $(d-1)$–variate Dirichlet probability density function on the $(d-1)$–dimensional unit simplex

$$h(\boldsymbol{p}; \boldsymbol{c}) = \frac{\Gamma(\sum_{i=1}^d c_i)}{\prod_{i=1}^d \Gamma(c_i)}\, p_1^{c_1-1} \cdots p_{d-1}^{c_{d-1}-1} (1 - \sum_{i=1}^{d-1} p_i)^{c_d - 1},$$

where $\boldsymbol{c} = (c_1, \dots, c_d) \in (0, \infty)^d$ and $\boldsymbol{p} = (p_1, \dots, p_{d-1})$.

Definition 1. [Ferguson, 1973]. *Let α be some finite and non–null measure on $\mathbb{Y}$ such that $\alpha(\mathbb{Y}) = a$. Suppose the random probability measure $\tilde{P}$ has distribution Q such that, for any choice of a (measurable) partition $\{A_1, \dots, A_d\}$ of $\mathbb{Y}$ and for any $d \ge 1$, one has*

$$Q(\{P : (P(A_1), \dots, P(A_d)) \in B\}) = \int_B h(\boldsymbol{p}; \boldsymbol{\alpha})\, \mathrm{d}p_1 \dots \mathrm{d}p_{d-1}, \tag{5}$$

where $\boldsymbol{\alpha} = (\alpha(A_1), \dots, \alpha(A_d))$. Then $\tilde{P}$ is termed a Dirichlet process *with base measure α.*

Note that $\alpha/a =: P_0$ defines a probability measure on $\mathbb{Y}$ and it coincides with the expected value of a Dirichlet process, i.e. $P_0 = \mathrm{E}[\tilde{P}]$, for this reason is often referred to as the prior guess at the shape of $\tilde{P}$. Henceforth we shall denote more conveniently the base measure α of $\tilde{P}$ as $a\,P_0$. Also note that the Dirichlet process has large support and it, thus, shares one of the properties that make the use of nonparametric priors attractive. Indeed, if the support of P_0 coincides with $\mathbb{Y}$, then the support of the Dirichlet process prior (in the weak convergence topology) coincides with the whole space $\mathscr{P}_{\mathbb{Y}}$. In other words, the Dirichlet process prior assigns positive probability to any (weak) neighborhood of any given probability measure in $\mathscr{P}_{\mathbb{Y}}$ thus making it a flexible model for Bayesian nonparametric inference.

As shown in Blackwell (1973), the Dirichlet process selects discrete distributions on $\mathbb{Y}$ with probability 1 and, hence, admits a representation of the form (4). An explicit construction of the $\tilde{p}_j$'s in (4) leading to the Dirichlet process has been provided by Sethuraman (1994) who relied on a stick–breaking procedure. This arises when the sequence of random probability masses $(\tilde{p}_j)_{j \ge 1}$ is defined as

$$\tilde{p}_1 = V_1, \qquad \tilde{p}_j = V_j \prod_{i=1}^{j-1} (1 - V_i) \quad j = 2, 3, \dots\ , \tag{6}$$

with the V_i's being iid and beta distributed with parameter $(1, a)$, and when the locations $(Z_j)_{j\geq 1}$ are iid from P_0. Under these assumptions (4) yields a random probability measure that coincides, in distribution, with a Dirichlet process with base measure $a\,P_0$.

A nice and well–known feature about the Dirichlet process is its conjugacy. Indeed, if $\tilde{P}$ in (3) is a Dirichlet process with base measure $a\,P_0$, then the posterior distribution of $\tilde{P}$, given the data $Y_1, \ldots, Y_n$, still coincides with the law of a Dirichlet process with parameter measure $(a+n)\,P_n$ where $P_n = a\,P_0/(a+n) + \sum_{i=1}^n \delta_{Y_i}/(a+n)$, where δ_y denotes a point mass at $y \in \mathbb{Y}$. On the basis of this result, one easily determines the predictive distributions associated to the Dirichlet process and for any A in $\mathbb{Y}$, one has

$$P\,[Y_{n+1} \in A \,|\, Y_1, \ldots, Y_n] = \frac{a}{a+n}\,P_0(A) + \frac{n}{a+n}\,\frac{1}{n}\sum_{j=1}^{r} n_j \delta_{Y_j^*}(A), \tag{7}$$

where, again, the Y_j^*'s with frequency n_j denote the $r \leq n$ distinct observations within the sample. Hence, the predictive distribution appears as a convex linear combination of the prior guess at the shape of $\tilde{P}$ and of the empirical distribution.

From (4) it is apparent that a decisive issue when defining a discrete nonparametric prior is the determination of the probability masses $\tilde{p}_j$'s, while at the same time preserving a certain degree of mathematical tractability. This is in general quite a challenging task. For instance, the stick–breaking procedure is useful to construct a wide range of discrete nonparametric priors as shown in Ishwaran and James (2001). However, only for a few of them is it possible to establish relevant distributional properties such as, e.g., the posterior or predictive structures. See Favaro, Lijoi and Prünster (2012) for a discussion on this issue. Also, as extensively discussed in Lijoi and Prünster (2010), a key tool for defining tractable discrete nonparametric priors (4) is given by completely random measures, a concept introduced in Kingman (1967). Since it is essential for the construction of the class of NRMIs considered in the paper, in the following section we concisely recall the basics and refer the interested reader to Kingman (1993) for an exhaustive account.

2.2 CRM and NRMI

Denote first by $\mathscr{M}_{\mathbb{Y}}$ the space of boundedly finite measures on $\mathbb{Y}$, this meaning that for any μ in $\mathscr{M}_{\mathbb{Y}}$ and any bounded set A in $\mathbb{Y}$ one has $\mu(A) < \infty$. Moreover, $\mathscr{M}_{\mathbb{Y}}$ can be endowed with a suitable topology that allows one to define the associated Borel σ–algebra. See Daley and Vere-Jones (2008) for technical details.

Definition 2. *A random element $\tilde{\mu}$, defined on $(\Omega, \mathscr{F}, P)$ and taking values in $\mathscr{M}_{\mathbb{Y}}$ is called* completely random measure *(CRM) if, for any $A_1, \ldots, A_n$ in $\mathbb{Y}$, with $A_i \cap A_j = \varnothing$ for any $i \neq j$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent.*

Hence, a CRM is simply a random measure, which gives rise to independent random variables when evaluated over disjoint sets. In addition, it is well–known that if $\tilde{\mu}$ is a CRM on $\mathbb{Y}$ then

$$\tilde{\mu} = \sum_{i\geq 1} J_i\,\delta_{Z_i} + \sum_{i=1}^{M} V_i\,\delta_{z_i},$$

where $(J_i)_{i\geq 1}$, $(V_i)_{i\geq 1}$ and $(Z_i)_{i\geq 1}$ are independent sequences of random variables and the jump points $\{z_1,\dots,z_M\}$ are fixed, with $M\in\{0,1,\dots\}\cup\{\infty\}$. If $M=0$, then $\tilde\mu$ has no fixed jumps and the Laplace transform of $\tilde\mu(A)$, for any A in $\mathbb{Y}$, admits the following representation

$$\mathrm{E}\left[\mathrm{e}^{-\lambda\tilde\mu(A)}\right]=\exp\left\{-\int_{\mathbb{R}^+\times A}\left[1-\mathrm{e}^{-\lambda v}\right]\nu(\mathrm{d}v,\mathrm{d}y)\right\}, \tag{8}$$

for any $\lambda>0$, with ν being a measure on $\mathbb{R}^+\times\mathbb{Y}$ such that

$$\int_B\int_{\mathbb{R}^+}\min\{v,1\}\,\nu(\mathrm{d}v,\mathrm{d}y)<\infty, \tag{9}$$

for any bounded B in $\mathbb{Y}$. The measure ν is referred to as the *Lévy intensity* of $\tilde\mu$ and, by virtue of (8), it characterizes the CRM $\tilde\mu$. This is extremely useful from an operational point of view since a single measure encodes all the information about the distribution of the jumps $(J_i)_{i\geq 1}$ and locations $(Z_i)_{i\geq 1}$ of $\tilde\mu$. The measure ν will be conveniently rewritten as

$$\nu(\mathrm{d}v,\mathrm{d}y)=\rho(\mathrm{d}v|y)\,\alpha(\mathrm{d}y), \tag{10}$$

where ρ is a transition kernel on $\mathbb{R}^+\times\mathbb{Y}$ controlling the jump intensity and α is a measure on $\mathbb{Y}$ determining the locations of the jumps. Two popular examples are *gamma* and *stable* processes. The former corresponds to the specification $\rho(\mathrm{d}v|y)=\mathrm{e}^{-v}\,\mathrm{d}v/v$ whereas the latter arises when $\rho(\mathrm{d}v|y)=\sigma\,v^{-1-\sigma}\,\mathrm{d}v/\Gamma(1-\sigma)$, for some $\sigma\in(0,1)$. Note that if $\tilde\mu$ is a gamma CRM, then, for any A, $\tilde\mu(A)$ is gamma distributed with shape parameter $\alpha(A)$ and scale 1. On the other hand, if $\tilde\mu$ is a stable CRM, then $\tilde\mu(A)$ has a positive stable distribution.

Since $\tilde\mu$ is a discrete random measure almost surely, one can then easily guess that discrete random probability measures (4) can be obtained by suitably transforming a CRM. The most obvious transformation is "normalization", which yields NRMIs. As a preliminary remark, it should be noted that "normalization" is possible when the denominator $\tilde\mu(\mathbb{Y})$ is positive and finite (almost surely). Such a requirement can be expressed in terms of the Lévy intensity, in particular, α being a finite measure and $\int_{\mathbb{R}^+}\rho(\mathrm{d}v|y)=\infty$ for any $y\in\mathbb{Y}$ are simple sufficient conditions for the normalization to be well defined. The latter condition essentially requires the CRM to jump infinitely often on any finite set and is sometimes referred to as *infinite activity*. See Regazzini, Lijoi and Prünster (2003) and James, Lijoi and Prünster (2009) for necessary and sufficient conditions. One can now provide the definition of a NRMI.

DEFINITION 3. *Let $\tilde\mu$ be a CRM with Lévy intensity (10) such that $0<\tilde\mu(\mathbb{Y})<\infty$ almost surely. Then, the random probability measure*

$$\tilde P=\frac{\tilde\mu}{\tilde\mu(\mathbb{Y})}, \tag{11}$$

is named normalized random measure with independent increments (NRMI).

It is apparent that a NRMI is uniquely identified by the Lévy intensity ν of the underlying CRM. If $\rho(\mathrm{d}v|y)$ in (10) does not depend on y, which means that the distribution of the jumps of $\tilde{\mu}$ are independent of their locations, then the CRM $\tilde{\mu}$ and the corresponding NRMI (11) are called *homogeneous*. Otherwise they are termed *non–homogeneous*. Moreover, it is worth pointing out that all NRMI priors share a support property analogous to the one recalled for the Dirichlet process prior. Specifically, if the support of the base measure coincides with $\mathbb{Y}$, then the corresponding NRMI has full weak support $\mathscr{P}_{\mathbb{Y}}$.

Note that the Dirichlet process can be defined as an NRMI: indeed it coincides, in distribution, with a normalized gamma CRM as shown in Ferguson (1973). If $\nu(\mathrm{d}v, \mathrm{d}y) = \mathrm{e}^{-v}\, v^{-1}\, \mathrm{d}v\, a\, P_0(\mathrm{d}y)$, then (11) yields a Dirichlet process with base measure $a\, P_0$. Another early use of (11) can be found in Kingman (1975), where the NRMI obtained by normalizing a stable CRM is introduced. The resulting random probability measure will be denoted as N–stable.

