The Annals of Statistics
2022, Vol. 50, No. 5, 2641–2663
https://doi.org/10.1214/22-AOS2201

IS INFINITY THAT FAR? A BAYESIAN NONPARAMETRIC PERSPECTIVE OF FINITE MIXTURE MODELS

BY RAFFAELE ARGIENTO[1,a] AND MARIA DE IORIO[2,b]

[1]*Department of Economics, Università degli Studi di Bergamo,* [a]*raffaele.argiento@unibg.it*
[2]*Yong Loo Lin School of Medicine, National University of Singapore,* [b]*mdi@nus.edu.sg*

Mixture models are one of the most widely used statistical tools when dealing with data from heterogeneous populations. Following a Bayesian nonparametric perspective, we introduce a new class of priors: the Normalized Independent Point Process. We investigate the probabilistic properties of this new class and present many special cases. In particular, we provide an explicit formula for the distribution of the implied partition, as well as the posterior characterization of the new process in terms of the superposition of two discrete measures. We also provide consistency results. Moreover, we design both a marginal and a conditional algorithm for finite mixture models with a random number of components. These schemes are based on an auxiliary variable MCMC, which allows handling the otherwise intractable posterior distribution and overcomes the challenges associated with the Reversible Jump algorithm. We illustrate the performance and the potential of our model in a simulation study and on real data applications.

1. Introduction. Mixture models are a very powerful and natural statistical tool to model data from heterogeneous populations. In a mixture model, observations are assumed to have arisen from one of M (finite or infinite) groups, each group being suitably modelled by a density. The density of each group is referred to as a component of the mixture, and is weighted by the relative frequency (weight) of the group in the population. This model offers a conceptually simple way of relaxing distributional assumptions and a convenient and flexible way to approximate distributions that cannot be modelled satisfactorily by a standard parametric family. Moreover, it provides a framework by which observations may be clustered together into groups for discrimination or classification. For a comprehensive review of mixture models and their applications see [14, 36] and [15]. Each observation is assumed to have arisen from one of $0 < M \leq \infty$ groups:

$$f_Y(y \mid P) = \int_\Theta f(y \mid \theta) P(d\theta) = \sum_{j=1}^{M} w_j f(y \mid \tau_j), \tag{1.1}$$

where $\{f(y \mid \theta), \theta \in \Theta \subset \mathbb{R}^d\}$ is a parametric family of densities on $\mathcal{Y}$, while P is an almost sure discrete measure on Θ, and it is referred to as *mixing measure*. Here $\{\tau_j, j = 1, \ldots, M\}$ is a collection of points in Θ, that defines the support of P. For each $j = 1, \ldots, M$, the density $f(y \mid \tau_j)$ is the kernel of the mixture, and is weighted by w_j, the relative frequency of the group in the population. In what follows M will denote the number of components in a mixture, that is, of possible clusters/subpopulations, while by number of clusters, k, we mean the number of allocated components, that is, components to which at least one observation has been assigned. What needs to be highlighted (see [49]) is that even when in a finite mixture model M is fixed, that is, the number of components (possible clusters) of the data

Received October 2019; revised May 2022.
MSC2020 subject classifications. Primary 62F15, 60G57; secondary 62G07, 92C20.
Key words and phrases. Bayesian mixture models, Bayesian clustering, Dirichlet process, mixture of finite mixtures, Markov chain Monte Carlo methods.

generating process is fixed, still we need to estimate k, the actual number of clusters in the sample (allocated components). Already [41] had pointed out this difference, noticing that the posterior distribution of the number of components M might assigns considerable probability to values greater than the number of allocated components.

In a Bayesian parametric framework (i.e., $M < \infty$ almost surely) the most popular approaches are (i) fix M and then focus mainly on density estimation (ii) treat M as a random parameter and make it the focus of inference. See, for instance, [40, 47, 50] and [34, 37] for more details. Although in the Bayesian paradigm there are approaches based on model choice criteria, such as DIC, it is usually preferable to perform full posterior inference on M as well, eliciting an appropriate prior. A fully Bayesian approach is often based on the reversible jump Markov chain Monte Carlo [8, 47] or, alternatively, on the marginal likelihood $p(\boldsymbol{y} \mid M)$. Both methods present significant computational challenges. On the other hand, in Bayesian nonparametrics M is set equal to infinity (i.e., $M = \infty$) and the focus of inference is only k.

In this work, we stress the importance of the distinction between M and k as it allows us to collocate nonparametric and parametric mixtures in exactly the same framework. This is achieved by exploiting the crucial observation by [50] that a finite mixture model is simply a realization of a stochastic process whose dimension is random and has an infinite dimensional support. Extending this approach, we introduce a new class of random measures, *Normalized Independent Finite Point Processes*, obtained by normalization of a point process and use it as mixing measures in Model (1.1). We derive the family of prior distributions induced on the data partition by providing a general formula for exchangeable partition probability functions [42]. The class we propose is rich and includes as particular case the popular finite Dirichlet mixture model. Finally, we characterize the posterior distribution of the Normalized Independent Finite Point Process. Our construction is exactly in the spirit of Bayesian nonparametrics, as it is based on the normalization of a point process, leading to an almost surely discrete measure.

Among the main achievements of this work, there is the construction of two Gibbs sampler schemes, a marginal and a conditional one, to simulate from the posterior distribution of the Normalized Finite Independent Point Process. While the conditional algorithm allows sampling all parameters in the model including the mixing measure, marginal schemes integrate out the random measure and rely on the predictive structure of the process. Both these schemes overcome many of the challenges associated with the Reversible Jump Markov chain Monte Carlo [21, 47] and the marginal algorithm recently proposed by [37]. The latter restricts the class of prior distributions for the weights and limits the analysis to linear functionals of the posterior distribution (see [19], for a discussion of these issues).

The key result (associated to the nonparametric construction of the process) is to be able to propose transdimensional moves which are automatic and naturally implied by the prior process (see [17] for a recent and related contribution).

In Section 2, we introduce the finite mixture model framework, highlighting the connection between parametric and nonparametric constructions. Section 3 introduces a working example which elucidates the main methodological contribution. In Section 4, we introduce the prior process, the Normalised Independent Finite Point Process, and discuss its clustering properties, while in Section 5 we characterise its posterior distribution. In Section 6, we briefly describe how the new prior can be used as a component in more complex hierarchies. In Section 7, we show how the new construction leads to efficient marginal and conditional algorithms. Section 8 provides consistency results. We demonstrate the proposed approach on a benchmark example, the Galaxy data, in Section 9, as well as on an application in population genetics in Section 10. Section 11 concludes the paper.

2. Finite mixture models. Let $Y_1, \ldots, Y_n$ be a set of observations taking values in an Euclidean space $\mathcal{Y}$. Exploiting the latent variable representation of a mixture model, we assume

$$
\begin{aligned}
Y_i|\theta_i &\overset{\text{ind}}{\sim} f(y \mid \theta_i), \quad i = 1, \ldots, n, \\
\theta_i|c_i, \boldsymbol{\tau} &\overset{\text{ind}}{\sim} \delta_{\tau_{c_i}}(d\theta_i), \\
\tau_m \mid M &\overset{\text{iid}}{\sim} P_0(d\tau), \quad m = 1, \ldots, M, \\
c_i \mid M, \boldsymbol{w} &\sim \text{Multinomial}_M(1, w_1, \ldots, w_M), \\
\boldsymbol{w} \mid M &\sim P_W(\boldsymbol{w} \mid M), \quad M \sim q_M,
\end{aligned}
\tag{2.1}
$$

where $f(y \mid \theta)$ is a parametric density on $\mathcal{Y}$, which depends on a vector of parameters θ. Here δ_τ is the Dirac measure assigning unit mass at location τ and $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_M\}$. The vector of parameters τ_m assumes values in $\Theta \subset \mathbb{R}^d$ and is assigned a nonatomic prior density p_0 corresponding to the probability measure P_0 on Θ. The number of components is given a prior q_M. Conditionally on M, the vector of weights $\boldsymbol{w} = (w_1, \ldots, w_M)$, which represents the probability of belonging to each mixture component, is given a prior probability P_W on the simplex of dimension $M - 1$. Finally, $\mathbf{c} = (c_1, \ldots, c_n)$ denote the latent allocation vector whose element c_i denotes to which component observation Y_i is assigned, $c_i \in \{1, \ldots, M\}$. Usually P_W is assumed to be a $\text{Dirichlet}_M(\gamma, \ldots, \gamma)$ distribution, while typical choices for q_M include a discrete uniform on some finite space, a Negative Binomial or a Poisson distribution. In this work, we propose a richer construction, where the prior on $\boldsymbol{w}$ is obtained by normalising a finite point process. Advantages of the proposed approach include: (i) extension of the family of prior distributions for the weights; (ii) full Bayesian inference on all the unknowns (in particular M and $\boldsymbol{w}$); (iii) possibility of inducing sparsity through appropriate choice of hyper-parameters; (iv) ease of interpretation; (v) possibility of extending the construction to covariate dependent weights and (vi) extension to more general processes.

The theoretical developments are based on the key observation that a realization M, $\boldsymbol{w}$, $\boldsymbol{\tau}$ from the prior on the mixture model parameters defined in equation (2.1) in terms of hierarchical parametric distribution q_M, P_W, P_0 defines an almost surely (a.s.) finite-dimensional random probability measure on the parameter space Θ, that is, $P(d\theta) = \sum_{m=1}^{M} w_m \delta_{\tau_m}(d\theta)$. This implies that the joint probability distribution on M, $\boldsymbol{w}$ and $\boldsymbol{\tau}$ induces a distribution on P, whose support is the space of the a.s. finite-dimensional random probability measures on Θ. Moreover, it is straightforward to prove (see [2]) that by letting $\theta_i = \tau_{c_i}$, as in equation (2.1), the variables $\theta_1, \ldots, \theta_n$ can be considered as a sample from P. From this observation, the link between infinite (nonparametric) and finite mixture models becomes evident as the model in equation (2.1) becomes

$$
\begin{aligned}
Y_1, \ldots, Y_n|\theta_1, \ldots, \theta_n &\overset{\text{ind}}{\sim} f(y; \theta_i), \\
\theta_1, \ldots, \theta_n|P &\overset{\text{iid}}{\sim} P, \\
P &= \sum_{m=1}^{M} w_m \delta_{\tau_m}(d\theta) \sim \mathcal{P},
\end{aligned}
\tag{2.2}
$$

where $\mathcal{P}$ is the law of P defined via q_M, P_W, P_0. The main theoretical contribution of this work is to give a constructive definition of $\mathcal{P}$, for which the weights w_m are the normalised jumps of a *finite point process* and the parameters τ_m are defined in terms of realisations of the same point process. Note that [50], while highlighting the connection between finite mixture models and point processes, defines the point process on the complex space of normalized weights. We refer also to Chapter 7 of [15] and Chapter 2 of [14] for further discussion about this link.

