Statistical Science
2013, Vol. 28, No. 3, 335–359
DOI: 10.1214/13-STS422
© Institute of Mathematical Statistics, 2013

MCMC for Normalized Random Measure Mixture Models

Stefano Favaro[1] and Yee Whye Teh

Abstract. This paper concerns the use of Markov chain Monte Carlo methods for posterior sampling in Bayesian nonparametric mixture models with normalized random measure priors. Making use of some recent posterior characterizations for the class of normalized random measures, we propose novel Markov chain Monte Carlo methods of both marginal type and conditional type. The proposed marginal samplers are generalizations of Neal's well-regarded Algorithm 8 for Dirichlet process mixture models, whereas the conditional sampler is a variation of those recently introduced in the literature. For both the marginal and conditional methods, we consider as a running example a mixture model with an underlying normalized generalized Gamma process prior, and describe comparative simulation results demonstrating the efficacies of the proposed methods.

Key words and phrases: Bayesian nonparametrics, hierarchical mixture model, completely random measure, normalized random measure, Dirichlet process, normalized generalized Gamma process, MCMC posterior sampling method, marginalized sampler, Algorithm 8, conditional sampler, slice sampling.

1. INTRODUCTION

Mixture models provide a statistical framework for modeling data where each observation is assumed to have arisen from one of k groups, with k possibly unknown, and each group being suitably modeled by a distribution function from some parametric family. The distribution function of each group is referred to as a component of the mixture model and is weighted by the relative frequency of the group in the population. Specifically, assuming k being fixed, a collection of observations $(Y_1, \ldots, Y_n)$ is modeled as independent draws from a mixture distribution function with k components, that is,

$$Y_i \overset{\text{ind}}{\sim} \sum_{j=1}^{k} \tilde{J}_j f(\cdot|\tilde{X}_j), \tag{1.1}$$

where $f(\cdot|\tilde{X})$ is a given parametric family of distribution functions indexed by a parameter $\tilde{X}$ and $(\tilde{J}_1, \ldots, \tilde{J}_k)$ are the mixture proportions constrained to be nonnegative and sum to unity. A convenient formulation of the mixture model (1.1) can be stated in terms of latent allocation random variables, namely, each observation Y_i is assumed to arise from a specific but unknown component Z_i of the mixture model. Accordingly, an augmented version of (1.1) can be written in terms of a collection of latent random variables $(Z_1, \ldots, Z_n)$, independent and identically distributed with probability mass function $\mathbb{P}[Z_i = j] = \tilde{J}_j$, such that the observations are modeled as

$$Y_i|Z_i \overset{\text{ind}}{\sim} f(\cdot|\tilde{X}_{Z_i}). \tag{1.2}$$

Integrating out the random variables $(Z_1, \ldots, Z_n)$ then yields (1.1). In a Bayesian setting the formulation of the mixture model (1.2) is completed by specifying

Stefano Favaro is Assistant Professor of Statistics, Department of Economics and Statistics, University of Torino, C.so Unione Sovietica 218/bis, 10134 Torino, Italy (e-mail: stefano.favaro@unito.it). Yee Whye Teh is Professor of Statistical Machine Learning, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX13TG, United Kingdom (e-mail: y.w.teh@stats.ox.ac.uk).
[1]Also affiliated with Collegio Carlo Alberto, Moncalieri, Italy.

suitable prior distributions for the unknown quantities that are objects of the inferential analysis: the parameter $(\tilde{X}_1, \ldots, \tilde{X}_k)$ and the vector of proportions $(\tilde{J}_1, \ldots, \tilde{J}_k)$. We refer to the monographs by Titterington et al. [83] and McLachlan and Basford [55] for accounts on mixture models with a fixed number of components. Markov chain Monte Carlo (MCMC) methods for Bayesian analysis of mixture models with a fixed number of components was presented in Dielbot and Robert [10].

As regards the general case where the number of components is unknown, a direct approach has been considered in Richardson and Green [79], who modeled the unknown k by mixing over the fixed k case, and made a fully Bayesian inference using the reversible jump MCMC methods proposed in Green [24]. See also Stephens [82] and references therein for some developments on such an approach, whereas different proposals can be found in the papers by Mengersen and Roberts [57], Raftery [74] and Roeder and Wasserman [81]. An early and fruitful approach, still in the context of mixture models with an unknown number k of components, was proposed in Escobar [11] who treated the problem in a Bayesian nonparametric setting by means of a prior distribution based on the Dirichlet process (DP) of Ferguson [16]. This approach arises as a major development of some earlier results in Lo [51] and it is nowadays the subject of a rich and active literature.

In this paper we deal with mixture models with an unknown number of components. In particular, we focus on a Bayesian nonparametric approach with the specification of a class of prior distributions generalizing the DP prior. In the Bayesian nonparametric setting the central role is played by a discrete random probability measure $\tilde{\mu}$ defined on a suitable measurable space $\mathbb{X}$, an example being the DP, whose distribution acts as a nonparametric prior. The basic idea is that since $\tilde{\mu}$ is discrete, it can be written as

$$\tilde{\mu} = \sum_{j \geq 1} \tilde{J}_j \delta_{\tilde{X}_j},$$

where $(\tilde{J}_j)_{j \geq 1}$ is a sequence of nonnegative random weights that add up to one and $(\tilde{X}_j)_{j \geq 1}$ is a sequence of $\mathbb{X}$-valued random locations independent of $(\tilde{J}_j)_{j \geq 1}$. Given $\tilde{\mu}$ and a collection of continuous observations $(Y_1, \ldots, Y_n)$, a Bayesian nonparametric mixture model admits a hierarchical specification in terms of a collection of independent and identically distributed latent random variables $(X_1, \ldots, X_n)$. Formally,

$$\begin{aligned} Y_i | X_i &\overset{\text{ind}}{\sim} F(\cdot | X_i), \\ X_i | \tilde{\mu} &\overset{\text{i.i.d.}}{\sim} \tilde{\mu}, \\ \tilde{\mu} &\sim P, \end{aligned} \tag{1.3}$$

where P denotes the nonparametric prior distribution and $F(\cdot|X_i)$ is a probability distribution parameterized by the random variable X_i and admitting a distribution function $f(\cdot|X_i)$. Note that, due to the discreteness of $\tilde{\mu}$, each random variable X_i will take on value $\tilde{X}_j$ with probability $\tilde{J}_j$ for each $j \geq 1$, and the hierarchical model (1.3) is equivalent to saying that observations $(Y_1, \ldots, Y_n)$ are independent and identically distributed according to a probability distribution F with random distribution function

$$f(\cdot) = \int_{\mathbb{X}} f(\cdot|x)\tilde{\mu}(dx) = \sum_{j \geq 1} \tilde{J}_j f(\cdot|\tilde{X}_j). \tag{1.4}$$

This is a mixture of distribution functions with a countably infinite number of components. The probability distribution $F(\cdot|X_i)$ is termed the mixture kernel, whereas the underlying distribution P is termed the mixing distribution or, alternatively, the mixing measure. Note that, since $\tilde{\mu}$ is discrete, each pair of the latent random variables $(X_1, \ldots, X_n)$ will take on the same value with positive probability, with this value corresponding to a component of the mixture model. In this way, the latent random variables allocate the observations $(Y_1, \ldots, Y_n)$ to a random number of components, thus naturally providing a model for the unknown number of components. Under the assumption of $\tilde{\mu}$ being a Dirichlet process, the model (1.4) was introduced by Lo [51] and it is known in Bayesian nonparametrics as the DP mixture model.

The reason of the success of the Bayesian nonparametric approach in the analysis of mixture models, as pointed out in the paper by Green and Richardson [25], is that it exploits the discreteness of $\tilde{\mu}$, thus providing a flexible model for clustering items of various kinds in a hierarchical setting without explicitly specifying the number of components. Bayesian nonparametrics is now the subject of a rich and active literature spanning applied probability, computational statistics and machine learning. Beyond mixture analysis, Bayesian nonparametrics has been applied to survival analysis by Hjort [29], to feature allocation models by Griffiths and Ghahramani [28] and Broderick et al. [6] and to regression (see the monograph by Rasmussen and Williams [77]), among others. The reader is referred to the comprehensive monograph edited by Hjort et al. [30] for a collection of reviews on recent developments in Bayesian nonparametrics.

Several MCMC methods have been proposed for posterior sampling from the DP mixture model. Early works exploited the tractable marginalization of $\tilde{\mu}$ with respect to the DP mixing distribution, thus removing the infinite-dimensional aspect of the inferential problem. The main references in this research area are represented by the sampling methods originally devised in Escobar [11, 12], MacEachern [52] and Escobar and West [13], and by the subsequent variants proposed in MacEachern [53] and MacEachern and Müller [54]. In Bayesian nonparametrics these MCMC methods are typically referred to as marginal samplers and, as noted by Ishwaran and James [31], apply to any mixture model for which the system of predictive distributions induced by $\tilde{\mu}$ is known explicitly. The reader is referred to Neal [61] for a detailed overview of marginal samplers for DP mixture models and for some noteworthy developments in this direction, such as the well-known Algorithm 8 which is now a gold standard against which other methods are compared.

An alternative family of MCMC methods for posterior sampling from the DP mixture model is typically referred to as conditional samplers and relies on the simulation from the joint posterior distribution, including sampling of the mixing distribution $\tilde{\mu}$. These methods do not remove the infinite-dimensional aspect of the problem and instead focus on finding appropriate ways for sampling a finite but sufficient number of the atoms of $\tilde{\mu}$. Ishwaran and James [31] proposed the use of a deterministic truncation level by fixing the number of atoms and then bounding the resulting truncation error introduced; the same authors also showed how to extend the proposed method to any mixing distribution $\tilde{\mu}$ in the class of the so-called stick-breaking random probability measures. Alternatively, Muliere and Tardella [58] proposed the use of a random truncation level that allows one to set in advance the truncation error. The idea of a random truncation has been recently developed by Papaspiliopoulos and Roberts [68] who proposed a Metropolis–Hastings sampling scheme, while Walker [85] proposed the use of a slice sampling scheme. See also Papaspiliopoulos [67] and Kalli et al. [39] for further noteworthy improvements and developments of conditional samplers with random truncation levels.

It is apparent that one can replace the DP mixing distribution with the distribution of any other discrete random probability measure. Normalized random measures (NRMs) form a large class of such random probability measures. This includes the DP as a special case, and was first proposed as a class of prior models in Bayesian nonparametrics by Regazzini et al. [78]. See also James [33]. Nieto-Barajas et al. [64] later proposed using NRMs as the mixing distribution in (1.4), while Lijoi et al. [46–48] investigated explicit examples of NRMs such as the generalized DP, the normalized σ-stable process, the normalized inverse Gaussian process (NIGP) and the normalized generalized Gamma process (NGGP). Various structural properties of the class of NRMs have been extensively investigated by James [34], Nieto-Barajas et al. [64], James et al. [35–37] and Trippa and Favaro [84]. Recently James et al. [36] described a slightly more general definition of NRMs in terms of the normalization of the so-called completely random measures (CRMs), a class of discrete random measures first introduced by Kingman [40]. We refer to Lijoi and Prünster [49] for a comprehensive and stimulating overview of nonparametric prior models defined within the unifying framework of CRMs.

In this paper we study MCMC methods of both marginal and conditional types for posterior sampling from the mixture model (1.4) with a NRM mixing distribution. We refer to such a model as a NRM mixture model. Historically, the first MCMC methods for posterior sampling from NRM mixture models are of the same type as those proposed by MacEachern [52] and Escobar and West [13] for DP mixture models: they rely on the system of predictive distributions induced by the NRM mixing distribution. See James et al. [36] for details. Typically these methods can be difficult to implement and computationally expensive due to the necessary numerical integrations. To overcome this drawback, we propose novel MCMC methods of marginal type for NRM mixture models. Our methods are generalizations of Neal's celebrated Algorithm 8 [61] to NRM mixture models, and represent, to the best of our knowledge, the first marginal type samplers for NRM mixture models that can be efficiently implemented and do not require numerical integrations. As opposed to MCMC methods of marginal type, conditional samplers for NRM mixture models have been well explored in the recent literature by Nieto-Barajas and Prünster [63], Griffin and Walker [27], Favaro and Walker [15] and Barrios et al. [2]. Here we propose some improvements to the existing conditional slice sampler recently introduced by Griffin and Walker [27].

For concreteness, throughout the present paper we consider as a running example the NGGP mixture model, namely, a mixture model of the form (1.4) with the specification of a NGGP mixing distribution. The NGGP is a recently studied NRM generalizing the DP and featuring appealing theoretical properties

which turns out to be very useful in the context of mixture modeling. We refer to Pitman [70], Lijoi et al. [48, 50] for an account on these properties with a view toward Bayesian nonparametrics. In particular, the NGGP mixture model has been investigated in depth by Lijoi et al. [48] who proposed a comprehensive and comparative study with the DP mixture model emphasizing the advantages of such a generalization.

The paper is structured as follows. Section 2 introduces NRMs and defines the induced class of NRM mixture models. In Section 3 we present the proposed MCMC methods, of both marginal type and conditional type, for posterior sampling from NRM mixture models. Section 4 reports on simulation results comparing the proposed methods on a NRM mixture model with an underlying NGGP mixing distribution. A final discussion is presented in Section 5.

2. NORMALIZED RANDOM MEASURES

We review the class of NRMs with particular emphasis on their posterior characterization recently provided by James et al. [36]. Such a characterization will be crucial in Section 3 for devising MCMC methods for posterior sampling from NRM mixture models.

2.1 Completely Random Measures

To be self-contained, we start with a description of CRMs. See the monograph by Kingman [41] and references therein for details on such a topic. Let $\mathbb{X}$ be a complete and separable metric space endowed with the corresponding Borel σ-algebra $\mathscr{X}$. A CRM on $\mathbb{X}$ is a random variable μ taking values on the space of boundedly finite measures on $(\mathbb{X}, \mathscr{X})$ and such that for any collection of disjoint sets $A_1, \ldots, A_n$ in $\mathscr{X}$, with $A_i \cap A_j = \varnothing$ for $i \neq j$, the random variables $\mu(A_1), \ldots, \mu(A_n)$ are mutually independent. Kingman [40] showed that a CRM can be decomposed into the sum of three independent components: a nonrandom measure, a countable collection of nonnegative random masses at nonrandom locations and a countable collection of nonnegative random masses at random locations. In this paper we consider CRMs consisting solely of the third component, namely, a collection of random masses $(J_j)_{j\geq 1}$ at random locations $(\tilde{X}_j)_{j\geq 1}$, that is,

$$\mu = \sum_{j\geq 1} J_j \delta_{\tilde{X}_j}. \tag{2.1}$$

The distribution of μ can be characterized in terms of the distribution of the random point set $(J_j, \tilde{X}_j)_{j\geq 1}$ as a Poisson random measure on $\mathbb{R}^+ \times \mathbb{X}$ with mean measure ν, which is typically referred to as the Lévy intensity measure. As an example, Figure 1 demonstrates a draw of a CRM along with its Lévy intensity measure.

