Statistics and Computing (2022) 32:40
https://doi.org/10.1007/s11222-022-10096-0

# Importance conditional sampling for Pitman–Yor mixtures

**Antonio Canale**[1] · **Riccardo Corradin**[2] · **Bernardo Nipoti**[3]

Received: 23 October 2021 / Accepted: 5 April 2022 / Published online: 17 May 2022
© The Author(s) 2022

**Abstract**
Nonparametric mixture models based on the Pitman–Yor process represent a flexible tool for density estimation and clustering. Natural generalization of the popular class of Dirichlet process mixture models, they allow for more robust inference on the number of components characterizing the distribution of the data. We propose a new sampling strategy for such models, named importance conditional sampling (ICS), which combines appealing properties of existing methods, including easy interpretability and a within-iteration parallelizable structure. An extensive simulation study highlights the efficiency of the proposed method which, unlike other conditional samplers, shows stable performances for different specifications of the parameters characterizing the Pitman–Yor process. We further show that the ICS approach can be naturally extended to other classes of computationally demanding models, such as nonparametric mixture models for partially exchangeable data.

**Keywords** Bayesian nonparametrics · Dependent Dirichlet process · Importance conditional sampling · Nonparametric mixtures · Pitman–Yor process · Sampling-importance resampling

**Mathematics Subject Classification** 65C60 · 62F15 · 65C40 · 62G07

## 1 Introduction

Bayesian nonparametric mixtures are flexible models for density estimation and clustering, nowadays a well-established modelling option for applied statisticians (Frühwirth-Schnatter et al. 2019). The first of such models to appear in the literature was the Dirichlet process (DP) (Ferguson 1973) mixture of Gaussian kernels by Lo (1984), a contribution that paved the way to the definition of a wide variety of nonparametric mixture models. In recent years, increasing interest has been dedicated to the definition of mixture models based on nonparametric mixing random probability measures that go beyond the DP (e.g. Nieto-Barajas et al. 2004; Lijoi et al. 2005a, b, 2007; Argiento et al. 2016). Among these measures, the Pitman–Yor process (PY) (Perman et al. 1992; Pitman 1995) stands out for conveniently combining mathematical tractability, interpretability, and modelling flexibility (see, e.g., De Blasi et al. 2015).

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an $n$-dimensional sample of observations defined on some probability space $(\Omega, \mathscr{A}, \mathrm{P})$ and taking values in $\mathbb{X}$, and $\mathscr{F}$ denote the space of all probability distributions on $\mathbb{X}$. A Bayesian nonparametric mixture model is a random distribution taking values in $\mathscr{F}$, defined as

$$\tilde{f}(x) = \int_{\Theta} \mathcal{K}(x; \theta) \mathrm{d}\tilde{p}(\theta), \tag{1}$$

where $\mathcal{K}(x; \theta)$ is a kernel and $\tilde{p}$ is a discrete random probability measure. In this paper we focus on $\tilde{p} \sim PY(\sigma, \vartheta; P_0)$, that is we assume that $\tilde{p}$ is distributed as a PY process with discount parameter $\sigma \in [0, 1)$, strength parameter $\vartheta > -\sigma$, and diffuse base measure $P_0 \in \mathscr{F}$. The DP is recovered as a special case when $\sigma = 0$. Model (1) can alternatively be written in hierarchical form as

✉ Bernardo Nipoti
bernardo.nipoti@unimib.it

Antonio Canale
canale@stat.unipd.it

Riccardo Corradin
riccardo.corradin@nottingham.ac.uk

[1] Department of Statistical Sciences, University of Padova, Padua, Italy

[2] School of Mathematical Sciences, University of Nottingham, Nottingham, UK

[3] Department of Economics, Management and Statistics, University of Milano Bicocca, Milan, Italy

$$
\begin{aligned}
X_i|\theta_i &\overset{\text{ind}}{\sim} \mathcal{K}(X_i;\theta_i), \qquad i = 1,\ldots,n \\
\theta_i|\tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \\
\tilde{p} &\sim PY(\sigma,\vartheta;P_0).
\end{aligned} \tag{2}
$$

The joint distribution of $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_n)$ is characterized by the predictive distribution of the PY, which, for any $i = 1, 2, \ldots$, is given by

$$
\begin{aligned}
\mathrm{P}(\theta_{i+1} \in \mathrm{d}t|\theta_1,\ldots,\theta_i) = &\frac{\vartheta + k_i\sigma}{\vartheta + i} P_0(\mathrm{d}t) \\
&+ \sum_{j=1}^{k_i} \frac{n_j - \sigma}{\vartheta + i} \delta_{\theta_j^*}(\mathrm{d}t),
\end{aligned} \tag{3}
$$

where $k_i$ is the number of distinct values $\theta_j^*$ observed in the first $i$ draws and $n_j$ is the number of observed $\theta_l$, for $l = 1,\ldots,i$, coinciding with $\theta_j^*$, such that $\sum_{j=1}^{k_i} n_j = i$.

Markov chain Monte Carlo (MCMC) sampling methods represent the gold standard for carrying out posterior inference based on nonparametric mixture models. Resorting to the terminology adopted by Papaspiliopoulos and Roberts (2008), most of the existing MCMC sampling methods for nonparametric mixtures can be classified into marginal and conditional, the two classes being characterized by different ways to deal with the infinite-dimensional random probability measure $\tilde{p}$. While marginal methods rely on the possibility of analytically marginalizing $\tilde{p}$ out, the conditional ones exploit suitable finite-dimensional summaries of $\tilde{p}$.

Marginal methods for nonparametric mixtures were first devised by Escobar (1988) and Escobar and West (1995), contributions that focused on DP mixtures of univariate Gaussian kernels. Extensions of such proposal include the works of Müller et al. (1996), MacEachern (1994), MacEachern and Müller (1998), Neal (2000), Barrios et al. (2013), Favaro and Teh (2013), and Lomelí et al. (2017). It is worth noting that, despite being the first class of MCMC methods for Bayesian nonparametric mixtures appeared in the literature, marginal methods are still routinely used in popular packages such as the `DPpackage` (Jara et al. 2011), the *de facto* standard software for many Bayesian nonparametric models. Alternatively, conditional methods rely on the use of summaries—of finite and possibly random dimension—of realizations of $\tilde{p}$. To this end, the stick-breaking representation for the PY (Pitman and Yor 1997) turns out to be very convenient. The almost sure discreteness of the PY allows $\tilde{p}$ to be written as an infinite sum of random jumps $\{p_j\}_{j=1}^{\infty}$ occurring at random locations $\{\tilde{\theta}_j\}_{j=1}^{\infty}$, that is

$$
\tilde{p} = \sum_{j=1}^{\infty} p_j \delta_{\tilde{\theta}_j}. \tag{4}
$$

The distribution of the locations is independent of that of the jumps and, while $\tilde{\theta}_j \overset{\text{iid}}{\sim} P_0$, the distribution of the jumps is characterized by the following construction:

$$
p_1 = V_1, \tag{5}
$$

$$
p_j = V_j \prod_{l=1}^{j-1} (1 - V_l), \tag{6}
$$

$$
V_j \overset{\text{ind}}{\sim} \text{Beta}(1-\sigma, \vartheta + j\sigma). \tag{7}
$$

A first example of conditional approach can be found in Ishwaran and James (2001) and Ishwaran and Zarepour (2002), contributions that consider a fixed truncation of the stick-breaking representation of a large class of random probability measures, and provide a bound for the introduced truncation error. Along similar lines, Muliere and Tardella (1998) and Arbel et al. (2019) make the truncation level of the DP and the PY, respectively, random so to make sure that the resulting error is smaller than a given threshold. Exact solutions that avoid introducing truncation errors are the slice samplers of Walker (2007) and Kalli et al. (2011), the improved slice sampler of Ge et al. (2015), and the retrospective sampler of Papaspiliopoulos and Roberts (2008). It is worth noticing that, although originally introduced for the case of DP mixture models, the ideas behind slice and retrospective sampling algorithms are naturally extended to the more general class of mixture models for which the mixing random probability measure admits a stick-breaking representation (Ishwaran and James 2001), thus including the PY mixture model as a special case. In this context Favaro and Walker (2013) propose a general framework for slice sampling the class of mixtures of $\sigma$-stable Poisson–Kingman model. Henceforth we will use the term slice sampling to refer to the proposals of Walker (2007) and Kalli et al. (2011), and not to the general definition of slice sampling.

Recent contributions have proposed hybrid strategies for posterior sampling nonparametric mixture models, which combine steps of marginal and conditional algorithms and therefore cannot be classified as either type of algorithm. Notable examples are the hybrid sampler of Lomelí et al. (2015) for the general class of Poisson–Kingman mixture models, and the hybrid approach proposed by Dubey et al. (2020) for a wide range of Bayesian nonparametric models based on completely random measures.

Marginal methods are appealing for their simplicity and for the fact that the number of random elements that must be drawn at each iteration of the sampler, i.e. the components of $\boldsymbol{\theta}$, is deterministic and thus bounded. At the same time, quantifying the posterior uncertainty, e.g. via posterior credible sets, by using the output of marginal methods is in general not straightforward since marginal methods do not generate

realizations of the posterior distribution of $\tilde{f}$, but only of its conditional expectation $\mathbb{E}[\tilde{f}|\boldsymbol{\theta}, \boldsymbol{X}]$, where the expectation is taken with respect to $\tilde{p}$. To this end, convenient strategies have been proposed, which typically exploit the possibility of sampling approximate realizations of $\tilde{p}$ conditionally on the values of $\boldsymbol{\theta}$ generated by the marginal algorithm (see discussions in Gelfand and Kottas 2002; Taddy and Kottas 2012; Arbel et al. 2016). Conditional methods, instead, produce approximate trajectories from the posterior distribution of $\tilde{f}$, which can be readily used to quantify posterior uncertainty. Moreover, by exploiting the conditional independence of the parameters $\theta_i$'s, given $\tilde{p}$ or a finite summary of it, conditional methods conveniently avoid sequentially updating the components of $\boldsymbol{\theta}$ at each iteration of the MCMC, thus leading to a fully parallelizable updating step within each iteration. On the other hand, the random truncation at the core of conditional methods such as slice and retrospective samplers makes the number of atoms and jumps that must be drawn at each iteration of the algorithm, random and unbounded. By confining our attention to the slice sampler of Walker (2007) and, equivalently, its dependent slice-efficient version (Kalli et al. 2011), we observe that, while its sampling routines are efficient and reliable when the DP case is considered, the same does not hold for the more general class of PY mixtures, specially when large values of $\sigma$ are considered. In practice, we noticed that, even for small sample sizes, the number of random elements that must be drawn at each iteration of the algorithm can be extremely large, often so large to make an actual implementation of the slice sampler for PY mixture models unfeasible. It is clear-cut that this limitation represents a major problem as the discount parameter $\sigma$ greatly impacts the robustness of the prior with respect to model-based clustering (see Lijoi et al. 2007; Canale and Prünster 2017). In order to shed some light on this aberrant behaviour, we investigate the distribution of the random number $N_n$ of jumps that must be drawn at each iteration of a slice sampler, implemented to carry out posterior inference based on a sample of size $n$. We can define—see Appendix A for details—a data-free lower bound for $N_n$, that is a random variable $M_n$ such that $N_n(\omega) \geq M_n(\omega)$ for every $\omega \in \Omega$ and for every sample of size $n$. $M_n$ is distributed as $\min\left\{l \geq 1 : \prod_{j\leq l}(1 - V_j) < B_n\right\}$, where the $V_j$'s are defined as in (5) and $B_n \sim \text{Beta}(1, n)$: studying the distribution of the lower bound $M_n$ will provide useful insight on $N_n$. Note that, in addition, $M_n$ coincides with the number of jumps to be drawn in order to generate a sample of size $n$ by adapting to the PY case the retrospective sampling idea introduced for the DP by Papaspiliopoulos and Roberts (2008).

Figure 1 shows the empirical distribution of $M_n$, with $n = 100$, for various combinations of $\vartheta$ and $\sigma$. The estimated median of the distribution of $M_n$ grows with $\sigma$ and, for any given value of $\sigma$, with $\vartheta$. It can be appreciated that the size of the values taken by $M_n$, and thus by $N_n$, explodes when $\sigma$ grows beyond 0.5, fact that leads to the aforementioned computational bottlenecks in routine implementations of the slice sampler. For example, when $\sigma = 0.8$, the estimated probability of $M_n$ exceeding $10^9$ is equal to 0.35, 0.42 and 0.63, for $\vartheta$ equal to 0.1, 1 and 10, respectively. From an analytic point of view, following Muliere and Tardella (1998), it is easy to show that in the DP case (i.e. $\sigma = 0$), $(M_n - 1) \sim \text{Poisson}(\vartheta \log(1/B_n))$. Beyond the DP case (i.e. $\sigma \in (0, 1)$), an application of Arbel et al. (2019) allows us to derive an analogous asymptotic result, which corroborates our empirical findings on the practical impossibility of using the slice sampler for PY mixtures with $\sigma \geq 0.5$. See Proposition 2 and related discussion in the Appendix.

