A moment-matching Ferguson & Klass algorithm

Julyan Arbel, Igor Prünster

A moment-matching Ferguson & Klass algorithm*

Julyan Arbel[1,2] **and Igor Prünster**[1]

[1]Bocconi University, Milan, Italy
[2]Collegio Carlo Alberto, Moncalieri, Italy

Abstract. Completely random measures (CRM) represent the key building block of a wide variety of popular stochastic models and play a pivotal role in modern Bayesian Nonparametrics. A popular representation of CRMs as a random series with decreasing jumps is due to Ferguson and Klass (1972). This can immediately be turned into an algorithm for sampling realizations of CRMs or more elaborate models involving transformed CRMs. However, concrete implementation requires to truncate the random series at some threshold resulting in an approximation error. The goal of this paper is to quantify the quality of the approximation by a moment-matching criterion, which consists in evaluating a measure of discrepancy between actual moments and moments based on the simulation output. Seen as a function of the truncation level, the methodology can be used to determine the truncation level needed to reach a certain level of precision. The resulting moment-matching Ferguson & Klass algorithm is then implemented and illustrated on several popular Bayesian nonparametric models.

Key words and phrases: Bayesian Nonparametrics, Completely random measures, Ferguson & Klass algorithm, Moment-matching, Normalized random measures, Posterior sampling.

1. INTRODUCTION

Independent increment processes or, more generally, completely random measures (CRMs) are ubiquitous in modern stochastic modeling and inference. They form the basic building block of countless popular models in, e.g., Finance, Biology, Reliability, Survival Analysis. Within Bayesian nonparametric statistics they play a pivotal role. The Dirichlet process, the cornerstone of the discipline introduced in Ferguson (1973), can be obtained as normalization or exponentiation of suitable CRMs (see Ferguson, 1974). Moreover, as shown in Lijoi and Prünster (2010), CRMs can be seen as the unifying concept of a wide variety of Bayesian nonparametric models. See also Jordan (2010). The concrete implementation of models based on CRMs often requires to simulate their realizations. Given they

*Julyan Arbel, Collegio Carlo Alberto, Via Real Collegio, 30, 10024 Moncalieri, Italy, julyan.arbel@carloalberto.org
Igor Prünster, Department of Decision Sciences, BIDSA and IGIER, Bocconi University, via Roentgen 1, 20136 Milan, Italy, igor@unibocconi.it.
Research supported by the European Research Council (ERC) through StG "N-BNP" 306406.

are discrete infinite objects, $\sum_{i\geq 1} J_i \delta_{Z_i}$, some kind of truncation is required, producing an approximation error $\sum_{i\geq M+1} J_i \delta_{Z_i}$. Among the various representations useful for simulating realizations of CRMs the method due to Ferguson and Klass (1972) and popularized by Walker and Damien (2000) stands out in that, for each realization, the weights J_i's are sampled in decreasing order. This clearly implies that for a given truncation level M the approximation error over the whole sample space is minimized. The appealing feature of decreasing jumps has lead to a huge literature exploiting the Ferguson & Klass algorithm. Limiting ourselves to recall contributions within Bayesian Nonparametrics we mention, among others, Argiento et al. (2016, 2015); Barrios et al. (2013); De Blasi et al. (2010); Epifani et al. (2003); Griffin and Walker (2011); Griffin (2016); Nieto-Barajas and Walker (2002); Nieto-Barajas et al. (2004); Nieto-Barajas and Walker (2004); Nieto-Barajas and Prünster (2009); Nieto-Barajas (2014). General references dealing with the simulation of Lévy processes include Rosiński (2001) and Cont and Tankov (2008), who review the Ferguson & Klass algorithm and the compound Poisson process approximation to a Lévy process.

However, the assessment of the quality of the approximation due to the truncation for general CRMs is limited to some heuristic criteria. For instance, Barrios et al. (2013) implement the Ferguson & Klass algorithm for mixture models by using the so called *relative error* index. The corresponding stopping rule prescribes to truncate when the relative size of an additional jump is below a pre-specified fraction of the sum of sampled jumps. The inherent drawbacks of such a procedure and related heuristic threshold-type procedures employed in the several of the above references is two-fold. On the one hand the threshold is clearly arbitrary without quantifying the total mass of the ignored jumps. On the other hand the total mass of the jumps beyond the threshold, i.e. the approximation error, can be very different for different CRMs or, even, for the same CRM with different parameter values; this implies that the same threshold can produce very different approximation errors in different situations. Starting from similar concerns about the quality of the approximation, the recent paper by Griffin (2016) adopts an algorithmic approach and proposes an adaptive truncation sampler based on sequential Monte Carlo for infinite mixture models based on normalized random measures and on stick-breaking priors. The measure of discrepancy that is used in order to assess the convergence of the sampler is based on the effective sample size (ESS) calculated over the set of particles: the algorithm is run until the absolute value of the difference between two consecutive ESS gets under a pre-specified threshold. Also motivated by the same concerns, Argiento et al. (2016, 2015) adopt an interesting approach to circumvent the problem of truncation by changing the model in the sense of replacing the CRM part of their model with a Poisson process approximation, which having an (almost surely) finite number of jumps can be sampled exactly. However, this leaves the question of the determination of the quality of approximation for truncated CRMs open. Another line of research, originated by Ishwaran and James (2001), is dedicated to validating the trajectories from the point of view of the marginal density of the observations in mixture models. In this context, the quality of the approximation is measured by the L_1 distance between the marginal densities under truncated and non-truncated priors. Recent interesting contributions in this direction include bounds for a Ferguson & Klass representation of the beta process (Doshi

et al., 2009) and bounds for the beta process, the Dirichlet process as well as for arbitrary CRMs in a *size biased representation* (Paisley et al., 2012; Campbell et al., 2015).

This paper faces the problem by a simple yet effective idea. In contrast to the above strategies, our approach takes all jumps of the CRMs into account and hence leads to select truncation levels in a principled way, which vary according to the type of CRM and its parameters. The idea is as follows: given moments of CRMs are simple to compute, one can quantify the quality of the approximation by evaluating some measure of discrepancy between the actual moments of the CRM at issue (which involve all its jumps) and the "empirical" moments, i.e. the moments computed based on the truncated sampled realizations of the CRM. By imposing such a measure of discrepancy not to exceed a given threshold and selecting the truncation level M large enough to achieve the desired bound, one then obtains a set of "validated" realizations of the CRM, or, in other terms, satisfying a moment-matching criterion. An important point to stress is that our validation criterion is all-purpose in spirit since it aims at validating the CRM samples themselves rather than samples of a transformation of the CRM. Clearly the latter type of validation would be ad hoc, since it would depend on the specific model. For instance, with the very same set of moment-matching realizations of a gamma process, one could obtain a set of realizations of the Dirichlet process via normalization and a set gamma mixture hazards by combination with a suitable kernel. Moreover, given moments of transformed CRMs are typically challenging to derive, a moment-matching strategy would not be possible in most cases. Hence, while the quantification of the approximation error does not automatically translate to transformed CRMs, one can still be confident that the moment-matching output at the CRM level produces good approximations. That this is indeed the case is explicitly shown in some practical examples both for prior and posterior quantities in Section 3.

The outline of the paper is as follows. In Sections 2.1-2.2 we recall the main properties of CRMs and provide expressions for their moments. In Sections 2.3-2.4 we describe the Ferguson & Klass algorithm and introduce the measure of discrepancy between moments used to quantify the approximation error due to truncation. Section 3 illustrates the moment-matching Ferguson & Klass algorithm for some popular CRMs and CRM-based Bayesian nonparametric models, namely normalized CRMs and the beta-stable Indian buffet process. Some probabilistic results, discussed in Section 2.3, are given in the Appendix.

2. COMPLETELY RANDOM MEASURES

2.1 Definition and main properties

Let $\mathscr{M}_{\mathbb{X}}$ be the set of boundedly finite measures on $\mathbb{X}$, which means that if $\mu \in \mathscr{M}_{\mathbb{X}}$ then $\mu(A) < \infty$ for any bounded set A. $\mathbb{X}$ is assumed to be a complete and separable metric space and both $\mathbb{X}$ and $\mathscr{M}_{\mathbb{X}}$ are equipped with the corresponding Borel σ-algebras. See Daley and Vere-Jones (2008) for details.

DEFINITION 1. *A random element $\tilde{\mu}$, defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $\mathscr{M}_{\mathbb{X}}$, is called a* completely random measure *(CRM) if, for any collection of pairwise disjoint sets $A_1, \ldots, A_n$ in $\mathbb{X}$, the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are mutually independent.*

An important feature is that a CRM $\tilde{\mu}$ selects (almost surely) discrete measures and hence can be represented as

$$\tilde{\mu} = \sum_{i \geq 1} J_i \delta_{Z_i} \tag{1}$$

where the jumps J_i's and locations Z_i's are random and independent. In (1) and throughout we assume there are no fixed points of discontinuity a priori. The main technical tool for dealing with CRMs is given by their Laplace transform, which admits a simple structural form known as Lévy–Khintchine representation. In fact, the Laplace transform of $\tilde{\mu}(A)$, for any A in $\mathbb{X}$, is given by

$$L_A(u) = \mathbb{E}[\mathrm{e}^{-\lambda \tilde{\mu}(A)}] = \exp\Big\{ - \int_{\mathbb{R}^+ \times A} [1 - \mathrm{e}^{-\lambda v}] \nu(\mathrm{d}v, \mathrm{d}x) \Big\} \tag{2}$$

for any $\lambda > 0$. The measure ν is known as *Lévy intensity* and uniquely characterizes $\tilde{\mu}$. In particular, there corresponds a unique CRM $\tilde{\mu}$ to any measure ν on $\mathbb{R}^+ \times \mathbb{X}$ satisfying the integrability condition

$$\int_B \int_{\mathbb{R}^+} \min\{v, 1\} \nu(\mathrm{d}v, \mathrm{d}x) < \infty \tag{3}$$

for any bounded B in $\mathbb{X}$. From an operational point of view this is extremely useful, since a single measure ν encodes all the information about the jumps J_i's and the locations Z_i's. The measure ν will be conveniently rewritten as

$$\nu(\mathrm{d}v, \mathrm{d}x) = \rho(\mathrm{d}v|x)\alpha(\mathrm{d}x), \tag{4}$$

where ρ is a transition kernel on $\mathbb{R}^+ \times \mathbb{X}$ controlling the jump intensity and α is a measure on $\mathbb{X}$ determining the locations of the jumps. If ρ does not depend on x, the CRM is said homogeneous, otherwise it is non-homogeneous.