3. A simple example. In this section, we illustrate the key ideas of this work using the popular *finite Dirichlet process mixture model* (FDMM):

$$\boldsymbol{w} \mid M \sim \text{Dirichlet}(\gamma, \dots, \gamma), \qquad M - 1 \sim \text{Poisson}(\lambda). \tag{3.1}$$

Let $S_m, m = 1, \dots, M$, be independent random variables with Gamma$(\gamma, 1)$ distribution, and let $T = \sum_{m=1}^{M} S_m$. Then w_m can be represented as S_m/T, that is, the weight vector $\boldsymbol{w}$ can be obtained through normalization of Gamma random variables. To perform Bayesian inference in the FDMM, we need the conditional law of $\boldsymbol{w}$ (or equivalently of $\mathbf{S} = (S_1, \dots, S_M)$) and $\boldsymbol{\tau} = (\tau_1, \dots, \tau_M)$ given a realization of M and $c_1, \dots, c_n$:

$$\begin{aligned} \mathcal{L}(\mathbf{S}, \boldsymbol{\tau} \mid M, c_1, \dots, c_n) &\propto \left(\prod_{i=1}^{n} \frac{S_{c_i}}{T} \right) \prod_{m=1}^{M} \text{Gamma}(dS_m; \gamma, 1) P_0(d\tau_m) \\ &= \frac{1}{T^n} \left(\prod_{i=1}^{n} S_{c_i} \right) \prod_{m=1}^{M} \text{Gamma}(dS_m; \gamma, 1) P_0(d\tau_m). \end{aligned} \tag{3.2}$$

The Gamma random variables are a priori independent as well as the cluster-specific parameter vectors. The main challenges when sampling from (3.2) are given by the fact that the S_m are dependent as they are normalised by their sum T and by the fact that their number M is random. We deal with the first problem by introducing an auxiliary variable U_n, and with the second by a marginalization trick, which requires collapsing $\mathbf{S}$ and $\boldsymbol{\tau}$.

Let $(c_1^\star, \dots, c_k^\star), k \le M$, be the unique values among the $c_1, \dots, c_n$. This implies that some of the M components in the mixture can be empty. Let $M^{(a)} = k$ be the number of allocated components and let $M^{(na)}$ be the number of unallocated components, with $M = M^{(a)} + M^{(na)}$. We denote with the superscripts $^{(a)}$ and $^{(na)}$ the variables corresponding to the allocated and unallocated components respectively, and with $\mathcal{M}^{(na)}$ the set of indices of the unallocated components. This implies that equation (3.2) can be rewritten as

$$\frac{1}{T^n} \left(\prod_{j=1}^{M^{(a)}} S_{c_j^\star}^{n_j} \right) \prod_{j=1}^{M^{(a)}} \text{Gamma}(dS_{c_j^\star}; \gamma, 1) P_0(d\tau_{c_j^\star}) \prod_{m \in \mathcal{M}^{(na)}} \text{Gamma}(dS_m; \gamma, 1) P_0(d\tau_m),$$

where $n_j = \#\{c_i = c_j^\star, i = 1, \dots, n\}$. Finally, we have

$$\frac{1}{T^n} \left(\prod_{j=1}^{M^{(a)}} S_{c_j^\star}^{n_j + \gamma - 1} e^{-S_{c_j^\star}} P_0(d\tau_{c_j^\star}) \, dS_{c_j^\star} \right) \prod_{m \in \mathcal{M}^{(na)}} S_m^{\gamma - 1} e^{-S_m} P_0(d\tau_m) \, dS_m,$$

where a Gamma kernel is recognizable for all the unnormalsied weights. Nevertheless, these variables are dependent with the dependence structure determined by their sum T, $M^{(a)}$ and $M^{(na)}$. Let $U_n \mid T \sim \text{Gamma}(n, T)$, so that

$$\frac{\Gamma(n)}{T^n} = \int_0^\infty e^{-Tu} u^{n-1} \, \mathrm{d}u.$$

The marginal distribution of U_n exists and can be derived by solving the integral $\int_0^\infty \text{Gamma}(u; n, t) \, \text{Gamma}(dt; n\gamma, 1)$. Then we can write the joint law of $\mathbf{S}$, $\boldsymbol{\tau}$, U_n as

$$\begin{aligned} &\mathcal{L}(\mathbf{S}, \boldsymbol{\tau}, U_n \mid M, c_1, \dots, c_n) \\ &\quad \propto e^{-Tu} \frac{u^{n-1}}{\Gamma(n)} \left(\prod_{j=1}^{M^{(a)}} S_{c_j^\star}^{n_j + \gamma - 1} e^{-S_{c_j^\star}} P_0(d\tau_{c_j^\star}) \, dS_{c_j^\star} \right) \\ &\qquad \times \prod_{m \in \mathcal{M}^{(na)}} S_m^{\gamma - 1} e^{-S_m} P_0(d\tau_m) \, dS_m \, du \end{aligned}$$

$$
\begin{aligned}
(3.3) \qquad & \propto e^{-u\sum_m S_m}\frac{u^{n-1}}{\Gamma(n)}\left(\prod_{j=1}^{M^{(a)}} S_{c_j^\star}^{n_j+\gamma-1} e^{-S_{c_j^\star}} P_0(d\tau_{c_j^\star})\, dS_{c_j^\star}\right) \\
& \quad \times \prod_{m\in\mathcal{M}^{(na)}} S_m^{\gamma-1} e^{-S_m} P_0(d\tau_m)\, dS_m\, du \\
& \propto \frac{u^{n-1}}{\Gamma(n)}\left(\prod_{j=1}^{M^{(a)}} S_{c_j^\star}^{n_j+\gamma-1} e^{-S_{c_j^\star}(1+u)} P_0(d\tau_{c_j^\star})\, dS_{c_j^\star}\right) \\
& \quad \times \prod_{m\in\mathcal{M}^{(na)}} S_m^{\gamma-1} e^{-(u+1)S_m} P_0(d\tau_m)\, dS_m\, du.
\end{aligned}
$$

Through the introduction of the latent variable U_n, we gain posterior conditional independence of the unnormalized weights S_j since the random variable T^n now appears in the form e^{uT}. In this way, the full conditional of the unnormalised weights factorises in a product of Gamma densities, whose hyperparameters depend only on the cluster numerosity and the latent variable U_n. This conjugacy clearly offers advantages when designing MCMC algorithms. Moreover, we can marginalize over all the S's and the τ's parameters in equation (3.3). This allows us to derive the full conditional of M and design a transdimensional collapsed Gibbs Algorithm (see Supplementary Material Appendix E [3]). Finally, we can marginalize also over M and this latter step is crucial to derive theoretical properties of the model as well as the full conditional distribution for U_n. For instance, it yields a closed form solution for the exchangeable product partition function of the Finite Dirichlet process mixture model. Further details are given in Section 4, while a full derivation is presented in Supplementary Material Appendix D.2.1. Note that in the same setting the telescopic sampling scheme recently proposed by [17] adopts a similar strategy, introducing a latent variable corresponding to $M^{(na)}$. This latter algorithm improves computational efficiency as compared to a Reversible Jump scheme, but still needs to resort to a truncation.

In the remainder of the paper, we show that this construction is general and applies to any (i.e., non-Dirichlet) mixture models where the weight w_m are set equal to S_m/T and the jumps S_m are realizations of a point process. Exploiting the augmentation trick, we cover a large class of $\mathcal{P}$ and are able to derive theoretical results which allow for efficient posterior inference.

4. Normalized independent finite point processes. A point process $X = \{\xi_1, \ldots, \xi_M\}$ is a set of unordered points of a complete separable metric space $\mathcal{X}$.

DEFINITION 4.1. Let $\nu(\cdot)$ and $q_M(m), m = 0, 1, \ldots$ be a density on $\mathcal{X}$ and a probability mass function respectively. X is an *independent finite point process*, $X \sim \text{IFPP}(\nu, q_M)$, if its Janossy density [7] can be written as

$$
(4.1) \qquad j(\xi_1, \ldots, \xi_m) = m! q_M(m) \prod_{j=1}^{m} \nu(\xi_j).
$$

If $\mathcal{X} = \mathbb{R}^d$, then $j(\xi_1, \ldots, \xi_m)\, d\xi_1 \ldots d\xi_M$ is the probability that there are exactly m points in the process, one in each of the distinct infinitesimal regions $(\xi_m, \xi_m + d\xi_m)$. In Supplementary Material Appendix B we review some concepts from point process theory. In our approach, we use the Janossy measure to assign a prior for P in equation (2.2). In particular, $q_M(\cdot)$ corresponds to the prior on the number of components M, while $\nu(\cdot)$ defines the joint prior for the unnormalised weights and the locations of the mixture.

Let $\Theta \subset \mathbb{R}^d$, for some positive integer d and let $\mathcal{X}$ be $\mathbb{R}^+ \times \Theta$. Here, Θ is the space of the mixture locations (i.e., the kernel parameters), while $\mathbb{R}^+$ is the space of the unnormalised weights. We denote with $\xi = (s, \tau)$ a point of $\mathcal{X}$. Let $\nu(s, \tau)$ be a density on $\mathcal{X}$ such that $\nu(s, \tau) = h(s)p_0(\tau)$, where $h(\cdot)$ is a density on $\mathbb{R}^+$ and $p_0(\cdot)$ is a density on Θ. The density $h(\cdot)$ defines the prior on the unnormalised weights and corresponds to the Gamma density in the example of Section 3, while $p_0(\cdot)$ is a nonatomic density which specifies a prior for the mixture locations. Finally, we assume that the prior probability of $M = 0$ is zero, that is, $q_M(0) = 0$. In what follows, it is easier to introduce a slight change of notation and define $\text{IFPP}(h, q_M, p_0) = \text{IFPP}(\nu, q_M)$ to highlight the dependence of the process also on $p_0(\cdot)$. We consider the independent finite point process $\tilde{P} = \{(S_1, \tau_1), \ldots, (S_M, \tau_M)\}$ with parameters h, p_0 and q_M, that is, $\tilde{P} \sim \text{IFPP}(h, q_M, p_0)$. Let $\mathcal{M} := \{1, \ldots, M\}$ be the set of indexes corresponding to the points of the process. Since we assume $q_M(0) = 0$, the random variable $T := \sum_{m \in \mathcal{M}} S_m$ is a.s. larger than 0 leading to the following definition.

DEFINITION 4.2. Let $\tilde{P} = \{(S_1, \tau_1), \ldots, (S_M, \tau_M)\} \sim \text{IFPP}(h, q_M, p_0)$, with $q_M(0) = 0$. A normalized independent finite point process (Norm-IFPP) with parameters h, p_0 and q_M is a discrete probability measure on Θ defined by

$$P(A) = \sum_{m \in \mathcal{M}} w_m \delta_{\tau_m}(A) \stackrel{d}{=} \sum_{m \in \mathcal{M}} \frac{S_m}{T} \delta_{\tau_m}(A), \tag{4.2}$$

where $T = \sum_{m \in \mathcal{M}} S_m$ and A denotes a measurable set of Θ. We refer to the process in equation (4.2) as $P \sim \text{Norm-IFPP}(h, q_M, p_0)$.

EXAMPLE 4.1 (Finite-Dirichlet process). Let h be the Gamma$(\gamma, 1)$ density, with shape parameter $\gamma > 0$ and rate 1. Then the Norm-IFPP is a finite Dirichlet process, as in equation (4.2). Conditionally on $M > 0$, the jump sizes $(w_1, \ldots, w_M)$ of P are a sample from the M-dimensional Dirichlet$_M(\gamma, \ldots, \gamma)$ distribution (see Section 3).

EXAMPLE 4.2 (Finite σ-Stable process). Let $\psi(u) = \exp(-u^\sigma)$, $u > 0$ with $0 < \sigma < 1$. From the Lévy–Khintchine formula in [29], $\psi(u)$ is the Laplace transform of the σ-Stable density [44]:

$$h(s; \sigma) = -\frac{1}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \sin(\pi \sigma k) \frac{\Gamma(\sigma k + 1)}{s^{\sigma k + 1}}.$$

Then the process P, whose jumps have density $h(s; \sigma)$ is defined as a (normalized) finite σ-Stable.