Herein, we propose a new sampling strategy, named importance conditional sampling (ICS), for PY mixture

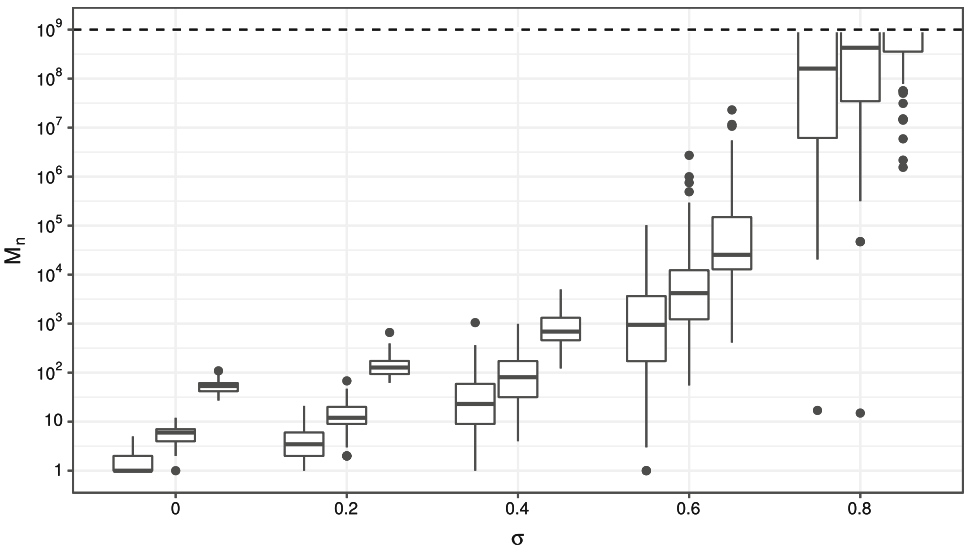


**Fig. 1** Boxplots for the empirical distributions of $M_n$, with $n = 100$, for $\sigma \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and different values of $\vartheta$, namely $\vartheta = 0.1$ (left), $\vartheta = 1$ (middle) and $\vartheta = 10$ (right). Results, based on 100 realizations of $M_n$, are truncated at $10^9$ (dashed line)

models, which combines the appealing features of both conditional and marginal methods, while avoiding their weaknesses, including the computational bottleneck depicted in Fig. 1. Like marginal methods, the ICS has a simple and interpretable sampling scheme, reminiscent of Blackwell-MacQueen's Pólya urn (Blackwell and MacQueen 1973), and allows to work with the update of a bounded number of random elements per iteration; at the same time, being a conditional method, it allows for fully parallelizable parameters update and it accounts for straightforward approximate posterior quantification. Our proposal exploits the posterior representation of the PY process, derived by Pitman (1996) in combination with an efficient sampling-importance resampling idea. The structure of Pitman (1996)'s representation makes it suitable for numerical implementations of PY based models, as indicated in Ishwaran and James (2001), and nicely implemented by Fall and Barat (2014).

The rest of the paper is organized as follows. The ICS is described in Sect. 2. Section 3 is dedicated to an extensive simulation study, comparing the performance of the ICS with state-of-the-art marginal and conditional sampling methods. Section 4 proposes (reports) an illustrative application, where the proposed algorithm is used to analyse a data set from the Collaborative Perinatal Project (Klebanoff 2009). In this context, Sect. 4.2 is dedicated to illustrate how the ICS approach can be extended to the case of nonparametric mixture models for partially exchangeable data. Section 5 concludes the paper with a discussion. Additional results are presented in the Appendix.

## 2 Importance conditional sampling

The random elements involved in a PY mixture model defined as in (2) are observations $\boldsymbol{X}$, latent parameters $\boldsymbol{\theta}$, and the PY random probability measure $\tilde{p}$. The joint distribution of $(\boldsymbol{X}, \boldsymbol{\theta}, \tilde{p})$ can be written as

$$p(\boldsymbol{X}, \boldsymbol{\theta}, \tilde{p}) = \prod_{i=1}^{n} \mathcal{K}(X_i; \theta_i) \prod_{j=1}^{k_n} \tilde{p}(\mathrm{d}\theta_j^*)^{n_j} Q(\tilde{p}), \tag{8}$$

where $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_{k_n}^*)$ is the vector of unique values in $\boldsymbol{\theta}$, with frequencies $(n_1, \ldots, n_{k_n})$ such that $\sum_{j=1}^{k_n} n_j = n$, and $Q$ is the distribution of $\tilde{p} \sim PY(\sigma, \vartheta; P_0)$. In line of principle, the full conditional distributions of all random elements can be derived from (8) and used to devise a Gibbs sampler. Given that the vector $\boldsymbol{X}$, conditionally on $\boldsymbol{\theta}$, is independent of $\tilde{p}$, the update of $\boldsymbol{\theta}$ is the only step of the Gibbs sampler which works conditionally on a realization of the infinite-dimensional $\tilde{p}$. The conditional distribution $p(\boldsymbol{\theta}|\boldsymbol{X}, \tilde{p})$ therefore will be the main focus of our attention: its study will allow us to identify a finite-dimensional summary of $\tilde{p}$, sufficient for the purpose of updating $\boldsymbol{\theta}$ from its full conditional distribution. As a result, as far as $\tilde{p}$ is concerned, only the update of its finite-dimensional summary will need to be included in the Gibbs sampler. Our proposal exploits a convenient representation of the posterior distribution of a PY process (Pitman 1996), reported in the next proposition.

**Proposition 1** (Corollary 20 in Pitman 1996). *Let $t_1, \ldots, t_n | \tilde{p} \sim \tilde{p}$ and $\tilde{p} \sim PY(\sigma, \vartheta; P_0)$, and denote by $(t_1^*, \ldots, t_{k_n}^*)$ and $(n_1, \ldots, n_{k_n})$ the set of $k_n$ distinct values and corresponding frequencies in $(t_1, \ldots, t_n)$. The conditional distribution of $\tilde{p}$, given $(t_1, \ldots, t_n)$, coincides with the distribution of*

$$p_0 \tilde{q}(\cdot) + \sum_{j=1}^{k_n} p_j \delta_{t_j^*}(\cdot),$$

*where $(p_0, p_1, \ldots, p_{k_n}) \sim Dirichlet(\vartheta + k_n\sigma, n_1 - \sigma, \ldots, n_{k_n} - \sigma)$ and $\tilde{q} \sim PY(\sigma, \vartheta + k_n\sigma; P_0)$ is independent of $(p_0, p_1, \ldots, p_{k_n})$.*

In the context of mixture models, Pitman's result implies that the full conditional distribution of $\tilde{p}$ coincides with the distribution of a mixture composed by a PY process $\tilde{q}$ with updated parameters, and a discrete random probability measure with $k_n$ fixed jump points at $\boldsymbol{t} = (t_1^*, \ldots, t_{k_n}^*)$. This means that, in the context of a Gibbs sampler, while, by conditional independence, the update of each parameter $\theta_i$ is done independently of the other parameters $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n)$, the distinct values $\boldsymbol{\theta}^*$ taken by the parameters at a given iteration, are carried on to the next iteration of the algorithm through $\tilde{p}$, in the form of fixed jump points $\boldsymbol{t}$. Specifically, if $\Theta^* = \Theta \setminus \{t_1^*, \ldots, t_{k_n}^*\}$, then, for every $i = 1, \ldots, n$, the full conditional distribution of the $i$-th parameter $\theta_i$ can be written as

$$\mathrm{P}(\theta_i \in \mathrm{d}t | X_i, \tilde{p}) \tag{9}$$

$$\propto p_0 \mathcal{K}(X_i; t)\tilde{q}(\mathrm{d}t) + \sum_{j=1}^{k_n} p_j \mathcal{K}(X_i; t_j^*)\delta_{t_j^*}(\mathrm{d}t), \tag{10}$$

where $\tilde{q}$ is the restriction of $\tilde{p}$ to $\Theta^*$, $p_0 = \tilde{p}(\Theta^*)$ and $p_j = \tilde{p}(t_j^*)$, for every $j = 1, \ldots, k_n$. The full conditional in (9) is reminiscent of the Blackwell-MacQueen urn scheme characterizing the update of the parameters in marginal methods: the parameter $\theta_i$ can either coincide with one of the $k_n$ fixed jump points of $\tilde{p}$ or take a new value from a distribution proportional to $\mathcal{K}(X_i; t)\tilde{q}(\mathrm{d}t)$. The key observation at the basis of the ICS is that, for the purpose of updating the parameters $\boldsymbol{\theta}$, there is no need to know the whole realization of $\tilde{p}$ but it suffices to know the vector $\boldsymbol{t}$ of fixed jump points of $\tilde{p}$, the value $\boldsymbol{p} = (p_0, p_1, \ldots, p_{k_n})$ taken by $\tilde{p}$ at

the partition $(\Theta^*, t_1^*, \ldots, t_{k_n}^*)$ of $\Theta$, and to be able to sample from a distribution proportional to $\mathcal{K}(X_i, t)\tilde{q}(\mathrm{d}t)$. For the latter task, we adopt a sampling-importance resampling approach (see, e.g., Smith and Gelfand 1992) with proposal distribution $\tilde{q}$. It is remarkable that such solution allows us to approximately sample from the target distribution while avoiding the daunting task of simulating a realization of $\tilde{q}$ itself. Indeed, for any $m \geq 1$, a vector $\boldsymbol{s} = (s_1, \ldots, s_m)$ such that $s_i | \tilde{q} \overset{\text{iid}}{\sim} \tilde{q}$ can be generated by means of an urn scheme exploiting (3). Given the almost sure discreteness of $\tilde{q}$, the generated vector will show ties with positive probability and thus will feature $r_m \leq m$ distinct values $(s_1^*, \ldots, s_{r_m}^*)$, with frequencies $(m_1, \ldots, m_{r_m})$ such that $\sum_{j=1}^{r_m} m_j = m$. In turn, importance weights for the resampling step are computed, for any $\ell = 1, \ldots, m$, as

$$w_\ell = \frac{\mathcal{K}(X_i, s_\ell)\tilde{q}(s_\ell)}{\tilde{q}(s_\ell)} = \mathcal{K}(X_i, s_\ell),$$

thus without requiring the evaluation of $\tilde{q}$. As a result, the full conditional (9) can be rewritten as

$$\mathrm{P}(\theta_i \in \mathrm{d}t | X_i, \tilde{p}) \propto p_0 \sum_{j=1}^{r_m} \frac{m_j}{m} \mathcal{K}(X_i; s_j^*)\delta_{s_j^*}(\mathrm{d}t) + \sum_{j=1}^{k_n} p_j \mathcal{K}(X_i; t_j^*)\delta_{t_j^*}(\mathrm{d}t). \tag{11}$$

Once more we highlight an interesting analogy between the conditional approach we propose and marginal methods: the introduction of the auxiliary random variables $s_1^*, \ldots, s_{r_m}^*$ reminds of the augmentation introduced in Algorithm 8 of Neal (2000), marginal algorithm proposed to deal with a non-conjugate specification of the mixture model. From (11) it is straightforward to identify $(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{p})$ as a finite-dimensional summary of $\tilde{p}$, sufficient for the purpose of updating the parameters $\theta_i$ from their full conditionals. This means that, as far as $\tilde{p}$ is concerned, only its summary $(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{p})$ must be included in the updating steps of the Gibbs sampler. To this end, Proposition 1 provides the basis for the update of $(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{p})$. Indeed, conditionally on $\boldsymbol{\theta}$, the fixed jump points $\boldsymbol{t}$ coincide with the $k_n$ distinct values appearing in $\boldsymbol{\theta}$, while the random vectors $\boldsymbol{p}$ and $\boldsymbol{s}$ are independent with $\boldsymbol{p} \sim \text{Dirichlet}(\vartheta + \sigma k_n, n_1 - \sigma, \ldots, n_{k_n} - \sigma)$ and the joint distribution of $\boldsymbol{s}$ characterized by the predictive distribution of a $\text{PY}(\sigma, \vartheta + \sigma k_n; P_0)$, that is, for any $\ell = 0, 1, \ldots, m-1$,

$$\mathrm{P}(s_{\ell+1} \in \mathrm{d}s | s_1, \ldots, s_\ell) = \frac{\vartheta + \sigma(k_n + r_\ell)}{\vartheta + \sigma k_n + \ell} P_0(\mathrm{d}s) + \sum_{j=1}^{r_\ell} \frac{m_j - \sigma}{\vartheta + \sigma k_n + \ell} \delta_{s_j^*}(\mathrm{d}s), \tag{12}$$

where $(s_1^*, \ldots, s_{r_\ell}^*)$ is the vector of $r_\ell$ distinct values appearing in $(s_1, \ldots, s_\ell)$, with corresponding frequencies $(m_1, \ldots, m_{r_\ell})$ such that $\sum_{j=1}^{r_\ell} m_j = \ell$.