We now introduce two popular examples of CRMs that we will serve as illustrations throughout the paper.

EXAMPLE 1. *The generalized gamma process introduced by Brix (1999) is characterized by a Lévy intensity of the form*

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\mathrm{e}^{-\theta v}}{\Gamma(1-\gamma) v^{1+\gamma}} \mathrm{d}v \, \alpha(\mathrm{d}x), \tag{5}$$

whose parameters $\theta \geq 0$ and $\gamma \in [0, 1)$ are such that at least one of them is strictly positive. Notable special cases are: (i) the gamma *CRM which is obtained by setting $\gamma = 0$; (ii) the inverse-Gaussian CRM, which arises by fixing $\gamma = 0.5$; (iii) the* stable *CRM which corresponds to $\theta = 0$. Moreover, such a CRM stands out for its analytical tractability. In the following we work with $\theta = 1$, a choice which excludes the stable CRM. This is justified in our setting because the moments of the stable process do not exist. See Remark 1.*

EXAMPLE 2. *The stable-beta process, or three-parameter beta process, was defined by Teh and Görür (2009) as an extension of the beta process (Hjort, 1990). Its jump sizes are upper-bounded by 1 and its Lévy intensity on $[0, 1] \times \mathbb{X}$ is given by*

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\Gamma(c+1)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} v^{-\sigma-1}(1-v)^{c+\sigma-1} \mathrm{d}v \, \alpha(\mathrm{d}x), \tag{6}$$

where $\sigma \in [0,1)$ is termed discount parameter and $c > -\sigma$ concentration parameter. When $\sigma = 0$, the stable-beta process reduces to the beta CRM of Hjort (1990). Moreover, if $c = 1 - \sigma$, it boils down to a stable CRM where the jumps larger than 1 are discarded.

2.2 Moments of a CRM

For any measurable set A of $\mathbb{X}$, the n-th (raw) moment of $\tilde{\mu}(A)$ is defined by

$$m_n(A) = \mathbb{E}[\tilde{\mu}^n(A)].$$

In the sequel the multinomial coefficient is denoted by $\binom{n}{k_1 \cdots k_n} = \frac{n!}{k_1! \ldots k_n!}$. In the next proposition we collect known results about moments of CRMs which are crucial for our methodology.

PROPOSITION 1. *Let $\tilde{\mu}$ be a CRM with Lévy intensity $\nu(\mathrm{d}v, \mathrm{d}x)$. Then the i-th cumulant of $\tilde{\mu}(A)$, denoted by $\kappa_i(A)$, is given by*

$$\kappa_i(A) = \int_{\mathbb{R}^+ \times A} v^i \nu(\mathrm{d}v, \mathrm{d}x),$$

which, in the homogeneous case $\nu(\mathrm{d}v, \mathrm{d}x) = \rho(\mathrm{d}v)\alpha(\mathrm{d}x)$, simplifies to

$$\kappa_i(A) = \alpha(A) \int_0^\infty v^i \rho(\mathrm{d}v).$$

The n-th moment of $\tilde{\mu}(A)$ is given by

$$m_n(A) = \sum_{(*)} \binom{n}{k_1 \cdots k_n} \prod_{i=1}^{n} \left(\kappa_i(A)/i!\right)^{k_i},$$

where the sum $(*)$ is over all n-tuples of nonnegative integers $(k_1, \ldots, k_n)$ satisfying the constraint $k_1 + 2k_2 + \cdots + nk_n = n$.

A proof is given in the Appendix.

In the following we focus on (almost surely) finite CRMs i.e. $\tilde{\mu}(\mathbb{X}) < \infty$. This is motivated by the fact that most Bayesian nonparametric models, but also models in other application areas, involve finite CRMs. Hence, we assume that the measure α in (3) is finite i.e. $\alpha(\mathbb{X}) := a \in (0, \infty)$. This is a sufficient condition for $\tilde{\mu}(\mathbb{X}) < \infty$ in the non-homogeneous case and also necessary in the homogeneous case (see e.g. Regazzini et al., 2003). A common useful parametrization of α is then given as aP^* with P^* a probability measure and a a finite constant. Note that, if $\tilde{\mu}(\mathbb{X}) = \infty$, one could still identify a bounded set of interest A and the whole following analysis carries over by replacing $\tilde{\mu}(\mathbb{X})$ with $\tilde{\mu}(A)$.

As we shall see in Section 2.3, the key quantity for evaluating the truncation error is given by the random total mass of the CRM, $\tilde{\mu}(\mathbb{X})$. Proposition 1 shows how the moments $m_n = m_n(\mathbb{X})$ can be obtained from the cumulants $\kappa_i = \kappa_i(\mathbb{X})$ and, in particular, the relations between the first four moments and the cumulants are

$$m_1 = \kappa_1,\ m_2 = \kappa_1^2 + \kappa_2,\ m_3 = \kappa_1^3 + 3\kappa_1\kappa_2 + \kappa_3,\ m_4 = \kappa_1^4 + 6\kappa_1^2\kappa_2 + 4\kappa_1\kappa_3 + 3\kappa_2^2 + \kappa_4.$$

CRM	Cumulants	Moments			
	κ_i	m_1	m_2	m_3	m_4
G	$a(i-1)!$	a	$a_{(2)}$	$a_{(3)}$	$a_{(4)}$
IG	$a(1/2)_{(i-1)}$	a	$a^2+\frac{1}{2}a$	$a^3+\frac{3}{2}a^2$ $+\frac{3}{4}a$	a^4+3a^3 $+\frac{15}{4}a^2+\frac{15}{8}a$
GG	$a(1-\gamma)_{(i-1)}$	a	$a^2+a_{(1-\gamma)}$	$a^3+3a^2{}_{(1-\gamma)}$ $+a_{(1-\gamma)_{(2)}}$	$a^4+6a^3{}_{(1-\gamma)}$ $+a^2{}_{(1-\gamma)(11-7\gamma)}+a_{(1-\gamma)_{(3)}}$
B	$a\frac{(i-1)!}{(c+1)_{(i-1)}}$	a	$a^2+\frac{a}{c+1}$	$a^3+\frac{3a^2}{c+1}$ $+\frac{2a}{(c+1)_{(2)}}$	$a^4+\frac{6a^3}{c+1}+\frac{8a^2}{(c+1)_{(2)}}$ $+\frac{3a^2}{(c+1)^2}+\frac{6a}{(c+1)_{(3)}}$
SB	$a\frac{(1-\sigma)_{(i-1)}}{(c+1)_{(i-1)}}$	a	$a^2+a\frac{1-\sigma}{c+1}$	$a^3+3a^2\frac{1-\sigma}{c+1}$ $+a\frac{(1-\sigma)_{(2)}}{(c+1)_{(2)}}$	$a^4+6a^3\frac{1-\sigma}{c+1}+4a^2\frac{(1-\sigma)_{(2)}}{(c+1)_{(2)}}$ $+3a^2\frac{(1-\sigma)^2}{(c+1)^2}+a\frac{(1-\sigma)_{(3)}}{(c+1)_{(3)}}$

TABLE 1

Cumulants and first four moments of the random total mass $\tilde{\mu}(\mathbb{X})$ for the gamma (G), inverse-Gaussian (IG), generalized gamma (GG), beta (B) and stable-beta (SB) CRMs.

With reference to the two examples considered in Section 2.1, in both cases the expected value of $\tilde{\mu}(\mathbb{X})$ is a, which explains the typical terminology *total mass parameter* attributed to a. For the generalized gamma CRM the variance is given by $\text{Var}(\tilde{\mu}(\mathbb{X}))=a(1-\gamma)$, which shows how the parameter γ affects the variability. Moreover, $\kappa_i = a(1-\gamma)_{(i-1)}$ with $x_{(k)} = x(x+1)\dots(x+k-1)$ denoting the ascending factorial. As for the stable-beta CRM, we have $\text{Var}(\tilde{\mu}(\mathbb{X})) = a\frac{1-\sigma}{c+1}$ with both discount and concentration parameter affecting the variability, and also $\kappa_i = a\frac{(1-\sigma)_{(i-1)}}{(1+c)_{(i-1)}}$. Table 1 summarizes the cumulants κ_i and moments m_n for the random total mass $\tilde{\mu}(\mathbb{X})$ for the generalized gamma (assuming as in Example 1 $\theta = 1$), stable-beta CRMs and some of their special cases.

REMARK 1. The *stable* CRM, which can be derived from the generalized gamma CRM by setting $\theta = 0$, does not admit moments. Hence, it cannot be included in our moment-matching methodology. However, the *stable* CRM with jumps larger than 1 discarded, derived from the stable-beta process by setting $c = 1-\sigma$, has all moments. Moreover, even when working with the standard stable CRM, posterior quantities typically involve an exponential updating of the Lévy intensity (see Lijoi and Prünster, 2010), which makes the corresponding moments finite. This then allows to apply the moment matching methodology to the posterior.

