Bayesian Analysis (2021) **TBA**, Number TBA, pp. 1–25

Perfect Sampling of the Posterior in the Hierarchical Pitman–Yor Process*

Sergio Bacallado†, Stefano Favaro‡,§, Samuel Power¶, and Lorenzo Trippa‖

Abstract. The predictive probabilities of the hierarchical Pitman–Yor process are approximated through Monte Carlo algorithms that exploits the Chinese Restaurant Franchise (CRF) representation. However, in order to simulate the posterior distribution of the hierarchical Pitman–Yor process, a set of auxiliary variables representing the arrangement of customers in tables of the CRF must be sampled through Markov chain Monte Carlo. This paper develops a perfect sampler for these latent variables employing ideas from the Propp–Wilson algorithm and evaluates its average running time by extensive simulations. The simulations reveal a significant dependence of running time on the parameters of the model, which exhibits sharp transitions. The algorithm is compared to simpler Gibbs sampling procedures, as well as a procedure for unbiased Monte Carlo estimation proposed by Glynn and Rhee. We illustrate its use with an example in microbial genomics studies.

Keywords: Bayesian nonparametrics, Gibbs sampling, hierarchical Pitman–Yor process, perfect sampling, species sampling, unbiased Monte Carlo estimation.

1 Introduction

The hierarchical Pitman–Yor process was introduced in Teh et al. (2006) and Teh (2006) as a nonparametric prior model for a collection of discrete distributions with heavy tails. See Teh and Jordan (2010) for a review on hierarchical nonparametric priors. In Bayesian nonparametrics, theoretical developments and applications of the hierarchical Pitman–Yor process have been considered in language modeling (Teh, 2006; Huang and Renals, 2007; Wood et al., 2009), infinite hidden Markov modeling (Beal et al., 2002; Van Gael et al., 2008; Blunsom and Cohn, 2011), species sampling with multiple populations (Battiston et al., 2018; Camerlenghi et al., 2019; Bassetti et al., 2020; Camerlenghi et al., 2019), clustering (Argiento et al., 2020), graphical modeling (Creamschi et al.,

*Sergio Bacallado and Samuel Power received funding from the *Cantab Capital Institute for the Mathematics of Information*. Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018–2022. Lorenzo Trippa has been supported by the *Claudia Adams Barr Program in Cancer Research* and the NIH grant 1R01LM013352-01A1.

†Statistical Laboratory, University of Cambridge, sergiobacallado@gmail.com

‡Department of Economics and Statistics, University of Torino and Collegio Carlo Alberto, stefano.favaro@unito.it

§Also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy)

¶Statistical Laboratory, University of Cambridge, sp825@cam.ac.uk

‖Department of Biostatistics, Dana–Farber Cancer Institute and Harvard School of Public Health, ltrippa@jimmy.harvard.edu

 https://doi.org/10.1214/21-BA1269

2020), image segmentation (Sudderth and Jordan, 2009), and topic models (Sato and Nakagawa, 2010; Araki et al., 2012; Lindsey et al., 2012). In this paper we evaluate and compare multiple computational strategies for posterior inference under the hierarchical Pitman–Yor process prior. In particular, we discuss: i) a novel conditional Gibbs sampling; ii) an application of coupling from the past for perfect sampling; iii) the application of a more recent framework studied by Glynn and Rhee (2014) with the aim of deriving unbiased posterior estimates from Markov chain sampling. The latter approach can be viewed as an intermediate solution between Gibbs sampling and perfect simulations.

1.1 The Pitman–Yor process

The Pitman–Yor process (Pitman, 1995) is a discrete random probability measure whose distribution depends on two parameters (β, θ), with $\beta \in [0, 1)$ and $\theta > 0$, and a probability measure ν. The parameter β is usually indicated as the *concentration parameter*, the parameter θ is referred to as the *mass parameter*, and the measure ν is the *base measure*. Among the various possible definitions, a simple and intuitive one follows from the so-called stick-breaking construction (Pitman, 2006). Let $(v_i)_{i\geq 1}$ be independent random variables such that v_i is distributed as a Beta distribution with parameter $(\theta+i\beta, 1-\beta)$, and define $\pi_1 = v_1$ and $\pi_i = v_i \prod_{1\leq j\leq i-1}(1-v_j)$ for $i \geq 1$. Also, let $(y_i)_{i\geq 1}$ be independent random variables with distribution ν. The discrete random probability measure

$$\mu = \sum_{i\geq 1} \pi_i \delta_{y_i},$$

where δ_y is the point mass at y, is a Pitman–Yor process with concentration β, mass θ and base distribution ν. For brevity we write $\mu \sim PY(\nu, \theta, \beta)$. The Dirichlet process (Ferguson, 1973) is recovered as a special case of the Pitman–Yor process for $\beta = 0$. See Sethuraman (1994).

Discrete random probability measures play a fundamental role in Bayesian nonparametrics, since their laws act as nonparametric priors for discrete distributions (Lijoi and Prünster, 2010). The Pitman–Yor process is arguably one of the most popular priors. In particular, several Bayesian nonparametric models include a collection of random variables $(X_1, \ldots, X_n)$ either observed or latent, from a Pitman–Yor process. That is,

$$\begin{aligned} X_i \mid \mu &\overset{iid}{\sim} \mu, \quad i = 1, 2, \ldots, \\ \mu &\sim PY(\nu, \theta, \beta). \end{aligned}$$

Here $(X_1, \ldots, X_n)$ are the first n coordinates of an exchangeable sequence $(X_i)_{i\geq 1}$ whose de Finetti measure (or prior) on the unknown distribution μ is the law of the Pitman–Yor process. Because of the discreteness of μ, $(X_1, \ldots, X_n)$ from μ presents $k \leq n$ distinct types, labelled by $X_1^*, X_2^*, \ldots, X_k^*$ with frequencies $(n_1, \ldots, n_k)$ such that $n_i \geq 1$ and $\sum_{1\leq i\leq k} n_i = n$. Here X_i^* is the i-th distinct value that appears in $(X_i)_{i\geq 1}$. For example if $X_1 = X_2$ and $X_2 \neq X_3$ then $X_2^* = X_3$. The number of distinct values k increases with the sample size n; we can therefore use the notation $k(n)$ if necessary to make the relationship explicit.

The predictive probabilities induced by the Pitman–Yor process (Pitman, 1995), i.e. the conditional distribution of X_{n+1} given $(X_1, \ldots, X_n)$, have the following explicit form

$$X_{n+1} \mid X_1, \ldots, X_n \sim \frac{\theta + k(n)\beta}{n+\theta}\nu + \sum_{i=1}^{k(n)} \frac{n_i - \beta}{n+\theta}\delta_{X_i^*} \tag{1.1}$$

for $n \geq 1$. Here to simplify the presentation we are assuming that ν is nonatomic, i.e. $\nu(\{x\}) = 0$ for every singleton $\{x\}$. The Chinese Restaurant Process (CRP) of Pitman (1995) gives an intuitive metaphorical description of the predictive probability (1.1). In particular, consider a sequence of customers entering a restaurant and sitting at various tables, each table serving a single dish. Customers select their table through a reinforced urn scheme, with balls sequentially drawn. The probability of selecting a ball is proportional to its weight. Initially the restaurant is empty and the urn contains only a black ball with weight θ. Customers seat sequentially to various tables accordingly to the following scheme. Whenever the black ball is selected, the next customer sits in a new table, and the dish served on this table is sampled from ν. In this case a new ball labeled by the table with weight $1 - \beta$ is added to the urn, and the weight of the black ball is increased by β. When, instead, a ball labelled by a table is selected, the next customer sits at the corresponding table, and the weight of the ball is increased by 1.

The parameter β has a critical role since it tunes the rate at which new dishes are generated. Indeed, when a new dish ($X_n \neq X_i,\ i = 1, \ldots, n-1$) is generated, a reinforcement equal to β is assigned to the continuous component ν (expression (1.1)), and it affects the probability of generating further new dishes. The larger β the stronger the reinforcement mechanism, which is absent in the Dirichlet process ($\beta = 0$). In different words, the expected number of unique dishes $k(n)$ increases with the value of the parameter β.

1.2 The hierarchical Pitman–Yor process

The hierarchical Pitman–Yor process is an extension of the Pitman–Yor process. It defines a nonparametric prior model for a collection of discrete distributions by means of a hierarchy of Pitman–Yor processes. Precisely, for any $R \geq 1$, the hierarchical Pitman–Yor process is a collection of dependent discrete random probability measures $(\mu_1, \ldots, \mu_R)$ defined as

$$\begin{aligned} \mu_r \mid \mu &\overset{iid}{\sim} PY(\mu, \theta, \beta), && \text{for } r = 1, \ldots, R, \\ \mu &\sim PY(\nu, \theta_0, \beta_0). \end{aligned}$$

The dependence among the random probability measures μ_r is induced by the common base measure μ. The hierarchical Dirichlet process (Teh et al., 2006) is recovered when $\beta = 0$ and $\beta_0 = 0$. As a generalization of the previous Bayesian nonparametric construction, let $\{(X_{r,1}, \ldots, X_{r,n_r})\}_{r=1,\ldots,R}$ be samples from the hierarchical Pitman–Yor process, i.e.,

$$X_{r,i} \mid \mu_1, \ldots, \mu_R, \mu \overset{iid}{\sim} \mu_r, \qquad \text{for } i = 1, \ldots, n_r \text{ and } r = 1, \ldots, R,$$

$$\begin{aligned} \mu_r \mid \mu &\overset{iid}{\sim} PY(\mu, \theta, \beta), && \text{for } r = 1, \dots, R, \\ \mu &\sim PY(\nu, \theta_0, \beta_0). \end{aligned}$$

By construction, conditional on the Pitman–Yor process μ, the μ_r's are independent Pitman–Yor processes. Here $(X_{r,1}, \dots, X_{r,n_r}; r = 1, \dots, R)$ includes the first n_r coordinates, for $r = 1, \dots, R$, of a partially exchangeable array $(X_{r,i})_{r \geq 1, i \geq 1}$ whose de Finetti (or prior) measure on the unknown $(\mu_1, \dots, \mu_R)$ is the law of the hierarchical Pitman–Yor process.