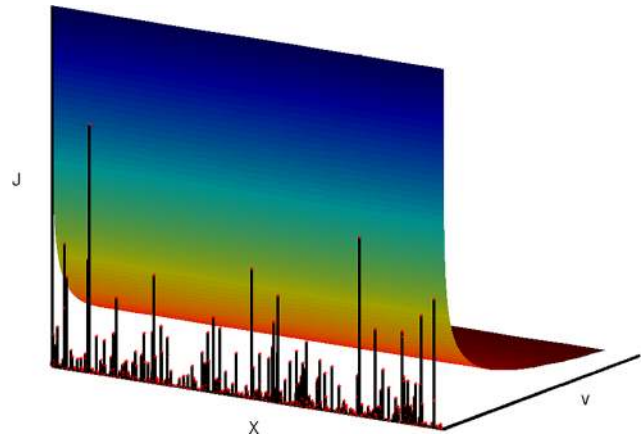

FIG. 1. *A draw $\sum_{j\geq 1} J_j \delta_{\tilde{X}_j}$ from a CRM. Each stick denotes an atom in the CRM, with mass given by its height J_j and location given by $\tilde{X}_j$. Behind the CRM is the density of its Lévy intensity measure ν. The random point set $\{(J_j, \tilde{X}_j)\}_{j\geq 1}$ is described by a Poisson process with intensity measure given by the Lévy measure ν.*

For our purposes we focus on the so-called homogeneous CRMs, namely, CRMs characterized by a Lévy intensity measure ν factorizing as $\nu(ds, dy) = \rho(ds)\mu_0(dy)$, for a nonnegative measure ρ absolutely continuous with respect to Lebesgue measure and a nonatomic probability measure μ_0 over $(\mathbb{X}, \mathscr{X})$. Such a factorization implies the independence between the random masses $(J_j)_{j\geq 1}$ and the random locations $(\tilde{X}_j)_{j\geq 1}$ in (2.1). Hence, without loss of generality, the random locations can be assumed to be independent and identically distributed according to the base distribution μ_0, while the distribution of the random masses $(J_j)_{j\geq 1}$ is governed by the Lévy measure ρ: it is distributed according to a Poisson random measure with intensity ρ.

2.2 Homogeneous Normalized Random Measures

Homogeneous CRMs provide a fundamental tool for defining almost surely discrete nonparametric priors via the so-called normalization approach. Specifically, consider a homogeneous CRM μ with Lévy intensity measure $\nu(ds, dy) = \rho(ds)\mu_0(dy)$ and denote by $T = \mu(\mathbb{X}) = \sum_{j\geq 1} J_j$ the corresponding total mass. Then one can define an almost surely discrete random probability measure on $\mathbb{X}$ as follows:

$$\tilde{\mu} = \frac{\mu}{T} = \sum_{j\geq 1} \tilde{J}_j \delta_{\tilde{X}_j}, \tag{2.2}$$

where $(\tilde{J}_j)_{j\geq 1}$ is a sequence of random probabilities defined by normalizing, with respect to T, the se-

quence of random masses $(J_j)_{j\geq 1}$. To ensure that the normalization in (2.2) is a well-defined operation, the random variable T has to be positive and finite almost surely; this is guaranteed by a well-known condition on the Lévy measure ρ, that is,

$$\begin{aligned} &\int_{\mathbb{R}^+} \rho(ds) = +\infty, \\ &\int_{\mathbb{R}^+} (1 - e^{-s})\rho(ds) < +\infty. \end{aligned} \tag{2.3}$$

The random probability measure $\tilde{\mu}$ is known from James et al. [36] as a homogeneous NRM with Lévy measure ρ and base distribution μ_0. See also Regazzini et al. [78] for an early definition of NRMs. The idea of normalizing CRMs, in order to define almost surely discrete nonparametric priors, is clearly inspired by the seminal paper of Ferguson [16] who introduced the DP as a normalized Gamma CRM.

EXAMPLE 2.1 (DP). A Gamma CRM is a homogeneous CRM with Lévy intensity measure of the form

$$\rho_a(ds)\mu_0(dy) = as^{-1}e^{-s}\,ds\mu_0(dy),$$

where $a > 0$. We denote a Gamma CRM by μ_a and its total mass by T_a. Note that the Lévy measure ρ_a satisfies the condition (2.3), thus ensuring that the NRM

$$\tilde{\mu}_a = \frac{\mu_a}{T_a}$$

is a well-defined random probability measure. Specifically, $\tilde{\mu}_a$ is a DP with concentration parameter a and base distribution μ_0.

Other examples of homogeneous NRMs have been introduced in the recent literature. Notable among these in terms of both flexibility and sufficient mathematical tractability is the normalized generalized Gamma process (NGGP). Such a process, first introduced by Pitman [70] and then investigated in Bayesian nonparametrics by Lijoi et al. [48], is defined by normalizing the so-called generalized Gamma CRM proposed by Brix [5]. Throughout this paper we will consider the NGGP as a running example.

EXAMPLE 2.2 (NGGP). A generalized Gamma CRM is a homogeneous CRM with Lévy intensity measure of the form

$$\begin{aligned} &\rho_{a,\sigma,\tau}(ds)\mu_0(dy) \\ &\quad = \frac{a}{\Gamma(1-\sigma)}s^{-\sigma-1}\mathrm{e}^{-\tau s}\,ds\,\mu_0(dy), \end{aligned} \tag{2.4}$$

where $a > 0$, $\sigma \in (0, 1)$ and $\tau \geq 0$. We denote a generalized Gamma CRM by $\mu_{a,\sigma,\tau}$ and its total mass by $T_{a,\sigma,\tau}$. Note that the Lévy measure $\rho_{a,\sigma,\tau}$ satisfies the condition (2.3), thus ensuring that the NRM

$$\tilde{\mu}_{a,\sigma,\tau} = \frac{\mu_{a,\sigma,\tau}}{T_{a,\sigma,\tau}}$$

is a well-defined random probability measure. Specifically, $\tilde{\mu}_{a,\sigma,\tau}$ is a NGGP with parameter (a, σ, τ) and base distribution μ_0.

The NGGP includes as special cases most of the discrete random probability measures currently applied in Bayesian nonparametric mixture modeling. The DP represents a special case of a NGGP given by $\tilde{\mu}_{a,0,1}$. Further noteworthy examples of NGGPs include: the normalized σ-stable process, given by $\tilde{\mu}_{a,\sigma,0}$, first introduced by Kingman et al. [42] in relation to optimal storage problems, and the normalized inverse Gaussian process (NIGP), given by $\tilde{\mu}_{a,1/2,\tau}$, recently investigated by Lijoi et al. [47] in the context of Bayesian nonparametric mixture modeling. As regards the celebrated two-parameter Poisson–Dirichlet process, introduced by Perman et al. [69], this is not a NRM. However, it can be expressed in terms of a suitable mixture of NGGPs. See Pitman and Yor [72] for details on such a representation.

It is worth pointing out that the parameterization of the Lévy intensity measure (2.4) is different from those proposed in the past by Brix [5], Pitman [70] and Lijoi et al. [48]. Such a parameterization uses three parameters rather than two parameters. This is so that our NGGP can easily encompass all the other NRMs mentioned above. The three-parameter formulation does not lead to a strict generalization of the two-parameter formulation since the a and τ parameters are in fact redundant. Indeed, rescaling $\mu_{a,\sigma,\tau}$ by a constant $c > 0$, which does not affect the resulting NRM, leads to a generalized Gamma CRM with parameters $(ac^{\sigma}, \sigma, \tau/c)$.

2.3 Normalized Random Measure Mixture Models

Given a set of n observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$, a NRM mixture model consists of a corresponding set of latent random variables $\mathbf{X} = (X_1, \ldots, X_n)$ conditionally independent and identically distributed given a NRM mixing measure $\tilde{\mu}$. According to the hierarchical formulation (1.3), a NRM mixture model can be stated as follows:

$$\begin{aligned} Y_i|X_i &\overset{\text{ind}}{\sim} F(\cdot|X_i), \\ X_i|\tilde{\mu} &\overset{\text{i.i.d.}}{\sim} \tilde{\mu}, \\ \tilde{\mu} &= \frac{\mu}{T}, \\ \mu &\sim \mathrm{CRM}(\rho, \mu_0), \end{aligned} \tag{2.5}$$

where CRM(ρ, μ_0) denotes the law of the CRM μ with Lévy measure ρ and base distribution μ_0. The rest of this section elaborates on some posterior and marginal characterizations for the NRM mixing measure $\tilde{\mu}$. These characterizations will be useful in deriving the MCMC methods for posterior sampling from the NRM mixture model (2.5).

Because $\tilde{\mu}$ is almost surely discrete, ties may occur among the latent random variables $\mathbf{X}$, so that $\mathbf{X}$ contains $k \leq n$ unique values. Hence, an equivalent representation of $\mathbf{X}$ can be given in terms of the random partition on $[n] := \{1, \ldots, n\}$ induced by the ties and the unique values. Let $\boldsymbol{\pi}$ be the induced random partition of $[n]$, that is, a family of random subsets of $[n]$ such that indices i and j belong to the same subset (cluster) if and only if $X_i = X_j$. For each cluster $c \in \boldsymbol{\pi}$, we denote the corresponding unique value by X_c^*. In the context of mixture modeling, the random partition $\boldsymbol{\pi}$ describes the assignment of observations to the various components, while the unique value X_c^* plays the role of the parameter associated with component c.

The random variables $\mathbf{X}$ are a sample from an exchangeable sequence directed by $\tilde{\mu}$ and, accordingly, the induced random partition $\boldsymbol{\pi}$ is also exchangeable, namely, the probability mass function of $\boldsymbol{\pi}$ depends only on the number of clusters $|\boldsymbol{\pi}|$ and the sizes of the clusters $\{|c| : c \in \boldsymbol{\pi}\}$. Such a probability mass function is known in the literature as the exchangeable partition probability function (EPPF). See the monograph by Pitman [71] and references therein for details on this topic. The EPPF induced by the NRM $\tilde{\mu}$ has been recently characterized by James et al. [36] using an auxiliary random variable U whose conditional distribution, given the total mass T, coincides with a Gamma distribution with shape n and inverse scale T. In particular, the joint conditional distribution of the random variables $\mathbf{X}$ and U, given μ, is

$$\begin{aligned}&\mathbb{P}\big[\boldsymbol{\pi} = \pi, \{X_c^* \in dx_c : c \in \pi\}, U \in du | \mu\big] \\ &\quad = \frac{1}{\Gamma(n)} u^{n-1} e^{-Tu}\, du \prod_{c \in \pi} \mu(dx_c)^{|c|}.\end{aligned} \tag{2.6}$$

The next propositions briefly summarize the posterior characterizations introduced by James et al. [36]. We start by considering the characterization of the EPPF and the system of predictive distributions induced by a NRM $\tilde{\mu}$. Note that such a characterization can be derived from the distribution (2.6) by means of an application of the so-called Palm formula for CRMs. See, for example, Daley and Vere-Jones [9].

PROPOSITION 2.1. *Let $\tilde{\mu}$ be a homogeneous NRM with Lévy measure ρ and base distribution μ_0. The induced joint distribution of $\mathbf{X}$ and U, with $\tilde{\mu}$ marginalized out, is given by*

$$\begin{aligned}&\mathbb{P}\big[\boldsymbol{\pi} = \pi, \{X_c^* \in dx_c : c \in \pi\}, U \in du\big] \\ &\quad = \frac{1}{\Gamma(n)} u^{n-1} e^{-\psi(u)}\, du \prod_{c \in \pi} \kappa_{|c|}(u) \mu_0(dx_c),\end{aligned} \tag{2.7}$$

where $\psi(\cdot)$ denotes the Laplace exponent of the underlying CRM μ and $\kappa_m(u)$ denotes the mth moment of the exponentially tilted Lévy measure $e^{-us}\rho(ds)$, that is,

$$\begin{aligned}\psi(u) &= \int_{\mathbb{R}^+} \big(1 - \mathrm{e}^{-us}\big)\rho(ds), \\ \kappa_m(u) &= \int_{\mathbb{R}^+} s^m e^{-us} \rho(ds).\end{aligned} \tag{2.8}$$

In particular, by marginalizing out the auxiliary random variable U, the EPPF of $\boldsymbol{\pi}$ has the following expression:

$$\mathbb{P}[\boldsymbol{\pi} = \pi] = \int_{\mathbb{R}^+} \frac{1}{\Gamma(n)} u^{n-1} e^{-\psi(u)} \prod_{c \in \pi} \kappa_{|c|}(u)\, du,$$

while the unique values $\{X_c^* : c \in \boldsymbol{\pi}\}$ are independent and identically distributed according to μ_0. Together these characterize the joint distribution of the latent variables $\mathbf{X}$. Accordingly,

$$\begin{aligned}&\mathbb{P}[X_{n+1} \in dx | U, \mathbf{X}] \\ &\quad \propto \kappa_1(U)\mu_0(dx) + \sum_{c \in \boldsymbol{\pi}} \frac{\kappa_{|c|+1}(U)}{\kappa_{|c|}(U)} \delta_{X_c^*}(dx)\end{aligned}$$

is the predictive distribution for a new sample $X_{n+1} \sim \tilde{\mu}$, given U and $\mathbf{X}$ and once $\tilde{\mu}$ is marginalized out.

Note that from the probability distribution (2.7) follows the posterior distribution of U given $\mathbf{X}$, that is,

$$\mathbb{P}[U \in du | \mathbf{X}] \propto u^{n-1} \mathrm{e}^{-\psi(u)}\, du \prod_{c \in \boldsymbol{\pi}} \kappa_{|c|}(u). \tag{2.9}$$

The next proposition completes the posterior characterization for NRMs by showing that the posterior distribution of a homogeneuos CRM μ, given $\mathbf{X}$ and U, is still a CRM.

PROPOSITION 2.2. *Let $\tilde{\mu}$ be a homogeneous NRM with Lévy measure ρ and base distribution μ_0. The posterior distribution of the underlying homogeneous CRM μ, given $\mathbf{X}$ and U, corresponds to*

$$\mu | U, \mathbf{X} \sim \mu' + \sum_{c \in \boldsymbol{\pi}} J_c' \delta_{X_c^*}, \tag{2.10}$$

where μ' is a homogeneous CRM with an exponential tilted Lévy intensity measure of the form

$$\nu'(ds, dy) = \mathrm{e}^{-Us}\rho(ds)\mu_0(dy)$$

and where the random masses $\{J'_c : c \in \boldsymbol{\pi}\}$ are independent of μ' and among themselves, with conditional distribution

$$\mathbb{P}\big[J'_c \in ds | U, \mathbf{X}\big] = \frac{1}{\kappa_{|c|}(U)} s^{|c|}\mathrm{e}^{-Us}\rho(ds).$$

The posterior distribution of the NRM $\tilde{\mu}$, given $\mathbf{X}$ and U, follows by normalizing the CRM $\mu|U, \mathbf{X}$.

We conclude this section by illuminating Propositions 2.1 and 2.2 via their applications to the DP and NGGP.