By combining the steps just described, as summarized in Algorithm 1, we can then devise a Gibbs sampler which we name ICS. In Algorithm 1 and henceforth, the superscript $(r)$ is used to denote the value taken by a random variable at the $r$-th iteration. In order to improve mixing, the ICS includes an acceleration step which consists in updating, at the end of each iteration, the distinct values $\boldsymbol{\theta}^*$ from their full conditional distributions. Namely, for every $j = 1, \ldots, k_n$,

$$\mathrm{P}(\theta_j^* \in \mathrm{d}t | \boldsymbol{X}) \propto P_0(\mathrm{d}t) \prod_{i \in C_j} \mathcal{K}(X_i; t), \tag{13}$$

where $C_j = \{i \in \{1, \ldots, n\} : \theta_i = \theta_j^*\}$.

Finally, a realization from the posterior distribution of $(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{p})$ defines an approximate realization $f$ of the posterior distribution of the random density defined in (1), that is

$$\tilde{f}_m(x) = p_0 \sum_{l=1}^{r_m} \frac{m_l}{m} \mathcal{K}(x; s_l^*) + \sum_{j=1}^{k_n} p_j \mathcal{K}(x; t_j^*). \tag{14}$$

If the algorithm is run for a total of $R$ iterations, the first $R_b$ of which discarded as burn-in, then the posterior mean is estimated by

$$\hat{f}(x) = \frac{1}{R - R_b} \sum_{r=R_b+1}^{R} \tilde{f}_m^{(r)}(x),$$

where $\tilde{f}_m^{(r)}$ denotes the approximate density sampled from the posterior at the $r$-th iteration. The set of densities $\tilde{f}_m^{(r)}$ can be also used to quantify posterior uncertainty. It is worth remarking though that any such quantification is based on realizations of a finite dimensional summary of the infinite-dimensional $\tilde{p}$ and thus is, by its nature, approximated. For a quantification of the approximating error one could resort to Arbel et al. (2019).

It is instructive to consider how the ICS works for the special case of DP mixture models, that is when $\sigma = 0$. In such case, the steps described in Algorithm 1 can be nicely interpreted by resorting to three fundamental properties characterizing the DP, namely conjugacy, self-similarity, and availability of finite-dimensional distributions. More specifically, when $\sigma = 0$, step 4 of Algorithm 1 consists in generating the random weights $\boldsymbol{p}$ from a Dirichlet distribution of parameters $(\vartheta, n_1, \ldots, n_{k_n})$. This follows by combining the conjugacy of the DP (Ferguson 1973), for which

**Algorithm 1:** ICS for PY mixture model

*Set admissible initial values* $\boldsymbol{\theta}^{(0)}$
**for** *each iteration* $r = 1, \ldots, R$ **do**
  **set** $\boldsymbol{t}^{(r)} = \boldsymbol{\theta}^{*(r-1)}$;
  **sample** $\boldsymbol{p}^{(r)}$ *from* $\boldsymbol{p}^{(r)} \sim \text{Dirichlet}(\vartheta + \sigma k_n^{(r-1)}, n_1^{(r-1)} - \sigma, \ldots, n_{k_n}^{(r-1)} - \sigma)$;
  **for** *each* $\ell = 0, \ldots, m-1$ **do**
    **let** $r_\ell^{(r)}$ *be the number of distinct values in* $(s_1^{(r)}, \ldots, s_\ell^{(r)})$, **sample** $s_{\ell+1}^{(r)}$ *from*

$$\mathrm{P}(s_{\ell+1}^{(r)} \in \cdot \,|\, s_1^{(r)}, \ldots, s_\ell^{(r)}) = \frac{\vartheta + \sigma(k_n^{(r-1)} + r_\ell^{(r)})}{\vartheta + \sigma k_n^{(r-1)} + \ell} P_0(\cdot) + \sum_{j=1}^{r_\ell^{(r)}} \frac{m_j^{(r)} - \sigma}{\vartheta + \sigma k_n^{(r-1)} + \ell} \delta_{s_j^{*(r)}}(\cdot);$$

  **let** $r_m^{(r)}$ *be the number of distinct values in* $\boldsymbol{s}^{(r)}$;
  **for** *each* $i = 1, \ldots, n$ **do**
    **sample** $\theta_i^{(r)}$ *from*

$$\mathrm{P}(\theta_i^{(r)} = t \,|\, \cdots) \propto \begin{cases} p_0^{(r)} \frac{m_\ell^{(r)}}{m} \mathcal{K}(X_i; s_\ell^{*(r)}) & \text{if } t \in \{s_1^{*(r)}, \ldots, s_{r_m^{(r)}}^{*(r)}\} \\ p_j^{(r)} \mathcal{K}(X_i; t_j^{*(r)}) & \text{if } t \in \{t_1^{*(r)}, \ldots, t_{k_n^{(r-1)}}^{*(r)}\} \\ 0 & \text{otherwise} \end{cases}$$

  **let** $\boldsymbol{\theta}^{*(r)} = (\theta_1^{*(r)}, \ldots, \theta_{k_n^{(r)}}^{*(r)})$ *be the vector of distinct parameters in* $\boldsymbol{\theta}^{(r)}$;
  **for** *each* $j = 1, \ldots, k_n^{(r)}$ **do**
    **let** $C_j^{(r)}$ *be the set of indexes* $i$ *such that* $\theta_i^{(r)} = \theta_j^{*(r)}$;
    **update** $\theta_j^{*(r)}$ *from* $\mathrm{P}(\theta_j^{*(r)} \in \mathrm{d}t \,|\, \cdots) \propto P_0(\mathrm{d}t) \prod_{i \in C_j^{(r)}} \mathcal{K}(X_i; t)$;
**end**

$$\tilde{p}|\boldsymbol{\theta} \sim DP\left(\vartheta + n; \frac{\vartheta}{\vartheta + n} P_0 + \sum_{j=1}^{k_n} \frac{n_j}{\vartheta + n} \delta_{\theta_j^*}\right),$$

with the availability of finite-dimensional distributions of DP (Ferguson 1973), which provides the distribution of $\boldsymbol{p}$, defined as the evaluation of the conditional distribution of $\tilde{p}$ on the partition of $\Theta$ induced by $\boldsymbol{\theta}$. Moreover, when $\sigma = 0$, according to the predictive distribution displayed in step 6 of Algorithm 1, the auxiliary random variables $\boldsymbol{s}$ are exchangeable from $\tilde{q} \sim DP(\vartheta; P_0)$, with $\tilde{q}$ independent of $\boldsymbol{p}$. This is nicely implied by the self-similarity of the DP (see, e.g., Ghosal 2010), according to which $\tilde{q} = \tilde{p}|_{\Theta^*}$ is independent of $\tilde{p}|_{\Theta \setminus \Theta^*}$, and therefore of $\boldsymbol{p}$, and is distributed as a $DP(\vartheta P_0(\Theta^*); P_0|_{\Theta^*})$, and by the diffuseness of $P_0$. As a result, in the DP case, the auxiliary random variables $\boldsymbol{s}$ are generated from the prior model.

## 3 Simulation study

We performed a simulation study to analyze the performance of the ICS algorithm and to compare it with marginal and slice samplers. For the latter, two versions proposed by Kalli et al. (2011) were considered, namely the dependent and the independent slice-efficient algorithms. The independent version of the algorithm requires the specification of a deterministic sequence $\xi_1, \xi_2, \ldots$, which in our implementation was set equal to $\mathbb{E}[p_1], \mathbb{E}[p_2], \ldots$, with the $p_j$'s defined in (5), in analogy with what was proposed by Kalli et al. (2011) for the DP (see Algorithm 5 in the Supplementary Material for more details). All algorithms were written in `C++` and are implemented in the `BNPmix` package (Corradin et al. 2021), available on CRAN. Aware that different implementations can lead to a biased comparison (see Kriegel et al. 2017, for an insightful discussion), we aimed at reducing such bias to a minimum by letting the four algorithms considered here share the same code for most sub-routines.

Throughout this section we consider synthetic data generated from a simple two-component mixture of Gaussians, namely $f_0(x) = 0.75\phi(x; -2.5, 1) + 0.25\phi(x; 2.5, 1)$, with $\phi(\cdot; \mu, \sigma^2)$ denoting the density of a Gaussian random variable with mean $\mu$ and variance $\sigma^2$. All data were analyzed by means of the nonparametric mixture model defined in (1) and specified by considering a univariate Gaussian kernel $\mathcal{K}(x, \theta) = \phi(x; \mu, \sigma^2)$, with $\theta = (\mu, \sigma^2)$, and by assuming a normal-inverse gamma base measure $P_0$ such that $\sigma^2 \sim IG(2, 1)$ and $\mu|\sigma^2 \sim N(0, 5\sigma^2)$. Different combinations of values for the parameters $\sigma$ and $\vartheta$, and for the sample size $n$ were considered. The results of this section

are then obtained as averages over a specified number of replicates. All algorithms were run for 1 500 iterations, of which the first 500 discarded as burn-in. Convergence of the chains was checked by visual inspection of the trace plots of randomly selected runs, which did not provide any evidence against it. The analysis was carried out by running `BNPmix` on R 4.0.3 on a 64-bit Windows machine with a 3.4-GHz Intel quad-core i7-3770 processor and 16 GB of RAM.

The first part of our investigation is dedicated to the role of $m$, the size of the auxiliary sample generated for the sampling-importance resampling step within the ICS. To this end, we considered two sample sizes, namely $n = 100$ and $n = 1\,000$, and generated 100 data sets per size. Such data were then analyzed by considering a combination of values for the PY parameters, namely $\sigma \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and $\vartheta \in \{1, 10\}$, and by running the ICS with $m \in \{1, 10, 100\}$. Estimated posterior densities, not displayed here, did not show any noticeable effect of $m$. More interesting findings were obtained when the analysis focused on the quality of the generated posterior sample: larger values for $m$ appear to lead to a better mixing of the Markov chain at the price of additional computational cost. These effects were measured by considering the effective sample size (ESS), computed by resorting to the `CODA` package (Plummer et al. 2006), and the ratio between runtime, in seconds, and ESS (time/ESS), both averaged over 100 replicates. Following the algorithmic performance analyses of Neal (2000), Papaspiliopoulos and Roberts (2008) and Kalli et al. (2011), the ESS was computed on the number of clusters—$k_n$ as far as the ICS is concerned—and on the deviance of the estimated density, with the latter defined as

$$\text{dev}(\boldsymbol{X}, \boldsymbol{\theta}^{(r)}) = -2 \sum_{i=1}^{n} \left( \sum_{j=1}^{k_n^{(r)}} \frac{n_j^{(r)}}{n} \mathcal{K}(X_i; \theta_j^{*(r)}) \right),$$

for the $r$-th MCMC draw. The ratio time/ESS takes into account both quality of the generated sample and computational cost, and can be interpreted as the average time needed to sample one independent draw from the posterior.

The results show that larger values of $m$ lead, on average, to a larger ESS, that is to better quality posterior samples. This is displayed in the top row of Fig. 2, which shows the estimated ESS for $k_n$. We observe that, when averaging over all the considered scenarios, the ESS obtained by setting $m = 100$ is 1.82 and 1.09 times larger than the average ESS obtained by setting $m = 1$ and $m = 10$, respectively. At the same time, larger values of $m$ require drawing more random objects per iteration and thus, as expected, lead to longer runtimes. In this sense, the bottom row of Fig. 2 clearly indicates that, as far as $k_n$ is concerned, the ratio time/ESS tends to be larger for larger values of $m$. This is particularly evident, for example, when $\sigma = 0.8$ as the ratio time/ESS corresponding to $m = 100$ is, on average, 1.81 and 1.06 times larger than the same ratio corresponding to $m = 1$ and $m = 10$, respectively. Similar conclusions can be drawn by looking at Fig.