The predictive probabilities induced by the hierarchical Pitman–Yor process can be described by means of the Chinese Restaurant Franchise (CRF), which extends the CRP. In particular, each sample $(X_{r,1}, \dots, X_{r,n_r})$ identifies the dishes of the n_r customers in restaurant r, for $r = 1, \dots, R$. Customers seating at the same table eat the same dish and, due to the discreteness of the common base measure μ, the same dish can be served at multiple tables within the same restaurant and in different restaurants. The assignment of the n_r customers to different tables is identical as in the CRP, and the assignment of customers to tables is independent in the R restaurants. Conditionally on μ, each table in the R restaurants is assigned a dish sampled from μ. We denote by I the number of distinct dishes, i.e. distinct values in the finite array $(X_{r,1}, \dots, X_{r,n_r}; r = 1, \dots, R)$, labelled by $\{X_1^*, \dots, X_I^*\}$, across the R restaurants, and by $n_{r,i} \geq 0$ the number of customers in restaurant r eating dish X_i^*. We can arbitrarily assign the indices $i = 1, \dots, I$ to the I distinct dishes in $(X_{r,1}, \dots, X_{r,n_r}; r = 1, \dots, R)$. Furthermore, we introduce a variable $k_{r,i}$ for the number of tables in restaurant r serving the dish X_i^*.

The distribution of X_{r,n_r+1}, for any index $1 \leq r \leq R$, conditioning on (i) $(X_{r,1}, \dots, X_{r,n_r}; r = 1, \dots, R)$, (ii) the number of tables in restaurant r occupied by the first n_r customers, $k_{r,\cdot} = \sum_i k_{r,i}$, and (iii) the number of tables occupied in the CRF, $k = \sum_1^R k_{r,\cdot}$ and (iv) $k_{\cdot,i} = \sum_i k_{r,i}$, can be represented as

$$\frac{\theta + \beta k_{r,\cdot}}{\theta + n_r} \left(\frac{\theta_0 + \beta_0 k}{\theta_0 + k} \nu + \sum_{i=1}^{I} \frac{k_{\cdot,i} - \beta_0}{\theta_0 + k} \delta_{X_i^*} \right) + \sum_{i : n_{r,i} > 0} \frac{n_{r,i} - k_{r,i}\beta}{\theta + n_r} \delta_{X_i^*}. \tag{1.2}$$

We refer to the recent works of Camerlenghi et al. (2019) and Bassetti et al. (2020) for a detailed study of the predictive probability (1.2) and related results.

1.3 Contributions

Teh et al. (2006) proposed a strategy for sampling from the posterior distribution of the hierarchical Dirichlet process and from the posterior distribution of the hierarchical Pitman–Yor process. In particular, they proposed a *collapsed* Gibbs sampler, which marginalizes out the random probability measures μ and $(\mu_1, \dots, \mu_R)$. Let $t_{r,j}$ be the index of the table assigned to customer j in restaurant r. The target of the Gibbs sampler is the joint distribution of $\{(t_{r,j}, \dots, t_{r,n_r})\}_{r=1,\dots,R}$ conditional on the observations $\mathbf{X} = \{(X_{r,1}, \dots, X_{r,n_r})\}_{r=1,\dots,R}$. Teh et al. (2006) also proposed a *conditional* Gibbs sampler which augments the set of latent variables with the measure μ. Van Gael et al. (2008) and Papaspiliopoulos and Roberts (2008) define similar conditional samplers which

augment the sample space with the random probability measures μ and $(\mu_1, \ldots, \mu_R)$, in the special case $\beta = \beta_0 = 0$.

In this paper we discuss the problem of posterior sampling from the hierarchical Pitman–Yor process. In particular, we employ an augmentation strategy which specifies the state of the urns in the CRF. We start by describing three Gibbs sampling algorithms, two of them based on the collapsed and conditional samplers of Teh et al. (2006), and a novel Gibbs algorithm which we call *doubly conditional* as it employs a larger augmentation. Then, as an alternative to Gibbs sampling, we propose a perfect sampling algorithm based on the coupling from the past method of Propp and Wilson (1996). Perfect sampling allows to perform posterior inference and to compute Monte Carlo approximations of predictive probabilities. Finally, as an intermediate solution between the Gibbs sampling algorithms and perfect simulations, we consider the application of a procedure for unbiased posterior estimation introduced by Glynn and Rhee (2014).

We present an evaluation of the convergence of the proposed Gibbs samplers and discuss the running time of the perfect sampling procedure. While perfect sampling yields exact samples from the posterior distribution, the algorithm has a random stopping time and it may not be practical in certain situations. An extensive simulation study is presented, revealing a dependence of the running time on the parameters (θ, β) and (θ_0, β_0) of the hierarchical Pitman–Yor process prior, with some parameter settings making it prohibitively time consuming to draw samples from the posterior. In contrast, the simulation-based evaluations of the doubly conditional Gibbs sampler suggest that the computing time to approximate the posterior does not vary substantially across parameterizations of the prior.

The paper is structured as follows. Section 2 contains the main contributions of the paper: three Gibbs sampling algorithms, the perfect sampling algorithm, and the procedure to compute unbiased approximations of posterior estimates. In Section 3 we present the simulation study for evaluating the convergence of the Gibbs samplers and the running time of the perfect sampling algorithm. Section 4 contains an application in microbial genomics. Section 5 concludes the paper with a discussion of related problems and open questions.

2 Posterior sampling from the hierarchical Pitman–Yor process

In the CRF metaphor, we have that: i) I denotes the number of dishes across the R samples $\mathbf{X} = \{(X_{r,1}, \ldots, X_{r,n_r})\}_{r=1,\ldots,R}$; ii) $n_{r,i} \geq 0$ denotes the number of customers in restaurant r and eating dish i; iii) $k_{r,i} \in \{1, \ldots, n_{r,i}\}$ denotes the number of tables serving dish i in restaurant r, and $k_{r,i} = 0$ if $n_{r,i} = 0$. Moreover, we use the notation $k_{\cdot,i} = \sum_r k_{r,i}$, $k_{r,\cdot} = \sum_i k_{r,i}$, and $k = \sum_{r,i} k_{r,i}$. We denote by $(x)_{n\uparrow a}$ the generalized factorial of $x \geq 0$ of order $n \in \mathbb{N}$ and increment $a \geq 0$, that is $(x)_{n\uparrow a} = \prod_{0\leq i\leq n-1}(x+ai)$ with the proviso $(x)_{0\uparrow a} = 1$.

We observe that, according to the predictive probabilities (1.2), the summaries

$\mathbf{n} = \{(n_{r,1}, \ldots, n_{r,I})\}_{r=1,\ldots,R}$ and $\mathbf{k} = \{(k_{r,1}, \ldots, k_{r,I})_{i\geq 1}\}_{r=1,\ldots,R}$ allow one to straightforwardly predict the dish assigned to future customers of the CRF. In particular in the CRF the joint probability of R partitions of customers into groups with sizes $\mathbf{n}$ allocated at $\mathbf{k}$ tables is

$$\frac{(\theta_0)_{I\uparrow\beta_0} \prod_i (1-\beta_0)_{k_{\cdot,i}-1\uparrow 1} \prod_{r,i} f(n_{r,i}, k_{r,i}) \prod_r (\theta)_{k_{r,\cdot}\uparrow\beta}}{(\theta_0)_{k\uparrow 1} \prod_r (\theta)_{n_r\uparrow 1}}, \tag{2.1}$$

where

$$f(n,s) = \sum_{\gamma \in B_{n,s}} \prod_{j=1}^{n-s} \gamma_j \tag{2.2}$$

and $\gamma \in B_{n,s} \subset \mathbb{R}^{n-s}$ satisfies $\gamma_j = (j-1) + \alpha_j(1-\beta)$, for integers α_i's such that $1 = \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{n-s} \leq s$. Precisely, the expression in (2.1) is obtained by summing, for every pair (r,i), over all possible times at which new tables in restaurant r serving dish i are created. Then the only factors depending on this order are gathered in $f(n_{r,i}, k_{r,i})$. The triangular array defined in equation (2.2) satisfies the following recursion

$$f(n,s) = \sum_{i=1}^{s} [(n-s-1) + i(1-\beta)] f(n-s+i-1, i), \tag{2.3}$$

for $n \neq s$, with $f(n,n) = 1$. This fact can be verified by induction using the definition of the function f in equation (2.2). We remark that the sampling algorithms in this section require memorizing the triangular array defined in equation (2.2), and our implementations compute it through the recursion 2.3. We wish to sample the conditional distribution

$$\begin{aligned} p(\mathbf{k} \mid \mathbf{n}) &\propto \frac{\prod_i (1-\beta_0)_{k_{\cdot,i}-1\uparrow 1} \prod_{r,i} f(n_{r,i}, k_{r,i}) \prod_r (\theta)_{k_{r,\cdot}\uparrow\beta}}{(\theta_0)_{k\uparrow 1}} \\ &\propto \frac{\prod_i (1-\beta_0)_{k_{\cdot,i}-1\uparrow 1} \prod_{r,i} f(n_{r,i}, k_{r,i}) \prod_r (\theta/\beta)_{k_{r,\cdot}\uparrow 1} \beta^{k_{r,\cdot}}}{(\theta_0)_{k\uparrow 1}}. \end{aligned} \tag{2.4}$$

2.1 Gibbs sampling algorithms

Markov chain Monte Carlo algorithms for the hierarchical Dirichlet process have been designed for models in which the dishes in the CRF analogy are not observed. In particular, dishes are assumed latent variables which determine, for example, component membership in a mixture model. Here, we focus on a different setting where the dishes are observed, which makes it difficult to directly compare with algorithms previously proposed in the literature. However, Markov chains in both cases can be characterized as either collapsed or conditional samplers, depending on whether the algorithm augments the sample space by the latent probability measure μ or not. We first define a Gibbs sampling algorithm of each type, and then introduce a third Gibbs sampler with a larger augmentation.