EXAMPLE 2.3 (DP). An application of Proposition 2.1 to the Lévy measure of the Gamma CRM shows that $\boldsymbol{\pi}$ is independent of U, and its distribution coincides with

$$\begin{aligned}\mathbb{P}[\boldsymbol{\pi} = \pi | U] &= \mathbb{P}[\boldsymbol{\pi} = \pi] \\ &= \frac{\Gamma(a)a^{|\pi|}}{\Gamma(a+n)} \prod_{c\in\pi} \Gamma(|c|).\end{aligned} \tag{2.11}$$

The corresponding predictive distributions are also independent of U and are of the form

$$X_{n+1}|U, \mathbf{X} \sim \frac{a}{a+n}\mu_0 + \sum_{c\in\boldsymbol{\pi}} \frac{|c|}{a+n}\delta_{X^*_c}. \tag{2.12}$$

An application of Proposition 2.2 shows that the posterior distribution of μ, given U and $\mathbf{X}$, corresponds to (2.10) with μ' a Gamma CRM with Lévy intensity measure

$$\nu'(ds, dy) = as^{-1}\mathrm{e}^{-s(U+1)}\, ds\, \mu_0(dy),$$

and random masses J'_c distributed according to a Gamma distribution with parameter $(|c|, U+1)$. Normalizing the posterior CRM, the resulting posterior random probability measure $\tilde{\mu}|U, \mathbf{X}$ does not depend on the scale $U+1$ and is still a DP, with updated base measure

$$\mu_n = a\mu_0 + \sum_{c\in\boldsymbol{\pi}} |c|\delta_{X^*_c}.$$

The law of the random partition $\boldsymbol{\pi}$ induced by the predictive distributions (2.12) is popularly known as the Chinese restaurant process. The metaphor is that of a sequence of customers entering a Chinese restaurant with an infinite number of round tables. The first customer sits at the first table, and each subsequent customer joins a new table with probability proportional to a, or a table with m previous customers with probability proportional to m. After n customers have entered the restaurant, the seating arrangement of customers around tables corresponds to the partition $\boldsymbol{\pi}$, with probabilities given by (2.11). Relating to $\mathbf{X}$, each table $c \in \boldsymbol{\pi}$ is served a dish X^*_c, with $X_i = X^*_c$ if customer i joined table c, that is, $i \in c$. See Blackwell and MacQueen [4] for a first characterization of the predictive distributions (2.12). See also Aldous [1] for details and Ewens [14] for an early account in population genetics.

EXAMPLE 2.4 (NGGP). An application of the formulae (2.8) to the Lévy measure of the generalized Gamma CRM leads to

$$\begin{aligned}\psi(u) &= \frac{a}{\sigma}\big((u+\tau)^\sigma - \tau^\sigma\big), \\ \kappa_m(u) &= \frac{a}{(u+\tau)^{m-\sigma}} \frac{\Gamma(m-\sigma)}{\Gamma(1-\sigma)}.\end{aligned} \tag{2.13}$$

The random partition $\boldsymbol{\pi}$ and U are not independent as in the DP, and has a joint distribution given by

$$\begin{aligned}&\mathbb{P}[\boldsymbol{\pi} = \pi, U \in du] \\ &\quad= \frac{a^{|\pi|}u^{n-1}}{\Gamma(n)(u+\tau)^{n-\sigma|\pi|}} e^{-(a/\sigma)((u+\tau)^\sigma - \tau^\sigma)}\, du \\ &\qquad\cdot \prod_{c\in\pi} \frac{\Gamma(|c|-\sigma)}{\Gamma(1-\sigma)},\end{aligned} \tag{2.14}$$

and the corresponding system of predictive distributions for X_{n+1}, given U and $\mathbf{X}$, is

$$\begin{aligned}&X_{n+1}|U, \mathbf{X} \\ &\quad\sim \frac{a(U+\tau)^\sigma}{a(U+\tau)^\sigma + n - \sigma|\boldsymbol{\pi}|}\mu_0 \\ &\qquad+ \sum_{c\in\boldsymbol{\pi}} \frac{|c|-\sigma}{a(U+\tau)^\sigma + n - \sigma|\boldsymbol{\pi}|}\delta_{X^*_c}.\end{aligned} \tag{2.15}$$

Finally, an application of Proposition 2.2 shows that the posterior distribution of μ, given U and $\mathbf{X}$, corresponds to

$$\mu|U, \mathbf{X} \sim \mu' + \sum_{c\in\boldsymbol{\pi}} J'_c\delta_{X^*_c}, \tag{2.16}$$

where μ' is a generalized Gamma CRM with parameters $(a, \sigma, U+\tau)$ and the random masses J'_c are independent among themselves and of μ', and distributed according to a Gamma distribution with parameter $(|c|-\sigma, U+\tau)$.

Note that the predictive distributions (2.15) provide a generalization of the Chinese restaurant process

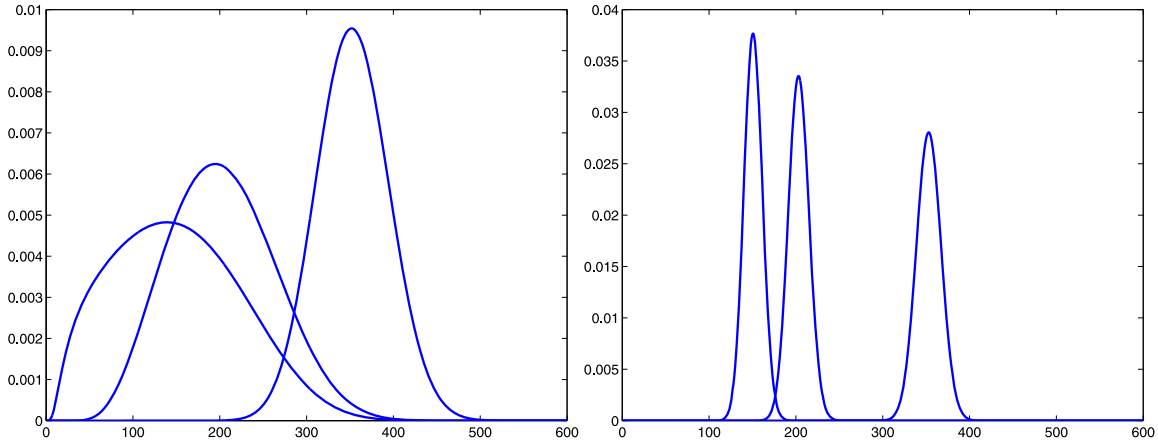

FIG. 2. *Left*: *prior distribution of the number of clusters with* $\sigma = 0.7$, $\tau = 1$, $a = 0.1, 1$ *and* 10 *and* $n = 1000$. *With increasing* a *the number of clusters increases*. *Right*: *distribution of the number of clusters with* $\sigma = 0.1$, $\tau = 1$, *and* $a = 38.5, 61.5$ *and* 161.8. *Values of* a *were chosen so that the mean number of clusters matches those in the left panel*. *With a smaller value of* σ *both the mean and the variance in the number of clusters decreases*, *which is why the values of* a *are increased from the left panel*.

metaphor for the DP. Conditionally on U, the probability of the $(n+1)$st customer joining a table with m existing customers is proportional to $m - \sigma$, with σ acting as a discount parameter. Note that the relative effect of σ is more pronounced for small values of m, which leads to larger proportions of small tables with larger σ and power-law behaviors in $\boldsymbol{\pi}$. On the other hand, the probability of joining a new table is proportional to an increasing function of all three parameters. Figure 2 shows how the distribution over the number of clusters is affected by the parameters, while Figure 3 shows how the distribution over the number of clusters grows with n for different values of the parameters.

Lijoi et al. [48] provided a detailed comparative study between the predictive structures of the NGGP and the DP in the context of mixture modeling. The advantage of specifying the NGGP mixing distribution with respect to the DP mixing distribution clearly relies on the availability of the additional parameter σ. In the DP mixture model the only free parameter which can be used to tune the distribution of the number of clusters is the mass parameter a: the bigger a, the larger the expected number of clusters. In the NGGP mixture

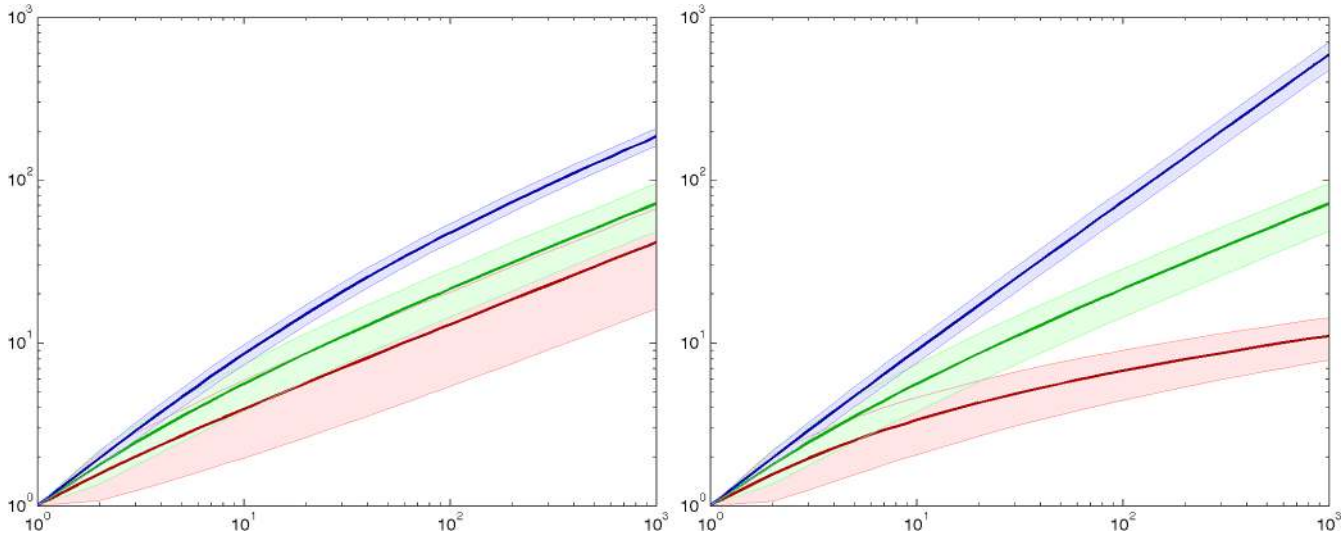

FIG. 3. *Mean and standard deviation of the number of clusters as a function of* n, *on a log–log plot*. *Left*: *with parameters* $\sigma = 0.5$, $\tau = 1$ *and* $a = 0.1, 1$ *and* 10. *Right*: *with parameters* $\sigma = 0.1, 0.5$ *and* 0.9, $\tau = 1$ *and* $a = 1$. *The growth rate with* n *follows a power-law with index* σ, *while* a *affects the number of clusters without affecting the power-law behavior*.

model the parameters a and τ play the same role as the mass parameter a in the DP mixture model. On the other hand, σ influences the grouping of the observations into distinct clusters and can be used to tune the variance of the number of clusters in the NGGP mixture model: the bigger σ, the larger the variance of the number of clusters. Further, σ also controls an interesting reinforcement mechanism that tends to reinforce significantly those clusters having higher frequencies. This turns out to be a very appealing feature in the context of mixture modeling. We refer to Lijoi et al. [48] for details on the prior elicitation for σ to control the reinforcement mechanisms induced by it.

3. MCMC POSTERIOR SAMPLING METHODS

In this section we develop some novel MCMC samplers of both marginal and conditional type for the NRM mixture models (2.5). In particular, we consider as a running example the NGGP mixing measure with parameter (a, σ, τ) and base distribution μ_0.

3.1 Conjugate Marginalized Sampler

We start with the simplest situation, when the base distribution μ_0 is conjugate to the mixture kernel F. In this case both the CRM μ and the cluster parameters $\{X_c^* : c \in \boldsymbol{\pi}\}$ can be marginalized out efficiently, leaving only the partition $\boldsymbol{\pi}$ and auxiliary variable U to be sampled. The joint distribution of $\boldsymbol{\pi}$ and U is given by (2.7), while the likelihood is

$$\mathbb{P}[\mathbf{Y}|\boldsymbol{\pi} = \pi] = \prod_{c\in\pi} f(\mathbf{Y}_c), \tag{3.1}$$

where $\mathbf{Y}_c = \{Y_i : i \in c\}$ and

$$f(\mathbf{Y}_c) = \int_{\mathbb{X}} \prod_{i\in c} f(Y_i|x)\mu_0(dx).$$

Since μ_0 is conjugate to F, the integral is assumed to be available in closed form and efficiently evaluated using the sufficient statistics of $\mathbf{Y}_c$. Moreover, since both the conditional distribution of $\boldsymbol{\pi}$ given U and the likelihood are in product partition form, the conditional distribution of $\boldsymbol{\pi}$ given $\mathbf{Y}$ and U is also in a product partition form.

We can update $\boldsymbol{\pi}$ using a form of Gibbs sampling whereby the cluster assignment of one data item Y_i is updated at a time. Let $\boldsymbol{\pi}_{\setminus i}$ be the partition with i removed. We denote the cluster assignment of Y_i with a variable z_i such that $z_i = c$ denotes the event that Y_i is assigned to cluster $c \in \boldsymbol{\pi}_{\setminus i}$, and $z_i = \varnothing$ denotes the event that it is assigned a new cluster. In order to update z_i, we can use formulae (2.7) and (3.1) to provide the conditional distribution of z_i, given $\boldsymbol{\pi}_{\setminus i}$, $\mathbf{Y}$ and U. Specifically,

$$\begin{aligned}&\mathbb{P}[z_i = c|\boldsymbol{\pi}_{\setminus i}, U, \mathbf{Y}]\\ &\quad\propto \begin{cases} \dfrac{\kappa_{|c|+1}(U)}{\kappa_{|c|}(U)}\dfrac{f(\{Y_i\}\cup\mathbf{Y}_c)}{f(\mathbf{Y}_c)}, & \text{for } c \in \boldsymbol{\pi}_{\setminus i},\\ \kappa_1(U) f(\{Y_i\}), & \text{for } c = \varnothing. \end{cases}\end{aligned} \tag{3.2}$$

Under the assumption that $\tilde{\mu}$ is a NGGP and using (2.13), the above simplifies to

$$\begin{aligned}&\mathbb{P}[z_i = c|\boldsymbol{\pi}_{\setminus i}, U, \mathbf{Y}]\\ &\quad\propto \begin{cases} (|c| - \sigma) f(Y_i|\mathbf{Y}_c), & \text{for } c \in \boldsymbol{\pi}_{\setminus i},\\ a(U+\tau)^\sigma f(Y_i|\varnothing), & \text{for } c = \varnothing, \end{cases}\end{aligned}$$

where

$$f(y|\mathbf{y}) = \frac{f(\{y\}\cup\mathbf{y})}{f(\mathbf{y})}.$$

We see that the update is a direct generalization of that for the DP which can be easily recovered by setting $\sigma = 0$. The probability of Y_i being assigned to a cluster is simply proportional to the product of a conditional prior probability of being assigned to the cluster and a conditional likelihood associated with the observation Y_i. See MacEachern [52] and Neal [60] for details on the DP case. In the next section we describe the updates for the parameters a, σ and τ, and for U, before proceeding to the marginalized and conditional samplers in the case when μ_0 is not conjugate.

