ISSN 2279-9362

Collegio Carlo Alberto

The Pitman-Yor multinomial process for mixture modeling

Antonio Lijoi, Igor Pruenster e Tommaso Rigon

No. 599
December 2019

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

The Pitman–Yor multinomial process for mixture modeling

Antonio Lijoi, Igor Prünster and Tommaso Rigon

Abstract

Discrete nonparametric priors play a central role in a variety of Bayesian procedures, most notably when used to model latent features as in clustering, mixtures and curve fitting. They are effective and well developed tools, though their infinite-dimensionality is unsuited to several applications. If one confines oneself to a finite-dimensional simplex, there are few nonparametric priors beyond the traditional Dirichlet-multinomial process, which is mainly motivated by conjugacy. Here we introduce an alternative based on the Pitman–Yor process, which ensures greater flexibility while preserving analytical tractability. Urn schemes and posterior characterizations are obtained in closed form, leading to exact sampling methods. In addition, our proposal can be used to accurately approximate the infinite-dimensional Pitman–Yor process, improving over existing truncation-based approaches. An application to convex mixture regression for quantitative risk assessment serves as an illustration of our results and allows comparisons with existing methodologies.

1 Introduction

The statistical investigation of discrete random structures has been a very lively area of research in recent years. Parametric and nonparametric Dirichlet priors have found wide applicability in numerous settings that include, among others, density estimation, model-based clustering, density regression, functional data analysis, and hidden Markov models. However, Dirichlet-based priors might be a restrictive modeling choice, for example, because the induced random partition is controlled by a single parameter. In the nonparametric case, the limitations of the Dirichlet process are successfully circumvented, for instance, by considering the more flexible Pitman–Yor process (Pitman and Yor, 1997) or a number of other alternatives broadly summarized in Lijoi and Prünster (2010). These remarkable advances in the nonparametric literature have not been paralleled by a similar wealth of proposals in the finite-dimensional setting, where only few alternatives are known beyond Dirichlet-like structures and most of them outside the Bayesian realm. For example, there is an interesting piece of literature on compositional data spurred by the pioneering work in Aitchison (1985) on general classes of distributions on the simplex. However, to the best of our knowledge, none of them has been used

as a prior distribution for modeling either the data or some latent feature. Indeed, the lack of deep theoretical results has prevented the development of more flexible classes of priors, as well as simple sampling algorithms that may facilitate their implementation. We fill this gap by introducing a novel random probability measure, the Pitman–Yor multinomial process. We show that this prior may be seen as a finite-dimensional analogue of a Pitman–Yor process with discount parameter $\sigma \in [0, 1)$ and naturally generalizes the Dirichlet multinomial process, a finite-dimensional prior which may be defined as a Pitman–Yor process with negative discount parameter. Besides gaining modeling flexibility, our proposal preserves analytical and computational tractability.

The Pitman–Yor multinomial process further serves as a very effective tool for computational purposes in a nonparametric setting. A popular class of Markov chain Monte Carlo algorithms for nonparametric mixture models, usually referred to as the blocked Gibbs sampler, relies on the truncation of a stick-breaking representation of the mixing Pitman–Yor process. If $(\nu_j)_{j\geq 1}$ is a sequence of independent random variables with $\nu_j \sim \text{Beta}(1-\sigma, \alpha+j\sigma)$, where $\sigma \in [0, 1)$ and $\alpha > -\sigma$, and P is a probability measure defined over Θ, then a prior on the space of density functions is the distribution of

$$\tilde{f}(y) = \int_{\Theta} \mathcal{K}(y;\theta)\, \tilde{p}_{\infty}(\mathrm{d}\theta), \qquad \tilde{p}_{\infty} = \sum_{h=1}^{\infty} \xi_h\, \delta_{\tilde{\phi}_h}, \qquad \tilde{\phi}_h \overset{\text{iid}}{\sim} P, \tag{1}$$

where $\xi_1 = \nu_1$, $\xi_h = \nu_h \prod_{j=1}^{h-1}(1-\nu_j)$ for $h \geq 2$ with $\mathcal{K}$ a transition kernel such that $\int_{\mathbb{R}} \mathcal{K}(y;\theta)\, \mathrm{d}y = 1$ for any $\theta \in \Theta$. When it comes to evaluating Bayesian inferences, the infinite series defining $\tilde{p}_{\infty}$ in (1) cannot be computed and one conveniently relies on a suitable finite-dimensional approximation obtained by truncating $\tilde{p}_{\infty}$ at some level H, i.e.

$$\tilde{p}_{H,\text{tr}} = \sum_{h=1}^{H-1} \xi_h\, \delta_{\tilde{\phi}_h} + (1 - |\xi^{(H-1)}|)\, \delta_{\tilde{\phi}_H}, \qquad |\xi^{(H-1)}| = \xi_1 + \cdots + \xi_{H-1}. \tag{2}$$

This approach has some limitations since one can hardly identify marginal probabilistic structures of interest such as, for example, the law of the induced exchangeable random partition, the probability distribution of the number of clusters or the prediction rule associated to (2). The results that are displayed in the next sections successfully address the above issues by replacing $\tilde{p}_{H,\text{tr}}$ with the Pitman–Yor multinomial process, which stands as an alternative finite-dimensional approximation of $\tilde{p}_{\infty}$ with H components. We show that for non-informative specifications of $\tilde{p}_{\infty}$, namely those corresponding to values of $\sigma > 1/2$, the Pitman–Yor multinomial process is a more accurate approximation of $\tilde{p}_{\infty}$ compared to $\tilde{p}_{H,\text{tr}}$. In addition, the distributional results we achieve allow for a straightforward implementation of novel generalised Blackwell–MacQueen sampling schemes as well as conditional algorithms.

On top of its computational relevance in Bayesian nonparametrics, the Pitman–Yor multinomial process has important applications in finite mixture modeling. As for the Dirichlet multinomial special

case, the value H represents a conservative upper bound for the number of mixture components. Such an approach has been motivated by asymptotic arguments in Rousseau and Mengersen (2011) and it has been supported by Malsiner-Walli et al. (2016) on the basis of more empirical arguments. As an alternative, following for example the works of Richardson and Green (1997); Gnedin and Pitman (2005) or, more recently, Miller and Harrison (2018); Argiento and De Iorio (2019), one may assume a hyperprior on H. Hence, the Pitman–Yor multinomial process stands as a natural generalisation of these methods and it translates the advantages of the Pitman–Yor process into finite-dimensional settings allowing a much finer control of the underlying random partition and a more robust specification of the cluster distribution, which is typically very informative in the Dirichlet setting (Lijoi et al., 2007).

Our proposal, and the related distributional results, are illustrated through a covariate-dependent mixture model that is reminiscent of the one proposed in Canale et al. (2018), who consider a toxicological study originally conducted by Longnecker et al. (2001).

2 The Pitman–Yor multinomial process

The Pitman–Yor multinomial process is built upon the Pitman–Yor process $\tilde{p}_\infty$ (Perman et al., 1992; Pitman and Yor, 1997) in equation (1). The probability distribution P of the atoms $\tilde{\phi}_h$ is often referred to as the baseline measure because $E(\tilde{p}_\infty) = P$. We will henceforth use the notation $\tilde{p}_\infty \sim \text{PY}(\sigma, \alpha; P)$ with P typically chosen to be diffuse, i.e. $P(\{\theta\}) = 0$ for any $\theta \in \Theta$. The Pitman–Yor multinomial process corresponds to the case where P is replaced by some discrete random probability measure with finitely many support points, as in the following definition.

Definition 1. A discrete random probability measure $\tilde{p}_H$ is a Pitman–Yor multinomial process with parameters $\sigma \in [0, 1)$ and $\alpha > -\sigma$ if it admits the hierarchical representation

$$\tilde{p}_H \mid \tilde{p}_{0,H} \sim \text{PY}(\sigma, \alpha; \tilde{p}_{0,H}), \qquad \tilde{p}_{0,H} = \frac{1}{H} \sum_{h=1}^{H} \delta_{\tilde{\theta}_h}, \qquad H \geq 1, \tag{3}$$

where $\tilde{\theta}_h$ are independent and identically distributed Θ-valued random variables with common distribution P. We will write $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$.

When $\sigma = 0$, the random probability measure $\tilde{p}_H$ in Definition 1 reduces to the Dirichlet multinomial process, which indeed admits such a hierarchical representation (Ishwaran and Zarepour, 2000), although it may also be regarded as a $\text{PY}(-\alpha/H, \alpha; P)$, where $\sigma = -\alpha/H < 0$. Though $\tilde{p}_H$ is finite-dimensional, one can give an alternative and equivalent definition in terms of the infinite-dimensional counterpart $\tilde{p}_\infty \sim \text{PY}(\sigma, \alpha; P)$. Indeed, for any finite and measurable partition $B_1, \ldots, B_d$ of Θ the vector $\{\tilde{p}_\infty(B_1), \ldots, \tilde{p}_\infty(B_{d-1})\}$ identifies a probability distribution on the simplex known as ratio-stable

(Carlton, 2002), so that

$$\{\tilde{p}_\infty(B_1), \ldots, \tilde{p}_\infty(B_{d-1})\} \sim \text{RS}\{\sigma, \alpha; P(B_1), \ldots, P(B_d)\},$$

where $\tilde{p}(B_i) = 0$ almost surely if $P(B_i) = 0$, for any $i = 1, \ldots, d$. Moment formulae for ratio-stable laws can be found in Carlton (2002). Unsurprisingly, the weights of a Pitman–Yor multinomial process follow a ratio-stable distribution, as summarized in the following proposition, whose proof is straightforward.

Proposition 1. *A Pitman–Yor multinomial process* $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$ *with parameters* $\sigma \in [0, 1)$ *and* $\alpha > -\sigma$ *admits the following marginal representation in distribution*

$$\tilde{p}_H = \sum_{h=1}^{H} \pi_h \delta_{\tilde{\theta}_h}, \qquad (\pi_1, \ldots, \pi_{H-1}) \sim \text{RS}(\sigma, \alpha; 1/H, \ldots, 1/H), \qquad \tilde{\theta}_h \overset{\text{iid}}{\sim} P. \tag{4}$$

The density function of the weights $(\pi_1, \ldots, \pi_{H-1})$ is generally not available in closed form, with some notable exceptions. The Dirichlet multinomial process is identified by $\sigma = 0$ and the weights have a symmetric Dirichlet distribution with parameters $(\alpha/H, \ldots, \alpha/H)$. When $\sigma = 1/2$ the density function is available in closed form, and it was firstly obtained by Carlton (2002). The lack of a closed form expression of the density function is not a concern for Bayesian inference, since a ratio-stable distribution can be sampled for any admissible value of σ and α both a priori and a posteriori, as detailed henceforth. The proposed algorithm will arise from a hierarchical representation of ratio-stable distributions in terms of tempered-stable and gamma random variables, which can be easily simulated. To this end, we briefly recall that a positive random variable J is *tempered-stable* if for some $c > 0$, $\sigma \in (0, 1)$ and $\kappa \geq 0$, its Laplace transform is $E(\mathrm{e}^{-\lambda J}) = \exp\left[-c\{(\lambda + \kappa)^\sigma - \kappa^\sigma\}\right]$, with $\lambda > 0$, and we shall use the notation $J \sim \text{TS}(c, \sigma, \kappa)$. Note that such a random variable can be efficiently sampled, for instance by means of the algorithm of Ridout (2009). When $\sigma = 1/2$, the random variable $J \sim \text{TS}(1/H, 1/2, \kappa)$ has inverse Gaussian distribution, while setting $\kappa = 0$ leads to the positive-stable distribution. The main result we will rely on for computational purposes is the following, that may also be obtained from Proposition 21 in Pitman and Yor (1997).

Proposition 2. *Let* $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$ *with* $\sigma \in (0, 1)$ *and* $\alpha \geq 0$. *Then the weights of* $\tilde{p}_H$ *in* (4) *admit the representation* $(\pi_1, \ldots, \pi_H) = (J_1/\sum_{h=1}^H J_h, \ldots, J_H/\sum_{h=1}^H J_h)$ *and*

$$J_h \mid U \overset{\text{iid}}{\sim} \text{TS}(1/H, \sigma, U), \qquad U^\sigma \sim \text{Ga}(\alpha/\sigma, 1),$$

where we agree that $U = 0$ *almost surely if* $\alpha = 0$.