The finite dimensional process defined in equation (4.2) belongs to the wide class of species sampling models (see [42]) and this allows us to use all the efficient machinery developed for such models. Let $(\theta_1, \ldots, \theta_n)$ be a sample from a Norm-IFPP. It is well known that sampling from a discrete probability measure induces ties among the θ_is and, therefore, a random partition of the observations. Let $\rho_n := \{C_1, \ldots, C_k\}$ indicate a partition of the set $\{1, \ldots, n\}$ in k subsets, where $C_j = \{i : \theta_i = \theta_j^\star\}$ for $j = 1, \ldots, k \le n$, and let $\{\theta_1^\star, \ldots, \theta_k^\star\}$ denote the set of distinct θ_is associated to each C_i. The marginal law of $(\theta_1, \ldots, \theta_n)$ has a unique characterization:

$$\mathcal{L}(\theta_1, \ldots, \theta_n) = \mathcal{L}(\rho_n, \theta_1^\star, \ldots, \theta_k^\star) = \pi(n_1, \ldots, n_k) \prod_{j=1}^{k} P_0(d\theta_j^\star),$$

where $n_j = \#(C_j)$, $\sum_{j=1}^{k} n_j = n$ and $\pi(\cdot)$ is the exchangeable partition probability function associated to the random probability P (see [42]). For each n, the eppf π is a probability law on the set of partitions of $\{1, \ldots, n\}$, which determines the (random) number of clusters k and the numerosity of each cluster C_j. The eppf is a key tool in Bayesian analysis as mixture models can be rewritten in terms of random partitions and such equivalence is often exploited to improve computational efficiency, in particular of marginal algorithms [33].

THEOREM 4.1. *Let $(n_1, \ldots, n_k)$ be a vector of positive integers such that $\sum_{j=1}^{k} n_j = n$. Then, the eppf associated with a* Norm-IFPP(h, q_M, p_0) *is*

$$\pi(n_1, \ldots, n_k) = \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)} \Psi(u, k) \prod_{j=1}^{k} \kappa(n_j, u)\, du, \tag{4.3}$$

where

$$\Psi(u, k) := \left\{ \sum_{m=0}^{\infty} \frac{(m+k)!}{m!} \psi(u)^m q_M(m+k) \right\},$$

$\psi(u)$ is the Laplace transform of the density $h(s)$, that is,

$$\psi(u) := \int_0^{\infty} e^{-us} h(s)\, ds \tag{4.4}$$

and

$$\kappa(n_j, u) := \int_0^{\infty} s^{n_j} e^{-us} h(s)\, ds = (-1)^{n_j} \frac{d}{du^{n_j}} \psi(u).$$

PROOF. See Supplementary Material Appendix C.1. □

The main challenges when computing the eppf of a Norm-IFFP are the evaluation of $\Psi(u, k)$ in equation (4.3) and of the Laplace transform ψ and its cumulants κ in Theorem 4.1.

EXAMPLE 4.3 (Finite-Dirichlet process, continued). Recall that the Laplace transform and its cumulants for a Gamma$(\gamma, 1)$ density are given by $\psi(u) = \frac{1}{(u+1)^{\gamma}}$, and $\kappa(n_j, u) = \frac{1}{(u+1)^{n_j+\gamma}} \frac{\Gamma(\gamma+n_j)}{\Gamma(\gamma)}$, $u > 0$, $n_j = 1, 2, \ldots$ Then, applying Theorem 4.1, we obtain that the eppf is

$$\begin{aligned} p(n_1, \ldots, n_k) &= \left\{ \sum_{m=0}^{\infty} \frac{(m+k)!}{m!} q_M(m+k) \frac{\Gamma((k+m)\gamma)}{\Gamma((k+m)\gamma + n)} \right\} \prod_{j}^{k} \frac{\Gamma(\gamma + n_j)}{\Gamma(\gamma)} \\ &= V(n, k) \prod_{j=1}^{k} \frac{\Gamma(\gamma + n_j)}{\Gamma(\gamma)}. \end{aligned} \tag{4.5}$$

See also Chapter 2 in [43] and [37].

EXAMPLE 4.4 (Finite σ-Stable, continued). By construction, the Laplace transform of the σ-stable density is $\psi(u) = \exp(-u^{\sigma})$, $u > 0$ with $0 < \sigma < 1$. From equation (13) in [12], we obtain its cumulants as

$$\kappa(n_j, u) = \frac{\mathrm{e}^{-u^{\sigma}}}{u^{n_j}} \sum_{l=1}^{n_j} u^{\sigma l} \left| \mathcal{C}(n_j, l; -\sigma) \right|,$$

where $\mathcal{C}(n, l; \sigma)$ denotes the central generalized factorial coefficient (see [6], formula (2.67) for details). Finally, the eppf is given by

$$p(n_1, \ldots, n_k) = \int_0^\infty \frac{1}{u\Gamma(n)} \sum_{m=0}^{\infty} \frac{(m+k)!}{m!} e^{-(m+k)u^\sigma} q_M(m+k) \prod_{j=1}^{k} \sum_{l=1}^{n_j} |\mathcal{C}(n_j, l; -\sigma)| \, du. \tag{4.6}$$

Note that for special choices of q_M, we are able to give an integral representation for the the infinite sum in equation (4.5) and find an analytical solution for the one in equation (4.6), allowing for efficient computations. See Supplementary Material Appendix D.1 for details.

The number of components of the finite mixture M is given by a realisation of the process in equation (4.2). On the other hand, k denotes the number of nonempty (allocated) components, with $k \leq M$. This difference has been noted before in the literature (see, e.g., [16, 37, 41]). Moreover, marginalising over the cluster sizes, it is also possible to derive the implied prior distribution on the number of clusters, k, which corresponds to the number of allocated components.

COROLLARY 4.1. *Under the assumptions of Theorem 4.1, the marginal prior probability of sampling a partition with k clusters is given by*

$$p_k^\star = \int_0^{+\infty} \frac{u^{n-1}}{\Gamma(n)} \left\{ \sum_{m=0}^{\infty} \frac{(m+k)!}{m!} \psi(u)^m q_M(m+k) \right\} B_{n,k}\big(\kappa(\cdot, u)\big) \, du, \tag{4.7}$$

where $k = 1, \ldots, n$, and $B_{n,k}(\kappa(\cdot, u))$ is the partial Bell polynomial [43] *over the sequence of coefficients $\{\kappa(n, u), n = 1, 2, \ldots\}$.*

PROOF. See Supplementary Material Appendix C.2. □

Moreover, from de Finetti's theorem, it follows that k converges almost surely to M as $n \to \infty$.

EXAMPLE 4.5 (Finite Dirichlet process, continued). Equation (4.5) implies that the finite Dirichlet process is a member of the family of Gibbs partition distributions [33, 42]. The Gibbs type structure allows us to simplify the prior for the number of allocated components given in equation (4.7), which becomes

$$p_k^\star = V(n, k)\gamma^k S_{n,k}^{-1,\gamma} = V(n, k)(-1)^n \mathcal{C}(n, k; -\gamma), \quad k = 1, \ldots, n, \tag{4.8}$$

where $S_{n,k}^{-1,\gamma}$ is the *Generalized Stirling number* computed for k compositions of n objects with parameters -1 and γ (see equation (1.20) in [43]), while for any nonnegative integer $n \geq 0$, $0 \leq k \leq n$ and real numbers α, $\mathcal{C}(n, k; \alpha)$ denotes the central generalized factorial coefficient. Here we mention that these indices can be easily computed using the recursive formula

$$\mathcal{C}(n, k; \alpha) = \alpha \mathcal{C}(n-1, k-1; \alpha) + (k\alpha - n + 1)\mathcal{C}(n-1, k; \alpha)$$

with $\mathcal{C}(1, 1, \alpha) = \alpha$.

Note that, when q_M is a shifted Poisson, if $\gamma = \alpha/\Lambda$, for $\alpha > 0$, and $\Lambda \to \infty$, then P converges in distribution to the Dirichlet process with mass parameter α (see Supplementary Material Appendix C.5 for a proof). Similarly, we recover the Dirichlet process when q_M assigns mass one to $\widetilde{M}$, $\gamma = \alpha/\widetilde{M}$ and $\widetilde{M}$ goes to infinity. This case has been extensively investigated in the Bayesian nonparametric literature from both computational and methodological perspective (see [28], for a thorough discussion).

5. Posterior characterization of a Norm-IFPP process. Let the random variable $U_n = \Gamma_n / T$, where $\Gamma_n \sim \text{Gamma}(n, 1)$, $T = \sum_{i \in \mathcal{M}} S_i$, Γ_n and T are independent. It is easy to show (see the Supplementary Material Appendix C.4) that if $P \sim \text{Norm-IFPP}(h, q_M, p_0)$ then, for any $n \geq 1$, the marginal density of U_n is given by

$$f_{U_n}(u; n) = \frac{u^{n-1}}{\Gamma(n)}(-1)^n \frac{d}{du^n}\mathbb{E}\big(\psi(u)^M\big), \tag{5.1}$$

where $\psi(u)$ is the Laplace transform of the density h, as defined in equation (4.4). The posterior distribution of U_n, given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, is crucial to perform posterior inference and allows us to derive the posterior distribution of the unnormalised process $\widetilde{P}$. To this end, we need to show that *a posteriori*, conditionally on U_n, $\widetilde{P}$ is the superposition (union) of two independent processes: an IFPP and a finite process with fixed locations at $(\theta_1^\star, \ldots, \theta_k^\star)$. Note that k corresponds to the number of allocated jumps $M^{(a)}$ and M is equal to the sum of k and the number $M^{(na)}$ of unallocated jumps, assuming values in $\mathbb{N} \cup \{0\}$. The process of unallocated jumps is a latent variable which links the parametric part of the model in P to a nonparametric process. This link is essential for computations as it will become clearer in Section 7, where we discuss the algorithm. The results below are conditional on the realizations of the random variable U_n, which is a typical strategy in the theory of normalised random measures, since working on the augmented space allows us to exploit the quasi-conjugacy of the process P (see [30]). We now present the main theoretical contribution of this work.

THEOREM 5.1. *If $P \sim \text{Norm-IFPP}(h, q_M, p_0)$, then the unnormalized process $\widetilde{P}$, given $\boldsymbol{\theta}^\star = (\theta_1^\star, \ldots, \theta_k^\star)$, $\boldsymbol{n} = (n_1, \ldots, n_k)$ and $U_n = u$, is the superposition of two processes*:

$$\widetilde{P} \stackrel{d}{=} \widetilde{P}^{(na)} \cup \widetilde{P}^{(a)},$$

where:

1. *The process of unallocated jumps $\widetilde{P}^{(na)}$ is an independent finite point process with Janossy density given by*

$$j_m\big((s_1, \tau_1), \ldots, (s_m, \tau_m)\big) = m! q_m^\star \prod_{j=1}^{m} h_u^\star(s_j) p_0(\tau_j),$$

where $h_u^\star(s) \propto \mathrm{e}^{-us} h(s)$, $q_m^\star \propto \frac{(m+k)!}{m!}\psi(u)^m q_M(m+k)$, $\psi(u)$ is the Laplace transform of h, and m is a realization of $M^{(na)}$, the number of unallocated jumps, taking values in $\{0, 1, 2, \ldots\}$.