Collapsed Gibbs sampler

One of the Gibbs sampling algorithms originally proposed by Teh et al. (2006) augments the sample space with dish assignments, which we assume to be observed, and a variable capturing the number of tables for each dish, in each restaurant, which is equivalent to the array $\mathbf{k}$ defined in the previous section. In particular, each iteration of the algorithm samples the distribution $p(k_{r,i} \mid \{k_{r',i'}; (r',i') \neq (r,i)\}, \mathbf{n})$ for each entry $k_{r,i}$ in the array $\mathbf{k}$. Then, the conditional distribution is known up to a constant. Specifically, we can write

$$p(k_{r,i} \mid \{k_{r',i'}; (r',i') \neq (r,i)\}, \mathbf{n}) \propto \frac{(1-\beta_0)_{k_{\cdot,i}-1\uparrow 1} f(n_{r,i}, k_{r,i})(\theta/\beta)_{k_{r,\cdot}\uparrow 1}\beta^{k_{r,\cdot}}}{(\theta_0)_{k\uparrow 1}},$$

for $k_{r,i} \in \{1, \ldots, n_{r,i}\}$ when $n_{r,i} > 0$.

Conditional Gibbs sampler

A different algorithm discussed by Teh et al. (2006) augments the space with the latent probability measure μ. The key fact is that the distribution of μ given $\mathbf{k}$ can be sampled retrospectively (Papaspiliopoulos and Roberts, 2008). That is, one only needs to sample the mass assigned to the atoms corresponding to the observed dishes in the probability measure μ, which we denote by $\mu(1), \ldots, \mu(I)$, and this can be done through the stick-breaking representation of the Dirichlet process (Sethuraman, 1994). The same strategy can be adopted in the context of the hierarchical Pitman–Yor process. In particular, conditionally on $(\mathbf{k}, \mathbf{n})$, the vector $(\mu(1), \ldots, \mu(I))$ is identically distributed to BD, where $B \sim \text{Beta}(k - I\beta_0, \theta_0 + I\beta_0)$ and $D \sim \text{Dirichlet}(k_{\cdot,1} - \beta_0, \ldots, k_{\cdot,I} - \beta_0)$ are independent. Then, conditionally on $(\mu, \mathbf{n})$ each row of $\mathbf{k}$ is independent and has distribution

$$p(k_{r,1}, \ldots, k_{r,I} \mid \mathbf{n}, \mu) \propto (\theta)_{k_{r,\cdot}\uparrow\beta} \prod_{i=1}^{I} f(n_{r,i}, k_{r,i})\mu(i)^{k_{r,i}}$$

for $r = 1, \ldots, R$. It is worth pointing out that, as the support of this distribution could be very large, it is convenient to sample one entry $k_{r,i}$ at a time conditional on all the others.

Doubly conditional Gibbs sampler

We introduce a novel Gibbs sampling algorithm, which augments the sample space by two vectors, one indexed by restaurants and one indexed by dishes. To eliminate the rising factorials in equation (2.4) for our target distribution $p(\mathbf{k} \,|\mathbf{n})$, here we apply a standard augmentation trick for normalized completely random probability measures (Favaro and Teh, 2013). In particular, we introduce independent random variables $(G_1, \ldots, G_R)$,

$$G_r \sim \text{Gamma}(\theta/\beta, 1)$$

for $r = 1, \dots, R$. Also, let

$$(D_1, \dots, D_I, D_+) \sim \text{Dirichlet}(1-\beta_0, \dots, 1-\beta_0, \theta_0 + I\beta_0)$$

be independent of $(G_1, \dots, G_R)$. Let p_G denote the distribution of $G = (G_1, \dots, G_R)$ and let p_D denote the distribution of $D = (D_1, \dots, D_I, D_+)$. Then, we consider the distribution

$$q(\mathbf{k}, d, g) \propto \prod_i \frac{1}{d_i} \prod_{r,i} f(n_{r,i}, k_{r,i}) (g_r d_i \beta)^{k_{r,i}} p_G(g) p_D(d). \tag{2.5}$$

The random variables $\mathbf{k}$ have the same support in (2.5) and in (2.4). Moreover, by integrating the right-hand side with respect to $g = (g_1, \dots, g_R)$ and $d = (d_1, \dots, d_I, d_+)$ we obtain

$$\begin{aligned}
&\int \prod_i \frac{1}{d_i} \prod_{r,i} f(n_{r,i}, k_{r,i}) (g_r d_i \beta)^{k_{r,i}} p_G(g) p_D(d) \mathrm{d}g \mathrm{d}d \qquad (2.6)\\
&\quad \propto \beta^k \prod_{r,i} f(n_{r,i}, k_{r,i}) \prod_r (\theta/\beta)_{k_{r,\cdot}\uparrow 1} \int \prod_i d_i^{k_{\cdot,i}-1} p_D(d) \mathrm{d}d \\
&\quad \propto \beta^k \prod_{r,i} f(n_{r,i}, k_{r,i}) \prod_r (\theta/\beta)_{k_{r,\cdot}\uparrow 1} \frac{\prod_i (1-\beta_0)_{(k_{\cdot,i}-1)\uparrow 1}}{(I+\theta_0)_{(k-I)\uparrow 1}} \\
&\quad \propto p(\mathbf{k} \mid \mathbf{n}).
\end{aligned}$$

This line of reasoning leads to the following Gibbs sampling algorithm for the distribution $q(\mathbf{k}, d, g)$ and for our target distribution $p(\mathbf{k} \mid \mathbf{n})$. First, conditionally on $\mathbf{k}$, the vectors d and g in (2.5) are Dirichlet distributed and Gamma distributed, respectively. Moreover, they are independent vectors. Second, conditionally on d and g, the variables $k_{r,i}$ in (2.5) are independent. In particular, since the random variables $k_{r,i}$ are discrete variables with a finite support, it is straightforward to sample their conditional distribution.

2.2 Perfect sampler

We start by introducing a collection of independent random variables (U, E, E', G_0, G) which will be used to define multiple coupled Markov chains. Let $U_{m,r,i}$ be independent random variables identically distributed as a Uniform distribution on $(0, 1)$, for $m \in \mathbb{Z}$, $1 \le i \le I$, $1 \le r \le R$. The index m will indicate time in the coupled Markov chains, while indices i and r indicate dishes and restaurants, as in previous sections. The random variables $U_{m,r,i}$ are combined with the variables $E_{m,\ell,h}$ and $E'_{m,\ell,h}$, which are independent and have Exponential distribution with parameter 1, for $m, \ell, h \in \mathbb{Z}$. Moreover, let $G_{0,\ell,m}$ be independent Gamma random variables with parameters $(1-\beta_0, 1)$, for $\ell, m \in \mathbb{Z}$, and let $G_{\ell,m}$ be independent Gamma random variables Gamma with parameters $(\theta/\beta, 1)$, for $\ell, m \in \mathbb{Z}$. For any $a > 0$, we define the distribution

$$p_a(\mathbf{k}) = Z(a) \prod_{r,i} f(n_{r,i}, k_{r,i}) \frac{\beta^k \prod_r (\theta/\beta)_{k_{r,\cdot}\uparrow 1} \prod_i (1-\beta_0)_{k_{\cdot,i}-1\uparrow 1}}{a^k},$$

where the support of $\mathbf{k}$ is the same as in equation (2.4), and $Z(a)$ is a normalization constant. We define an augmentation of this distribution similar to (2.5),

$$p_{a,D,G}(\mathbf{k}, g_0, g) \propto \prod_{r,i} f(n_{r,i}, k_{r,i}) \frac{\prod_r (g_r \beta)^{k_{r\cdot}} \prod_i (g_{0,i})^{k_{\cdot,i}-1}}{a^k} p_G(g) p_{G_0}(g_0), \tag{2.7}$$

where

i) p_G indicates the distribution of independent Gamma random variables with parameter $(\theta/\beta, 1)$;

ii) p_{G_0} indicates the distribution of independent Gamma random variables with parameter $(1-\beta_0, 1)$.

Integrating (2.7) with respect to g and g_0 yields $p_a(\mathbf{k})$. In (2.7) the array entries $k_{r,i}$ are independent conditional on G and G_0, which leads to a natural Gibbs sampler similar to the doubly conditional algorithm of the previous section. We will apply coupling from the past to sample from $p_a(\mathbf{k})$ exactly, and subsequently extend the procedure in order to sample $p(\mathbf{k} \mid \mathbf{n})$.

We construct a coupling of Gibbs samplers where we generate each Markov transition by using the set of random numbers (U, E, E', G_0, G). Let $\phi_{a,r,i}^{g_0,g}$ be the inverse marginal cumulative distribution function of $k_{r,i}$ after we condition on (g_0, g) in (2.7). For $j \leq m$, define

$$k_{r,i}^{j,m} = \begin{cases} \phi_{a,r,i}^{G_{0,a}^{j,m-1}, G_a^{j,m-1}}(U_{m,r,i}) & \text{if } j < m, \\ k_{r,i}^{\text{init}} & \text{if } j = m, \end{cases} \tag{2.8}$$

where $k_{r,i}^{\text{init}}$ are fixed integers and

$$G_{0,a}^{j,m-1} = \left(\left[G_{0,1,m} + \sum_{h=1}^{k_{\cdot 1}^{j,m-1}-1} E_{m,1,h} \right], \ldots, \left[G_{0,I,m} + \sum_{h=1}^{k_{\cdot I}^{j,m-1}-1} E_{m,I,h} \right] \right), \tag{2.9}$$

$$G_a^{j,m-1} = \left(\left[G_{1,m} + \sum_{h=1}^{k_{1\cdot}^{j,m-1}} E'_{m,1,h} \right], \ldots, \left[G_{R,m} + \sum_{h=1}^{k_{R\cdot}^{j,m-1}} E'_{m,R,h} \right] \right). \tag{2.10}$$

In the above construction, the sequences $(\mathbf{k}^{j,j}, \mathbf{k}^{j,j+1}, \mathbf{k}^{j,j+2}, \ldots)$, where $\mathbf{k}^{j,m} = \{k_{r,i}^{j,m};\ r = 1, \ldots, R,\ i = 1, \ldots, I\}$, defined by equation (2.8), equation (2.9) and equation (2.10), for distinct values of j are copies of the same Gibbs Markov chain with stationary distribution $p_a(\mathbf{k})$. In particular, the Markov chains are coupled through the use of common random numbers (U, E, E', G_0, G). Furthermore, this coupling is monotone with respect to the following partial order: $\mathbf{k} \succeq \tilde{\mathbf{k}}$ if and only if $k_{r,i} \geq \tilde{k}_{r,i}$ for all $r = 1, \ldots, R$, and $i = 1, \ldots, I$. That is, $\mathbf{k}^{j,m} \succeq \mathbf{k}^{j',m}$ implies $\mathbf{k}^{j,m+1} \succeq \mathbf{k}^{j',m+1}$ with probability 1.