Although the above hierarchical representation holds only for $\alpha \geq 0$, one can make it fully general through the following argument. For any α one can conveniently represent the distribution of weights

in Proposition 1 as $(\pi_1, \ldots, \pi_H) = W(G_1, \ldots, G_H) + (1 - W)(\pi_1^*, \ldots, \pi_H^*)$, where the random variable $W \sim \text{BETA}(1 - \sigma, \alpha + \sigma)$, and the random vectors $(G_1, \ldots, G_H) \sim \text{MULTINOM}(1/H, \ldots, 1/H)$ and $(\pi_1^*, \ldots, \pi_{H-1}^*) \sim \text{RS}(\sigma, \alpha + \sigma; 1/H, \ldots, 1/H)$ are mutually independent. See Carlton (2002). Since $\alpha + \sigma$ is positive, the simulation of $(\pi_1^*, \ldots, \pi_H^*)$ can be addressed by means of Proposition 2 and this allows sampling of $(\pi_1, \ldots, \pi_H)$ for any $\alpha > -\sigma$.

Remark 1. The simulation of each J_h in Proposition 2 might be the source of numerical issues, for values of σ close to 0. This occurs because the distribution of U places mass on very large values that might cause overflows. However, such a problem can be circumvented by considering the rescaled random variables $\tilde{J}_h = J_h / \{(\sigma/\alpha)^{1/\sigma}\}$ whose distribution is $\tilde{J}_h \mid \tilde{U} \sim \text{TS}\{\alpha/(\sigma H), \sigma, \tilde{U}\}$ with $\tilde{U}^\sigma \sim \text{Ga}(\alpha/\sigma, \alpha/\sigma)$. This leads to more stable algorithms because $E(\tilde{U}^\sigma) = 1$. The rescaling constant cancels in the normalization and therefore one has equivalently that $(\pi_1, \ldots, \pi_H) = (\tilde{J}_1 / \sum_{h=1}^H \tilde{J}_h, \ldots, \tilde{J}_H / \sum_{h=1}^H \tilde{J}_h)$.

The hierarchical representation of Proposition 2 is a useful practical tool for simulating ratio-stable random vectors. However, a further and extremely useful characterisation of the random variables $J_1, \ldots, J_H$ is available as their law can be obtained via polynomial tilting of a collection of independent and identically distributed positive-stable random variables. Such a construction is reminiscent of the change of measure formula given in Pitman and Yor (1997), for the infinite-dimensional case. The connection is of great theoretical importance for the derivation of posterior characterisations, as detailed in the Supplementary Material.

Proposition 3. *Let $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$ with $\sigma \in (0, 1)$ and $\alpha > -\sigma$. Then the vector of jumps $(J_1, \ldots, J_H)$ identifying the weights of $\tilde{p}_H$ in Proposition 2 is such that*

$$E\left\{\exp\left(-\sum_{h=1}^H \lambda_h J_h\right)\right\} = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha/\sigma+1)} E\left\{\left(\sum_{h=1}^H J_h^{(\sigma)}\right)^{-\alpha} \exp\left(-\sum_{h=1}^H \lambda_h J_h^{(\sigma)}\right)\right\},$$

for any $\lambda_1, \ldots, \lambda_H > 0$, where $J_h^{(\sigma)} \overset{\text{iid}}{\sim} \text{TS}(1/H, \sigma, 0)$.

3 Distributional properties

Due to its discreteness, $\tilde{p}_H$ defines an attractive prior if one is interested in investigating the clustering structure featured by the data, with a finite maximum number H of groups in the population being allowed. In other terms, if $\theta_1, \ldots, \theta_n \mid \tilde{p}_H \overset{\text{iid}}{\sim} \tilde{p}_H$, with $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$, the n–sample $\theta^{(n)} = (\theta_1, \ldots, \theta_n)$ will display ties with positive probability. Let $K_{n,H} = k \leq \min\{n, H\}$ be the number of distinct values, say $\theta_1^*, \ldots, \theta_k^*$ in $\theta^{(n)}$ and $n_1, \ldots, n_k$ their respective frequencies, so that $\sum_{j=1}^k n_j = n$. This is reflected in a random partition $\Psi_{n,H}$ of $[n] = \{1, \ldots, n\}$ into k sets $C_1, \ldots, C_k$ such that i and j are in the same set if and only if $\theta_i = \theta_j$. The probability distribution of $\Psi_{n,H}$ is known as *exchangeable*

partition probability function and is defined by

$$\Pi_H(n_1,\dots,n_k) = \mathrm{pr}(\Psi_{n,H} = \{C_1,\dots,C_k\}) = \sum_{i_1\neq\cdots\neq i_k} E\left(\prod_{j=1}^{k}\pi_{i_j}^{n_j}\right),$$

where $n_j = \mathrm{card}(C_j)$ and the sum runs over all the positive and distinct integers $(i_1,\dots,i_k)$ in $\{1,\dots,H\}$. As discussed in Pitman (1996), when P is diffuse Π_H characterises the underlying random probability measure and yields, as a by-product, the system of predictive distributions. Such a construction clearly parallels the well-known infinite-dimensional case, namely when $\phi_1,\dots,\phi_n \mid \tilde{p}_\infty \overset{\text{iid}}{\sim} \tilde{p}_\infty$, with $\tilde{p}_\infty \sim \mathrm{PY}(\sigma,\alpha;P)$ and P is diffuse. The corresponding exchangeable partition probability function is $\Pi_\infty(n_1,\dots,n_k) = \{\prod_{j=1}^{k-1}(\alpha+j\sigma)\}\{\prod_{j=1}^{k}(1-\sigma)_{n_j-1}\}/(\alpha+1)_{n-1}$, where $(a)_n = a(a+1)\cdots(a+n-1)$ for any real a and integer $n \geq 1$ is the ascending factorial, with $(a)_0 = 1$. Moreover, if one conditions on $\phi^{(n)} = (\phi_1,\dots,\phi_n)$ featuring k distinct values $\phi_1^*,\dots,\phi_k^*$, the predictive distribution is $\mathrm{pr}(\phi_{n+1}\in A \mid \phi^{(n)}) = \{(\alpha+k\sigma)/(\alpha+n)\}P(A) + \sum_{j=1}^{k}\{(n_j-\sigma)/(\alpha+n)\}\delta_{\phi_j^*}(A)$. The next theorem provides the finite-dimensional counterpart to Π_∞ and is expressed in terms of the generalized factorial coefficients $\mathscr{C}(n,k;\sigma) = (k!)^{-1}\sum_{j=0}^{k}(-1)^j k!\{j!(k-j)!\}^{-1}(-j\sigma)_n$. Henceforth, we shall further assume that P is diffuse and $\sigma\in(0,1)$, so that the well-known Dirichlet multinomial case might be obtained by taking the limit as $\sigma\to 0$.

Theorem 1. *The exchangeable partition probability function induced by a Pitman–Yor multinomial process* $\tilde{p}_H \sim \mathrm{PYM}(\sigma,\alpha,H;P)$ *is*

$$\Pi_H(n_1,\dots,n_k) = \frac{H!}{(H-k)!}\frac{1}{(\alpha+1)_{n-1}}\sum_{(\ell_1,\dots,\ell_k)}\frac{\Gamma(\alpha/\sigma+|\ell^{(k)}|)}{\sigma\Gamma(\alpha/\sigma+1)}\prod_{j=1}^{k}\frac{\mathscr{C}(n_j,\ell_j;\sigma)}{H^{\ell_j}},$$

where the sum runs over all the vectors $\ell^{(k)} = (\ell_1,\dots,\ell_k)$ *such that* $\ell_j\in\{1,\dots,n_j\}$ *and* $|\ell^{(k)}| = \ell_1+\cdots+\ell_k$.

Based on this result, one may determine the system of predictive distributions corresponding to the Pitman–Yor multinomial process and the related urn-scheme. This admits a tractable form if one conditions on $\ell^{(k)} = (\ell_1,\dots,\ell_k)$ that will act as latent variables, thus simplifying computations. Firstly, it can be easily noted that

$$\mathrm{pr}(\ell_1 = l_1,\dots,\ell_k = l_k \mid \theta^{(n)}) \propto \Gamma(\alpha/\sigma+|l^{(k)}|)\prod_{j=1}^{k}\frac{\mathscr{C}(n_j,l_j;\sigma)}{H^{l_j}}, \tag{5}$$

for any $l^{(k)} = (l_1,\dots,l_k)$ such that $l_j\in\{1,\dots,n_j\}$, and is zero elsewhere. These latent random variables can be interpreted in terms of the multiroom Chinese restaurant metaphor (Camerlenghi et al., 2018), but we do not pursue the discussion here. An efficient algorithm for sampling independent

values from (5) is derived and presented in Section 4. This is very useful since it enables the Monte Carlo approximation of its expectation.

Theorem 2. *Let* $\theta_1, \ldots, \theta_n \mid \tilde{p}_H \overset{\text{iid}}{\sim} \tilde{p}_H$ *and* $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$. *If* $\theta^{(n)} = (\theta_1, \ldots, \theta_n)$ *displays* k *distinct values* $\theta_1^*, \ldots, \theta_k^*$ *with frequencies* $n_1, \ldots, n_k$, *then*

$$\begin{aligned} \text{pr}(\theta_{n+1} \in A \mid \theta^{(n)}) = \left(1 - \frac{k}{H}\right) \left(\frac{\alpha + |\bar{\ell}^{(k)}|\sigma}{\alpha + n}\right) P(A) \\ + \sum_{j=1}^{k} \left(\frac{1}{H} \frac{\alpha + |\bar{\ell}^{(k)}|\sigma}{\alpha + n} + \frac{n_j - \bar{\ell}_j \sigma}{\alpha + n}\right) \delta_{\theta_j^*}(A), \end{aligned} \tag{6}$$

having set $\bar{\ell}^{(k)} = (\bar{\ell}_1, \ldots, \bar{\ell}_k) = E(\ell^{(k)} \mid \theta^{(n)})$, $|\bar{\ell}^{(k)}| = \bar{\ell}_1 + \cdots + \bar{\ell}_k$ *and* $\ell^{(k)} = (\ell_1, \ldots, \ell_k)$ *is the vector of integer-valued random variables whose distribution is described in* (5).

The well-known predictive distribution of the Dirichlet multinomial process is recovered with $\sigma = 0$. In this special case, the predictive law (6) does not depend on the conditional expectations of the underlying latent variables. Moreover, as $H \to \infty$ one can easily see that $\bar{\ell}_j \to 1$ and $|\bar{\ell}^{(k)}| \to k$ in probability. This unsurprisingly implies that, as H increases, the predictive distributions of the infinite-dimensional Pitman–Yor and (6) get closer.

The closed form expression of $\Pi_H(n_1, \ldots, n_k)$ in Theorem 2 is essential for determining the probability distribution of $K_{n,H}$. Besides its theoretical relevance, the law of $K_{n,H}$ is often of great importance in applications, e.g. for mixture modeling or clustering. Compared to the Dirichlet multinomial case, which is recovered as $\sigma \to 0$, the Pitman–Yor multinomial process induces a richer parametrization for $K_{n,H}$, hence enhancing model flexibility.

