Bayesian nonparametric inference beyond the Gibbs–type framework

Federico Camerlenghi*
Department of Economics, Management and Statistics,
University of Milano-Bicocca,
via Bicocca degli Arcimboldi 8, 20126 Milano, Italy.
E-mail: federico.camerlenghi@unimib.it

Antonio Lijoi†
Department of Decision Sciences, BIDSA and IGIER,
Bocconi University,
via Röntgen 1, 20136 Milano, Italy.
E-mail: antonio.lijoi@unibocconi.it

Igor Prünster
Department of Decision Sciences, BIDSA and IGIER,
Bocconi University,
via Röntgen 1, 20136 Milano, Italy.
E-mail: igor@unibocconi.it

Abstract

The definition and investigation of general classes of nonparametric priors has recently been an active research line in Bayesian Statistics. Among the various proposals, the Gibbs–type family, which includes the Dirichlet process as a special case, stands out as the most tractable class of nonparametric priors for exchangeable sequences of observations. This is the consequence of a key simplifying assumption on the learning mechanism, which however has justification except that of ensuring mathematical tractability. In this paper we remove such an assumption and investigate a general class of random probability measures going beyond the Gibbs–type framework. More specifically, we present a nonparametric hierarchical structure based on transformations of completely random measures, which extends the popular hierarchical Dirichlet process. This class of priors preserves a good degree of tractability, given we are able to determine the fundamental quantities for Bayesian inference. In particular, we derive the induced partition structure and the prediction rules, and also characterize the posterior distribution. These theoretical results are also crucial to devise both a marginal and a conditional algorithm for posterior inference. An illustration concerning prediction in Genomic sequencing is also provided.

Keywords: Bayesian Nonparametrics; hierarchical process; completely random measure; exchangeability; normalized random measure; partition probability function; species sampling

*Also affiliated to BIDSA, Bocconi University, Milano and to the Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Torino, Italy.

†Also affiliated to the Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Torino, Italy.

1 Introduction

The proposal of novel classes of priors for Bayesian nonparametric inference typically aims at increasing modeling flexibility, while preserving analytical or computational tractability. In fact, the ubiquitous Dirichlet process stands out in terms of tractability due to its conjugacy property but it appears quite limited in terms of flexibility for a number of relevant applied problems. These limitations spurred a considerable stream of papers in the last 15 years. Recent accounts on this subject can be found in the monographs by Hjort *et al.* (2010), Müller *et al.* (2015) and Phadia (2013). Among the proposed generalizations, Gibbs–type priors play a prominent role as shown in De Blasi *et al.* (2015), where an account of their possible uses for Bayesian inference is given. They are defined through a system of predictive distributions, whose weights depend on a sequence of non–negative numbers that solve a simple system of recursive equations and on a parameter $\sigma < 1$. Importantly, they include, as special cases, several instances of popular random probability measures such as, e.g., the Dirichlet process itself and the Pitman–Yor process.

The present contribution shows that even moving beyond the Gibbs–type setting, one can still identify classes of priors that are amenable to a full Bayesian analysis. Previous attempts in this direction focus on specific cases and are clearly affected by the challenging analytical hurdles that arise. See, e.g., Lijoi *et al.* (2005) and Favaro *et al.* (2011). In order to provide a preliminary description of the problem, suppose that $\mathbb{X}$ is a Polish space and $\mathscr{X}$ is the associated Borel σ–algebra. We denote by $\mathsf{P}_{\mathbb{X}}$ the space of all probability measures on $(\mathbb{X}, \mathscr{X})$, which is assumed to be equipped with the topology of weak convergence and $\mathscr{P}_{\mathbb{X}}$ represents the corresponding Borel σ–algebra. Then consider an exchangeable sequence of $\mathbb{X}$–valued observations $\boldsymbol{X} = \{X_i\}_{i\geq 1}$ such that

$$\begin{aligned} X_i|\tilde{p} &\overset{\text{iid}}{\sim} \tilde{p}, \quad i \geq 1 \\ \tilde{p} &\sim \mathcal{Q}. \end{aligned} \tag{1}$$

Hence, the prior $\mathcal{Q}$ is a probability measure on $(\mathsf{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$ and most popular choices are such that $\mathcal{Q}$ selects with probability 1 a discrete distribution on $\mathbb{X}$. The Dirichlet process (Ferguson, 1973), the Pitman–Yor process (Pitman & Yor, 1997) and normalized random measures with independent increments (Regazzini *et al.*, 2003) are notable examples. Almost sure discreteness of $\tilde{p}$ entails that ties within the data may occur with positive probability, i.e. $\mathbb{P}[X_i = X_j] > 0$ for any $i \neq j$. Hence, a vector of n observations $\boldsymbol{X}^{(n)} = (X_1, \ldots, X_n)$ will display $k \leq n$ distinct values, say $x_1^*, \ldots, x_k^*$, with respective frequencies $n_1, \ldots, n_k$. If P_0 is a non–atomic distribution, we say that $\mathcal{Q}$ is a Gibbs–type prior with parameter $\sigma < 1$ and base measure P_0 if

$$\mathbb{P}[X_{n+1} \in A \,|\, \boldsymbol{X}^{(n)}] = \frac{V_{n+1,k+1}}{V_{n,k}} P_0(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma)\, \delta_{x_j^*}(A) \tag{2}$$

for any $A \in \mathscr{X}$. The coefficients $\{V_{n,k} : n \geq 1,\ 1 \leq k \leq n\}$ are solutions of the system of recursive equation $V_{n,k} = V_{n+1,k+1} + (n - k\sigma)\, V_{n+1,k}$, with initial condition $V_{1,1} := 1$. Gibbs–type priors have been introduced in Gnedin & Pitman (2005), where one can also find a characterization of the $V_{n,k}$'s corresponding to different values of $\sigma < 1$, and rephrased in predictive terms as (2) in Lijoi *et al.* (2007). For our purposes, it is also important to stress that such a specification amounts to assuming that the pair $(\boldsymbol{X}, \mathcal{Q})$ induces a random partition Ψ of $\mathbb{N}$ such that, for any $n \geq 1$,

$$\mathbb{P}[\Psi_n = \{C_1, \ldots, C_k\}] = V_{n,k} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1} \tag{3}$$

identifies a random partition of $[n] = \{1, \ldots, n\}$, $n_j = \sharp(C_j)$ is the cardinality of C_j and $(a)_q = \Gamma(a+q)/\Gamma(a)$ for any $a > 0$ and integer $q \geq 0$. In this case, one has that $i, j \in [n]$ are in the same cluster of Ψ_n if and only if $X_i = X_j$. Note that the probability of observing a value not contained in $\{x_1^*, \ldots, x_k^*\}$, namely a "new" observation, is $V_{n+1,k+1}/V_{n,k}$: it depends on n and k but not on the frequencies $n_1, \ldots, n_k$. The same obviously holds for the probability of sampling any of the "old" observations, which is $1 - V_{n+1,k+1}/V_{n,k}$. This simplifying assumption on the learning mechanism (2) is the key reason for the mathematical tractability of Gibbs–type priors but clearly represents a restriction from an inferential point of view. See De Blasi *et al.* (2015) for a discussion. Here we remove this assumption and study random probability measures that lead to a more general predictive distribution, where the weights make use of all the information contained in the sample. To this end, we focus on a class of priors $\mathcal{Q}$ such that $(\boldsymbol{X}, \mathcal{Q})$ induces a random partition of $\mathbb{N}$ that is obtained as a mixture of random partitions on $\mathbb{N}$ having a simpler structure as they are induced by a wide class of discrete random probability measures with diffuse base measure. This representation is the key for obtaining an expression suitable to carry out posterior inference. This construction corresponds to the popular hierarchical processes, that are deeply related to the huge probabilistic literature on coagulation and fragmentation processes. See Kingman (1982) and Bertoin (2006). From a statistical standpoint, nice works connected to our contribution, though in a different dependence setup for the observations, can be found in Teh & Jordan (2010), Wood *et al.* (2011), Nguyen (2015) and Camerlenghi et al. (2016). It is finally worth mentioning another considerable body of work in the literature where instances of non Gibbs–type priors have been proposed and investigated. However, unlike our contribution, in all these papers the exchangeability assumption is dropped. Examples can be found in Fuentes-García *et al.* (2010), Bassetti *et al.* (2010), Müller *et al.* (2011), Navarrete & Quintana (2011), Airoldi *et al.* (2014), Quintana *et al.* (2015) and Dahl *et al.* (2016).

Section 2 recalls some basic notions on completely random measures, which are then used to define normalized random measures with independent increments and the Pitman–Yor process. In Section 3 we introduce a prior $\mathcal{Q}$ that induces a system of predictive distributions, for which, in contrast to (2), the probability of a new observation depends on the cluster frequencies $n_1, \ldots, n_k$ and hence use the full sample information. This generality clearly complicates their analysis, but we are still able to determine the probability distribution of the induced random partition. Moreover, in Section 4 we prove a posterior characterization, given the data and the latent variables. These findings, in addition to their theoretical interest, are also relevant for computational purposes. Indeed, we devise both a marginal algorithm that extends the standard Blackwell–MacQueen urn scheme and a conditional algorithm that simulates approximate realizations of $\tilde{p}$ from its posterior distribution. See Section 5. Finally, resorting to the marginal algorithm some prediction problems related to genomic data are faced in Section 6. Proofs are deferred to the Appendix.

2 Completely random measures and discrete nonparametric priors

Most commonly used priors on infinite–dimensional spaces can be described in terms of suitable transformations of completely random measures, which can be seen as a unifying concept of the field as argued in Lijoi & Prünster (2010). Also our construction relies on completely random measures and, hence, we briefly recall a few definitions and introduce some relevant notation that occurs throughout.

Let $\mathsf{M}_{\mathbb{X}}$ denote the space of all boundedly finite measures on $(\mathbb{X}, \mathscr{X})$, namely $m(A) < \infty$ for any $m \in \mathsf{M}_{\mathbb{X}}$ and for any bounded Borel set $A \in \mathscr{X}$. The space $\mathsf{M}_{\mathbb{X}}$ is assumed to be endowed with the Borel σ–field $\mathscr{M}_{\mathbb{X}}$ (see Daley & Vere-Jones, 2008). A random measure $\tilde{\mu}$ is a measurable map from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, taking values in $(\mathsf{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$. If one further has that the random variables $\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_n)$ are independent for any choice of bounded disjoint measurable sets $A_1, \ldots, A_n \in \mathscr{X}$, and any $n \geq 1$, then $\tilde{\mu}$ is termed *completely random measure* (CRM) on $(\mathbb{X}, \mathscr{X})$. See Kingman (1993) for an exhaustive account. If a CRM $\tilde{\mu}$ does not have jumps at fixed points in $\mathbb{X}$, then $\tilde{\mu} = \sum_{i \geq 1} J_i \delta_{Y_i}$, where $\{J_i\}_{i \geq 1}$ are independent random jumps and $\{Y_i\}_{i \geq 1}$ are independent and identically distributed (iid) random atoms. In this case, there exists a measure ν, termed Lévy intensity, on $\mathbb{R}^+ \times \mathbb{X}$ such that $\int_{\mathbb{R}^+ \times B} \min\{1, s\} \, \nu(\mathrm{d}s, \mathrm{d}x) < \infty$, for any bounded $B \in \mathscr{X}$, and

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

for any measurable function $f : \mathbb{X} \to \mathbb{R}^+$. In the following we consider almost surely finite CRMs and, for the sake of illustration, assume the jumps $\{J_i\}_{i \geq 1}$ and the locations $\{Y_i\}_{i \geq 1}$ to be independent which amounts to having a Lévy intensity of the form

$$\nu(\mathrm{d}s, \mathrm{d}x) = \rho(s)\mathrm{d}s c P_0(\mathrm{d}x)$$

for some constant $c > 0$ and a non–atomic probability measure P_0 on $(\mathbb{X}, \mathscr{X})$. We will write $\tilde{\mu} \sim \mathrm{CRM}(\rho, c; P_0)$.