Following the usual coupling from the past construction, we define two arrays $\overline{\mathbf{k}^{j,m}}$ and $\underline{\mathbf{k}^{j,m}}$ for $j \geq m$, through the recursive equations (2.8)–(2.10) where, in the first case, the initial state $\mathbf{k}^{\text{init}}$ is set to the maximum of the partial order, $k_{r,i}^{\text{init}} = n_{r,i}$ for all r, i, and in the second case to the minimum, $k_{r,i}^{\text{init}} = \mathbb{1}(n_{r,i} > 0)$ for all r, i. Theorem A.1 guarantees that

$$\lim_{j\to\infty} \overline{\mathbf{k}^{-j,0}} = \lim_{j\to\infty} \underline{\mathbf{k}^{-j,0}},$$

almost surely, and the distribution of this limit is $p_a(\mathbf{k})$. This theorem is based on the coupling from the past approach introduced in the work of Propp and Wilson (1996). We shall denote by $\mathbf{k}^{(a)} = \{k_{r,i}^{(a)};\ r = 1, \ldots, R,\ i = 1, \ldots, I\}$ the limiting random element with distribution $p_a(\mathbf{k})$.

In the described construction $\mathbf{k}^{(a)}$ is a function of (U, E, E', G_0, G), and if $a' \geq a$, $p(\mathbf{k}^{(a)} \succeq \mathbf{k}^{(a')}) = 1$. Therefore, if $e_0 \sim \text{Gamma}(\theta_0, 1)$, $(E_i^*)_{i\geq 1} \overset{iid}{\sim} \text{Exponential}(1)$, and $E_j^{**} = e_0 + \sum_{i=1}^{j} E_i^*$, there is at most one integer H such that

$$k^{(E_H^{**})} = H, \tag{2.11}$$

where $k^{(E_H^{**})} = \sum_{r,i} k_{r,i}^{(E_H^{**})}$. This is because the right hand side is strictly increasing in H and the left hand side is a.s. monotone decreasing in H. Define the random variable H as the solution to this equation, when it exists, with $H = -1$ when it does not. Then, the distribution of $\mathbf{k}^{(E_H^{**})}$ conditional on the event that the equation above has a solution matches the target posterior $p(\mathbf{k} \mid \mathbf{n})$. Indeed we can write the following

$$\begin{aligned}
\Pr(\mathbf{k}^{(E_H^{**})} = \mathbf{k} \mid H > -1) &\propto \int_0^\infty \Pr(E_k^{**} = a)\Pr(\mathbf{k}^{(a)} = \mathbf{k})\mathrm{d}a \\
&= \int_0^\infty \frac{a^{k+\theta_0-1}e^{-a}}{\Gamma(k+\theta_0)} \\
&\times Z(a) \prod_{r,i} f(n_{r,i}, k_{r,i}) \frac{\beta^k \prod_r (\theta/\beta)_{k_{r,\cdot}\uparrow 1} \prod_i (1-\beta_0)_{k_{\cdot,i}-1\uparrow 1}}{a^k} \mathrm{d}a \\
&\propto p(\mathbf{k} \mid \mathbf{n}).
\end{aligned}$$

Algorithm 1 provides a schematic view of a perfect sampler for the distribution $p(\mathbf{k} \mid \mathbf{n})$. The pseudocode outputs one sample of the posterior distribution. Note that the inner while loop terminates when we find an integer H which satisfies equation (2.11), in which case we return the array $\mathbf{k}^{(E_H^{**})}$, or alternatively, when we can verify that the equation has no integer solution, in which case we restart the outer loop. The routine called in Line 7 computes, for the given value of a, the arrays $\overline{\mathbf{k}^{-j,0}}$ and $\underline{\mathbf{k}^{-j,0}}$ for increasing values of $j \in \{2, 2^2, 2^3, \ldots\}$, until they become equal and therefore converge to $\mathbf{k}^{(a)}$. The computational cost of each iteration of the inner while loop is $\mathcal{O}(RIj_{\max})$, where R is the number of restaurants, I is the number of dishes, and $j_{\max}$ is the number of steps of the coupling required in Line 7. The next section will quantify this computational burden by simulations. Line 3 is purely schematic, as E, E', U, G_0, G are infinite arrays. These pseudorandom numbers may be memorized or recomputed from a random seed as needed, which ensures that memory requirements do not increase with $j_{\max}$.

Algorithm 1 Perfect sampler for the hierarchical Pitman–Yor process.

Input array $\mathbf{n}$
while true do
  Sample arrays $E, E', U, G_0, G, E^*, e_0$
  $H \leftarrow n$
  while true do
    $a \leftarrow E_H^{**}$
    Simulate coupling from the past using E, E', U, G_0, G to obtain $\mathbf{k}^{(a)}$
    if $k^{(a)} = H$ **then**
      return $\mathbf{k}^{(a)}$
    else if Verified $k^{(E_H^{**})} \neq H$ for all $H \in \mathbb{N}$ **then**
      break while
    else
      Set H to a new higher or lower value depending on whether $k^{(a)} > H$ or $k^{(a)} < H$.
    end if
  end while
end while

2.3 Unbiased estimation through Markov chain couplings

The inference objective in a Bayesian analysis is usually a posterior moment of the form $h^* = \mathbb{E}_{\mathbf{k}|\mathbf{n}}\, h(\mathbf{k}, \mathbf{n})$ for a function h of interest. In what follows we will assume that the function h is bounded; examples include $h(\mathbf{k}, \mathbf{n}) = k_{r,i}$ and predictive probabilities in species sampling problems with multiple populations (Camerlenghi et al., 2019). An MCMC estimator derived from a Markov chain $\mathbf{k}_1, \ldots, \mathbf{k}_M$,

$$\hat{h} = \frac{1}{M} \sum_{m=1}^{M} h(\mathbf{k}_m, \mathbf{n}),$$

is generally biased. The bias is a function of the starting point and the length of the chain. Importance sampling can also be applied to estimate posterior moments, but estimates will be biased if the posterior density is only known up to a constant. On the other hand, if the sequence $\mathbf{k}_1, \ldots, \mathbf{k}_M$ consists of perfect samples of the posterior of $\mathbf{k}$ given $\mathbf{n}$, then the estimator above is unbiased.

Glynn and Rhee (2014) proposed an alternative strategy to estimate posterior expectations without bias. Concretely, let $(h_m)_{m\geq 0}$ be a sequence of estimators for h^* which is asymptotically unbiased, that is, $\mathbb{E}(h_m) \to h^*$ as $m \to \infty$. The Glynn–Rhee procedure constructs unbiased estimators from this sequence. While this yields one of the desirable properties of perfect sampling, the unbiased estimators produced by this method don't need to be bounded, even when the function h is, and indeed constructing bounded estimators is not always possible (Jacob and Thiery, 2015).

To define the estimator, let Δ_m for $m \geq 0$ be random variables satisfying $\mathbb{E}(\Delta_m) = \mathbb{E}(h_m - h_{m-1})$ with the convention that $h_{-1} = 0$. Let T be a random positive integer

independent of $(\Delta_m)_{m\geq 0}$ with $p(T \geq m) > 0$ for all $m \geq 0$. Then, by Fubini's theorem, the estimator

$$\hat{h}_{\mathrm{GR}} = \sum_{m=0}^{T} \frac{\Delta_m}{p(T \geq m)} = \sum_{m=0}^{\infty} \frac{\mathbb{1}(T \geq m)}{p(T \geq m)} \Delta_m \tag{2.12}$$

satisfies $\mathbb{E}\hat{h}_{\mathrm{GR}} = h^*$, provided that $\sum_{m=0}^{\infty}(\mathbb{E}|\Delta_m|/p(T \geq m)) < \infty$. This last condition implies that $\Delta_m \to 0$ weakly. The Glynn–Rhee estimator $\hat{h}_{\mathrm{GR}}$ has finite variance if the variables Δ_m tend to vanish quickly. Previous work on the outlined method includes two constructions of sequences $(\Delta_m)_{m\geq 0}$, based on Markov chain couplings. For both cases Glynn and Rhee (2014) discussed sufficient conditions to verify that $\hat{h}_{\mathrm{GR}}$ has finite variance. The recent work of Jacob et al. (2020) reviews and extends the discussion on these sufficient conditions. See also Theorem 1 in Rhee and Glynn (2015) for a self-contained analysis. Glynn and Rhee (2014) derived also a number of useful properties for averages of the form

$$\check{h}_{\mathrm{GR}} = \frac{1}{L} \sum_{\ell=1}^{L} \hat{h}_{\mathrm{GR}}^{(\ell)},$$

where $\hat{h}_{\mathrm{GR}}^{(1)}, \ldots, \hat{h}_{\mathrm{GR}}^{(L)}$ are i.i.d. copies of $\hat{h}_{\mathrm{GR}}$. Notably, under certain conditions on the coupling and the variable T, the estimator $\check{h}_{\mathrm{GR}}$ satisfies a Central Limit Theorem.

The first construction in Glynn and Rhee (2014) involves defining

$$\Delta_m = h(\tilde{\mathbf{k}}_m, \mathbf{n}) - h(\mathbf{k}_{m-1}, \mathbf{n}) \quad \text{for } m \geq 1, \quad \text{and} \quad \Delta_0 = h(\tilde{\mathbf{k}}_0, \mathbf{n}), \tag{2.13}$$

where $(\mathbf{k}_m)_{m\geq 0}$ and $(\tilde{\mathbf{k}}_m)_{m\geq 0}$ are identically distributed, coupled Markov chains with stationary distribution $p(\mathbf{k} \mid \mathbf{n})$, in which $\|\tilde{\mathbf{k}}_{m+1} - \mathbf{k}_m\|$ decreases quickly with high probability as $m \to \infty$. The requirements on this coupling are less stringent than the monotonicity required by the Propp–Wilson algorithm. Thus, there is significant flexibility in how to define it. By way of example, we propose using two copies of the doubly conditional Gibbs sampler introduced in Section 2.1. This construction is used in Section 4 to produce unbiased posterior moments in an example from microbial genomics.