2.3 Ferguson & Klass algorithm

For notational simplicity we present the Ferguson & Klass algorithm for the case $\mathbb{X} = \mathbb{R}$. However, note that it can be readily extended to more general Euclidean spaces (see e.g. Orbanz and Williamson, 2012). Given a CRM

$$\tilde{\mu} = \sum_{i=1}^{\infty} J_i \delta_{Z_i}, \tag{7}$$

the Ferguson & Klass representation consists in expressing random jumps J_i occurring at random locations Z_i in terms of the underlying Lévy intensity.
In particular, the random locations Z_i, conditional on the jump sizes J_i, are obtained from the distribution function $F_{Z_i|J_i}$ given by

$$F_{Z_i|J_i}(s) = \frac{\nu(\mathrm{d}J_i, (-\infty, s])}{\nu(\mathrm{d}J_i, \mathbb{R})}.$$

In the case of a homogeneous CRM with Lévy intensity $\nu(\mathrm{d}v, \mathrm{d}x) = \rho(\mathrm{d}v)\, aP^*(\mathrm{d}x)$, the jumps are independent of the locations and, therefore $F_{Z_i|J_i} = F_{Z_i}$ implying that the locations are i.i.d. samples from P^*.
As far as the random jumps are concerned, the representation produces them in decreasing order, that is, $J_1 \geq J_2 \geq \cdots$. Indeed, they are obtained as $\xi_i = N(J_i)$, where $N(v) = \nu([v, \infty), \mathbb{R})$ is a decreasing function, and $\xi_1, \xi_2, \ldots$ are jump times of a standard Poisson process (PP) of unit rate i.e. $\xi_1, \xi_2 - \xi_1, \ldots \overset{\text{i.i.d.}}{\sim} \text{Exp}(1)$. Therefore, the J_i's are obtained by solving the equations $\xi_i = N(J_i)$. In general, this is achieved by numerical integration, e.g., relying on quadrature methods (see, e.g. Burden and Faires, 1993). For specific choices of the CRM, it is possible to make the equations explicit or at least straightforward to evaluate. For instance, if $\tilde{\mu}$ is a generalized gamma process (see Example 1), the function N takes the form

$$N(v) = \frac{a}{\Gamma(1-\gamma)} \int_v^\infty \mathrm{e}^{-u} u^{-(1+\gamma)}\, \mathrm{d}u = \frac{a}{\Gamma(1-\gamma)} \Gamma(v; -\gamma), \tag{8}$$

with $\Gamma(\,\cdot\,;\,\cdot\,)$ indicating an incomplete gamma function. If $\tilde{\mu}$ is the stable-beta process, one has

$$N(v) = a \frac{\Gamma(c+1)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \int_v^1 u^{-\sigma-1}(1-u)^{c+\sigma-1}\, \mathrm{d}u = a \frac{\Gamma(c+1)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} B(1-v; c+\sigma, -\sigma), \tag{9}$$

where $B(\,\cdot\,;\,\cdot\,,\,\cdot\,)$ denotes the incomplete beta function.

Hence, the Ferguson & Klass algorithm can be summarized as follows.

Algorithm 1 Ferguson & Klass algorithm

1: Sample $\xi_i \sim \text{PP}$ for $i = 1, \ldots, M$
2: Define $J_i = N^{-1}(\xi_i)$ for $i = 1, \ldots, M$
3: Sample $Z_i \sim P^*$ for $i = 1, \ldots, M$
4: Approximate $\tilde{\mu}$ by $\sum_{i=1}^M J_i \delta_{Z_i}$

Since it is impossible to sample an infinite number of jumps, approximate simulation of $\tilde{\mu}$ is in order. This becomes a question of determining the number M of jumps to sample in (7) leading to the truncation

$$\tilde{\mu} \approx \tilde{\mu}_M = \sum_{i=1}^M J_i \delta_{Z_i}, \tag{10}$$

with approximation error in terms of the un-sampled jumps equal to $\sum_{i=M+1}^\infty J_i$. The Ferguson & Klass representation has the key advantage of generating the

jumps in decreasing order implicitly minimizing such an approximation error. Then, the natural path to determining the truncation level M would be the evaluation of the Ferguson & Klass tail sum

$$\sum_{i=M+1}^{\infty} N^{-1}(\xi_i). \tag{11}$$

Brix (1999, Theorem A.1) provided an upper bound for (11) in the generalized gamma case. In Proposition 4 of Appendix B we derive also an upper bound for the tail sum of the stable-beta process. However, both bounds are far from sharp and therefore of little practical use as highlighted in Appendix B. This motivates the idea of looking for a different route and our proposal consists in the moment-matching technique detailed in the next section.

2.4 Moment-matching criterion

Our methodology for assessing the quality of approximation of the Ferguson & Klass algorithm consists in comparing the actual distribution of the random total mass $\tilde{\mu}(\mathbb{X})$ with its empirical counterpart, where by empirical distribution we mean the distribution obtained by the sampled trajectories, i.e. by replacing random quantities by Monte Carlo averages of their sampled trajectories. In particular, based on the fact that the first K moments carry much information about a distribution, theoretical and empirical moments of $\tilde{\mu}(\mathbb{X})$ are compared.

The infinite vector of jumps is denoted by $\boldsymbol{J} = (J_i)_{i=1}^{\infty}$ and a vector of jumps sampled by the Ferguson & Klass algorithm by $\boldsymbol{J}^{(l)} = (J_1^{(l)}, \ldots, J_M^{(l)})$. Here, $l = 1, \ldots, N_{\text{FK}}$ stands for the l-th iteration of the algorithm, i.e. for the l-th sampled realization. We then approximate the expectation $\mathbb{E}$ of a statistic of the jumps, say $S(\boldsymbol{J})$, by the following empirical counterpart, denoted by $\mathbb{E}_{\text{FK}}$,

$$\mathbb{E}[S(\boldsymbol{J})] \approx \mathbb{E}_{\text{FK}}[S(\boldsymbol{J})] := \frac{1}{N_{\text{FK}}} \sum_{l=1}^{N_{\text{FK}}} S(\boldsymbol{J}^{(l)}). \tag{12}$$

Note that there are two layers of approximation involved in (12): first, only a finite number of jumps M is used; second, the actual expected value is estimated through an empirical average which typically conveys on Monte Carlo error. The latter is not the focus of the paper, so we take a large enough number of trajectories, $N_{\text{FK}} = 10^4$, in order to insure a limited Monte Carlo error of the order of 0.01. We focus on the first approximation inherent to the Ferguson & Klass algorithm.

More specifically, as far as moments are concerned, $\boldsymbol{m}_K = (m_1, \ldots, m_K)$ denotes the first K moments of the random total mass $\tilde{\mu}(\mathbb{X}) = \sum_{i=1}^{\infty} J_i$ provided in Section 2.2 and $\hat{\boldsymbol{m}}_K = (\hat{m}_1, \ldots, \hat{m}_K)$ indicates the first K empirical moments given by

$$\hat{m}_n = \mathbb{E}_{\text{FK}}\left[\left(\sum_{i=1}^{M} J_i\right)^n\right]. \tag{13}$$

As measure of discrepancy between theoretical and empirical moments, a natural choice is given by the mean squared error between the vectors of moments

or, more precisely, between the n-th roots of theoretical and empirical moments

$$\ell = \ell(\boldsymbol{m}_K, \hat{\boldsymbol{m}}_K) = \left(\frac{1}{K} \sum_{n=1}^{K} \left(m_n^{1/n} - \hat{m}_n^{1/n} \right)^2 \right)^{1/2}. \tag{14}$$

When using the Ferguson & Klass representation for computing the empirical moments the index ℓ depends on the truncation level M and we highlight such a dependence by using the notation ℓ_M. Of great importance is also a related quantity, namely the number of jumps necessary for achieving a given level of precision, which essentially consists in inverting ℓ_M and is consequently denoted by $M(\ell)$.

The index of discrepancy (14) clearly also depends on K, the number of moments used to compute it and $1/K$ in (14) normalizes the indices in order to make them comparable as K varies. A natural question is then about the sensitivity of (14) w.r.t. K. It is desirable for ℓ_M to capture fine variations between the theoretical and empirical distributions, which is assured for large K. In extensive simulation studies not reported here we noted that increasing K in the range $\{1, \ldots, 10\}$ makes the index increase and then plateau and this holds for all processes and parameter specifications used in the paper. Recalling also the whole body of work by Pearson on eponymous curves, which shows that the knowledge of four moments suffices to cover a large number of known distributions, we adhere to his rule of thumb and choose $K = 4$ in our analyses. On the one hand it is a good compromise between targeted precision of the approximation and speed of the algorithm. On the other hand it is straightforward to check the results as K varies in specific applications; for the ones considered in the following sections the differences are negligible.

In the literature several heuristic indices based on the empirical jump sizes around the level of truncation have been discussed (cf Remark 3 in Barrios et al., 2013). Here, in order to compare such procedures with our moment criterion, we consider the relative error index which is based on the jumps themselves. It is defined as the expected value of the relative error between two consecutive partial sums of jumps. Its empirical counterpart is denoted by e_M and given by

$$e_M = \mathbb{E}_{\mathrm{FK}}\left[\frac{J_M}{\sum_{i=1}^{M} J_i} \right]. \tag{15}$$

3. APPLICATIONS TO BAYESIAN NONPARAMETRICS

In this section we concretely implement the proposed moment-matching Ferguson & Klass algorithm to several Bayesian nonparametric models. The performance in terms of both a priori and a posteriori approximation is evaluated. A comparison of the quality of approximation resulting from using (15) as benchmark index is provided.

3.1 A priori simulation study

We start by investigating the performance of the proposed moment-matching version of the Ferguson & Klass algorithm w.r.t. the CRMs defined in Examples 1 and 2, namely the generalized gamma and stable-beta processes. Figure 1 displays the behaviour of both the moment-matching distance ℓ_M (left panel) and the relative jumps' size index e_M (right panel) as the truncation level M increases.

The plots, from top to bottom, correspond to: the generalized gamma process with varying γ and $a = 1$ fixed; the inverse-Gaussian process with varying total mass a (which is a generalized gamma process process with $\gamma = 0.5$); the stable-beta process with varying discount parameter σ and $a = 1$ fixed.