2. *The process of allocated jumps $\widetilde{P}^{(a)}$ is the unordered set of points $(S_1, \tau_1), \ldots, (S_k, \tau_k)$, such that, for $j = 1, \ldots, k$, $\tau_j = \theta_j^\star$ and the distribution of S_j is proportional to $s^{n_j}\mathrm{e}^{-us}h(s)$.*

3. *Conditionally on $\mathcal{M}^{(a)}$ and $U_n = u$, $\widetilde{P}^{(a)}$ and $\widetilde{P}^{(na)}$ are independent.*

Moreover, the posterior law of U_n given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ depends only on the partition ρ_n and has density on the positive reals given by

$$f_{U_n}(u \mid \rho_n) \propto \frac{u^{n-1}}{\Gamma(n)} \left\{ \sum_{m=0}^{\infty} \frac{(m+k)!}{m!} \psi(u)^m q_M(m+k) \right\} \prod_{j=1}^{k} \kappa(n_j, u).$$

PROOF. See Supplementary Material Appendix C.3. □

The result in Theorem 5.1 is the finite dimensional counterpart of Theorem 1 in [30] for normalised completely random measure. This theorem allows building an efficient block

Gibbs sampler for finite mixture models. Since the order in which the points of a point process arise is not important, without loss of generality, given a realization of the posterior process $\widetilde{P}$, we assume that, in $\widetilde{P} = \{S_1, \ldots, S_M\}$, $M = k + M^{(na)}$, that is, the first k points $\{S_1, \ldots, S_k\}$ correspond to the allocated jumps, while the last $M^{(na)}$ to the unallocated ones.

EXAMPLE 5.1 (Finite Dirichlet process, continued). Conditionally on $U_n = u$, $h_u^\star$ is still a Gamma density (see Section 3) with parameters γ and $u + 1$. Moreover, $q^\star$ has closed form for particular choices of q_M as detailed in Supplementary Material Appendix D.1. Thus, the process corresponding to the unallocated jumps is still a Finite Dirichlet process with updated parameters. The unnormalised weights, S_j, of the k allocated jumps are a posteriori independent and have a Gamma distribution with parameters $n_j + \gamma$ and $(u + 1)$.

EXAMPLE 5.2 (Finite σ-Stable, continued). In this case, conditionally on $U_n = u$, $h_u^\star$ is an Exponential tilted stable density $\mathrm{e}^{-us}h(s)$ (or generalized Gamma), while the density of the the unnormalised weights, S_j, of the k allocated jumps is a Gamma tilted density $s^{-n_j}\mathrm{e}^{-us}h(s)$. Properties of the Exponential tilted density are presented in [9] and [24], while Gamma tilted σ-stable densities are discussed in [12].

In general, there are two possible alternatives to define a Norm-IFPP: either to choose a parametric density as h or to select the Laplace transform of h, ψ. In Supplementary Material Appendix D, we extensively discuss special choices of h and ψ which allow for analytical derivation of the eppf and of the posterior distribution. In addition to the two examples discussed in the manuscript, we also describe: (i) uniform weights; (ii) Gamma approximation; (iii) the Bessel process. Finally, in the same Appendix, choices of q_M that allow for easy computations are discussed.

6. Norm-IFFP hierarchical mixture models. Most real world applications of discrete random measures involve an additional layer in the model hierarchy and convolve the random measure with a continuous kernel as described in equation (2.2). In our context, the random measure is a Norm-IFFP. This leads to models of the form

$$\begin{aligned} Y_1, \ldots, Y_n \mid \theta_1, \ldots, \theta_n &\overset{\text{ind}}{\sim} f(y \mid \theta_i), \\ \theta_1, \ldots, \theta_n \mid P &\overset{\text{iid}}{\sim} P, \\ P &\sim \text{Norm-IFPP}(h, q_M, p_0), \end{aligned} \tag{6.1}$$

where $f(y \mid \theta_i)$ is a parametric density on $\mathcal{Y}$, for all $\theta \in \Theta \subset \mathbb{R}^d$. We point out that p_0 is the density of a nonatomic probability measure P_0 on Θ, such that $\mathbb{E}(P(A)) = P_0(A)$ for all $A \in \mathcal{B}(\Theta)$. Model (6.1) will be addressed here as a *Norm-IFFP hierarchical mixture* model. The model can be extended by specifying appropriate hyperpriors. It is well known that this model is equivalent to assuming that the Y_i's, conditional on P, are independently distributed according to the random density (1.1). We point out that Model (6.1) admits as a special case the popular finite Dirichlet mixture model (see [37, 40, 47, 50]) discussed in more details in Section D.2.1. The posterior characterization given in Theorem 5.1, as well as the analytical expression for the eppf given in Theorem 4.1, allow us to devise conditional or marginal algorithms to perform inference under Model (6.1) as discussed in Section 7. Moreover, Building upon [23], in Section 8 we prove consistency results for the number of components as well deriving an optimal contraction rates for component parameters the class of Norm-IFPP under minimal assumptions on the density $h(s)$ of the unnormalized weights.

7. Posterior inference. To perform posterior inference tailored MCMC algorithms need to be devised. The two most popular strategies in Bayesian nonparametrics are marginal [38] and conditional algorithms [27, 31]. Our construction allows for straightforward extension of such strategies to the finite mixture case, offering a convenient alternative to the often inefficient and labour intensive reversible jump. To implement marginal algorithms it is desirable (although not necessary, but at the cost of extra computations) to be able to compute the eppf of the process, and, therefore, the sum in equation (4.3) to obtain the probability of a random partition. On the other hand, for conditional algorithms we need to sample from the posterior distribution of a Norm-IFPP which requires a closed form expression for the Laplace transform in Theorem 5.1. More specifically, it is essential to be able to sample from the posterior distribution of the number of the nonallocated jumps, $q_m^\star$, as well as from the distribution of the allocated and unallocated jumps, that is, the densities proportional to $e^{-us}h(s)$ (Exponential tilted) and $s^{n_j}e^{-us}h(s)$ (Gamma tilted). Specific solutions for well-known processes will be presented in Supplementary Material Appendix D. Here we give a general outline of both algorithms.

7.1. *Marginal algorithm.* As mentioned before, a sample $\theta_1, \ldots, \theta_n$ from P induces a partition of the set of the data indexes, denoted by $\rho_n = \{C_1, \ldots, C_k\}$, such that $i \in C_j$ implies that datum i belongs to cluster j. Marginal algorithms rely on the fact that, by integrating out the measure P, the only parameters left in equation (2.2) are the random partition ρ_n and the cluster specific parameters $\theta_1^\star, \ldots, \theta_k^\star$. Posterior sampling strategies for ρ_n are based on the Chinese restaurant process [1], which describes the (a priori) predictive generative process for ρ_n, and relies on the evaluation of the eppf associated with P. Nevertheless, when P corresponds to the Norm-IFPP model, this evaluation can be computationally burdensome due to the integral with respect to u in equation (4.3). To design efficient algorithms we adopt a *disintegration* technique following a strategy similar to the one suggested by [30] and [13] for NRMI. Indeed, we can define the joint law of the random partition ρ_n and the latent variable U_n defined in Section 5 as follows:

$$\mathcal{L}(\rho_n, U_n) = \pi(n_1, \ldots, n_k; u)\frac{u^{n-1}}{\Gamma(n)}\Psi(u,k)\prod_{j=1}^{k}\kappa(n_j, u)\,du, \tag{7.1}$$

where $\Psi(u,k) := \{\sum_{m=0}^{\infty}\frac{(m+k)!}{m!}\psi(u)^m q_M(m+k)\}$. Equation (7.1) is a well-defined joint distribution since the marginal density of U_n exists and is given in equation (5.1). This enables us to give a generalized Chinese restaurant representation of the generative process of the partition ρ_n jointly with the latent variable U_n, in which the observations are represented as *customers* in a restaurant with infinite many *tables* representing the clusters.

The predictive probability (conditionally on $U_{n+1} = u$) that *customer* $n+1$ seats to a new unoccupied *table* C_{k+1} is

$$\mathbb{P}(n+1 \in C_{k+1} | u, \rho_n) \propto \frac{\pi(n_1, \ldots, n_k, 1; u)}{\pi(n_1, \ldots, n_k; u)} = \frac{\Psi(u, k+1)}{\Psi(u,k)}\kappa(1,u) \tag{7.2}$$

while the predictive probability that it seats to an existing *table* is

$$\mathbb{P}(n+1 \in C_j | u, \rho_n) \propto \frac{\pi(n_1, \ldots, n_j+1, \ldots, n_k; u)}{\pi(n_1, \ldots, n_j, \ldots, n_k; u)} = \frac{\kappa(n_j+1, u)}{\kappa(n_j, u)} \tag{7.3}$$

for $j = 1, \ldots, k$. Note that, for each new customer i, a variable U_i is drawn. After n customers have entered the restaurant, the seating arrangement of customers around tables corresponds to a partition ρ_n of $\{1, \ldots, n\}$ with numerosity $(n_1, \ldots, n_k)$, $n_j = \#C_j$. The main difference with the standard Chinese process consists in updating the cluster allocation of customer $n+1$

conditional on the U_{n+1}. The strategy of conditioning on a sequence of auxiliary variables to generalise the Chinese restaurant process was introduced for infinite dimensional measures by [30]. Here, we have derived the finite dimensional counterpart.

We can now describe the marginal algorithm based on the generalised Chinese process, by exploiting the exchangeability of the partition ρ_n obtained under such process. As a consequence, for each i, we can assume that the ith observation is the last *customer*. Then we can update ρ_n using Gibbs sampling whereby the cluster assignment of one datum i is updated one at a time. Let $\rho_n^{-i} = \{C_1^{-i}, \ldots, C_{k^{-i}}^{-i}\}$ be the partition in k^{-i} clusters obtained from the partition ρ_n when the ith datum is removed. We denote the cluster assignment of the ith observation to cluster C_j^{-i} of size n_j^{-i}, with the event $\{i \in C_j^{-i}\}$, while with $\{i \in C_{k^{-i}+1}^{-i}\}$ we denote the event that the ith observation is assigned to a new (empty) cluster. The full conditional of the allocation events given ρ_n^{-i}, $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $U_n = u$ is obtained modifying (7.2)–(7.3), as shown in the top panel of Figure 1, where $\boldsymbol{y}_{C_j}$ denotes the vector of observations y_l such that $l \in C_j$ and $\mathcal{M}(\boldsymbol{y}_{C_j}) = \int_\Theta \prod_{l \in C_j} f(y_l \mid \theta) p_0(\theta)\, d\theta$ is the marginal distribution of the data within cluster C_j with sampling model $f(y_l \mid \theta)$ and prior $p_0(\theta)$.

We can also perform posterior inference on M under the marginal algorithm by sampling $M^{(na)}$ from

$$q_m^\star \propto \frac{(m+k)!}{m!} \psi(u)^m q_M(m+k).$$

The algorithm is summarised in Figure 1 and relies on the conjugacy of the kernel $f(y|\theta)$ and the density $p_0(\tau)$. Nevertheless the algorithm can be easily extended to the nonconjugate case

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

1. Sample $\rho_n^{(g)}$, reallocating each individual $i = 1, \ldots, n$ according to the following full conditionals

$$\mathbb{P}(i \in C_{k^{-i}+1}^{-i} | u, \rho_n^{-i}, \boldsymbol{y}) \propto \frac{\Psi(u, k+1)}{\Psi(u,k)} \kappa(1,u) \mathcal{M}(y_i) \tag{7.4}$$

and for $j = 1, \ldots, k^{-i}$

$$\mathbb{P}(i \in C_j^{-i} | u, \rho_n) \propto \frac{\kappa(n_j^{-i}+1, u)}{\kappa(n_j^{-i}, u)} \frac{\mathcal{M}(\boldsymbol{y}_{C_j^{-i} \cup i})}{\mathcal{M}(\boldsymbol{y}_{C_j^{-i}})} \tag{7.5}$$

2. Update $U_n^{(g)}$ from the full conditional

$$\mathbb{P}(U_n = du | rest) \propto u^{n-1} \Psi(u,k) \prod_{j=1}^{k} \kappa(n_j, u) du \tag{7.6}$$

3 If assumed random, sample the hyperparametrs $\eta_1^{(g)}$ of the density h from

$$\mathbb{P}(\eta_1 = d\eta_1 \mid \text{rest}) \propto \Psi(u,k) \prod_{j=1}^{k} \kappa(n_j, u) \pi_1(\eta_1) d\eta_1$$

where $\pi_1(\eta_1)$ denotes the prior density for η_1.