In particular, let $\tilde{\mathbf{k}}_0 = \mathbf{k}_0$ be a fixed array. Each step of the coupled Gibbs samplers involves sampling the variables g, d, and $\mathbf{k}$ (see expression (2.6)). We define a Markov coupling similar to the one used to define the perfect sampler, in which the same set of random variables is utilized to define two Markov chains $(\mathbf{k}_m)_{m\geq 0}$ and $(\tilde{\mathbf{k}}_m)_{m\geq 0}$. Each transition $\mathbf{k}_{m-1} \to \mathbf{k}_m$ is coupled to the transition $\tilde{\mathbf{k}}_m \to \tilde{\mathbf{k}}_{m+1}$ for $m \geq 1$, in such a way that if $\mathbf{k}_j = \tilde{\mathbf{k}}_{j+1}$, then for all $m > j$, $\mathbf{k}_m = \tilde{\mathbf{k}}_{m+1}$. In each of these transitions, conditional sampling of g and d requires generating Gamma and Dirichlet random variables (expression (2.6)), and conditional on g and d the entries of $\mathbf{k}$ are generated using the inverse cumulative distribution. In the construction of the Markov chains, at each transition time Gamma random variables with different parameters are coupled by using shared arrays of exponential random variables. That is, we obtain Gamma distributed variables by summation of independent exponential random variables. The

Dirichlet random variables are coupled similarly, by generating coupled Gamma random variables and normalizing them. In this case we exploit the fact that a Dirichlet vector can be sampled by normalizing independent Gamma random variables. Finally, at each transition, the discrete variables ($k_{r,i}$ and $\tilde{k}_{r,i}$) are coupled by applying the conditional inverse cumulative distribution functions to the same Uniform$(0,1)$ variables.

As the support of the latent array $\mathbf{k}$ is finite, the coupling satisfies $p(\mathbf{k}_m = \tilde{\mathbf{k}}_{m+1} \mid \mathbf{k}_{m-1}, \tilde{\mathbf{k}}_m) > c_1$, for some $c_1 > 0$. The inequality follows from the fact that for any configuration of $(\mathbf{k}_{m-1}, \tilde{\mathbf{k}}_m)$ the conditional probability that both $\mathbf{k}_m$ and $\tilde{\mathbf{k}}_{m+1}$ take minimal value is strictly positive. Therefore, the coupling coalesces at a geometric rate, and as h is assumed to be bounded, we can deduce that $\mathbb{E}(\Delta_{m+1}^2) = \mathbb{E}([h(\mathbf{k}_m) - h(\tilde{\mathbf{k}}_{m+1})]^2) < c_2 c_3^m$, for some $c_2 > 0$ and $c_3 \in (0,1)$. The same argument implies that $\mathbb{E}(|\Delta_{m_1}\Delta_{m_2}|) < c_2 c_3^{\max(m_1,m_2)}$.These inequalities will be applied in the sequel to show that, for some choices of the random variable T, the estimator $\hat{h}_{\mathrm{GR}}$ has finite variance and the computing time has finite expectation.

The second construction in Glynn and Rhee (2014) employs a coupling from the past of Markov chain samplers for the target posterior $p(\mathbf{k} \mid \mathbf{n})$. To be precise, let π be a random mapping from the state space of the array $\mathbf{k}$ to itself, such that if $\mathbf{k}' \sim p(\mathbf{k} \mid \mathbf{n})$ and π is independent of $\mathbf{k}'$, then $\pi(\mathbf{k}') \sim p(\mathbf{k} \mid \mathbf{n})$. Letting $(\pi_m)_{m \geq 1}$ be i.i.d. copies of π, define $\mathbf{k}_m = \pi_1 \circ \pi_2 \circ \cdots \circ \pi_m(\mathbf{k}_0)$, for $\mathbf{k}_0$ fixed. We can define a Glynn–Rhee estimator through equation (2.12), with

$$\Delta_m = h(\mathbf{k}_m, \mathbf{n}) - h(\mathbf{k}_{m-1}, \mathbf{n}) \quad \text{for } m \geq 1, \quad \Delta_0 = h(\mathbf{k}_0, \mathbf{n}). \tag{2.14}$$

By way of example, we could use the doubly conditional Gibbs sampler (Subsection 2.1) to specify π. The random function π_m, which maps points from the support of $\mathbf{k}$ into the same finite set, can be specified using arrays of Gamma and uniform random variables. The definition of π_m is nearly identical to the transitions in the inner loop of Algorithm 1 (line 6) where, given the arrays of Gamma and uniform random variables indexed by a transition time m, the transitions from one point in the support of $\mathbf{k}$ to the subsequent one become deterministic. In different words, arrays of Gamma and uniform random variables indexed by transition times $(m = 1, 2, \ldots)$ are used to define π_m and the grand coupling $(\pi_1 \circ \pi_2 \circ \cdots \circ \pi_m;\ m \geq 1)$.

As in the first coupling, in this second construction the conditional probability $p(\cap_{j \geq 0}\{\mathbf{k}_{m+j} = \mathbf{k}_{m+j+1}\} \mid \mathbf{k}_0, \ldots, \mathbf{k}_{m-1}) > c_1$ for some strictly positive c_1. This follows from the fact that, with strictly positive probability, the random function π_m can take the same minimal (or maximal) value at every point in the function domain. In this case the inequality implies that $\mathbb{E}(|\Delta_m|^2) < c_2 c_3^m$, and $\mathbb{E}(|\Delta_{m_1}\Delta_{m_2}|) < c_2 c_3^{\max(m_1,m_2)}$ for some $c_2 > 0$ and $c_3 \in (0,1)$.

For both of the couplings defined above, we can directly bound the variance of the estimator $\hat{h}_{\mathrm{GR}}$,

$$\mathrm{Var}(\hat{h}_{\mathrm{GR}}) \leq \mathbb{E}\hat{h}_{\mathrm{GR}}^2 \leq \mathbb{E} \sum_{m_1 \geq 0} \sum_{m_2 \geq 0} \frac{|\Delta_{m_1}|}{p(T \geq m_1)} \frac{|\Delta_{m_2}|}{p(T \geq m_2)}$$

and for several choices of $p(T \geq m)$, for example $p(T \geq m) \propto m^{c_4}$, the right hand of the expression is finite. If the tails are exponential, i.e. $p(T \geq m) \propto c_4^m$, then the constant c_4 must be large enough for the bound to be finite. As explained in Glynn and Rhee (2014) the required constant is determined by the geometric rate at which the coupling coalesces.

The simulations in Section 4 are limited to the two constructions in equations (2.13) and (2.14), both tailored to the doubly conditional Gibbs sampler in Subsection 2.1. However, the other two Gibbs samplers presented in Section 2 can be coupled in a similar way, and there is great flexibility on how to specify the dependent Markov chains, as monotonicity is not required. This is a key difference between perfect sampling and the Glynn–Rhee estimation procedure. The Propp–Wilson algorithm requires a grand coupling with monotonicity, which severely restricts its applicability, whereas the Glynn–Rhee estimator can use any coupling where the two copies of the Markov chain tend to coalesce quickly. Our couplings are based on the divisibility property of gamma variables and inverse integral transformations. Recent papers by Jacob et al. (2020) and Jacob and Heng (2019) propose several constructions beyond those we considered to define coupled Markov chains in popular MCMC algorithms, such as Gibbs sampling and Metropolis–Hastings. One example is a maximal coupling (see Jacob et al. (2020)) which at each transition maximizes the probability of coalescence.

Jacob et al. (2020) and Rhee and Glynn (2015) discuss unbiased estimators which are similar to $\hat{g}_{\mathrm{GR}}$ but may have significantly lower variance. For example, Jacob et al. (2020) considered an estimator which does not use a truncation variable T independent of the sequence $(\Delta_m)_{m\geq 0}$. This is defined by

$$\hat{h}_{\mathrm{GR2}} = \sum_{m=0}^{\tau} \Delta_m,$$

where $\tau = \inf\{m : \Delta_{m+j} = 0 \text{ for every } j \geq 0\}$. This and other modifications are discussed in more detail in Appendix C, in relation to the application in Section 4.

3 Simulation study

The Gibbs samplers and perfect sampler of Section 2 were implemented (code available at `http://www.github.com/bacallado/hpy`). This section describes numerical experiments with the aim of evaluating the efficiency of the three Gibbs samplers, and the running time of the perfect sampler, for a range of parameter settings.

3.1 Testing the correctness of the implementation

The testing procedure described by Geweke (2004) was used to verify the correctness of the algorithms and their implementation. To be precise, for a fixed setting of parameters $(\theta, \theta_0, \beta, \beta_0)$, define the following routine:

1. Sample a pair $(\mathbf{n}, \mathbf{k}')$ from the CRF, with $R = 5$ populations, and $n_{r,\cdot} = 100$ samples for each population $r = 1, \ldots, 5$.

2. Sample an array $\mathbf{k}''$ from one of the algorithms in the following list with target distribution $p(\mathbf{k} \mid \mathbf{n})$:

   (a) 100 iterations of the collapsed Gibbs sampler, initialised at $\mathbf{k}'$,

   (b) 100 iterations of the conditional Gibbs sampler, initialised at $\mathbf{k}'$,

   (c) 100 iterations of the doubly conditional Gibbs sampler, initialised at $\mathbf{k}'$, or

   (d) perfect sampling.

In each case (a–d), if the implementation is correct, the pair $(\mathbf{k}', \mathbf{k}'')$ should be exchangeable. For a specific setting of the parameters, namely $\beta_0 = \beta = 0.1$ and $\theta_0 = \theta = 1$, we sample 200 independent copies $(\mathbf{k}'_m, \mathbf{k}''_m)_{m=1}^{200}$ of the pair, and test exchangeability through a permutation test using the following test statistics

$$s_1 = \frac{1}{200} \sum_{m=1}^{200} \text{mean}(\mathbf{k}'_m) - \text{mean}(\mathbf{k}''_m), \tag{3.1}$$

$$s_2 = \frac{1}{200} \sum_{m=1}^{200} \max(\mathbf{k}'_m) - \max(\mathbf{k}''_m).$$

The test was performed for both statistics and all four algorithms. Using a Bonferroni correction to maintain the familywise error below 1% for the two tests and all algorithms, the null hypothesis was not rejected in any case.