3.1.1 *Updates for NGGP parameters and U.* For U, note that given $\boldsymbol{\pi}$, U is independent of $\mathbf{Y}$ with conditional distribution (2.9). In particular, in the case of the NGGP, the conditional distribution simplifies to

$$\mathbb{P}[U \in du|\boldsymbol{\pi}] \propto \frac{u^{n-1}}{(u+\tau)^{n-a|\boldsymbol{\pi}|}} e^{-(a/\sigma)((u+\tau)^\sigma - \tau^\sigma)}\, du.$$

A variety of updates can be used here. We have found that a change of variable $V = \log(U)$ leads to better behaved algorithms, since the conditional density $f_{V|\boldsymbol{\pi}}(v)$ of V given $\boldsymbol{\pi}$, that is,

$$\begin{aligned}\mathbb{P}[V \in dv|\boldsymbol{\pi}] &\propto \frac{e^{vn}}{(e^v+\tau)^{n-a|\boldsymbol{\pi}|}} e^{-(a/\sigma)((e^v+\tau)^\sigma - \tau^\sigma)}\, dv\\ &= f_{V|\boldsymbol{\pi}}(v)\, dv,\end{aligned}$$

is log concave. We use a simple Metropolis–Hastings update with a Gaussian proposal kernel with mean V and variance $1/4$, although slice sampling by Neal [62] or, alternatively, adaptive rejection sampling by Gilks and Wild [21] can also be employed.

For the NGGP, we can easily derive the updates for the parameters a, σ and τ using (2.14) and given prior specifications for the parameters. See Lijoi et al. [48] for a detailed analysis on prior specification in the context of Bayesian nonparametric mixture modeling. As regards a, we can simply use a Gamma prior distribution with parameter (α_a, β_a). Then the conditional distribution of a, given σ, τ, U and $\boldsymbol{\pi}$, is simply a Gamma distribution, that is,

$$\mathbb{P}[da|\sigma,\tau,U,\boldsymbol{\pi}] \propto a^{\alpha_a+|\boldsymbol{\pi}|-1}e^{-a(\beta_a+((U+\tau)^\sigma-\tau^\sigma)/\sigma)}\,da.$$

For τ we can again use a Gamma prior distribution with parameter $(\alpha_\tau, \beta_\tau)$. Then the conditional distribution of τ, given a, σ, U and $\boldsymbol{\pi}$, is

$$\mathbb{P}[d\tau|a,\sigma,U,\boldsymbol{\pi}] \propto \tau^{\alpha_\tau-1}e^{-\tau\beta_\tau}\frac{e^{-(a/\sigma)((U+\tau)^\sigma-\tau^\sigma)}}{\tau^{\sigma|\boldsymbol{\pi}|}(U+\tau)^{n-\sigma|\boldsymbol{\pi}|}}\,d\tau.$$

We update τ in its logarithmic domain, using the same procedure as for U described above. Finally, for σ we can use a Beta prior distribution with parameter $(\alpha_\sigma, \beta_\sigma)$. Then the conditional distribution of σ, given a, τ, U and $\boldsymbol{\pi}$, corresponds to

$$\mathbb{P}[d\sigma|a,\tau,U,\boldsymbol{\pi}] \propto \sigma^{\alpha_\sigma-1}(1-\sigma)^{\beta_\sigma-1}\frac{e^{-(a/\sigma)((U+\tau)^\sigma-\tau^\sigma)}}{\tau^{\sigma|\boldsymbol{\pi}|}(U+\tau)^{n-\sigma|\boldsymbol{\pi}|}} \cdot\prod_{c\in\pi}\frac{\Gamma(|c|-\sigma)}{\Gamma(1-\sigma)}\,d\sigma.$$

We can easily update σ using slice sampling by Neal [62].

3.2 Nonconjugate Marginalized Samplers

The main drawback of the previous algorithm is the assumption of conjugacy, which limits its applicability since nonconjugate priors are often desirable in order to increase modeling flexibility. For DP mixture models a number of marginalized algorithms for the nonconjugate setting have been proposed and investigated in the literature. The review of Neal [61] provides a detailed overview along with two novel algorithms. One of these algorithms, the so-called Algorithm 8, is simple to implement, has been demonstrated to provide excellent mixing speed, and has a tunable parameter to trade off computation cost against speed of convergence.

3.2.1 *Generalizing Neal's Algorithm* 8. In this section we provide a straightforward generalization of Neal's Algorithm 8 to the class of NRM mixture models with a nonconjugate base distribution. Here, the cluster parameters X_c^* cannot be easily marginalized out. Instead we include them into the state of the MCMC algorithm, so that the state now consists of the partition $\boldsymbol{\pi}$, $\{X_c^* : c \in \boldsymbol{\pi}\}$ and the random variable U, and we sample the cluster parameters along with $\boldsymbol{\pi}$ and U. Note that the parameters for existing clusters X_c^* can be updated with relative ease, using any MCMC update whose stationary distribution is the conditional distribution of X_c^* given everything else, that is,

$$\mathbb{P}\big[X_c^* \in dx|\boldsymbol{\pi},U,\mathbf{Y}\big] \propto \mu_0(dx)\prod_{i\in c}f(Y_i|x).$$

The difficulty with a nonconjugate marginalized sampler is the introduction of new clusters (along with their parameters) when Gibbs sampling the cluster assignments. Following Neal [61], we conceptualize our update in terms of an augmented state with additional temporarily existing variables, such that the marginal distribution of the permanent variables once the temporary ones are integrated out is the appropriate posterior distribution.

Consider updating the cluster assignment variable z_i given the existing clusters in $\boldsymbol{\pi}_{\setminus i}$. We introduce an augmented space with C empty clusters, with parameters $X_1^{\mathrm{e}}, \ldots, X_C^{\mathrm{e}}$ that are independent of $\boldsymbol{\pi}_{\setminus i}$ and independent and identically distributed according to μ_0. The state space of z_i is augmented as well to include both existing clusters $\boldsymbol{\pi}_{\setminus i}$ and the new ones $[C] = \{1, \ldots, C\}$, with conditional distribution

$$\mathbb{P}[z_i = c \in \boldsymbol{\pi}_{\setminus i}|\boldsymbol{\pi}_{\setminus i}] \propto \frac{\kappa_{|c|+1}(U)}{\kappa_{|c|}(U)}$$

and

$$\mathbb{P}\big[z_i = k \in [C]|\boldsymbol{\pi}_{\setminus i}\big] \propto \frac{\kappa_1(U)}{C},$$

respectively. Identifying z_i being in any of the additional clusters as assigning Y_i to a new cluster, we see that the total probability for Y_i being assigned to a new cluster is proportional to the first moment $\kappa_1(U)$, which is the same as in (2.7) and (3.2).

The update can be derived by first initializing the augmentation variables given the current state of the Markov chain, updating z_i, then discarding the augmentation variables. If Y_i is currently assigned to a cluster which contained another data item, then $z_i = c$

for some $c \in \boldsymbol{\pi}_{\setminus i}$, and the empty cluster parameters are simply drawn independently and identically according to μ_0. On the other hand, if Y_i is currently assigned to a cluster containing only itself, say, with parameter $X^*_\varnothing$, then in the augmented space z_i has to be one of the new clusters, say, $z_i = k$ for some $k \in [C]$ with $X^{\mathrm{e}}_k = X^*_\varnothing$. The actual value of k is unimportant, for convenience we may use $k = 1$. The other empty clusters then have parameters drawn independently and identically according to μ_0. We can now update z_i by sampling from its conditional distribution given Y_i and the parameters of all existing and empty clusters. Specifically,

$$\begin{aligned}&\mathbb{P}[z_i = c|\boldsymbol{\pi}_{\setminus i}, U, \mathbf{Y}]\\ &\propto \begin{cases} \dfrac{\kappa_{|c|+1}(U)}{\kappa_{|c|}(U)} f\big(Y_i|X^*_c\big), & \text{for } c \in \boldsymbol{\pi}_{\setminus i},\\ \dfrac{\kappa_1(U)}{C} f\big(Y_i|X^{\mathrm{e}}_c\big), & \text{for } c \in [C]. \end{cases}\end{aligned} \tag{3.3}$$

Under the assumption that $\tilde{\mu}$ is a NGGP, (3.3) again simplifies to

$$\begin{aligned}&\mathbb{P}[z_i = c|\boldsymbol{\pi}_{\setminus i}, U, \mathbf{Y}]\\ &\propto \begin{cases} (|c| - \sigma) f\big(Y_i|X^*_c\big), & \text{for } c \in \boldsymbol{\pi}_{\setminus i},\\ \dfrac{a}{C}(U + \tau)^\sigma f\big(Y_i|X^{\mathrm{e}}_c\big), & \text{for } c \in [C]. \end{cases}\end{aligned}$$

If the new value of z_i is $c \in [C]$, this means that Y_i is assigned to a new cluster with parameter X^{e}_c; the other empty clusters are discarded to complete the update. On the other hand, if the new value is $c \in \boldsymbol{\pi}_{\setminus i}$, then Y_i is assigned to an existing cluster c, and all empty clusters are discarded. Finally, the random variable U and any hyperparameters may be updated using those in Section 3.1.1.

3.2.2 *The Reuse algorithm.* In the above algorithm, each update to the cluster assignment of an observation is associated with a set of temporarily existing variables which has to be generated prior to the update and discarded afterward. As a result, many independent and identically distributed samples from the base distribution have to be generated throughout the MCMC run, and in our experiments this actually contributes a significant portion of the overall computational cost. We can mitigate this wasteful generation and discarding of clusters by noting that after updating the cluster assignment of each observation, the parameters of any unused empty clusters are in fact already independently and identically distributed according to the base distribution. Thus, we can consider reusing them for updating the next observation. However, note that as a result the parameters of the empty clusters used in different updates will not be independent, and the justification of correctness of Neal's Algorithm 8 (as Gibbs sampling in an augmentation scheme) is no longer valid.

In this section we develop an algorithm that does reuse new clusters, and show using a different technique that it is valid with stationary distribution given by the posterior. For the new algorithm, we instead augment the MCMC state space with a permanent set of C empty clusters, so the augmented state space now consists of the partition $\boldsymbol{\pi}$, the latent variable U, the parameters $\{X^*_c : c \in \boldsymbol{\pi}\}$ of existing clusters and the parameters $\{X^{\mathrm{e}}_k : k \in [C]\}$ of the auxiliary empty clusters. Further, we develop the cluster assignment updates as Metropolis–Hastings updates instead of Gibbs updates.

In the following we use the superscript $'$ in order to denote variables and values associated with the new proposed state of the Markov chain. Suppose we wish to update the cluster assignment of observation Y_i. Again we introduce the variable z_i, which takes value $c \in \boldsymbol{\pi}_{\setminus i}$ if Y_i is assigned to a cluster containing other observations, and takes values $k \in [C]$ uniformly at random if Y_i is assigned to a cluster by itself. If $z_i = c \in \boldsymbol{\pi}_{\setminus i}$, then the proposal distribution $\mathbb{Q}$ is described by a two-step algorithm:

1. Sample the variable z'_i from the conditional distribution (3.3) as before.

2a. If $z'_i = c' \in \boldsymbol{\pi}_{\setminus i}$, then we simply assign Y_i to the existing cluster c'.

2b. If $z'_i = k'$ for one of the empty clusters $k' \in [C]$ with $X^{\mathrm{e}}_{k'} = x$, then:
   (i) we assign Y_i to a newly created cluster with parameter $X^{*\prime}_\varnothing := x$;
   (ii) set $X^{\mathrm{e}\prime}_{k'} := x' \sim \mu_0$ with a new draw from the base distribution.

On the other hand, if Y_i is currently assigned to a cluster all by itself, say, with parameter $X^*_\varnothing = x_0$, then z_i will initially take on each value $k \in [C]$ uniformly with probability $1/C$. We start by setting the value for the kth empty cluster parameter $X^{\mathrm{e}}_k := x_0$ (its old value is discarded) and then removing the singleton cluster that Y_i is currently assigned to. Then the two-step algorithm above is carried out.

It is important to point out that the proposal described above is reversible. For example, the reverse of moving Y_i from an existing cluster c to a new cluster with parameter $X^*_\varnothing = x$, where x is the previous value of $X^{\mathrm{e}}_{k'}$ with its new value being a draw x' from μ_0, is exactly the reverse of the proposal moving Y_i from a singleton cluster with parameter $X^*_\varnothing = x$ to

the cluster c, while replacing the previous value x' of $X^{\mathrm{e}}_{k'}$ with x. We denote the two proposals as $(c \Rightarrow k')$ and $(k' \Rightarrow c)$. Analogously, the reverse of $(c \Rightarrow c')$ is $(c' \Rightarrow c)$ and the reverse of $(k \Rightarrow k')$ is $(k' \Rightarrow k)$.

Note also that the proposals are trans-dimensional since the number of clusters in the partition $\boldsymbol{\pi}$ (and particularly the number of cluster parameters) can change. See Green [24] and Richardson and Green [79] for approaches to trans-dimensional MCMC. Fortunately they are dimensionally balanced. In fact, we can show that the acceptance probability is simply always one. For example, for the $(c \Rightarrow k')$ proposal, the joint probability of the initial state and the proposal probability are, respectively, proportional to

$$\mathbb{P}\big[z_i = c, X^{\mathrm{e}}_{k'} \in dx \cdots\big] \propto \frac{\kappa_{|c|+1}(U)}{\kappa_{|c|}(U)} f\big(Y_i|X^*_c\big)\mu_0(dx)$$

and

$$\begin{aligned}\mathbb{Q}\big[z'_i = k', X^{*\prime}_{\varnothing} \in dx_0, X^{\mathrm{e}\prime}_{k'} \in dx' \cdots \,|\\ z_i = c, X^{\mathrm{e}}_{k'} \in dx \cdots\big]\\ \propto \frac{\kappa_1(U)}{C} f(Y_i|x)\delta_x(dx_0)\mu_0\big(dx'\big).\end{aligned}$$

We have suppressed listing all other variables for brevity. For the reverse proposal $(k' \Rightarrow c)$, the probabilities are, respectively,

$$\begin{aligned}\mathbb{P}\big[z'_i = k', dX^{*\prime}_{\varnothing} \in dx_0, X^{\mathrm{e}\prime}_{k'} \in dx' \cdots\big]\\ \propto \frac{1}{C}\kappa_1(U)\mu_0(dx_0) f(Y_i|x)\mu_0\big(dx'\big)\end{aligned}$$

and

$$\begin{aligned}\mathbb{Q}\big[z_i = c, X^{\mathrm{e}}_{k'} \in dx \cdots | z'_i = k', X^{*\prime}_{\varnothing} \in dx_0 \cdots\big]\\ \propto \frac{\kappa_{|c|+1}(U)}{\kappa_{|c|}(U)} f\big(Y_i|X^*_c\big)\delta_{x_0}(dx).\end{aligned}$$

Note that the normalization constants arising from the conditional distributions (3.3) for proposals in both directions are the same, so they can be ignored. We see that the product of the probabilities for the $(c \Rightarrow k')$ proposal is the same as that for the reverse $(k' \Rightarrow c)$, so the Metropolis–Hastings acceptance ratio is simply one. Similarly, the acceptance ratios of other proposal pairs are also equal to one.