4 If assumed random, sample the hyperparametrs $\eta_2^{(g)}$ of the density q_M from

$$\mathbb{P}(\eta_2 = d\eta_2 \mid \text{rest}) \propto \Psi(u,k) \prod_{j=1}^{k} \kappa(n_j, u) \pi_2(\eta_2) d\eta_2$$

where $\pi_2(\eta_2)$ denotes the prior density for η_2.

5 Draw $\theta_j^{\star(g)}$, for each $j = 1, \ldots, k$, from

$$\mathbb{P}(\theta_j^\star = d\theta_j^\star \mid \text{rest}) \propto \left\{ \prod_{i \in C_j} f(y_i \mid \theta_j^\star) \right\} p_0(\theta_j^\star) d\theta_j^\star$$

FIG. 1. *Marginal Gibbs sampler scheme*; *the conditioning arguments of all full conditionals have been omitted to simplify notation. We point out that steps* 4 *and* 5 *are needed only if the parameters of the density of the unnormalised jumps* h *and of the prior on the number of components* q_M *are assumed to be random.*

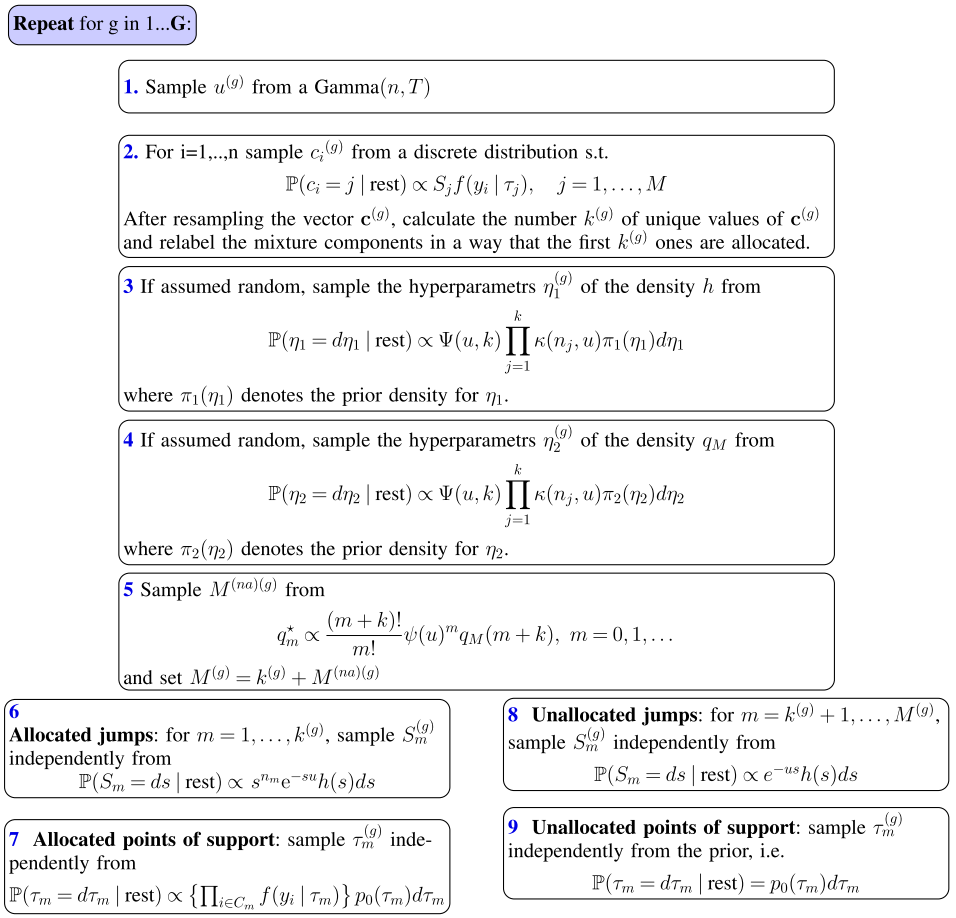

FIG. 2. *Blocked Gibbs sampler scheme*; *the conditioning arguments of all full conditionals have been omitted to simplify notation. We point out that steps* 4 *and* 5 *are needed only if the parameter of the density of unnormalized jumps* h *and of the prior on the number of components* q_M *are assumed random.*

following a similar strategy to Algorithm 8 of [38] (see also [13]). The marginal algorithm for the Finite Dirichlet process is detailed in Supplementary Material Appendix E.2.

7.2. *Conditional algorithm.* Conditional algorithms allow us to draw from the joint distribution of $(M, \boldsymbol{\tau}, \mathbf{S}, \mathbf{c})$ in equation (2.1), where $w_i = S_i / T$, which in turn defines a draw of the random probability measure on Θ: $P(d\theta) = \sum_{m=1}^{M} w_m \delta_{\tau_m}(d\theta)$. As the algorithm samples from the posterior distribution of the random measure, we are able to perform full posterior inference, at least numerically, on any functional of such distribution [19]. Moreover, it is simple to make inference on the hyper-parameters of the distributions of M and $\mathbf{S}$. An outline of the MCMC algorithm is given in Figure 2. The scheme follows directly from Theorem 5.1, adapted to the mixture case. Note that in Step 2 of the algorithm, the relabelling of the mixture components is essential so that the nonempty components correspond to the first k components. Moreover, Step 7 of the algorithm requires a standard Bayesian update of the cluster-specific parameters, which can be performed using any MCMC strategy in case of lack of conjugacy. The conditional algorithm for the FDMM is described in details in Supplementary Material Appendix E.1 and it is implemented in the R-package `AntMAN` [4].

8. Consistency. In a recent paper, [23] show that in the particular case of the FDMM with random number of components it is possible to obtain both a consistent estimate of the

number of mixture components, and more notably, an optimal posterior contraction rate for component parameters, under a general set of conditions. It is worth emphasizing that all these results are possible only under the assumption that the model is well-specified, that is, the true but unknown population density lies in the support of the prior process. In this section we extend the results by [23] to the class of Norm-IFPP under minimal assumptions on the density $h(s)$ of the unnormalized weights.

We assume that the data $Y_1, \ldots, Y_n$ are an i.i.d. sample from a mixture density $f_Y(y|P^\star) = \int f(y|\theta)\, dP^\star(\theta)$, where $P^\star$ is a discrete mixing measure with *unknown* number of support points $m^\star \leq \infty$ residing in the compact $\Theta \subset \mathbb{R}^d$. Moreover, we make the following assumptions (see Supplementary Material Appendix C.6 for details and definitions):

P.1 The parameter space Θ is compact, while the kernel density f is first-order identifiable and admits the uniform Lipschitz property up to the first order.

P.2 The base distribution P_0 is approximately uniform, that is, $\min_{\tau\in\Theta} p_0(\tau) > c_0 > 0$, where p_0 is the density of P_0.

P.3 There exists $\epsilon^\star > 0$ such that $\int f(y|P^\star)^2/f(y|P)\, dy \leq M(\epsilon^\star)$ as long as $W_1(P, P^\star) \leq \epsilon^\star$ for any $P \in \mathcal{O}_{m^\star}$ where $M(\epsilon^\star)$ depends only on $\epsilon^\star$, $P^\star$, and Θ. Here W_1 denotes the Wasserstein distance of order one and $P^\star$ indicates the true density.

P.4 The prior q_M places positive mass on the set of natural numbers, that is, $q_M(m) > 0$ for all $m = 1, 2, \ldots$

P.5 For each $\tau \in \mathbb{R}^d$, let $B(\tau, \epsilon) = \{\tau' \in \mathbb{R}^d : \|\tau - \tau'\| < \epsilon\}$ the l_2 ball of radius ϵ, then

$$\lim_{\epsilon\to 0}\left(\min_{\tau\in\Theta} \frac{\mu(B(\tau,\epsilon)\cap\Theta)}{\mu(B(\tau,\epsilon))}\right) \geq c_1 > 0.$$

This condition essentially requires that a $\mathbb{R}^d$-balls with a center in Θ, irrespective of its radius, has at least $c_1\%$ of mass in Θ.

P.6 There exists a $t^\star > 0$ such that, for each $0 < \delta < t^\star$, the density h is bounded away from 0 in the interval $[\delta, t^\star]$.

THEOREM 8.1. *Let $\mathcal{L}(Y|P^\star)$ denote the probability law of the infinite sequence $Y_1, Y_2, \ldots$ which forms an i.i.d. sample from $f_Y(y|P^\star)$ where $f_Y(y|P^\star) = \int f(y\mid\theta)\, dP^\star(\theta)$. Under the assumptions P.1–P.6 for a mixture model with Norm-IFPP as mixing measure, we have that*

A-1

$$\mathbb{P}(M = m^\star|Y_1, \ldots, Y_n) \to 1 \quad a.s.\ \mathcal{L}(Y|P^\star), \tag{8.1}$$

where $m^\star$ is the true number of components.

A-2

$$\mathbb{P}\big(P \in \bar{\mathcal{G}}(\theta) : W_1(P, P^\star) \lesssim (\log n/n)^{1/2}|Y_1, \ldots, Y_n\big) \to 1 \tag{8.2}$$

in $\mathcal{L}(Y|P^\star)$-probability,

where $\bar{\mathcal{G}}(\theta)$ denotes the space of all discrete measures including those with countably infinite support on Θ.

PROOF. See Supplementary Material Appendix C.6. □

Result A-1 has been proved by [40], while A-2 is an extension of a result in [23]. The Theorem provides further support for employing finite mixtures when the number of mixture components is unknown and object of inference. Nevertheless, the Bayesian nonparametric

approach is appealing especially from a computational point of view when there are many components with small probabilities.

We conclude this section with a final consideration. When the true mixing distribution $P^\star$ has infinite support points, consistency results are available for the case of Dirichlet Process location mixtures (see Theorem 6 in [39]). Our conjecture is that similar results hold for the class of Norm-IFPP. To this end, we would need to prove that the assumptions in Theorem 2 in [39] hold. Furthermore, we need to show that the prior is fairly diffused on the space of discrete distributions. This latter property implies that there exists a lower bound for the probability that the Wasserstein distance between the process $P \sim$ Norm-IFPP and the true $P^\star$ is small. Lemma 5 in [39] proves this property for the infinite Dirichlet process and Lemma 4.3 in [23] for the finite Dirichlet process with γ depending on M. For models beyond location mixture, we believe the question remains open and this is object of current research.