2.1 Transformations of CRMs

The first class of priors we consider has been introduced in Regazzini *et al.* (2003) and it is obtained through a normalization of a CRM. Indeed, if $\tilde{\mu} \sim \mathrm{CRM}(\rho, c; P_0)$ and ρ is such that $\int_0^\infty \rho(s)\,\mathrm{d}s = \infty$, then $\tilde{p} = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$ defines a random probability measure on $(\mathbb{X}, \mathscr{X})$ which is termed *normalized random measure with independent increments* (NRMI), in symbols $\tilde{p} \sim \mathrm{NRMI}(\rho, c; P_0)$. Such a construction includes, as special cases, several noteworthy priors. In particular, one obtains the Dirichlet process with base measure cP_0 when $\rho(s) = s^{-1}\,\mathrm{e}^{-s}$ and the normalized σ–stable process (Kingman, 1975), arising from the normalization of a σ–stable CRM, specified by the choice $c = 1$ and $\rho(s) = \sigma\, s^{-1-\sigma}/\Gamma(1-\sigma)$ for some $\sigma \in (0, 1)$.

The other popular class of random probability measures we are going to consider is the Pitman–Yor process, which may be defined in different ways. A relevant construction for our purposes is as follows. Let $\tilde{\mu}_{0,\sigma}$ be a σ–stable CRM with base measure P_0 and denote with P_σ its probability distribution. Introduce, now, another random measure $\tilde{\mu}_{\sigma,\theta}$ whose probability distribution $P_{\sigma,\theta}$ on $(\mathsf{M}_{\mathbb{X}}, \mathscr{M}_{\mathbb{X}})$ is absolutely continuous with respect to P_σ and has Radon–Nikodym derivative

$$\frac{\mathrm{d}P_{\sigma,\theta}}{\mathrm{d}P_\sigma}(m) = m(\mathbb{X})^{-\theta}\frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)} \tag{4}$$

Then the random probability measure $\tilde{p} = \tilde{\mu}_{\sigma,\theta}/\tilde{\mu}_{\sigma,\theta}(\mathbb{X})$ is termed Pitman–Yor process or two–parameter Poisson–Dirichlet process. See Pitman & Yor (1997). Henceforth we use the notation $\tilde{p} \sim \mathrm{PY}(\sigma, \theta; P_0)$.

2.2 Exchangeable random partitions

Both NRMIs and the Pitman–Yor process are discrete random probability measures. Hence, when used to model exchangeable data as in (1), they induce a random partition Ψ of $\mathbb{N}$, whose restriction

Ψ_n on $[n] = \{1, \ldots, n\}$ is such that any two integers i and j in $[n]$ are in the same partition set if and only if $X_i = X_j$. From a probabilistic standpoint, such a partition is characterized by the so–called *exchangeable partition probability function* (EPPF). See Pitman (2006). It also plays a pivotal role to infer the clustering structure featured by the data and to carry out posterior inference. It is defined as

$$\Phi_k^{(n)}(n_1, \ldots, n_k) := \int_{\mathbb{X}^k} \mathbb{E}[\tilde{p}^{n_1}(\mathrm{d}x_1) \cdots \tilde{p}^{n_k}(\mathrm{d}x_k)] \tag{5}$$

for any $n \geq 1$, $k \in \{1, \ldots, n\}$ and vector $(n_1, \ldots, n_k)$ of positive integers such that $\sum_{i=1}^k n_i = n$. This is nothing but the probability of having a partition $\Psi_n = \{C_1, \ldots, C_k\}$ of $[n]$ into k sets with frequencies $n_j = \sharp(C_j)$ and, relative to (1), is the probability of observing a specific sample $\boldsymbol{X}^{(n)}$ featuring k distinct values with respective frequencies $(n_1, \ldots, n_k)$. Hence, for Gibbs–type priors, the form of $\Phi_k^{(n)}$ is readily available from (3) and it displays a simple product form. For example, for the Pitman–Yor process with parameters $\sigma \in (0, 1)$ and $\theta > -\sigma$ one has

$$\Phi_k^{(n)}(n_1, \ldots, n_k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta + 1)_{n-1}} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1}. \tag{6}$$

Also for NRMIs the expression of $\Phi_k^{(n)}$ is known. Indeed, if $\tilde{p} \sim \mathrm{NRMI}(\rho, c; P_0)$ in (1), with non–atomic P_0, one has

$$\Phi_k^{(n)}(n_1, \ldots, n_k) = \frac{c^k}{\Gamma(n)} \int_0^\infty u^{n-1} \, \mathrm{e}^{-c\psi(u)} \prod_{j=1}^{k} \tau_{n_j}(u) \, \mathrm{d}u \tag{7}$$

with $\psi(u) = \int_0^\infty [1 - \mathrm{e}^{-uv}] \, \rho(v) \, \mathrm{d}v$ and $\tau_r(u) = \int_0^\infty v^r \, \mathrm{e}^{-uv} \, \rho(v) \, \mathrm{d}v$ for any integer $r \geq 1$. For example, when $\rho(s) = s^{-1} \, \mathrm{e}^{-s}$, then $\Phi_k^{(n)}(n_1, \ldots, n_k) = c^k \prod_{i=1}^k (n_i - 1)!/(c)_n$ which is the EPPF corresponding to the Dirichlet process. On the other hand, if $\rho(s) = \sigma \, s^{-1-\sigma}/\Gamma(1 - \sigma)$ it is easily checked from (7) that $\Phi_k^{(n)}(n_1, \ldots, n_k) = \sigma^{k-1} \, \Gamma(k) \prod_{i=1}^k (1 - \sigma)_{n_i - 1}/\Gamma(n)$, which is the EPPF of a normalized σ–stable process.

It is worth noting that these three examples display the structure of a Gibbs–type prior and they are characterized by the specific weights $V_{n,k}$ and the value of σ. Hence, all these models yield a predictive distribution (2) for which the probability of observing a new value at the $(n+1)$–th step of the process, conditional on the first n values, does not depend on cluster frequencies $(n_1, \ldots, n_k)$ in $\boldsymbol{X}^{(n)}$. Next we remove such a simplifying condition, which is not justified from a modeling perspective, and work out explicit expressions of the EPPF that arise as mixtures of simpler EPPFs of the form displayed in this section.

3 Hierarchical processes

An effective strategy for obtaining a more complex structure than the one displayed by Gibbs–type priors is the use a mixture model. Let $Q(\cdot \,|\, P_0)$ be a probability distribution on $\mathsf{P}_{\mathbb{X}}$ such that $\int_{\mathsf{P}_{\mathbb{X}}} p(A) Q(\mathrm{d}p | P_0) = P_0(A)$, for any $A \in \mathscr{X}$. $\mathcal{Q}$ in (1) may then be defined as

$$\tilde{p} \,|\, \tilde{p}_0 \;\sim\; Q(\cdot \,|\, \tilde{p}_0), \qquad \tilde{p}_0 \;\sim\; Q'(\cdot \,|P_0) \tag{8}$$

for some non–atomic distribution P_0 on $\mathbb{X}$. Natural specifications for $Q(\cdot \,|\, \tilde{p}_0)$ and $Q'(\cdot \,|P_0)$ are then the probability distributions of either a NRMI or a Pitman–Yor process. Alternatively, the

model can be viewed as a random partition of $\mathbb{N}$ obtained as a mixture over a space of partitions induced by $\tilde{p}_0$. Such an interpretation will become apparent when stating our results. The model in (8) corresponds to a hierarchical process introduced in a multi–sample setting in Teh *et al.* (2006), with both $Q(\cdot\,|\,\tilde{p}_0)$ and $Q'(\cdot\,|P_0)$ being the probability distributions of Dirichlet processes, and known as hierarchical Dirichlet process.
In order to gain an intuitive insight on the induced partition structure we resort to a variation of the popular Chinese restaurant franchise metaphor in Teh *et al.* (2006), which we term *multi–room Chinese restaurant.* According to this scheme, the restaurant has a menu of infinitely many dishes (with labels generated by the non–atomic base measure P_0 in (8)) and rooms, with each room containing infinitely many tables, where people may be seated. Each person in the same room, regardless of the table she is seated at, eats the same dish in the menu. Different rooms serve different dishes. The generative construction is as follows: the first customer picks a dish from the menu, is assigned to a room and seated at one of its tables. The n–th customer entering the restaurant may either select a dish already chosen from the menu by at least one of the previous $n-1$ customers or a new dish. In the former case, she will be seated in the room serving her dish of choice, which hosts all other people, among the previous $n-1$, who have selected that very same dish. She may be seated either at a new table or at an existing table. If a new dish is chosen, the n–th customer will sit in a new room and at a new table. This scheme implies that distinct tables may share the same dish (as long as they are in the same room, which is the only one serving that dish). The key difference w.r.t. the standard Chinese restaurant process description of the Dirichlet process and of other Gibbs–type priors is that here we allow the same dish to be served at different tables. In the sequel X_i represents the dish eaten by the i–th customer of the restaurant, for $i = 1, \ldots, n$, whereas the tables are latent variables which denoted by $T_1, \ldots, T_n$. Moreover, if k distinct dishes are served in the restaurant, with respective labels $X_1^*, \ldots, X_k^*$, the frequency n_j in (5) represents the total number of customers seating in room j or, equivalently, eating dish j. Then $\ell_j \in \{1, \ldots, n_j\}$ is the number of occupied tables in room j, and $q_{j,t}$ is the number of customers seating at table t in room j, under the obvious constraint $\sum_{t=1}^{\ell_j} q_{j,t} = n_j$. Hence, the actual random partition has a probability distribution that is obtained by mixing with respect to all tables' configurations $\{(q_{j,t}, \ell_j) : t = 1, \ldots, \ell_j\}$, for $j \in \{1, \ldots, k\}$. This description clearly hints at an immediate connection with the theory of coagulation and fragmentation processes as nicely accounted for in the monograph Bertoin (2006).

3.1 Hierarchies of NRMIs

We first examine the following class of random probability measures

$$\begin{aligned} \tilde{p}\,|\,\tilde{p}_0 &\sim \text{NRMI}(\rho, c; \tilde{p}_0) \\ \tilde{p}_0 &\sim \text{NRMI}(\rho_0, c_0; P_0) \end{aligned} \tag{9}$$

and, for notational convenience, in the sequel we set $|\boldsymbol{x}| = \sum_{i=1}^{k} x_i$, for any $\boldsymbol{x} = (x_1, \ldots, x_k) \in \mathbb{R}^k$ and for any $k \geq 1$.

Theorem 1. *Suppose $\{X_n\}_{n\geq 1}$ is an exchangeable sequence of $\mathbb{X}$–valued random elements according to (1), where $\tilde{p} \sim \mathcal{Q}$ is such that*

$$\tilde{p}\,|\,\tilde{p}_0 \sim \text{NRMI}(\rho, c; \tilde{p}_0), \qquad \tilde{p}_0 \sim \text{NRMI}(\rho_0, c_0; P_0).$$

Then the EPPF that characterizes the random partition induced by $(X_1, \ldots, X_n)$*, for any* $n \geq 1$*, coincides with*

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \sum_{\boldsymbol{\ell}} \Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\ell_1, \ldots, \ell_k) \prod_{j=1}^{k} \frac{1}{\ell_j!} \sum_{\boldsymbol{q}_j} \binom{n_j}{q_{j,1}, \ldots, q_{j,\ell_j}} \times \Phi_{|\boldsymbol{\ell}|}^{(n)}(q_{1,1}, \ldots, q_{1,\ell_1}, \ldots, q_{k,1}, \ldots, q_{k,\ell_k}) \tag{10}$$

where $\Phi_{\cdot}^{(\cdot)}$ *and* $\Phi_{\cdot,0}^{(\cdot)}$ *indicate the EPPFs* (7) *associated to NRMIs with parameters* (ρ, c) *and* (ρ_0, c_0)*, respectively. Moreover, the first sum runs over all vectors of* $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ *such that* $\ell_i \in \{1, \ldots, n_i\}$ *and the* j*–th of the other* k *sums runs over all* $\boldsymbol{q}_j = (q_{j,1}, \ldots, q_{j,\ell_j})$ *such that* $q_{j,i} \geq 1$ *and* $|\boldsymbol{q}_j| = n_j$*.*

The expression in (10) can be readily rewritten in an equivalent form which discloses a nice interpretation of exchangeable random partitions induced by hierarchical NRMIs. One can represent $\Pi_k^{(n)}$ as a mixture over spaces of partitions, which can be described in terms of the multi–room Chinese restaurant metaphor: (i) the n customers are partitioned into t tables according to the NRMI with parameters (c, ρ); (ii) these t tables are are further allocated into k distinct rooms in the restaurant according to the NRMI with parameters (c_0, ρ_0) and this step corresponds to gathering the tables being served the same dish. The following Corollary effectively describes this structure.