In addition to updating the cluster assignments of all observations as above, we also need to update the parameters of the C empty clusters. We do this by marginalizing them out before updating U and the hyperparameters according to Section 3.1.1, and replacing them afterward with new independent and identically distributed draws from the base distribution. Note that the resulting Metropolis–Hastings updates are very similar to the augmentation scheme Gibbs updates described in Section 3.2.1. The only difference is the way the parameters of the empty clusters are managed and retained across cluster assignment updates of multiple observations.

3.3 Conditional Slice Sampler

In the so-called marginalized samplers the CRM μ is marginalized out while the latent variables $\mathbf{X}$ representing the partition structure and the cluster parameters are sampled. In a conditional sampler we instead alternatively Gibbs sample μ given $\mathbf{X}$ and $\mathbf{X}$ given μ. Proposition 2.2 provides the conditional distribution for μ given $\mathbf{X}$, while the conditional of $\mathbf{X}$ given μ is straightforward. What is not straightforward is the fact that since μ has an infinite number of atoms we cannot explicitly sample all of it on a computer with finite resources. Thus, it is necessary to truncate μ and work only with a finite number of atoms.

In this section we will describe a conditional sampler based on a slice sampling strategy for truncation. See Walker [85] for the slice sampler in DP mixture models, and Griffin and Walker [27] and Griffin et al. [26] for slice samplers in NRM mixture models on which our sampler is based. Recall from (2.5) that each observation Y_i is assigned to a cluster parametrized by an atom X_i of μ. We augment the state with an additional slice variable S_i, whose conditional distribution is a Uniform distribution taking values between 0 and the mass of atom X_i in μ, that is,

$$S_i|X_i, \mu \sim \mathrm{Uniform}\big(0, \mu\big(\{X_i\}\big)\big). \tag{3.4}$$

Marginalizing out S_i, the joint distribution of the other variables reduces back to the desired posterior distribution. On the other hand, conditioned on S_i, X_i can only take on values corresponding to atoms in μ with mass at least S_i. Since S_i is almost surely positive, this set of atoms is finite, and so S_i effectively serves as a truncation level for μ in the sense that only these finitely many atoms are needed when updating X_i. Over the whole data set, only the (finitely many) atoms in μ with mass at least $S = \min_{i\in[n]} S_i > 0$ are required when updating the set of latent variables $\mathbf{X}$ given μ and the slice variables $\mathbf{S} = \{S_i : i \in [n]\}$.

The state space of our sampler thus consists of the latent variables $\mathbf{X}$, the slice variables $\mathbf{S}$, the CRM μ and the auxiliary variable U introduced in Section 2.3. At a high level, our sampler is simply a Gibbs sampler, iterating among updates to $\mathbf{X}$, U, and both μ and $\mathbf{S}$ jointly.

First consider updating $\mathbf{X}$. It is easy to see that conditioned on U, μ and $\mathbf{S}$ the X_i's are mutually independent. For each $i \in [n]$, the conditional probability of X_i taking on value $x \in \mathbb{X}$ is proportional to the product of the probability $\mu(\{x\})/\mu(\mathbb{X})$ of x under the NRM $\tilde{\mu}$, the conditional distribution function $f(Y_i|x)$ of observation Y_i, and the conditional density of S_i given $X_i = x$, which is simply $1/\mu(\{x\})$ when $0 < S_i < \mu(\{x\})$ and 0 otherwise. The resulting conditional distribution of X_i simplifies to

$$\mathbb{P}[X_i = x|\mu, Y_i, S_i] \propto \begin{cases} f(Y_i|x), & \text{if } S_i < \mu(\{x\}), \\ 0, & \text{otherwise.} \end{cases}$$

This is a discrete distribution, with positive probability of $X_i = x$ only when x coincides with the location of an atom in μ with mass greater than S_i. Note that there almost surely are only a finite number of such atoms in μ since $S_i > 0$, so that updating X_i is computationally feasible.

Now consider updating U. We will perform this update conditioned only on the partition described by $\mathbf{X}$, with the random measure μ and the slice variables $\mathbf{S}$ marginalized out. We can also update any hyperparameters of the CRM and of the base distribution μ_0 at this step as well. For example, if $\tilde{\mu}$ is a NGGP, we can update both U and the parameters (a, σ, τ) using those described in Section 3.1.1, which makes these Metropolis-within-Gibbs updates.

Finally, consider updating μ and $\mathbf{S}$ jointly. Note that this update needs to be performed right after the U update since μ and $\mathbf{S}$ were marginalized out when updating U. The conditional distribution of μ given U and $\mathbf{X}$ is given by Proposition 2.2, which shows that μ will contain a finite number of fixed atoms located at the unique values $\{X_c^* : c \in \boldsymbol{\pi}\}$ among $\mathbf{X}$, and a countably infinite number of randomly located atoms corresponding to the unused clusters in the NRM mixture model. Given μ, the slice variables are independent with distributions given by (3.4); in particular, note that they depend only on the masses of the fixed atoms of μ. On the other hand, as noted above, we only need the random atoms of μ with masses above the overall truncation level $S = \min_{i \in [n]} S_i$. Therefore, a sufficient method for sampling both μ and $\mathbf{S}$ is to first sample the fixed atoms of μ, followed by $\mathbf{S}$, and finally the random atoms with masses above S.

For the fixed atoms of μ, Proposition 2.2 states that each of them corresponds to a unique value among $\mathbf{X}$ and that their masses are mutually independent and independent from the random atoms. For each such unique value X_c^*, $c \in \boldsymbol{\pi}$, the conditional distribution of its mass J_c' is

$$\mathbb{P}[J_c' \in ds|U, \mathbf{X}] \propto s^{|c|} e^{-Us} \rho(ds), \tag{3.5}$$

where $|c|$ is the number of observations allocated to the cluster c, that is, with $X_i = X_c^*$. Under the assumption of $\tilde{\mu}$ being a NGGP, the density in (3.5) simplifies to $s^{|c|-\sigma-1} e^{-(U+\tau)s}$, a Gamma density. We also update the locations of the fixed atoms as well using an acceleration step as in Bush and MacEachern [7]. The conditional distribution function of X_c^* is proportional to its prior distribution function times the likelihoods of observations assigned to the cluster, that is,

$$\mathbb{P}[X_c^* \in dx|\mathbf{Y}] \propto \mu_0(dx) \prod_{i \in c} f(Y_i|x),$$

where $i \in c$ indicates indices of those observations assigned to the cluster c. Note that any ergodic Markov kernel with the above as its stationary distribution suffices.

Once the fixed atoms are updated, the slice variables are updated by sampling each S_i independently from its conditional distribution (3.4). Finally, the random atoms of μ with mass above the overall truncation level S can be sampled using Proposition 2.2. As we work only with homogeneous CRMs here, the locations are simply independent and identically distributed draws from μ_0, while their masses are distributed according to a Poisson random measure on $[S, \infty)$ with an exponentially tilted intensity measure $\rho'(ds) = e^{-Us}\rho(ds)$.

We propose an adaptive thinning approach (see Ogata [66]) to sample from the Poisson random measure which is computationally efficient but applies only to certain classes of intensity measures which can be adaptively bounded in the following sense. Let $v'(s)$ be the density of $\rho'(ds)$ with respect to the Lebesgue measure and assume that for each $t \in \mathbb{R}_+$ there is a function $w_t(s)$ such that $w_t(t) = v'(t)$ and $w_t(s) \geq w_{t'}(s) \geq v'(s)$ for every $s, t' \geq t$. See Figure 4. In particular, for the NGGP one has

$$v'(s) = \frac{a}{\Gamma(1-\sigma)} s^{-1-\sigma} e^{-s(\tau+U)},$$

and we can use the family of adaptive bounds

$$w_t(s) = \frac{a}{\Gamma(1-\sigma)} t^{-1-\sigma} e^{-s(\tau+U)},$$

with the inverse of the integral given by

$$W_t^{-1}(r) = t - \frac{1}{\tau+U} \log\left(1 - \frac{r(\tau+U)\Gamma(1-\sigma)}{a t^{-1-\sigma} e^{-t(\tau+U)}}\right).$$

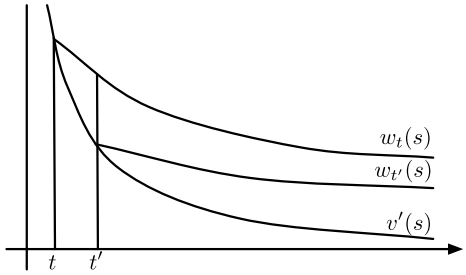

FIG. 4. *Adaptive bounds for simulating from a Poisson random measure with intensity $v'(s)$.*

Note that both $w_t(s)$ and the inverse of the map $W_t(s) = \int_t^s w_t(s')\,ds'$ are analytically tractable, with $\int_t^\infty w_t(s')\,ds' < \infty$.

The method is based on the idea of thinning by Lewis and Shedler [45], a method to simulate from a Poisson random measure by first proposing points according to a proposal Poisson random measure with higher intensity than the desired one. Each point is then accepted with probability given by the ratio of intensities under the proposal and desired Poisson random measures. The idea of adaptive thinning is that we can propose points iteratively from left to right starting at S, and after each proposed point t the bound w_t is used as the intensity of the proposed Poisson random measure from which the next point is drawn. As t increases, the bound tightens, so rejections are reduced. Further, as $\int_s^\infty w_t(s')\,ds' < \infty$, the iteration will terminate after a finite number of points are proposed. Specifically, the sampling scheme is described as follows:

1. set $N := \varnothing$, $t := S$;
2. iterate until termination:
   (i) let r be a draw from an Exponential distribution with parameter 1;
   (ii) if $r > W_t(\infty)$, terminate; else set $t' := W_t^{-1}(r)$;
   (iii) with probability $v'(t')/w_t(t')$ accept sample: set $N := N \cup \{t'\}$;
   (iv) set $t := t'$ and continue to next iteration;
3. return N as a draw from the Poisson random measure with intensity v' on $[S, \infty)$.

The returned N constitutes the set of masses for the random atoms in μ with masses above the overall truncation level S.

3.4 Some Remarks

There is a rich literature on conditional sampling schemes for nonparametric mixture models. In the DP mixture model case, the use of the stick-breaking representation for $\tilde{\mu}$, as proposed by Ishwaran and James [31], Papaspiliopoulos and Roberts [68] and Walker [85], is very simple since it involves a sequence of random variables that are independently Beta distributed a priori as well as a posteriori conditioned on other variables. However, this simplicity comes at a cost of slower mixing due to the label-switching problem discussed in Jasra et al. [38]. Papaspiliopoulos and Roberts [68] noted that while the likelihood is invariant to the ordering of atoms, the stick-breaking prior has a weak preference for atoms to be sorted by decreasing mass, resulting in multiple modes in the posterior. Then, they proposed Metropolis–Hastings moves that interchange pairs of atoms to improve mixing. A more sophisticated approach that avoids the weak identifiability altogether is to use the natural unordered representation stated in Proposition 2.2. This approach was taken in Griffin and Walker [27], and we used it here as well.

There are a few alternative methods for sampling from the Poisson random measure governing the masses of the random atoms. Griffin and Walker [27] proposed first sampling the number of atoms from a Poisson with rate $\rho'([S, \infty))$, then sampling the masses independently and identically distributed according to a distribution obtained by normalizing ρ'. Another possibility proposed by Barrios et al. [2] and Nieto-Barajas and Prünster [63] is to use the representation proposed by Ferguson and Klass [17], which involves using the mapping theorem for Poisson random measures to sample the masses in order starting from the largest to the smallest.

Our slice sampler follows Griffin and Walker [27] in introducing a slice variable S_i for each observation i. Another approach described in Griffin and Walker is to introduce a single slice variable S_{all} for all observations, with conditional distribution

$$S_{\text{all}} \sim \text{Uniform}\Big(0, \min_{i \in [n]} \mu(\{X_i\})\Big).$$

Griffin and Walker [27] found that either method may work better than the other in different situations. We preferred the method described here, as it is simpler and the updates for the latent variables, which form the most time consuming part of the algorithm, can be trivially parallelized to take advantage of recent parallel computation hardware architectures.

Slice samplers have the advantages that they can technically be exact in the sense that they target the true posterior distribution. This is opposed to alternative truncations which introduce approximations by ignoring atoms with low masses, for example, Ishwaran

and James [31] and Barrios et al. [2]. However, a difficulty with slice samplers is that although the number of random atoms in μ above the truncation level S is finite with probability one, the actual number generated can occasionally be extremely large, for example, in case of NGGPs when S is small and σ is large. In our implementation our program can occasionally terminate as it runs out of memory. We fix this by introducing an approximation where we only generate atoms with masses above 10^{-8} and only keep a maximum of the 10^6 atoms with largest masses. Griffin and Walker [27] and Barrios et al. [2] have also made similar approximations. Of course this approximation effectively nullifies the advantage of slice samplers being exact, though we have found in experiments that the approximation introduced is minimal.

Comparing the computational requirements of the proposed marginalized and conditional samplers, we expect the marginalized samplers to produce chains with less autocorrelation since they marginalize more latent variables out. Further, their computational costs per iteration are controllable and more stable since each involves introducing a fixed number of empty clusters. Concluding, while the conditional sampler is easily parallelizable, the marginalized samplers are not.

4. NUMERICAL ILLUSTRATIONS

In this section we illustrate the algorithms on a number of data sets: three simple and well-studied data sets, the galaxy, acidity and the Old Faithful geyser data sets, as well as a more complex data set of neuronal spike waveforms. The galaxy data set consists of the velocities at which 82 galaxies are receding away from our own and the acidity data set consists of the log acidity measurements of 155 lakes in Wisconsin; both are one-dimensional. The geyser data set is two-dimensional, consisting of 272 durations of eruptions along with the waiting times since the last one. The spikes data set[2] consists of a total of 14,802 neuronal spike waveforms recorded using tetrodes. Each of the four electrodes contributes 28 readings sampled at 32 kHz, so that each waveform is 112-dimensional. Prototypical waveforms are shown in Figure 10. To reduce computation time, in the following we first used PCA to reduce the data set down to six dimensions, which preserved approximately 80% of the variance and sufficient information for the mixture model to recover distinct clusters.

We analyzed the data sets by means of NRM mixtures of (multivariate) Gaussian distributions. Let D be the number of dimensions of the data set. The base distribution over the Gaussian means and covariance matrices is factorized as follows:

$$\mu_0(dm, d\Sigma) = \mathcal{N}_D(dm; m_0, S_0)\mathcal{IW}_D(d\Sigma; \alpha_0, \Sigma_0),$$

where $\mathcal{N}_D$ denotes a D-dimensional Gaussian distribution with given mean and covariance matrix and $\mathcal{IW}_D$ denotes an inverse Wishart over $D \times D$ positive definite matrices with given degree of freedom and scale matrix.