Theorem 3. *If* $\theta_1, \ldots, \theta_n \mid \tilde{p}_H \overset{\text{iid}}{\sim} \tilde{p}_H$ *and* $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$, *the probability distribution of the number of distinct values* $K_{n,H}$ *in* $\theta^{(n)}$ *equals*

$$\text{pr}(K_{n,H} = k) = \frac{H!}{(H-k)!} \frac{1}{(\alpha+1)_{n-1}} \sum_{\ell=k}^{n} \frac{1}{H^\ell} \frac{\Gamma(\alpha/\sigma + \ell)}{\Gamma(\alpha/\sigma + 1)} \mathscr{S}(\ell, k) \mathscr{C}(n, \ell; \sigma),$$

for any $k \leq \min\{H, n\}$, *where* $\mathscr{S}(\ell, k) = 1/k! \sum_{r=0}^{k} (-1)^{k-r} k! \{r!(k-r)!\}^{-1} r^\ell$ *is the Stirling number of the second kind.*

The parameter α controls the location of $K_{n,H}$, while σ regulates both the location and the variability. As suggested by Figure 1, the choice $\sigma = 0$ leads to very informative prior distributions, implying that the choice of the location parameter α in the Dirichlet multinomial process is tricky and heavily influences the inferential results. Clearly, one might mitigate this issue by placing a prior distribution on α. However, this would lead to a specification for which Bayesian inferences of interest lack closed

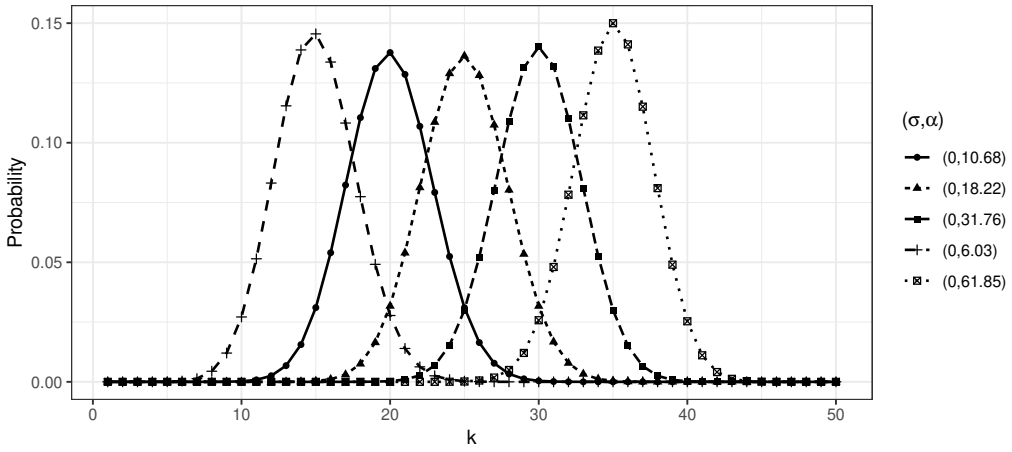

Figure 1: Distribution of the number of clusters $\text{pr}(K_{n,H} = k)$ in the Dirichlet case ($\sigma = 0$), when $n = 100$, $H = 50$, and for various choices of the location parameter α.

form expressions and in addition it would require a Metropolis step in the posterior sampling. In contrast, the Pitman–Yor multinomial process, with $\sigma > 0$, exhibits a degree of flexibility that does not require the specification of hyperpriors on (σ, α) and, as a result, the Gibbs sampler we rely on features a much faster mixing. Additionally, in view of our results it is apparent that the presence of $\sigma > 0$ does not impact the analytical tractability of this more general model. In practice, the parameters (σ, α) can be selected by fixing the mean $E(K_{n,H})$ and the variance $\text{var}(K_{n,H})$ equal to some pre-specified constants, and then numerically solving this system of equations in the variables (σ, α). Alternatively, one may choose the parameters (σ, α) by direct inspection of the distribution in Theorem 3. As an illustration, we depict in Figure 2 the distribution of $K_{n,H}$ for different choices of (σ, α), keeping its expectation fixed. As the discount parameter σ increases, the law of $K_{n,H}$ becomes less informative. This is further reflected in a higher degree of flexibility and robustness of the Pitman–Yor multinomial process compared to the Dirichlet case. See also De Blasi et al. (2015) and Canale and Prünster (2017) for further discussions on the robustness issue.

4 Posterior distribution and latent variables sampling

The use of the Pitman–Yor multinomial process in applications is greatly facilitated by the availability of its posterior distribution. Indeed, the posterior law of a Pitman–Yor multinomial, conditionally on the set of latent variables (5), is available in closed form and it equals the distribution of a linear combination of Dirichlet and ratio-stable distributions. Such a representation parallels the quasi-conjugate posterior characterization of the Pitman–Yor process (Lijoi et al., 2008) in the infinite-dimensional case. When the sample $\theta^{(n)} = (\theta_1, \ldots, \theta_n)$ displays $k < H$ distinct values $\theta_1^*, \ldots, \theta_k^*$, we let $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ represent the point masses in $\tilde{p}_H$ that are not included in $\theta^{(n)}$, up to a permutation.

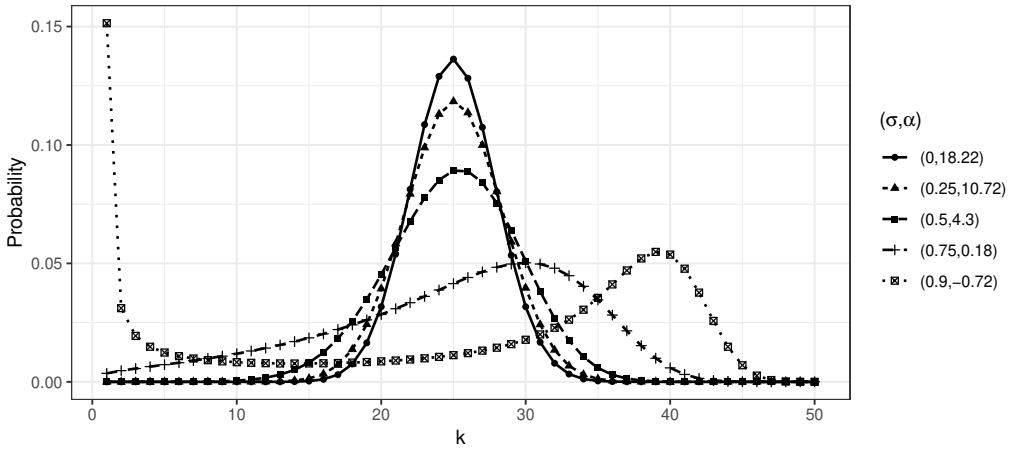

Figure 2: Distribution of the number of clusters $\text{pr}(K_{n,H} = k)$ in the Pitman–Yor multinomial case when $n = 100$, $H = 50$, and for various choices of (σ, α) so that the expected value $\mathbb{E}(K_{n,H}) = 25$ is fixed.

Theorem 4. *Let* $\theta_1, \ldots, \theta_n \mid \tilde{p}_H \overset{\text{iid}}{\sim} \tilde{p}_H$ *and* $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$ *with* P *diffuse. Moreover, let* $\ell^{(k)} = (\ell_1, \ldots, \ell_k)$ *be a collection of random variables having distribution* (5)*. Then, the posterior distribution of* $\tilde{p}_H$ *conditional on* $\theta^{(n)}$ *and* $\ell^{(k)}$ *is*

$$\tilde{p}_H \mid \theta^{(n)}, \ell^{(k)} = \sum_{j=1}^{k} (W_j + W_{k+1} R_j) \delta_{\theta_j^*} + W_{k+1} \sum_{j=k+1}^{H} R_j \delta_{\bar{\theta}_j},$$

where $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ *are independent and identically distributed from* P*. Moreover, the two vectors* $(W_1, \ldots, W_k)$ *and* $(R_1, \ldots, R_{H-1})$ *are conditionally independent, given* $(\theta^{(n)}, \ell^{(k)})$*, and*

$$\begin{aligned}(W_1, \ldots, W_k \mid \theta^{(n)}, \ell^{(k)}) &\sim \text{DIR}(n_1 - \ell_1\sigma, \ldots, n_k - \ell_k\sigma, \alpha + |\ell^{(k)}|\sigma),\\ (R_1, \ldots, R_{H-1} \mid \theta^{(n)}, \ell^{(k)}) &\sim \text{RS}(\sigma, \alpha + |\ell^{(k)}|\sigma; 1/H, \ldots, 1/H).\end{aligned}$$

The ratio-stable distribution appearing in Theorem 4 is such that $\alpha + |\ell^{(k)}|\sigma > 0$, for any σ and α, implying that the hierarchical representation of $(R_1, \ldots, R_H \mid \theta^{(n)}, \ell^{(k)})$ in terms of tempered-stable random variables, as for Proposition 2, can be used for simulation purposes. It is easy to check that as $\sigma \to 0$ the posterior distribution of the Dirichlet multinomial is recovered, while also being independent on $(\ell_1, \ldots, \ell_k)$.

Therefore, provided that we can simulate independent values from (5), we can obtain independent posterior samples for $\tilde{p}_H$ without the need of Markov chain Monte Carlo. To this end, note that the law of $(\ell_1, \ldots, \ell_k)$ in (5) is discrete with finite support, meaning that in principle one could directly sample from it. However, standard strategies are computationally feasible only in very simple cases, because the number of support points rapidly increases with n and $(n_1, \ldots, n_k)$. We address this issue with

a data-augmentation step. Indeed, by expanding over the gamma integral in (5), we recognize that, conditionally on a latent variable V, the discrete random variables $(\ell_1, \ldots, \ell_k)$ become independent and therefore much easier to simulate. Specifically, we have

$$\mathrm{pr}(\ell_1 = l_1, \ldots, \ell_k = l_k \mid \theta^{(n)}, V) \propto \prod_{j=1}^{k} \left(\frac{V}{H}\right)^{l_j} \mathscr{C}(n_j, l_j; \sigma),$$

where V is a positive random variable on $(0, \infty)$ having density, conditional on $\theta^{(n)}$, given by

$$f(v) \propto e^{-v} v^{\alpha/\sigma - 1} \prod_{j=1}^{k} \sum_{\ell_j=1}^{n_j} \left(\frac{v}{H}\right)^{\ell_j} \mathscr{C}(n_j, \ell_j; \sigma).$$

A draw from the above density can be simulated using acceptance-rejection strategies. We obtained very good empirical performance with the classic ratio-of-uniform acceptance-rejection algorithm applied on the logarithmic scale $\log V$; see e.g. Devroye (1986). We remark that the generalized factorial coefficients appearing in the above distributions should not be computed directly from their definition, since one can more efficiently rely on the recursive relationships $\mathscr{C}(n+1, k; \sigma) = \mathscr{C}(n, k; \sigma)(n - k\sigma) + \sigma\mathscr{C}(n, k-1; \sigma)$, with initial conditions $\mathscr{C}(0, 0; \sigma) = 1$, $\mathscr{C}(n, 0; \sigma) = 0$ for $n > 0$ and $\mathscr{C}(n, k; \sigma) = 0$ for $k > n$.

Remark 2. For the sake of exposition, we have confined ourselves to highlighting the posterior distribution of the random probability measure $\tilde{p}_H$. However, Theorem 4 also yields, with the obvious modifications, the posterior distribution of the random probabilities $(\pi_1, \ldots, \pi_H)$ under multinomial sampling. In such a setting, the frequencies $n_1, \ldots, n_k$ correspond to the k occupied cells in a multinomial distribution having probability vector $(\pi_1, \ldots, \pi_H)$ and a ratio-stable prior. Then the posterior of $(\pi_1, \ldots, \pi_H)$ will coincide with the weights of $\tilde{p}_H$ in Theorem 4.

5 Weak limit representation of the Pitman–Yor process

In this section we draw a sharp connection between the Pitman–Yor multinomial process and the infinite-dimensional Pitman–Yor, which is recovered as limiting case when $H \to \infty$. This formal relationship sheds some further light on the interpretation of (σ, α), while motivating the use of $\tilde{p}_H \sim \mathrm{PYM}(\sigma, \alpha, H; P)$ as an approximation of the infinite-dimensional process $\tilde{p}_\infty \sim \mathrm{PY}(\sigma, \alpha; P)$. Our next theorem relies on the notion of weak convergence for random measures (Daley and Vere-Jones, 2008) and on the convergence of the exchangeable partition probability functions. However, one may alternatively adapt the arguments used in Kingman (1975), who proved the analogous of our theorem in the Dirichlet case.

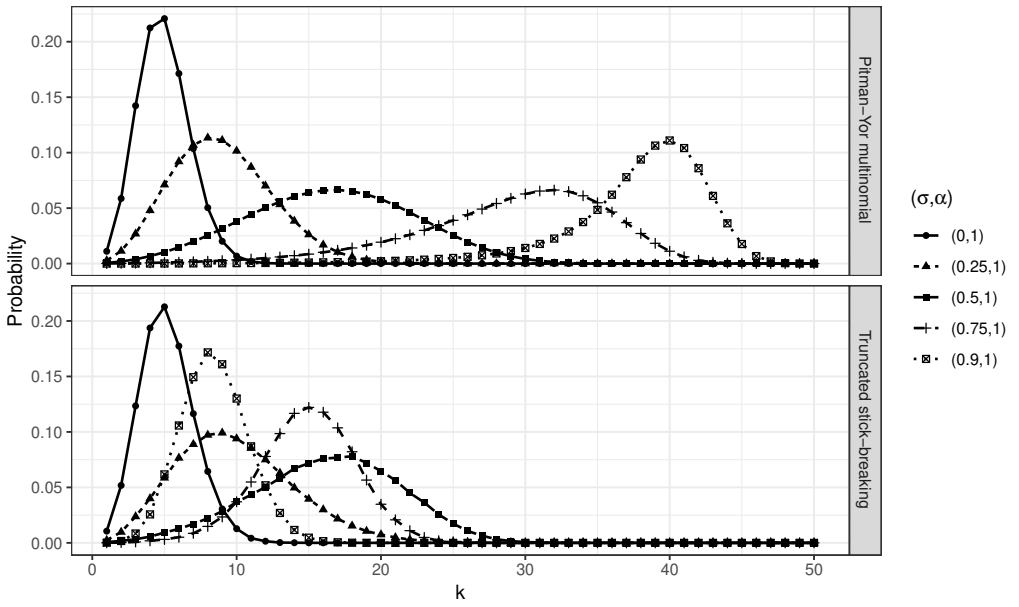

Figure 3: Distribution of the number of clusters $\text{pr}(K_{n,H} = k)$ in the Pitman–Yor multinomial case (upper plot), and in the truncated stick-breaking case (lower plot), when $n = 100$, $H = 50$, and $\alpha = 1$, for various choices of σ. The distribution of the truncated stick-breaking is obtained averaging over 10^4 Monte Carlo simulations.