9. Galaxy data. We illustrate our model using the Galaxy dataset [48], which offers a standard benchmark for mixture models. It contains $n = 82$ measurements on velocities of different galaxies from six well-separated conic sections of space. Values are expressed in km/s, scaled by a factor of 10^{-3}. We fit Model (6.1), using a Gaussian density $\mathcal{N}(\mu, \sigma^2)$ on $\mathbb{R}$ as $f(y \mid \tau)$, $\tau = (\mu, \sigma^2)$. We specify the following prior $p_0(\mu, \sigma^2) = \mathcal{N}(\mu; m_0, \sigma^2/\kappa_0) \times$ Inv-gamma$(\sigma^2; \nu_0/2, \nu_0/2\sigma_0^2)$. Here Inv-Gamma$(a, b)$ denotes the Inverse-Gamma distribution with mean $b/(a-1)$ (if $a > 1$). We set $m_0 = \bar{x}_n = 20.8315$, $\kappa_0 = 0.01$, $\nu_0 = 4$, $\sigma_0^2 = 0.5$ (see [10]). Finally, we assume a shifted Poisson(Λ) as prior on M and a Gamma$(\gamma, 1)$ as a prior for S_m (i.e., a finite Dirichlet process as mixing distribution). We implement the conditional algorithm described in Supplementary Material Appendix E.1 to perform posterior inference, which offers an efficient alternative to the Reversible Jump. In particular, we focus on density estimation and inference on the number of mixture components and clusters.

We fit the model with Λ and γ fixed, with the aim of comparing the performance of our algorithm with the reversible jump sampler of [47] as implemented in the `mixAK` R-package [32]. We set the prior hyperparamters for the model implemented in `mixAK` R-package in such a way that their prior specification closely match ours, with the only difference that the scale parameters in the Inverse Gamma prior is treated as random in `mixAK`, while for us it is fixed.

Our conditional algorithm has been implemented in the R package `AntMAN` [4], with post processing of the MCMC results in R. For each MCMC run, we have discarded the first 5000 iterations as burn-in and thinned every 10, obtaining a final sample size of 5000. We have considered different scenarios, and in Figure 3 we show the predictive density with 95% credible bounds for one of them.

First of all, we fix the hyperparameters γ and Λ in equation (2.1) in such a way that the prior mean for the number of clusters is (A) $\mathbb{E}(k) = 1$; (B) $\mathbb{E}(k) = 5$; (C) $\mathbb{E}(k) = 10$. In order to compare the conditional algorithm with the Reversible Jump, we compute the integrated autocorrelation time (IAC) and the effective sample size (ESS) for the number M of components for the all combinations of hyper-parameters. Posterior results are summarised in Table 1: it is evident that our algorithm outperforms the reversible jump in terms of both the IAC and ESS. In the same table, we report as well the running times in seconds.

Note that, through an appropriate choice of (Λ, γ), we are able to introduce in the model a desired level of sparsity. In Figure 4, we report the posterior distribution of the number of clusters (allocated components) for the same combinations of hyper-parameters in Table 1. The results are in line with previous analyses of the same data (see, for instance, [22], Table 1 and Figure 5). It is clear that the posterior distribution of the number of clusters is robust to the choice of hyper-parameters within each scenario (A, B and C), since the prior mean on

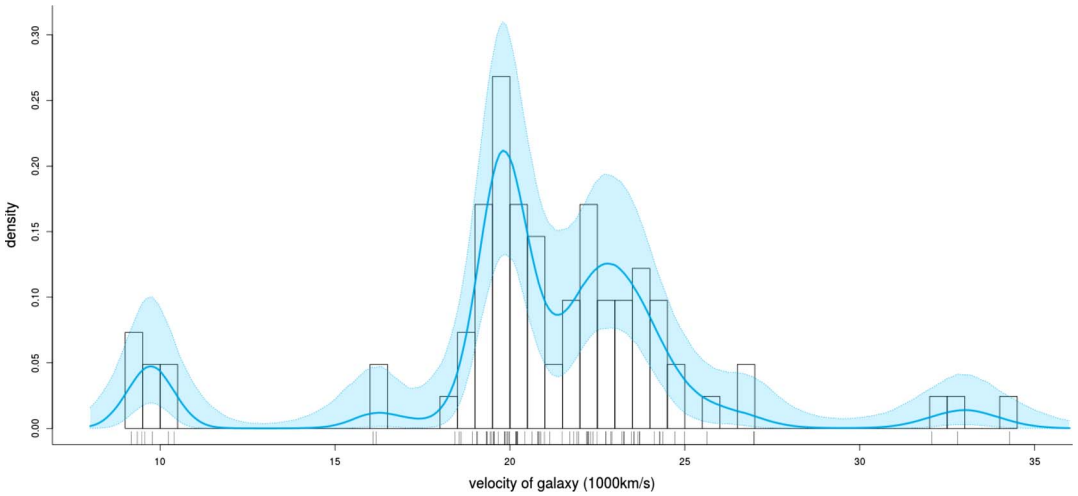

FIG. 3. *Density estimation for* $m_0 = \bar{x}_n = 20.8315$, $\kappa_0 = 0.01$, $\nu_0 = 4$, $\sigma_0^2 = 0.5$. *The hyperparameter settings of the mixing distribution are specified in simulation scenario in* $D.1$ *corresponding to the optimal value of the LPML index.*

the number of allocated components is constant. To gain more insight, in Figure 5 we show the posterior distribution of $M^{(na)}$, the number of unallocated components. We highlight: (i) these posteriors are more concentrated on large number for large values of Λ (ii) for the same value of Λ the level of sparsity increases for small values of γ (see variations within columns). Large values of Λ and small values for γ favour a posterior distribution for $M^{(na)}$ centred on large values. We conclude that Λ controls the number of unallocated clusters, while γ controls degree of sparsity of the mixture. In Supplementary Material Appendix G, we present further results obtained setting a prior on γ and Λ.

Finally, in Supplementary Material Appendix F an extensive simulation study is carried out to compare the performance of the marginal algorithm for FDMM with the marginal algorithm for the finite σ-Stable mixture process, as well as with the performance of the conditional algorithm for the FDMM. Furthermore, in Section F.5, we investigate the robustness of posterior inference with respect to different prior specifications on M.

TABLE 1

Posterior mean of M, *integrated autocorrelation times* $\hat{\rho}$ *and running time in seconds for the Marginal Gibbs sampler (GS) in Supplementary Material Appendix E.*1 *and the Reversible Jump (RJ) MCMC implemented in the R-package* `mixAK`

		GS				*RJ*			
	(Λ, γ)	$\mathbb{E}(M\|\text{data})$	ESS	IAC	sec.	$\mathbb{E}(M\|\text{data})$	ESS M	IAC M	sec.
A	$(100, 2e^{-4})$	102.96	4637.69	1.50	7.74	103.23	2.32	727.32	43.43
	$(10, 2e^{-3})$	13.69	4235.77	0.54	2.52	13.05	18.04	253.28	33.01
	$(1, 10e^{-2})$	4.19	913.42	2.70	1.64	4.17	230.76	11.80	33.36
B	$(100, 1e^{-2})$	103.49	4602.13	0.54	9.81	100.45	11.57	190.16	68.08
	$(10, 0.143)$	13.56	2166.03	1.31	2.58	15.27	301.55	8.21	72.46
	$(5, 0.5)$	8.63	1019.06	2.32	2.07	8.06	1034.58	2.30	67.08
C	$(1000, 2.8e^{-3})$	1001.07	5000.00	1.51	110.58	995.09	1.21	890.70	80.47
	$(100, 3.2e^{-2})$	101.48	4172.39	0.55	10.34	103.33	23.05	76.70	75.90
	$(10, 1.8)$	11.51	888.67	2.85	2.85	11.04	921.14	2.52	69.20

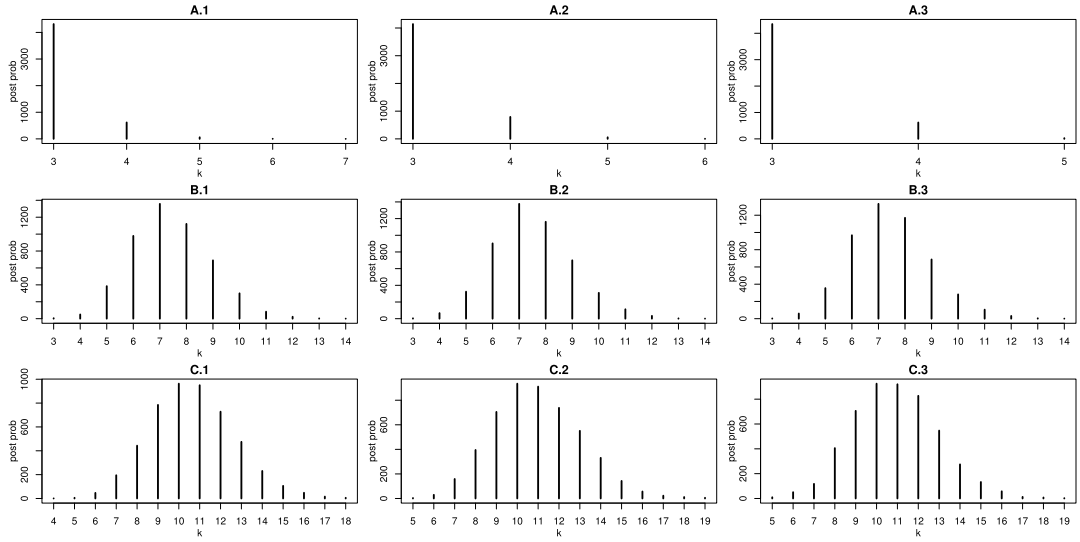

FIG. 4. *Posterior distribution of k for the three scenarios.*

10. Population structure: Taita thrush data. In population genetics, population structure refers to the presence of systematic differences in genetic markers' allele frequencies between subpopulations due to variation in ancestry. This phenomenon arises from the biogeographical distribution of species, due to the fact that either natural populations occupy a vast geographic area and cannot act as randomly mating or geographical barriers reduce migration between different regions. Consequently population structure affects the dynamics of alleles in populations and impacts the type of statical analysis to perform in many applications, for example, in genetic association studies. A variety of statistical approaches have been proposed to infer population structure. Arguably the most widely used method is the one proposed by [45] based on Bayesian mixture models and implemented in the software STRUCTURE [46]. [45] assume that individuals come from one of M (fixed) subpopulations and population membership and population specific allele frequencies are jointly estimated from the data. Independent priors on the allelic profile parameters of each population are

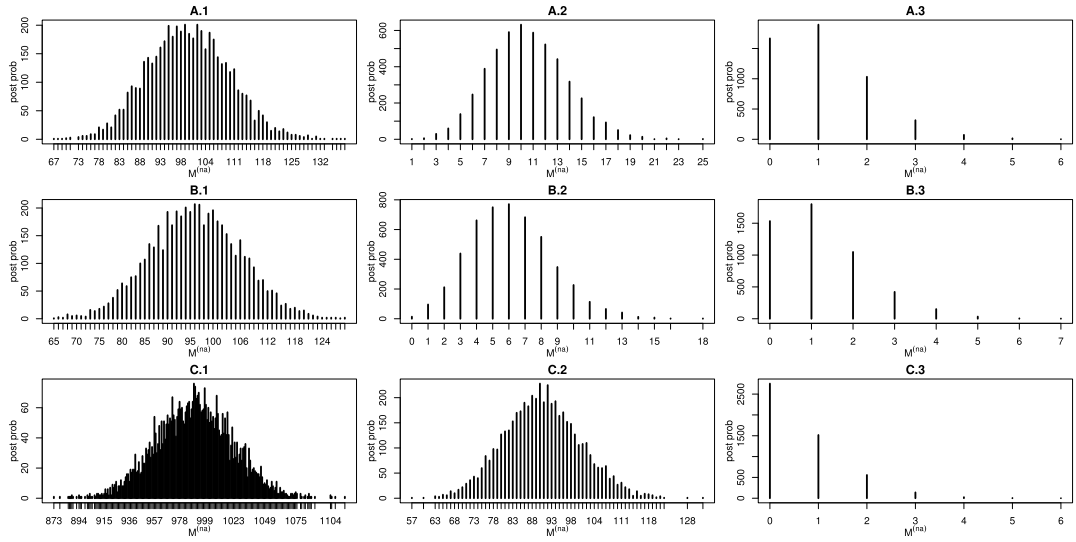

FIG. 5. *Posterior distribution of $M^{(na)}$ for the three scenarios.*

specified and posterior inference is performed through MCMC. In [45], the number of mixture components is fixed and their method clusters individual in one of a fixed number of populations. Determination of the number of populations in a sample is achieved using a model selection criteria based on MCMC estimates of the log marginal probabilities of the data and the Bayesian deviance information criterion, though it has been noted by [11] that such estimates are highly sensitive to prior specifications regarding the relatedness of the populations. To avoid such model selection, [25] propose a method for the analysis of population structure based on a Dirichlet process mixture model and implemented in the software `Structurama` [26], which does not require the specification of a fixed and finite M.