A number of authors have advocated the use of weakly informative priors for mixtures of Gaussian distributions. See Nobile [65], Raftery [75] and Richardson and Green [79]. We follow the approach advocated by Richardson and Green [79], generalizing it to the multivariate setting. In particular, we assume knowledge of a likely range over which the data lies, with the range in the ith dimension being $[m_{0i} - s_i, m_{0i} + s_i]$. We set S_0 to be a diagonal matrix with ith diagonal entry being s_i^2 so that the prior over component means is rather flat over the range. We set $\alpha_0 = D + 3$, and set a hierarchical prior $\Sigma_0 \sim \mathcal{IW}_D(\beta_0, \gamma_0 S_0)$ where β_0 is chosen to be $D - 0.6$. These degrees of freedom express the prior belief that component covariances are generally similar without being informative about their absolute scales. We choose γ_0 so that $\mathbb{E}[\Sigma] = S_0/50$, that is, that the a priori range of each component is approximately $\sqrt{50} \approx 7$ times smaller than the range set by S_0, although the model is not sensitive to this prior range since Σ_0 is random and allowed to adapt to the data in its posterior. In the one-dimensional setting this prior reduces to the same one used by Richardson and Green. A detailed study of prior specifications for mixtures of multivariate Gaussian distributions is beyond the scope of this paper and the interested reader is referred to Müller et al. [59] and Fraley and Raftery [18] for alternative specifications.

In the one-dimensional setting we also considered a conjugate prior so that we can compare the samplers with and without component parameters marginalized out. We use a similar weakly informative prior in the conjugate case as well, with base distribution given by

$$\mu_0(dm, d\Sigma) = \mathcal{N}_1(dm; m_0, S_0\Sigma_0^{-1}\Sigma)\mathcal{IW}_1(d\Sigma; \alpha_0, \Sigma_0),$$

[2]We thank Görür and Rasmussen [23] for providing us with the data set.

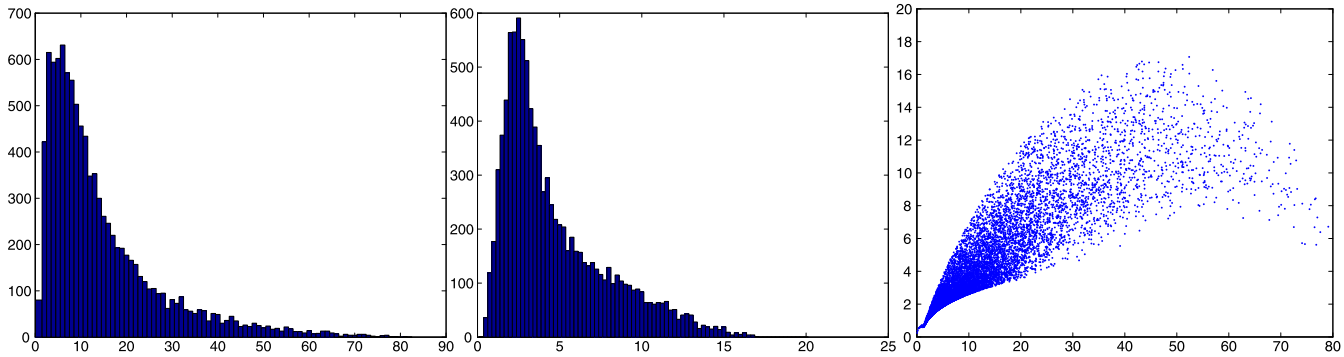

FIG. 5. *Visualizing the induced prior on the number of clusters with* $n = 82$ *corresponding to the size of the galaxy data set. Left*: *histogram of the mean number of clusters. Center*: *histogram of the standard deviation of the number of clusters. Right*: *scatter plot of standard deviation vs mean of the number of clusters.* 10,000 *draws from the prior for* a *and* σ *were used.*

where the one-dimensional inverse Wishart with parameter (a, s) is simply an inverse gamma with parameter $(a/2, s/2)$. We used the same α_0 and hierarchical prior for Σ_0 as for the nonconjugate prior, while the a priori expected value for the variance of m can be seen to be $\mathbb{E}[S_0\Sigma_0^{-1}\Sigma] = S_0$, which is independent of Σ_0 and matches the nonconjugate case. In both cases we updated the Σ_0 by Gibbs sampling.

The parameters a and τ of the NGGP are redundant (see Section 2 for details), so we simply set $\tau = 1$ in the simulations. We place a gamma $(1, 1)$ prior on a, while σ is given a beta prior with parameters $(1, 2)$. We can visualize the induced prior on the partition structure by drawing samples of a and σ from their prior and for each sample calculating the mean and standard deviation of the prior over the number of clusters. Figure 5 shows the result for $n = 82$, corresponding to the size of the galaxy data set. We see that the prior gives support over a wide range of values for the mean and standard deviation of the number of clusters, with higher probability for the mean number of clusters to be in the region between 1 and 20.

4.1 One-Dimensional Data Sets: Galaxy and Acidity

In the conjugate case, we applied both the conjugate marginalized sampler of Section 3.1 and the conditional slice sampler of Section 3.3 (but with mixture component parameters marginalized out). To investigate the difference between marginalizing out the component parameters and not, we also applied the generalization of Neal's Algorithm 8 in Section 3.2 and the Reuse algorithm of Section 3.2.2, both with $C \in \{1, 2, 3, 4, 5\}$ and the conditional slice sampler to the conjugate model (sampling the parameters instead of marginalizing them out). In the nonconjugate case we applied the conditional slice sampler and the two nonconjugate marginalized samplers with $C \in \{1, 2, 3, 4, 5\}$. For all samplers in both conjugate and nonconjugate models, the initial 10,000 iterations were discarded as burn-in, followed by 200,000 iterations, from which we collected 10,000 samples.

Figure 6 shows some aspects of the posterior distribution on the galaxy data set for the nonconjugate model obtained using the conditional slice sampler, while Figure 7 shows the same for the acidity data set. The marginalized samplers produce the same results, while the posterior for the conjugate model is similar and not shown. The co-clustering probabilities are computed as follows: the color at location (x, y) indicates the posterior probability that observations Y_i and Y_j belong to the same components, where Y_i is the largest observed value smaller than $\min(x, y)$ and Y_j is the smallest observed value larger than $\max(x, y)$. The posterior distribution of the number of components used is shown in the top half of Figure 8. The posterior distributions are consistent with those obtained by previous authors, for example, Richardson and Green [79], Escobar and West [13], Griffin and Walker [27] and Roeder [80].

In Table 1 we compared the samplers in terms of both their run times (in seconds, excluding time required to compute predictive probabilities) and their effective sample sizes (ESSs) of the number of components (as computed using the R Coda package). By marginalizing out the mixture component parameters, we see that the samplers mix more effectively with higher ESSs. The conditional slice sampler and the nonconjugate marginalized samplers were effective at handling mixture component parameters that were sampled instead of marginalized out, but the ESSs were a little lower, as expected. Among the marginalized samplers, with increasing C both the computational

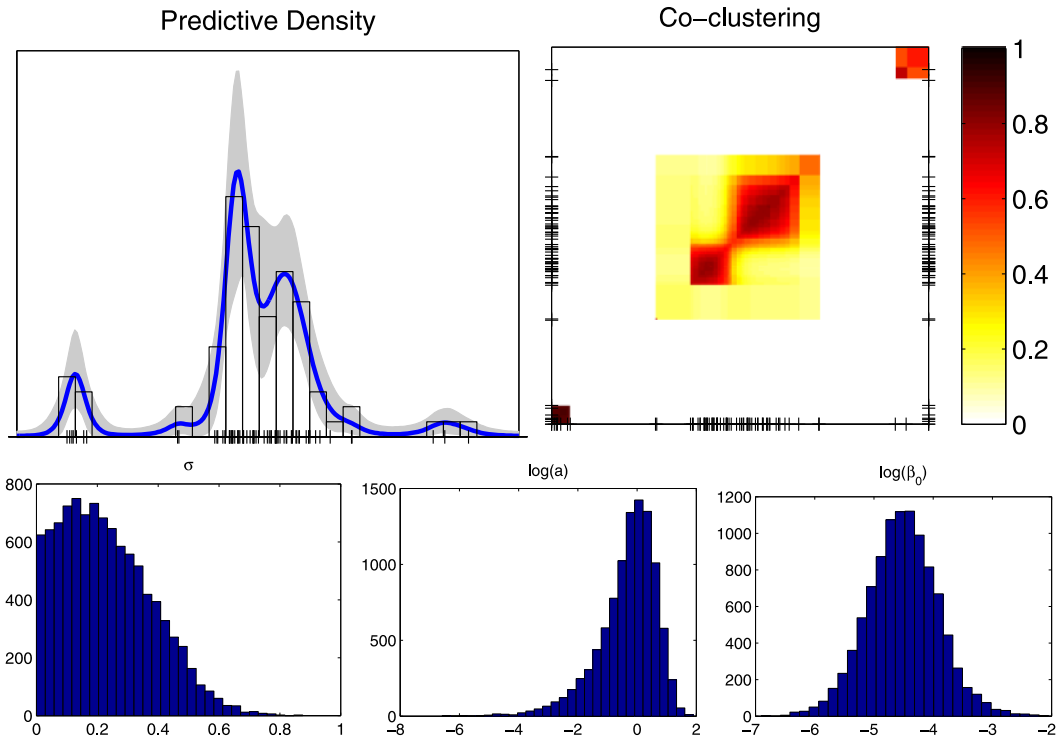

FIG. 6. *Visualizations of the posterior distribution of the nonconjugate NGGP mixture model on the galaxy data set. Top-left: posterior mean and 95% credible interval (pointwise) of the density function. Top-right: co-clustering probabilities, whiskers at edges denote observations. Bottom: histograms of the posteriors of* σ, $\log(a)$ *and* $\log(\beta_0)$, *respectively.*

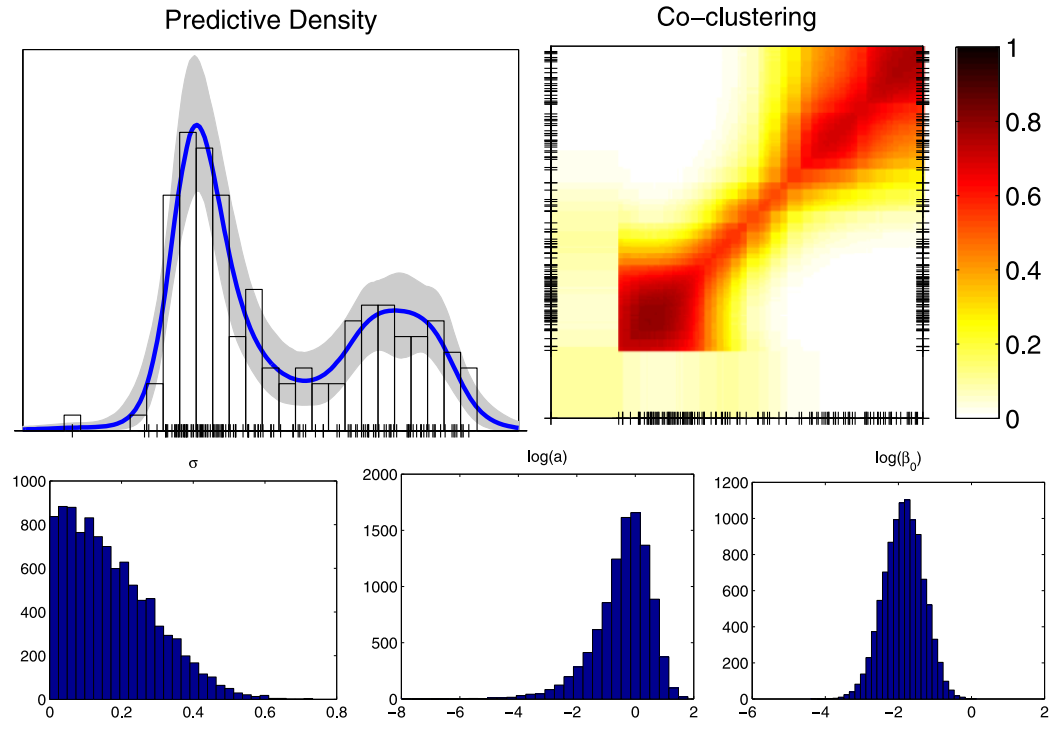

FIG. 7. *Visualizations of the posterior distribution of the nonconjugate NGGP mixture model on the acidity data set. Top-left: posterior mean and 95% credible interval (pointwise) of the density function. Top-right: co-clustering probabilities, whiskers at edges denote observations. Bottom row: histograms of the posterior of* σ, $\log(a)$ *and* $\log(\beta_0)$, *respectively.*

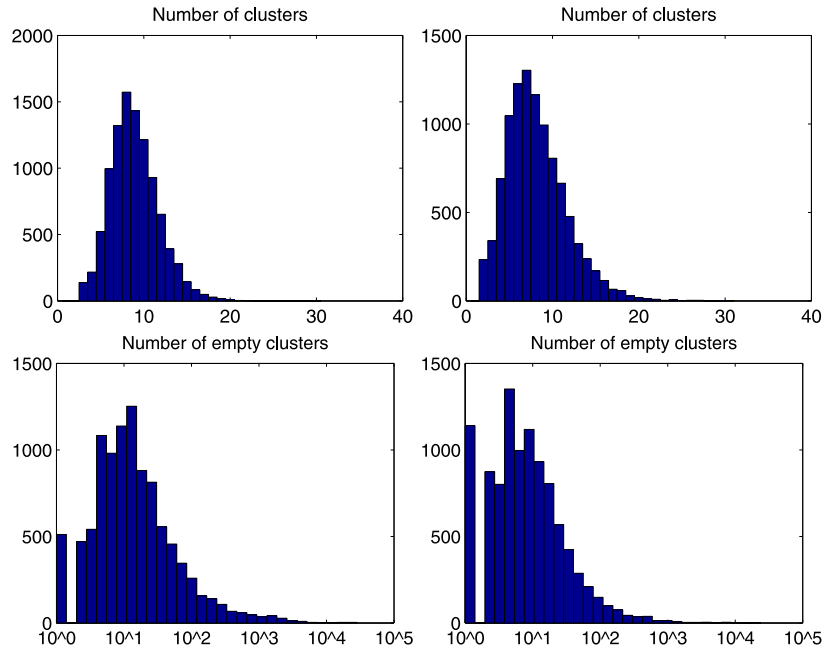

FIG. 8. *Top*: *Distribution of the number of components used, for the nonconjugate NGGP mixture model for the galaxy* (*left*) *and the acidity* (*right*) *data sets, respectively. Bottom*: *distribution of the number of empty clusters instantiated by the conditional slice sampler at each iteration* (*on the logarithmic scale*) *for the galaxy* (*left*) *and the acidity* (*right*) *data sets.*

costs and the ESS generally increase, with the computational cost of Neal's Algorithm 8 increasing more rapidly, as expected.