Corollary 1. *If* $\{X_n\}_{n\geq 1}$ *is as in* Theorem 1*, then*

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \sum_{t=k}^{n} \sum_{\pi \in \mathscr{P}_{n,t}} \Big\{ \Phi_t^{(n)}(\sharp A_1^{\pi}, \ldots, \sharp A_t^{\pi}) \times \sum_{\pi' \in \mathscr{P}_{t,k}} \Big[\Phi_{k,0}^{(t)}(\sharp A_{0,1}^{\pi'}, \ldots, \sharp A_{0,k}^{\pi'}) \prod_{j=1}^{k} \mathbb{1}_{\{n_j\}}\Big(\sum_{i \in A_{0,j}^{\pi'}} \sharp A_i^{\pi} \Big) \Big] \Big\}$$

where $\mathscr{P}_{n,t}$ *and* $\mathscr{P}_{t,k}$ *are the spaces of all partitions of* $[n] = \{1, \ldots, n\}$ *and* $[t] = \{1, \ldots, t\}$ *into* t *and* k *sets, respectively, the sets* $A_1^{\pi}, \ldots, A_t^{\pi}$ *and* $A_{0,1}^{\pi'}, \ldots, A_{0,k}^{\pi'}$ *identify the partitions* π *and* π'*. Moreover, for any set* A*,* $\mathbb{1}_A$ *is the indicator function of* A*.*

Note that the product of indicators on the right–hand side of the equality displayed in Corollary 1 implies that we are summing over all partitions π and π' such that $\sum_{i \in A_{0,j}^{\pi'}} \sharp A_i^{\pi} = n_j$, for any $j = 1, \ldots, k$. The multi–room Chinese restaurant interpretation amounts to saying that the frequency of customers eating dish j at different tables in room j of the restaurant is exactly n_j.

Having characterized the distribution of the random partition induced by $\{X_n\}_{n\geq 1}$, it is then natural to look at the marginal distribution of the number of clusters K_n out of n observations $\boldsymbol{X}^{(n)} = (X_1, \ldots, X_n)$. This is a key quantity in a number of applications. For instance, it serves as a prior distribution on the number of components in nonparametric mixture models or on the number of distinct species in a sample of size n in species sampling problems. It is identified by the following

Corollary 2. *If* $\{X_n\}_{n\geq 1}$ *is as in* Theorem 1 *and* K_n *is the number of distinct values in* $\boldsymbol{X}^{(n)}$*, then*

$$\mathbb{P}[K_n = k] = \sum_{t=k}^{n} \frac{1}{t!} \sum_{\boldsymbol{\nu} \in \Delta_{t,n}} \binom{n}{\nu_1, \cdots, \nu_t} \Phi_t^{(n)}(\nu_1, \cdots, \nu_t) \times \frac{1}{k!} \sum_{\boldsymbol{\zeta} \in \Delta_{k,t}} \binom{t}{\zeta_1, \cdots, \zeta_k} \Phi_{k,0}^{(t)}(\zeta_1, \cdots, \zeta_k) \quad (11)$$

where $\Delta_{t,n} = \{\boldsymbol{\nu} = (\nu_1, \ldots, \nu_t) : \nu_i \geq 1, |\boldsymbol{\nu}| = n\}$ *for any* $n \geq 1$ *and* $t \in \{1, \ldots, n\}$.

In general, the sums involved in the mixtures (10) and (11) cannot be evaluated in closed form. However, as shown in the sequel, expressions like (10) form the backbone for devising suitable algorithms that allow to sample exchangeable random elements from $\tilde{p}$ and evaluate, among others, an approximation of the distribution of K_n. The following two examples illustrate the previous general results.

Example 1. Suppose $\rho(s) = \rho_0(s) = s^{-1}\,\mathrm{e}^{-s}$, so that (9) corresponds to a hierarchical Dirichlet process (HDP). See Teh *et al.* (2006) for its version in a multi–sample setting. A straightforward application of Theorem 1 yields a novel closed form expression of the EPPF

$$\Pi_k^{(n)}(n_1, \ldots, n_k) = \frac{c_0^k}{(c)_n} \sum_{\boldsymbol{\ell}} \frac{c^{|\boldsymbol{\ell}|}}{(c_0)_{|\boldsymbol{\ell}|}} \prod_{j=1}^{k} (\ell_j - 1)!\, |\mathfrak{s}(n_j, \ell_j)|. \quad (12)$$

where $|\mathfrak{s}(n,k)|$ are the signless Stirling numbers of the first kind and the sum runs over all vectors $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ such that $\ell_i \in \{1, \ldots, n_i\}$. The predictive distributions can be determined from (12). Indeed, using the recursive relationship $|\mathfrak{s}(n_j + 1, \ell_j)| = n_j |\mathfrak{s}(n_j, \ell_j)| + |\mathfrak{s}(n_j, \ell_j - 1)|$ and some algebra, one finds out that

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, \boldsymbol{X}^{(n)}] = \frac{c}{(c+n)} \sum_{\boldsymbol{\ell}} \frac{c_0}{(c_0 + |\boldsymbol{\ell}|)}\, \pi(\boldsymbol{\ell} \,|\, \boldsymbol{X}^{(n)}) P_0(\,\cdot\,) + \sum_{j=1}^{k} \left[\frac{n_j}{c+n} + \frac{c}{(c+n)} \sum_{\boldsymbol{\ell}} \frac{\ell_j}{(c_0 + |\boldsymbol{\ell}|)}\, \pi(\boldsymbol{\ell} \,|\, \boldsymbol{X}^{(n)}) \right] \delta_{x_j^*}(\,\cdot\,)$$

where $\pi(\boldsymbol{\ell} \,|\, \boldsymbol{X}^{(n)})$ is the posterior distribution of the latent variables $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k)$ coinciding with

$$\pi(\boldsymbol{\ell} \,|\, \boldsymbol{X}^{(n)}) \propto \frac{c_0^k\, c^{|\boldsymbol{\ell}|}}{(c_0)_{|\boldsymbol{\ell}|} \prod_{j=1}^{k} (c)_{n_j}} \prod_{j=1}^{k} (\ell_j - 1)!\, |\mathfrak{s}(n_j, \ell_j)|\, \mathbb{1}_{\{1, \ldots, n_j\}}(\ell_j).$$

Notice that the predictive probability mass associated to the j–th distinct observation x_j^*'s is $n_j/(c+n)$, which appears also in the Dirichlet prediction rule, and an additional term generated by the discreteness of $\tilde{p}_0$. In other terms, a dish x_j^* is picked by the $(n+1)$–th customer either because it is the one served at that particular table she seats at or because she seats at a new table while picking x_j^* from the menu. The latter term does not appear in the standard Chinese restaurant process associated to the Dirichlet case, where tables and dishes are in one–to–one correspondence.

One can also give an interesting integral representation of the EPPF corresponding to the HDP (12). Indeed, if Δ_k is the k–dimensional simplex, the definition of the signless Stirling number of

the first kind entails $\sum_{\ell_j=1}^{n_j} a^{\ell_j-1}|\mathfrak{s}(n_j,\ell_j)| = (a+1)_{n_j-1}$ and from (12) we deduce

$$\begin{aligned}\Pi_k^{(n)}(n_1,\ldots,n_k) &= \frac{(cc_0)^k}{(c)_n}\sum_{\boldsymbol{\ell}}\int_{\Delta_k}(1-|\boldsymbol{p}|)^{c_0-1}\prod_{j=1}^{k}(cp_j)^{\ell_j-1}|\mathfrak{s}(n_j,\ell_j)|\mathrm{d}\boldsymbol{p}\\ &= \frac{c^k}{(c)_n}\frac{c_0^k}{(c_0)_k}\int_{\Delta_k} D_k(\mathrm{d}\boldsymbol{p};1,\ldots,1,c_0)\prod_{j=1}^{k}(c\,p_j+1)_{n_j-1}\end{aligned} \tag{13}$$

where $D_k(\,\cdot\,;\alpha_1,\ldots,\alpha_{k+1})$ is the k–variate Dirichlet distribution with parameters $(\alpha_1,\ldots,\alpha_{k+1})$. Finally, note that (13) admits an interesting interpretation by noting that if $(Y_1,\ldots,Y_k)|\boldsymbol{p} \sim \mathrm{D}_k(cp_1+1,\ldots,cp_k+1,c(1-|\boldsymbol{p}|)$, then

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{(cc_0)^k}{(c)_k(c_0)_k}\int_{\Delta_k}\mathbb{E}[Y_1^{n_1}\cdots Y_k^{n_k}\,|\,\boldsymbol{p}]\,D_k(\mathrm{d}\boldsymbol{p};1,\ldots,1,c_0)$$

From (13) one can also readily obtain the corresponding predictive distributions.

Example 2. Let $\rho(s) = \sigma\, s^{-1-\sigma}/\Gamma(1-\sigma)$ and $\rho_0(s) = \sigma_0\, s^{-1-\sigma_0}/\Gamma(1-\sigma_0)$, with σ and σ_0 in $(0,1)$. In view of this specification, $\tilde{p}$ is a hierarchical normalized σ–stable process whose base measure is itself a normalized σ_0–stable process. Since the total masses c and c_0 are redundant under normalization, we set $c = c_0 = 1$ with no loss of generality. Based on Theorem 1 one has

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \sum_{\boldsymbol{\ell}}\frac{\sigma_0^{k-1}\,\Gamma(k)}{\Gamma(|\boldsymbol{\ell}|)}\,\frac{\sigma^{|\boldsymbol{\ell}|-1}\,\Gamma(|\boldsymbol{\ell}|)}{\Gamma(n)}\prod_{j=1}^{k}(1-\sigma_0)_{\ell_j-1}\,\frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\sigma^{\ell_j}}$$

where $\mathscr{C}(n,k;\sigma) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(-i\sigma)_n$ is the generalized factorial coefficient, for any $k \le n$ (see Charalambides (2002)), and the sum runs over all vectors $\boldsymbol{\ell} = (\ell_1,\ldots,\ell_k)$ such that $\ell_i \in \{1,\ldots,n_i\}$. Now rewrite $\Pi_k^{(n)}(n_1,\ldots,n_k)$ as

$$\begin{aligned}\Pi_k^{(n)}(n_1,\ldots,n_k) &= \frac{\sigma_0^{k-1}\,\Gamma(k)}{\sigma\Gamma(n)}\prod_{j=1}^{k}\sum_{\ell_j=1}^{n_j}\mathscr{C}(n_j,\ell_j;\sigma)\,(1-\sigma_0)_{\ell_j-1}\\ &= \frac{(\sigma\,\sigma_0)^{k-1}\,\Gamma(k)}{\Gamma(n)}\prod_{j=1}^{k}(1-\sigma\,\sigma_0)_{n_j-1}\end{aligned} \tag{14}$$

where the last equality follows immediately by the definition of generalized factorial coefficients. We observe that the EPPF in (14) is the one of a normalized $(\sigma\sigma_0)$–stable process. From a probabilistic point of view, the hierarchical construction can also be seen as subordination and, as such, (14) is an important and well–known formula originally derived in Pitman (1999), which has found several applications in Population Genetics with reference to coagulation phenomena.

3.2 Hierarchies of Pitman–Yor processes

The results displayed in Section 3.1 can be extended to the case where $Q(\,\cdot\,|\tilde{p}_0)$ and $Q'(\,\cdot\,|P_0)$ in (8) are the probability distributions of Pitman–Yor process, namely

$$\begin{aligned}\tilde{p}\,|\,\tilde{p}_0 &\sim \mathrm{PY}(\sigma,\theta;\tilde{p}_0)\\ \tilde{p}_0 &\sim \mathrm{PY}(\sigma_0,\theta_0;P_0)\end{aligned} \tag{15}$$

being $\sigma,\sigma_0 \in (0,1)$, $\theta > 0$ and $\theta_0 > 0$. We also refer to $\tilde{p}$ in (15) as hierarchical Pitman–Yor process and use the acronym HPYP. The following result provides the corresponding partition structure.

Theorem 2. *Let $\{X_n\}_{n\geq 1}$ be an exchangeable sequence of $\mathbb{X}$–valued random elements as in (1), with $\tilde{p}$ being a HPYP. Then the EPPF that characterizes the random partition of $\mathbb{N}$ induced by $\{X_n\}_{n\geq 1}$ is*

$$\Pi_k^{(n)}(n_1,\ldots,n_k) = \frac{\prod_{r=1}^{k-1}(\theta_0+r\sigma_0)}{(\theta+1)_{n-1}} \sum_{\boldsymbol{\ell}} \frac{\prod_{s=1}^{|\boldsymbol{\ell}|-1}(\theta+s\sigma)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j,\ell_j;\sigma)}{\sigma^{\ell_j}} (1-\sigma_0)_{\ell_j-1} \tag{16}$$

where the sum runs over all vectors of positive integers $\boldsymbol{\ell}=(\ell_1,\ldots,\ell_k)$ such that $\ell_i\in\{1,\ldots,n_i\}$, for each $i=1,\ldots,k$.