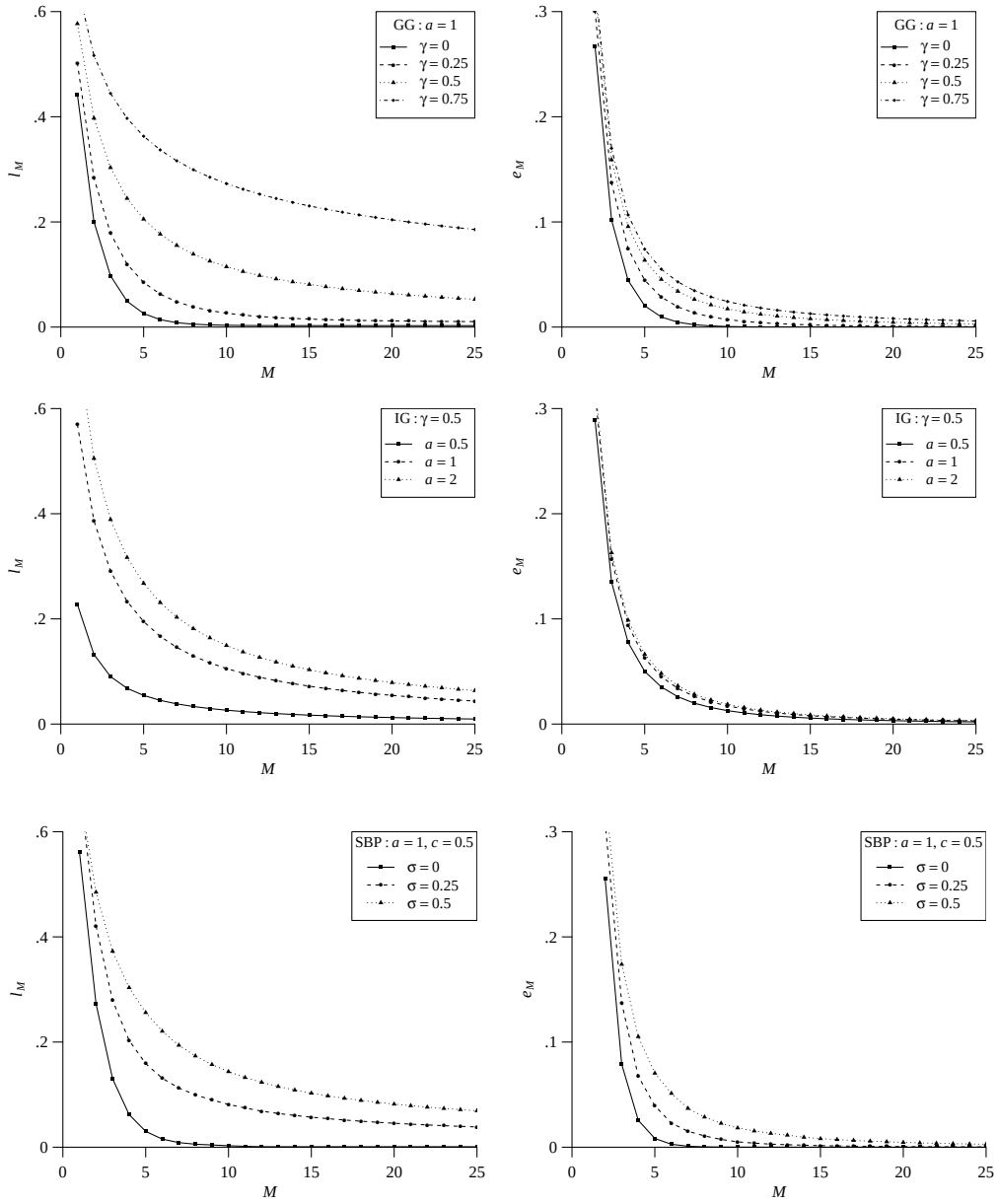
First consider the behaviour of the indices as the parameter specifications vary. It is apparent that, for any fixed truncation level M, the indices ℓ_M and e_M increase as each of the parameters a, γ or σ increases. For instance, roughly speaking, a total mass parameter a corresponds to sampling trajectories defined on the interval $[0, a]$ (see Regazzini et al., 2003), and a larger interval worsens the quality of approximation for any given truncation level. Also it is natural that γ and σ impact in similar way ℓ_M and e_M given they stand for the "stable" part of the Lévy intensity. See first and third rows of Figure 1.

As far as the comparison between ℓ_M and e_M is concerned, it is important to note that e_M consistently downplays the error of approximation related to the truncation. This can be seen by comparing the two columns of Figure 1. ℓ_M is significantly more conservative than e_M for both the generalized gamma and the stable-beta processes, especially for increasing values of the parameters γ, a or σ. This indicates quite a serious issue related to e_M as a measure for the quality of approximation and one should be cautious when using it. In contrast, the moment-matching index ℓ_M matches more accurately the known behaviour of these processes as the parameters vary.

By reversing the viewpoint and looking at the truncation level $M(\ell)$ needed for achieving a certain error of approximation ℓ in terms of moment-match, the results become even more intuitive. We set $\ell = 0.1$ and computed $M(\ell)$ on a grid of size 20×20 with equally-spaced points for the parameters $(a, \gamma) \in (0, 2) \times (0, 0.8)$ for the generalized gamma process and $(a, c) \in (0, 2) \times (0, 30)$ for the beta process. Figure 2 displays the corresponding plots. In general, it is interesting to note that a limited number of jumps is sufficient to achieve good precision levels. Analogously to Figure 1, larger values of the parameters require a larger number of jumps to achieve a given precision level. In particular, when $\gamma > 0.5$, one needs to sample a significantly larger number of jumps. For instance, in the generalized gamma process case, with $a = 1$, the required number of jumps increases from 28 to 53 when passing from $\gamma = 0.5$ to $\gamma = 0.75$. It is worth noting that for the normalized version of the generalized gamma process, to be discussed in Section 3.2 and quite popular in applications, the estimated value of γ rarely exceeds 0.75 in species sampling, whereas it is typically in the range $[0.2, 0.4]$ in mixture modeling.

3.2 Normalized random measures with independent increments

Having illustrated the behaviour of the moment-matching methodology for plain CRMs we now investigate it on specific classes of nonparametric priors, which typically involve a transformation of the CRM. Moreover, given their posterior distributions involve updated CRMs it is important to test the moment-matching Ferguson & Klass algorithm also on posterior quantities. The first class of models we consider are normalized random measures with independent increments (NRMI) introduced by Regazzini et al. (2003). Such nonparametric priors

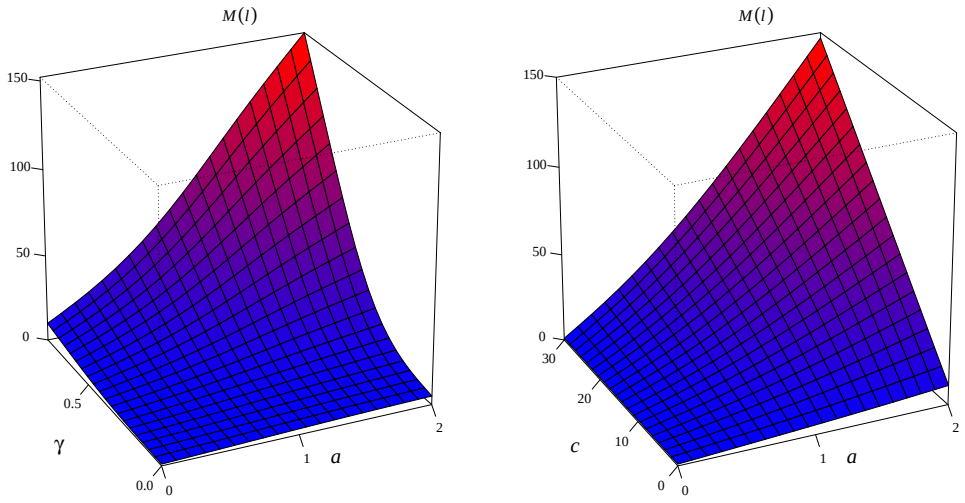


(a) ℓ_M (b) e_M

Fig 1: Left panel: ℓ_M as M varies; right panel: e_M as M varies. Top row: generalized gamma process (GG) with varying γ and $a = 1$ fixed; middle row: inverse-Gaussian process (IG), $\gamma = 0.5$, with varying total mass a; bottom row: stable-beta process (SBP) with $a = 1$, $c = 0.5$ fixed and varying discount parameter σ. The points are connected by straight lines only for visual simplification.

have been used as ingredients of a variety of models and in several application contexts. Recent reviews can be found in Lijoi and Prünster (2010); Barrios et al. (2013).

If $\tilde{\mu}$ is a CRM with Lévy intensity (4) such that $0 < \tilde{\mu}(\mathbb{X}) < \infty$ (almost surely),



(a) generalized gamma process (b) beta process

Fig 2: Number of jumps $M(\ell)$ required to achieve a precision level of $\ell = 0.1$ for ℓ_M. Left panel: generalized gamma process for $a \in (0, 2)$ and $\gamma \in (0, 0.8)$. Right panel: beta process for $a \in (0, 2)$ and $c \in (0, 30)$.

then an NRMI is defined as

$$\tilde{P} = \frac{\tilde{\mu}}{\tilde{\mu}(\mathbb{X})}. \tag{16}$$

Particular cases of NRMI are then obtained by specifying the CRM in (16). For instance, by picking the generalized gamma process defined in Example 1 one obtains the normalized generalied gamma process, denoted by NGG, and first used in a Bayesian context by Lijoi et al. (2007).