3.2 Convergence diagnostics for Gibbs samplers

Simulations of the three Gibbs samplers were run for a fixed dataset and two settings of the parameters $(\theta, \theta_0, \beta, \beta_0) = (1, 1, 0.1, 0.1)$ or $(1, 1, 0.7, 0.1)$. We use a dataset from the human vaginal microbiome study discussed in Ravel et al. (2011). In the CRF analogy, there are 900 restaurants, 134 distinct dishes observed, and we subsample 1,000 customers per restaurant.

For each algorithm and each parameter setting, we simulate a total of 16 chains, which are initialized at the extremes of the state space of $\mathbf{k}$, namely $k_{r,i} = n_{r,i}$ and $k_{r,i} = \mathbb{1}(n_{r,i} > 0)$. Each chain is 2,000 steps long. Figure 5 in the Supplementary Material (Bacallado et al., 2021) shows trace plots for two functions of $\mathbf{k}$, organized by algorithm (columns) and initial state (rows), with $(\theta, \theta_0, \beta, \beta_0) = (1, 1, 0.7, 0.1)$. The quantities examined were the mean of the array, $(RI)^{-1} \sum_{r,i} k_{r,i}$, and a specific entry, $k_{758,1}$, which had the largest mean across simulations. A visual inspection of the plots suggests rapid mixing for all three algorithms, and the figure is similar for the parameters $(\theta, \theta_0, \beta, \beta_0) = (1, 1, 0.1, 0.1)$. In addition, two convergence diagnostics were computed using the R package CODA (Plummer et al., 2006), the autocorrelation function and the potential scale reduction factor of Gelman and Rubin (1992). In every case, the three Gibbs samplers appear to mix in just 20 iterations, with potential scale reduction factors below 1.02 with 95% confidence, and autocorrelations after 20 steps below 0.043 in absolute value.

Grids of parameter vectors with two parameters fixed
$\{(\theta, \theta_0, \beta, \beta_0)\ ;\ \theta = 1, \theta_0 = 1, \beta \in B, \beta_0 \in B\}$
$\{(\theta, \theta_0, \beta, \beta_0)\ ;\ \beta = 0.1, \beta_0 = 0.1, \theta \in \Theta, \theta_0 \in \Theta\}$
$\{(\theta, \theta_0, \beta, \beta_0)\ ;\ \theta = 1, \beta = 0.1, \theta_0 \in \Theta, \beta_0 \in B\}$
$\{(\theta, \theta_0, \beta, \beta_0)\ ;\ \theta_0 = 1, \beta_0 = 0.1, \theta \in \Theta, \beta \in B\}$

Table 1: Parameter settings for simulation experiments. We define four grids with two parameters fixed in each case, and the other ones varying over the ranges $\Theta = \{1, 5, 10, 15, 20\}$ or $B = \{0.1, 0.2, 0.3, 0.4, 0.45, 0.5, 0.55, 0.6, 0.7, 0.8, 0.9\}$.

3.3 Testing the doubly conditional Gibbs sampler's convergence

To further evaluate the convergence of the doubly conditional Gibbs sampler, for a range of parameter values summarised in Table 1, we apply a test similar to that of Section 3.1. For each setting of the parameters $(\theta, \theta_0, \beta, \beta_0)$, we iterate the following steps:

1. Sample a pair $(\mathbf{n}, \mathbf{k}')$ from the CRF, with $R = 5$ populations, and $n_{r\cdot} = 100$ samples for each population $r = 1, \ldots, 5$.
2. Obtain a pseudo-sample $\mathbf{k}''$ from the distribution $p(\mathbf{k} \mid \mathbf{n})$ by simulating 1000 steps of the Gibbs sampler from the initial state $\mathbf{k}^{\text{init}}$ with $k^{\text{init}}_{r,i} = \mathbb{1}(n_{i,r} > 0)$.

The difference between this routine and that of Section 3.1 is the initial state of the Markov chain. The routine was iterated 200 times to obtain pairs $(\mathbf{k}'_m, \mathbf{k}''_m)_{1 \leq m \leq 200}$. If the Markov chain produced a perfect sample from the target distribution, then $\mathbf{k}'_m$ and $\mathbf{k}''_m$ would be exchangeable. Thus, we test the null hypothesis of good mixing through a permutation test, using the statistics in equation (3.1) which have mean zero under the null.

With a Bonferroni correction to maintain the familywise error rate at 1% across the list of parameter values described in Table 1, we find that the test using either s_1 or s_2 does not reject the null hypothesis of good mixing.

3.4 Evaluation of perfect sampler's running time

The perfect sampler of Section 2 has a random running time. The running time of the sampler for a given simulation setup will be characterized in terms of two quantities: (i) the total number of Markov coupled transitions that had to be simulated in order to obtain a sample, and (ii) the number of attempts required until equation (2.11) is satisfied. The first quantity is linearly related to the CPU time required to obtain one sample.

For a fixed set of parameters $(\theta, \theta_0, \beta, \beta_0)$, we define the following routine:

1. Sample a pair $(\mathbf{n}, \mathbf{k}')$ from the CRF, with $R = 5$ populations, and $n_{r,\cdot} = 100$ samples for each population $r = 1, \ldots, 5$.

2. Sample an array $\mathbf{k}''$ from the distribution $p(\mathbf{k} \mid \mathbf{n})$ using the perfect sampler, recording the two running time statistics listed above.

This routine allows us to evaluate the running time of the algorithm in a well-specified setting; i.e. when the data $\mathbf{n}$ is sampled from the same model used for posterior inference.

For each combination of the parameters in Table 1, the routine above was iterated for a total of 3 CPU hours. The most striking results emerged from the experiment which fixed θ_0 and β_0 and varied θ and β; they are plotted in Figure 1. Large values of β and small values of θ dramatically increase the number of coupling steps required to obtain a sample. In fact, for the range of parameters θ considered in this simulation study, the algorithm almost never output a sample when $\beta > 0.5$. In addition, the number of attempts required until equation (2.11) is satisfied also increased with large β and small θ. This is despite the fact that convergence diagnostics suggest good mixing of the doubly conditional Gibbs sampler for all the parameter values in Table 1. Figure 1 illustrates also that the average computing time of the exact sampler, for fixed values of θ and β, tend to decrease with respect to θ_0 and slightly increases with β_0.

To explain this phenomenon we provide some heuristics. In the posterior distribution $p(\mathbf{k} \mid \mathbf{n})$ of equation (2.4), the dependence between variables $(k_{r,i}; 1 \leq r \leq R, 1 \leq i \leq I)$ may be represented by a factor graph with potential energy function

$$\log p(\mathbf{k} \mid \mathbf{n}) = \sum_{r,i} [\log f(n_{r,i}, k_{r,i}) + k_{r,i} \log \beta] + \sum_{i} \log \Gamma(k_{\cdot i} - \beta_0) + \sum_{r} \log \Gamma(\theta/\beta + k_{r\cdot}) - \log \Gamma(\theta_0 + k) + \text{constant}. \tag{3.2}$$

The potentials collected in the first term each depend on a single variable $k_{r,i}$, so the correlation between the variables is due to the other terms. In particular, the interaction terms of the form $\log \Gamma(\theta/\beta + k_{r\cdot})$ will be nearly linear in $k_{r\cdot}$ when the crucial parameter θ/β is large. Figure 2 provides a graphical illustration of this fact. Approximating these terms by linear functions of $k_{r\cdot} = \sum_i k_{r,i}$ breaks the dependence. Thus, one might speculate that a large value of θ/β makes computational inference of this posterior distribution easier. Now, considering the perfect sampling algorithm of Section 2, and the distribution $p_a(\mathbf{k})$ in particular, the same argument implies that a large value of θ/β makes the columns of $\mathbf{k}$ approximately independent. This could lead to faster mixing of the conditional Gibbs sampler for $p_a(\mathbf{k})$, and a faster running time for coupling from the past.

4 Application

This section illustrates the inference methods of Sections 2 and 2.3 using data from Ravel et al. (2011) on the vaginal microbiome. The data are collected in a contingency table $\mathbf{n}$ where rows represent 900 different biological samples from pregnant women, and columns represent 134 bacterial species. We split the table into a training set $\mathbf{n}_{\text{train}}$, with one hundred counts per biological sample, and a test set $\mathbf{n}_{\text{test}}$, such that $\mathbf{n}_{\text{train}} + \mathbf{n}_{\text{test}} = \mathbf{n}$.