Remark 1. For the special case of $\theta=\theta_0\sigma$ in (15), and proceeding as for hierarchies of normalized σ–stable processes in Example 2, the EPPF (16) reduces to that of a simple PY process, or in other terms, $\tilde{p}\sim \mathrm{PY}(\sigma\sigma_0,\,\theta_0\sigma;P_0)$ and this can be easily checked through (16). Such a coagulation property was first established in Pitman (1999). See also Bertoin (2006).

Having the EPPF (16) at hand, one can derive the predictive distributions in a similar vein to the HDP case considered in Example 1. In fact,

$$\begin{aligned}\mathbb{P}[X_{n+1}\in\cdot\,|\,\boldsymbol{X}^{(n)}] = &\sum_{\boldsymbol{\ell}} \left(\frac{\theta+|\boldsymbol{\ell}|\sigma}{\theta+n}\right)\left(\frac{\theta_0+k\sigma_0}{\theta_0+|\boldsymbol{\ell}|}\right)\pi(\boldsymbol{\ell}\,|\,\boldsymbol{X}^{(n)})P_0(\,\cdot\,)\\ &+\sum_{j=1}^{k}\sum_{\boldsymbol{\ell}}\left\{\frac{n_j-\ell_j\sigma}{\theta+n}+\left(\frac{\theta+|\boldsymbol{\ell}|\sigma}{\theta+n}\right)\left(\frac{\ell_j-\sigma_0}{\theta_0+|\boldsymbol{\ell}|}\right)\right\}\pi(\boldsymbol{\ell}\,|\,\boldsymbol{X}^{(n)})\delta_{X_j^*}(\,\cdot\,)\end{aligned}$$

where $\pi(\boldsymbol{\ell}\,|\,\boldsymbol{X}^{(n)})$ is the posterior distribution of the latent variables $\boldsymbol{\ell}=(\ell_1,\ldots,\ell_k)$ coinciding with

$$\pi(\boldsymbol{\ell}\,|\,\boldsymbol{X}^{(n)}) \propto \frac{\prod_{s=1}^{|\boldsymbol{\ell}|-1}(\theta+s\sigma)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}} \prod_{i=1}^{k}\frac{\mathscr{C}(n_i,\ell_i;\sigma)}{\sigma^{\ell_i}}(1-\sigma_0)_{\ell_i-1}.$$

It is then immediately clear that a simulation algorithm can be devised by augmenting with respect to the latent variables ℓ_i's. Also note that an analogous structural interpretation to the one given for the HDP holds also here. Similarly to what we have shown for hierarchical NRMIs, one can establish the probability distribution of K_n.

Corollary 3. *Let $\{X_n\}_{n\geq 1}$ be as in* Theorem 2. *Then*

$$\mathbb{P}[K_n=k] = \frac{\prod_{r=1}^{k-1}(\theta_0+r\sigma_0)}{(\theta+1)_{n-1}}\,\frac{1}{\sigma\,\sigma_0^k}\sum_{q=k}^{n}\frac{\left(\frac{\theta}{\sigma}+1\right)_{q-1}}{(\theta_0+1)_{q-1}}\,\mathscr{C}(n,q;\sigma)\,\mathscr{C}(q,k;\sigma_0) \tag{17}$$

It is worth noting that in both cases covered by Corollary 2 and Corollary 3 one has the following equality

$$\mathbb{P}[K_n=k] = \sum_{t=k}^{n}\mathbb{P}[K_n'=t]\,\mathbb{P}[K_{t,0}=k] \tag{18}$$

where K_n' and $K_{n,0}$ denote the number of distinct values, out of n exchangeable observations driven by $\tilde{p}$ and $\tilde{p}_0$, respectively. In terms of the multi–room Chinese restaurant process representation, this amounts to saying that the n customers are seated at t tables and, conditional on having t tables, these are allocated into k different rooms each being identified by a specific distinct dish.

It is apparent from (18) that $K_n \overset{\mathrm{d}}{=} K_{K_n',0}$. Such an equality is the obvious starting point to investigate the asymptotic behaviour of K_n, which is described in the following

Proposition 1. *Assume that K_n is the number of distinct values in a sample $\boldsymbol{X}^{(n)}$ from an exchangeable sequence $\{X_i\}_{i\geq 1}$, governed by a HPYP defined as in (15). Then, as $n \to \infty$,*

$$\frac{K_n}{n^{\sigma\sigma_0}} \to \tilde{S}(\sigma, \sigma_0, \theta, \theta_0)$$

almost surely, where $\tilde{S}(\sigma, \sigma_0, \theta, \theta_0)$ is a finite and non–negative random variable.

The previous proposition follows from the asymptotic behaviors concerning $K_{n,0}$ and K'_n, derived in Pitman (2006). Asymptotic results analogous to Proposition 1 may be derived also for hierarchical NRMIs (see also Camerlenghi et al. (2016) for a general treatment).

Some other insight on the distribution of K_n in (17) can be gained by means of Figures 1–2, which correspond to a sample size $n = 500$. For the plain PY$(\sigma, \theta; P_0)$ it is well–known (see Lijoi *et al.*, 2007) that the parameter σ tunes the flatness: the larger σ and the flatter the distribution of K_n. On the other hand, θ has a direct impact on the location and a slight influence on variability. For the HPYP the interaction between the two hierarchies has an important effect. Indeed, increasing just one of σ and σ_0 is not enough to achieve a non–informative situation. One has to increase at least one parameter for each level of the hierarchy. This is apparent when comparing the two curves both in right and left panels in Figure 1. In other terms, if σ is low (high) and σ_0 is high (low), one should fix a high value of θ (θ_0) to obtain a less informative prior. In both the cases, the increase of either θ or θ_0 induces an expected shift of the distribution of K_{500} towards a large number of (a priori) clusters. Figure 2 shows that a large value of $\sigma\sigma_0$ has the effect of flattening the prior, regardless of the values of θ and θ_0.

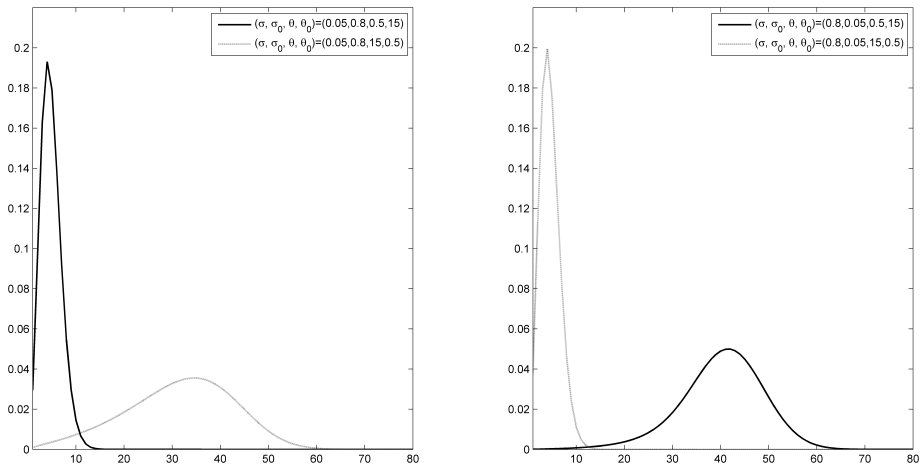


Figure 1: Prior distribution of K_n, for $n = 500$ and for different choices of the parameters $\sigma, \sigma_0, \theta, \theta_0$.

Remark 2. In general, hierarchical processes are not of Gibbs–type, since the probability of sampling a new value depends explicitly also on $n_1, \ldots, n_k$. The only exceptions we are aware of are

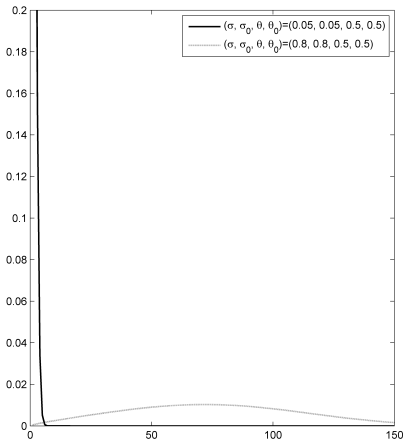


Figure 2: Prior distribution of K_n, for $n = 500$ and for different values of (σ, σ_0), when $\theta = \theta_0 = 0.5$ are fixed.

the σ–stable hierarchies (see Example 2) and a very specific HPYP (see Remark 1). Nonetheless, they still preserve a good degree of tractability and the displayed examples represent the most explicit instances of nonparametric priors outside the Gibbs framework.

Moreover, note that by using the characterization of Gibbs–type priors as σ–stable Poisson–Kingman models one could define hierarchies of Gibbs–type priors in a straightforward way. Also Theorem 2 and Corollary 3 could be extended with minor conceptual adaptations but a significant additional notational burden.

Remark 3. The previous results concern hierarchical processes involving random probability measures which are structurally the same, i.e. $Q(\cdot|\tilde{p}_0)$ and $Q'(\cdot|P_0)$ in (8) are both either NRMIs or Pitman–Yor processes. Nonetheless, the techniques used to prove Theorems 1–2 can be easily adapted to yield the EPPFs and the prediction rules of hierarchical processes where $Q(\cdot|\tilde{p}_0)$ and $Q'(\cdot|P_0)$ are the probability distribution of a NRMI and of a Pitman–Yor process, respectively, or vice versa. We omit detailed derivation of the relevant formulas in these cases as they easily follow from the previous treatment.

4 Posterior characterizations

One of the reasons of the popularity of the Dirichlet process is its conjugacy, which makes full posterior inference feasible. When moving beyond the Dirichlet process, one typically has to give up conjugacy but it is still possible to retain a considerable degree of analytical tractability as shown in James *et al.* (2009) for the large class of NRMIs. As we will show the same holds for hierarchical processes in the sense that we are able to derive analytically their posterior distributions. This completes the picture of their main distributional properties in view of Bayesian inference. As for the partition structure, the displayed results refer to cases where both levels of the hierarchy

involve a NRMI or a Pitman–Yor process, but they can be easily extended to mixed models where one hierarchy is identified by a NRMI and the other by a Pitman–Yor process. See also Remark 3.

The posterior representations we establish can be used to devise conditional algorithms, which are important in order to estimate non–linear functionals. See Section 5.2.

4.1 Hierarchical NRMI posterior

Assume that $\{X_j\}_{j\geq 1}$ is a sequence of $\mathbb{X}$–valued exchangeable observations such that $X_i|\tilde{p} \overset{\text{iid}}{\sim} \tilde{p}$, with $\tilde{p}\,|\,\tilde{p}_0 \overset{\text{iid}}{\sim} \text{NRMI}(\rho, c; \tilde{p}_0)$, $\tilde{p}_0 \sim \text{NRMI}(\rho_0, c_0; P_0)$ and P_0 is a non–atomic probability measure on $(\mathbb{X}, \mathscr{X})$. Recall that $x_1^*, \ldots, x_k^*$ denote the distinct observations featured by the sample $\boldsymbol{X}^{(n)}$ and assume that U_0 is a positive random variable whose density function, conditional on the observations and the tables, denoted as $\boldsymbol{T}^{(n)} = (T_1, \ldots, T_n)$, equals

$$f_0(u|\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}) \propto u^{|\boldsymbol{\ell}|-1} e^{-c_0\psi_0(u)} \prod_{j=1}^{k} \tau_{\ell_j,0}(u) \tag{19}$$

where both the ℓ_j's and $q_{j,t}$'s are functions of $(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)})$. The latent variable U_0 plays a role similar to U_n in the posterior representation of NRMIs obtained in James *et al.* (2009). In the current hierarchical setting, the posterior characterization is described in terms of: (i) the posterior of $\tilde{p}_0$, at the root of the hierarchy, given the data; (ii) the posterior of $\tilde{p}$, conditional on the data and on $\tilde{p}_0$. As it will be apparent from the next Theorems 3–4, the main results are stated in terms of the CRMs $\tilde{\mu}_0$ and $\tilde{\mu}$.