In the sequel particular attention will be devoted to generalized gamma NRMIs (Lijoi, Mena and Prünster, 2007) since they are analytically tractable and include many well–known priors as special cases. This class of NRMIs is obtained by normalizing generalized gamma CRMs that were introduced in Brix (1999) and are characterized by a Lévy intensity of the form

$$\rho(\mathrm{d}v)\alpha(\mathrm{d}x) = \frac{\mathrm{e}^{-\kappa v}}{\Gamma(1-\gamma)\, v^{1+\gamma}} \mathrm{d}v\, aP_0(\mathrm{d}x), \tag{12}$$

whose parameters $\kappa \geq 0$ and $\gamma \in [0, 1)$ are such that at least one of them is strictly positive and with base measure $\alpha = aP_0$, where $a \in (0, \infty)$ and P_0 is a probability distribution on $\mathbb{Y}$. The corresponding generalized gamma NRMI will be denoted as $\tilde{P} \sim \mathrm{NGG}(a, \kappa, \gamma; P_0)$. Within this class of priors one finds the following special cases: (i) the Dirichlet process which is a $\mathrm{NGG}(a, 1, 0; P_0)$ process; (ii) the normalized inverse Gaussian (N–IG) process (Lijoi, Mena and Prünster, 2005), which corresponds to a $\mathrm{NGG}(1, \kappa, 1/2; P_0)$ process; (iii) the N–stable process (Kingman, 1975) which arises as $\mathrm{NGG}(1, 0, \gamma; P_0)$. As a side remark, we observe that the parameterization $\mathrm{NGG}(a, \kappa, \gamma; P_0)$ is redundant in the sense that either κ or a can be fixed according to one's convenience. Loosely speaking, this is due to the fact that the normalization operation implies the loss of "one degree of freedom" as a reference to the Dirichlet process might clarify. For example, we mentioned that the Dirichlet case arises when κ is set equal to 1, but this choice is only due to convenience. Indeed, a Dirichlet process is obtained, as long as $\gamma = 0$, whatever the value κ takes on. See Pitman (2003) and Lijoi, Mena and Prünster (2007) for detailed explanations. For our purposes it is worth sticking to the redundant parameterization since it allows us to recover immediately all three specific cases listed above, which would be cumbersome with the alternative parameterization usually adopted, i.e. $\mathrm{NGG}(1, \beta, \gamma; P_0)$ with $\beta = \kappa^\gamma/\gamma$. The role of these parameters is best understood by looking at the induced (prior) distribution of the number of distinct values R_n in an sample $Y_1, \ldots, Y_n$. Indeed, one has that κ (or, equivalently, a) affects the location: a larger κ (or a) shifts the distribution of R_n to the right implying a larger expected number of distinct values. In contrast, γ allows to tune the flatness of the distribution of R_n: the bigger γ, the flatter is the distribution of R_n so that a large value of γ corresponds to a less informative prior for the number of distinct values in $Y_1, \ldots, Y_n$. This also explains why the Dirichlet process, which corresponds to $\gamma = 0$, yields the most highly–peaked distribution for R_n. See also Lijoi, Mena and Prünster (2007) for a graphical display of these behaviors.

Also variations of NRMI have already appeared in the literature. In Nieto-Barajas, Prünster and Walker (2004) weighted versions of NRMI are considered. To be more specific, letting h be some non–negative function defined on $\mathbb{Y}$, a normalized weighted CRM is obtained, for any B in $\mathbb{Y}$, as

$$\tilde{P}(B) = \frac{\int_B h(y)\tilde{\mu}(\mathrm{d}y)}{\int_{\mathbb{Y}} h(y)\tilde{\mu}(\mathrm{d}y)}.$$

The function h can be seen as a perturbation of the CRM and in Nieto-Barajas and Prünster (2009) the sensitivity of posterior inference with respect to (w.r.t.) h is examined. Another related class is represented by Poisson–Kingman models Pitman (2003), where one essentially conditions on $\tilde{\mu}(\mathbb{Y})$ and then mixes with respect to some probability measure on $\mathrm{I\!R}^+$.

Remark 1. If $\mathbb{Y} = \mathrm{I\!R}^m$, one can also consider the càdlàg random distribution function induced by $\tilde{\mu}$, namely $\tilde{M} := \{\tilde{M}(s) = \tilde{\mu}((-\infty, s_1] \times \ldots \times (-\infty, s_m]) : s = (s_1, \ldots, s_m) \in \mathrm{I\!R}^m\}$, known in the literature as *increasing additive process* or *independent increment process*. See Sato (1990) for details. One can then associate to the NRMI random probability measure in (11) the corresponding NRMI random cumulative distribution function

$$\tilde{F}(s) = \frac{\tilde{M}(s)}{T} \qquad \text{for any } s \in \mathrm{I\!R}^m, \tag{13}$$

where $T := \lim_{s\to\infty} \tilde{M}(s)$ and the limit is meant as componentwise. The original definition of NRMI in Regazzini, Lijoi and Prünster (2003) was given in terms of increasing additive processes. The definition on more abstract spaces adopted here, and used also, e.g. in James, Lijoi and Prünster (2009), allows us to by–pass some tedious technicalities. Nonetheless, we preserve the term NRMI, although on abstract spaces one should refer to normalized CRM rather than to “increments”. □

Remark 2. Although the previous examples deal with homogeneous CRMs and NRMIs, also non–homogeneous CRM are very useful for the construction of nonparametric priors. This is apparent in contributions to Bayesian nonparametric inference for survival analysis. See Lijoi and Prünster (2010). Hence, given the importance of non–homogeneous structures in some other contexts it seems worth including these in our treatment. □

2.3 Posterior distribution of a NRMI

The posterior distribution associated to an exchangeable model as in (3) is a preliminary step for attaining Bayesian inferential results of interest and, therefore, represents an object of primary importance. In the case of NRMI, the determination of the posterior distribution is a challenging task since one cannot rely directly on Bayes’ theorem (the model is not dominated) and, with the exception of the Dirichlet process, NRMIs are not conjugate as shown in James, Lijoi and Prünster (2006). Nonetheless, a posterior characterization has been established in James, Lijoi and Prünster (2009) and it turns out that, even though NRMIs are not conjugate, they still enjoy a sort of “conditional conjugacy”. This means that, conditionally on a suitable latent random variable, the posterior distribution of a NRMI coincides with the distribution of a NRMI having fixed points of discontinuity located at the observations. Such a simple structure suggests that when working with a general NRMI, instead of the Dirichlet process, one faces only one additional layer of difficulty represented by the marginalization with respect to the conditioning latent variable.

Before stating the main result we recall that, due to the discreteness of NRMIs, ties will appear with positive probability in $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ and therefore the sample information can be encoded by the $R_n = r$ distinct observations $(Y_1^*, \dots, Y_r^*)$ with frequencies $(n_1, \dots, n_r)$ such that $\sum_{j=1}^r n_j = n$. Moreover, introduce the non–negative random variable U such that the distribution of $[U|\boldsymbol{Y}]$ has density, w.r.t. the Lebesgue measure, given by

$$f_{U|\boldsymbol{Y}}(u) \propto u^{n-1} \exp\{-\psi(u)\} \prod_{j=1}^r \tau_{n_j}(u|Y_j^*), \tag{14}$$

where $\tau_{n_j}(u|Y_j^*) = \int_0^\infty v^{n_j} e^{-uv} \rho(\mathrm{d}v|Y_j^*)$ and ψ is the Laplace exponent of $\tilde{\mu}$ as in (8). Finally, in the following we assume the probability measure P_0 defining the base measure of a NRMI to be non–atomic.

Theorem 1 (James, Lijoi and Prünster (2009)). *Let $(Y_n)_{n\geq 1}$ be as in (3) where $\tilde{P}$ is a NRMI defined in (11) with Lévy intensity as in (10). Then the posterior distribution of the unnormalized CRM $\tilde{\mu}$, given a sample $\boldsymbol{Y}$, is a mixture of the distribution of $[\tilde{\mu}|U, \boldsymbol{Y}]$ with respect to the distribution of $[U|\boldsymbol{Y}]$. The latter is identified by (14), whereas $[\tilde{\mu}|U, \boldsymbol{Y}]$ is equal in distribution to a CRM with fixed points of discontinuity at the distinct observations Y_j^**

$$\tilde{\mu}^* + \sum_{j=1}^r J_j^* \delta_{Y_j^*} \tag{15}$$

such that

a) *$\tilde{\mu}^*$ is a CRM characterized by the Lévy intensity*

$$\nu^*(dv, dy) = \mathrm{e}^{-uv} \rho(dv|y) \alpha(dy). \tag{16}$$

b) *the jump height J_j^* corresponding to Y_j^* has density, w.r.t. the Lebesgue measure, given by*

$$f_j^*(v) \propto v^{n_j} e^{-uv} \rho(dv|Y_j^*). \tag{17}$$

c) *$\tilde{\mu}^*$ and J_j^*, $j = 1, \dots, r$ are independent.*

Moreover, the posterior distribution of the NRMI $\tilde{P}$, conditional on U, is given by

$$[\tilde{P} \,|\, U, \boldsymbol{Y}] \overset{d}{=} w \frac{\tilde{\mu}^*}{\tilde{\mu}^*(\mathbb{X})} + (1 - w) \frac{\sum_{i=1}^r J_i^* \delta_{Y_i^*}}{\sum_{l=1}^r J_l^*}, \tag{18}$$

where $w = \tilde{\mu}^*(\mathbb{X}) / (\tilde{\mu}^*(\mathbb{X}) + \sum_{l=1}^r J_l^*)$.

In order to simplify the notation, in the statement we have omitted explicit reference to the dependence on $[U|\boldsymbol{Y}]$ of both $\tilde{\mu}^*$ and $\{J_i^* : i = 1, \ldots, r\}$. However, such a dependence is apparent from (16) and (17). From Theorem 1 it is apparent that the only quantity needed for deriving explicit expressions for particular cases of NRMI is the Lévy intensity (10). For instance, in the case of normalized generalized gamma NRMI, $\mathrm{NGG}(a, \kappa, \gamma; P_0)$ one has that the un–normalized posterior CRM $\tilde{\mu}^*$ in (15) is characterized by a Lévy intensity of the form

$$(19) \qquad \nu^*(\mathrm{d}v, \mathrm{d}y) = \frac{\mathrm{e}^{-(\kappa+u)v}}{\Gamma(1-\gamma)\, v^{1+\gamma}} \mathrm{d}v\, a P_0(\mathrm{d}y),$$

Moreover, the distribution of the jumps (17) corresponding to the fixed points of discontinuity Y_i^*'s in (15) reduce to a gamma distribution with density

$$(20) \qquad f_j^*(v) = \frac{(\kappa+u)^{n_j-\gamma}}{\Gamma(n_j-\gamma)}\, v^{n_j-\gamma-1}\, \mathrm{e}^{-(\kappa+u)v}$$

Finally, the conditional distribution of the latent variable U given $\boldsymbol{Y}$ (14) is given by

$$(21) \qquad f_{U|\boldsymbol{Y}}(u) \propto u^{n-1}(u+\kappa)^{r\gamma-n} \exp\left\{-\frac{a}{\gamma}(u+\kappa)^{\gamma}\right\},$$

for $u > 0$. The availability of this posterior characterization makes it then possible to determine several important quantities such as the predictive distributions and the induced partition distribution. See James, Lijoi and Prünster (2009) for general NRMI and Lijoi, Mena and Prünster (2007) for the subclass of generalized gamma NRMI.

2.4 NRMI mixture models

Discrete nonparametric priors are particularly effective when used for modelling latent variables within hierarchical mixtures. The most popular of these models is the DPM due to Lo (1984) and displayed in (2). Its most natural generalization corresponds to allowing any NRMI to act as a nonparametric mixing measure. In view of the result on the posterior characterization of NRMIs, such a program is also feasible from a practical perspective.