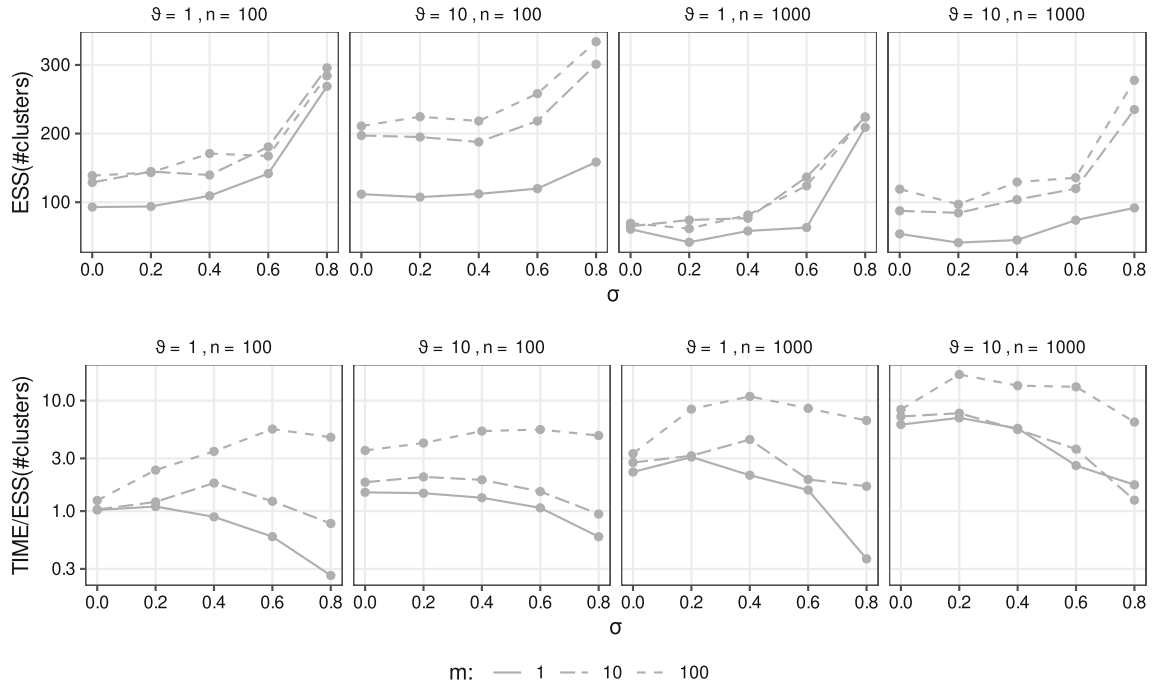


**Fig. 2** Simulated data. ICS: ESS computed on the random variable number of clusters (top row) and ratio between runtime (in seconds) and ESS for the same random variable (bottom row). Results are averaged over 100 replicates

8, presented in Appendix B and displaying time/ESS for the deviance of the estimated densities.

When implementing the ICS, the value of $m$ can be tuned based on the desired algorithm performance in terms of quality of mixing and runtime. As for the rest of the paper, and for the ease of illustration, we will work with $m = 10$, chosen as a sensible compromise between good mixing and controlled computational cost.

The second part of the simulation study compares the performance of ICS, marginal sampler, dependent and independent slice-efficient samplers. For the sake of clarity, pseudo-code of the implemented algorithms is provided as supplementary material. We considered the sample sizes $n = 100$, $n = 250$ and $n = 1\,000$, and generated 10 data sets per size from $f_0$. These data were then analyzed by considering a combination of values for the PY parameters, namely $\sigma \in \{0, 0.2, 0.4, 0.6, 0.8\}$ and $\vartheta \in \{1, 10, 25\}$. The results we report are obtained, for each scenario, by averaging over the 10 replicates. As for the two slice samplers, due to the aforementioned explosion of the number of drawings per iteration when $\sigma$ takes large values, our analysis was forcefully confined to the case $\sigma \leq 0.4$. Moreover, the results referring to the case $\sigma = 0.4$ are approximate as they were obtained by constraining the slice sampler to draw at most $10^5$ components at each iteration: such limitation of our study could not be avoided, given the otherwise unmanageable computational burden associated with this specific setting. Table 1 in Appendix B shows that such bound was reached more often when large data sets were analyzed. For example, while for $n = 100$ the bound was reached on average 12% and 15% of the iterations, for independent and dependent slice-efficient samplers respectively, the same happened on average 26% and 40% of the iterations when $n = 1\,000$. For this reason, these specific results must be considered approximated and, as far as the runtime is concerned, conservative. The four algorithms were compared by using the same measures adopted in the first part of the simulation study, namely the ESS for the number of clusters, the ESS for the deviance of the estimated density, and the corresponding ratios time/ESS.

A clear trend can be appreciated in Fig. 3 where the focus is on the ESS for the number of clusters: the marginal sampler displays, on average, a larger ESS than ICS, whose ESS appears, in turn, uniformly larger than the ones characterizing the two slice samplers. As for the latter two, while the displayed trend is similar, it can be appreciated that the independent algorithm is uniformly characterized by a better mixing. Results referring to the ratio time/ESS, for the variable number of clusters, are displayed in Fig. 4. ICS and marginal sampler show in general similar performances. It is interesting to notice though that, while the ratio time/ESS for the marginal algorithm is rather stable over the values of $\sigma$ considered in the study, the same quantity for ICS indicates a slightly better performance when $\sigma$ takes large values. On the other hand, the efficiency of the two slice samplers is heavily affected by the value of $\sigma$, with time/ESS exploding when $\sigma$ moves from 0 to 0.4 and when $\vartheta$ increases. On the basis of this study, the slice samplers appear competitive options when $\sigma \in \{0, 0.2\}$ and a small $\vartheta$ are considered. On the contrary, it is apparent that larger values of $\sigma$ make the two slice samplers less efficient than ICS and marginal sampler. Similar considerations can be drawn when analyzing the performance in terms of deviance of the estimated densities,

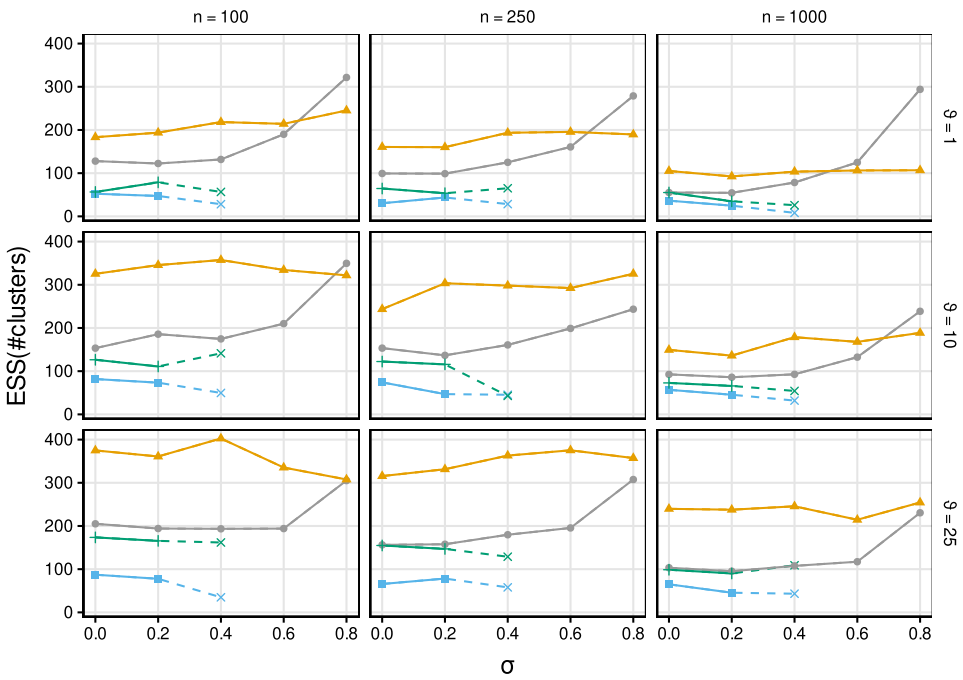


**Fig. 3** Simulated data. ESS computed on the random variable number of clusters, for ICS (gray), marginal sampler (orange), independent slice-efficient sampler (green) and dependent slice-efficient sampler (blue). Results are averaged over 10 replicates. The $\times$-shaped marker for the two slice samplers indicates that, when $\sigma = 0.4$, the value of the ESS is obtained with an arbitrary upper bound at $10^5$ for the number of jumps drawn per iteration

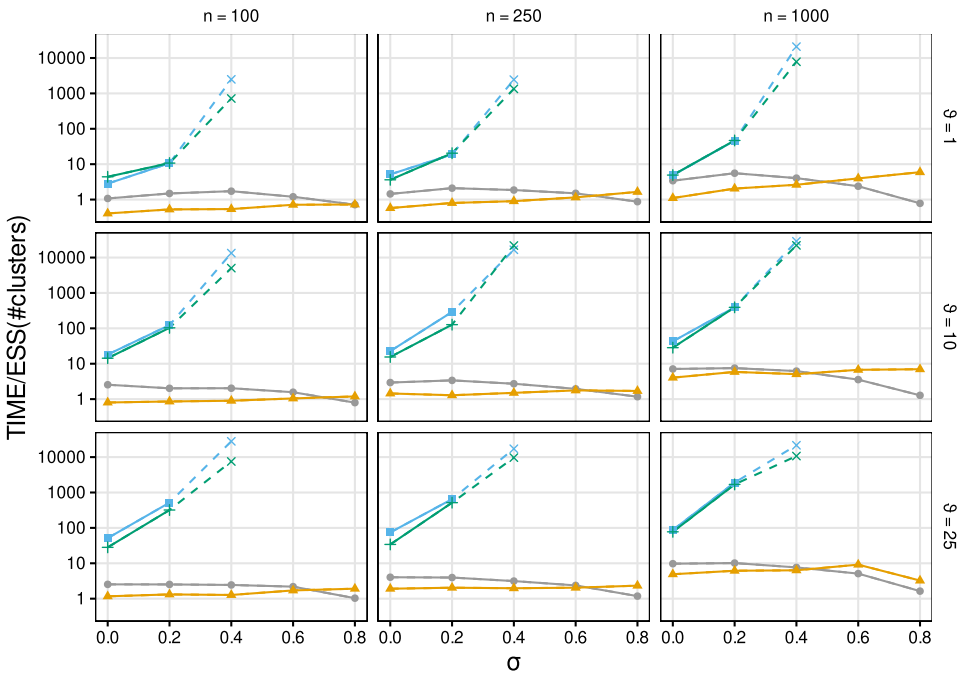


**Fig. 4** Simulated data. Ratio of runtime (in seconds) over ESS, in log-scale, computed for the number of clusters, for ICS (gray), marginal sampler (orange), independent slice-efficient sampler (green) and dependent slice-efficient sampler (blue). Results are averaged over 10 replicates. The $\times$-shaped marker for the two slice samplers indicates that, when $\sigma = 0.4$, the value of time/ESS is obtained with an arbitrary upper bound at $10^5$ for the number of jumps drawn per iteration

with the plots for ESS and the ratio time/ESS displayed in Figs. 9 and 10 in Appendix B.

## 4 Illustrations

We consider a data set from the Collaborative Perinatal Project (CPP), a large prospective study of the cause of neurological disorders and other pathologies in children in the United States. Pregnant women were enrolled between 1959 and 1966 when they showed up for prenatal care at one of 12 hospitals. While several measurements per pregnancy are available, our attention focuses on two main quantities: the gestational age (in weeks) and the logarithm of the concentration level of DDE in $\mu g/l$, a persistent metabolite of the pesticide DDT, known to have adverse impact on the gestational age (Longnecker et al. 2001). Our analysis has a two-fold goal. First, we focus on estimating and comparing the joint density of gestational age and DDE for two groups of women, namely smokers and non-smokers. This will also allow us to assess how the probability of premature birth varies conditionally on the level of DDE. Adopting a nonparametric mixture model will allow us to investigate the presence of clusters within the data. Second, we consider the data set partitioned in the 12 hospitals of the study and focus on the estimation of the hospital-specific distribution of the gestational age, by accounting for possible association across subsamples collected at different hospitals. For this analysis we adopt a nonparametric mixture model for partially exchangeable data and propose an extension of the ICS approach presented in Sect. 2.