Theorem 3. *Suppose that $\{X_i\}_{i\geq 1}$ is an $\mathbb{X}$–valued sequence of exchangeable observations as in (9), with $\tilde{p} = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$ and $\tilde{p}_0 = \tilde{\mu}_0/\tilde{\mu}_0(\mathbb{X})$. Then the conditional distribution of $\tilde{\mu}_0$, given $(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, U_0)$, equals the distribution of the CRM*

$$\tilde{\mu}_0^* := \eta_0^* + \sum_{j=1}^{k} I_j \delta_{X_j^*} \tag{20}$$

where the two summands on the right–hand–side of (20) are independent. Moreover, η_0^* is a CRM with intensity $\nu_0(\mathrm{d}s, \mathrm{d}x) = \mathrm{e}^{-U_0 s}\rho_0(s)\mathrm{d}s\, c_0\, P_0(\mathrm{d}x)$ and the I_j's are independent and non–negative jumps with density

$$f_j(s|\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, U_0) \propto s^{\ell_j} e^{-sU_0}\rho_0(s)$$

It is now possible to provide a posterior characterization of $\tilde{\mu}$, which is expressed in terms of a non–negative random variable U_n whose density function is

$$f(u|\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}) \propto u^{n-1} e^{-c\psi(u)} \prod_{j=1}^{k} \prod_{t=1}^{\ell_j} \tau_{q_{j,t}}(u).$$

The main theorem of the section can now be stated as follows.

Theorem 4. *Suppose that $\{X_i\}_{i\geq 1}$ is an $\mathbb{X}$–valued sequence of exchangeable observations as in (9), with $\tilde{p} = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$ and $\tilde{p}_0 = \tilde{\mu}_0/\tilde{\mu}_0(\mathbb{X})$. Then*

$$\tilde{\mu}|(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, U_n, \tilde{\mu}_0^*) \overset{d}{=} \tilde{\mu}^* + \sum_{j=1}^{k} \sum_{t=1}^{\ell_j} J_{j,t}\delta_{X_j^*}, \tag{21}$$

where the two summands on the right–hand–side of (21) *are independent. Moreover,* $\tilde{\mu}^*$ *is a CRM that has intensity* $\nu(\mathrm{d}s, \mathrm{d}x) = e^{-Us}\rho(s)\mathrm{d}s\, c\, \tilde{p}_0^*(\mathrm{d}x)$*, with* $\tilde{p}_0^* = \tilde{\mu}_0^*/\tilde{\mu}_0^*(\mathbb{X})$ *and the jumps* $J_{j,t}$ *are independent and non–negative random variables whose density equals*

$$f_{j,t}(s|\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, U) \propto e^{-Us} s^{q_{j,t}}\, \rho(s).$$

Finally, $\tilde{\mu}_0$ *and* U_n *are conditionally independent, given* $(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)})$.

We point out that the CRM $\tilde{\mu}^*$ in Theorem 4 may be thought of as a *hierarchical CRM*, in the sense that, conditional on $\tilde{\mu}_0^*$, its base measure is random and equals $\tilde{p}_0^*$, while the posterior distribution of $\tilde{\mu}_0^*$ is specified in Theorem 3. Moreover, note that the expressions in Theorem 4 are somehow reminiscent of the ones provided in James *et al.* (2009), with the only difference that here we have an additional level of hierarchy. Though the following illustrative example refers to the Dirichlet process, the general results can be easily adapted to determine the posterior distribution of other specific hierarchical processes based on NRMIs.

Example 3. Let $\rho(s) = \rho_0(s) = e^{-s}/s$. In such a case, we have $\psi(u) = \psi_0(u) = \log(1+u)$ and $\tau_q(u) = \tau_{q,0}(u) = \Gamma(q)/(1+u)^q$. This implies that $U_0/(1+U_0)|(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}) \sim \mathrm{Beta}(|\boldsymbol{\ell}|, c_0)$. Next, by virtue of Theorem 3, one can specialize the posterior of $\tilde{\mu}_0$ in (20), by noting that

(a) η_0^* is a gamma CRM with intensity $\mathrm{e}^{-(1+U_0)s} s^{-1}\, \mathrm{d}s\, c_0\, P_0(\mathrm{d}x)$,

(b) $I_j \overset{\text{ind}}{\sim} \mathrm{Ga}(\ell_j, 1+U_0)$, namely

$$f_j(s|\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, U_0) = \frac{(1+U_0)^{\ell_j}}{\Gamma(\ell_j)}\, x^{\ell_j - 1} e^{-(1+U_0)x} \mathbb{1}_{(0,\infty)}(x)$$

Since the normalized distributions in (a) and (b) do not depend on the scale U_0, it follows that

$$\tilde{p}_0^* \overset{\mathrm{d}}{=} \tilde{p}_0|(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}) \sim \mathscr{D}(c_0 P_0 + \sum_{j=1}^{k} \ell_j \delta_{x_j^*})$$

where $\mathscr{D}(\alpha)$ stands for the Dirichlet process with base measure α on $\mathbb{X}$. Theorem 4, in turn, implies that the conditional distribution of $\tilde{\mu}$, given $(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, U_n)$, equals the distribution of the random measure $\tilde{\mu}^* + \sum_{j=1}^{k} H_j \delta_{x_j^*}$ where

(a') $\tilde{\mu}^*$ is a gamma CRM having intensity $\mathrm{e}^{-(1+U)s} s^{-1}\, \mathrm{d}s\, c\, \tilde{p}_0^*(\mathrm{d}x)$

(b') $H_j \sim \mathrm{Ga}(n_j, U+1)$.

Moreover, note that $U_n/(1+U_n)|(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}) \sim \mathrm{Beta}(c, n)$. Hence one has

$$\tilde{p}|(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, \tilde{p}_0^*) \sim \mathscr{D}(c\tilde{p}_0^* + \sum_{j=1}^{k} n_j \delta_{x_j^*})$$

and it depends on the latent configuration of $\boldsymbol{T}^{(n)}$ only through $\tilde{p}_0^*$. □

4.2 Hierarchical PY posterior

Results similar to those stated in Theorems 3–4 can also be given for hierarchies of Pitman–Yor processes and the proofs rely on similar techniques, so we omit them. In this section, we assume $\{X_j\}_{j\geq 1}$ is a sequence of $\mathbb{X}$–valued exchangeable random elements as in (1), with $\tilde{p}$ being a HPYP defined as in (15). Moreover, we set $\tilde{p}_0 = \tilde{\mu}_0/\tilde{\mu}_0(\mathbb{X})$ and $\tilde{p} = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$ and, despite $\tilde{\mu}_0$ and $\tilde{\mu}$ are not CRMs as apparent from (4), we are still able to establish a posterior characterization for $\tilde{p}$.

Theorem 5. *Let V_0 be such that $V_0^{\sigma_0} \sim \mathrm{Ga}(k+\theta_0/\sigma_0, 1)$. Then $\tilde{\mu}_0|(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, V_0)$ equals, in distribution, the random measure $\tilde{\mu}_0^* := \eta_0^* + \sum_{j=1}^k I_j \delta_{X_j^*}$, where η_0^* is a generalized gamma CRM whose intensity is*

$$\frac{\sigma_0}{\Gamma(1-\sigma_0)} \frac{\mathrm{e}^{-V_0 s}}{s^{1+\sigma_0}} \mathrm{d}s P_0(\mathrm{d}x),$$

$\{I_j : j = 1, \ldots, k\}$ and η_0^* are independent and $I_j \overset{\text{ind}}{\sim} \mathrm{Ga}(\ell_j - \sigma_0, V_0)$, for $j = 1, \ldots, k$.

One can now state a posterior characterization of $\tilde{\mu}$, which is the key for determining the posterior of a HPYP.

Theorem 6. *Let V be such that $V^\sigma \sim \mathrm{Ga}(|\boldsymbol{\ell}| + \theta/\sigma, 1)$. Then*

$$\tilde{\mu}|(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)}, V, \tilde{\mu}_0^*) \overset{d}{=} \tilde{\mu}^* + \sum_{j=1}^k H_j \delta_{X_j^*} \tag{22}$$

where the two summands in the above expression are independent. Moreover, $\tilde{\mu}^*$ is a generalized gamma CRM with intensity

$$\frac{\sigma}{\Gamma(1-\sigma)} \frac{\mathrm{e}^{-Vs}}{s^{1+\sigma}} \mathrm{d}s\, \tilde{p}_0^*(\mathrm{d}x)$$

$\tilde{p}_0^* = (\eta_0^* + \sum_{j=1}^k I_j \delta_{x_j^*})/(\eta_0^*(\mathbb{X}) + \sum_{j=1}^k I_j)$ and $H_j \overset{\text{ind}}{\sim} \mathrm{Ga}(n_j - \ell_j\sigma, V)$.

The posterior distribution of $\tilde{p}$ can be, finally, deduced by normalizing the random measure (22) and one can further simplify the resulting representation by integrating out V_0 and V. This is effectively illustrated by the following.

Theorem 7. *The posterior distribution of $\tilde{p}_0$, conditional on $(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)})$, equals the distribution of the random probability measure*

$$\sum_{j=1}^k W_j \delta_{X_j^*} + W_{k+1}\, \tilde{p}_{0,k} \tag{23}$$

where $(W_1, \ldots, W_k)$ is a k–variate Dirichlet random vector with parameters $(\ell_1 - \sigma_0, \ldots, \ell_k - \sigma_0, \theta_0 + k\sigma_0)$, $W_{k+1} = 1 - \sum_{i=1}^k W_i$ and $\tilde{p}_{0,k} \sim \mathrm{PY}(\sigma_0, \theta_0 + k\sigma_0; P_0)$. Moreover, conditional on $(\tilde{p}_0, \boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)})$, the posterior distribution of $\tilde{p}^* = (\tilde{\mu}^* + \sum_{j=1}^k H_j \delta_{x_j^*})/(\tilde{\mu}^*(\mathbb{X}) + \sum_{j=1}^k H_j)$ equals the distribution of the random measure

$$\sum_{j=1}^k W_j^*\, \delta_{X_j^*} + W_{k+1}^*\, \tilde{p}_k \tag{24}$$

where $(W_1^*, \ldots, W_k^*)$ is a k–variate Dirichlet random vector with parameters $(n_1 - \ell_1\sigma, \ldots, n_k - \ell_k\sigma, \theta + |\boldsymbol{\ell}|\sigma)$, $W_{k+1}^* = 1 - \sum_{j=1}^k W_j^*$ and $\tilde{p}_k \,|\, \tilde{p}_0 \sim \mathrm{PY}(\sigma, \theta + |\boldsymbol{\ell}|\sigma; \tilde{p}_0)$.

5 Algorithms

The previous theoretical findings are crucial for establishing computational algorithms which allow an effective implementation of hierachical processes providing approximations to posterior inferences and to the quantification of uncertainty associated to them. Here we provide two sets of algorithms. The first one is based on the marginalization of $\tilde{p}$ and is in the spirit of the traditional Blackwell–MacQueen urn scheme. The second algorithm, on the contrary, simulates the trajectories of $\tilde{p}$ from its posterior distribution and falls within the category of so–called conditional algorithms.

5.1 Blackwell–MacQueen urn schemes

Here we devise a generalized Blackwell–MacQueen urn scheme that allows to generate elements from an exchangeable sequence governed by a hierarchical prior as in (8). Such a tool becomes of great importance if one wants to address prediction problems. Indeed, conditional on observed data $\boldsymbol{X}^{(n)}$ one may be interested in predicting specific features of additional and unobserved samples

$$\boldsymbol{X}^{(m|n)} = (X_{n+1}, \ldots, X_{n+m})$$

such as, e.g., the number of new distinct values or the number of distinct values that have appeared r times in the observed sample $\boldsymbol{X}^{(n)}$ that will be recorded in $\boldsymbol{X}^{(m|n)}$. Such an algorithm suits also density estimation problems based on a mixture model, where the predictives are to be considered for the case $m = 1$ and to be combined with a kernel. For the sake of illustration we confine ourselves to the HPYP prior (15), though the algorithms easily carry over also to the class of hierarchical NRMIs, with suitable adaptations.