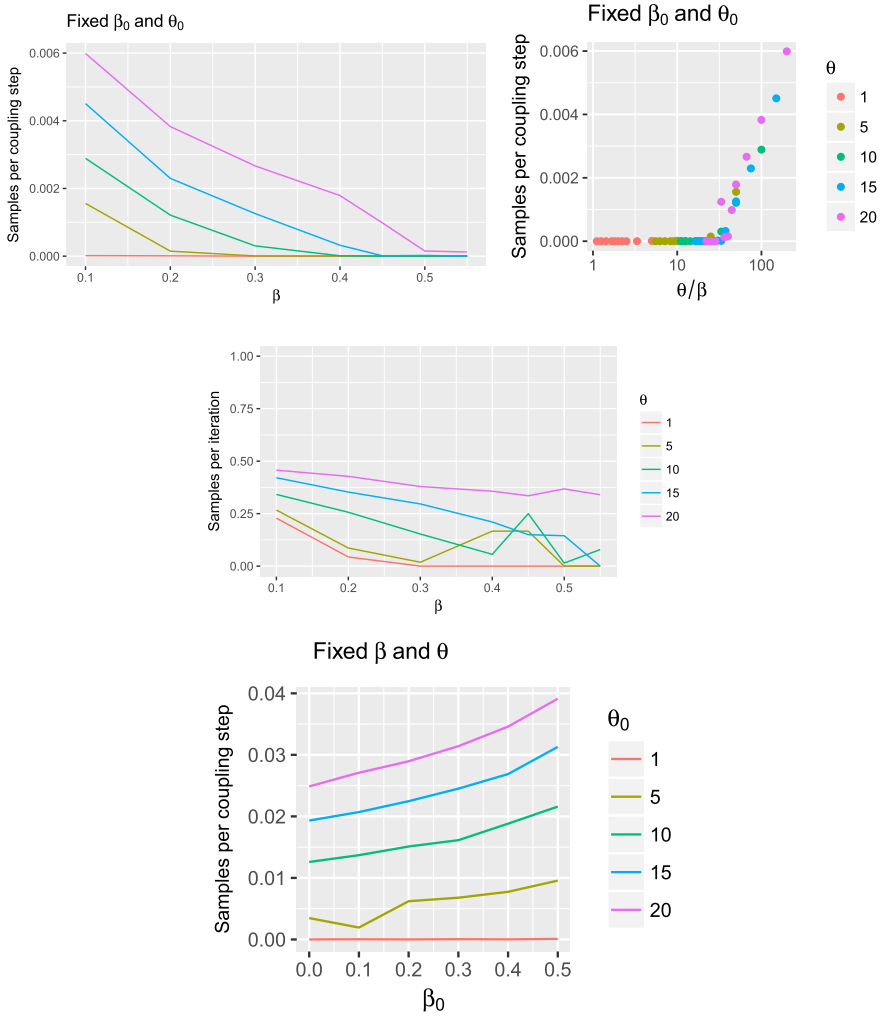

Figure 1: Computing time. We considered the grid of parameters in Table 1. Each point in the panels represents statistics recorded during 3 CPU hours. First and second row: simulations with $\beta_0 = 0.1$, $\theta_0 = 1$. The first row shows the number of samples obtained per coupling step simulated, and it illustrates trends when we vary β (left panel) and θ/β (right panel). The second row shows the number of samples obtained per iteration of the outer while loop of Algorithm 1. Third row, simulations with $\beta = 0.1$, $\theta = 1$. It illustrates variations of the number of samples per coupling step across combinations of β_0 and θ_0 values.

Perfect sampling was used to draw posterior samples of the latent variable $\mathbf{k}$ given $\mathbf{n}_{\text{train}}$, which were then used to obtain Monte Carlo estimates of two Bayesian estimators for quantities of interest: (i) the *missing-in-sample mass* and (ii) the *missing mass* for

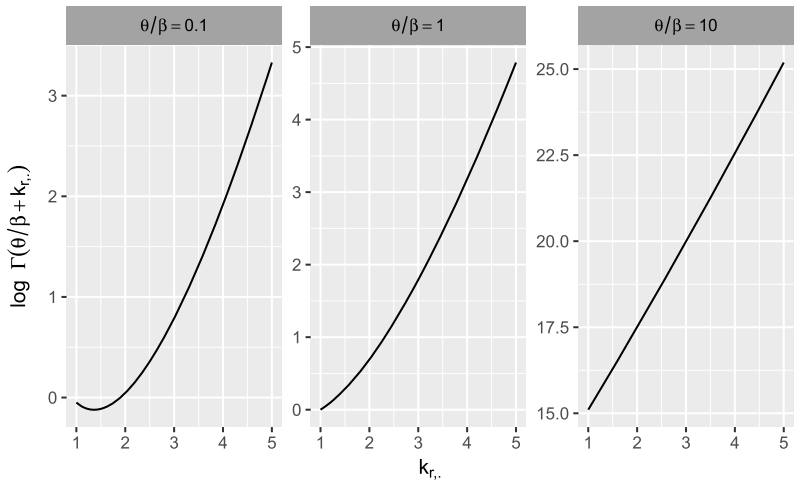

Figure 2: An illustration of the interaction potential $\log \Gamma(\theta/\beta + k_{r\cdot})$ for different values of θ/β. Larger values of θ/β reduce the curvature of the function.

one of the biological samples. To define these quantities, let $(X_1^*, X_2^*, \dots)$ be the atoms of the measure μ, and let $\{X_i^*; i \in \mathcal{I}_r\}$ be the set of species observed in sample r. The missing-in-sample mass for sample r is defined by

$$\sum_{i \in \mathbb{N} \setminus \mathcal{I}_r} \mu_r(\{X_i^*\}),$$

i.e. the total probability of all species which have not been observed in the sample r. The missing mass for sample r is defined by

$$\sum_{i \in \mathbb{N} \setminus (\mathcal{I}_1 \cup \cdots \cup \mathcal{I}_R)} \mu_r(\{X_i^*\}),$$

i.e. the total probability of all species which have not been observed in any of the samples. Estimators for the missing mass, such as the Good–Turing estimator, are classical in the literature on species sampling and Bayesian nonparametric counterparts have been studied (Good, 2000; Favaro et al., 2016).

We observe that the posterior expectation of the missing-in-sample and missing mass are simply predictive probabilities, respectively, the probability that the next observation from the biological sample in question is a new species for the sample, or a new species overall. Given the data $\mathbf{n}_{\text{train}}$ and the latent variable $\mathbf{k}$, these probabilities are available in closed form. Therefore, given perfect samples from the posterior of $\mathbf{k}$ given $\mathbf{n}_{\text{train}}$, we can estimate the missing-in-sample and missing mass by averaging the corresponding predictive probabilities.

Table 2 contrasts the estimates for missing-in-sample and missing mass for different settings of the model parameters (θ, β). The estimates derived via perfect sampling

are compared to those obtained through Markov chain Monte Carlo with the doubly conditional Gibbs sampler of Section 2.1, as well as unbiased estimates obtained with the method of Glynn and Rhee based on the same Gibbs sampler. In particular, we use the first construction of the unbiased estimator defined in equation (2.13), with a stopping time $T = 1000 + \text{Geometric}(0.1)$. This choice was based on the observation that the coupled Gibbs samplers tend to coalesce before 1000 transitions, making the estimator identical with high probability to the estimator $\hat{h}_{\text{GR2}}$ described by Jacob et al. (2020), which does not require a truncation variable T. Other versions of the Glynn–Rhee estimator are compared in Appendix C.

The cost of computing the confidence intervals in Table 2 was different for each method, so one should not interpret the width of the intervals as a measurement of efficiency. For completeness, the cost of each method is quantified in Table 3 in Appendix B. As one would expect, perfect sampling was between 100 and 1000 times as expensive as the other two methods.

The 95% confidence intervals displayed in Table 2 for the perfect sampling and Glynn–Rhee estimators are constructed via normal approximation. In the case of $\beta = 0.7$ and $\theta = 1$, perfect sampling becomes computationally infeasible, whereas we can derive missing mass estimators through Glynn–Rhee method. In this example, we expect Glynn–Rhee confidence intervals to have good coverage (see simulations in Appendix C).

It is possible to simulate the posterior predictive distribution given $\mathbf{n}_{\text{train}}$, by drawing $\mathbf{k}$ from its posterior and then simulating the CRF urn scheme. This two-step procedure was applied to sample replicates of $\mathbf{n}_{\text{test}}$, with the same row margins. Figure 3 shows the posterior distribution of the number of new species discovered, for two different settings of the model parameters. The observed statistic in $\mathbf{n}_{\text{test}}$ is marked on the plot. As expected the parameters affect the prediction.

5 Conclusions

The hierarchical Pitman–Yor process is a popular model for dependent random probability measures. However, the behavior of MCMC methods for posterior inference is not well-understood. This paper proposes three new inference algorithms –a Gibbs sampler, a method for unbiased estimation, and a perfect sampler– which provide different levels of reliability. It is difficult to rigorously evaluate the convergence of the Markov chain sampler. The Glynn–Rhee method eliminates the bias of traditional MCMC estimators. The perfect sampler, on the other hand, provides the strongest guarantees but has a random running time.

Availability of multiple computational methods for inference under the hierarchical Pitman–Yor process facilitates the use of the model in data analyses. Our simulation study suggests the use of the doubly-collapsed Gibbs sampler and the application of the Glynn-Rhee framework, among the procedures that we tested in our manuscript, as effective tools for data analyses like those presented in Section 4. The comparison of data analyses repeated leveraging different algorithms to approximate the posterior, is a viable solution to validate their implementations. This includes the evaluation of the

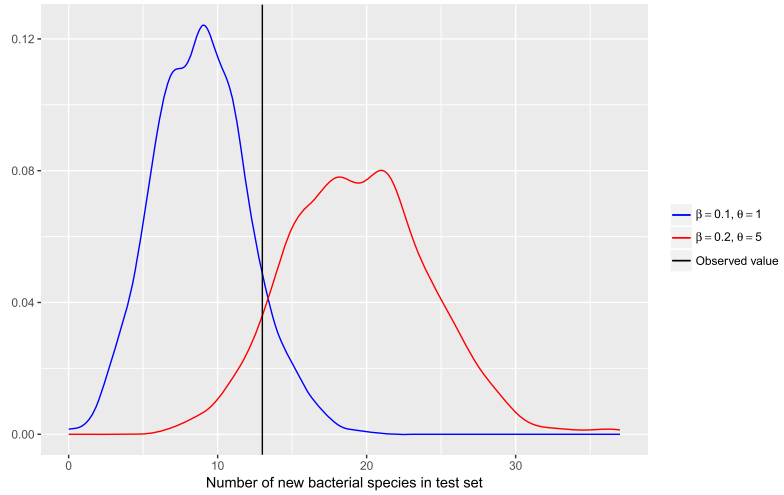

Figure 3: Posterior distribution of the number of species discovered in the test set under two different parameter settings. In each case, $\theta_0 = 1$ and $\beta_0 = 0.1$. The value observed in the test set is marked with a vertical line.

number of MCMC iterations necessary to approximate the posterior and the assessment of potential bias of the MCMC estimates. A major advantage of the coupling from the past sampler is that it permit straightforward comparisons between exact samples and other approaches to approximate the posterior distribution.

Also, MCMC algorithms can be easily modified, for example to include a prior distribution on the hierarchical Pitman–Yor process parameters $(\theta, \beta, \theta_0, \beta_0)$, or to produce posterior inference under other variants of the hierarchical model. In these cases, as we discussed, the Glynn-Rhee methodology, appears more flexible compared to coupling of Markov chains for exact sampling. Indeed, the application of this methodology allows the analyst to build on MCMC procedures, without the restrictive monotonicity requirements of the Propp and Wilson approach (Propp and Wilson, 1996).