We start by describing the NRMI mixture model in some detail. First, let us introduce a change in the notation. In order to highlight that the law of a NRMI acts as the de Finetti measure at a latent level, we denote the elements of the exchangeable sequence by θ_i instead of Y_i, for $i = 1, 2, \ldots$. Then, consider a NRMI $\tilde{P}$ and convolute it with a suitable density kernel $k(\,\cdot\,|\theta)$ thus obtaining the random mixture density $\tilde{f}(x) = \int_{\mathbb{Y}} k(x|\theta)\tilde{P}(\mathrm{d}\theta)$. This can equivalently be written in a hierarchical form as

$$(22) \qquad \begin{array}{rcll} X_i\,|\theta_i & \overset{\text{ind}}{\sim} & k(\,\cdot\,|\theta_i), & i = 1, \ldots, n \\ \theta_i|\tilde{P} & \overset{\text{iid}}{\sim} & \tilde{P}, & i = 1, \ldots, n \\ \tilde{P} & \sim & \text{NRMI.} & \end{array}$$

In the sequel, we take kernels defined on $\mathbb{X} \subseteq \mathbb{R}$ and NRMIs defined on $\mathbb{Y} = \Theta \subseteq \mathbb{R}^m$. Consequently, instead of describing the results in terms of the random measures $\tilde{\mu}$ and $\tilde{P}$, we will work with corresponding distribution functions, $\tilde{M}$ and $\tilde{F}$, respectively, for the sake of simplicity in the presentation (see Remark 1). It is worth noting that the derivations presented here carry over to general spaces in a straightforward way.

As for the base measure of the NRMI P_0 on Θ, we denote its density (w.r.t. the Lebesgue measure) by f_0. When P_0 depends on a further hyperparameter ϕ, we will use the symbol $f_0(\,\cdot\,|\phi)$. The case $m = 2$ typically corresponds to the specification of a nonparametric model for the location and scale parameters of the mixture, i.e. $\theta = (\mu, \sigma)$. This will be used to illustrate the algorithm in Section 4, where we apply our proposed modeling to simulated and real datasets. In order to distinguish the hyperparameters for location and scale we will use the notation $f_0(\mu, \sigma|\phi) = f_0^1(\mu|\sigma, \varphi)\, f_0^2(\sigma|\varsigma)$. In applications a priori independence between μ and σ is commonly assumed.

The most popular uses of mixtures of discrete random probability measures, as the one displayed in (22), relate to density estimation and data clustering. The former can be addressed by evaluating

$$\hat{f}_n(x) = \mathrm{E}(\tilde{f}(x) \,|\, X_1, \ldots, X_n), \tag{23}$$

for any x in $\mathbb{X}$. As for the latter, if R_n is the number of distinct latent values $\theta_1^*, \ldots, \theta_{R_n}^*$ out of a sample of size n, one can deduce a partition of the observations such that any two X_i and X_j belong to the same cluster if the corresponding latent variables θ_i and θ_j coincide. Then,it is interesting to determine an estimate $\hat{R}_n$ of the number of clusters into which the data are grouped. In the examples we will illustrate $\hat{R}_n$ is set equal to the mode of $R_n|\boldsymbol{X}$, with $\boldsymbol{X} := (X_1, \ldots, X_n)$ representing the observed sample. Both estimation problems can be faced by relying on the simulation algorithm that will be detailed in the next section.

3. POSTERIOR SIMULATION OF NRMI MIXTURES

Our main aim is to provide a general algorithm to draw posterior inferences with the mixture model (22), for any choice of the mixing NRMI and of the kernel. A further by-product of our algorithm is the possibility of determining credible intervals. The main block of the conditional algorithm presented in this section is the posterior representation provided in Theorem 1. In fact, in order to sample from the posterior distribution of the random mixture model (22), given a sample $X_1, \ldots, X_n$, a characterization of the posterior distribution of the mixing measure at the higher stage of the hierarchy is needed. We rely on the posterior representation, conditional on the unobservable variables $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n)$, of the un–normalized process $\tilde{M}$, since the normalization can be carried out within the algorithm.

For the implementation of a Gibbs sampling scheme we use the distributions of

$$[\tilde{M}|\boldsymbol{X}, \boldsymbol{\theta}] \;\text{ and }\; [\boldsymbol{\theta}|\boldsymbol{X}, \tilde{M}]. \tag{24}$$

For illustration we shall detail the algorithm when $\tilde{P} \sim \mathrm{NGG}(a, \kappa, \gamma; P_0)$ and provide explicit expressions for each of the distributions in (24). Nonetheless, as already recalled the algorithm can be implemented for any NRMI: one just needs to plug–in the corresponding Lévy intensity.

Due to conditional independence properties, the conditional distribution of $\tilde{M}$ given $\boldsymbol{X}$ and $\boldsymbol{\theta}$ does not depend on $\boldsymbol{X}$, that is $[\tilde{M}|\boldsymbol{X},\boldsymbol{\theta}] = [\tilde{M}|\boldsymbol{\theta}]$. Now, by Theorem 1, the posterior distribution function $[\tilde{M}|\boldsymbol{\theta}]$ is characterized as a mixture in terms of a latent variable U, i.e. through $[\tilde{M}|U,\boldsymbol{\theta}]$ and $[U|\boldsymbol{\theta}]$. Specifically, the conditional distribution of $\tilde{M}$ given U and $\boldsymbol{\theta}$ is another CRM with fixed points of discontinuity at the distinct θ_i's, namely $\{\theta_1^*, \ldots, \theta_r^*\}$, given by

$$\tilde{M}_+^*(s) := \tilde{M}^*(s) + \sum_{j=1}^{r} J_j^* \, \mathbb{I}_{(-\infty,s]}(\theta_j^*), \tag{25}$$

where $(-\infty, s] = \{y \in \mathrm{I\!R}^m : y_i \leq s_i, \ \ i = 1, \ldots, m\}$. Recall that in the $\mathrm{NGG}(a, \kappa, \gamma; P_0)$ case, $\tilde{M}^*$ has Lévy intensity as in (19) and the density of the jumps J_j^* is (20). Finally, the conditional distribution of U given $\boldsymbol{\theta}$, is then (21).

The second conditional distribution $[\boldsymbol{\theta}|\boldsymbol{X},\tilde{M}]$ involved in the Gibbs sampler in (24) consists of conditional independent distributions for each θ_i, whose density is given by

$$f_{\theta_i|X_i,\tilde{M}}(s) \propto k(X_i|s)\tilde{M}_+^*\{s\}, \tag{26}$$

for $i = 1, \ldots, n$, where the set $\{s\}$ corresponds to the m-variate jump locations $s \in \mathrm{I\!R}^m$ of the posterior process $\tilde{M}_+^*$.

In the following we will provide a way of simulating from each of the distributions (25), (21) and (26).

3.1 Simulating $[\tilde{M}|U,\boldsymbol{\theta}]$.

Since the distribution of the process $\tilde{M}$ given U and $\boldsymbol{\theta}$ is the distribution function associated to a CRM, we need to sample its trajectories. Algorithms for simulating such processes usually rely on inverse Lévy measure techniques as is the case for the algorithms devised in Ferguson and Klass (1972) and in Wolpert and Ickstadt (1998). According to Walker and Damien (2000) the former is more efficient in the sense that it has a better performance with a small number of simulations. Therefore, for simulating from the conditional distribution of $\tilde{M}$ we follow the Ferguson and Klass device. Their idea is based on expressing the part without fixed points of discontinuity of the posterior $\tilde{M}_+^*$, which in our case is $\tilde{M}^*$, as an infinite sum of random jumps J_j that occur at random locations $\vartheta_j = (\vartheta_j^{(1)}, \ldots, \vartheta_j^{(m)})$, i.e.

$$\tilde{M}^*(s) = \sum_{j=1}^{\infty} J_j \mathbb{I}_{(-\infty,s]}(\vartheta_j). \tag{27}$$

The positive random jumps are ordered, i.e., $J_1 \geq J_2 \geq \cdots$, since the J_j's are obtained as $\xi_j = N(J_j)$, where $N(v) = \nu^*([v,\infty), \mathrm{I\!R}^m)$ and $\xi_1, \xi_2, \ldots$ are jump times of a standard Poisson process of unit rate, i.e., $\xi_1, \xi_2 - \xi_1, \ldots \overset{\text{iid}}{\sim} \mathrm{ga}(1,1)$. Here $\mathrm{ga}(a,b)$ denotes a gamma distribution with shape and scale parameters a and b. The random locations ϑ_j conditional on the jump sizes J_j, are obtained from the distribution function $F_{\vartheta_j|J_j}$, given by

$$F_{\vartheta_j|J_j}(s) = \frac{\nu^*(\mathrm{d}J_j, (-\infty, s])}{\nu^*(\mathrm{d}J_j, \mathrm{I\!R}^m)}.$$

Therefore, the J_j's can be obtained by solving the equations $\xi_i = N(J_i)$. This can be accomplished by combining quadrature methods to approximate the integral (see, for example, Burden and Faires, 1993) and a numerical procedure to solve the equation. Moreover, when one is dealing with a homogeneous NRMI the jumps are independent of the locations and therefore $F_{\vartheta_j|J_j} = F_{\vartheta_j}$ does not depend on J_j, implying that the locations are iid samples from P_0. For an extension of the Ferguson–Klass device to general space see Orbanz and Williamson (2011).
In our specific case where $\tilde{M}$ is a generalized gamma process, the functions N and F_ϑ take on the form

$$\begin{aligned} N(v) &= \frac{a}{\Gamma(1-\gamma)} \int_v^\infty e^{-(\kappa+u)x} x^{-(1+\gamma)} \mathrm{d}x, \\ F_\vartheta(s) &= \int_{(-\infty,s]} P_0(\mathrm{d}y), \end{aligned} \tag{28}$$

and all above described steps become straightforward.

As for the part of $\tilde{M}_+^*$ concerning the fixed points of discontinuity, the distribution of the jumps at the fixed locations will depend explicitly on the underlying Lévy intensity as can be seen from (17). In the NGG case they reduce to the gamma distributions displayed in (20).

Now, combining the two parts of the process, with and without fixed points of discontinuity, the overall posterior representation of the process $\tilde{M}$ will be

$$\tilde{M}_+^*(s) = \sum_j \bar{J}_j \mathbb{I}_{(-\infty,s]}(\bar{\vartheta}_j),$$

having set $\{\bar{J}_j\}_{j\ge1} = \{J_1^*, \dots, J_r^*, J_1, \dots\}$ and also $\{\bar{\vartheta}_j\}_{j\ge1} = \{\theta_1^*, \dots, \theta_r^*, \vartheta_1, \dots\}$.

Remark 3. A fundamental merit of Ferguson and Klass' representation, compared to similar algorithms, is the fact that the random heights J_i are obtained in a descending order. Therefore one can truncate the series (27) at a certain finite index ℓ in such a way that the relative error between $\sum_{j\le\ell} J_j$ and $\sum_{j\le\ell+1} J_j$ is smaller than ϵ, for any desired $\epsilon > 0$. This guarantees, on the one hand, that the highest jumps are not left out and, on the other hand, allows us to control the size of the ignored jumps. Argiento, Guglielmi and Pievatolo (2010) provide an upper bound for the ignored jump sizes. □

As mentioned before, the generalized gamma NRMI defines a wide class of processes which include gamma, inverse Gaussian and stable processes. To appreciate better the difference between these processes, consider the function N in (28). This function is depicted in Figure 1 for the three cases with parameters fixed in such a way that the corresponding NRMIs (Dirichlet, normalized inverse Gaussian and normalized stable) have the same mean and variance structures. See Lijoi, Mena and Prünster (2005) and James, Lijoi and Prünster (2006) for the relevant explicit expressions needed to fix the parameters. In particular, Figure 1 is displayed in two panels which represent close up views to the upper left and bottom right tails of the graph.