While the conditional slice sampler is typically faster than the marginalized samplers, they also produce lower ESSs. An important difference between the slice sampler and the marginalized samplers is that the slice sampler we proposed uses one slice variable per observation, so typically a significantly smaller number of components are considered at each update of the cluster assignment variables, and thus the algorithm is faster and has lower ESSs. If a single slice variable is used instead, as proposed in Griffin and Walker [27], or if a nonslice conditional sampler like Barrios et al. [2] is used, then all instantiated components will be considered at each update. This can result in not only higher ESSs but also higher computational overheads since the number of empty components introduced can be very large. The bottom half of Figure 8 shows the distribution of the number of empty components for one of the runs for the nonconjugate case (on the logarithmic scale). The mean numbers of empty components are 76.6 and 31.4 for the galaxy and acidity data sets, respectively. Other runs and the conjugate case are similar and not shown. For comparison, the top panels of Figure 8 show the posterior distribution of the number of nonempty components, which are smaller. As a further note, we have found that the truncation of the slice variables at 10^{-8} described in Section 3.3 is essential to the program working properly, as otherwise it will sometimes generate far too many atoms, causing the program to run out of memory. Table 2 shows the number of times the truncation came into effect during each MCMC run. We did not find cases in which the 10^6 limit on the number of atoms was reached among these runs.

4.2 Multidimensional Data Sets: Geyser and Spikes

We have also explored the efficacies of the algorithms on the geyser and spikes data sets. For the spikes data set we reduced the size of the data set by randomly selecting 500 spike waveforms to reduce the overall computation time for the experiments. In preliminary experiments this does not affect the qualitative conclusions drawn from the results. We did not include the generalization of Neal's Algorithm 8 in these experiments, as we have found in initial explorations that it took significantly more computation time without producing substantially higher ESSs than the Reuse algorithm. The setups of the experiments are similar as for the one-dimensional setting, with each algorithm producing ten independent runs, each consisting of 10,000 burn-in iterations followed by 10,000

TABLE 1

Comparison of sampler efficiencies on the one-dimensional galaxy and acidity data sets. Each of 10 *runs produces* 10,000 *samples, at intervals of* 20 *iterations, after an initial burn-in period of* 10,000 *iterations. Each entry reports the average and standard error over the* 10 *runs. In the first column, C indicates conjugate prior specification, N for nonconjugate, while M indicates component parameters are marginalized and S means they are sampled*

Model	Sampler	Galaxy		Acidity	
		Runtime (s)	ESS	Runtime (s)	ESS
CM	Cond Slice	239.1 ± 4.2	2004 ± 178	196.5 ± 1.0	910 ± 142
CM	Marg ($C = 1$)	215.7 ± 1.4	7809 ± 87	395.5 ± 1.7	5236 ± 181
CS	Cond Slice	133.0 ± 3.2	1594 ± 117	77.4 ± 0.7	1099 ± 49
CS	Marg Neal 8 ($C = 1$)	74.4 ± 0.6	5815 ± 145	133.3 ± 1.8	4175 ± 85
CS	Marg Neal 8 ($C = 2$)	87.9 ± 0.6	6292 ± 94	163.8 ± 1.5	4052 ± 158
CS	Marg Neal 8 ($C = 3$)	101.9 ± 0.7	6320 ± 137	188.2 ± 1.1	4241 ± 99
CS	Marg Neal 8 ($C = 4$)	115.9 ± 0.6	6283 ± 86	216.6 ± 1.7	4266 ± 122
CS	Marg Neal 8 ($C = 5$)	130.0 ± 0.6	6491 ± 203	243.8 ± 2.0	4453 ± 123
CS	Marg Reuse ($C = 1$)	64.3 ± 0.3	4451 ± 79	114.6 ± 2.0	3751 ± 65
CS	Marg Reuse ($C = 2$)	67.6 ± 0.5	5554 ± 112	123.1 ± 1.9	4475 ± 110
CS	Marg Reuse ($C = 3$)	71.3 ± 0.5	5922 ± 157	128.2 ± 2.2	4439 ± 158
CS	Marg Reuse ($C = 4$)	74.9 ± 0.5	6001 ± 101	140.1 ± 1.6	4543 ± 108
CS	Marg Reuse ($C = 5$)	78.7 ± 0.6	6131 ± 124	147.7 ± 1.5	4585 ± 116
NS	Cond Slice	75.5 ± 1.2	939 ± 92	50.9 ± 0.5	949 ± 70
NS	Marg Neal 8 ($C = 1$)	65.0 ± 0.5	4313 ± 172	110.9 ± 0.8	4144 ± 64
NS	Marg Neal 8 ($C = 2$)	78.6 ± 0.4	4831 ± 168	139.2 ± 1.8	4290 ± 125
NS	Marg Neal 8 ($C = 3$)	92.5 ± 0.5	4785 ± 97	162.7 ± 0.9	4368 ± 72
NS	Marg Neal 8 ($C = 4$)	106.3 ± 0.5	4849 ± 120	187.6 ± 1.1	4234 ± 142
NS	Marg Neal 8 ($C = 5$)	119.7 ± 0.6	5029 ± 89	215.4 ± 1.3	4144 ± 213
NS	Marg Reuse ($C = 1$)	55.2 ± 0.5	3830 ± 103	91.3 ± 0.9	4007 ± 122
NS	Marg Reuse ($C = 2$)	58.7 ± 0.5	4286 ± 101	98.1 ± 0.9	4192 ± 138
NS	Marg Reuse ($C = 3$)	62.4 ± 0.6	4478 ± 124	105.1 ± 0.9	4260 ± 136
NS	Marg Reuse ($C = 4$)	66.1 ± 0.5	4825 ± 63	112.3 ± 1.0	4191 ± 139
NS	Marg Reuse ($C = 5$)	69.8 ± 0.6	4755 ± 141	121.0 ± 1.8	4186 ± 121

samples collected at intervals of 20 iterations. In addition to $C = 1, \ldots, 5$, we also explored higher values of $C = 10$, 15 and 20.

The run times and ESSs are reported in Table 3. The trends observed for the one-dimensional setting hold here as well: that the slice sampler is faster but produces lower ESSs, and that with increasing C the marginalized sampler produces higher ESSs at higher computational costs. As expected, the algorithms mix more slowly on the higher-dimensional spikes data set, with significantly lower ESSs. For the spikes data set the nonconjugate marginalized samplers with higher values of C have significantly higher ESSs. In fact, they had better ESSs per unit of run time than for lower values of C or for the slice sampler. This contrasts with the other simpler data sets, where lower values of C worked very well, probably because the additional complexity of higher C values was not needed. Figure 9 shows the posterior distributions over the number of clusters, $\log(a)$ and σ.

Finally, we illustrate the clustering structure among spike waveforms discovered by the NGGP mixture model. 2000 spike waveforms were selected at random from the data set and the Reuse algorithm with $C = 20$ is run as before, with 10,000 burn-in iterations followed by 10,000 samples collected every 20 itera-

TABLE 2

Average number of times the slice threshold S was less than the 10^{-8} *truncation level over the* 10 *conditional slice sampling runs. The total number of iterations of each run is* 210,000. *Each entry reports the average and standard error over* 10 *runs*

Model	Galaxy	Acidity	Geyser	Spikes
CM	4476 ± 440	6143 ± 1148	–	–
CS	4597 ± 385	4385 ± 394	–	–
NS	3712 ± 222	8017 ± 1180	$15{,}180 \pm 980$	5621 ± 475

TABLE 3

Comparison of sampler efficiencies on the geyser (2D) and spikes (6D) data sets. Each of 10 runs produces 10,000 samples, at intervals of 20 iterations, after an initial burn-in period of 10,000 iterations. Each entry reports the average and standard error over the 10 runs

		Geyser		Spikes	
Model	**Sampler**	**Runtime (s)**	**ESS**	**Runtime (s)**	**ESS**
NS	Cond Slice	142.6 ± 1.1	574 ± 36	732.6 ± 8.1	17.1 ± 2.3
NS	Marg Reuse ($C = 1$)	208.0 ± 1.3	2770 ± 209	1120.3 ± 8.8	35.7 ± 2.4
NS	Marg Reuse ($C = 2$)	225.3 ± 1.4	3236 ± 73	1164.5 ± 5.4	46.9 ± 2.9
NS	Marg Reuse ($C = 3$)	241.5 ± 1.3	3148 ± 71	1204.1 ± 7.3	57.0 ± 3.9
NS	Marg Reuse ($C = 4$)	257.7 ± 1.7	3291 ± 145	1238.5 ± 7.8	61.4 ± 3.3
NS	Marg Reuse ($C = 5$)	274.8 ± 1.7	3144 ± 70	1291.8 ± 7.9	69.8 ± 4.9
NS	Marg Reuse ($C = 10$)	356.3 ± 2.5	3080 ± 135	1513.8 ± 11.9	90.8 ± 5.6
NS	Marg Reuse ($C = 15$)	446.6 ± 4.9	3312 ± 154	1746.3 ± 10.7	95.9 ± 4.2
NS	Marg Reuse ($C = 20$)	550.4 ± 3.5	3336 ± 109	1944.0 ± 14.7	114.5 ± 8.4

tions. We use co-clustering probabilities to summarize the clustering structure. For each pair (i, j) of spikes let p_{ij} be the (estimated) posterior probability that the two spikes were assigned to the same cluster. We use average linkage to organize the spikes into a hierarchy, where the distance between spikes i and j is defined to be $1 - p_{ij}$. This is then used to reorder the co-clustering matrix. The hierarchy and reordered matrix are shown on the upper panels of Figure 10. We see that most spikes belong to six large clusters, two of which have significant overlap and merged into one, while a subset of waveforms formed smaller clusters which may or may not overlap with other clusters. In the bottom panels of Figure 10 we visualize the various clusters found by thresholding the hierarchy at 0.95 and ignoring clusters of size less than 10.

We can interpret the clusters found here in the context of spike sorting, an important process in experimental neuroscience of detecting spikes from neural recordings and determining the neuron corresponding to each spike from the shape of its waveform (as well as the number of neurons) using a variety of manual or automated clustering techniques, with each cluster interpreted as a unique neuron. See Quiroga [73] and Lewicki [44] for reviews of spike sorting methods, and also Görür and Rasmussen [23], Wood and Black [86] and Gasthaus et al. [20] for Bayesian nonparametric mixture modeling approaches to spike sorting. We find that the 5 largest clusters found (1, 4, 6, 7 and 8) all correspond to well-defined waveforms with distinctive shapes, and expect each of 1, 4, 6 and 8 to correspond to a single neuron. Spikes in cluster 5 have similar waveforms, as 6 and the two clusters are in fact merged at a threshold of 0.99, though spikes in 5 lacked refractory periods; they may either correspond to the same or distinct neurons. Clusters 2 and 3 consist of outliers, false detections or waveforms formed by the superposition of two consecutive spikes. We note that a number of waveforms in other clusters are also superpositions as well. Finally, analyzing the two subclusters of 7, we see that although their shapes are very similar, the waveforms in the first two subpanels of 7a seem to be slightly smaller than those in 7b, though it is unclear if the subclustering is due to two neurons or is an artefact of the mixture components not being flexible enough to capture spike waveform variability.

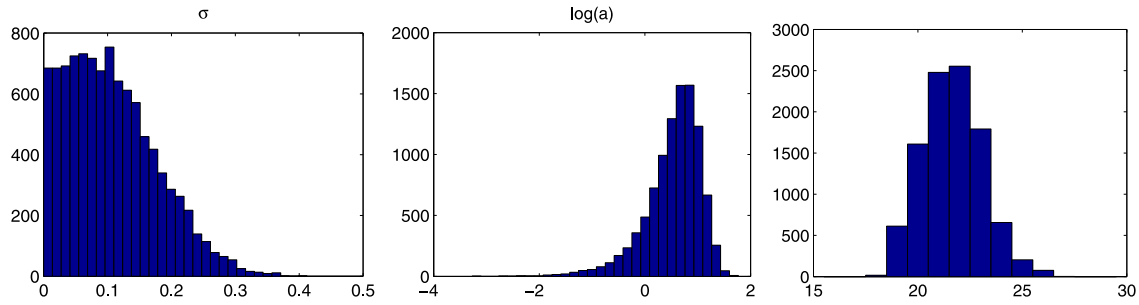

FIG. 9. *Histograms of the posterior distribution of σ, $\log(a)$ and the number of clusters, respectively, for the spikes data.*

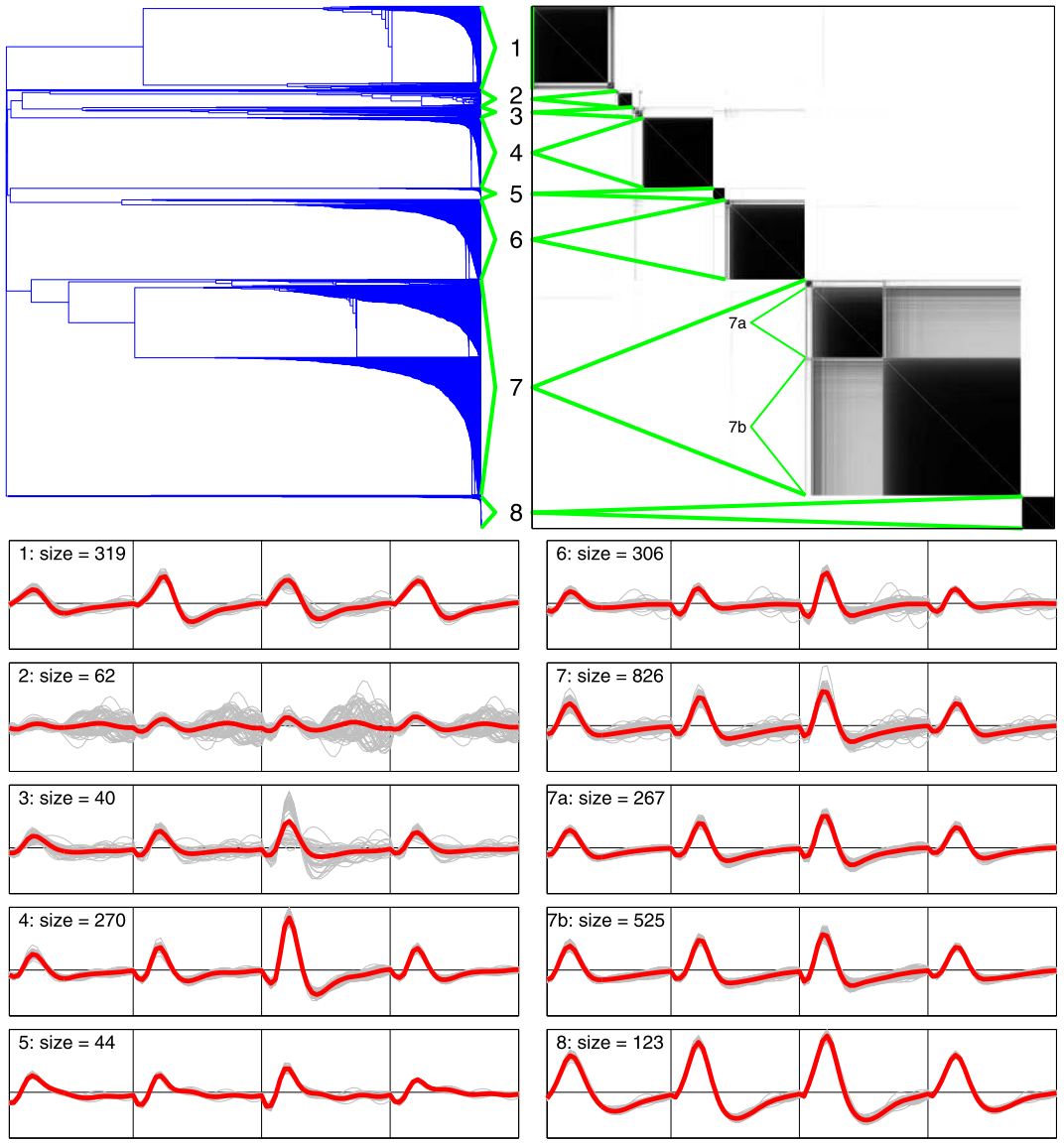

FIG. 10. *Top*: *hierarchical organization of spike waveforms obtained by average linkage and the corresponding reordered co-clustering matrix*. *Bottom*: *clusters found by thresholding at* 0.95. *Each panel consists of four subpanels, each corresponding to the waveforms recorded by an electrode. Each waveform in the cluster is plotted in light grey and their mean in dark grey.*

The approach taken here is simply to use an agglomerative linkage algorithm to help us visualize and explore the posterior over partition structures under the mixture model. An alternative approach is to summarize the posterior using a single partition, for example, using the maximum a posteriori partition or one that minimizes the posterior expectation of a loss function like Binder's loss. The issue of how best to analyze and interpret the posterior partition structure of Bayesian models for clustering is still an open question and beyond the scope of this paper. We refer the interested reader to Binder [3], Medvedovic and Sivaganesan [56], Dahl [8], Lau and Green [43], Fritsch and Ickstadt [19] and Rasmussen et al. [76] for classical and recent efforts in this regard.