We now illustrate the performance of our method in a population structure problem, using an empirical data set of $n = 237$ Taita thrushes kindly made available by Dr P. Galbusera. A previous smaller version of these data [18] has been analysed by [45] and [25] as benchmark example. The Taita Hills in Kenya represent the northernmost part of the Eastern Arc Mountains biodiversity hotspot of Kenya and Tanzania. They are isolated from other highlands by over 80 km of semiarid plains in either direction. During the last 200 years, indigenous forest cover in the Taita Hills has decreased by circa 98% and the critically endangered Taita thrush, endemic to the Taita Hills, is currently restricted to the fragments of Mbololo, Ngangao and Chawia [5]. These fragments are separated from each other by cultivated areas and human settlements. Each bird was sampled at $L = 6$ microsatellite loci. The Taita thrush is diploid, that is, has two sets of chromosomes and for each locus we have genotype data. At locus l, we observe J_l unique alleles. The number of copies of allele j at locus l in individual i is denoted by $Y_{ilj} \in \{0, 1, 2\}$ and the number of copies of all alleles observed at locus l in individual i is denoted by $Y_{il} = \sum_{j=1}^{J_l} Y_{ilj}$. The allelic information for individual i at locus l is contained in the vector $Y_{il} = (Y_{il1}, Y_{il2}, \ldots, Y_{ilJ_l})$, with the constrain $\sum_{j=1}^{J_l} Y_{ilj} = 2$. Given M possible populations, let τ_{mlj} denote the frequency of allele j al locus l in population m, let $\tau_{ml} = (\tau_{ml1}, \ldots, \tau_{mlJ_l})$ be the vector of allele frequencies at locus l in population m and let $\tau_m = (\tau_{m1}, \ldots, \tau_{mL})$. Finally, let $c_i \in \{1, \ldots, M\}$ be the allocation variable of bird i, that is, $c_i = m$ if the bird comes from population m. Following [25], we assume that

$$f(y_{il}|\tau_{ml}) = \mathbb{P}(Y_{il} = y_{il} \mid \tau_{ml}, c_i = m) \propto \prod_{j=1}^{J_l} \tau_{mlj}^{y_{ilj}}, \quad y_{ilj} \in 0, 1, 2.$$

We assume independence across loci, so that, if $Y_i = (Y_{i1}, \ldots, Y_{iL})$ is the multidimensional array of the allelic information at the L loci for individual i, we have

$$f(y_i \mid \tau_m, c_i = m) = \mathbb{P}(Y_i = y_i \mid \tau_m, c_i = m) = \prod_{l=1}^{L} f(y_{il} \mid \tau_{ml}). \tag{10.1}$$

We fit Model (2.1), with the sampling model defined in equation (10.1). The mixing measure is a finite Dirichlet process as in Supplementary Material Appendix D.2.1, with the following prior specification: M has a shifted Poisson prior distribution with parameter Λ, P_0 is the convolution of L independent Dirichlet distributions with parameter 1, γ in the finite Dirichlet process has a Gamma prior with parameter $(0.1, 0.1)$, Λ has a Gamma prior with parameter $(3/2, 1/2)$. For the parameter γ we have specified a vague prior distribution, while the hyperparameters in the prior for Λ are chosen so that the prior mean is 3, corresponding to the three geographical fragments, and the prior variance is large. We employ the conditional algorithm described in Supplementary Material Appendix E to perform posterior inference. The mode of the posterior distribution for k is at 3 ($\mathbb{E}(k \mid \text{data}) = 3$, $\text{Var}(k \mid \text{data}) = 0.03$), as well as the one of the posterior of M ($\mathbb{E}(M \mid \text{data}) = 3.12$, $\text{Var}(M \mid \text{data}) = 0.42$). From Figure 6 it is evident that the three clusters coincide with the three geographical fragments, except in

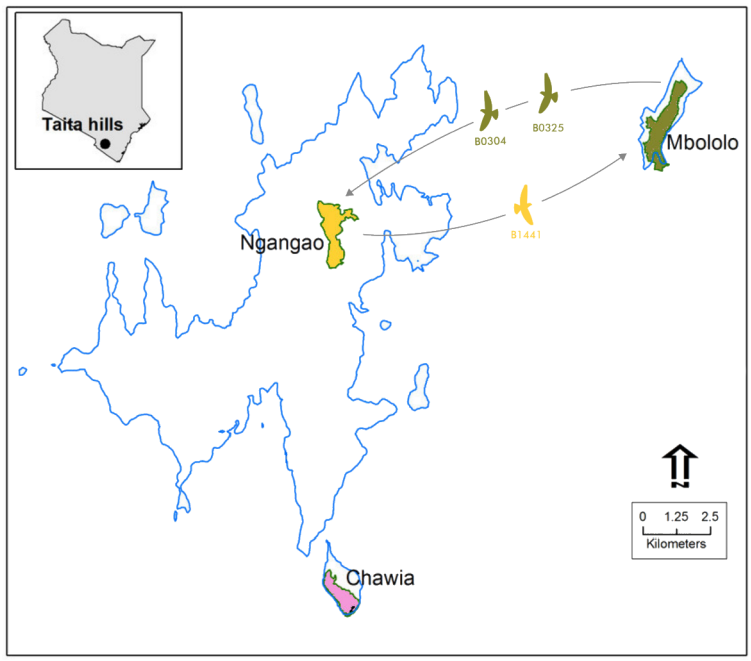

FIG. 6. *Posterior estimate of the clustering allocation*: *each colour correspond to a cluster. Note that he two green thrushes have been captured in Ngangao, but have the genetic profile of the Mbololo birds. The opposite is true for thrush B*1441.

three cases where the birds appear to be out of the obvious clusters. This could be due to rare migration events [18]. We have also fitted the same model using the marginal algorithm for FDMM and for the σ-Stable mixture model. In the latter case the prior for σ is Uniform(0, 1), for a fairer comparison, as we specify a vague prior distribution on γ. The clustering results and posterior inference on k and M are very similar, with the marginal algorithm for the σ-Stable process approximately two times slower than the other two. Moreover, looking at the the posterior distribution of the hyperparameters Λ of q_M, γ for the Dirichlet prior and σ for the σ-Stable process, we notice that the correlation between the hyper-parameters is higher for the marginal algorithm for the FDMM and smallest for the σ-Stable. We find that the Spearman's correlation coefficient between Λ and γ is -0.47, and -0.19 for Λ and σ (see Figure 7). High correlation between parameters deteriorates the mixing of the algorithm, for instance, the effective sample size of Λ is equal to 2945 for the marginal algorithm for FDMM, 3925 for the conditional algorithm for FDMM, 4006 for the marginal algorithm for the σ-Stable.

An important goal of population structure analysis is not only to uncover the group structure of the observations, but also to identify variables that best distinguish the different populations. The results could lead to a better understanding of the evolutionary patterns of population differentiation. To this end we would like to identify the microsatellite loci that most influence the clustering structure. Variable selection for clustering is a challenging problem since there is no observed response to inform the selection and the inclusion of unnecessary variables could complicate or mask the recovery of the clusters. As such there are few contributions in the literature. Here we opt for a model choice method proposed by [20] in the generalised linear model framework, which we adapt to our context. The approach of [20] focuses on the predictive properties of a model and, employing the Kullback–Leibler distance as discrepancy measure, aims to assess the relevance of some restriction on the parameter Θ (leading to a simpler model) with respect to a full model described by a density

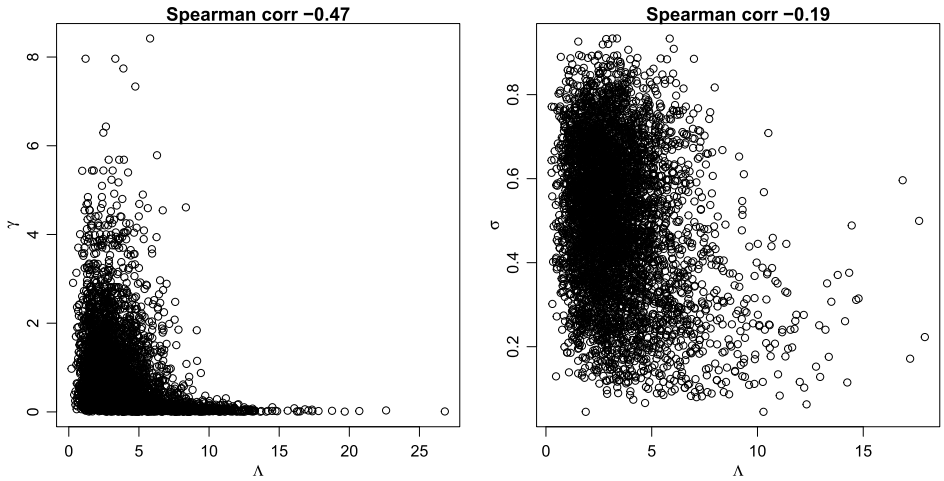

FIG. 7. *Scatterplot between posterior samples of γ v.s. Λ for the finite Dirichlet process* (*right panel*) *and the σ v.s. Λ for the σ-Stable* (*left panel*). *The samples are obtained in both cases from marginal algorithm.*

$f(y \mid \theta)$. More in details, for each locus l, let $Y_l = (Y_{1l}, \ldots, Y_{nl})$, $\theta_{il} = \tau_{ml}$ if $c_i = m$ and $\theta_l = (\theta_{1l}, \ldots, \theta_{nl})$. Let $f(y_l \mid \theta_l)$ be the full general mixture model:

$$f(y_l \mid \theta_l) \propto \prod_{i=1}^{n} \prod_{j=1}^{J_l} \theta_{ilj}^{y_{ijl}}. \tag{10.2}$$

We define a model choice hypothesis H_0 through a restriction on the parameter space, that is, $\theta_l \in \Theta_0 \subset \Theta$, where Θ_0 is the subset of the parameter space such that $\theta_{ilj} = \widetilde{\theta}_{lj}$ for each i. In our application, H_0 represents a fully parametric model for locus l. [20] define the projection $\theta_l^{\perp}$ of θ_l according to the Kullback–Leibler distance d to be the point in Θ_0 that achieves the infimum

$$d\{f(\cdot \mid \theta_l), f(\cdot \mid \theta_l^{\perp})\} = \inf_{\widetilde{\theta}_l \in \Theta_0} d\{f(\cdot \mid \theta_l), f(\cdot \mid \widetilde{\theta}_l)\},$$

where $\widetilde{\theta}_l = (\widetilde{\theta}_{l1}, \ldots, \widetilde{\theta}_{lJ_l})$ and $f(\cdot \mid \theta_l^{\perp})$ is the projection of $f(\cdot \mid \theta_l)$. Obviously small values of d support H_0. We opt for this approach because, instead of phrasing the problem in terms of the classical dichotomy between null and alternative hypothesis, it interprets model choice in terms of the approximation efficacy of a more parsimonious model, focusing on whether or not θ_l is far away from the subspace Θ_0. In Figure 8, we show the posterior distribution (under the FDMM) of $d\{f(\cdot \mid \theta_l), f(\cdot \mid \theta_l^{\perp})\}$ for each locus l. It is evident that locus PC3 contributes the least to the clustering structure as the distance is concentrated near zero, implying the its allele frequencies are similar across Taita thrush populations. The other loci, in particular PAT43, present allele frequency differences among the three groups, which in our case well correspond to geographical locations.