### 4.1 Cross-hospital analysis

Smokers and non-smokers samples have size of $n_1 = 1023$ and $n_2 = 1290$, respectively. For the two groups we independently model the joint distribution of gestational age and DDE by means of a PY mixture model (2) with bivariate Gaussian kernel function $\mathcal{K}(x, \boldsymbol{\theta}) = \phi(x, \boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and with conjugate normal-inverse Wishart base measure $P_0 = N\text{-}IW(\boldsymbol{m}_0, k_0, \nu_0, \boldsymbol{S}_0)$. In absence of precise prior information on the density to be estimated, we specify a vague base measure following an empirical Bayes approach. Specifically we let $\boldsymbol{m}_0$ be equal to the sample average, $\boldsymbol{S}_0$ be equal to three times the empirical covariance, $k_0 = 1/10$, and $\nu_0 = 5$. These settings are equivalent to assuming that the scale parameter of the generic mixture component coincides with 1.5 times the empirical covariance, while the location parameter is centered on the sample mean with prior variance equal to 10 times the scale parameter. Next, we set the parameters $\vartheta$ and $\sigma$ on the basis of the prior distribution they imply on the number of clusters $k_n$, within each group. Specifically, we set the prior expectation and prior standard deviation for $k_n$ equal to 10 and 20, respectively. Our choice implies that a small probability ($\approx 0.05$) is assigned to the event $k_n \geq 50$. This argument leads to set $(\sigma, \vartheta)$ equal to $(0.548, -0.485)$ and $(0.5295, -0.4660)$ for the groups of smokers and non-smokers, respectively. The values specified for $\sigma$ are thus larger than 0.5, a situation that is conveniently tackled by the ICS, as displayed by the simulation study of Sect. 3. An alternative modelling strategy is achieved by introducing a hyperprior distribution for both $\sigma$ and $\theta$. While not explored in this illustration, it is worth stressing that this

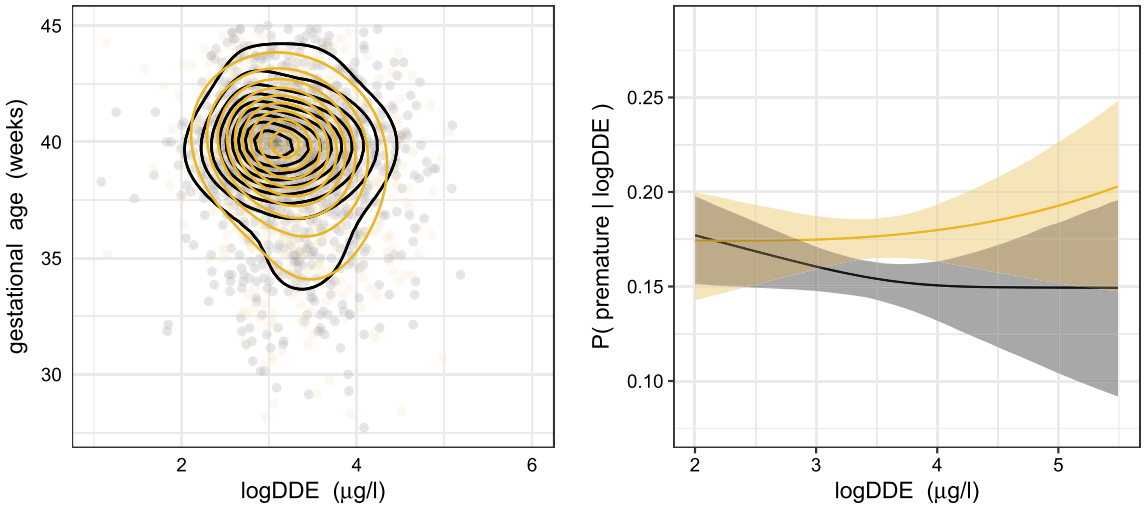


**Fig. 5** CPP cross-hospital data. Left: observations and contour curves of the estimated joint posterior density of gestational age and DDE, for smokers (yellow dots and curves) and non-smokers (black dots and curves). Right: estimated probability of premature birth (gestational age below 37 weeks), conditionally on the level of DDE, for smokers (yellow curves) and non smokers (black curves), and associated pointwise 90% quantile-based posterior credible bands (filled areas)

strategy might be conveniently implemented by adopting the ICS: if the prior on $\sigma$ is defined on $(0, 1)$, an implementation of the model requires a sampler whose efficiency is not compromised by the specific values of $\sigma$ explored by the chain.

The analysis of both samples was carried out by running the ICS for 12 000 iterations, with the first 7 000 discarded as burn-in. Convergence of the chain was assessed as satisfactory by visually investigating the trace plots and by means of the Geweke's diagnostics (Geweke 1992). Running the analysis of the two samples took less than two minutes in total. It is important to stress that, given the model specification, the same analysis could not be carried out by implementing the two versions of the slice samplers we considered (see Algorithms 4 and 5 in the Supplementary Material), as the value of $\sigma$ would make computations prohibitive. We could instead implement the marginal sampler (Algorithm 3 in the Supplementary Material) which, as expected, took considerably longer than the ICS (about 11 minutes), due to the moderately large sample sizes.

The contour curves of the estimated joint densities of gestational age and DDE for the two groups are displayed in the left panel of Fig. 5 and suggest different distributions between smokers and non-smokers, specially when large values for DDE are considered. Differences between the two groups are further highlighted by the right panel of Fig. 5, which shows the estimated probability—along with corresponding pointwise 90% posterior credible bands—of premature birth (i.e. gestational age smaller than 37 weeks), conditionally on the value taken by DDE, for the two groups. Once again, a difference between smokers and non-smokers can be appreciated for large levels of DDE, although a sizeable uncertainty is associated with posterior estimates, as displayed by the large credible bands. Although the difference between the estimated densities for smokers and non-smokers is narrow, Fig. 5 suggests that being a smoker might be a risk factor: smoking mothers face a higher risk of premature birth, more apparently for large levels of concentration of DDE, and their average gestational age is overall slightly smaller than the one of non-smokers.

## 4.2 Multi-hospital analysis

The same data set as in the previous section is considered here, with observations classified according to both smoking habits of women and the hospitals where they were enrolled. This leads to two samples stratified into $L = 12$ strata, with cardinalities summarized by the vectors

$$\begin{aligned}\boldsymbol{n}_1 &= (n_{1,1}, n_{2,1}, \ldots, n_{12,1}) \\ &= (236, 51, 59, 38, 92, 56, 67, 51, 61, 187, 81, 44)\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{n}_2 &= (n_{1,2}, n_{2,2}, \ldots, n_{12,2}) \\ &= (245, 73, 91, 39, 113, 98, 74, 90, 56, 197, 70, 144)\end{aligned}$$

for smokers and non-smokers, respectively. The focus of the analysis is modelling the distribution of gestational age.

### 4.2.1 A mixture model for partially exchangeable data

Smokers and non-smokers data are analyzed independently. For each group, heterogeneity across hospitals suggests to assume that data are partially exchangeable in the sense of de Finetti (1938). To account for this assumption, we consider

a mixture model for partially exchangeable data, where the stratum-specific mixing random probability measures form the components of a dependent Dirichlet process. Within this flexible class of processes (see Foti and Williamson 2015, and references therein), we consider the Griffiths-Milne dependent Dirichlet processes (GM-DDP), as defined and studied in Lijoi et al. (2014a, b). For an allied approach see Griffin et al. (2013). Let $X_{i,l}$ be the gestational age of the $i$-th woman in the $l$-th hospital, and $\boldsymbol{\theta}_l$ be the vector of latent variables $\theta_{i,l}$ referring to the $l$-th hospital. The mixture model can be represented in its hierarchical form as

$$\begin{aligned} X_{i,l}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L &\overset{\text{ind}}{\sim} \mathcal{K}(x_{i,l}; \theta_{i,l}) \\ \theta_{i_l,l}|(\tilde{p}_1, \ldots, \tilde{p}_L) &\overset{\text{iid}}{\sim} \tilde{p}_l \\ (\tilde{p}_1, \ldots, \tilde{p}_L) &\sim \text{GM-DDP}(\vartheta, z; P_0), \end{aligned} \tag{15}$$

with $l = 1, \ldots, L, i = 1, \ldots, n_l$, $\vartheta > 0$, $z \in (0, 1)$, $P_0$ is a probability distribution on $\mathbb{R} \times \mathbb{R}^+$, and the GM-DDP distribution of the vector $(\tilde{p}_1, \ldots, \tilde{p}_L)$ coincides with the distribution of the vector of random probability measures whose components are defined, for every $l = 1, \ldots, L$, as

$$\tilde{p}_l = \gamma_l\, w_l + \gamma_0\, (1 - w_l),$$

where $\gamma_1, \ldots, \gamma_L \overset{\text{iid}}{\sim} DP(\vartheta z; P_0)$ and $\gamma_0 \sim DP(\vartheta(1 - z); P_0)$ is independent of $\gamma_l$, for any $l = 1, \ldots, L$. Moreover, the vector of random weights $\boldsymbol{w} = (w_1, \ldots, w_L)$, taking values in $[0, 1]^L$, is distributed as a multivariate beta of parameters $(\vartheta z, \ldots, \vartheta z, \vartheta(1 - z))$, as defined in Olkin and Liu (2003), and its components are independent of the random probability measures $\gamma_0, \gamma_1, \ldots, \gamma_L$. As a result, the random probabilities $\tilde{p}_l$ are, marginally, identically distributed with $\tilde{p}_l \sim DP(\vartheta; P_0)$ (see Lijoi et al. 2014a, for details).

#### 4.2.2 ICS for GM-DDP mixture model and its application

The ICS can be easily adapted to a variety of models. For example, it naturally fits the partially exchangeable framework of model (15). The ICS algorithm for GM-DDP mixture models is described in Algorithm 2 in Appendix C, and consists of three main steps.

First, conditionally on the allocation of observations to clusters referring to either the idiosyncratic process $\gamma_l$, with $l = 1, \ldots, L$, or the common process $\gamma_0$, summaries of all the processes, that is $(\boldsymbol{s}_l, \boldsymbol{t}_l, \boldsymbol{p}_l)$, for $l = 0, \ldots, L$, are updated as done in Sect. 2 for a single process, with the proviso that $\sigma = 0$. Second, the latent variables $\theta_{i,l}$ are updated for every $l = 1, \ldots, L$ and $1 \leq i \leq n_l$; and, third, the components of $\boldsymbol{w}$ are sampled. The full conditional distributions for $\theta_{i,l}$ and $\boldsymbol{w}$ are provided in Appendix C. Model (15) is specified by assuming a univariate Gaussian kernel and normal-inverse gamma base measure $P_0 = N\text{-}IG(0, 5, 4, 1)$. Moreover, the specification $\vartheta = 1$ and $z = 0.5$ is adopted, with the latter choice corresponding to equal prior weights assigned to idiosyncratic and common components $\gamma_l$ and $\gamma_0$. The ICS algorithm for the GM-DDP mixture model was run for 10 000 iterations, the first 5 000 of which were discarded as burn-in. Estimating posterior densities for smokers and non-smokers required a total runtime of less than two and a half minutes. Convergence of the chains was assessed by visually investigating the trace plots, which did not provide any evidence against it.

Figure 6 shows the estimated densities of the gestational age, for each stratum, with a comparison between smokers and non-smokers. The distribution for smokers is globally more skewed and shifted to the left than the one for non-smokers, indicating an expected more adverse effect of smoking on gestational age.

## 5 Discussion

We proposed a new sampling strategy for PY mixture models, named ICS, which combines desirable properties of existing marginal and conditional methods: the ICS shares easy interpretability with marginal methods, while allowing, likewise conditional samplers, for a parallelizable update of the latent parameters $\boldsymbol{\theta}$, and for a straightforward quantification of posterior uncertainty. The simulation study of Sect. 3 showed that the ICS overtakes some of the computational bottlenecks characterizing the conditional methods considered in the comparison. Specifically, the ICS can be implemented for any value of the discount parameter $\sigma$, with its efficiency being stable to the specification of $\sigma$. This is appealing as the discount parameter plays a crucial modelling role when PY mixture models are used for model-based clustering: the ICS allows for an efficient implementation of such models, without the need of setting artificial constraints on the value of $\sigma$. As far as the comparison of the performances of ICS and other algorithms is concerned, it is important to remark that the independent slice-efficient algorithm proposed by Kalli et al. (2011) is more general than the one considered in Sect. 3 as other specifications of the deterministic sequence $\xi_1, \xi_2, \ldots$ are possible. As nicely discussed by Kalli et al. (2011), the choice of such sequence "is a delicate issue and any choice has to balance efficiency and computational time". Alternative specifications of the sequence may be explored on a case-by-case basis but, in our experience, the computational time can be reduced only at the cost of worsening the mixing of the algorithm. It is also worth remarking that the ICS does not rely on any assumption of conjugacy between base measure and kernel, and thus it can be considered by all means a valid alternative to the celebrated Algorithm 8 of Neal (2000) when non-conjugate mixture models are to be implemented.