5. DISCUSSION

NRMs provide a large class of flexible nonparametric priors beyond the standard DP, but their more common use is currently hindered by a lack of understanding and of good algorithms for posterior simulation. This work provides a review of NRMs for easy access to the extensive literature, as well as novel algorithms for posterior simulation that are efficient and easy to use. We will also provide open source Java

software implementing all four algorithms described in Section 3 so that others might more easily explore them.

All the samplers proposed in this paper are basic samplers that make changes to the cluster assignment of one observation at a time. Samplers that make more complex changes, for example, those based on split-merge Metropolis–Hastings moves by Jain and Neal [32], can be significantly more efficient at exploring multiple posterior modes. Such samplers can be derived in both marginalized and conditional forms, using the characterizations reviewed in this paper, and are an interesting avenue of future research. Beyond the algorithms described in Section 3, there are many variants possible with both marginalized and conditional samplers for NRM mixture models. While conditional samplers have been well explored in the literature, ours are the first tractable marginalized samplers for mixture models with a homogeneous NRM prior. In addition, a number of samplers based on the system of predictive distributions of NRMs have been proposed by James et al. [36] and by Lijoi et al. [46–48], but these sampling methods can be computationally expensive in the nonconjugate setting due to numerical integrations needed for computing the probabilities associated to new clusters, and convergence is slow, requiring additional acceleration steps. See, for example, Bush and MacEachern [7] for details.

A random probability measure that is in popular use but conspicuously not within the class of NRMs is the two-parameter Poisson–Dirichlet process (otherwise known as the Pitman–Yor process) by Perman et al. [69]. See also Pitman and Yor [72] and Ishwaran and James [31] for details. It is instead in an even larger class known as the Poisson–Kingman processes introduced by Pitman [70], which are obtained by allowing the total mass of the otherwise completely random measure underlying the NRM to have a different distribution. Poisson–Kingman processes represent the largest known class of random probability measures that are still mathematically tractable. In addition to NRMs, they also include random probability measures induced by the so-called Gibbs type exchangeable random partitions introduced by Gnedin and Pitman [22]. The marginalized and conditional samplers we have developed may be extended to the Poisson–Kingman processes as well.

Throughout this paper we have used the NGGP as a running example to illustrate the various properties and formulae, because of its tractability and because it includes many well-known NRMs as examples. It has been shown by Lijoi et al. [50] that the NGGP is the only NRM that is also of Gibbs type. Beyond the NGGP, the formulae derived tend to become intractable and require numerical integrations. A notable exception is the class of NRMs whose Lévy intensity measure are mixtures of those for the generalized Gamma CRM, first proposed by Trippa and Favaro [84] who also showed that they form a dense subclass of the NRMs. It is straightforward to extend the algorithms and the software derived in this paper to this larger class.

As a final remark, the study of random probability measures underpins a large body of work spanning probability, statistics, combinatorics and mathematical genetics. They also form the core of many Bayesian nonparametric models that are increasingly popular in applied statistics and machine learning. By expanding the class of tractable random probability measures beyond the DP to NRMs, we hope that our work will increase both the range and flexibility of the models in use now and in the future.

ACKNOWLEDGMENTS

The authors are grateful to the Editor, an Associate Editor and two anonymous referees for their constructive comments and suggestions. This work was supported by the European Research Council (ERC) through StG "N-BNP" 306406.

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

[2] BARRIOS, E., LIJOI, A., NIETO-BARAJAS, L. E. and PRÜENSTER, I. (2012). Modeling with normalized random measure mixture models. Unpublished manuscript.

[3] BINDER, D. A. (1978). Bayesian cluster analysis. *Biometrika* **65** 31–38. MR0501592

[4] BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist*. **1** 353–355. MR0362614

[5] BRIX, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. in Appl. Probab.* **31** 929–953. MR1747450

[6] BRODERICK, T., JORDAN, M. I. and PITMAN, J. (2012). Clusters and features from combinatorial stochastic processes. Available at arXiv:1206.5862 [math.ST].

[7] BUSH, C. A. and MACEACHERN, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83** 275–285.

[8] DAHL, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* (K. Do, P. Müller and M. Vannucci, eds.). Cambridge Univ. Press, Cambridge.

[9] DALEY, D. J. and VERE-JONES, D. (2002). *An Introduction to the Theory of Point Processes*. Springer, New York. MR0950166

[10] DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **56** 363–375. MR1281940

[11] ESCOBAR, M. D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Ph.D. thesis, Yale Univ. MR2637324

[12] ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268–277. MR1266299

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

[14] EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biology* **3** 87–112; erratum, ibid. **3** (1972), 240, 376. MR0325177

[15] FAVARO, S. and WALKER, S. G. (2013). Slice sampling σ-stable Poisson–Kingman mixture models. *J. Comput. Graph. Statist.* To appear.

[16] FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[17] FERGUSON, T. S. and KLASS, M. J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Statist.* **43** 1634–1643. MR0373022

[18] FRALEY, C. and RAFTERY, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *J. Classification* **24** 155–181. MR2415725

[19] FRITSCH, A. and ICKSTADT, K. (2009). Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal.* **4** 367–391. MR2507368

[20] GASTHAUS, J., WOOD, F., GÖRÜR, D. and TEH, Y. W. (2009). Dependent Dirichlet process spike sorting. In *Advances in Neural Information Processing Systems 21* 497–504.

[21] GILKS, W. R. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.* **41** 337–348.

[22] GNEDIN, A. and PITMAN, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.* **138** 5674–5684.

[23] GÖRÜR, D., RASMUSSEN, C. E., TOLIAS, A. S., SINZ, F. and LOGOTHETIS, N. K. (2004). Modelling spikes with mixtures of factor analysers. In *Proceedings of the Conference of the German Association for Pattern Recognition (DAGM)*.

[24] GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

[25] GREEN, P. J. and RICHARDSON, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scand. J. Stat.* **28** 355–375. MR1842255

[26] GRIFFIN, J. E., KOLOSSIATIS, M. and STEEL, M. F. J. (2013). Comparing distributions using dependent normalized random measure mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75** 499–529.

[27] GRIFFIN, J. E. and WALKER, S. G. (2011). Posterior simulation of normalized random measure mixtures. *J. Comput. Graph. Statist.* **20** 241–259. MR2816547

[28] GRIFFITHS, T. L. and GHAHRAMANI, Z. (2011). The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.* **12** 1185–1224. MR2804598

[29] HJORT, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18** 1259–1294. MR1062708

[30] HJORT, N. L., HOLMES, C., MÜLLER, P. and WALKER, S. G., eds. (2010). *Bayesian Nonparametrics. Cambridge Series in Statistical and Probabilistic Mathematics* **28**. Cambridge Univ. Press, Cambridge. MR2722987

[31] ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729

[32] JAIN, S. and NEAL, R. M. (2000). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Unpublished manuscript.

[33] JAMES, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. Available at arXiv:math/0205093v1.

[34] JAMES, L. F. (2003). A simple proof of the almost sure discreteness of a class of random measures. *Statist. Probab. Lett.* **65** 363–368. MR2039881

[35] JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Stat.* **33** 105–120. MR2255112

[36] 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

[37] JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2010). On the posterior distribution of classes of random means. *Bernoulli* **16** 155–180. MR2648753

[38] JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20** 50–67. MR2182987

[39] KALLI, M., GRIFFIN, J. E. and WALKER, S. G. (2011). Slice sampling mixture models. *Stat. Comput.* **21** 93–105. MR2746606

[40] KINGMAN, J. F. C. (1967). Completely random measures. *Pacific J. Math.* **21** 59–78. MR0210185

[41] KINGMAN, J. F. C. (1993). *Poisson Processes. Oxford Studies in Probability* **3**. Clarendon Press, Oxford. MR1207584

[42] KINGMAN, J. F. C., TAYLOR, S. J., HAWKES, A. G., WALKER, A. M., COX, D. R., SMITH, A. F. M., HILL, B. M., BURVILLE, P. J. and LEONARD, T. (1975). Random discrete distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **37** 1–22. MR0368264

[43] LAU, J. W. and GREEN, P. J. (2007). Bayesian model-based clustering procedures. *J. Comput. Graph. Statist.* **16** 526–558. MR2351079

[44] LEWICKI, M. S. (1998). A review of methods for spike sorting: The detection and classification of neural action potentials. *Network* **9** 53–78.

[45] LEWIS, P. A. W. and SHEDLER, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Res. Logist. Quart.* **26** 403–413. MR0546120

[46] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2005). Bayesian nonparametric analysis for a generalized Dirichlet process prior. *Stat. Inference Stoch. Process.* **8** 283–309. MR2177315

[47] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.* **100** 1278–1291. MR2236441
[48] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 715–740. MR2370077
[49] LIJOI, A. and PRÜNSTER, I. (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics* (N. L. Hjort, C. C. Holmes, P. Müller and S. G. Walker, eds.) 80–136. Cambridge Univ. Press, Cambridge. MR2730661
[50] LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Investigating nonparametric priors with Gibbs structure. *Statist. Sinica* **18** 1653–1668. MR2469329
[51] LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519
[52] MACEACHERN, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23** 727–741. MR1293996
[53] MACEACHERN, S. N. (1998). Computational methods for mixture of Dirichlet process models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.). *Lecture Notes in Statist.* **133** 23–43. Springer, New York. MR1630074
[54] MACEACHERN, S. N. and MÜLLER, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7** 223–238.
[55] MCLACHLAN, G. J. and BASFORD, K. E. (1988). *Mixture Models*: *Inference and Applications to Clustering*. *Statistics*: *Textbooks and Monographs* **84**. Dekker, New York. MR0926484
[56] MEDVEDOVIC, M. and SIVAGANESAN, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18** 1194–1206.
[57] MENGERSEN, K. L. and ROBERT, C. P. (1996). Testing for mixtures: A Bayesian entropic approach. In *Bayesian Statistics*, 5 (*Alicante*, 1994) (J. O. Berger, J. M. Bernardo, A. P. Dawid, D. V. Lindley and A. F. M. Smith, eds.) 255–276. Oxford Univ. Press, New York. MR1425410
[58] MULIERE, P. and TARDELLA, L. (1998). Approximating distributions of random functionals of Ferguson–Dirichlet priors. *Canad. J. Statist.* **26** 283–297. MR1648431
[59] MÜLLER, P., ERKANLI, A. and WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79. MR1399156
[60] NEAL, R. M. (1992). Bayesian mixture modeling. In *Proceedings of the* 11*th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, *Seattle*. Kluwer, Dordrecht.
[61] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804
[62] NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. MR1994729
[63] NIETO-BARAJAS, L. E. and PRÜNSTER, I. (2009). A sensitivity analysis for Bayesian nonparametric density estimators. *Statist. Sinica* **19** 685–705. MR2514182
[64] NIETO-BARAJAS, L. E., PRÜNSTER, I. and WALKER, S. G. (2004). Normalized random measures driven by increasing additive processes. *Ann. Statist.* **32** 2343–2360. MR2153987
[65] NOBILE, A. (1994). Bayesian analysis of finite mixture distributions. Ph.D. thesis, Carnegie Mellon Univ. MR2692049
[66] OGATA, Y. (1981). On Lewis' simulation method for Point processes. *IEEE Trans. Inform. Theory* **27** 23–31.
[67] PAPASPILIOPOULOS, O. (2008). A note on posterior sampling from Dirichlet mixture models. Working Paper 20, Centre for Research in Statistical Methodology, Univ. Warwick.
[68] PAPASPILIOPOULOS, O. and ROBERTS, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95** 169–186. MR2409721
[69] PERMAN, M., PITMAN, J. and YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92** 21–39. MR1156448
[70] PITMAN, J. (2003). Poisson–Kingman partitions. In *Statistics and Science*: *A Festschrift for Terry Speed* (D.R. Goldstein, ed.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **40** 1–34. IMS, Beachwood, OH. MR2004330
[71] PITMAN, J. (2006). *Combinatorial Stochastic Processes*. *Lecture Notes in Math.* **1875**. Springer, Berlin. MR2245368
[72] PITMAN, J. and YOR, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855–900. MR1434129
[73] QUIROGA, R. Q. (2007). Spike sorting. *Scholarpedia* **2** 3583.
[74] RAFTERY, A. E. (1996). Hypothesis testing and model selection. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.). Chapman & Hall, London.
[75] RAFTERY, A. E. (1996). Hypothesis testing and model selection via posterior simulation. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds.). Chapman & Hall, London.
[76] RASMUSSEN, C. E., DE LA CRUZ, B. J., GHAHRAMANI, Z. and WILD, D. L. (2009). Modeling and visualizing uncertainty in gene expression clusters using Dirichlet process mixtures. *IEEE/ACM Trans. Comput. Biol. and Bioinform.* **6** 615–628.
[77] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. MR2514435
[78] REGAZZINI, E., LIJOI, A. and PRÜNSTER, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31** 560–585. MR1983542
[79] RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59** 731–792. MR1483213
[80] ROEDER, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89** 487–495. MR1294074
[81] ROEDER, K. and WASSERMAN, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92** 894–902. MR1482121
[82] 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

[83] TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester. MR0838090

[84] TRIPPA, L. and FAVARO, S. (2012). A class of normalized random measures with an exact predictive sampling scheme. *Scand. J. Stat.* **39** 444–460. MR2971631

[85] WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.* **36** 45–54. MR2370888

[86] WOOD, F. and BLACK, M. J. (2008). A nonparametric Bayesian alternative to spike sorting. *Journal of Neuroscience Methods* **173** 1–12.