The function N defines the height of the jumps in the part of the process without fixed points of discontinuity, i.e. $J_i = N^{-1}(\xi_i)$. To help intuition, imagine horizontal lines going up in Figure 1. The values in the y-axis correspond to

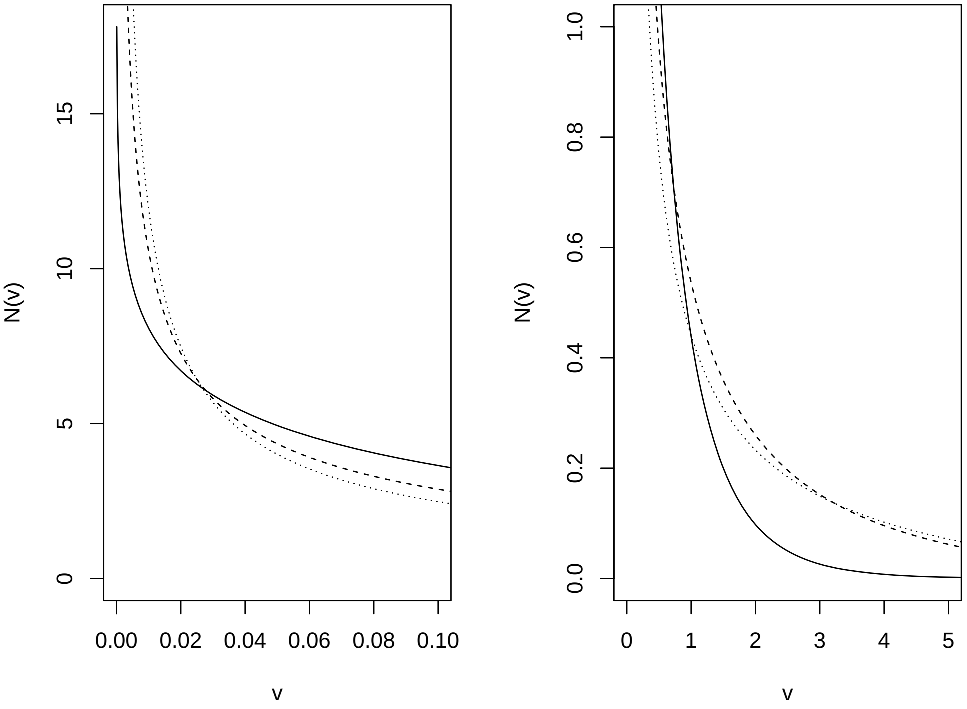

FIGURE 1. *Function N in (28) for three special cases of generalized gamma process: gamma process with $(a, \kappa, \gamma) = (2, 1, 0)$ (solid line); inverse Gaussian process with $(a, \kappa, \gamma) = (1, 0.126, 0.5)$ (dashed line); and stable process with $(a, \kappa, \gamma) = (1, 0, 0.666)$ (dotted line). In all the cases, the prior mean and variance of $\tilde{P}(A)$ obtained by normalization are the same, for any A.*

the Poisson process jumps and, for each of them, there is a value in the x-axis corresponding to the jump sizes of the process. Looking at the right panel in Figure 1 we can see that the stable process has the largest jumps followed closely by the inverse Gaussian process. On the other hand, the left panel shows the concentration of the sizes of the jumps of the (un–normalized) CRMs around the origin. Hence, the stable CRM tends to have a larger number of jumps of "small" size when compared to the Dirichlet process, with the N-IG process again in an intermediate position. As shown in Kingman (1975), this different behavior also impacts the normalized weights. To grasp the idea, let the J_i's be the jump sizes of the CRM and $\tilde{p}_i = J_i / \sum_{k\geq 1} J_k$ are the normalized jumps. Moreover, $(\tilde{p}_{(j)})_{j\geq 1}$ is the sequence obtained by considering the $\tilde{p}_j$'s in decreasing order so that $\tilde{p}_{(1)} > \tilde{p}_{(2)} > \cdots$. One then has $\tilde{p}_{(j)} \sim \exp\{-j/a\}$ as $j \to \infty$, almost surely, in the Dirichlet case; whereas $\tilde{p}_{(j)} \sim \xi(\gamma) j^{-1/\gamma}$ as $j \to \infty$, almost surely, in the N–stable case. Here $\xi(\gamma)$ is a positive random variable. Hence, for j large enough the atom $Z_{(j)}$ associated to the weight $\tilde{p}_{(j)}$ is less likely to be observed in the Dirichlet case rather than in the N–stable case. These arguments can be suitably adapted and the conclusion can be extended to the case where the N–stable is replaced by a NGG(a, κ, γ, P_0) process, for any $\gamma \in (0, 1)$. An important well–known implication of this different behavior concerns the distribution of the number of distinct values R_n: clearly for both the Dirichlet and the NGG(a, κ, γ, P_0) (with $\gamma > 0$) processes R_n diverges as n diverges; however, the rate at which the number of clusters R_n increases is slower in the Dirichet than in the NGG case being, respectively, $\log(n)$ and n^γ. Moreover, in order to gain a full understanding of the role of γ in determining the clustering structure featured by models defined either as in (3) or (22), one has to consider the influence γ has on the sizes of the clusters. To this end, it is useful to recall that when $\gamma > 0$ a reinforcement mechanism of larger clusters takes place. A concise description is as follows: Consider a configuration reached after sampling n values, denote by n_i and n_j the sizes of the i–th and j–th cluster, respectively, with $n_i > n_j$. Then, the ratio of the probabilities that the $(n+1)$–th sampled value will belong to the i–th or j–th clusters coincides with $(n_i - \gamma)/(n_j - \gamma)$, an increasing function of γ, with its lowest value corresponding to the Dirichlet process, i.e. $\gamma = 0$. For instance, if $n_i = 2$ and $n_j = 1$, the probability of sampling a value belonging to the i–th cluster is twice the probability of getting a value belonging to the j–th cluster in the Dirichlet case, whereas it is three times larger for $\gamma = 1/2$ and five times larger for $\gamma = 3/4$. This implies that as γ increases, the clusters tend to be much more concentrated with a very large number of small clusters and very few groups having large frequencies. In other words, a mass re-allocation occurs and it penalizes clusters with smaller sizes while reinforcing larger clusters, which are interpreted as those having stronger empirical evidence. On the other hand, κ (or a) does not have any significant impact on the balancedness of the partition sets. This mechanism is far from being a drawback and Lijoi, Mena and Prünster (2007) have shown that it is beneficial when drawing inference on the number of components in a mixture. Finally, it is worth stressing that, in general, the unevenness of partition configurations is an unavoidable aspect of a nonparametric models beyond the specific cases we are considering here. This is due to the fact, that with discrete nonparametric priors, R_n increases indefinitely with n. Hence, for any n there will always be a positive probability that a new value is generated and, even if at different rates, new values will be continuously added making it impossible to obtain models with (a priori) balanced partitions. If one needs balancedness even a priori, a finite–dimensional model is more appropriate.

3.2 Simulating $[\boldsymbol{U}|\boldsymbol{\theta}]$

Since the conditional density of U given in (21), is univariate and continuous, there are several ways of drawing samples from it. Damien, Wakefield and Walker (1999), for instance, propose to introduce uniform latent variables to simplify the simulation. However, in our experience, this procedure increases the autocorrelation in the chain thus leading to a slower mixing. Additionally, the values of this conditional density explode for sample sizes larger than 100. An alternative procedure consists of introducing a Metropolis-Hastings (M–H) step (see, e.g., Tierney, 1994). M–H steps usually work fine as long as the proposal distribution is adequately chosen, and since they rely only on ratios of the desired density, this solves the overflow problem for large values of n.

In our approach we propose to use a M–H step with proposal distribution that follows a random walk. Since U takes only positive values, we use a gamma proposal distribution centered at the previous value of the chain and with coefficient of variation equal $1/\sqrt{\delta}$. Specifically, at iteration $[t+1]$ simulate $u' \sim \mathrm{ga}(\delta, \delta/u^{[t]})$ and set $u^{[t+1]} = u'$ with acceptance probability given by

$$q_1(u', u^{[t]}) = \min\left\{1, \frac{f_{U|\boldsymbol{\theta}}(u')\mathrm{ga}(u^{[t]}|\delta, \delta/u')}{f_{U|\boldsymbol{\theta}}(u^{[t]})\mathrm{ga}(u'|\delta, \delta/u^{[t]})}\right\}, \tag{29}$$

where $\mathrm{ga}(\,\cdot\,|a, b)$ denotes the density function of a gamma random variable whose expected value is a/b. The parameter δ controls the acceptance rate of the M–H step being higher for larger values. It is suggested to use $\delta \geq 1$.

3.3 Re-sampling the unique values $\boldsymbol{\theta}_j^*$

It is well known that discrete nonparametric priors, as is the case of NRMIs, induce some effect when carrying out posterior inference via simulation. This is called by some authors "sticky clusters effect". Bush and MacEachern (1996) suggested an important acceleration step to overcome this problem by re–sampling the location of the fixed jumps θ_j^* from its conditional distribution given the cluster configuration (c.c.), which in this case takes on the form

$$f_{\theta_j^*|\boldsymbol{X},\mathrm{c.c.}}(s) \propto f_0(s) \prod_{i \in C_j} k(X_i|s), \tag{30}$$

where $C_j = \{i : \theta_i = \theta_j^*\}$. Also recall that $\theta_j^* = (\theta_{j1}^* \ldots, \theta_{jm}^*) \in \mathrm{I\!R}^m$ with $m \geq 1$. For the case $m = 2$ of location–scale mixture, i.e. $\theta = (\mu, \sigma)$, we suggest to use a M–H step with joint proposal distribution for the pair (μ, σ) whose density we denote in general by g. In particular, at iteration $[t+1]$ one could sample $\theta^{*'} = (\mu_j^{*'}, \sigma_j^{*'})$ by first taking $\sigma_j^{*'} \sim \mathrm{ga}(\delta, \delta/\sigma_j^{*[t]})$ and then, conditionally on $\sigma_j^{*'}$, take $\mu_j^{*'}$ from the marginal base measure on μ, f_0^1, specified in such a way that its mean coincides with $\bar{X}_j$ and its standard deviation with $\eta\sigma_j^{*'}/\sqrt{n_j}$, where $\bar{X}_j = \frac{1}{n_j}\sum_{i \in C_j} X_i$. Finally set $\theta_j^{*[t+1]} = \theta_j^{*'}$ with acceptance probability given by

$$q_2(\theta^{*'}, \theta^{*[t]}) = \min\left\{1, \frac{f_{\theta_j^*|\boldsymbol{X},\mathrm{c.c.}}(\theta')}{f_{\theta_j^*|\boldsymbol{X},\mathrm{c.c.}}(\theta^{*[t]})}\,\frac{g(\theta^{*[t]})}{g(\theta')}\right\}. \tag{31}$$

For the examples considered in this paper we use $\delta = 4$ and $\eta = 2$ to produce a moderate acceptance probability.

3.4 Simulating $[\boldsymbol{\theta}|\boldsymbol{X}, \tilde{\boldsymbol{M}}]$

Since $\tilde{M}_+^*$ is a pure jump process, the support of the conditional distribution of θ_i are the locations of the jumps of $\tilde{M}_+^*$, that is $\{\bar{\vartheta}_j\}$, and therefore

$$f_{\theta_i|X_i,\tilde{M}}(s) \propto \sum_j k(X_i|s)\bar{J}_j\delta_{\bar{\vartheta}_j}(\mathrm{d}s). \tag{32}$$

Simulating from this conditional distribution is straightforward: one just needs to evaluate the right hand side of the expression above and normalize.

3.5 Updating the hyper-parameters of P_0

As pointed out by one of the Referees, in general the hyper-parameters ϕ of the base measure density $f_0(\,\cdot\,|\phi)$ affect the performance of nonparametric mixtures. For the location–scale mixture case, i.e. $m = 2$ with $\theta = (\mu, \sigma)$ and $f_0(\mu, \sigma|\phi) = f_0^1(\mu|\sigma, \varphi)f_0^2(\sigma|\varsigma)$, it turns out that the subset of parameters φ pertaining the locations μ have a higher impact. By assuming in addition a priori independence between μ and σ, the conditional posterior distribution of φ, given the observed data and the rest of the parameters, only depends on the distinct μ_i's, say μ_j^*, for $j = 1, \dots, r$. The simplest way to proceed is to consider a conjugate prior $f(\varphi)$ for a sample $\mu_1^*, \dots, \mu_r^*$ from $f_0^1(\mu|\varphi)$. Clearly such a prior depends on the particular choice of f_0^1 and some examples will be considered in Section 4.