Let $\boldsymbol{T}^{(n)}$ be an $\mathbb{X}^n$–valued vector of latent tables associated to the observations $\boldsymbol{X}^{(n)}$. They are from an exchangeable sequence $\{T_i\}_{i\geq 1}$ such that $T_i \,|\, \tilde{q} \overset{\text{iid}}{\sim} \tilde{q}$ and $\tilde{q} \sim \text{PY}(\sigma, \theta; P_0)$. In terms of the multi–room Chinese restaurant process description, T_i can be thought of as the label of the table where the i–th costumer X_i is seated. It is further assumed that $\boldsymbol{X}^{(n)}$ features k distinct values $x_1^*, \ldots, x_k^*$, with respective frequencies $n_1, \ldots, n_k$, and $\boldsymbol{T}^{(n)}$ has $|\boldsymbol{\ell}|$ distinct values

$$t_{j,1}^*, \ldots, t_{j,\ell_j}^* \quad \text{with} \quad q_{j,r} = \sharp\{i:\ T_i = t_{j,r}\} \quad r = 1, \ldots, \ell_j$$

for $j = 1, \ldots, k$. From Theorem 2, the marginal probability distribution of $(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)})$ on $(\mathbb{X}^{2n}, \mathscr{X}^{2n})$ is equivalently identified by the joint probability distribution of the partitions induced by $(\boldsymbol{X}^{(n)}, \boldsymbol{T}^{(n)})$ and of the distinct values associated to such random partitions, which boils down to

$$\prod_{j=1}^{k} P_0(\mathrm{d}x_j^*) \prod_{i=1}^{\ell_j} P_0(\mathrm{d}t_{j,i}^*) \frac{\prod_{r=1}^{k-1}(\theta_0 + r\sigma_0)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}} \frac{\prod_{r=1}^{|\boldsymbol{\ell}|-1}(\theta + r\sigma)}{(\theta+1)_{n-1}} \prod_{j=1}^{k}(1-\sigma_0)_{\ell_j - 1} \prod_{i=1}^{\ell_j}(1-\sigma)_{q_{j,i}-1} \tag{25}$$

because of non–atomicity of P_0. The full conditionals of the Gibbs sampler we are going to propose can now be determined from (25). If V is a variable that is a function of $(T_1, \ldots, T_{n+m})$ and of $(X_{n+1}, \ldots, X_{n+m})$, use $V^{(-r)}$ to denote the generic value of the variable V after removal of T_r, for $r = 1, \ldots, n$, and of (X_r, T_r), for $r = n+1, \ldots, n+m$.

(1) At $s = 0$, start from an initial configuration $X_{n+1}^{(0)}, \ldots, X_{n+m}^{(0)}$ and $T_1^{(0)}, \ldots, T_{n+m}^{(0)}$.

(2) At iteration $s \geq 1$

(2.a) With $X_r = X^*_h$ generate latent variables $T^{(s)}_r$, for $r = 1, \dots, n$,

$$\mathbb{P}[T_r = \text{“new”} | \cdots] \propto w_{h,r} \, \frac{(\theta + |\boldsymbol{\ell}^{(-r)}|\sigma)}{(\theta_0 + |\boldsymbol{\ell}^{(-r)}|)}$$

$$\mathbb{P}[T_r = T^{*,(-r)}_{h,\kappa} | \cdots] \propto (q^{(-r)}_{h,\kappa} - \sigma) \qquad \text{for } \kappa = 1, \cdots, \ell^{(-r)}_h$$

where $w_{h,r} = (\ell^{(-r)}_h - \sigma_0)$ if $\ell^{(-r)}_h \geq 1$ and $w_{h,r} = 1$ otherwise.

(2.b) For $r = n+1, \dots, n+m$, generate $(X^{(s)}_r, T^{(s)}_r)$ from the following predictive distributions

$$\mathbb{P}[X_r = \text{“new”}, T_r = \text{“new”} | \cdots] = \frac{(\theta_0 + (k + j^{(-r)})\sigma_0)}{(\theta_0 + |\boldsymbol{\ell}^{(-r)}|)} \, \frac{(\theta + |\boldsymbol{\ell}^{(-r)}|\sigma)}{(\theta + n + m - 1)}$$

while, for any $h = 1, \cdots, k + j^{(-r)}$ and $\kappa = 1, \cdots, \ell^{(-r)}_h$,

$$\mathbb{P}[X_r = X^{*,(-r)}_h, T_r = \text{“new”} | \cdots] = \frac{(\ell^{(-r)}_h - \sigma_0)}{(\theta_0 + |\boldsymbol{\ell}^{(-r)}|)} \, \frac{(\theta + |\boldsymbol{\ell}^{(-r)}|\sigma)}{(\theta + n + m - 1)}$$

$$\mathbb{P}[X_{n+1} = X^{*,(-r)}_h, T_{n+1} = T^{*,(-r)}_{h,\kappa} | \cdots] = \frac{q^{(-r)}_{h,\kappa} - \sigma}{\theta + n + m - 1}$$

where $T^{*,(-r)}_{h,\kappa}$ stands for the distinct values in the sample after the removal of T_r. When $n \geq 1$, this algorithm yields approximate samples $(X^{(s)}_{n+1}, \dots, X^{(s)}_{n+m})$, for $s = 1, \dots, S$, from the exchangeable sequence $\{X_i\}_{i \geq 1}$ that can be used, e.g., for addressing prediction problems on specific features of the additional sample $X_{n+1}, \dots, X_{n+m}$ as detailed in Section 6.

5.2 Simulation of $\tilde{p}$ from its posterior distribution

The posterior representations determined in Section 4 are the main ingredient for simulating the trajectories of $\tilde{p}$ from its posterior distribution. Here we address this issue, focusing on the Pitman–Yor process and the obvious adaptations may be easily deduced to sample posterior hierarchical NRMIs. The merit of this sampling scheme, compared to the one discussed in Section 5.1, relies on the possibility to estimate non–linear functionals of $\tilde{p}$ such as, for example, credible intervals that quantify the uncertainty associated to proposed point estimators.

For the sake of clarity, assume that $\mathbb{X} = \mathbb{R}^+$. The Ferguson–Klass representation of a CRM provided in Ferguson & Klass (1972), combined with Theorems 5–6, entails

$$\eta^*_0((0,t]) = \sum_{h=1}^{\infty} J^{(0)}_h \mathbb{1}\{M^{(0)}_h \leq P_0((0,t])\} \tag{26}$$

with $M^{(0)}_1, M^{(0)}_2, \dots \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$, the jumps' heights $J^{(0)}_h$ are decreasing and may be recovered from the identity

$$S^{(0)}_h = \frac{\sigma_0}{\Gamma(1-\sigma_0)} \int_{J^{(0)}_h}^{\infty} e^{-V_0 s} s^{-1-\sigma_0} \mathrm{d}s \tag{27}$$

where $S^{(0)}_1, S^{(0)}_2, \dots$ are the points of a standard Poisson process on $\mathbb{R}^+$, i.e. $S^{(0)}_h - S^{(0)}_{h-1}$ are exponential random variables with unit mean. Similarly one obtains that

$$\tilde{\mu}^*((0,t]) = \sum_{h=1}^{\infty} J_h \mathbb{1}\{M_h \leq P_0((0,t])\}. \tag{28}$$

As in (27), the J_h's are ordered and solve the equation

$$S_h = \frac{\sigma}{\Gamma(1-\sigma)} \int_{J_h}^{\infty} e^{-Vs} s^{-1-\sigma_0} \mathrm{d}s \tag{29}$$

with $S_1, S_2, \ldots$ denoting, again, the points of a standard Poisson process on $\mathbb{R}^+$ with unit mean. The terms defining the series representations (26) and (28) can be sampled and, hence, an approximate sampled trajectory of $\tilde{p}$ from its posterior distribution can be determined according to the following procedure:

`(1)` Fix $\varepsilon > 0$ and generate an approximate trajectory of $\tilde{p}_0$ from the posterior distribution derived in Theorem 5:

`(1.a)` generate a random variable $V_{0,\sigma_0} \sim \mathrm{Ga}(k + \theta_0/\sigma_0, 1)$ and put $V_0 = (V_{0,\sigma_0})^{1/\sigma_0}$;

`(1.b)` generate the weights $I_j \sim \mathrm{Ga}(\ell_j - \sigma_0, V_0)$, for $j = 1, \ldots, k$, independently;

`(1.c)` for any $h \geq 1$ generate $S_h^{(0)}$ from the Poisson process $\left\{S_h^{(0)}\right\}_{h\geq 1}$ with unit rate on $\mathbb{R}^+$;

`(1.d)` determine $J_h^{(0)}$ from (27);

`(1.e)` if $J_h^{(0)} \leq \varepsilon$ stop and set $h = H_0$, otherwise $h \to h+1$ and go to `(1.c)`;

`(1.f)` generate the random variables $M_1^{(0)}, \ldots, M_{H_0}^{(0)} \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$.

An approximate draw of $\tilde{p}_0$ would, then, be

$$\tilde{p}_0^*((0,t]) \approx \frac{\sum_{h=1}^{H_0} J_h^{(0)} \mathbb{1}\{M_h^{(0)} \leq P_0((0,t])\} + \sum_{j=1}^{k} I_j \delta_{x_j^*}((0,t])}{\sum_{h=1}^{H_0} J_h^{(0)} + \sum_{j=1}^{k} I_j}.$$

Hence, conditional on this approximate realization of $\tilde{p}_0^*$, one can proceed to the second step of the algorithm

`(2)` Fix $\varepsilon > 0$ and sample an approximate trajectory of $\tilde{p}$ from its posterior distribution derived in Theorem 6:

`(2.a)` generate a random variable $V_\sigma \sim \mathrm{Ga}(|\boldsymbol{\ell}| + \theta/\sigma, 1)$ and put $V = (V_\sigma)^{1/\sigma}$;

`(2.b)` generate the weights $H_j \sim \mathrm{Ga}(n_j - \ell_j\sigma, V)$, for $j = 1, \ldots, k$, independently;

`(2.c)` for any $h \geq 1$ sample S_h from the Poisson process $\{S_h\}_{h\geq 1}$ with unit rate on $\mathbb{R}^+$;

`(2.d)` determine J_h from (29);

`(2.e)` if $J_h \leq \varepsilon$ stop and set $h = H_0$, otherwise $h \to h+1$ and go to `(1.c)`;

`(1.f)` generate the random variables $M_1, \ldots, M_{H_0} \overset{\text{iid}}{\sim} \mathrm{U}(0,1)$.

An approximate trajectory of $\tilde{p}$ from the corresponding posterior distribution can be obtained as follows

$$\tilde{p}((0,t]) \approx \frac{\sum_{h=1}^{H} J_h \mathbb{1}\{M_h \leq \tilde{p}_0^*((0,t])\} + \sum_{j=1}^{k} H_j \delta_{x_j^*}((0,t])}{\sum_{h=1}^{H} J_h + \sum_{j=1}^{k} H_j}.$$

6 Application to a species sampling problem

A natural application area of hierarchical processes is represented by species sampling problems. Consider a population of individuals consisting of species, possibly infinite, labeled $\boldsymbol{Z} = \{Z_i\}_{i\geq 1}$, with unknown proportions $\boldsymbol{p} = \{p_i\}_{i\geq 1}$. If $\{X_n\}_{n\geq 1}$ is the sequence of observed species labels, it is then natural to take

$$X_i|(\boldsymbol{p}, \boldsymbol{Z}) \overset{\text{iid}}{\sim} \sum_{i\geq 1} p_i\, \delta_{Z_i}.$$

Given a basic sample of size n, namely $\boldsymbol{X}^{(n)} = (X_1, \cdots, X_n)$, containing $k \leq n$ distinct species with respective labels $x_1^*, \ldots, x_k^*$ and frequencies $N_{1,n}, \ldots, N_{k,n}$, possible inferential problems of interest are

(i) The estimation of the *sample coverage*, namely the proportion of the population species that have been observed in $\boldsymbol{X}^{(n)}$.

(ii) The *prediction* of $K_m^{(n)} = K_{n+m} - K_n$, namely the number of the *new species* that will be seen in further additional sample of size m.

(iii) The estimation of the (m, r)–*discovery probability*, which coincides with the probability of discovering a species at the $(n+m+1)$–th draw that has been observed r times in $\boldsymbol{X}^{(n+m)}$, given the initial sample of size n and without observing the outcomes of the additional sample $\boldsymbol{X}^{(m|n)}$ of size m. If $\Delta_{r,n+m} = \{j : \sum_{i=1}^{n+m} \mathbb{1}_{\{Z_j\}}(X_i) = r\}$ is the number of species with frequency r in $\boldsymbol{X}^{(n+m)}$, then the discovery probability equals

$$D_m^0 := \sum_{i\in\Delta_{0,n+m}} p_i$$

for $m \geq 0$. For $r = 0$ this is also known as the m–*discovery* probability.