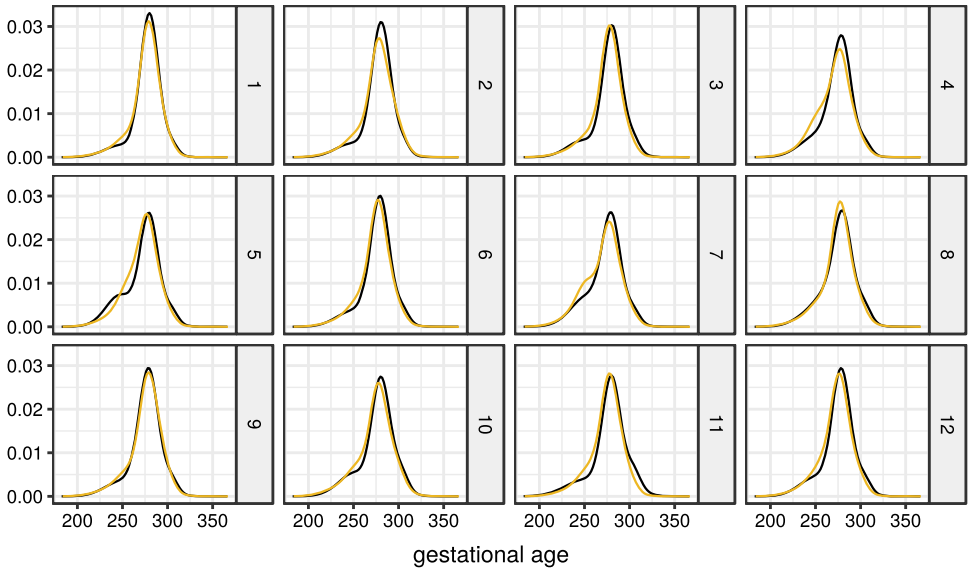


**Fig. 6** CPP multi-hospital data. Estimated densities of the gestational age for the 12 hospitals, with comparison between smokers (yellow curves) and non-smokers (black curves)

Finally, while originally introduced to overtake computational problems arising in the implementation of algorithms for PY mixture models, the idea behind the ICS approach can be naturally extended to other classes of computationally demanding models. As an example, we implemented the same idea to deal with posterior inference based on a flexible class of mixture models for partially exchangeable data. Other extensions are also possible and are currently subject of ongoing research.

## Supplementary Information

Details on the implementation of the algorithms considered in Sect. 3 are provided as supplementary material.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s11222-022-10096-0.

**Acknowledgements** The first author is supported by the University of Padova under the STARS Grant. The second and the third authors are grateful to the DEMS Data Science Lab for supporting this work by providing computational resources.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## Appendix A On the number of jumps to be drawn with the slice sampler

Let $N_n$ be the random number of jumps that need to be drawn at each iteration of a slice sampler (Walker 2007) or, equivalently, its dependent slice-efficient version (Kalli et al. 2011), implemented to carry out posterior inference based on a sample of size $n$. Conditionally on the cluster assignment variables $c_1, \ldots, c_n$ and on the weights $p_{c_1}, \ldots, p_{c_n}$ of the non-empty components of the mixture, $N_n$ is given by

$$N_n = \min \left\{ l \geq 1 \, : \, \sum_{j \leq l} p_j > 1 - \min(U_1 p_{c_1}, \ldots, U_n p_{c_n}) \right\},$$

where the random weights $p_j$'s are defined as in (5) and $U_1, \ldots, U_n$ are independent uniform random variables, independent of the weights $p_j$'s. We next define a second random variable $M_n$, function of the same uniform random variables $U_1, \ldots, U_n$, as

$$\begin{aligned} M_n &= \min \left\{ l \geq 1 \, : \, \sum_{j \leq l} p_j > 1 - \min(U_1, \ldots, U_n) \right\} \\ &\leq \min \left\{ l \geq 1 \, : \, \sum_{j \leq l} p_j > 1 - \min(U_1 p_{c_1}, \ldots, U_n p_{c_n}) \right\} \\ &= N_n. \end{aligned}$$

The random number $M_n$ is thus a data-free lower bound for $N_n$, where the inequality $M_n(\omega) \leq N_n(\omega)$ holds for every $\omega \in \Omega$. Studying the distribution of $M_n$ will shed light on the distribution of its upper bound $N_n$. Interestingly, $M_n$ represents also the random number of jumps to be drawn in order to generate a sample of size $n$ from a PY by

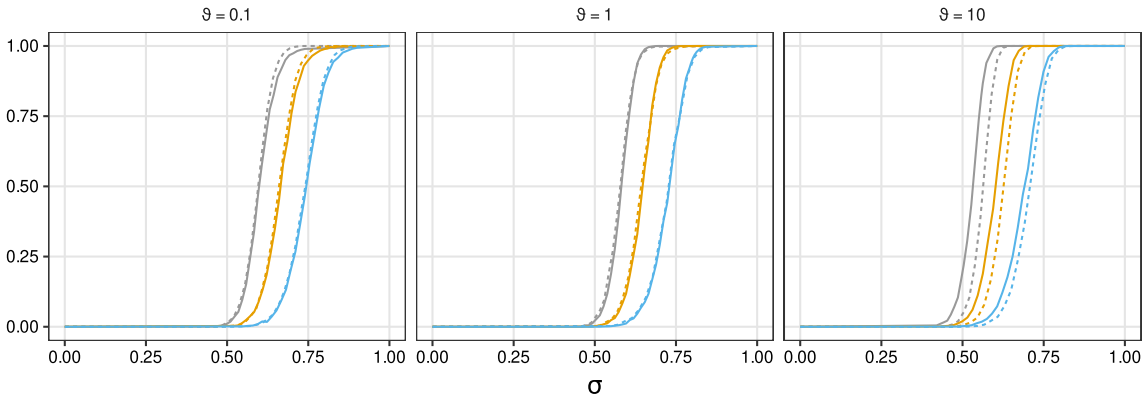


**Fig. 7** Estimated values for $P(M_n > 10^6)$ (solid curves) and $P(L_n > 10^6)$ (dashed curves) as a function of $\sigma \in (0, 1)$, for $n = 100$ (blue), $n = 1\,000$ (orange), $n = 10\,000$ (gray), and for $\theta = 0.1$ (left panel), $\theta = 1$ (middle panel), $\theta = 10$ (right panel)

adapting the retrospective sampling idea of Papaspiliopoulos and Roberts (2008), described in their Sect. 2 for the DP case. The distribution of $M_n$ coincides with the distribution of $\min\left\{l \geq 1 : \prod_{j\leq l}(1 - V_j) < B_n\right\}$, where the stick-breaking variables $\{V_j\}_{j=1}^{\infty}$ are defined as in (5) and $B_n$ is a beta random variable with parameters 1 and $n$. Following Muliere and Tardella (1998), it is easy to show that, when $\sigma = 0$, then $M_n - 1$ is distributed as a mixture of Poisson distributions, specifically $(M_n - 1) \sim \text{Poisson}(\vartheta \log(1/B_n))$. This leads to $\mathbb{E}[M_n] = \vartheta H_n + 1$, where $H_n = \sum_{l=1}^{n} l^{-1}$ is the $n$-th harmonic number. It is worth noting that, for $n \to \infty$, $\mathbb{E}[M_n] \approx \vartheta \log(n)$, that is the growth is logarithmic in $n$, while the contribution of $\vartheta$ is linear. As for the PY process, we resort to Arbel et al. (2019), where the asymptotic distribution of the minimum number of jumps of a PY, needed to guarantee that the truncation error is smaller than a deterministic threshold, is studied. We introduce the notation $a_n \overset{\text{a.s.}}{\sim} b_n$ to indicate that $P(\lim_{n\to\infty} a_n/b_n = 1) = 1$ and, by exploiting Theorem 2 in Arbel et al. (2019), we prove the following proposition.

**Proposition 2** *Let* $M_n = \min\left\{l \geq 1 : \prod_{j\leq l}(1 - V_j) < B_n\right\}$ *where the sequence* $(V_j)_{j\geq 1}$ *is defined as in* (5) *and* $B_n$ *is a beta random variable with parameters* 1 *and* $n$. *Then, for* $n \to \infty$,

$$M_n - 1 \overset{\text{a.s.}}{\sim} \left(\frac{B_n T_{\sigma,\vartheta}}{\sigma}\right)^{-\sigma/(1-\sigma)}, \tag{16}$$

*where* $T_{\sigma,\vartheta}$, *independent of* $B_n$, *is a polynomially tilted stable random variable* (Devroye 2009), *with probability density function proportional to* $t^{-\vartheta} f_\sigma(x)$, *where* $f_\sigma$ *is the density function of a unilateral stable random variable with Laplace transform equal to* $\exp\{-\lambda^\sigma\}$.

***Proof*** Define $M(\epsilon) = \min\left\{l \geq 1 : \prod_{j\leq l}(1 - V_j) < \epsilon\right\}$. Following Arbel et al. (2019),

$$M(\epsilon) - 1 \overset{\text{a.s.}}{\sim} \left(\frac{\epsilon T_{\sigma,\vartheta}}{\sigma}\right)^{-\sigma/(1-\sigma)}, \tag{17}$$

as $\epsilon \to 0$. Observe that $M_n = M(B_n)$ and that $B_n \overset{\text{a.s.}}{\sim} 0$ as $n \to \infty$. We then define the events

$$A = \left\{\omega \in \Omega : M(\epsilon) \asymp_{\text{a.s.}} \left(\epsilon T_{\sigma,\vartheta}/\sigma\right)^{-\sigma/(1-\sigma)} \text{ as } \epsilon \to 0\right\}$$
$$B = \{\omega \in \Omega : B_n \asymp_{\text{a.s.}} 0 \text{ as } n \to \infty\}$$
$$C = \left\{\omega \in \Omega : M_n \asymp_{\text{a.s.}} \left(B_n T_{\sigma,\vartheta}/\sigma\right)^{-\sigma/(1-\sigma)} \text{ as } n \to \infty\right\}$$

and observe that $C \subset A \cup B$. Which implies that $P(C) \leq P(A \cup B) \leq P(A) + P(B) = 0$. □

If we define $L_n = \left(B_n T_{\sigma,\vartheta}/\sigma\right)^{-\sigma/(1-\sigma)}$, for any positive integer $n$, the statement of Proposition 2 is tantamount to $M_n - 1 \overset{\text{a.s.}}{\sim} L_n$ as $n \to \infty$. The random variable $L_n$ has finite mean if and only if $\sigma \in (0, 1/2)$, case in which $\mathbb{E}[L_n] = c_{\sigma,\vartheta}\Gamma(n+1)/\Gamma(n+2-1/(1-\sigma))$, where

$$c_{\sigma,\vartheta} = \sigma^{\sigma/(1-\sigma)} \frac{\Gamma(2 - 1/(1-\sigma))\Gamma(1 + \vartheta/\sigma + 1/(1-\sigma))}{\Gamma(\vartheta + 1/(1-\sigma))},$$

which implies that $\mathbb{E}[L_n] \approx c_{\sigma,\theta} n^{\sigma/(1-\sigma)}$, when $n \to \infty$. A simple simulation experiment was run to empirically investigate the quality of the asymptotic approximation of $M_n$ provided by $L_n$. The random variable $T_{\sigma,\vartheta}$ appearing in the defintion of $L_n$ was sampled by resorting to Hofert (2011). Figure 7 displays the estimated probability of the events $M_n > 10^6$ and $L_n > 10^6$, as a function of $\sigma \in (0, 1)$, for $\vartheta \in \{0.1, 1, 10\}$ and for different sample sizes $n \in \{100, 1\,000, 10\,000\}$.