3.6 Computing a path of $\tilde{f}(x)$

Once we have a sample from the posterior distribution of the process $\tilde{M}$, the desired path from the posterior distribution of the random density $\tilde{f}$, given in (22), can be expressed as a discrete mixture of the form

$$\tilde{f}(x|\tilde{M}_+^*, \phi) = \sum_j k(x|\bar{\vartheta}_j)\frac{\bar{J}_j}{\sum_l \bar{J}_l}. \tag{33}$$

3.7 General algorithm

An algorithm for simulating from the posterior distributions (24) can be summarized as follows. Given the starting points $\theta_1^{[0]}, \dots, \theta_n^{[0]}$, with the corresponding unique values $\theta_j^{*[0]}$ and frequencies $n_j^{[0]}$, for $j = 1, \dots, r$, and given $u^{[0]}$, at iteration $[t + 1]$:

1. Sample the latent $U|\boldsymbol{\theta}$: simulate a proposal value $U^* \sim \mathrm{ga}(\delta, \delta/U^{[t]})$ and take $U^{[t+1]} = U^*$ with probability $q_1(U^*, U^{[t]})$, otherwise take $U^{[t+1]} = U^{[t]}$, where the acceptance probability q_1 is given in (29).
2. Sample trajectories of the part of the process without fixed points of discontinuity $\tilde{M}^*$: simulate $\zeta_j \sim \mathrm{ga}(1, 1)$ and find $J_j^{[t+1]}$ by solving numerically the equation $\sum_{l=1}^j \zeta_l = N(J_j)$; simulate $\vartheta_j^{[t+1]}$ from P_0. The function N is given in (28). Stop simulating when $J_{\ell+1}/\sum_{j=1}^{\ell} J_j < \epsilon$, say $\epsilon = 0.0001$.
3. Re–sample the unique values $\{\theta_j^*\}$: record the unique values $\theta_j^{*[t]}$ from $\{\theta_1^{[t]}, \dots, \theta_n^{[t]}\}$ and their frequency $n_j^{[t]}$. If $m = 2$ with $k(\,\cdot\,|\theta)$ parameterized in terms of mean and standard deviation ($\theta = (\mu, \sigma)$), simulate a pair

$(\mu_j^{*\backprime}, \sigma_j^{*\backprime})$ from a joint proposal (see Section 3.3 and then set $\theta_j^{*[t+1]}$ equal to $\theta_j^{*\backprime}$ with probability $q_2(\theta^{*\backprime}, \theta^{*[t]})$. Otherwise take $\theta_j^{*[t+1]} = \theta_j^{*[t]}$. The acceptance probability q_2 is given in (31).

4. Sample the fixed jumps of the process, $\{J_j^*\}$: for each $\theta_j^{*[t+1]}$ with frequency $n_j^{[t+1]}$, $j = 1, \ldots, r$, sample the jump $J_j^{*[t+1]} \sim \mathrm{ga}(n_j^{[t+1]} - \gamma, \kappa + u^{[t+1]})$.
5. Update the hyper-parameters $\boldsymbol{\phi}$ of $f_0(\theta|\boldsymbol{\phi})$: in particular, for the case of $m = 2$ with $\theta = (\mu, \sigma)$ simulate a value $\boldsymbol{\varphi}^{[t+1]}$ from its conditional posterior distribution as described in Section 3.5.
6. Sample the latent vector $\boldsymbol{\theta}$: for each $i = 1, \ldots, n$, sample $\theta_i^{[t+1]}$ from its discrete conditional density given in (32) by evaluating the kernel $k(X_i|\cdot)$ at the different jump locations $\{\bar{\vartheta}_j^{[t+1]}\} = \{\theta_1^{*[t+1]}, \ldots, \theta_r^{*[t+1]}, \vartheta_j^{[t+1]}, \ldots\}$ and weights $\{\bar{J}_j^{[t+1]}\} = \{J_1^{*[t+1]}, \ldots, J_r^{*[t+1]}, J_1^{[t+1]}, \ldots\}$.
7. Compute a path of the desired random density function $\tilde{f}(x|(\tilde{M}_+^*)^{[t+1]})$ as in (33).

Repeat steps 1 to 7 for $t = 1, \ldots, T$. Note that the values of δ and η can be used to tune the acceptance probability in the M-H steps. The values suggested here are those considered more appropriate according to our experience. The performance of this algorithm depends on the particular choices of the density kernel, the NRMI driving measure and the dataset at hand. In order to assess the mixing of the chains one can resort to the effective sample size (ESS) implemented in the R package library `coda`. In our context the natural parameter to consider for assessing the mixing is given by the total jump sizes of the NRMI process $\sum_j \bar{J}_j$. First note that the conjugacy of the Dirichlet process yields a simpler posterior representation (independent of the latent variable U) and recall also that the jumps are independent of the locations. Therefore the samples are independent and the ESS coincides with the number of iterations of the chain. For the other NRMIs this is not the case: the posterior representation depends on the latent variable U and, moreover, the distribution of the jumps depends on the θ_j^*'s. For instance, for the two real datasets considered in Section 4.1, for chains of length 4,500 (obtained from 20,000 iterations with burn–in of 2,000 and keeping every 4th iteration), the ESS was around 1,250 for the N-IG process and for the associated latent variable U the value of the ESS was 1,500.

4. COMPARING NRMI MIXTURES

In this section we provide a comprehensive illustration of NRMI mixtures using the R package `BNPdensity`, which implements the general algorithm outlined in Section 3.7. The aim of such a study is two–fold: on the one hand it illustrates the potential and flexibility of NRMI mixture models in terms of fitting and capturing the appropriate number of clusters in a dataset for different choices of kernels and mixing NRMI; on the other hand, we also compare the performance of NRMI mixtures with respect to other alternative density estimates.

To implement the algorithm described in the previous section, we first specify the mixture kernel $k(\cdot|\theta)$. We will consider, in total, a set of four kernels parameterized in terms of mean μ and standard deviation σ such that $\theta = (\mu, \sigma)$. Two of these kernels have support $\mathbb{R}$ and the other two have support $\mathbb{R}^+$. They are:

i. Normal kernel:

$$k(x|\mu,\sigma)=\frac{1}{\sqrt{2\pi}b}\exp\left\{-\frac{1}{2b^2}(x-a)^2\right\}\mathbb{I}_{\mathbb{R}}(x),$$

with $a=\mu$ and $b=\sigma$.

ii. Double exponential kernel:

$$k(x|\mu,\sigma)=\frac{1}{2b}\exp\left\{-\frac{1}{b}|x-a|\right\}\mathbb{I}_{\mathbb{R}}(x),$$

with $a=\mu$ and $b=\sigma/\sqrt{2}$.

iii. Gamma kernel:

$$k(x|\mu,\sigma)=\frac{b^a}{\Gamma(a)}x^{a-1}e^{-bx}\mathbb{I}_{\mathbb{R}^+}(x),$$

with $a=\mu^2/\sigma^2$ and $b=\mu/\sigma^2$.

iv. Log-normal kernel:

$$k(x|\mu,\sigma)=\frac{1}{x\sqrt{2\pi}b}\exp\left\{-\frac{1}{2b^2}(\log x-a)^2\right\}\mathbb{I}_{\mathbb{R}^+}(x),$$

with $a=\log\left(\frac{\mu}{\sqrt{1+\sigma^2/\mu^2}}\right)$ and $b=\sqrt{\log\left(1+\frac{\sigma^2}{\mu^2}\right)}$

As for the NRMI mixing measure, we will resort to different members of the class NGG$(a,\kappa,\gamma;\,P_0)$: the Dirichlet process NGG$(a,1,0;P_0)$, the N–IG process NGG$(1,\kappa,1/2;P_0)$, the N–stable process NGG$(1,0,\gamma;\,P_0)$. Their parameters will be fixed to obtain mixtures with a prior expected number of components $\mathrm{E}(R_n)$ equal to any desired number $c\in\{1,\ldots,n\}$, where n denotes the sample size. This strategy allows one to effectively compare different priors given they induce a priori the same expected number of mixture components. See Lijoi, Mena and Prünster (2007) for details on this procedure. As for the base measure P_0 of the NRMIs to be considered, we will assume a priori independence between μ and σ so that $f_0(\mu,\sigma|\phi)=f_0^1(\mu|\varphi)f_0^2(\sigma|\varsigma)$. In particular, we will take $f_0^2(\sigma|\varsigma)=\mathrm{ga}(\sigma|\varsigma_1,\varsigma_2)$, with shape ς_1 and scale ς_2 fixed a priori to specify a certain knowledge in the degree of smoothness. For f_0^1 we will consider two options with support $\mathbb{R}$ and $\mathbb{R}^+$, respectively. These are:

a. Normal base measure for μ:

$$f_0^1(\mu|\varphi)=\mathrm{N}(\mu|\varphi_1,\varphi_2),$$

where φ_1 and φ_2 are the mean and precision, respectively. The conjugate prior distribution for φ is then $f(\varphi)=\mathrm{N}(\varphi_1|\psi_1,\psi_2\varphi_2)\mathrm{ga}(\varphi_2|\psi_3,\psi_4)$ and the (conditional) posterior distribution, needed for the hyperparameter updating (see Section 3.5) are given by

$$f(\varphi|\mu^*)=\mathrm{N}\left(\varphi_1\left|\frac{\psi_2\psi_1+r\bar{\mu}^*}{\psi_2+r}\right.,\,(\psi_2+r)\varphi_2\right)\times$$

$$\mathrm{ga}\left(\varphi_2\left|\psi_3+\frac{r}{2}\right.,\,\psi_4+\frac{1}{2}\sum_{j=1}^{r}(\mu_j^*-\bar{\mu}^*)^2+\frac{\psi_2 r(\bar{\mu}^*-\psi_1)^2}{2(\psi_2+r)}\right).$$

b. Gamma base measure for μ:

$$f_0^1(\mu|\varphi) = \text{ga}(\mu|1, \varphi),$$

where φ corresponds to the scale parameter. The conjugate prior for φ is $f(\varphi) = \text{ga}(\varphi|\psi_1, \psi_2)$ and the (conditional) posterior distribution is $f(\varphi|\mu^*) = \text{ga}(\varphi|\psi_1 + r, \psi_2 + \sum_{j=1}^r \mu_j^*)$. Clearly this choice is reasonable only for experiments leading to positive outcomes.

Since we aim at comparing the performance of NRMI mixtures in terms of density estimates, we also need to specify measures of goodness of fit. We will use two different measures for the real data and the simulated data. In the former case, we resort to the conditional predictive ordinates (CPOs) statistics, which are now widely used in several contexts for model assessment. See for example, Gelfand, Dey and Chang (1992). For each observation i, the CPO statistic is defined as follows

$$\text{CPO}_i = \tilde{f}(x_i|D^{(-i)}) = \int k(x_i|\theta)\tilde{P}(\text{d}\theta|D^{(-i)}),$$

with $D^{(-i)}$ the data with the i^{th} case excluded and $\tilde{P}(\text{d}\theta|D^{(-i)})$ the posterior density of the model parameters θ based on data $D^{(-i)}$. By re-writing the statistic CPO_i as

$$\text{CPO}_i = \left(\int \frac{1}{k(x_i|\theta)}\tilde{P}(d\theta|D)\right)^{-1},$$

it can be easily approximated by Monte Carlo as

$$\widehat{\text{CPO}_i} = \left(\frac{1}{T}\sum_{t=1}^{T} \frac{1}{k(x_i|\theta^{[t]})}\right)^{-1},$$

where $\{\theta^{[t]}, \, t = 1, 2, \ldots, T\}$ is an MCMC sample from $\tilde{P}(\theta|D)$. We will summarize the CPO_i, $i = 1, \ldots, n$, values in two ways, as an average of the logarithm of CPOs (ALCPO) and as the median of the logarithm of CPOs (MLCPO). The average of log-CPOs is also called average of log-pseudo marginal likelihood and is denoted by ALPML.