These issues were first addressed in Good (1953); Good & Toulmin (1956), where the authors derived frequentist estimators also known in the literature as *Good–Turing* and *Good–Toulmin* estimators. More recently Lijoi *et al.* (2007) have introduced a Bayesian nonparametric approach to this problem, with further developments in Favaro *et al.* (2012, 2013). The key idea is to randomize the proportions p_i's and consider them as generated by a Gibbs–type prior. This leads to closed form expressions for the estimator of the (m, r)–discover probability and for $K_m^{(n)}$. See De Blasi *et al.* (2015) for a review of these methods. Besides these investigations, very little has been explored in a non–Gibbs context. The only exceptions we are aware of are Lijoi *et al.* (2005) and Favaro *et al.* (2011). Here we want to employ hierarchical processes to face this type of problem. Though we are not able to identify closed form expressions of the estimators for addressing (i)–(ii) as for Gibbs–type priors, we can still determine approximations based on the algorithm developed in Section 5.1.

As an illustrative example, we focus on a genomic application even if one can easily think of similar problems arising in ecological applications, where one has populations of animals or plants, or in economics, linguistics, topic modeling, and so forth. The specific application we are interested in involves Expressed Sequence Tags (EST) data. Such data are generated by partially sequencing randomly isolated gene transcripts that have been converted into cDNA. EST have been playing an important role in the identification, discovery and characterization of organisms as they are a cost–effective tool in genomic technologies. The resulting transcript sequences and their corresponding abundances are the main focus of interest as they identify the distinct genes

and their expression levels. Indeed, an EST sample of size n, consists of K_n distinct genes with frequencies, or expression levels, $N_{1,n}, \ldots, N_{K_n,n}$ where $\sum_{j=1}^{K_n} N_{j,n} = n$. Based on these data, one is interested in assessing some features of the whole library. For the sake of comparison, we use the same dataset as in Lijoi *et al.* (2007), which is obtained from a cDNA library made from the 0 mm to 3 mm buds of tomato flowers (see also Mao (2004)). It consists of $n = 2586$ EST showing $K_n = 1825$ distinct values or genes.

We assume the data are exchangeable and that $\tilde{p}$ is a hierarchical Pitman–Yor random probability measure as in (15). We assign a non–informative prior to $(\sigma, \sigma_0, \theta, \theta_0)$ of the form

$$(\sigma, \sigma_0, \theta, \theta_0) \sim \mathrm{U}(0,1) \times \mathrm{U}(0,1) \times \mathrm{Gam}(20, 50^{-1}) \times \mathrm{Gam}(20, 50^{-1}).$$

For the tomato flower dataset, considering an additional sample of size m, the quantities of interest are the $(m,0)$–discovery probability and the expected number of new distinct genes, $K_n^{(m)}$, recorded in the additional sample. To this end trajectories of the additional unobserved sample $\boldsymbol{X}^{(m|n)}$, are generated conditional on $\boldsymbol{X}^{(n)} = (X_1, \ldots, X_n)$. This is done by resorting to the algorithm devised in Section 5.1 with the addition of a Metropolis–Hastings step to update the model parameters.

If $\boldsymbol{X}_s^{(m|n)} = (X_{n+1}^{(s)}, \ldots, X_{n+m}^{(s)})$ denotes a realization of $\boldsymbol{X}^{(m|n)}$, one the can evaluate the discovery probability as

$$\widehat{D}_m^0 \approx \frac{1}{S} \sum_{s=1}^{S} \mathbb{P}\Big[X_{n+m+1} = \text{“new”} \,\Big|\, \boldsymbol{X}^{(n)}, \boldsymbol{X}_s^{(m|n)}\Big],$$

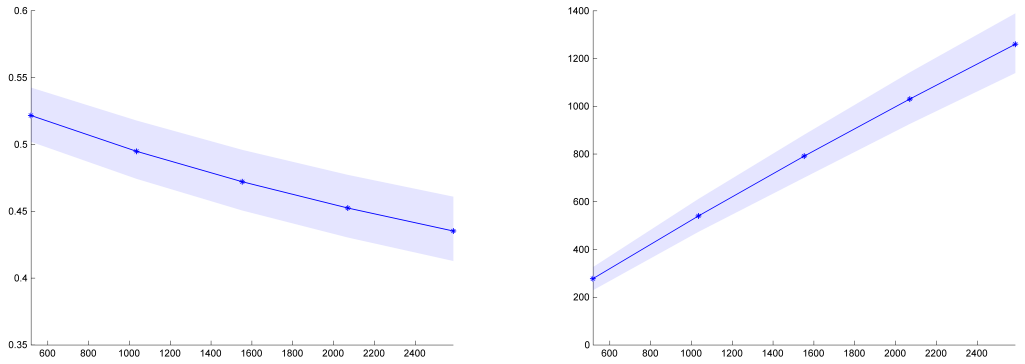
which is easy to calculate, since $\mathbb{P}[X_{n+m+1} = \text{“new”} | \boldsymbol{X}^{(n)}, \boldsymbol{X}_s^{(m|n)}]$ is a one–step prediction. As for the estimation of the new distinct genes recorded in an additional sample of size m one has:

$$\hat{K}_n^{(m)} \approx \frac{1}{S} \sum_{s=1}^{S} \sum_{r=1}^{m} \mathbb{1}_{\{X_1, \ldots, X_{n+r-1}\}^c}(X_{n+r}^{(s)}).$$

The numerical outputs are based on $S = 15,000$ iterations of the Gibbs sampler after $5,000$ burn–in sweeps. We choose $m \in \{517, 1034, 1552, 2069, 2586\}$, which correspond to $20\%, 40\%, 60\%, 80\%$ and 100% of the size of the basic sample in order to make direct comparison with the results contained in Lijoi *et al.* (2007). There the prior distribution of $\tilde{p}$ is a Poisson–Dirichlet process, which is of Gibbs–type and the two parameters are selected on the basis of a maximum likelihood procedure.

Table 1 displays the expected number of new species, the discovery probabilities and the corresponding 95% highest posterior density (HPD) intervals. Figure 3(a) shows the decay of the discovery probability as m increases, and Figure 3(b) the number of new genes detected for different sizes of the additional sample. The point estimates we obtain are very similar to those contained in Lijoi *et al.* (2007) thus showing that HPYPs are effective competitors w.r.t. to the standard Pitman–Yor process. In contrast, the HPD intervals are significantly wider for the HPYP model. This is due in part to the different prior specification for the parameters but also hints to the fact that the HPYP yields a more flexible prediction scheme. In a future applied paper the performances of the two models will be compared in detail in both species sampling and mixture setups.

Finally the posterior means of parameters are $\mathbb{E}[(\theta_0, \sigma_0, \theta, \sigma) \,|\, \boldsymbol{X}^{(n)}] = (520.8, 0.9128, 1179.7, 0.6155)$.



(a) (b)

Figure 3: Tomato flower: (a) decay of the discovery probability as a function of m; (b) number of new genes detected in additional samples, for different values of m.

m	$\widehat{K}_n^{(m)}$	HPD (95%)	$\widehat{D}_m^0$	HPD (95%)
517	277.90	(229.38,326.13)	0.5217	(0.5020 , 0.5426)
1034	540.62	(474.05, 611.90)	0.4949	(0.4743 , 0.5179)
1552	791.24	(701.10, 880.85)	0.4721	(0.4506 , 0.4960)
2069	1030.26	(926.00, 1142.00)	0.4525	(0.4304 , 0.4773)
2586	1259.82	(1139.10, 1389.15)	0.4353	(0.4128 , 0.4609)

Table 1: Tomato flower: posterior expected number of new species, discovery probabilities and 95% highest posterior density intervals, for different values of m

7 Concluding remarks

In this paper we have introduced and investigated a broad class of nonparametric priors with a more general and flexible predictive structure than the one implied by Gibbs–type priors. Although the higher degree of flexibility is not for free, we have shown that hierarchical NRMIs and HPYPs are still analytical tractable. Indeed, we obtained closed form expressions for the probability function of the exchangeable random partition induced by the data and posterior characterizations that make them viable alternatives to Gibbs–type priors in several applied settings, even beyond the species sampling setting considered in Section 6.

The most natural development of our work is the extension to a multiple–sample framework, where data are recorded under different experimental conditions. These arise in several applied settings that involve, for example, multicenter studies, change–point analysis, clinical trials, topic modeling and so on. In all these cases, a source of heterogeneity affects data arising from different

samples: the exchangeability assumption is, thus, not realistic and is replaced by the weaker condition of *partial exchangeability*. In other words, instead of a single exchangeable sequence $\{X_i\}_{i\geq 1}$, one considers a collection of d sequences $\boldsymbol{X}_j = \{X_{i,j}\}_{i\geq 1}$, for $j = 1, \ldots, d$, corresponding to d different, though related, experiments. They are such that each $\boldsymbol{X}_j$ is itself exchangeable while exchangeability does not hold true across any two samples $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$, for any $j \neq k$. Any two distinct samples are conditionally independent but are not identically distributed. This would be along the lines of Teh *et al.* (2006) and would lead to a generalization of the popular HDP to the case of normalized random measures. A first study in this direction can be found in Camerlenghi et al. (2016).

A Appendix

A.1 Proof of Theorem 1

Recall that the n–th order derivative of $\mathrm{e}^{-m\psi(u)}$, with $\psi(u) = \int_0^\infty [1 - \mathrm{e}^{-uv}]\, \rho(v)\, \mathrm{d}v$, is given by

$$\frac{\mathrm{d}^n}{\mathrm{d}u^n}\mathrm{e}^{-m\psi(u)} = \sum_{\pi} \frac{\mathrm{d}^{|\pi|}}{\mathrm{d}x^{|\pi|}}(\mathrm{e}^{-mx})\Big|_{x=\psi(u)} \prod_{B\in\pi} \frac{\mathrm{d}^{|B|}}{\mathrm{d}u^{|B|}}\psi(u) \tag{30}$$

where $m \in \mathbb{R}$, and the sum is extended over all partitions π of $[n] = \{1, \ldots, n\}$. See, e.g., (10) in Hardy (2006). Clearly (30) can be rewritten as

$$\frac{\mathrm{d}^n}{\mathrm{d}u^n}\mathrm{e}^{-m\psi(u)} = \sum_{i=1}^{n}(-m)^i \mathrm{e}^{-m\psi(u)} \sum_{\pi:|\pi|=i} \prod_{B\in\pi} \frac{\mathrm{d}^{|B|}}{\mathrm{d}u^{|B|}}\psi(u).$$

Since to each unordered partition π of size i there correspond $i!$ ordered partitions of the set $[n]$ into i components, which are obtained by permuting the elements of π in all the possible ways, one has

$$\begin{aligned}\frac{\mathrm{d}^n}{\mathrm{d}u^n}\mathrm{e}^{-m\psi(u)} &= \sum_{i=1}^{n}(-m)^i \mathrm{e}^{-m\psi(u)} \frac{1}{i!} \sum_{\pi:|\pi|=i}^{\mathrm{o}} \prod_{B\in\pi} \frac{\mathrm{d}^{|B|}}{\mathrm{d}u^{|B|}}\psi(u) \\ &= \sum_{i=1}^{n}(-m)^i \mathrm{e}^{-m\psi(u)} \frac{1}{i!} \sum_{(*)} \binom{n}{q_1, \cdots, q_i} \frac{\mathrm{d}^{q_1}}{\mathrm{d}u^{q_1}}\psi(u) \cdots \frac{\mathrm{d}^{q_i}}{\mathrm{d}u^{q_i}}\psi(u)\end{aligned}$$

where $\sum^{\mathrm{o}}$ is the sum over the ordered partitions of the set $[n]$ while the sum $(*)$ runs over all vectors $(q_1, \cdots, q_i)$ of positive integers such that $\sum_{j=1}^{i} q_j = n$. The second equality follows upon noting that the derivative of ψ depends only on the number of elements within each component of the partition π and that the number of partitions π of the set $[n]$ containing i elements $(B_1, \cdots, B_i)$, with $(|B_1|, \cdots, |B_i|) = (q_1, \cdots, q_i)$, equals the multinomial coefficient above. It is now easy to see that

$$(-1)^n \frac{\mathrm{d}^n}{\mathrm{d}u^n} e^{-m\psi(u)} = e^{-m\psi(u)} \sum_{i=1}^{n} m^i \xi_{n,i}(u) \tag{31}$$

where we have set

$$\xi_{n,i}(u) = \sum_{(*)} \frac{1}{i!} \binom{n}{q_1, \cdots, q_i} \tau_{q_1}(u) \cdots \tau_{q_i}(u) \tag{32}$$