Theorem 5. *Let* $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha, H; P)$ *and* $\tilde{p}_\infty \sim \text{PY}(\sigma, \alpha; P)$. *Then the law of the process* $\tilde{p}_H$ *weakly converges to the law of* $\tilde{p}_\infty$ *as* $H \to \infty$.

In view of the above theorem, it is quite natural to draw a comparison between the Pitman–Yor multinomial process and the truncated stick-breaking representation $\tilde{p}_{H,\text{tr}}$ in (2). Despite its popularity, the truncated stick-breaking construction lacks a deep theoretical understanding as one cannot rely on results analogous to the ones that we have displayed in the previous sections Specifically, the exchangeable partition probability function, the associated predictive schemes and the distribution of the number of clusters are not available. Hence, the use of $\tilde{p}_{H,\text{tr}}$ as prior law can be motivated only when H is large enough, so that one can consider sampled trajectories of $\tilde{p}_{H,\text{tr}}$, conditional on the data, as reasonable approximations of the realizations of the posterior infinite-dimensional process. In contrast, the Pitman–Yor multinomial can be studied and used even for small values of H.

In terms of quality of the approximation, there is a rather striking argument in favor of the Pitman–Yor multinomial process. It is well-known that the truncation level H required to be reasonably close to $\tilde{p}_\infty$ might be exceptionally large, especially when the discount parameter σ approaches 1. In practice, one would typically choose the largest truncation level H which maintains computations feasible. However, this might lead to very poor approximations of the infinite process if the truncated stick-breaking $\tilde{p}_{H,\text{tr}}$ were employed. Suppose we are given a sample of $n = 100$ observations and a conservative truncation level $H = 50$ is selected. Then, one might expect that a higher value of σ

implies an increase in the expected number of clusters, thus paralleling the behavior of the Pitman–Yor process. Unfortunately, this is not the case with $\tilde{p}_{H,\mathrm{tr}}$, as shown in Figure 3. Indeed, the mode of the distribution of $K_{n,H}$ initially increases but then decreases as a function of σ and a similar mechanism would hold also for the parameter α. Broadly speaking, this occurs because of the stick-breaking truncation: large values of either σ or α push, on average, the mass of $\tilde{p}_{H,\mathrm{tr}}$ towards the last atom, possibly forcing $\tilde{p}_{H,\mathrm{tr}}$ to collapse to a single random mass. This is a strongly undesirable behavior which has no modeling justification, and furthermore it undermines one of the most appealing property of the probabilistic structure of the Pitman–Yor process prior, namely the ability of controlling the variability of the cluster distribution. On the other hand, the Pitman–Yor multinomial process preserves the peculiar features of the Pitman–Yor process, as shown in Figure 3, while still being computationally tractable.

We now conduct a formal comparison between $\tilde{p}_H$ and $\tilde{p}_{H,\mathrm{tr}}$ within the context of mixture modeling. If the data $Y_1, \ldots, Y_n$ are conditionally independent and identically distributed draws from the random density $\tilde{f}$ in (1), then the marginal density of the sample is

$$m_\infty(Y^{(n)}) = E\left\{\prod_{i=1}^{n} \tilde{f}(Y_i)\right\} = E\left\{\prod_{i=1}^{n} \int_\Theta \mathcal{K}(Y_i;\theta)\tilde{p}_\infty(\mathrm{d}\theta)\right\}, \tag{7}$$

where $Y^{(n)} = (Y_1, \ldots, Y_n)$ and the expected value is taken with respect to the prior law of $\tilde{p}_\infty$. Similarly, we define the marginal densities m_H and $m_{H,\mathrm{tr}}$ as in (7), having replaced $\tilde{p}_\infty$ with the approximations $\tilde{p}_H$ and $\tilde{p}_{H,\mathrm{tr}}$, respectively. Upper-bounds for the total variation distance between these marginal densities were obtained by Ishwaran and James (2001) in the truncated Pitman–Yor case and Ishwaran and Zarepour (2000, 2002) in the Dirichlet multinomial case. When $\sigma = 0$, the the total variation distance between m_∞ and $m_{H,\mathrm{tr}}$ vanishes exponentially fast. On the basis of this result, Ishwaran and Zarepour (2000) argued that the truncated stick-breaking representation $\tilde{p}_{H,\mathrm{tr}}$ might constitute a better approximation than $\tilde{p}_H$ in the Dirichlet case. However, the aforementioned exponential decay does not occur for general values of σ and furthermore the quality of the truncated stick-breaking approximation deteriorates as σ increases. These aspects are clarified in the following proposition.

Proposition 4. *Let m_∞, m_H and $m_{H,\mathrm{tr}}$ be the marginal densities defined in (7). If $\sigma \in (0,1)$ and P is diffuse then*

$$d_{\mathrm{TV}}\left\{m_{H,\mathrm{tr}}, m_\infty\right\} \leq 2\left[1 - \left\{1 - \frac{\left(\frac{\alpha}{\sigma}+1\right)_{H-1}}{\left(\frac{\alpha}{\sigma}+\frac{1}{\sigma}\right)_{H-1}}\right\}^n\right] = \mathcal{O}(H^{-\frac{1}{\sigma}+1}), \quad H \to \infty.$$

If $\Psi_{n,H}$ and $\Psi_{n,\infty}$ are the random partitions associated to $\tilde{p}_H$ and $\tilde{p}_\infty$ respectively, then

$$d_{\text{TV}}\{m_H, m_\infty\} \leq d_{\text{TV}}\{\text{pr}(\Psi_{n,H}), \text{pr}(\Psi_{n,\infty})\} = \mathcal{O}\left(\frac{1}{H}\right), \quad H \to \infty.$$

The total variation distance $d_{\text{TV}}\{\text{pr}(\Psi_{n,H}), \text{pr}(\Psi_{n,\infty})\}$ can be obtained explicitly, although the actual computation could be cumbersome, since it requires the summation over the space of the partitions of $[n]$. The proportionality constants relative to the above convergence rates are known and they are reported in the Supplementary Material.

Hence, when $\sigma > 1/2$ the convergence rate of $d_{\text{TV}}\{\text{pr}(\Psi_{n,H}), \text{pr}(\Psi_{n,\infty})\}$ is linear regardless the value of σ. In contrast, the upper-bound in the truncated stick-breaking case displays slower converge rates as σ increases, and it is not anymore exponential when $\sigma > 0$. This fact, together with the qualitative findings illustrated in Figure 3, suggests that the Pitman–Yor multinomial prior might be preferable especially when σ is large. When $\sigma < 1/2$ the truncated stick-breaking approximation might behave better than the Pitman–Yor multinomial in terms of convergence rates, but the unappealing behavior of $\tilde{p}_{H,\text{tr}}$ highlighted in Figure 3 will still be present.

6 Simulation study

The additional flexibility of the Pitman–Yor multinomial prior is illustrated on a synthetic dataset of $n = 3000$ independent and identically distributed observations from the mixture

$$\frac{1}{4}\mathcal{N}(y; -2, 0.2^2) + \frac{1}{8}\mathcal{N}(y; -1, 0.2^2) + \frac{1}{4}\mathcal{N}(y; 0, 0.2^2) + \frac{1}{8}\mathcal{N}(y; 1, 0.2^2) + \frac{1}{4}\mathcal{N}(y; 2, 0.2^2),$$

with $\mathcal{N}(y; \mu, \sigma^2)$ denoting the Gaussian density function with mean μ and variance σ^2. The model is given by $\tilde{f}(y) = \sum_{h=1}^{H} \pi_h \mathcal{K}(y; \tilde{\theta}_h)$, where the random weights and the random locations have Pitman–Yor multinomial distribution. The true number of mixture components is 5 and the goal is to infer it from the data. To this purpose, we rely on the approach advocated by Rousseau and Mengersen (2011). In such a setting H is assumed to be large enough and the optimal number of clusters is inferred by inspecting the posterior distribution of $K_{n,H}$. As discussed in Section 3, in the Dirichlet multinomial case the results may depend on the choice of α and therefore we leverage the Pitman–Yor multinomial prior to allow for a more robust specification. This is achieved by increasing the prior variability of $K_{n,H}$, which is low in the Dirichlet case.

Let the kernel $\mathcal{K}(y; \theta)$ be a Gaussian density having mean μ and variance σ^2, with $\theta = (\mu, \sigma^2)$. We choose conditionally conjugate priors for the atoms $\tilde{\theta}_h = (\tilde{\theta}_{1h}, \tilde{\theta}_{2h})$ $(h = 1, \ldots, H)$, and in particular we assume independent gamma $\text{GA}(a_\sigma, b_\sigma)$ priors for the precisions $\tilde{\theta}_{2h}^{-1}$ and independent Gaussian priors for the locations $\tilde{\theta}_{11}, \ldots, \tilde{\theta}_{1H}$, with mean μ_μ and variance σ_μ^2. We set the hyperparameters consistently

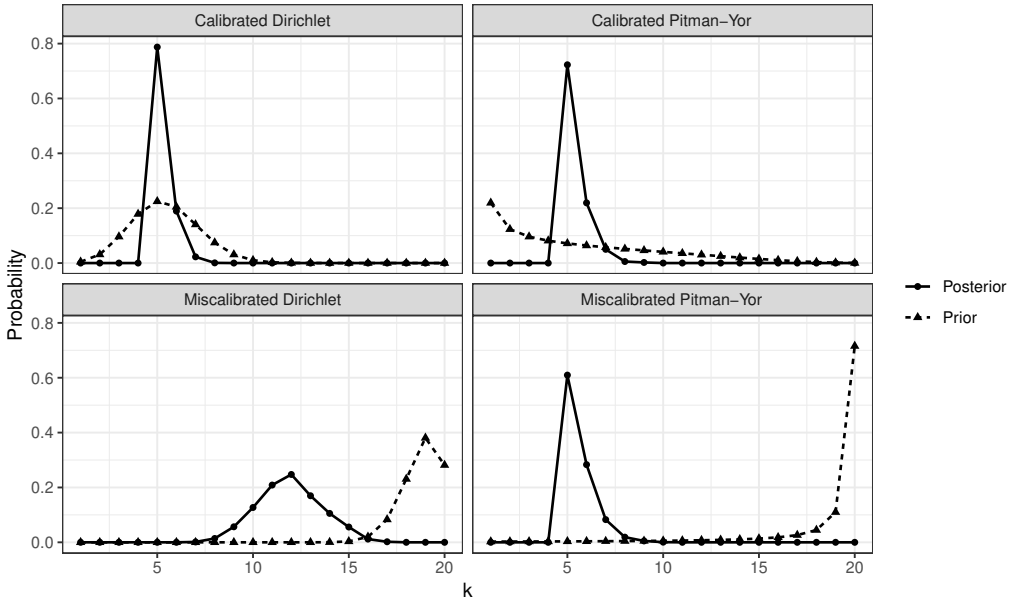

Figure 4: Prior and posterior distributions of the number of clusters $\text{pr}(K_{n,H} = k)$ corresponding to the four scenarios described in Table 1. The a posteriori distributions are obtained averaging over 5000 Markov chain Monte Carlo samples.

with the data generating process to make the prior distributions centered on the true values but still relatively vague. More precisely, we set $\mu_\mu = 0$ and $\sigma^2_\mu = 1000$, whereas we let $a_\sigma = 2.5$ and $b_\sigma = 0.1$.

We let $H = 20$, a fairly conservative upper bound for the true number of mixture components, which is 5 in our simulation study. We consider four different prior specifications for the ratio-stable parameters σ and α, as summarized in Table 1. In two of these scenarios, the discount parameter σ is set to 0, corresponding to the Dirichlet multinomial, which we aim at comparing with the Pitman–Yor prior. As displayed in Table 1, two hyperparameters settings are well calibrated, in the sense that the expected values of the number of clusters $K_{n,H}$ are both close to 5 a priori. Under these calibrated choices, we expect the model to be able to recover the correct number of clusters a posteriori. However, in real applications one does not know the true number of components and therefore she might inadvertently adopt a miscalibrated prior for the data at hand. This scenario is mimicked by considering hyperparameters that lead to a priori expectations for $K_{n,H}$ close to the upper bound, leading to an "overfitted" mixture model.