In contrast, when considering simulated data, the true model, say f^*, is known and hence, it is possible to use the mean integrated squared error (MISE) for model comparison. If we denote by $\hat{f}_n$ the density estimate conditional on a sample of size n from f^*, then the MISE is defined as

$$\text{MISE} = \text{E}\left\{\int \{\hat{f}_n(x) - f^*(x)\}^2 \text{d}x\right\}.$$

Like in other approaches to density estimation (see, e.g., Müller and Vidakovic, 1998; Roeder and Wasserman, 1997) the standard method to compare with is the kernel density estimator (Silverman, 1986). Therefore, instead of the MISE, we report the relative MISE (RMISE) defined as the ratio of the MISE obtained with the NRMI mixture model and the MISE obtained with the kernel density estimator with standard bandwidth.

We are now in a position to illustrate our methodology. We first provide the analysis of two real datasets popular in the mixture modeling literature, namely the galaxy data and the enzyme data. See Richardson and Green (1997). Then, we perform an extensive simulation study by considering the models dealt with in Marron and Wand (1992). In analyzing the real data we focus on the performance of different NRMI mixtures, by varying kernel and mixing NRMI and illustrate the flexibility of the algorithm. Later, through the simulation study we aim at comparing NRMI mixtures with other methods used in the literature. For this purpose we fix a single NRMI mixture. Such a choice, based on the results of the real data examples and on our previous experience, exhibits good and robust performances hus making it a valid default model.

4.1 Real data

4.1.1 Galaxy data For illustration of the algorithm and analysis of NRMI mixtures we start with some real data. The first dataset we consider is the widely studied galaxy dataset. Data consist of velocities of 82 distant galaxies diverging from our own galaxy. Typically this density has been estimated by considering mixtures of normal kernels (Escobar and West, 1995; Richardson and Green, 1997; Lijoi, Mena and Prünster, 2005): given the data range from 9.2 to 34, clearly away from zero, it is possible to use kernels with support $\mathbb{R}$. Here, we compare the normal kernel with another kernel with real support, namely the double exponential kernel. These two kernels are written in mean and standard deviation parameterization as in cases (i) and (ii) above. In terms of mixing measures we compare two options: the Dirichlet process with specifications $\mathrm{NGG}(3.641, 1, 0; P_0)$ and the N–IG process with specifications $\mathrm{NGG}(1, 0.015, 1/2; P_0)$. The prior parameters of the two processes were determined so to obtain an expected number of a priori components equal to 12, roughly twice the typically estimated number of components, which is between 4 and 6. It is worth noting that with such a prior specification the N-stable process would correspond to a $\mathrm{NGG}(1, 0, 0.537; P_0)$. This essentially coincides with the above N-IG specification which indeed has a small value of κ and $\gamma = 1/2$, and is therefore omitted.

For the base measure P_0 we took $f_0^2(\sigma|\varsigma) = \mathrm{ga}(\sigma|\varsigma_1, \varsigma_2)$ with two specifications for $(\varsigma_1, \varsigma_2)$, namely $(1, 1)$ and $(0.1, 0.1)$, and the gamma specification in case (b) above for $f_0^1(\mu|\varphi)$ with a vague hyperprior on the scale parameter φ, namely $\psi_1 = \psi_2 = 0.01$. In neither case P_0 is conjugate w.r.t. the kernel and in addition to the standard deviations, it forces also the means of the mixture components to be positive as required. The Gibbs sampler was run for $20,000$ iterations with a burn-in of $2,000$ sweeps. One simulation every 4^{th} after burn-in was kept resulting in $4,500$ iterations to compute the estimates.

Table 1 provides the ALCPO statistics, the MLCPO statistics and the mode of posterior distribution of the number of components, $R_n|\boldsymbol{X}$, for the $8 = 2 \times 2 \times 2$ combinations of kernel–NRMI–$(\varsigma_1, \varsigma_2)$. Recall that the ALCPO and MLCPO statistics are the average and the median of the CPOs in log scale respectively. First note that starting from an "incorrect" prior specification of the number of components R_n the N–IG process mixture is able to detect the typically estimated number of components regardless of the choice of the kernel and the other parameters. In contrast DPMs are not able to overcome completely the wrong prior specification and tend to overestimate the number of components. As one would expect, given on the one hand a distribution can always be fitted with more components than necessary and on the other the kernel smooths out differences in the mixing measures, the differences between the two processes in terms of the density estimates are much less evident. Considering the ALCPO goodness of fit statistics, the best fitting is obtained with the normal DPM with $(\varsigma_1, \varsigma_2) = (1, 1)$. However, the differences w.r.t. other specifications are not particularly remarkable. If, instead, we consider the MLCPO statistic, the best fitting is achieved by the N–IG normal mixture with $(\varsigma_1, \varsigma_2) = (1, 1)$ and the superior performance starts becoming significant being 0.1 better than any DPM specification. The overall behavior of the CPO is illustrated by Figure 2, where box–plots of the logarithm of the CPO values corresponding to normal mixtures with $(\varsigma_1, \varsigma_2) = (1, 1)$ for both Dirichlet and N-IG processes are depicted. Coherently with the values of the ALCPO and MLCPO, the logarithm of the CPOs produced by the DPM are more dispersed: for some trajectories it produces the best ordinates, which, once averaged lead to a slightly better ALCPO; however, if we consider a more robust summary, like the median, the N–IG mixture produces a significantly better result.

Measure	Kernel	$(\varsigma_1, \varsigma_2)$	ALCPO	MLCPO	Mode$(R_n\|\boldsymbol{X})$
Dirichlet	Normal	$(1, 1)$	**-2.581**	-2.250	7
		$(0.1, 0.1)$	-2.619	-2.205	6
	Dble.Exp.	$(1, 1)$	-2.597	-2.303	7
		$(0.1, 0.1)$	-2.620	-2.305	6
N–IG	Normal	$(1, 1)$	-2.608	**-2.099**	5
		$(0.1, 0.1)$	-2.647	-2.154	3
	Dble.Exp.	$(1, 1)$	-2.600	-2.258	5
		$(0.1, 0.1)$	-2.637	-2.260	4

Table 1

Galaxy dataset: Summaries of log-conditional predictive ordinates [average (ALCPO) and median (MLCPO)] and mode of the posterior distribution of the number of components in the mixture, $R_n|bmX$, for different prior specifications. Bold numbers denote best fitting according to the corresponding statistic.

Figure 3 displays the density estimates together with 95% pointwise credible intervals when using the Dirchlet and N–IG process mixtures with normal and double exponential kernels. In accordance to the above results there is not much difference in terms of the chosen nonparametric prior. However, it is interesting to note how the double exponential kernel, while exhibiting poorer performance in terms of CPO, produces significantly sharper estimates than the normal kernel. This feature which singles out possible modes may be desirable in certain situations.

4.1.2 Enzyme Data The second example consists of 245 measurements of the enzymatic activity in the blood of unrelated patients. The values of this dataset are all positive and close to zero, ranging from 0.021 to 2.9. Richardson and Green (1997) analyzed this dataset and applied a finite mixture of normals model to estimate the density, even though the data are fairly close to zero. Instead of working with real support kernels, we perform our analysis with positive support kernels to be more consistent with the nature of the data. In particular, we take the gamma density kernel and the log-normal density kernel, both with the mean and standard deviation parameterizations as displayed in cases (iii) and (iv) at

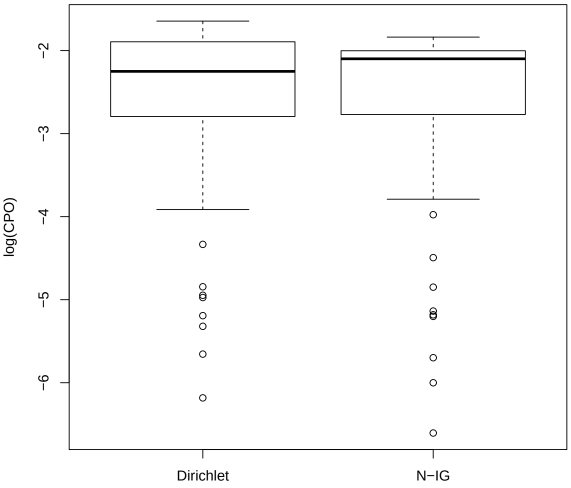

FIGURE 2. *Galaxy dataset: Box–plot of the logarithm of the conditional predictive ordinates for DPM and N–IG mixtures, with normal kernel and* $(\varsigma_1, \varsigma_2) = (1, 1)$.

the beginning of the section.

As for the nonparametric mixing measures, we consider the Dirichlet process $\mathrm{NGG}(4.977, 1, 0;\ P_0)$ and the N–IG process $\mathrm{NGG}(1, 0.007, 1/2; P_0)$. The prior parameters were fixed so to obtain an expected number of a priori components equal to 20. Again, the specification of the corresponding N–stable process $\mathrm{NGG}(1, 0, 0.523; P_0)$ essentially coincides with the above N-IG process and is therefore omitted. Note that such a value for the prior expected number of components is much larger than the typically 2 or 3 components estimated for this dataset. As for the base measure P_0, we took $f_0^2(\sigma|\varsigma) = \mathrm{ga}(\sigma|\varsigma_1, \varsigma_2)$ with two possible sets of values for the hyperparameters that is $(\varsigma_1, \varsigma_2) = (4, 1)$ and $(\varsigma_1, \varsigma_2) = (0.5, 0.5)$. Moreover, for μ the gamma specification in (b) is adopted with a vaguely informative hyperprior on the scale, namely $\psi_1 = \psi_2 = 0.01$. We remark that, as in the previous example, these choices give rise to base measures that are not conjugate for the kernel. The Gibbs sampler was run for $20,000$ iterations with a burn-in of $2,000$ sweeps, keeping one simulation of every 4^{th}, ending up with $4,500$ iterations to compute the estimates.