One advantage of the perfect sampler is that it allows us to evaluate the running time by numerical simulations. We found that there are regions of the parameter space in which the algorithm terminates quickly and others in which the computational burden to obtain posterior samples increases substantially. These regions are separated by a relatively sharp boundary. The problematic regions correspond to values of the parameters for which we expect posterior inference to be harder, based on the analytic expression of $p(\mathbf{k} \mid \mathbf{n})$.

We focused on applications where the dishes assigned to each customer, in the CRF analogy, are directly observed, as in the human microbiome dataset of Section 4. In many applications of the hierarchical Pitman–Yor distribution, the dishes assigned to each customer are not observed and instead represent latent variables, such as cluster membership indicators in a mixture model. The collapsed Gibbs sampler and the

	Missing-in-sample mass		
	Perfect sampling	Gibbs sampling	Glynn–Rhee
$\theta = 1$, $\beta = 0.1$	1.40×10^{-2}	1.40×10^{-2}	1.33×10^{-2}
	$(\pm 1.71 \times 10^{-4})$	$(\pm 1.50 \times 10^{-4})$	$(\pm 6.09 \times 10^{-5})$
$\theta = 5$, $\beta = 0.2$	6.59×10^{-2}	6.79×10^{-2}	5.74×10^{-2}
	$(\pm 2.60 \times 10^{-3})$	$(\pm 7.67 \times 10^{-4})$	$(\pm 2.59 \times 10^{-4})$
$\theta = 5$, $\beta = 0.7$	–	1.38×10^{-1}	7.75×10^{-2}
		$(\pm 4.17 \times 10^{-3})$	$(\pm 8.26 \times 10^{-4})$
	Missing mass		
	Perfect sampling	Gibbs sampling	Glynn–Rhee
$\theta = 1$, $\beta = 0.1$	2.84×10^{-5}	2.83×10^{-5}	2.92×10^{-5}
	$(\pm 3.48 \times 10^{-7})$	$(\pm 3.04 \times 10^{-7})$	$(\pm 1.34 \times 10^{-7})$
$\theta = 5$, $\beta = 0.2$	8.69×10^{-5}	8.95×10^{-5}	1.07×10^{-4}
	$(\pm 3.35 \times 10^{-6})$	$(\pm 1.01 \times 10^{-6})$	$(\pm 5.16 \times 10^{-7})$
$\theta = 5$, $\beta = 0.7$	–	1.52×10^{-4}	1.40×10^{-4}
		$(\pm 4.63 \times 10^{-6})$	$(\pm 1.59 \times 10^{-6})$

Table 2: Bayes estimators of the missing-in-sample and missing mass in the first biological sample with 95% confidence intervals. Two parameters are fixed at $\theta_0 = 1$ and $\beta_0 = 0.1$.

conditional Gibbs sampler in Section 2 were originally designed with such models in mind. Appendix D describes two examples of algorithms that leverage and adapt the doubly conditional Gibbs sampler for posterior inference with a mixture model, where the assigned dishes in the CRF representation are latent variables. Although the perfect sampler defined here could be employed within a larger Gibbs sampler, in practice this would not be efficient due to the large computational cost compared to the Markov chain sampler it is based on. On the other hand, using coupled Markov chains for unbiased estimation in the context of mixture models appears as an attractive complement to MCMC for posterior inference, for example on the number of clusters, predictive probabilities, and other summaries of interest.

Supplementary Material

Supplementary Material (DOI: 10.1214/21-BA1269SUPP; .pdf).

References

Araki, T., Nakamura, T., Nagai, T., Nagasaka, S., Taniguchi, T. and Iwahashi, N. (2012). Online learning of concepts and words using multimodal LDA and hierarchical Pitman–Yor language model. In *Proceeding of the International Conference on Intelligent Robots and Systems*, 1623–1630. 2

Argiento, R., Cremaschi, A. and Vannucci, M. (2020). Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical*

Association, in press. MR4078466. doi: https://doi.org/10.1080/01621459.2019.1594833. 1

Bacallado, S., Favaro, S., Power, S., and Trippa, L. (2021). "Supplementary Material of "Perfect Sampling of the Posterior in the Hierarchical Pitman–Yor Process"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1269SUPP. 15

Bassetti, F., Casarin, R. and Rossini, L. (2020). Hierarchical species sampling models. *Bayesian Analysis*, in press. MR4132651. doi: https://doi.org/10.1214/19-BA1168. 1, 4

Battiston, M., Favaro, S. and Teh, Y. W. (2018). Multi-armed bandits for species discovery: a Bayesian nonparametric approach. *Journal of the American Statistical Association* **113**, 455–466. MR3803478. doi: https://doi.org/10.1080/01621459.2016.1261711. 1

Beal, M. J., Ghahramani, Z. and Rasmussen, C. E. (2002). The infinite hidden Markov model. In *Proceedings of Advances in Neural Information Processing Systems*, 577–584. 1

Blunsom, P. and Cohn, T. (2011). A hierarchical Pitman–Yor process HMM for unsupervised part of speech induction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 865–874. 1

Camerlenghi, F., Lijoi, A., Orbanz, P. and Prünster, I. (2019). Distribution theory for hierarchical processes. *Annals of Statistics* **47**, 67–92. MR3909927. doi: https://doi.org/10.1214/17-AOS1678. 1, 4, 11

Camerlenghi, F., Dumitrascu, B., Ferrari, F., Engelhardt, B. E. and Favaro, S. (2019). Nonparametric Bayesian multi-armed bandits for single cell experiment design. *Preprint arXiv:1910.05355*. MR4194258. doi: https://doi.org/10.1214/20-AOAS1370. 1

Cremaschi, A., Argiento, R., Shoemaker, K., Peterson, C. B. and Vannucci, M. (2020). Hierarchical normalized completely random measures for robust graphical modeling. *Bayesian Analysis*, in press. MR4044853. doi: https://doi.org/10.1214/19-BA1153. 1

Favaro, S., Nipoti, B. and Teh, Y. W. (2016). Rediscovery of Good–Turing estimators via Bayesian nonparametrics. *Biometrics* **72**, 136–145. MR3500582. doi: https://doi.org/10.1111/biom.12366. 19

Favaro, S., and Teh, Y. W. (2013). MCMC for normalized random measure mixture models. *Statistical Science* **28**, 335–359. MR3135536. doi: https://doi.org/10.1214/13-STS422. 7

Ferguson (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230. MR0350949. 2

Gelman, A., Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472. 15

Geweke, J. (2004). Getting it right: joint distribution tests of posterior simulators. *Journal of the American Statistical Association* **99**, 799–804. MR2090912. doi: https://doi.org/10.1198/016214504000001132. 14

Glynn, P. W. and Rhee, C. (2014). Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability* **51**, 377–389. MR3317370. doi: https://doi.org/10.1239/jap/1417528487. 2, 5, 11, 12, 13, 14

Good, I. J. (2000). Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma. *Journal of Statistical Computation and Simulation* **66**, 101–111. MR1807533. doi: https://doi.org/10.1080/00949650008812016. 19

Huang, S. and Renals, S. (2007). Hierarchical Pitman–Yor language models for ASR in meetings. In *Proceeding of the Workshop on Automatic Speech Recognition and Understanding*, 124–129. 1

Jacob, P. and Heng, J. (2019). Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika* **106**, 287–302. MR3949304. doi: https://doi.org/10.1093/biomet/asy074. 14

Jacob, P. and Thiery, A. (2015). On nonnegative unbiased estimators. *The Annals of Statistics* **43**, 769–784. MR3319143. doi: https://doi.org/10.1214/15-AOS1311. 11

Jacob, P., O'Leary, J. and Atchadé, Y. (2020). Unbiased Markov chain Monte Carlo with couplings. *Journal of the Royal Statistical Society Series B*, in press. MR4112777. doi: https://doi.org/10.1111/rssb.12336. 12, 14, 20

Lijoi, A. and Prünster, I (2010). Models beyond the Dirichlet process. In *Bayesian Nonparametrics*. Cambridge University Press, Cambridge. MR2730661. 2

Lindsey, R. V., Headden, W. P. and Stipicevic, M. J. (2012). A phrase-discovering topic model using hierarchical Pitman–Yor processes. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 214–222. 2

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169–186. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 4, 7

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 2, 3

Pitman, J. (2006). *Combinatorial Stochastic Processes* Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002. Springer. MR2245368. 2

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R news* **6**, 7–11. 15

Propp, J. G. and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9**, 223–252. MR1611693. doi: https://doi.org/10.1002/(SICI)1098-2418(199608/09)9:1/2<223::AID-RSA14>3.3.CO;2-R. 5, 10, 21

Ravel, J., Gajer, P., Abdo, Z., Schneider, G. M., Koenig, S. S., McCulle, S. L., Karlebach, S., Gorle, R., Russell, J., Tacket, C. O. et al. (2011). Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences* **108**, 4680–4687. doi: https://doi.org/10.1073/pnas.1002611107. 15, 17

Rhee, C. H. and Glynn, P. W. (2015). Unbiased estimation with square root convergence for SDE models. *Operations Research* **63**, 1026–1043. MR3422533. doi: https://doi.org/10.1287/opre.2015.1404. 12, 14

Sato, I. and Nakagawa, H. (2010). Topic models with power-law using Pitman–Yor process. In *Proceedings of the International Conference on Knowledge discovery and data mining*, 673–682. 2

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650. MR1309433. 2, 7

Sudderth, E. B. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman–Yor processes. In *Proceeding of Advances in Neural Information Processing Systems*, 1585–1592. 2

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman–Yor processes. In *Proceedings of the International Conference on Computational Linguistics*, 985–992. 1

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*. Cambridge University Press, Cambridge. MR2730663. 1

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* **101**, 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 1, 3, 4, 5, 7

Van Gael, J., Saatci, Y., Teh, Y. W. and Ghahramani, Z. (2008). Beam sampling for the infinite hidden Markov model. In *Proceedings of the International Conference on Machine Learning*, 1088–1095. 1, 4

Wood, F., Archambeau, C., Gasthaus, J., James, L. F. and Teh, Y. W. (2009). A stochastic memoizer for sequence data. In *Proceedings of the International Conference on Machine Learning*, 1129–1136. 1

Acknowledgments

The authors are grateful to an Associate Editor and two anonymous Referees for all their comments, corrections, and suggestions which improved remarkably the paper.