	Calibrated Dirichlet	Calibrated PY	Miscalibrated Dirichlet	Miscalibrated PY
α	1	-0.18	20	-0.02
σ	0	0.40	0	0.80
$E(K_{n,H})$	5.42	5.42	18.81	18.81

Table 1: Hyperparameter settings for the simulation study

Posterior samples for $K_{n,H}$ in each scenario can be obtained via Markov chain Monte Carlo. The details are given in the Supplementary Material. We run the algorithm for 7000 iterations and hold out the first 2000 as burn-in period. From the results in Figure 4 it is evident that both the calibrated choices are able to recover the correct number of clusters. Conversely, in the overfitted scenarios the differences are marked: under the Dirichlet multinomial specification the distribution of $K_{n,H}$ struggles to deviate from the prior, whereas in the Pitman–Yor multinomial case the posterior law of $K_{n,H}$ correctly recovers the true number of mixture components. This behavior motivates the use of the Pitman–Yor multinomial to robustify mixture modeling, with applications beyond the convex mixture regression modeling presented in this paper.

7 Convex mixture regression modeling

Dichlorodiphenyldichloroethylene (DDE) is a persistent metabolite of the pesticide DDT, and it is measured in the maternal serum during the third quarter of pregnancy. The presence of DDT has been linked to preterm birth, a major contributor to infant mortality (Longnecker et al., 2001). Hence, there is strong interest in relating the dose level of the DDE to the corresponding risk of premature delivery. We assume that the gestational ages at delivery $Y_1, \ldots, Y_n$ for $n = 2312$ are conditionally independent draws from the mixture density

$$\tilde{f}_x(y) = \{1 - \beta(x)\} \sum_{h=1}^{H} \pi_h \mathcal{K}(y; \tilde{\theta}_h) + \beta(x) \mathcal{K}(y; \tilde{\theta}_\infty), \qquad x \geq 0, \tag{8}$$

where x represents the DDE covariate associated to each woman and $\tilde{\theta}_1, \ldots, \tilde{\theta}_H, \tilde{\theta}_\infty$ are random variables in Θ. We extend the approach of Canale et al. (2018) by assuming a ratio stable law for the mixing weights, namely $(\pi_1, \ldots, \pi_{H-1}) \sim \text{RS}(\sigma, \alpha; 1/H, \ldots, 1/H)$. The smooth transition between the two components in (8) is regulated by a nondecreasing bounded function $\beta(x) \in [0, 1]$ defined for each DDE level $x \geq 0$ with $\beta(0) = 0$. In quantitative risk assessment one customarily assumes that $\beta(x)$ can be expressed as a linear combination of pre-specified basis functions $\beta(x) = \sum_{m=1}^{M} \psi_m(x)\gamma_m$ where $\psi_1(x), \ldots, \psi_M(x)$ are nondecreasing and such that $\psi_m \in [0, 1]$ $(m = 1, \ldots, M)$. Under this choice, the constraints on $\beta(x)$ are satisfied if $0 \leq \gamma_m \leq 1$ and $\sum_{m=1}^{M} \gamma_m = 1$. Hence, we let $(\gamma_1, \ldots, \gamma_{M-1}) \sim \text{RS}(\sigma_\gamma, \alpha_\gamma; 1/M, \ldots, 1/M)$. Although several alternatives are available for $\psi_1(x), \ldots, \psi_M(x)$, a tractable default choice is the I-splines basis (Ramsay, 1988), with the knots placed on the empirical quantiles of the DDE. This is slightly different from the approach of Canale et al. (2018), who consider I-splines with equally spaced knots. Moreover, we set $\psi_M(x) = 0$ for any $x \geq 0$ to allow asymptotes in $\beta(x)$.

Compared to more complex covariate-dependent approaches, the convex mixture regression model is appealing for quantitative risk assessment because of its intuitive interpretation. Specifically, when

	Low V Dirichlet	Low V Pitman–Yor	High V Dirichlet	High V Pitman–Yor
α	1	0	1	0
σ	0	0.3	0	0.3
α_γ	20	1.1	2	-0.7
σ_γ	0	0.9	0	0.9
$\text{sd}(\gamma_m)$	0.07	0.07	0.17	0.17

Table 2: Hyperparameter settings for the convex mixture regression model

there is no exposure to DDE ($x = 0$), observations are drawn from a mixture model $\sum_{h=1}^{H} \pi_h \mathcal{K}(y; \tilde{\theta}_h)$ because $\beta(0) = 0$. Conversely, at high exposure levels ($x \to \infty$), observations smoothly shift towards a more adverse health profile, represented by $\mathcal{K}(y; \tilde{\theta}_\infty)$. Such a transition is regulated by the function $\beta(x)$, which has an explicit interpretation. Let $\tilde{F}_x(y)$ be the cumulative distribution function associated to the density $\tilde{f}_x(y)$ in equation (8). A common risk assessment measure is the additional risk function $\tilde{F}_x(a) - \tilde{F}_0(a)$, which is evaluated in some fixed clinical threshold a. Hence, one can show that $\beta(x) \propto \tilde{F}_x(a) - \tilde{F}_0(a)$, implying that $\beta(x)$ constitutes a standardized measure of risk.

We estimate the convex mixture regression model under different hyperparameters settings, which are reported in Table 2. Moreover, we let $H = M = 10$, to parallel the choices of Canale et al. (2018). Since the true data generating mechanism is unknown, there is no clear notion of calibrated model. Hence, to emphasize the differences between Dirichlet and Pitman–Yor priors, we consider two different variability levels for the vector $(\gamma_1, \ldots, \gamma_M)$ appearing in the specification of $\beta(x)$. The variance of each weight γ_m can be obtained from the formula $\text{var}(\gamma_m) = H^{-1}(1 - H)^{-1}(1 - \sigma_\gamma)/(\alpha_\gamma + 1)$; see Carlton (2002). In the low variance scenarios, the prior is centered around its prior expectation, and vice versa in the high variance ones.

Consistently with Canale et al. (2018), let the kernel function $\mathcal{K}(y; \theta)$ be a Gaussian density having mean μ and variance σ^2, with $\theta = (\mu, \tau)$. Under this choice, the prior distributions for each atom $\tilde{\theta}_h = (\tilde{\theta}_{1h}, \tilde{\theta}_{2h})$ $(h = 1, \ldots, H)$, in equation (8) and for the adverse health profile atom $\tilde{\theta}_\infty = (\tilde{\theta}_{1\infty}, \tilde{\theta}_{2\infty})$ can be chosen conditionally conjugate. In first place, we assume independent gamma $\text{GA}(a_\sigma, b_\sigma)$ priors for the precisions $\tilde{\theta}_{2h}^{-1}$, independently also on $\tilde{\theta}_{2\infty} \sim \text{GA}(a_\sigma, b_\sigma)$. Moreover, we specify truncated independent Gaussian distributions for the locations $\tilde{\theta}_{11}, \ldots, \tilde{\theta}_{1H}$ and $\tilde{\theta}_{1\infty}$ with mean μ_μ and variance σ_μ^2. The truncations are imposed to met an adversity health profile property, namely that $\tilde{\theta}_{1\infty} < \tilde{\theta}_{1h}$ $(h = 1, \ldots, H)$ almost surely. Broadly speaking, such a constraint enforces large values of the DDE ($x \to \infty$) to be associated on average with a greater risk of premature birth. We set the hyperparameters consistently with previous works, so that $a_\sigma = b_\sigma = 2$ and $\sigma_\mu^2 = 10$, whereas we set $\mu_\mu = 39.27$.

Markov chain Monte Carlo is required for approximating the posterior distribution of model (8). Our algorithm resembles the one of Canale et al. (2018), with adjustments in the updates of $(\pi_1, \ldots, \pi_H)$

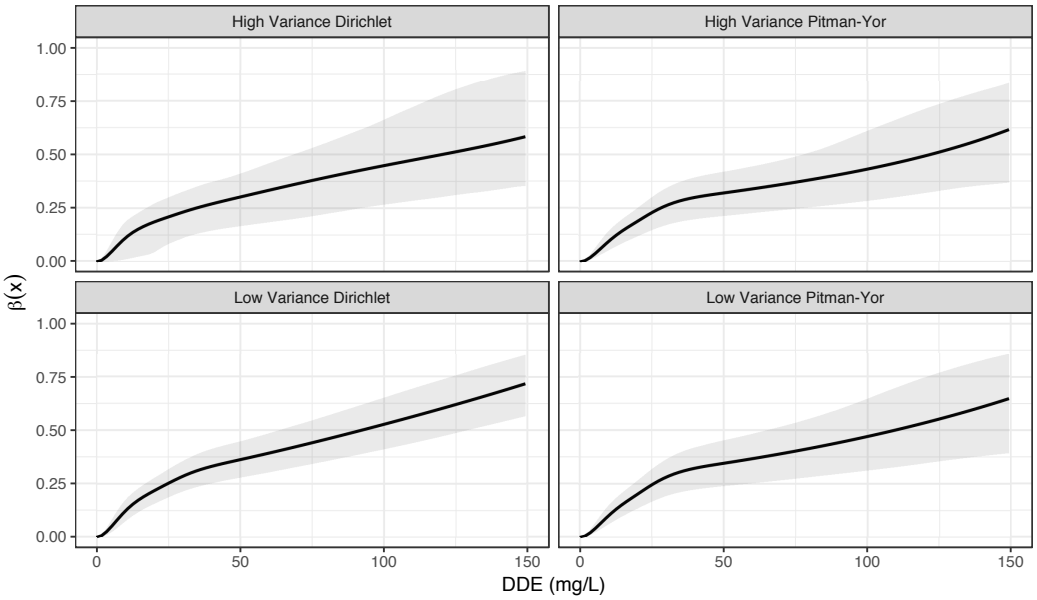

Figure 5: Posterior summaries of $\beta(x)$ in the four scenarios described in Table 2. Solid lines correspond to the posterior means, whereas the shaded areas denote 95% pointwise credible intervals.

and $(\gamma_1, \ldots, \gamma_M)$, according to Theorem 4. We run the Gibbs sampler for 60000 iterations with the first 10000 samples discarded as burn in. The estimated curves $\beta(x)$ and related credible bands, under the four scenarios of Table 2, are depicted in Figure 5. In the low variance case the Dirichlet multinomial prior slightly overestimates the function $\beta(x)$ compared to the other estimates and to the results in Canale et al. (2018), while also underestimating its variability. Conversely, the Pitman–Yor multinomial prior has the same prior variability as the Dirichlet, but recovers a posteriori essentially the same variability level of the high variance scenarios. This is due to the robustness of the Pitman–Yor multinomial prior discussed in Section 3.

A Supplementary material

Background on the Pitman–Yor process

Throughout the Supplementary Material, we will make extensive use of an alternative construction of the Pitman–Yor process based on completely random measures. Refer to Lijoi and Prünster (2010) for a review on nonparametric priors using completely random measures as a unifying concept. Any homogeneous and almost surely finite completely random measures without fixed points of discontinuity is characterized by the Laplace functional

$$E\left\{\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}(\mathrm{d}\theta)}\right\} = \exp\left[-\int_{\mathbb{R}^+\times\Theta}\left\{1-\mathrm{e}^{-sf(\theta)}\right\}\rho(s)\,cP(\mathrm{d}\theta)\,\mathrm{d}s\right],$$

for any $f : \Theta \to \mathbb{R}^+$ and with $\rho(s)\, cP(\mathrm{d}\theta)\, \mathrm{d}s$ the Lévy intensity function associated to $\tilde{\mu}$. The σ-stable process (Kingman, 1975) is identified by setting $c = 1$, $\rho(s) = \sigma s^{-1-\sigma}/\Gamma(1-\sigma)$, for some $\sigma \in (0,1)$, and letting $\mathbb{P}_\sigma$ denote its probability distribution, that is obviously defined on the space of boundedly finite measure on Θ. Let $\mathbb{P}_{\sigma,\alpha}$ be another probability measure which is absolutely continuous with respect to $\mathbb{P}_\sigma$ and such that

$$\frac{\mathrm{d}\mathbb{P}_{\sigma,\alpha}}{\mathrm{d}\mathbb{P}_\sigma}(g) = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha/\sigma+1)} g^{-\alpha}(\Theta). \tag{9}$$

The resulting random measure $\tilde{\mu}_{\sigma,\alpha}$ with distribution $\mathbb{P}_{\sigma,\alpha}$ is almost surely discrete while not completely random. Clearly when $\alpha = 0$ then $\tilde{\mu}_\sigma = \tilde{\mu}_{\sigma,0}$ is a σ-stable completely random measure. Moreover, $\tilde{p}_\infty = \tilde{\mu}_{\sigma,\alpha}/\tilde{\mu}_{\sigma,\alpha}(\Theta)$ is a Pitman–Yor process $\tilde{p}_\infty \sim \textsc{py}(\sigma, \alpha; P)$.