Table 2 provides the ALCPO statistics, the MLCPO statistics and the mode of the posterior distribution of the number of components for the $8 = 2 \times 2 \times 2$ combinations of kernel–NRMI–$(\varsigma_1, \varsigma_2)$, respectively. Let us first focus on the estimated number of components. In this case, starting from a "strongly incorrect" prior specification of the number of components, the ability of N–IG mixtures to overcome misspecifications becomes even more apparent. Indeed, it can be seen that the N–IG mixture estimates at least 3 fewer components than the DPM, for any choice of the kernels and of the base measures hyperparameters. Having established the better performance of the N–IG mixtures, we have a closer look at the impact of the kernels and hyperparameter specifications in Figure 4 we display the corresponding complete posterior distributions of the number of com-

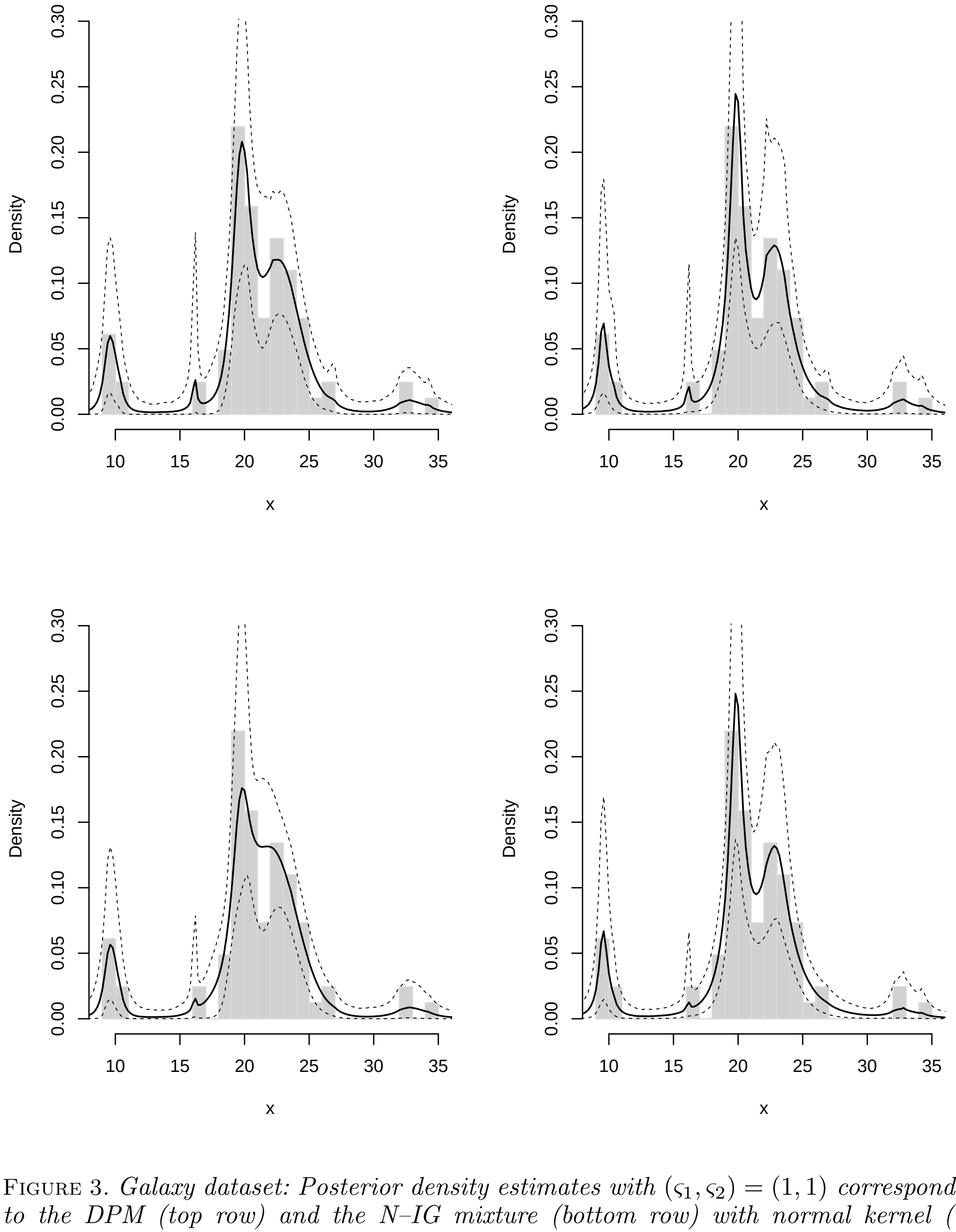

FIGURE 3. *Galaxy dataset: Posterior density estimates with* $(\varsigma_1, \varsigma_2) = (1, 1)$ *corresponding to the DPM (top row) and the N–IG mixture (bottom row) with normal kernel (left column) and double exponential kernel (right column).*

ponents. The gamma kernel displays a better performance in locating the number of components with, additionally, a lower variability, regardless of the hyperparameters choice. With respect to the choice of hyperparameters in the distribution of σ, the ones generating larger values with higher variability are superior. When looking at the density estimates the differences are, as in the previous example, less apparent. In terms of the ALCPO goodness of fit statistics, the best fitting is obtained through the DPM with lognormal kernel and $(\varsigma_1, \varsigma_2) = (0.5, 0.5)$, but the differences with respect to the other specifications are minimal. Nonetheless it is worth pointing out that this corresponds to the case which has the worst behavior in terms of estimation of the number of components. On the one side, this confirms that using more components than necessary does not impact the fit in terms of density estimation. On the other hand it represents an indication that goodness of fit summaries have to be handled with some care to understand the numerical output. If we consider the MLCPO statistic, the best fitting is achieved by the model one would actually expect on the basis of the analysis of the posterior distribution of the number of components, namely the N–IG process mixture with gamma kernel and $(\varsigma_1, \varsigma_2) = (4, 1)$. Moreover, its superiority is quite significant w.r.t. all other specifications. This enforces our previous comment concerning the care needed in drawing conclusions from numerical summaries of the fit.

Measure	Kernel	$(\varsigma_1, \varsigma_2)$	ALCPO	MLCPO	Mode$(R_n \mid \boldsymbol{X})$
Dirichlet	Gamma	$(4, 1)$	-0.227	0.204	5
		$(0.5, 0.5)$	-0.218	0.126	13
	Log.N.	$(4, 1)$	-0.216	0.054	8
		$(0.5, 0.5)$	**-0.205**	0.006	14
N–IG	Gamma	$(4, 1)$	-0.217	**0.275**	2
		$(0.5, 0.5)$	-0.213	0.233	5
	Log.N.	$(4, 1)$	-0.210	0.065	5
		$(0.5, 0.5)$	-0.208	0.048	8

Table 2

Enzyme dataset: Summaries of log-conditional predictive ordinates [average (ALCPO) and median (MLCPO)] and mode of the posterior distribution of the number of components in the mixture, $R_n \mid \boldsymbol{X}$, for different prior specifications. Bold numbers denote best fitting according to the corresponding statistic.

4.2 Simulation Study

We now provide an extensive simulation study and use it also for comparing the performance of NRMI mixtures with other density estimation methods. Marron and Wand (1992) considered a set of 15 densities with different behaviors, which are challenging to estimate. These densities are either unimodal, multimodal, symmetric and/or skewed. According to Marron and Wand (1992) the last 5 densities are strongly multimodal and are difficult to recover with moderate sample sizes. Therefore, we concentrate on their first 10 densities to test the performance of NRMI mixtures. For each of the 10 models, the simulation study was based on $N = 40$ simulation experiments and for each experiment a sample of size $n = 250$ was drawn from the model.

We considered NRMI mixtures with a normal kernel (i) and a N–stable process

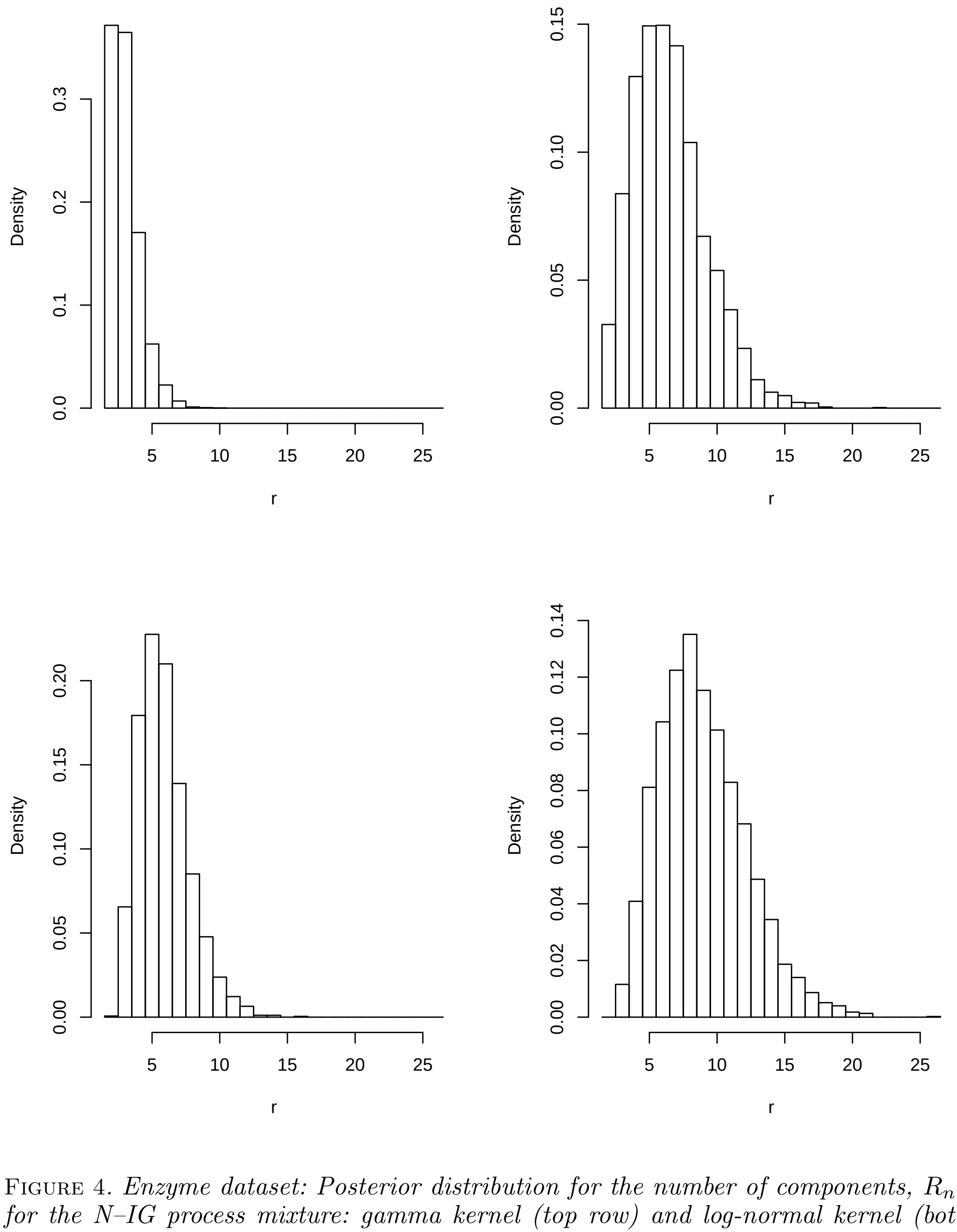

FIGURE 4. *Enzyme dataset: Posterior distribution for the number of components,* $R_n|\boldsymbol{X}$*, for the N–IG process mixture: gamma kernel (top row) and log-normal kernel (bottom row) with* $(\varsigma_1, \varsigma_2) = (4, 1)$ *(left column) and* $(\varsigma_1, \varsigma_2) = (0.5, 0, 5)$ *(right column).*

NGG$(1, 0, 0.396; P_0)$ as mixing measure. This choice of the parameter $\gamma = 0.396$ implies that the a priori expected number of components is equal to 10, which seems a reasonable default choice. As for the base measure P_0, we took $f_0^2(\sigma|\varsigma) =$ ga$(\sigma|1, 1)$, whereas for μ we adopted the normal specification in (a). As for the latter the hyperparameters of the normal-gamma prior on (φ_1, φ_2) are $\psi_1 = 0$, $\psi_2 = 0.01$, $\psi_3 = 0.1$ and $\psi_4 = 0.1$. It is important to note that these prior specifications were the same for all 10 models and, hence, all experiments: the idea is to verify its performance as a default choice rather than tailoring the model on each specific example. As we mentioned at the beginning of the section, since these are simulation experiments, one can compute the relative mean integrated squared error (RMISE) as a measure of goodness of fit. As benchmarking nonparametric kernel density estimator, w.r.t. which the RMISE is computed, we considered the optimal bandwidth given in Silverman (1986) which is $\sigma = s^2(1.06)^2 n^{-2/5}$, with s^2 being the sample variance. For each case the Gibbs sampler was run for 10,000 iterations with a burn-in of 1,000 sweeps and one simulation every 4th was taken for computing the estimates.

Table 3 summarizes the results in terms of RMISE. For comparison purposes we have also included the RMISE obtained by Müller and Vidakovic (1998) using Bayesian wavelets and those obtained by Roeder and Wasserman (1997) using finite mixture of normals. In a private communication, Müller and Vidakovic informed us of a minor problem with the RMISE values originally reported in Müller and Vidakovic (1998): the values in Table 3 are the correct ones obtained from their model. Figure 5 displays the true density (solid line) and the estimated densities resulting from our NRMI mixture (dashed line) and the kernel density estimates with optimal bandwith (dotted line) for models $1 - 10$. The numbers reported in Table 3 and the density estimates in Figure 5 are averages over the 40 experiments.