3.2.1 Posterior Distribution of an NRMI The basis of any Bayesian inferential procedure is represented by the posterior distribution. In the case of NRMIs, 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 et al. (2006). Nonetheless, a posterior characterization has been established in James et al. (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 an NRMI coincides with the distribution of an 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 posterior characterization to be used with our algorithm, we need to introduce some notation and basic facts. Let $(Y_n)_{n\geq 1}$ be an exchangeable

sequence directed by an NRMI, i.e.

$$
\begin{aligned}
&Y_i|\tilde{P} \overset{\text{i.i.d.}}{\sim} \tilde{P}, \qquad \text{for } i = 1, \dots, n, \\
&\tilde{P} \sim Q,
\end{aligned}
\tag{17}
$$

with Q the law of NRMI, and set $\mathbf{Y} = (Y_1, \dots, Y_n)$. Due to the discreteness of NRMIs, ties will appear with positive probability in $\mathbf{Y}$ and, therefore, the sample information can be encoded by the $K_n = k$ distinct observations $(Y_1^*, \dots, Y_k^*)$ with frequencies $(n_1, \dots, n_k)$ such that $\sum_{j=1}^k n_j = n$. Moreover, introduce the nonnegative random variable U such that the distribution of $[U|\mathbf{Y}]$ has density, w.r.t. the Lebesgue measure, given by

$$
f_{U|\mathbf{Y}}(u) \propto u^{n-1} \exp\{-\psi(u)\} \prod_{j=1}^{k} \tau_{n_j}(u|Y_j^*),
\tag{18}
$$

where $\tau_{n_j}(u|Y_j^*) = \int_0^\infty v^{n_j} \mathrm{e}^{-uv} \rho(\mathrm{d}v|Y_j^*)$ and ψ is the Laplace exponent of $\tilde{\mu}$ defined by $\psi(u) = -\log\left(L_{\mathbb{X}}(u)\right)$, cf (2). Finally, assume $P^* = \mathbb{E}[\tilde{P}]$ to be nonatomic.

Proposition 2 (James et al., 2009). *Let $(Y_n)_{n\geq 1}$ be as in (17) where $\tilde{P}$ is an NRMI defined in (16) with Lévy intensity as in (4). Then the posterior distribution of the unnormalized CRM $\tilde{\mu}$, given a sample $\mathbf{Y}$, is a mixture of the distribution of $[\tilde{\mu}|U, \mathbf{Y}]$ with respect to the distribution of $[U|\mathbf{Y}]$. The latter is identified by (18), whereas $[\tilde{\mu}|U = u, \mathbf{Y}]$ is equal in distribution to a CRM with fixed points of discontinuity at the distinct observations Y_j^*,*

$$
\tilde{\mu}^* + \sum_{j=1}^{k} J_j^* \delta_{Y_j^*}
\tag{19}
$$

such that:

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

$$
\nu^*(\mathrm{d}v, \mathrm{d}x) = \mathrm{e}^{-uv} \nu(\mathrm{d}v, \mathrm{d}x),
\tag{20}
$$

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

$$
f_j^*(v) \propto v^{n_j} \mathrm{e}^{-uv} \rho(\mathrm{d}v|Y_j^*),
\tag{21}
$$

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

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

$$
[\tilde{P}|U, \mathbf{Y}] \overset{d}{=} w \frac{\tilde{\mu}^*}{\tilde{\mu}^*(\mathbb{X})} + (1 - w) \frac{\sum_{k=1}^k J_j^* \delta_{Y_j^*}}{\sum_{l=1}^k J_l^*},
\tag{22}
$$

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

In order to simplify the notation, in the statement we have omitted explicit reference to the dependence on $[U|\mathbf{Y}]$ of both $\tilde{\mu}^*$ and $\{J_j^* : j = 1, \ldots, k\}$, which is apparent from (20) and (21). A nice feature of the posterior representation of Proposition 2 is that the only quantity needed for deriving explicit expressions for particular cases of NRMI is the Lévy intensity (4). For instance, in the case of the generalized gamma process, the CRM part $\tilde{\mu}^*$ in (19) is still a generalized gamma process characterized by a Lévy intensity of the form of (5)

$$\nu^*(\mathrm{d}v, \mathrm{d}y) = \frac{\mathrm{e}^{-(1+u)v}}{\Gamma(1-\gamma)v^{1+\gamma}}\,\mathrm{d}v\, aP^*(\mathrm{d}y). \tag{23}$$

Moreover, the distribution of the jumps (21) corresponding to the fixed points of discontinuity Y_j^*'s in (19) reduces to a gamma distribution with density

$$f_j^*(v) = \frac{(1+u)^{n_j-\gamma}}{\Gamma(n_j-\gamma)} v^{n_j-\gamma-1}\mathrm{e}^{-(1+u)v}. \tag{24}$$

Finally, the conditional distribution of the non-negative latent variable U given $\mathbf{Y}$ (18) is given by

$$f_{U|\mathbf{Y}}(u) \propto u^{n-1}(u+1)^{k\gamma-n} \exp\left\{-\frac{a}{\gamma}(u+1)^{\gamma}\right\}. \tag{25}$$

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 et al. (2009) for general NRMI and Lijoi et al. (2007) for the subclass of normalized generalized gamma processes. See also Argiento et al. (2016) for another approach to approximate the NGG with a finite number of jumps.

3.2.2 Moment-matching for posterior NRMI From (19) it is apparent that the posterior of the unnormalized CRM $\tilde{\mu}$, conditional on the latent variable U, is composed of the independent sum of a CRM $\tilde{\mu}^*$ and fixed points of discontinuity at the distinct observations Y_j^*. The part which is at stake here is obviously $\tilde{\mu}^*$ for which only approximate sampling is possible. As for the fixed points of discontinuities, they are independent from $\tilde{\mu}^*$ and can be sampled exactly, at least in special cases.

We focus on the case of the NGG process. By (20) the Lévy intensity of $\tilde{\mu}^*$ is obtained by exponentially tilting the Lévy intensity of the prior $\tilde{\mu}$. Hence, the Ferguson & Klass algorithm applies in the same way as for the prior. The sampling of the fixed points jumps is straightforward from the gamma distributions (24). As far as the moments are concerned, key ingredient of our algorithm, the cumulants of $\tilde{\mu}^*$ are equal to $\kappa_i^* = a\frac{(1-\gamma)_{(i-1)}}{(u+1)^{i-\gamma}}$ and the corresponding moments are then obtained via Proposition 1.

Our simulation study is based on a sample of size $n = 10$. Such a small sample size is challenging in the sense that the data provide rather few information and the CRM part of the model is still prevalent. We examine three possible clustering configurations of the observations Y_i^*s: (i) $k = 1$ group, with $n_1 = 10$, (ii) $k = 3$ groups, with $n_1 = 1$, $n_2 = 3$, $n_3 = 6$, and (iii) $k = 10$ groups, with $n_j = 1$ for $j = 1, \ldots, 10$. First let us consider the behaviour of $f_{U|\mathbf{Y}}$, which is illustrated in Figure 3 for $n = 10$ and $k \in \{1, 2, \ldots, 10\}$. It is clear that the smaller the number of clusters, the more $f_{U|\mathbf{Y}}$ is concentrated on small values, and vice versa.

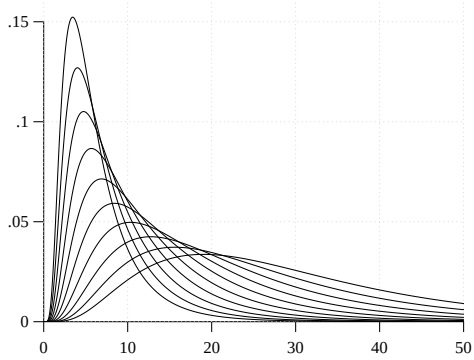


Fig 3: NGG posterior: density $f_{U|\mathbf{Y}}$ with $n = 10$ observations, $a = 1$, $\gamma = 0.5$, and number of clusters $k \in \{1, \ldots, 10\}$; $k = 1$ corresponds to the most peaked density and $k = 10$ to the flattest.

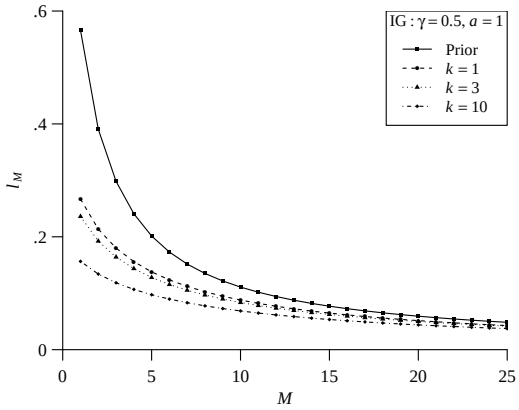
Now we consider $\tilde{\mu}^*(\mathbb{X})$, the random total mass corresponding to the CRM part of the posterior only given in (22). Such a quantity depends on U whose distribution is driven by the data $\mathbf{Y}$. In order to keep the presentation as neat as possible, and in the same time to remain consistent with the data, we choose to condition on $U = u$ for u equal to the mean of $f_{U|\mathbf{Y}}$, the most natural representative value. Given this, it is possible to run the Ferguson & Klass algorithm on the CRM part $\tilde{\mu}^*$ of the posterior and compute moment-matching index ℓ_M as the number of jumps varies. Figure 4 shows these results for the inverse-Gaussian CRM, a special case of the generalized gamma process corresponding to $\gamma = 0.5$. Such posteriors were sampled under the above mentioned $\mathbf{Y}$ clustering configuration scenarios (i)-(iii), which led to mean values of $U|\mathbf{Y}$ of, respectively, 6.3, 8.9 and 25.1. The plot also displays a comparison to the prior values of ℓ_M and indicates that for a given number of jumps the approximation error, measured in terms of ℓ_M, is smaller for the posterior CRM part $\tilde{\mu}^*$ w.r.t. to the prior CRM $\tilde{\mu}$.