Proofs of Propositions 1,2 and 3

If we consider the partition $\{(\tilde{\theta}_1), \ldots, (\tilde{\theta}_{H-1}), \Theta \setminus (\tilde{\theta}_1, \ldots, \tilde{\theta}_{H-1})\}$ of Θ, by definition one has

$$(\pi_1, \ldots, \pi_{H-1} \mid \tilde{p}_{0,H}) = \{\tilde{p}_H(\{\tilde{\theta}_1\}), \ldots, \tilde{p}_H(\{\tilde{\theta}_{H-1}\}) \mid \tilde{p}_{0,H}\} \sim \textsc{rs}(\sigma, \alpha; 1/H, \ldots, 1/H),$$

since $\tilde{p}_{0,H}(\{\tilde{\theta}_h\}) = 1/H$ $(h = 1, \ldots, H)$. This entails that

$$\tilde{p}_H = \sum_{h=1}^{\infty} \xi_h \delta_{\tilde{\phi}_h} = \sum_{h=1}^{H} \pi_h \delta_{\tilde{\theta}_h}, \qquad \pi_h = \sum_{j:\tilde{\phi}_j = \tilde{\theta}_h} \xi_j,$$

and the vectors $(\pi_1, \ldots, \pi_{H-1})$ and $(\tilde{\theta}_1, \ldots, \tilde{\theta}_H)$ are independent, thus proving Proposition 1.

Let, now, $\sigma \in (0,1)$ and $\alpha > 0$. By virtue of (9), conditionally on $\tilde{p}_{0,H}$ the Pitman–Yor multinomial process can be represented as $\tilde{\mu}_{\sigma,\alpha}/\tilde{\mu}_{\sigma,\alpha}(\Theta)$, where the Laplace functional transform of $\tilde{\mu}_{\sigma,\alpha} \mid \tilde{p}_{0,H}$ is

$$\begin{aligned} E\left\{\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_{\sigma,\alpha}(\mathrm{d}\theta)} \mid \tilde{p}_{0,H}\right\} &= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha/\sigma+1)} E\left\{\tilde{\mu}_\sigma(\Theta)^{-\alpha} \mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_\sigma(\mathrm{d}\theta)} \mid \tilde{p}_{0,H}\right\} \\ &= \frac{\alpha}{\Gamma(\alpha/\sigma+1)} \int_0^\infty u^{\alpha-1} \mathrm{e}^{-u^\sigma} E\left\{\mathrm{e}^{-\int_\Theta f(\theta)\tilde{\mu}_\sigma^{(u)}(\mathrm{d}\theta)} \mid \tilde{p}_{0,H}\right\} \mathrm{d}u, \end{aligned} \tag{10}$$

for any measurable function $f : \Theta \to \mathbb{R}^+$ where $\tilde{\mu}_\sigma^{(u)} \mid \tilde{p}_{0,H}$ is a generalised gamma process, namely a completely random measure with Lévy intensity $\rho^{(u)}(s) = \sigma/\Gamma(1-\sigma) s^{-1-\sigma} \mathrm{e}^{-us}$ and baseline probability measure $\tilde{p}_{0,H}$. Moreover, conditionally on $(U, \tilde{p}_{0,H})$, the random variable $\tilde{\mu}_\sigma^{(u)}(A)$ is tempered stable for any set A such that $\tilde{p}_{0,H}(A) > 0$. Hence

$$(\pi_1, \ldots, \pi_{H-1} \mid U, \tilde{p}_{0,H}) = \left(\frac{J_1}{\sum_{h=1}^H J_h}, \ldots, \frac{J_{H-1}}{\sum_{h=1}^H J_h} \mid U, \tilde{p}_{0,H}\right),$$

with $U^\sigma \sim \text{Ga}(\alpha/\sigma, 1)$ and $J_h \mid U, \tilde{p}_{0,H} = \tilde{\mu}_\sigma^{(u)}(\{\tilde{\theta}_h\}) \mid U, \tilde{p}_{0,H} \sim \text{TS}(1/H, \sigma, U)$ independently and identically distributed for $h = 1, \ldots, H$, which concludes the proof of Proposition 2 for $\alpha > 0$. The case where $\alpha = 0$ s straightforward since the Laplace functional of $\tilde{\mu}_\sigma \mid \tilde{p}_{0,H}$ is already that of a σ-stable completely random measure.

The proof of Proposition 3 is a direct consequence of the first equality in equation (10).

Proof of Theorem 1

The exchangeable partition probability function by definition is

$$\Pi_H(n_1, \ldots, n_k) = \sum_{i_1 \neq \cdots \neq i_k} E\left(\prod_{j=1}^k \pi_{i_j}^{n_j}\right) = \frac{H!}{(H-k)!} E\left(\prod_{j=1}^k \pi_j^{n_j}\right),$$

where the sum runs over all the vectors $(i_1, \ldots, i_k)$ of distinct positive integers such that $i_j \in \{1, \ldots, H\}$, whereas the second equality is a consequence of the symmetry of the weights π_j. By virtue of (9), one further has

$$\begin{aligned} E\left(\prod_{j=1}^k \pi_j^{n_j}\right) &= E\left\{\prod_{j=1}^k \frac{J_j^{n_j}}{(\sum_{h=1}^H J_h)^{n_j}}\right\} \\ &= \frac{1}{\Gamma(\alpha/\sigma + 1)} \frac{1}{(\alpha+1)_{n-1}} \int_0^\infty u^{\alpha+n-1} e^{-u^\sigma} \prod_{j=1}^k \mathcal{V}_{n_j}(u) \mathrm{d}u, \end{aligned}$$

where each $\mathcal{V}_m(u)$ is defined, for every $m \geq 1$ and $u > 0$, as

$$\mathcal{V}_m(u) = \left\{(-1)^m \frac{\partial^m}{\partial u^m} e^{-u^\sigma/H}\right\} e^{u^\sigma/H} = \sum_{\ell=1}^m H^{-\ell} \mathscr{C}(m, \ell; \sigma) u^{-m+\ell\sigma}. \tag{11}$$

The last equality may be proved through combinatorial arguments similar to those used in Lemma 1 in Appendix A.3 and the results in Appendix A.4 of the supplementary material of Camerlenghi et al. (2019). Hence

$$\begin{aligned} \Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} &\frac{1}{\Gamma(\alpha/\sigma + 1)} \frac{1}{(\alpha+1)_{n-1}} \\ &\times \sum_{(\ell_1, \ldots, \ell_k)} \int_0^\infty u^{\alpha + |\ell^{(k)}|\sigma} e^{-u^\sigma} \mathrm{d}u \prod_{j=1}^k H^{-\ell_j} \mathscr{C}(n_j, \ell_j; \sigma), \end{aligned}$$

where the sum runs over $\ell^{(k)} = (\ell_1, \ldots, \ell_k)$ for $\ell_j \in \{1, \ldots, n_j\}$. The change of variable $v = u^\sigma$ in the above integral yields the desired result.

Proof of Theorem 2

By definition, the predictive distribution can be determined as follows

$$\mathrm{pr}(\theta_{n+1} \in A \mid \theta^{(n)}) \propto \Pi_H(n_1, \ldots, n_k, 1)P(A) + \sum_{j=1}^{k} \Pi_H(n_1, \ldots, n_j + 1, \ldots, n_k)\delta_{\theta_j^*}(A)$$

and the exchangeable partition probability functions acting as coefficients are mixtures over $\ell^{(k)}$. Indeed, if $\mathscr{V}_{n,k} = \prod_{j=1}^{k-1}(\alpha + j\sigma)/(\alpha+1)_{n-1}$, from Theorem 1

$$\Pi_H(n_1, \ldots, n_k, 1) = \frac{H!}{(H-k-1)!} \sum_{(\ell_1,\ldots,\ell_k)} \frac{\mathscr{V}_{n+1,|\ell^{(k)}|+1}}{H^{|\ell^{(k)}|+1}} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{\sigma^{\ell_j}}.$$

Moreover, the recursion $\mathscr{C}(n_j+1, \ell_j; \sigma) = \mathscr{C}(n_j, \ell_j; \sigma)(n_j - \ell_j\sigma) + \sigma\mathscr{C}(n_j, \ell_j - 1; \sigma)$ (Charalambides, 2002), combined with some algebraic manipulations, yields the following representation for $\Pi_H(n_1, \ldots, n_j + 1, \ldots, n_k)$

$$\frac{H!}{(H-k)!} \sum_{(\ell_1,\ldots,\ell_k)} \left\{ \frac{\mathscr{V}_{n+1,|\ell^{(k)}|+1}}{H^{|\ell^{(k)}|+1}} \prod_{i=1}^{k} \frac{\mathscr{C}(n_i, \ell_i; \sigma)}{\sigma^{\ell_i}} + \frac{\mathscr{V}_{n+1,|\ell^{(k)}|}}{H^{|\ell^{(k)}|}} (n_j - \ell_j\sigma) \prod_{i=1}^{k} \frac{\mathscr{C}(n_i, \ell_i; \sigma)}{\sigma^{\ell_i}} \right\}.$$

Then, by augmenting over the set of random variables $(\ell_1, \ldots, \ell_k)$ and after normalization, one obtains

$$\begin{aligned} \mathrm{pr}(\theta_{n+1} \in A \mid \theta^{(n)}, \ell^{(k)}) = \left(1 - \frac{k}{H}\right) \left(\frac{\alpha + |\ell^{(k)}|\sigma}{\alpha + n}\right) P(A) \\ + \sum_{j=1}^{k} \left(\frac{1}{H} \frac{\alpha + |\ell^{(k)}|\sigma}{\alpha + n} + \frac{n_j - \ell_j\sigma}{\alpha + n}\right) \delta_{\theta_j^*}(A), \end{aligned} \tag{12}$$

and the desired predictive distribution follows after taking the expectation with respect to (5).

Proof of Theorem 3

For any $k \leq \min\{H, n\}$ the probability mass function of $K_{n,H}$ is

$$\begin{aligned} \mathrm{pr}(K_{n,H} = k) &= \frac{1}{k!} \sum_{n^{(k)} \in \Delta_n} \binom{n}{n_1, \ldots, n_k} \Pi_H(n_1, \ldots, n_k) \\ &= \frac{H!}{(H-k)!} \sum_{t=k}^{n} \frac{\mathscr{V}_{n,t}}{\sigma^t H^t} \sum_{n^{(k)} \in \Delta_n} \sum_{\ell^{(k)} \in \Delta_t(n^{(k)})} \frac{1}{k!} \binom{n}{n_1, \ldots, n_k} \prod_{j=1}^{k} \mathscr{C}(n_j, \ell_j; \sigma), \end{aligned}$$

where the first sum $n^{(k)} \in \Delta_n$ runs over all vectors of positive integers $n^{(k)} = (n_1, \ldots, n_k)$ such that $|n^{(k)}| = n$, while $\ell^{(k)} \in \Delta_t(n^{(k)})$ denotes the sum over all the integers $\ell^{(k)} = (\ell_1, \ldots, \ell_k)$ such that

$\ell_j \in \{1, \ldots, n_j\}$ and $|\ell^{(k)}| = t$. Then, by interchanging the order of the summation and exploiting well-known combinatorial identities, we obtain

$$\begin{aligned}
\mathrm{pr}(K_{n,H} = k) &= \frac{H!}{(H-k)!} \sum_{t=k}^{n} \frac{\mathscr{V}_{n,t}}{\sigma^t H^t} \sum_{\ell^{(k)} \in \Delta_t} \frac{1}{k!} \sum_{n^{(k)} \in \Delta_n(\ell^{(k)})} \binom{n}{n_1, \ldots, n_k} \prod_{j=1}^{k} \mathscr{C}(n_j, \ell_j; \sigma) \\
&= \frac{H!}{(H-k)!} \sum_{t=k}^{n} \frac{\mathscr{V}_{n,t}}{\sigma^t H^t} \sum_{\ell^{(k)} \in \Delta_t} \frac{1}{k!} \binom{t}{\ell_1, \ldots, \ell_k} \mathscr{C}(n, t; \sigma) \\
&= \frac{H!}{(H-k)!} \sum_{t=k}^{n} \frac{\mathscr{V}_{n,t}}{\sigma^t H^t} \mathscr{S}(t, k) \mathscr{C}(n, t; \sigma),
\end{aligned}$$

where $n^{(k)} \in \Delta_n(\ell^{(k)})$ denotes the summation over all the integers $(n_1, \ldots, n_k)$ such that $n_j \in \{\ell_j, \ldots, n\}$ and $|n^{(k)}| = n$, whereas $\ell^{(k)} \in \Delta_t$ denotes the summation over all the integers $(\ell_1, \ldots, \ell_k)$ such that $|\ell^{(k)}| = t$. The result now follows from some simple algebra.