11. Conclusions. In this work, we contribute to the growing understanding of mixture models by providing an unifying framework which encompasses both finite and infinite mixtures. The construction we propose differs from this previous attempts (see [34, 35]), as it is based on the normalization of a point process, which is a standard trick in Bayesian nonparametrics. We introduce the Norm-IFFP prior process and we provide theoretical results characterizing the induced prior on the partition of the observations and the posterior distribution of this process. This construction is very general, allowing the definition of prior processes beyond the finite Dirichlet mixture model. We give consistency results for both the

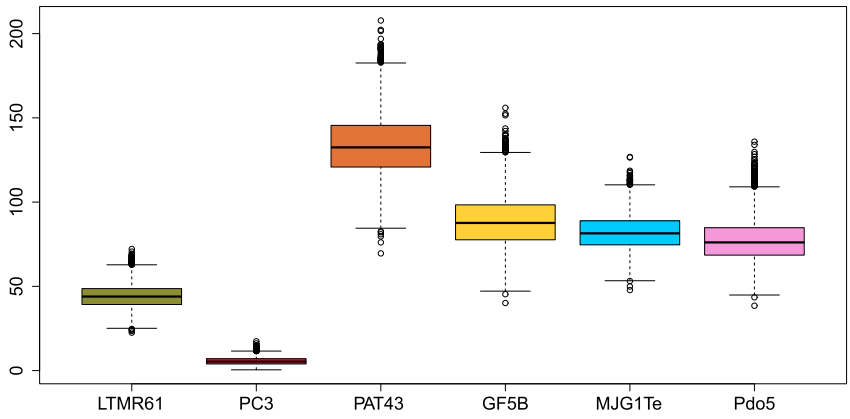

FIG. 8. *Posterior distribution of the KL divergence for each microsatellite locus.*

number of components and of clusters. Our framework allows for efficient computations (inherited from the nonparametric construction) and for data driven estimation of both number of clusters and components, as well as of any functional of interest.

Acknowledgement. We would like to thank Dr Peter Galbusera at the Royal Zoological Society of Antwerp for sharing the enriched Taita Thrush Dataset. We are also grateful to Judith Rousseau, Igor Prünster and XuanLong Nguyen for their advice.

Maria de Iorio is also affiliated to the Department of Statistical Science at University College London (UK). Raffaele Argiento is also affiliated to Collegio Carlo Alberto Torino (IT).

Funding. Dr Argiento is grateful to National University of Singapore for the funding provided.

SUPPLEMENTARY MATERIAL

Supplement to "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models." (DOI: 10.1214/22-AOS2201SUPP; .pdf). We provide an extensive simulations study, further results for the Galaxy dataset, important examples, details of the MCMC algorithms and proofs of the theorems in the main manuscript.

REFERENCES

[1] ALDOUS, D. J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII*—1983. *Lecture Notes in Math.* **1117** 1–198. Springer, Berlin. MR0883646 https://doi.org/10.1007/BFb0099421

[2] ARGIENTO, R., CREMASCHI, A. and VANNUCCI, M. (2020). Hierarchical normalized completely random measures to cluster grouped data. *J. Amer. Statist. Assoc.* **115** 318–333. MR4078466 https://doi.org/10.1080/01621459.2019.1594833

[3] ARGIENTO, R. and DE IORIO, M. (2022). Supplement to "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models." https://doi.org/10.1214/22-AOS2201SUPP

[4] BODIN, B., IORIO, M. D. and ARGIENTO, R. (2020). AntMAN: Anthology of Mixture ANalysis tools.

[5] CALLENS, T., GALBUSERA, P., MATTHYSEN, E., DURAND, E. Y., GITHIRU, M., HUYGHE, J. R. and LENS, L. (2011). Genetic signature of population fragmentation varies with mobility in seven bird species of a fragmented Kenyan cloud forest. *Mol. Ecol.* **20** 1829–1844.

[6] CHARALAMBIDES, C. A. (2005). *Combinatorial Methods in Discrete Distributions. Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2131068 https://doi.org/10.1002/0471733180

[7] DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II: General Theory and Structure*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR2371524 https://doi.org/10.1007/978-0-387-49835-5

[8] DELLAPORTAS, P. and PAPAGEORGIOU, I. (2006). Multivariate mixtures of normals with unknown number of components. *Stat. Comput.* **16** 57–68. MR2224189 https://doi.org/10.1007/s11222-006-5338-6

[9] DEVROYE, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Trans. Model. Comput. Simul.* **19** 18.

[10] ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

[11] FALUSH, D., STEPHENS, M. and PRITCHARD, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164** 1567–1587.

[12] FAVARO, S., NIPOTI, B. and TEH, Y. W. (2015). Random variate generation for Laguerre-type exponentially tilted α-stable distributions. *Electron. J. Stat.* **9** 1230–1242. MR3355756 https://doi.org/10.1214/15-EJS1033

[13] FAVARO, S. and TEH, Y. W. (2013). MCMC for normalized random measure mixture models. *Statist. Sci.* **28** 335–359. MR3135536 https://doi.org/10.1214/13-STS422

[14] FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models. Springer Series in Statistics*. Springer, New York. MR2265601

[15] FRÜHWIRTH-SCHNATTER, S., CELEUX, G. and ROBERT, C. P., eds. (2019). *Handbook of Mixture Analysis. Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3889980

[16] FRÜHWIRTH-SCHNATTER, S. and MALSINER-WALLI, G. (2019). From here to infinity: Sparse finite versus Dirichlet process mixtures in model-based clustering. *Adv. Data Anal. Classif.* **13** 33–64. MR3935190 https://doi.org/10.1007/s11634-018-0329-y

[17] FRÜHWIRTH-SCHNATTER, S., MALSINER-WALLI, G. and GRÜN, B. (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Anal.* **16** 1279–1307. MR4381135 https://doi.org/10.1214/21-BA1294

[18] GALBUSERA, P., LENS, L., SCHENCK, T., WAIYAKI, E. and MATTHYSEN, E. (2000). Genetic variability and gene flow in the globally, critically-endangered Taita thrush. *Conserv. Genet.* **1** 45–55.

[19] GELFAND, A. E. and KOTTAS, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *J. Comput. Graph. Statist.* **11** 289–305. MR1938136 https://doi.org/10.1198/106186002760180518

[20] GOUTIS, C. and ROBERT, C. P. (1998). Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika* **85** 29–37. MR1627250 https://doi.org/10.1093/biomet/85.1.29

[21] GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810 https://doi.org/10.1093/biomet/82.4.711

[22] GRÜN, B., MALSINER-WALLI, G. and FRÜHWIRTH-SCHNATTER, S. (2022). How many data clusters are in the Galaxy data set? *Adv. Data Anal. Classif.* **16** 325–349. MR4440853 https://doi.org/10.1007/s11634-021-00461-8

[23] GUHA, A., HO, N. and NGUYEN, X. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli* **27** 2159–2188. MR4303879 https://doi.org/10.3150/20-BEJ1275

[24] HOFERT, M. (2011). Sampling exponentially tilted stable distributions. *ACM Trans. Model. Comput. Simul.* **22** Art. 3, 11. MR2955859 https://doi.org/10.1145/2043635.2043638

[25] HUELSENBECK, J. P. and ANDOLFATTO, P. (2007). Inference of population structure under a Dirichlet process model. *Genetics* **175** 1787–1802.

[26] HUELSENBECK, J. P., ANDOLFATTO, P. and HUELSENBECK, E. T. (2011). Structurama: Bayesian inference of population structure. *Evol. Bioinform.* **7** 55–59. https://doi.org/10.4137/EBO.S6761

[27] ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729 https://doi.org/10.1198/016214501750332758

[28] ISHWARAN, H. and ZAREPOUR, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canad. J. Statist.* **30** 269–283. MR1926065 https://doi.org/10.2307/3315951

[29] JACOD, J. and SHIRYAEV, A. N. (1987). *Limit Theorems for Stochastic Processes. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **288**. Springer, Berlin. MR0959133 https://doi.org/10.1007/978-3-662-02514-7

[30] JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36** 76–97. MR2508332 https://doi.org/10.1111/j.1467-9469.2008.00609.x

[31] KALLI, M., GRIFFIN, J. E. and WALKER, S. G. (2011). Slice sampling mixture models. *Stat. Comput.* **21** 93–105. MR2746606 https://doi.org/10.1007/s11222-009-9150-y

[32] KOMÁREK, A. (2009). A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Comput. Statist. Data Anal.* **53** 3932–3947. MR2744295 https://doi.org/10.1016/j.csda.2009.05.006

[33] LIJOI, A. and PRÜNSTER, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*. *Camb. Ser. Stat. Probab. Math.* **28** 80–136. Cambridge Univ. Press, Cambridge. MR2730661

[34] MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Stat. Comput.* **26** 303–324. MR3439375 https://doi.org/10.1007/s11222-014-9500-2

[35] MALSINER-WALLI, G., FRÜHWIRTH-SCHNATTER, S. and GRÜN, B. (2017). Identifying mixtures of mixtures using Bayesian estimation. *J. Comput. Graph. Statist.* **26** 285–295. MR3640186 https://doi.org/10.1080/10618600.2016.1200472

[36] MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

[37] MILLER, J. W. and HARRISON, M. T. (2018). Mixture models with a prior on the number of components. *J. Amer. Statist. Assoc.* **113** 340–356. MR3803469 https://doi.org/10.1080/01621459.2016.1255636

[38] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804 https://doi.org/10.2307/1390653

[39] NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422 https://doi.org/10.1214/12-AOS1065

[40] NOBILE, A. (1994). *Bayesian Analysis of Finite Mixture Distributions*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Carnegie Mellon Univ. MR2692049

[41] NOBILE, A. (2004). On the posterior distribution of the number of components in a finite mixture. *Ann. Statist.* **32** 2044–2073. MR2102502 https://doi.org/10.1214/009053604000000788

[42] PITMAN, J. (1996). Blackwell–Macqueen urn scheme. In *Statistics*, *Probability*, *and Game Theory*: *Papers in Honor of David Blackwell* **30** 245.

[43] PITMAN, J. (2006). *Combinatorial Stochastic Processes*. *Lecture Notes in Math.* **1875**. Springer, Berlin. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. MR2245368

[44] POLLARD, H. (1946). The representation of $e^{-x^{\lambda}}$ as a Laplace integral. *Bull. Amer. Math. Soc.* **52** 908–910. MR0018286 https://doi.org/10.1090/S0002-9904-1946-08672-3

[45] PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.

[46] PRITCHARD, J. K. and WEN, W. (2003). Documentation for STRUCTURE software: Version 2.3.X.

[47] 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. MR1483213 https://doi.org/10.1111/1467-9868.00095

[48] ROEDER, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85** 617–624.

[49] ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454 https://doi.org/10.1111/j.1467-9868.2011.00781.x

[50] STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903 https://doi.org/10.1214/aos/1016120364