Additionally, instead of considering only the CRM part $\tilde{\mu}^*$ of the posterior, one may be interested in the quality of the full posterior which includes also the fixed discontinuities. For this purpose we consider an index which is actually of interest in its own. In particular, we evaluate the relative importance of the CRM part w.r.t. the part corresponding to the fixed points of discontinuity in terms of the ratio $\mathbb{E}(\sum_{j=1}^{k} J_j^*)/\mathbb{E}(\tilde{\mu}^*(\mathbb{X}))$. Loosely speaking one can think of the numerator as the expected weight of the data and the denominator as the expected weight of the prior. Recall that in the NGG case, for a given pair (n, k) and conditional on $U = u$, the sum of fixed location jumps is a gamma$(n - k\gamma, u + 1)$. Hence, the index becomes

$$\frac{\mathbb{E}(\sum_{j=1}^{k} J_j^* | U = u)}{\mathbb{E}(\tilde{\mu}^*(\mathbb{X})|U = u)} = \frac{(n - k\gamma)/(u + 1)}{a/(u + 1)^{1-\gamma}} = \frac{n - k\gamma}{a(u + 1)^{\gamma}}. \tag{26}$$

By separately mixing the conditional expected values in (26) over $f_{U|\mathbf{Y}}$ (we use an adaptive rejection algorithm to sample from $f_{U|\mathbf{Y}}$) we obtained the results summarized in the table of Figure 4. We can appreciate that the fixed part

typically overcomes (or is at least of the same order than) the CRM part, a phenomenon which uniformly accentuates as the sample size n increases. Returning to the original problem of measuring the quality of approximation in terms of moment matching, these findings make it apparent that the comparative results of Figure 4 between prior and posterior are conservative. In fact, if performing the moment-match on the whole posterior, i.e. including the fixed jumps which can be sampled exactly, the corresponding moment-matching index would, for any given truncation level M, indicate a better quality of approximation w.r.t. the index based solely on $\tilde{\mu}^*$. Note that computing the moments of $\tilde{\mu}^*(\mathbb{X}) + \sum_{i=1}^{k} J_i$ straightforward given the independence between $\tilde{\mu}^*$ and the fixed jumps J_i's and also among the jumps themselves. From a practical point of view the findings of this section suggest that a given quality of approximation ℓ in terms of moment-match for the prior represents an upper bound for the quality of approximation in the posterior.



	$\mathbb{E}\left(\sum_{j=1}^{k} J_j^*\right)/\mathbb{E}\left(\tilde{\mu}^*(\mathbb{X})\right)$		
$k \setminus n$	10	30	100
1	3.34	7.30	13.50
$\sqrt{n}$	2.65	4.68	6.05
n	0.89	0.98	0.99

Fig 4: Inverse-Gaussian process ($\gamma = 0.5$) with $a = 1$. Left: Moment-matching errors ℓ_M as the number of jumps M varies. ℓ_M corresponding to prior $\tilde{\mu}$ (continuous line) and posterior $\tilde{\mu}^*$ under $\mathbf{Y}$ clustering scenarios (i) (dashed line), (ii) (dotted line), (iii) (dotted-dashed line). Right: Index of relative importance $\mathbb{E}(\sum_{j=1}^{k} J_j^*)/\mathbb{E}(\tilde{\mu}^*(\mathbb{X}))$ for varying (n, k).

3.2.3 A note on the inconsistency for diffuse distributions In the context of Gibbs-type priors, of which the normalized generalized gamma process is a special case, De Blasi et al. (2012) showed that, if the data are generated from a "true" P_0, the posterior of $\tilde{P}$ concentrates at a point mass which is the linear combination

$$bP^*(\cdot) + (1-b)P_0(\cdot)$$

of the prior guess $P^* = \mathbb{E}(\tilde{P})$ and P_0. The weight b depends on the prior and, indirectly, on P_0, since P_0 dictates the rate at which the distinct observations k are generated. For a diffuse P_0, all observations are distinct and $k = n$ (almost surely). In the NGG case this implies that $b = \gamma$ and hence the posterior is inconsistent since it does not converge to P_0. For the inverse-Gaussian process, i.e. with $\gamma = 0.5$, the posterior distribution gives asymptotically the same weight to P^* and P_0. The last row of the table of Figure 4, which displays the ratio

$\mathbb{E}(\sum_{j=1}^{k} J_j^*)/\mathbb{E}(\tilde{\mu}^*(\mathbb{X}))$ for $k = n$, is an illustration of this inconsistency result since the ratio gets close to 1 as n grows. In contrast, when P_0 is discrete, which implies that k increases at a slower rate than n, one always has consistency. This is illustrated by the first two rows of the table of Figure 4, where one can appreciate that the ratio $\mathbb{E}(\sum_{j=1}^{k} J_j^*)/\mathbb{E}(\tilde{\mu}^*(\mathbb{X}))$ increases as n increases, giving more and more weight to the data. These findings suggest that consistency issues for general NRMI could be explored from new perspectives based on the study of the asymptotic behavior of $f_{U|\mathbf{Y}}$, which will be subject to future work.

3.3 Stable-beta Indian buffet process

The Indian buffet process (IBP), introduced in Ghahramani and Griffiths (2005), is one of the most popular models for feature allocation and is closely connected to the beta process discussed in Example 2. In fact, when marginalizing out the Dirichlet process and considering the resulting partition distribution one obtains the well known Chinese restaurant process. Likewise, as shown in Thibaux and Jordan (2007), when integrating out a beta process in a Bernoulli process (BeP) model one obtains the IBP. Recall that a Bernoulli process, with an atomic base measure $\tilde{\mu}$, is a stochastic process whose realizations are collections of atoms of mass 1, with possible locations given by the atoms of the base measure $\tilde{\mu}$. Such an atom is element of the collection with probability given by the jump size in $\tilde{\mu}$. Later, Teh and Görür (2009) generalized the construction and defined the stable-beta Indian buffet process as

$$\begin{aligned} Y_i|\tilde{\mu} &\overset{\text{i.i.d.}}{\sim} \text{BeP}(\tilde{\mu}) \quad \text{for } i = 1, \dots, n, \\ \tilde{\mu}|c, \sigma, aP^* &\sim \text{SBP}(c, \sigma, aP^*). \end{aligned} \tag{27}$$

Given the construction involves a CRM, it is clear that any conditional simulation algorithm will need to rely on some truncation for which we use our moment-matching Ferguson & Klass algorithm.

3.3.1 Posterior distribution in the IBP Let us consider a conditional iid sample $\mathbf{Y} = (Y_1, \dots, Y_n)$ as in (27). Note that due to the discreteness of $\tilde{\mu}$, ties appear with positive probability. We adopt the same notations for the ties Y_j^* and frequencies n_j as in Section 3.2. Then we can state the following result which highlights the posterior structure of the stable-beta process in the Indian buffet process.

Proposition 3 (Teh and Görür, 2009). *Let $(Y_n)_{n\geq 1}$ be as in (27). Then the posterior distribution of $\tilde{\mu}$ conditional on $\mathbf{Y}$ is given by the distribution of*

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

where

(a) $\tilde{\mu}^*$ is a stable-beta process characterized by the Lévy intensity

$$\nu^*(\mathrm{d}v, \mathrm{d}x) = (1 - v)^n \nu(\mathrm{d}v, \mathrm{d}x),$$

(b) the jump height J_j^* corresponding to Y_j^* is beta distributed

$$J_j^* \sim \text{beta}(n_j - \sigma, c + \sigma + n - n_j),$$

(c) $\tilde{\mu}^*$ *and* J_j^*, $j = 1, \ldots, k$, *are independent.*

Note that due to the polynomial tilting of ν by $(1-u)^n$ in *(a)* above, the CRM part $\tilde{\mu}^*$ is still a stable-beta process with updated parameters

$$c^* = c + n \text{ and } a^* = a\frac{(c+\sigma)_{(n)}}{(c+1)_{(n)}},$$

while the discount parameter σ remains unchanged.

3.3.2 Moment-matching for the IBP In order to implement the moment-matching methodology we first need to evaluate the posterior moments of the random total mass. For this purpose, we rely on the moments characterization in terms of the cumulants provided in Proposition 1. The cumulants κ_i^* of the CRM part $\tilde{\mu}^*(\mathbb{X})$ are obtained from Table 1 with the appropriate parameter updates which leads to

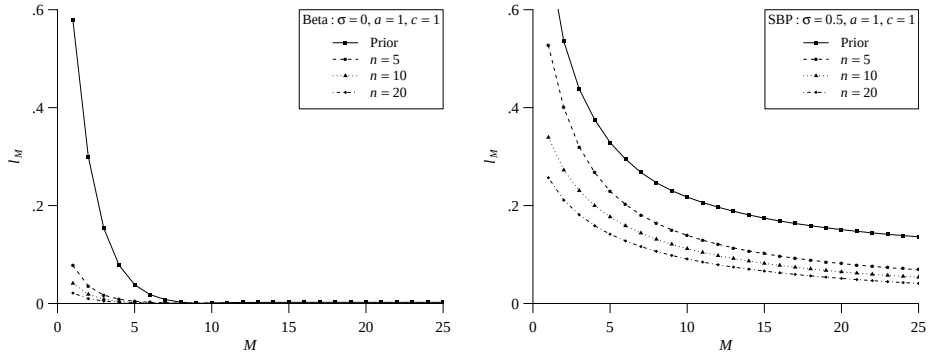
$$\kappa_i^* = a^*\frac{(1-\sigma)_{(i-1)}}{(1+c^*)_{(i-1)}} = a\frac{(1-\sigma)_{(i-1)}(c+\sigma)_{(n)}}{(1+c)_{(n+i-1)}}.$$

We consider two stable-beta processes: the beta process prior $\tilde{\mu} \sim \text{SBP}(c = 1, \sigma = 0, a = 1)$ and the stable-beta process prior $\tilde{\mu} \sim \text{SBP}(c = 1, \sigma = 0.5, a = 1)$. We let n vary in $\{5, 10, 20\}$. In contrast to the NRMI case, there is no need to work under different scenarios for the clustering profile of the data, since the posterior CRM $\tilde{\mu}^*$ is not affected by them with only the sample size entering the updating scheme. We compare the prior moment-match for $\tilde{\mu}$ with the posterior moment-match for $\tilde{\mu}^*$ in terms of our discrepancy index ℓ_M and the results are displayed in Figure 5. The comparison shows that there is a gain in precision between prior and posterior distributions in terms of ℓ_M suggesting that the a priori error level ℓ represents an upper bound for the posterior approximation error.