## Appendix B Additional details on the simulation study

This section provides additional results of the simulation study presented in Sect. 3. Table 1 reports on the number of times the upper bound for the number of jumps drawn at each iteration of dependent and independent slice-efficient samplers was reached. Figures 8, 9 and 10 focus on the functional

**Table 1** Relative frequency of the of times that the bound $10^5$ on the number of jumps is reached when $\sigma = 0.4$ for the independent slice-efficient algorithm (I-SE) and the dependent slice-efficient algorithm (D-SE)

| | | I-SE | D-SE |
|---|---|---|---|
| $\vartheta = 1$ | $n = 100$ | 0.00 | 0.00 |
| | $n = 250$ | 0.00 | 0.00 |
| | $n = 1000$ | 0.00 | 0.00 |
| $\vartheta = 10$ | $n = 100$ | 0.06 | 0.07 |
| | $n = 250$ | 0.8 | 0.14 |
| | $n = 1000$ | 0.16 | 0.31 |
| $\vartheta = 25$ | $n = 100$ | 0.30 | 0.39 |
| | $n = 250$ | 0.39 | 0.54 |
| | $n = 1000$ | 0.63 | 0.89 |

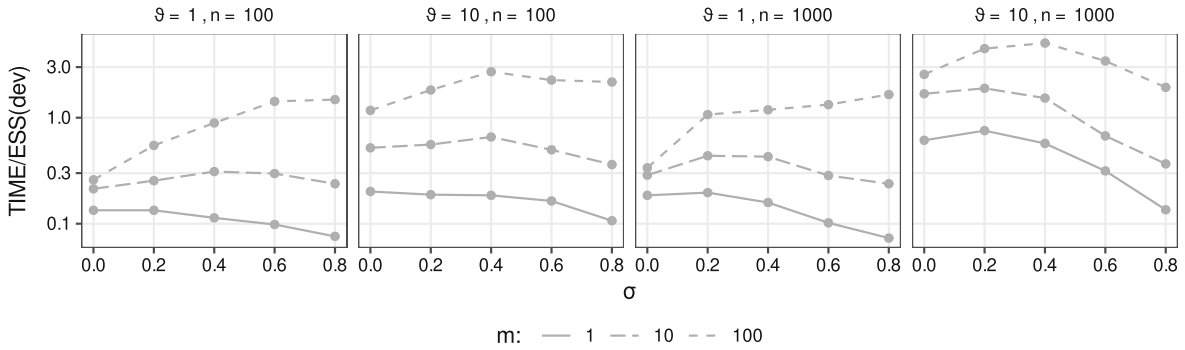


**Fig. 8** Simulated data. ICS: ratio between runtime (in seconds) and ESS computed on the deviance on a log-scale. Results are averaged over 100 replicates

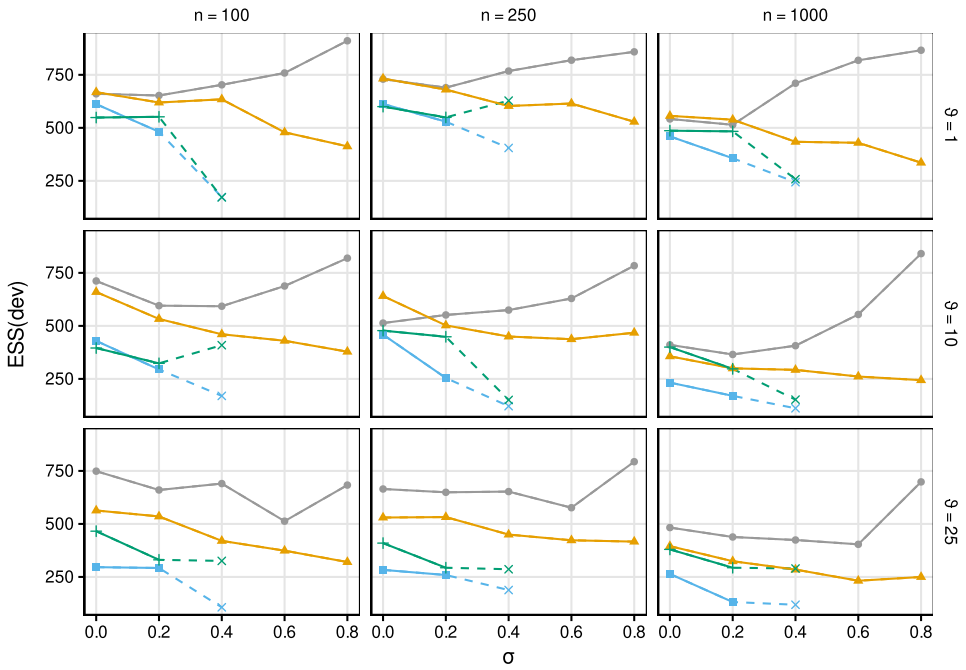


**Fig. 9** Simulated data. ESS computed on the deviance, for ICS (gray), marginal sampler (orange), independent slice-efficient sampler (green) and dependent slice-efficient sampler (blue). Results are averaged over 10 replicates. The $\times$-shaped marker for the two slice samplers indicates that, when $\sigma = 0.4$, the value of the ESS is obtained with an arbitrary upper bound at $10^5$ for the number of jumps drawn per iteration

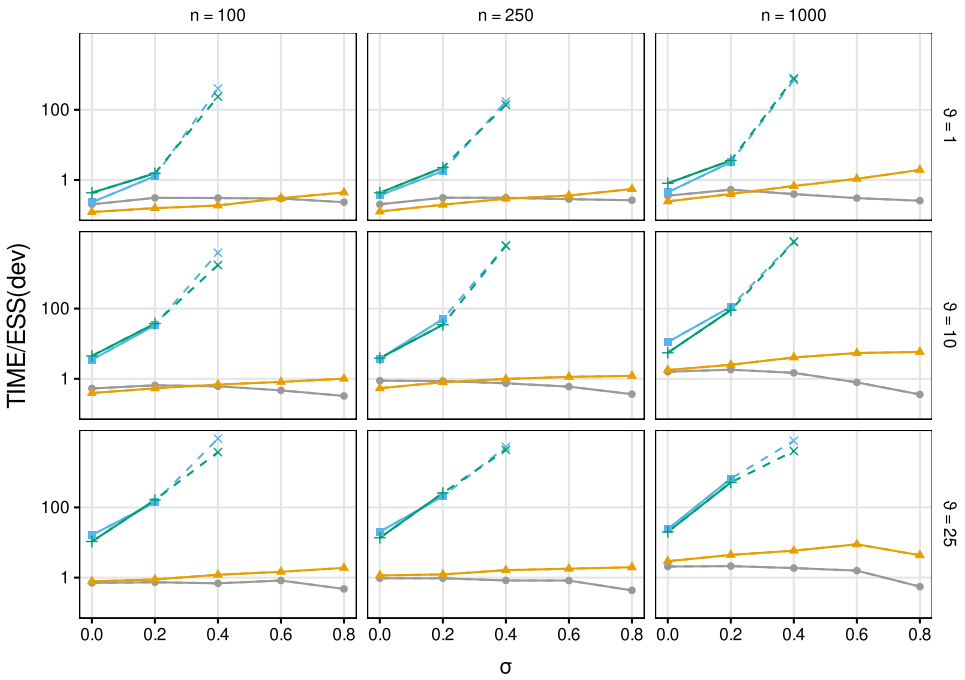


**Fig. 10** Simulated data. Ratio of runtime (in seconds) over ESS computed on the deviance, in log-scale, for ICS (gray), marginal sampler (orange), independent slice-efficient sampler (green) and dependent slice-efficient sampler (blue). Results are averaged over 10 replicates. The $\times$-shaped marker for the two slice samplers indicates that, when $\sigma = 0.4$, the value of time/ESS is obtained with an arbitrary upper bound at $10^5$ for the number of jumps drawn per iteration

deviance and display results analogous to those presented in Sect. 3 for the random variable number of clusters.

## Appendix C ICS for GM-DDP

In order to describe the full conditional distributions of $\theta_{i,l}$ and $\boldsymbol{w}$, and to provide the pseudo-code of the ICS for the GM-DDP mixture model, some notation needs to be introduced. Let $r_{m,0}$ and $r_{m,l}$, for $l = 1, \ldots, L$, represent the number of distinct values $s^*_{j,0}$ and $s^*_{j,l}$ appearing in the vectors $\boldsymbol{s}_0$ and $\boldsymbol{s}_l$, respectively. The corresponding frequencies are given by $m_{j,0}$ and $m_{j,l}$, and are such that $\sum_{j=1}^{k_{m,l}} m_{j,l} = m$, for every $l = 0, 1, \ldots, L$. Let $\boldsymbol{\theta}^*_0$ be the vector of distinct values appearing in $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L)$ coinciding with either the $k_{\mathbf{n},0}$ fixed jump points $\boldsymbol{t}_0$ of the common process $\gamma_0$ or with any of the $r_{m,0}$ values appearing in $\boldsymbol{s}_0$. Similarly, for any $l = 1, \ldots, L$, $\boldsymbol{\theta}^*_l$ denotes the vector of distinct values appearing in $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L)$ coinciding with either the $k_{\mathbf{n},l}$ fixed jump points $\boldsymbol{t}_l$ of the idiosyncratic process $\gamma_l$ or with any of the $r_{m,l}$ values appearing in $\boldsymbol{s}_l$. Finally, we let $\mathcal{C}_{j,0} = \{(i,l) \ : \ \theta_{i,l} = \theta^*_{j,0}\}$ and, for $l = 1, \ldots, L$, $\mathcal{C}_{j,l} = \{i \ : \ \theta_{i,l} = \theta^*_{j,l}\}$.

The full conditional distribution of $\theta_{i,l}$, for every $l = 1, \ldots, L$ and $1 \leq i \leq n_l$, is given, up to a proportionality constant, by

$$\begin{aligned}
\mathrm{P}(\theta_{i,l} \in \mathrm{d}t | \ldots) \propto w_l \Bigg( & p_{0,l} \sum_{j=1}^{r_{m,l}} \frac{m_{j,l}}{m} \mathcal{K}(X_{i,l}, s^*_{j,l}) \delta_{s^*_{j,l}}(\mathrm{d}t) \\
& + \sum_{j=1}^{k_{\boldsymbol{n},l}} p_{j,l} \mathcal{K}(X_{i,l}, t^*_{j,l}) \delta_{t^*_{j,l}}(\mathrm{d}t) \Bigg) \\
& + (1 - w_l) \Bigg( p_{0,0} \sum_{j=1}^{r_{m,0}} \frac{m_{j,0}}{m} \mathcal{K}(X_{i,l}, s^*_{j,0}) \delta_{s^*_{j,0}}(\mathrm{d}t) \\
& + \sum_{j=1}^{k_{\boldsymbol{n},0}} p_{j,0} \mathcal{K}(X_{i,l}, t^*_{j,0}) \delta_{t^*_{j,0}}(\mathrm{d}t) \Bigg).
\end{aligned}$$

The full conditional for $\boldsymbol{w}$ is given, up to a proportionality constant, by

$$\begin{aligned}
& \mathrm{P}(\boldsymbol{w} = (v_1, \ldots, v_L) | \ldots) \\
& \quad \propto \prod_{l=1}^{L} \frac{v_l^{\vartheta z - 1}}{(1 - v_l)^{\vartheta z + 1}} \prod_{i=1}^{n_l} \left( v_l q^{(l)}_{i,l} + (1 - v_l) q^{(0)}_{i,l} \right) \\
& \quad \times \left( 1 + \sum_{l=1}^{L} \frac{v_l}{1 - v_l} \right)^{-L\vartheta z - \theta(1-z)}
\end{aligned} \tag{18}$$