Model	RMISE		
	MRMI	M&V	R&W
1	0.39	1.99	0.07
2	0.76	0.98	0.34
3	0.18	0.28	2.91
4	0.09	0.25	1.67
5	0.05	0.43	0.44
6	0.81	1.62	0.31
7	0.13	0.38	0.23
8	0.73	1.72	0.74
9	0.86	1.42	0.54
10	0.81	0.83	2.76

TABLE 3

RMISE statistic for the first 10 *models in Marron and Wand (1992): column two displays the RMISE values for the NRMI normal mixture model based on a N–stable process; columns three and four report the RMISE values for the methods of Müller and Vidakovic (1998) and Roeder and Wasserman (1997), respectively.*

From Table 3 we can observe that the approach of Roeder and Wasserman (1997) improves on the kernel density estimator in 7 of the 10 models. In particular, they fail to provide a good fit for those densities that are quite spiky (models

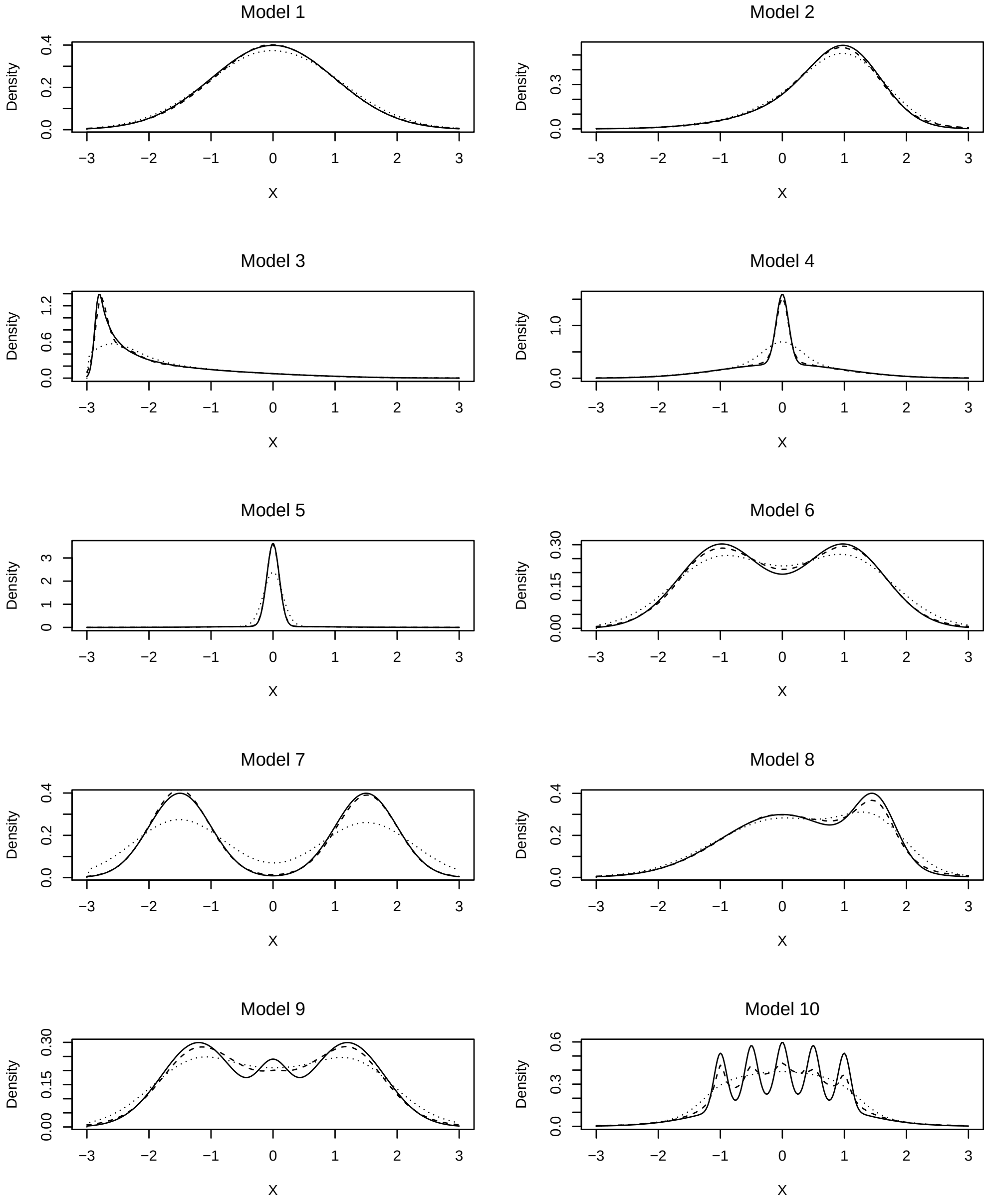

FIGURE 5. *Posterior density estimates for the first 10 models of Marron and Wand (1992): true density (solid line), NRMI normal mixture estimate based on the N–stable process (dashed line), and kernel density estimate with optimal bandwith (dotted line). The estimates have been obtained as averages over the $N = 40$ simulation experiments.*

3, 4 and 10). Also, the wavelets approach of Müller and Vidakovic (1998) have the best behavior precisely for these spiky models producing the smallest RMISE. The NRMI normal mixtures performs significantly better than the kernel density estimator in all 10 models, the highest RMISE being 0.86. This is also apparent in Figure 5. Moreover, it reaches the smallest RMISE in 6 of the 10 models compared to all its competitors. However, rather than focusing on best performances it is important to stress that the estimates yielded by the approaches of R&W and M&V are, in some cases, significantly worse than the kernel density estimator. Hence, NRMI mixtures give the best result in 6 cases (models $3-5$, 7, 8 and 10), but more importantly yield at least second–best results in all the other cases and there is always quite some gap between its RMISE and the one of the worse estimate. In summary, the flexibility of the NRMI mixtures makes it a valuable alternative to more standard methods. In particular, the N–stable mixtures could be considered as a default model, which works reasonably well regardless whether the density is unimodal, multimodal, spiky or flat.

Remark 4. NRMI mixtures with nonparametric specification of both location and scale parameters considered in this section correspond to the `MixNRMI2` function in R–package `BNPdensity`. Additionally, the package also includes semi–parametric NRMI mixtures, in which the location and the scale are modeled, respectively, according to an NRMI and a parametric distribution. Such a specification corresponds to a common value of the smoothing parameter σ for all mixture components and to locations μ_j's generated by the NRMI. This is called `MixNRMI1` function in the package. Extensive simulation studies, not reported here indicate that semi–parametric mixtures are more sensitive w.r.t. wrong prior specifications, in the sense that they tend to get stuck on wrong values for the number of mixture components. Moreover, as one would expect given the lack of flexibility in controlling the dispersion some over–smoothing typically would appear.

Remark 5. Although for comparison purposes it is more convenient to work with simple NRMI mixtures as done here, extensions to more general settings have been provided in the literature. For example, Lijoi, Nipoti and Prünster (2011) define vectors of dependent NRMIs, where the dependence originates from a suitable construction of the underlying Poisson random measures: such models are readily implementable in two–sample problems and meta–analysis. More general regression problems can also be obtained starting from simple NRMI mixtures. For instance, a generalization of the ANOVA dependent Dirichlet process model (DeIorio et al., 2004) to NRMI can be written via the hierarchical representation (22). In the normal case the first equation becomes

$$X_i|\theta_i, Z_i, \sigma \overset{\text{iid}}{\sim} \mathrm{N}(\theta_i' Z_i, \sigma^2),$$

where Z_i is the covariate vector. The second and third equations remain the same together with a prior specification for σ. Suitable modifications of the simulation algorithm, and thus on the `BNPdensity` package, can be implemented to cover this regression case.

ACKNOWLEDGEMENTS

The authors are grateful to an Associate Editor and three Referees for insightful comments and suggestions. The participants of the 2011 Bayesian Nonparametrics Research Programme in Veracruz, Mexico, where a preliminary version of this paper and the corresponding R–package were presented, are also gratefully acknowledged for several stimulating discussions. Special thanks are due to Alejandro Jara, Peter Müller and Steven MacEachern for their helpful suggestions which lead to remarkable improvements of the R–package.

REFERENCES

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

Berry, D.A. and Christensen, R. (1979). Empirical Bayes estimation of a binomial parameter via mixtures of Dirichlet processes. *Ann. Statist.* **7**, 558–568.

Blackwell, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1**, 356–358.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Prob.* **31**, 929–953.

Burden R.L. and Faires, J.D. (1993). *Numerical analysis.* PWS Publishing Company, Boston.

Bush, C.A. and MacEachern, S.N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

Daley, D.J. and Vere-Jones, D. (2008). *An introduction to the theory of point processes. Volume II: General Theory and Structure.* Springer, New York.

Damien, P., Wakefield, J. and Walker, S.G. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *J. R. Stat. Soc. Ser. B* **61**, 331-344.

De Iorio, M., Müller, P., Rosner, G. L. and MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99**, 205-215.

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

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

Favaro, S. and Teh, Y.W. (2012). MCMC for normalized random measure mixture models. *Preprint.*

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

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615–629.

Ferguson, T.S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent advances in statistics*, 287–302, Academic Press, New York.

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

Gelfand, A.E., Dey, D.K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian statistics, 4*, Oxford Univ. Press, New York, 147-167.

Gelfand, A.E. and Sahu, S.K. (1994). On Markov chain Monte Carlo acceleration. *J. Comput. Graph. Statist.* **3**, 261–276.

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

James, L.F., Lijoi, A. and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.* **33**, 105–120.

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

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

Kingman, J.F.C. (1975). Random discrete distributions. *J. Roy. Statist. Soc. Ser. B* **37**, 1–22.

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

Lijoi, A., Mena, R. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *J. Amer. Statist. Assoc.* **100**, 1278–1291.

Lijoi, A., Mena, R.H. and Prünster, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 715–740.

Lijoi, A., Nipoti, B. and Prünster, I. (2011). Bayesian inference with dependent normalized completely random measures. *Preprint.*

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

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

MacEachern, S.N. and Müller, P. (1998). Estimating mixtures of Dirichlet process models. *J. Comput. Graph. Statist.* **7**, 223-238.

MacEachern, S.N. and Müller, P. (2000). Efficient MCMC schemes for robust model extensions using encompassing Dirichlet process mixture models. In *Robust Bayesian Analysis. Lecture Notes in Statist. 152*, 295316. Springer, New York.

Marron, J. S. and Wand, M.P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712–736.

Müller, P. and Quintana, F.A. (2004). Nonparametric Bayesian data analysis, *Statist. Sci.* **19**, 95–110.

Müller, P. and Vidakovic, B. (1998). Bayesian inference with wavelets: Density estimation. *J. Comput. Graph. Statist.* **7**, 456–468.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Can. J. Stat.* **26**, 283–297.

Nieto–Barajas, L.E. and Prünster, I. (2009). A sensitivity analysis for Bayesian nonparametric density estimators. *Statist. Sinica* **19**, 685-705.

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

Orbanz, P. and Williamson, S. (2011). Unit–rate Poisson representations of completely random measures. *Tech. Report.*

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

Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: A Festschrift for Terry Speed* (Ed. Goldstein, D.R.). Lecture Notes, Monograph Series, **40**, 1–35. Institute of Mathematical Statistics, Hayward.

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31**, 560–585.

Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59**, 731-792.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92**, 894–902.

Sato, K. (1990). *Lévy processes and infinitely divisible distributions.* Cambridge University Press, Cambridge.

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

Silverman, B. W. (1986) *Density estimation for statistics and data analysis.* Chapman & Hall, London.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22**, 1701-1762.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simul. Computat.* **36**, 45–54.

Walker, S.G. and Damien, P. (2000). Representations of Lévy processes without Gaussian components. *Biometrika* **87**, 477–483.

Wolpert, R. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–267.