As in Section 3.2, we also evaluate the relative weights of fixed jumps and posterior CRM or, roughly, of the data w.r.t. the prior. Recalling that fixed location jumps J_j^* are independent and beta$(n_j - \sigma, c + \sigma + n - n_j)$ and some algebra allow to re-write the ratio of interest as

$$\frac{\mathbb{E}(\sum_{j=1}^k J_j^*)}{\mathbb{E}(\tilde{\mu}^*(\mathbb{X}))} = \frac{(n - k\sigma)(c+1)_{(n-1)}}{a(c+\sigma)_{(n)}}.$$

Table 2 displays the corresponding values for different choices of n and k. As in the NRMI case, the fixed part overcomes the CRM part, which means that the data dominate the prior, and, moreover, their relative weight increases as n increases. In terms of moment-matching this shows that, if one looks at the overall posterior structure, the approximation error connected to the truncation is further dampened.



(a) beta process ($\sigma = 0$) (b) stable beta process ($\sigma = 0.5$)

Fig 5: Moment-matching errors ℓ_M as the number of jumps M varies for the stable-beta process with $c = 1$, $a = 1$, and, respectively, $\sigma = 0$ (left panel) and $\sigma = 0.5$ (right panel). ℓ_M corresponding to prior $\tilde{\mu}$ (continuous line) and the posterior $\tilde{\mu}^*$ given with $n = 5$ (dashed line) and $n = 10$ (dotted line) and $n = 20$ (dashed-dotted line) observations.

	$\mathbb{E}\big(\sum_{j=1}^{k} J_j^*\big)/\mathbb{E}\big(\mu^*(\mathbb{X})\big)$		
$k \quad \backslash \quad n$	10	30	100
1	2.57	4.71	8.79
n^σ	2.28	4.36	8.39
n	1.35	2.40	4.41

TABLE 2

Stable-beta process with $\sigma = 0.5$, $c = 1$ and $a = 1$: Index of relative importance $\mathbb{E}\big(\sum_{j=1}^{k} J_j^*\big)/\mathbb{E}\big(\mu^*(\mathbb{X})\big)$ for varying (n, k).

3.4 Practical use of the moment-matching criterion

We illustrate the use of the moment-matching strategy by implementing it within location-scale NRMI mixture models, which can be represented in hierarchical form as

$$\begin{aligned} Y_i|\mu_i, \sigma_i &\overset{\text{ind}}{\sim} k(\cdot|\mu_i, \sigma_i), \quad i = 1, \ldots, n, \\ (\mu_i, \sigma_i)|\tilde{P} &\overset{\text{i.i.d.}}{\sim} \tilde{P}, \quad i = 1, \ldots, n, \\ \tilde{P} &\sim \text{NRMI}, \end{aligned}$$

where k is a kernel parametrized by $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+$ and the NRMI $\tilde{P}$ is defined in (16). Under this framework, density estimation is carried out by evaluating the posterior predictive density. Specifically, we consider the Gaussian kernel $k(x|\mu, \sigma) = \mathcal{N}(x|\mu, \sigma)$ and NGG on locations and scales with a normal base measure P_0, parameter $\theta = 1$ in Equation (5), and varying stability parameter $\gamma \in \{0, 0.25, 0.5, 0.75\}$.

The dataset we consider is the popular `Galaxy` dataset, which consists of velocities of 82 distant galaxies diverging from our own galaxy. Since the data are

clearly away from zero (range from 9.2 to 34), Gaussian kernels, although having the whole real line as support, are typically employed in its analysis.

As far as the simulation algorithm is concerned, based on Sections 3.1 to 3.3, the following moment-matching Ferguson & Klass posterior sampling strategy is implemented: (1) evaluate the threshold $M(\ell)$ which validates trajectories of the CRM using Algorithm 1 on the prior distribution; (2) implement Algorithm 1 on the posterior distribution using the threshold $M(\ell)$. More elaborate and suitably tailored moment-matching strategies can be devised for specific models. However, to showcase the generality and simplicity of our proposal we do not pursue this here.

In particular, we set $\ell_M = 0.01$. We compare the output to the Ferguson & Klass algorithm with heuristic relative error e_M criterion, which consists of step (2) only with truncation dictated by the relative error for which we set $e_M \in \{0.1, 0.05, 0.01\}$. For both algorithms the Gibbs sampler is run for $20,000$ iterations with a burn-in of $4,000$, thinned by a factor of 5.

In order to compare the results, we compute the Kolmogorov–Smirnov distance $d_{KS}(\hat{F}_{\ell_M}, \hat{F}_{e_M})$ between associated estimated cumulative distribution functions (cdf) $\hat{F}_{\ell_M}$ and $\hat{F}_{e_M}$ under, respectively, the moment-match and the relative error criteria. The results are displayed in Table 3. The estimated cdf $\hat{F}_{\ell_M}$ with $\ell_M = 0.01$ can be seen as a reference estimate since the truncation error is controlled uniformly across the different values of γ by the moment-match at the CRM level. First, one immediately notes that the smaller e_M, the closer the two estimates become (in the d_{KS} distance). Second, and more importantly, the numerical values of the distances heavily depend on the particular choice of the parameter γ for any given e_M. In fact, $\hat{F}_{\ell_M}$ and $\hat{F}_{e_M}$ are significantly further apart for large values of γ than for small ones. This clearly shows that the quality of approximation with the heuristic criterion of the relative index is highly variable in terms of a single parameter; in passing from $\gamma = 0$ to $\gamma = 0.75$ the distance increases by at least a factor of 2. This means that for comparing correctly CRM based models with different parameters one would need to pick different relative indices for each value of the parameter. However, there is no way to guess such thresholds without the guidance of an analytic criterion. And, this already happens by varying a single parameter, let alone when changing CRMs for which the same e_M could imply drastically different truncation errors. This seems quite convincing evidence supporting the abandonment of heuristic criteria for determining the truncation threshold and the adoption of principled approaches such as the moment-matching criterion proposed in this paper.

APPENDIX A: PROOF OF PROPOSITION 1

For any measurable set A of $\mathbb{X}$, the n-th moment of $\tilde{\mu}(A)$, if it exists, is given by $m_n(A) = (-1)^n L_A^{(n)}(0)$, where $L_A^{(n)}(0)$ denotes the n-th derivative of the Laplace transform L_A in (2) evaluated at 0. The result is proved by applying Faà di Bruno's formula to (2) for obtaining the derivatives.

APPENDIX B: EVALUATION OF THE TAIL SUM OF THE STABLE-BETA PROCESS

Here we provide an evaluation of the tail sum (11) in the case of the stable-beta process. We start by stating a lemma useful for upper bounding the tail sum.

γ	$e_M = 0.1$	$e_M = 0.05$	$e_M = 0.01$
0	19.4	15.5	9.2
0.25	31.3	23.7	15.1
0.5	42.4	28.9	18.3
0.75	64.8	41.0	23.2

TABLE 3

`Galaxy` dataset. Kolmogorov–Smirnov distance $d_{KS}(\hat{F}_{\ell_M}, \hat{F}_{e_M})$ between estimated cdfs $\hat{F}_{\ell_M}$ and $\hat{F}_{e_M}$ under, respectively, the moment-match (with $\ell_M = 0.01$) and the relative error (with $e_M = 0.1, 0.05, 0.01$) criteria. The mixing measure of normal mixture is the normalized generalized gamma process with varying $\gamma \in \{0, 0.25, 0.5, 0.75\}$.

LEMMA 1. *Let function $N(\,\cdot\,)$ be as in (9) for the stable-beta process. Then for any $\xi > 0$*

$$N^{-1}(\xi) \leq \begin{cases} \mathrm{e}^{\frac{1-\xi/a}{c}} & \text{if } \sigma = 0, \\ (\alpha\xi + \beta)^{-1/\sigma} & \text{if } \sigma \in (0,1), \end{cases}$$

where $\alpha = \sigma\Gamma(1-\sigma)\frac{\Gamma(c+\sigma)}{a\Gamma(c+1)}$ and $\beta = 1 - \frac{\sigma}{c+\sigma}\Gamma(1-\sigma)$.

PROOF. For $\sigma = 0$, from $u^{-1}(1-u)^{c-1} \leq u^{-1} + (1-u)^{c-1}$ one obtains $\int_v^1 u^{-1}(1-u)^{c-1}\mathrm{d}u \leq 1/c - \log v$. Hence, $N(v)/a \leq 1 - c\log v$ and $N^{-1}(\xi) \leq \mathrm{e}^{(1-\xi/a)/c}$. The argument for $\sigma \neq 0$ follows along the same lines starting from $u^{-1-\sigma}(1-u)^{\sigma+c-1} \leq \Gamma(1-\sigma)u^{-\sigma-1} + (1-u)^{\sigma+c-1}$. □

PROPOSITION 4. *Let $(\xi_j)_{j\geq 1}$ be the jump times for a homogeneous Poisson process on $\mathbb{R}^+$ with unit intensity. Define the tail sum of the stable-beta process as*

$$T_M = \sum_{j=M+1}^{\infty} N^{-1}(\xi_j),$$

where $N(\,\cdot\,)$ is given by (9). Then for any $\epsilon \in (0,1)$,

$$\mathbb{P}\Big(T_M \leq t_M^\epsilon\Big) \geq 1 - \epsilon, \text{ for } t_M^\epsilon = \begin{cases} \frac{C_1}{\epsilon}\mathrm{e}^{\frac{1}{c} - \frac{\epsilon M}{C_1}} & \text{if } \sigma = 0, \\ \frac{\sigma}{1-\sigma}\frac{(C_2\epsilon)^{1/\sigma}}{(M+\beta C_2/\epsilon)^{1/\sigma - 1}} & \text{if } \sigma \in (0,1), \end{cases}$$

where $C_1 = 2ace$ and $C_2 = 2\mathrm{e}/\alpha$ do not depend on ϵ.