**Algorithm 2:** ICS for GM-DDP mixture model

**set** *admissible initial values for* $\boldsymbol{\theta}_l^{(0)}$, *for* $l = 1, \ldots, L$;
**for** *each iteration* $r = 1, \ldots, R$ **do**
  **set** $\boldsymbol{t}_0^{(r)} = \boldsymbol{\theta}_0^{*(r-1)}$;
  **sample** $\boldsymbol{p}_0^{(r)}$ *from* $\boldsymbol{p}_0^{(r)} \sim \text{Dirichlet}(c(1-z), n_{1,0}^{(r-1)}, \ldots, n_{k_{\mathbf{n},0},0}^{(r-1)})$;
  **sample** $\boldsymbol{s}_0^{(r)}$ *from a* $DP(c(1-z); P_0)$;
  **for** *each urn* $l = 1, \ldots, L$ **do**
    **set** $\boldsymbol{t}_l^{(r)} = \boldsymbol{\theta}_l^{*(r-1)}$;
    **sample** $\boldsymbol{p}_l^{(r)}$ *from* $\boldsymbol{p}_l^{(r)} \sim \text{Dirichlet}(cz, n_{1,l}^{(r-1)}, \ldots, n_{k_{\mathbf{n},l},l}^{(r-1)})$;
    **sample** $\boldsymbol{s}_l^{(r)}$ *from a* $DP(cz; P_0)$;
  **sample** $\boldsymbol{w}^{(r)}$ *from (18)*;
  **for** *each* $i = 1, \ldots, n_l;\ l = 1, \ldots, L$ **do**
    **sample** $\theta_{i,l}^{(r)}$ *from*

$$\mathrm{P}(\theta_{i,l}^{(r)} = t | \cdots) \propto \begin{cases} w_l p_{0,l}^{(r)} \frac{m_{j,l}^{(r)}}{m} k(X_{i,l}; s_{j,l}^{*(r)}) & \text{if } t \in \{s_{1,l}^{*(r)}, \ldots, s_{r_{m,l},l}^{*(r)}\} \\ w_l p_{j,l}^{(r)} k(X_{i,l}; t_{j,l}^{*(r)}) & \text{if } t \in \{t_{1,l}^{*(r)}, \ldots, t_{k_{\mathbf{n},l}^{(r-1)},l}^{*(r)}\} \\ (1-w_l) p_{0,0}^{(r)} \frac{m_{j,0}^{(r)}}{m} k(X_{i,l}; s_{j,0}^{(r)}) & \text{if } t \in \{s_{1,0}^{*(r)}, \ldots, s_{r_{m,0}^{(r)},0}^{*(r)}\} \\ (1-w_l) p_{j,0}^{(r)} k(X_{i,l}; t_{j,0}^{*(r)}) & \text{if } t \in \{t_{1,0}^{*(r)}, \ldots, t_{k_{\mathbf{n},0}^{(r-1)},0}^{*(r)}\} \\ 0 & \text{otherwise} \end{cases}$$

  **for** *each element* $\theta_{j,0}^{*(r)}$ *in* $\boldsymbol{\theta}_0^{*(r)}$ **do**
    **let** $\mathcal{C}_{j,0}^{(r)}$ *be the set of pairs* $(i,l)$ *such that* $\theta_{i,l}^{(r)} = \theta_{j,0}^{*(r)}$;
    **update** $\theta_{j,0}^{*(r)}$ *from*

$$\mathrm{P}(\theta_{j,0}^{*(r)} \in \mathrm{d}t | \cdots) \propto P_0(\mathrm{d}t) \prod_{(i,l) \in \mathcal{C}_{j,0}^{(r)}} \mathcal{K}(X_{i,l}; t);$$

  **for** *each element* $\theta_{j,l}^{*(r)}$ *in* $\boldsymbol{\theta}_l^{*(r)}$, $l = 1, \ldots, L$ **do**
    **let** $\mathcal{C}_{j,l}^{(r)}$ *be the set of index pairs indexes* $i$ *such that* $\theta_{i,l}^{(r)} = \theta_{j,l}^{*(r)}$;
    **update** $\theta_{j,l}^{*(r)}$ *from*

$$\mathrm{P}(\theta_{j,l}^{*(r)} \in \mathrm{d}t | \cdots) \propto P_0(\mathrm{d}t) \prod_{i \in \mathcal{C}_{j,l}^{(r)}} \mathcal{K}(X_{i,l}; t);$$

**end**

where

$$q_{i,l}^{(l)} = p_{0,l} \sum_{j=1}^{r_{m,l}} \frac{m_{j,l}}{m} \mathcal{K}(X_{i,l}, s_{j,l}^*) \delta_{s_{j,l}^*}(\mathrm{d}t) + \sum_{j=1}^{k_{\boldsymbol{n},l}} p_{j,l} \mathcal{K}(X_{i,l}, t_{j,l}^*) \delta_{t_{j,l}^*}(\mathrm{d}t),$$

$$q_{i,l}^{(0)} = p_{0,0} \sum_{j=1}^{r_{m,0}} \frac{m_{j,0}}{m} \mathcal{K}(X_{i,l}, s_{j,0}^*) \delta_{s_{j,0}^*}(\mathrm{d}t) + \sum_{j=1}^{k_{\boldsymbol{n},0}} p_{j,0} \mathcal{K}(X_{i,l}, t_{j,0}^*) \delta_{t_{j,0}^*}(\mathrm{d}t).$$

The pseudo-code of the ICS for the GM-DDP mixture model is presented in Algorithm 2.

## References


Arbel, J., De Blasi, P., Prünster, I.: Stochastic approximations to the Pitman–Yor process. Bayesian Anal. **14**(4), 1201–1219 (2019)

Arbel, J., Lijoi, A., Nipoti, B.: Full Bayesian inference with hazard mixture models. Comput. Stat. Data Anal. **93**, 359–372 (2016)

Argiento, R., Bianchini, I., Guglielmi, A.: Posterior sampling from $\varepsilon$-approximation of normalized completely random measure mixtures. Electron. J. Stat. **10**(2), 3516–3547 (2016)

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

Blackwell, D., MacQueen, J.B.: Ferguson distributions via Polya urn schemes. Ann. Stat. **1**(2), 353–355 (1973)

Canale, A., Prünster, I.: Robustifying Bayesian nonparametric mixtures for count data. Biometrics **73**(1), 174–184 (2017)

Corradin, R., Canale, A., Nipoti, B.: BNPmix: an R package for Bayesian nonparametric modelling via Pitman–Yor mixtures. J. Stat. Softw. **100**(15), 1–33 (2021)

De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I., Ruggiero, M.: Are Gibbs-type priors the most natural generalization of the Dirichlet process? IEEE Trans. Pattern Anal. Mach. Intell. **37**(2), 212–229 (2015)

de Finetti, B.: Sur la condition d'equivalence partielle. Actualités Sci. Ind. **739**, 5–18 (1938)

Devroye, L.: Random variate generation for exponentially and polynomially tilted stable distributions. ACM Trans. Model. Comput. Simul. **19**(4), 18:1-18:20 (2009)

Dubey, K.A., Zhang, M., Xing, E., Williamson, S.: Distributed, partially collapsed MCMC for Bayesian nonparametrics. In: Proceedings of Machine Learning Research, vol. 108, pp. 3685–3695 (2020)

Escobar, M.D.: Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. PhD thesis, Department of Statistics, Yale University (1988)

Escobar, M.D., West, M.: Bayesian density estimation and inference using mixtures. J. Am. Stat. Assoc. **90**(430), 577–588 (1995)

Fall, M.D., Barat, E.: Gibbs sampling methods for Pitman–Yor mixture models. hal-00740770v2 (2014)

Favaro, S., Teh, Y.W.: MCMC for normalized random measure mixture models. Statist. Sci. **28**(3), 335–359 (2013)

Favaro, S., Walker, S.G.: Slice sampling $\sigma$-stable Poisson–Kingman mixture models. J. Comput. Graph. Stat. **22**(4), 830–847 (2013)

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

Foti, N., Williamson, S.: A survey of non-exchangeable priors for Bayesian nonparametric models. IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015)

Frühwirth-Schnatter, S., Celeux, G., Robert, C.P.: Handbook of Mixture Analysis. Chapman and Hall/CRC (2019)

Ge, H., Chen, Y., Wan, M., Ghahramani, Z.: Distributed inference for Dirichlet process mixture models. In: Proceedings of Machine Learning Research, vol. 37, pp. 2276–2284 (2015)

Gelfand, A.E., Kottas, A.: A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. J. Comput. Graph. Stat. **11**(2), 289–305 (2002)

Geweke, J.: Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics, vol. 4. Oxford University Press, Oxford (1992)

Ghosal, S.: The Dirichlet Process, Related Priors and Posterior Asymptotics, pp. 35–79. Cambridge University Press (2010)

Griffin, J.E., Kolossiatis, M., Steel, M.F.: Comparing distributions by using dependent normalized random-measure mixtures. J. Roy. Stat. Soc. Ser. B (Statistical Methodology) **75**(3), 499–529 (2013)

Hofert, M.: Efficiently sampling nested Archimedean copulas. Comput. Stat. Data Anal. **55**(1), 57–70 (2011)

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

Ishwaran, H., Zarepour, M.: Exact and approximate sum representations for the Dirichlet process. Can. J. Stat. **30**(2), 269–283 (2002)

Jara, A., Hanson, T., Quintana, F., Müller, P., Rosner, G.: DPpackage: Bayesian semi- and nonparametric modeling in R. J. Stat. Softw. **40**(5), 1–30 (2011)

Kalli, M., Griffin, J.E., Walker, S.G.: Slice sampling mixture models. Stat. Comput. **21**(1), 93–105 (2011)

Klebanoff, M.A.: The collaborative perinatal project: a 50-year retrospective. Paediatr. Perinat. Epidemiol. **23**(1), 2 (2009)

Kriegel, H.-P., Schubert, E., Zimek, A.: The (black) art of runtime evaluation: are we comparing algorithms or implementations? Knowl. Inf. Syst. **52**(2), 341–378 (2017)

Lijoi, A., Mena, R.H., Prünster, I.: Bayesian nonparametric analysis for a generalized Dirichlet process prior. Stat. Infer. Stoch. Process. **8**(3), 283–309 (2005)

Lijoi, A., Mena, R.H., Prünster, I.: Hierarchical mixture modeling with normalized inverse-Gaussian priors. J. Am. Stat. Assoc. **100**(472), 1278–1291 (2005)

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

Lijoi, A., Nipoti, B., Prünster, I.: Bayesian inference with dependent normalized completely random measures. Bernoulli **20**(3), 1260–1291 (2014)

Lijoi, A., Nipoti, B., Prünster, I.: Dependent mixture models: clustering and borrowing information. Comput. Stat. Data Anal. **71**, 417–433 (2014)

Lo, A.Y.: On a class of Bayesian nonparametric estimates: I. Density estimates. Ann. Stat. **12**(1), 351–357 (1984)

Lomelí, M., Favaro, S., Teh, Y.W.: A hybrid sampler for Poisson–Kingman mixture models. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 28. Curran Associates Inc (2015)

Lomelí, M., Favaro, S., Teh, Y.W.: A marginal sampler for $\sigma$-stable Poisson–Kingman mixture models. J. Comput. Graph. Stat. **26**(1), 44–53 (2017)

Longnecker, M.P., Klebanoff, M.A., Zhou, H., Brock, J.W.: Association between maternal serum concentration of the ddt metabolite dde and preterm and small-for-gestational-age babies at birth. Lancet **358**(9276), 110–114 (2001)

MacEachern, S.N.: Estimating normal means with a conjugate style Dirichlet process prior. Commun. Stat. Simul. Comput. **23**(3), 727–741 (1994)

MacEachern, S.N., Müller, P.: Estimating mixture of Dirichlet process models. J. Comput. Graph. Stat. **7**(2), 223–238 (1998)

Muliere, P., Tardella, L.: Approximating distributions of random functionals of Ferguson–Dirichlet priors. Can. J. Stat. **26**(2), 283–297 (1998)

Müller, P., Erkanli, A., West, M.: Bayesian curve fitting using multivariate normal mixtures. Biometrika **83**(1), 67–79 (1996)

Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. **9**(2), 249–265 (2000)

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

Olkin, I., Liu, R.: A bivariate beta distribution. Stat. Probab. Lett. **62**(4), 407–412 (2003)

Papaspiliopoulos, O., Roberts, G.O.: Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. Biometrika **95**(1), 169–186 (2008)

Perman, M., Pitman, J., Yor, M.: Size-biased sampling of Poisson point processes and excursions. Probab. Theory Relat. Fields **92**(1), 21–39 (1992)

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

Pitman, J.: Some developments of the Blackwell–MacQueen urn scheme. Lect. Notes Monogr. Ser. **30**, 245–267 (1996)

Pitman, J., Yor, M.: The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. Ann. Probab. **25**(2), 855–900 (1997)

Plummer, M., Best, N., Cowles, K., Vines, K.: CODA: convergence diagnosis and output analysis for MCMC. R News **6**(1), 7–11 (2006)

Smith, A.F., Gelfand, A.E.: Bayesian statistics without tears: a sampling-resampling perspective. Am. Stat. **46**(2), 84–88 (1992)

Taddy, M.A., Kottas, A.: Mixture modeling for marked Poisson processes. Bayesian Anal. **7**(2), 335–362 (2012)

Walker, S.G.: Sampling the Dirichlet mixture model with slices. Commun. Stat. Simul. Comput. **36**(1), 45–54 (2007)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.