Proof of Theorem 4

We first state, without proof, the following technical lemma concerning the posterior distribution of $\tilde{p}_{0,H}$. The proof is based on elementary properties of species sampling models.

Lemma 1. *Let $\theta^{(n)} = (\theta_1, \ldots, \theta_n)$ be a draw from an exchangeable sequence directed by a Pitman–Yor multinomial process and let P be diffuse. Then the following equality in distribution holds true*

$$\tilde{p}_{0,H} \mid \theta^{(n)} = \frac{1}{H} \left(\sum_{j=1}^{k} \delta_{\theta_j^*} + \sum_{j=k+1}^{H} \delta_{\bar{\theta}_j} \right),$$

where $\bar{\theta}_{k+1}, \ldots, \bar{\theta}_H$ are independent and identically distributed draws from P and $\theta_1^*, \ldots, \theta_k^*$ are the distinct values in $\theta^{(n)}$.

Because of the symmetry of the weights, we can assume without loss of generality that the distinct values $\theta_1^*, \ldots, \theta_k^*$ are associated to the first k random weights $\pi_1, \ldots, \pi_k$ of $\tilde{p}_H$. Recalling representation (9), for any function $f : \Theta \to \mathbb{R}^+$, the Laplace functional of $\tilde{\mu}_{\sigma,\alpha} \mid \tilde{p}_{0,H}$ given the observations is

$$E\left\{ \mathrm{e}^{-\tilde{\mu}_{\sigma,\alpha}(f)} \mid \theta^{(n)}, \tilde{p}_{0,H} \right\} = \frac{E\left\{ \mathrm{e}^{-\int_\Theta f(\theta) \tilde{\mu}_{\sigma,\alpha}(\mathrm{d}\theta)} \prod_{j=1}^{k} \pi_j^{n_j} \mid \tilde{p}_{0,H} \right\}}{E\left(\prod_{j=1}^{k} \pi_j^{n_j} \mid \tilde{p}_{0,H} \right)},$$

where $\tilde{\mu}_{\sigma,\alpha}(f) = \int_\Theta f(\theta)\tilde{\mu}_{\sigma,\alpha}(\mathrm{d}\theta)$. Hence, following the same steps as for Theorem 1, the above Laplace functional transform may be written as

$$E\left\{\mathrm{e}^{-\tilde{\mu}_{\sigma,\alpha}(f)} \mid \theta^{(n)}, \tilde{p}_{0,H}\right\} = \\ = \frac{\int_0^\infty u^{\alpha+n-1}\mathrm{e}^{-\frac{1}{H}\sum_{j=1}^k\{f(\theta_j^*)+u\}^\sigma}\mathrm{e}^{-\frac{1}{H}\sum_{j=k+1}^H\{f(\tilde{\theta}_j)+u\}^\sigma}\prod_{j=1}^k \mathcal{V}_{n_j}\{f(\theta_j^*)+u\}\mathrm{d}u}{\int_0^\infty u^{\alpha+n-1}\mathrm{e}^{-u^\sigma}\prod_{j=1}^k \mathcal{V}_{n_j}(u)\mathrm{d}u}.$$

where each $\mathcal{V}_m(u)$ is defined as in (11). Hence, by augmenting the above Laplace functional over the set of latent variables $(\ell_1, \ldots, \ell_k)$ with distribution function (5), we obtain that

$$E\left\{\mathrm{e}^{-\tilde{\mu}_{\sigma,\alpha}(f)} \mid \theta^{(n)}, \ell^{(k)}, \tilde{p}_{0,H}\right\} \propto \\ \propto \int_0^\infty u^{\alpha+|\ell^{(k)}|\sigma-1}\,\mathrm{e}^{-\frac{1}{H}\sum_{j=k+1}^H\{f(\tilde{\theta}_j)+u\}^\sigma}\mathrm{e}^{-\frac{1}{H}\sum_{j=1}^k\{f(\theta_j^*)+u\}^\sigma}\prod_{j=1}^k\left\{1+\frac{f(\theta_j^*)}{u}\right\}^{-n_j+\ell_j\sigma}\mathrm{d}u.$$

Hence, after normalization

$$E\left\{\mathrm{e}^{-\tilde{\mu}_{\sigma,\alpha}(f)} \mid \theta^{(n)}, \ell^{(k)}, \tilde{p}_{0,H}\right\} = \int_0^\infty \prod_{j=k+1}^H E\left\{\mathrm{e}^{-f(\tilde{\theta}_j)J_j^*}\right\}\prod_{j=1}^k E\left\{\mathrm{e}^{-f(\theta_j^*)(J_j^*+I_j)}\right\} q_\infty(u)\mathrm{d}u,$$

where $q_\infty(u) = \sigma/\Gamma(\alpha/\sigma + |\ell^{(k)}|)u^{\alpha+|\ell^{(k)}|\sigma-1}\mathrm{e}^{-u^\sigma}I_{(0,\infty)}(u)$ is a density function and the corresponding random variable, say U, is such that $U^\sigma \sim \textsc{ga}(\alpha/\sigma + |\ell^{(k)}|, 1)$. Hence, conditionally on U and by marginalizing over $\tilde{p}_{0,H}$ as for Lemma 1, we get the following posterior representation for the unormalized Pitman–Yor multinomial process

$$\tilde{\mu}_{\sigma,\alpha} \mid \theta^{(n)}, \ell^{(k)}, U = \sum_{j=1}^k (J_j^* + I_j)\delta_{\theta_j^*} + \sum_{j=k+1}^H J_j^*\delta_{\bar{\theta}_j}, \tag{13}$$

where $\bar{\theta}_j$ are independent and identically distributed random variables from P and in addition

$$J_h^* \mid \theta^{(n)}, \ell^{(k)}, U \overset{\text{iid}}{\sim} \textsc{ts}(1/H, \sigma, U), \qquad (h = 1, \ldots, H),$$

whereas

$$I_j \mid \theta^{(n)}, \ell^{(k)}, U \overset{\text{ind}}{\sim} \textsc{ga}(n_j - \ell_j\sigma, U), \qquad (j = 1, \ldots, k).$$

Equation (13) already leads to a posterior representation of $\tilde{p}_H$. The normalization and the subsequent marginalization with respect to the random variable U leads to the final representation. In first place, set

$$W_j = \frac{I_j}{\sum_{j'=1}^k I_{j'} + \sum_{h=1}^H J_h^*}, \qquad (j = 1, \ldots, k), \qquad W_{k+1} = \frac{\sum_{j=1}^k J_j^*}{\sum_{j=1}^k I_j + \sum_{h=1}^H J_h^*},$$

whereas set $R_h = J_h^* / \sum_{h'=1}^H J_{h'}^*$ $(h = 1, \ldots, H)$. Note that the distribution of $(W_1, \ldots, W_k, W_{k+1})$ can be represented as follow

$$(\tilde{W}_1(1 - W_{k+1}), \ldots, \tilde{W}_k(1 - W_{k+1}), W_{k+1}), \quad \tilde{W}_j = \frac{I_j}{\sum_{j'=1}^k I_{j'}}, \quad (j = 1, \ldots, k),$$

and therefore the vector $(\tilde{W}_1, \ldots, \tilde{W}_k)$, given $\theta^{(n)}$ and $\ell^{(k)}$, has Dirichlet distribution and it is independent on $(R_1, \ldots, R_H)$, on W_{k+1} and on U . Indeed, the random variable U cancels in the ratio $I_j / \sum_{j'=1}^k I_{j'}$, whereas the independence on W_{k+1} is a consequence of the independence of the Dirichlet distribution with its own total mass $\sum_{j'=1}^k I_{j'}$. Because of Proposition 2, we recognize that $(R_1, \ldots, R_H \mid \theta^{(n)}, \ell^{(k)})$ follows a ratio-stable distribution. We now prove that, given $\theta^{(n)}$ and $\ell^{(k)}$, the random vector $(R_1, \ldots, R_H)$ is independent on W_{k+1}, which in turns is shown to be beta distributed.

Let $f(s_1, \ldots, s_H, t, u)$ be the density function associated to the random variables $(J_1^*, \ldots, J_H^*)$, $\sum_{j=1}^k I_j$ and U, respectively, given the observations $\theta^{(n)}$ and the latent variables $\ell^{(k)}$. Representation (13) implies that such a density factorizes as

$$f(s_1, \ldots, s_H, t, u) = q_\infty(u) f_u(t) \prod_{h=1}^H f_u^{(\sigma)}(s_h),$$

where $q_\infty(u)$ is defined as before, where $f_u(t)$ is the density of a gamma random variable, and where each $f_u^{(\sigma)}(s_h)$ represents the density of a tempered stable distribution. Now consider the change of variable $r_h = s_h / (s_1 + \cdots + s_H)$ $(h = 1, \ldots, H-1)$, $s = s_1 + \cdots + s_H$, $w = (s_1 + \cdots + s_H)/\{t + (s_1 + \cdots + s_H)\}$ and $u = u$. The resulting density is given by

$$\begin{aligned}
f(r_1, \ldots, r_{H-1}, s, w, u) &= s^H w^{-2} q_\infty(u) f_u\{s(1-w)w^{-1}\} \prod_{h=1}^H f_u^{(\sigma)}(s r_h), \\
&\propto u^{n+\alpha-1} s^{n+H-|\ell^{(k)}|\sigma-1} \frac{(1-w)^{n-|\ell^{(k)}|\sigma-1}}{w^{n-|\ell^{(k)}|\sigma+1}} \mathrm{e}^{-u\{s(1-w)w^{-1}\} - u^\sigma} \prod_{h=1}^H f_u^{(\sigma)}(s r_h), \\
&\propto u^{n+\alpha-1} \mathrm{e}^{-us/w} s^{n+H-|\ell^{(k)}|\sigma-1} \frac{(1-w)^{n-|\ell^{(k)}|\sigma-1}}{w^{n-|\ell^{(k)}|\sigma+1}} \prod_{h=1}^H f^{(\sigma)}(s r_h),
\end{aligned}$$

where $r_H = 1 - (r_1 + \cdots + r_{H-1})$ and where it has been exploited the relationship $f_u^{(\sigma)}(s r_h) = \mathrm{e}^{u^\sigma / H} \mathrm{e}^{-u s r_h} f_\sigma(s r_h)$, with $f^{(\sigma)}(s r_h)$ denoting the density of a positive stable distribution. Then, the integration over u and s leads to $f(r_1, \ldots, r_{H-1}, w) = \int_0^\infty \int_0^\infty f(r_1, \ldots, r_{H-1}, s, w, u) \mathrm{d}u \mathrm{d}s$ and

$$f(r_1, \ldots, r_{H-1}, w) \propto w^{\alpha+|\ell^{(k)}|\sigma-1} (1-w)^{n-|\ell^{(k)}|\sigma-1} \int_0^\infty s^{H-|\ell^{(k)}|\sigma-\alpha-1} \prod_{h=1}^H f^{(\sigma)}(s r_h) \mathrm{d}s.$$

This concludes the proof since $(W_{k+1} \mid \theta^{(n)}, \ell^{(k)}) \sim \text{BETA}(\alpha + |\ell^{(k)}|\sigma, n - |\ell^{(k)}|\sigma)$ and it is independent on $(R_1, \ldots, R_H \mid \theta^{(n)}, \ell^{(k)})$.