PROOF. The proof follows along the same lines as the proof of Theorem A.1. in Brix (1999). Let q_j denote the $\epsilon 2^{M-j}$ quantile, for $j = M+1, M+2, \ldots$, of a gamma distribution with mean and variance equal to j. Then

$$\mathbb{P}\Big(\sum_{j=M+1}^{\infty} N^{-1}(\xi_j) \leq \sum_{j=M+1}^{\infty} N^{-1}(q_j)\Big) \geq 1 - \epsilon.$$

An upper bound on $\tilde{t}^{\epsilon}_M = \sum_{j=M+1}^{\infty} N^{-1}(q_j)$ is then found by resorting to Lemma 1 along with the inequality $q_j \geq \frac{\epsilon}{2\mathrm{e}} j$. If $\sigma = 0$

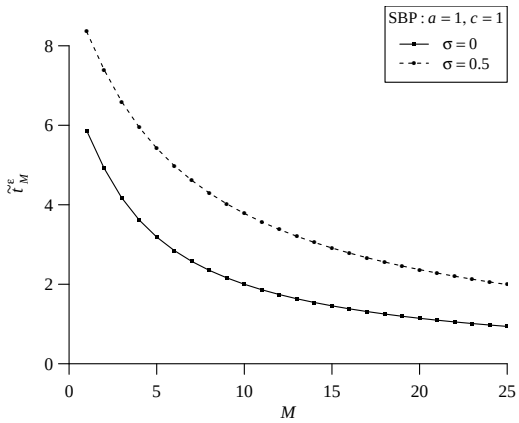
$$\tilde{t}^{\epsilon}_M \leq \mathrm{e}^{1/c} \sum_{j=M+1}^{\infty} \mathrm{e}^{-\frac{q_j}{ac}} \leq \mathrm{e}^{1/c} \sum_{j=M+1}^{\infty} \mathrm{e}^{-\frac{\epsilon j}{2ace}} \leq \mathrm{e}^{1/c} \frac{2ace}{\epsilon} \mathrm{e}^{-\frac{\epsilon M}{2ace}},$$

whereas if $\sigma \neq 0$

$$\tilde{t}^{\epsilon}_M \leq \sum_{j=M+1}^{\infty} (\alpha q_j + \beta)^{-\frac{1}{\sigma}} \leq \sum_{j=M+1}^{\infty} \left(\frac{\alpha \epsilon j}{2\mathrm{e}} + \beta \right)^{-\frac{1}{\sigma}} = \left(\frac{2\mathrm{e}}{\alpha\epsilon} \right)^{-\frac{1}{\sigma}} \sum_{j=M+1}^{\infty} \left(j + \frac{2\mathrm{e}\beta}{\alpha\epsilon} \right)^{-\frac{1}{\sigma}}.$$

The result follows by bounding the last sum by $\int_M^{\infty} \left(x + \frac{2\mathrm{e}\beta}{\alpha\epsilon} \right)^{-\frac{1}{\sigma}} \mathrm{d}x$. □

The bound t^{ϵ}_M obtained in Proposition 4 is exponential when $\sigma = 0$ and polynomial when $\sigma \neq 0$, but it is very conservative as already pointed out by Brix (1999). This finding is further highlighted in the table associated to Figure 6, where the bound t^{ϵ}_M is computed with appropriate constants derived from the proof. In contrast, the bound $\tilde{t}^{\epsilon}_M$ obtained by direct calculation of the quantiles q_j (instead of resorting to a lower bound on them) is much sharper. Figure 6 displays the sharper bound $\tilde{t}^{\epsilon}_M$. Inspection of the plot demonstrates a decrease pattern in this bound in probability which is reminiscent of the ones for the indices ℓ_M and e_M studied in the paper. This observation is a further indication that the Ferguson & Klass algorithm is a tool with well-behaved approximation error.



	M	25	100	500
t^{ϵ}_M	$\sigma = 0$	1411	1230	589
	$\sigma = 0.5$	1554	1250	612
$\tilde{t}^{\epsilon}_M$	$\sigma = 0$	0.942	0.230	0.003
	$\sigma = 0.5$	1.998	0.534	0.008

Fig 6: Stable-beta process with parameters $\sigma = 0$ and $\sigma = 0.5$. Left: Bound in probability $\tilde{t}^{\epsilon}_M$ of the tail sum T_M obtained by direct calculation of the quantiles q_j with $\epsilon = 10^{-2}$ as the truncation level M increases. Right: Bounds t^{ϵ}_M (provided in Proposition 4) and $\tilde{t}^{\epsilon}_M$ (obtained by direct calculation of the quantiles q_j) of the tail sum after M jumps with $\epsilon = 10^{-2}$.

REFERENCES

Argiento, R., Bianchini, I., and Guglielmi, A. (2015). A priori truncation method for posterior sampling from homogeneous normalized completely random measure mixture models. *arXiv preprint arXiv:1507.04528*.

Argiento, R., Bianchini, I., and Guglielmi, A. (2016). A blocked Gibbs sampler for NGG-mixture models via a priori truncation. *Stat. Comput.*, 26(3):641–661.

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statist. Sci.*, 28(3):313–334.

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

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

Campbell, T., Huggins, J., Broderick, T., and How, J. (2015). Truncated completely random measures. In *Bayesian Nonparametrics: The Next Generation (NIPS workshop)*.

Cont, R. and Tankov, P. (2008). *Financial modelling with jump processes*. Chapman & Hall / CRC Press, London.

Daley, D. J. and Vere-Jones, D. (2008). An introduction to the theory of point processes. Vol. II. General theory and structure. Probability and its Applications.

De Blasi, P., Favaro, S., and Muliere, P. (2010). A class of neutral to the right priors induced by superposition of beta processes. *J. Stat. Plan. Inference*, 140(6):1563–1575.

De Blasi, P., Lijoi, A., and Prünster, I. (2012). An asymptotic analysis of a class of discrete nonparametric priors. *Stat. Sin.*, 23:1299–1322.

Doshi, F., Miller, K., Gael, J. V., and Teh, Y. W. (2009). Variational inference for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 137–144.

Epifani, I., Lijoi, A., and Prünster, I. (2003). Exponential functionals and means of neutral-to-the-right priors. *Biometrika*, 90(4):791–808.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.*, 2(4):615–629.

Ferguson, T. S. and Klass, M. J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Stat.*, 43(5):1634–1643.

Ghahramani, Z. and Griffiths, T. L. (2005). Infinite latent feature models and the Indian buffet process. In *Adv. Neur. In.*, pages 475–482.

Griffin, J. E. (2016). An adaptive truncation method for inference in bayesian nonparametric models. *Stat. Comput.*, 26(1-2):423–441.

Griffin, J. E. and Walker, S. G. (2011). Posterior simulation of normalized random measure mixtures. *J. Comput. Graph. Stat.*, 20(1):241–259.

Hjort, N. L. (1990). Nonparametric bayes estimators based on beta processes in models for life history data. *Ann. Statist.*, 18(3):1259–1294.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.*, 96(453):161–173.

James, L. F., Lijoi, A., and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.*, 33(1):105–120.

James, L. F., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.*, 36(1):76–97.

Jordan, M. I. (2010). Hierarchical models, nested models and completely random measures. *Frontiers of Statistical Decision Making and Bayesian Analysis: in Honor of James O. Berger. New York: Springer*, pages 207–218.

Lijoi, A., Mena, R. H., and Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. Roy. Stat. Soc. B Met.*, 69(4):715–740.

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

Nieto-Barajas, L. E. (2014). Bayesian semiparametric analysis of short- and long-term hazard ratios with covariates. *Comput. Stat. Data Anal.*, 71(0):477–490.

Nieto-Barajas, L. E. and Prünster, I. (2009). A sensitivity analysis for Bayesian nonparametric density estimators. *Stat. Sin.*, 19(2):685.

Nieto-Barajas, L. E., Prünster, I., and Walker, S. G. (2004). Normalized random measures driven by increasing additive processes. *Ann. Statist.*, 32(6):2343–2360.

Nieto-Barajas, L. E. and Walker, S. G. (2002). Markov Beta and Gamma Processes for Modelling Hazard Rates. *Scand. J. Statist.*, 29(3):413–424.

Nieto-Barajas, L. E. and Walker, S. G. (2004). Bayesian nonparametric survival analysis via

Lévy driven Markov processes. *Stat. Sin.*, 14(4):1127–1146.

Orbanz, P. and Williamson, S. (2012). Unit-rate poisson representations of completely random measures. *Tech. report*.

Paisley, J. W., Blei, D. M., and Jordan, M. I. (2012). Stick-breaking beta processes and the Poisson process. In *International Conference on Artificial Intelligence and Statistics*, pages 850–858.

Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.*, 31(2):560–585.

Rosiński, J. (2001). Series representations of Lévy processes from the perspective of point processes. In *Lévy processes*, pages 401–415. Springer.

Teh, Y. W. and Görür, D. (2009). Indian buffet processes with power-law behavior. In *Adv. Neur. In.*, pages 1838–1846.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, pages 564–571.

Walker, S. G. and Damien, P. (2000). Miscellanea. Representations of Lévy processes without Gaussian components. *Biometrika*, 87(2):477–483.