Proof of Theorem 5

The proof relies on the convergence of the exchangeable partition probability function in Theorem 1 to that of the Pitman–Yor process, which can be easily checked. Indeed, for any collection of measurable subsets $A_1, \ldots, A_n$ of Θ it holds

$$\begin{aligned}\text{pr}(\theta_1 \in A_1, \ldots, \theta_n \in A_n) &= \sum_{\Psi} \Pi_H(n_1, \ldots, n_k) \prod_{j=1}^{k} P(\cap_{i \in C_j} A_i) \\ &\to \sum_{\Psi} \Pi_\infty(n_1, \ldots, n_k) \prod_{j=1}^{k} P(\cap_{i \in C_j} A_i) = \text{pr}(\phi_1 \in A_1, \ldots, \phi_n \in A_n), \quad H \to \infty,\end{aligned}$$

with $(\theta_n)_{n \geq 1}$ and $(\phi_n)_{n \geq 1}$ being two exchangeable sequences directed by $\tilde{p}_H$ and $\tilde{p}_\infty$, respectively, and where the sum runs over the space of partitions with $\Psi = \{C_1, \ldots, C_k\}$. Using de Finetti representation theorem, the latter is equivalent to

$$E\left\{\prod_{i=1}^{n} \tilde{p}_H(A_i)\right\} \to E\left\{\prod_{i=1}^{n} \tilde{p}_\infty(A_i)\right\}, \qquad H \to \infty.$$

Since the above convergence holds for any collection of sets $A_1, \ldots, A_n$ and any $n \geq 1$, we can write equivalently

$$E\left\{\prod_{j=1}^{k} \tilde{p}_H(B_j)^{n_j}\right\} \to E\left\{\prod_{j=1}^{k} \tilde{p}_\infty(B_j)^{n_j}\right\}, \qquad H \to \infty, \tag{14}$$

for any measurable collection $B_1, \ldots, B_k$. The random vector $\{\tilde{p}_\infty(B_1), \ldots, \tilde{p}_\infty(B_k)\}$ is positive and bounded, thus being fully determined by its cross-moments. Hence, the vector version of Theorem 30.2 in Billingsley (1995) ensures that convergence of the cross-moments (14) implies the convergence in distribution, namely

$$\{\tilde{p}^{(H)}(B_1), \ldots, \tilde{p}^{(H)}(B_k)\} \to \{\tilde{p}^{(\infty)}(B_1), \ldots, \tilde{p}^{(\infty)}(B_k)\}, \quad H \to \infty,$$

in the sense of weak convergence. The weak convergence of the whole process is then guaranteed by Theorem 11.1.VII in Daley and Vere-Jones (2008).

Proof of Proposition 4

From Theorem 1 and Theorem 2 in Ishwaran and James (2001) it follows that

$$d_{\text{TV}}\left\{m_{H,\text{tr}}, m_{\infty}\right\} \leq 2\left[1-\left\{1-\frac{\left(\frac{\alpha}{\sigma}+1\right)_{H-1}}{\left(\frac{\alpha}{\sigma}+\frac{1}{\sigma}\right)_{H-1}}\right\}^{n}\right].$$

We will say that two sequences of real numbers $(a_H)_{H\geq 1}$ and $(b_H)_{H\geq 1}$ such that $\lim_{H\to\infty} a_H = \lim_{H\to\infty} b_H = 0$ are asymptotically equivalent, written $a_H \approx b_H$, if $\lim_{H\to\infty} a_H/b_H = 1$. The order of convergence stated in Proposition 4 is a direct consequence of the standard properties of the Gamma function. Indeed, note that

$$\frac{\left(\frac{\alpha}{\sigma}+1\right)_{H-1}}{\left(\frac{\alpha}{\sigma}+\frac{1}{\sigma}\right)_{H-1}} \approx \frac{\Gamma\left(\frac{\alpha}{\sigma}+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{\alpha}{\sigma}+1\right)} H^{-\frac{1}{\sigma}+1}, \quad H\to\infty,$$

from which it follows that

$$2\left[1-\left\{1-\frac{\left(\frac{\alpha}{\sigma}+1\right)_{H-1}}{\left(\frac{\alpha}{\sigma}+\frac{1}{\sigma}\right)_{H-1}}\right\}^{n}\right] \approx 2n\frac{\Gamma\left(\frac{\alpha}{\sigma}+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{\alpha}{\sigma}+1\right)} H^{-\frac{1}{\sigma}+1} = \mathcal{O}(H^{-\frac{1}{\sigma}+1}), \quad H\to\infty.$$

Moreover, following the same steps as for Theorem 4 in Ishwaran and Zarepour (2002) we get

$$\begin{aligned} d_{\text{TV}}\left\{m_H, m_{\infty}\right\} &\leq d_{\text{TV}}\{\text{pr}(\Psi_{n,H}), \text{pr}(\Psi_{n,\infty})\} \\ &\leq \frac{1}{2}\sum_{\Psi} \left|\Pi_H(n_1,\ldots,n_k) - \Pi_{\infty}(n_1,\ldots,n_k)\right| \\ &\leq \frac{1}{2}\sum_{\Psi} \Pi_{\infty}(n_1,\ldots,n_k)\left|1-\frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_{\infty}(n_1,\ldots,n_k)}\right|, \end{aligned}$$

where the sum runs over all the partitions Ψ of $[n]$ with cardinalities $n_1,\ldots,n_k$. From the proof of Theorem 1 and Theorem 2 it follows that

$$\frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_{\infty}(n_1,\ldots,n_k)} = \frac{H!}{H^k(H-k)!}\sum_{(\ell_1,\ldots,\ell_k)} \frac{\mathscr{V}_{n,|\ell^{(k)}|}}{\mathscr{V}_{n,k}} \frac{1}{H^{|\ell^{(k)}|-k}} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\sigma^{\ell_j}} \frac{1}{(1-\sigma)_{n_j-1}},$$

from which we obtain

$$\lim_{H\to\infty} H\left|1-\frac{\Pi_H(n_1,\ldots,n_k)}{\Pi_{\infty}(n_1,\ldots,n_k)}\right| = \sum_{j=1}^{k} \frac{\mathscr{V}_{n,k+1}}{\mathscr{V}_{n,k}} \frac{\mathscr{C}(n_j,2;\sigma)}{\sigma^2} \frac{1}{(1-\sigma)_{n_j-1}},$$

since the term not vanishing in the summation over $(\ell_1,\ldots,\ell_k)$ are those for which $\ell_1+\cdots+\ell_k = k+1$, meaning that each $\ell_j = 1$ but one equal to 2, and recalling also that $\mathscr{C}(n_j,1;\sigma) = \sigma(1-\sigma)_{n_j-1}$. Hence,

one get

$$d_{\text{TV}}\{\text{pr}(\Psi_{n,H}), \text{pr}(\Psi_{n,\infty})\} \approx \frac{1}{2H} \sum_{\Psi} \Pi_{\infty}(n_1, \ldots, n_k) \sum_{j=1}^{k} \frac{\mathscr{V}_{n,k+1}}{\mathscr{V}_{n,k}} \frac{\mathscr{C}(n_j, 2; \sigma)}{\sigma^2 (1-\sigma)_{n_j-1}} = \mathcal{O}\left(\frac{1}{H}\right),$$

which concludes the proof.

Gibbs sampling algorithm

We describe here a Gibbs sampling algorithm for posterior computation of the model described in the simulation study of Section 6, for a general kernel $\mathcal{K}(y;\theta)$. Let $S_i \in \{1, \ldots, H\}$ be an indicator function denoting to which mixture component each unit Y_i is allocated, for $i = 1, \ldots, n$. The symbol "$-$" denotes the conditioning to all the other variables whereas $f(\theta)$ represents the prior density associated to the baseline measure P. The Gibbs sampling algorithm alternates between the following full conditional steps.

Step 1. Update the cluster indicators $S_i \in \{1, \ldots, H\}$ from their full conditional categorical random variables

$$\text{pr}(S_i = h \mid -) = \frac{\pi_h \mathcal{K}(y_i; \tilde{\theta}_h)}{\sum_{h'=1}^{H} \pi_{h'} \mathcal{K}(y_i; \tilde{\theta}_{h'})}, \qquad h = 1, \ldots, H,$$

independently for any $i = 1, \ldots, n$.

Step 2. Update the within-cluster parameters $\tilde{\theta}_h$ independently from their full conditional distribution having density proportional to

$$f(\tilde{\theta}_h \mid -) \propto f(\tilde{\theta}_h) \prod_{i:S_i=h} \mathcal{K}(y_i; \tilde{\theta}_h)$$

for $h = 1, \ldots, H$. This is a standard step in mixture modeling and it is straightforward to simulate if conjugate priors are chosen.

Step 3. Update the weights $(\pi_1, \ldots, \pi_H)$ from their full conditional distribution, by exploiting the posterior characterization of Theorem 4. For example, if $\sigma = 0$ one has that

$$(\pi_1, \ldots, \pi_{H-1} \mid -) \sim \text{DIR}\left(\alpha/H + \sum_{i=1}^{n} I(S_i = 1), \ldots, \alpha/H + \sum_{i=1}^{n} I(S_H = 1)\right),$$

where $I(\cdot)$ denotes the indicator function. Note that the frequencies $n_1, \ldots, n_k$ in Theorem 4 corresponds to the positive values of the vector $\sum_{i=1}^{n} I(S_i = 1), \ldots, \sum_{i=1}^{n} I(S_i = H)$. In general, the full conditional of $(\pi_1, \ldots, \pi_H)$ will be obtained by first drawing the latent variables $(\ell_1, \ldots, \ell_k)$, as described in Section 4, and then by leveraging the quasi-conjugate representation of Theorem 4.

References

Aitchison, J. (1985). A general class of distributions on the simplex. *J. R. Statist. Soc. B 47*(1), 136–146.

Argiento, R. and M. De Iorio (2019). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *arXiv:1904.09733*.

Billingsley, P. (1995). *Probability and Measure* (Third ed.). New York: John Wiley & Sons.

Camerlenghi, F., A. Lijoi, P. Orbanz, and I. Prünster (2019). Distribution theory for hierarchical processes. *Ann. Statist. 47*(1), 67–92.

Camerlenghi, F., A. Lijoi, and I. Prünster (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scand. J. Statist. 45*, 1062–91.

Canale, A., D. Durante, and D. Dunson (2018). Convex mixture regression for quantitative risk assessment. *Biometrics 74*, 1331–40.

Canale, A. and I. Prünster (2017). Robustifying Bayesian nonparametric mixtures for count data. *Biometrics 73*(1), 174–84.

Carlton, M. A. (2002). A family of densities derived from the three-parameter Dirichlet process. *J. Appl. Prob. 39*, 764–74.

Charalambides, C. A. (2002). *Enumerative Combinatorics*. New York: Chapman and Hall / CRC.

Daley, D. J. and D. Vere-Jones (2008). *An Introduction to the Theory of Point Processes* (Second ed.), Volume II: General Theory and Structure. New York: Springer.

De Blasi, P., S. Favaro, A. Lijoi, R. H. Mena, I. Prünster, and M. Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pat. Anal. Mach. Intel. 37*(2), 212–29.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer.

Gnedin, A. and J. Pitman (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov, POMI 325*, 83–102.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *J. Am. Statist. Assoc. 96*(453), 161–73.

Ishwaran, H. and M. Zarepour (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika 87*(2), 371–90.

Ishwaran, H. and M. Zarepour (2002). Exact and approximate sum representation for the Dirichlet process. *Canad. J. Statist. 30*(2), 269–283.

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

Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *J. R. Statist. Soc. B 69*(4), 715–40.

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

Lijoi, A., I. Prünster, and S. G. Walker (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Prob. 18*(4), 1519–47.

Longnecker, M. P., M. A. Klebanoff, H. Zhou, and J. W. Brock (2001). Association between maternal serum concentration of the DDT metabolite DDE and preterm and small-for-gestational-age babies at birth. *Lancet 358*, 110–4.

Malsiner-Walli, G., S. Frühwirth-Schnatter, and B. Grün (2016). Model-based clustering based on sparse finite Gaussian mixtures. *Statist. Comp. 26*(1-2), 303–24.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *J. Am. Statist. Assoc. 113*(521), 340–56.

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

Pitman, J. (1996). Some developments of the Blackwell-Macqueen urn scheme. In T. S. Ferguson, L. S. Shapley, and J. B. MacQueen (Eds.), *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, Volume 30 of *IMS Lecture notes, Monograph Series*, pp. 245–67. Hayward: Institute of Mathematical Statistics.

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

Ramsay, J. O. (1988). Monotone regression splines in action. *Statist. Sc. 3*(4), 425 – 61.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B 59*(4), 731–92.

Ridout, M. S. (2009). Generating random numbers from a distribution specified by its Laplace transform. *Statist. Comp. 19*(4), 439–50.

Rousseau, J. and K. Mengersen (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Statist. Soc. B 73*(5), 689–710.