with $\tau_q(u) = \int_0^\infty v^q \, \mathrm{e}^{-uv} \, \rho(v) \, \mathrm{d}v$. In view of this, for any $x_1 \neq \cdots \neq x_k$ one can now evaluate

$$M_{n_1,\ldots,n_k}(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) = \mathbb{E} \prod_{j=1}^{k} \tilde{p}^{n_j}(\mathrm{d}x_j),$$

then the EPPF $\Pi_k^{(n)}$ will be recovered integrating this quantity over $\mathbb{X}^k$ with respect to $x_1, \ldots, x_k$. Indeed, if one sets $A_{j,\varepsilon} = B(x_j; \varepsilon)$ a ball of radius ε around x_j, with $\varepsilon > 0$ small enough so that $A_{i,\varepsilon} \cap A_{j,\varepsilon} = \varnothing$ for any $i \neq j$, then

$$\begin{aligned}
M_{n_1,\ldots,n_k}(A_{1,\varepsilon} \times \cdots \times A_{k,\varepsilon}) &= \mathbb{E}\,\mathbb{E}\Big[\prod_{j=1}^{k} \tilde{p}^{n_j}(A_{j,\varepsilon}) \Big| \tilde{p}_0 \Big] \\
&= \frac{1}{\Gamma(n)} \, \mathbb{E} \int_0^\infty u^{n-1} \mathbb{E}[\mathrm{e}^{-u\tilde{\mu}(\mathbb{X}_\varepsilon^*)} | \tilde{p}_0] \prod_{j=1}^{k} \mathbb{E}[\mathrm{e}^{-u\tilde{\mu}(A_{j,\varepsilon})} \tilde{\mu}^{n_j}(A_{j,\varepsilon}) | \tilde{p}_0] \, \mathrm{d}u \\
&= \frac{1}{\Gamma(n)} \, \mathbb{E} \int_0^\infty u^{n-1} \mathrm{e}^{-c\psi(u)\tilde{p}_0(\mathbb{X}_\varepsilon^*)} \prod_{j=1}^{k} \Big((-1)^{n_j} \frac{\mathrm{d}^{n_j}}{\mathrm{d}u^{n_j}} e^{-c\psi(u)\tilde{p}_0(A_{j,\varepsilon})} \Big) \, \mathrm{d}u
\end{aligned}$$

where $\mathbb{X}_\varepsilon^* = \mathbb{X} \setminus (\cup_{j=1}^k A_{j,\varepsilon})$. If $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_k) \in \times_{i=1}^k \{1, \ldots, n_i\}$, by virtue of (31) one obtains

$$\begin{aligned}
&M_{n_1,\ldots,n_k}(A_{1,\varepsilon} \times \cdots \times A_{k,\varepsilon}) \\
&= \frac{1}{\Gamma(n)} \int_0^\infty u^{n-1} \, \mathrm{e}^{-c\,\psi(u)} \sum_{\boldsymbol{\ell}} c^{|\boldsymbol{\ell}|} \left(\mathbb{E} \prod_{j=1}^{k} \tilde{p}_0^{\ell_j}(A_{j,\varepsilon}) \, \xi_{n_j,\ell_j}(u) \right) \mathrm{d}u \\
&= \frac{1}{\Gamma(n)} \sum_{\boldsymbol{\ell}} M^0_{\ell_1,\ldots,\ell_k}(A_{1,\varepsilon} \times \cdots \times A_{k,\varepsilon}) \, c^{|\boldsymbol{\ell}|} \int_0^\infty u^{n-1} \, \mathrm{e}^{-c\psi(u)} \left(\prod_{j=1}^{k} \xi_{\ell_j,n_j}(u) \right) \mathrm{d}u
\end{aligned}$$

where $M^0_{\ell_1,\ldots,\ell_k}(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) = \mathbb{E} \prod_{j=1}^k \tilde{p}_0^{\ell_j}(\mathrm{d}x_j)$. Since P_0 in (9) is non–atomic, Proposition 3 in James *et al.* (2009) yields

$$M^0_{\ell_1,\ldots,\ell_k}(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) = \left(\prod_{j=1}^{k} P_0(\mathrm{d}x_j) \right) \Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\ell_1, \ldots, \ell_k)$$

for any $(x_1, \ldots, x_k) \in \mathbb{X}^k$ such that $x_1 \neq \cdots \neq x_k$, where $\Phi_{\cdot,0}^{(\cdot)}$ is the EPPF of a NRMI with parameter (c_0, ρ_0) recalled in (7). Hence, by letting $\varepsilon \to 0$, one has

$$\begin{aligned}
M_{n_1,\ldots,n_k}(\mathrm{d}x_1, \ldots, \mathrm{d}x_k) = \frac{\prod_{j=1}^k P_0(\mathrm{d}x_j)}{\Gamma(n)} \sum_{\boldsymbol{\ell}} c^{|\boldsymbol{\ell}|} \Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\ell_1, \ldots, \ell_k) \\
\times \int_0^\infty u^{n-1} \, \mathrm{e}^{-c\psi(u)} \left(\prod_{j=1}^{k} \xi_{\ell_j,n_j}(u) \right) \mathrm{d}u.
\end{aligned}$$

The result follows by taking into account the definition of $\xi_{n,\ell}$ in (32). □

A.2 Proof of Theorem 2

The proof works along similar lines as the proof of Theorem 1, the key difference being that one has to take into account the appropriate change of measure in (4). Indeed, in this case one has

$$\mathbb{E}\prod_{j=1}^{k}\tilde{p}^{n_j}(\mathrm{d}x_j) = \mathbb{E}\,\mathbb{E}\Big[\prod_{j=1}^{k}\tilde{p}^{n_j}(\mathrm{d}x_j)\,\Big|\,\tilde{p}_0\Big]$$

$$= \frac{\sigma\,\Gamma(\theta)}{\Gamma(\theta/\sigma)}\,\frac{1}{\Gamma(\theta+n)}\int_0^{\infty} u^{\theta+n-1}\mathbb{E}\,\mathbb{E}\Big[\mathrm{e}^{-u\tilde{\mu}(\mathbb{X})}\prod_{j=1}^{k}\tilde{\mu}^{n_j}(\mathrm{d}x_j)\,\Big|\,\tilde{p}_0\Big]$$

where, conditional on $\tilde{p}_0$, $\tilde{\mu}$ is a σ–stable CRM with $\mathbb{E}[\tilde{p}\,|\,\tilde{p}_0]=\tilde{p}_0$. Hence, the proof easily follows upon recalling that in this case $\tau_q(u)=\sigma\,(1-\sigma)_{q-1}$, for any $q\geq 1$, and

$$\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\ell_1,\ldots,\ell_k) = \frac{\prod_{i=1}^{k-1}(\theta_0+i\sigma_0)}{(\theta_0+1)_{|\boldsymbol{\ell}|-1}}\prod_{j=1}^{k}(1-\sigma_0)_{\ell_j-1}.$$

□

A.3 Proof of Corollary 2

We first recall that

$$\mathbb{P}[K_n=k] = \frac{1}{k!}\sum_{(n_1,\ldots,n_k)\in\Delta_{k,n}}\binom{n}{n_1,\ldots,n_k}\Pi_k^{(n)}(n_1,\ldots,n_k)$$

with $\Pi_k^{(n)}$ as in (10). If $\Delta_{k,t}(\boldsymbol{n})=\{\boldsymbol{\ell}=(\ell_1,\ldots,\ell_k)\in\times_{i=1}^{k}\{1,\ldots,n_i\}\ \text{s.t.}\ |\boldsymbol{\ell}|=t\}$, one can rewrite $\Pi_k^{(n)}$ as follows

$$\Pi_k^{(n)}(n_1,\cdots,n_k) = \sum_{t=k}^{n}\sum_{\boldsymbol{\ell}\in\Delta_{k,t}(\boldsymbol{n})}\Phi_{k,0}^{(t)}(\ell_1,\cdots,\ell_k)\prod_{j=1}^{k}\frac{1}{\ell_j!}\sum_{\boldsymbol{q}_j}\binom{n_j}{q_{j,1},\cdots,q_{j,\ell_j}}$$

$$\times\,\Phi_t^{(n)}(q_{1,1},\cdots,q_{1,\ell_1},\cdots,q_{k,1},\cdots,q_{k,\ell_k}).$$

In view of this one has

$$\mathbb{P}[K_n=n] = \frac{1}{k!}\sum_{t=k}^{n}\sum_{\boldsymbol{n}\in\Delta_{k,n}}\sum_{\boldsymbol{\ell}\in\Delta_{k,t}(\boldsymbol{n})}\binom{n}{n_1,\cdots,n_k}\Phi_{k,0}^{(t)}(\ell_1,\cdots,\ell_k)$$

$$\times\prod_{j=1}^{k}\frac{1}{\prod_{j=1}^{k}\ell_j!}\sum_{\boldsymbol{q}_j\in\Delta_{\ell_j,n_j}}\binom{n_j}{q_{j,1},\cdots,q_{j,\ell_j}}\Phi_t^{(n)}(q_{1,1},\cdots,q_{1,\ell_1},\cdots,q_{k,1},\cdots,q_{k,\ell_k})$$

$$= \frac{1}{k!}\sum_{t=k}^{n}\sum_{\boldsymbol{n}\in\Delta_{k,n}}\sum_{\boldsymbol{\ell}\in\Delta_{k,t}(\boldsymbol{n})}\binom{t}{\ell_1,\cdots,\ell_k}\Phi_{k,0}^{(t)}(\ell_1,\cdots,\ell_k)$$

$$\times\,\frac{1}{t!}\sum_{\boldsymbol{q}}\binom{n}{q_{1,1},\cdots,q_{1,\ell_1},\cdots,q_{k,1},\cdots,q_{k,\ell_k}}\Phi_t^{(n)}(q_{1,1},\cdots,q_{1,\ell_1},\cdots,q_{k,1},\cdots,q_{k,\ell_k}).$$

At this point, if $\Delta^*_{k,n}(\boldsymbol{\ell}) = \{\boldsymbol{n} = (n_1, \ldots, n_k) \in \times_{i=1}^k \{\ell_i, \ldots, n\} \text{ s.t. } |\boldsymbol{n}| = n\}$ one has

$$\begin{aligned}\sum_{\boldsymbol{n} \in \Delta^*_{k,n}(\boldsymbol{\ell})} \sum_{\boldsymbol{q}} \binom{n}{q_{1,1}, \cdots, q_{1,\ell_1}, \cdots, q_{k,1}, \cdots, q_{k,\ell_k}} \Phi_t^{(n)}(q_{1,1}, \cdots, q_{1,\ell_1}, \cdots, q_{k,1}, \cdots, q_{k,\ell_k}) \\ = \sum_{\boldsymbol{\nu} \in \Delta_{t,n}} \binom{n}{\nu_1, \cdots, \nu_t} \Phi_t^{(n)}(\nu_1, \cdots, \nu_t),\end{aligned} \tag{33}$$

now the result easily follows by interchanging the sum over $\boldsymbol{n}$ with the sum over $\boldsymbol{\ell}$ and taking into account (33).

□

A.4 Proof of Corollary 3

The proof for the hierarchical Pitman–Yor process (15) follows along the same lines of Corollary 2 and relies on the following identity for generalized factorial coefficients

$$\sum_{\boldsymbol{n} \in \Delta^*_{k,n}(\boldsymbol{\ell})} \binom{n}{n_1, \cdots, n_k} \prod_{j=1}^k \mathscr{C}(n_j, \ell_j; \sigma) = \binom{|\boldsymbol{\ell}|}{\ell_1, \cdots, \ell_k} \mathscr{C}(n, |\boldsymbol{\ell}|; \sigma).$$

Therefore, we have

$$\begin{aligned}\mathbb{P}[K_n = k] &= \frac{1}{k!} \sum_{(n_1, \ldots, n_k) \in \Delta_{n,\mathbf{1}}} \binom{n}{n_1, \cdots, n_k} \Pi_k^{(n)}(n_1, \ldots, n_k) \\ &= \frac{1}{k!} \sum_{\boldsymbol{\ell}} \binom{|\boldsymbol{\ell}|}{\ell_1, \cdots, \ell_k} \Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\ell_1, \ldots, \ell_k) \frac{\prod_{s=1}^{|\boldsymbol{\ell}|-1}(\theta + s\sigma)}{(\theta+1)_{n-1}} \frac{\mathscr{C}(n, |\boldsymbol{\ell}|; \sigma)}{\sigma^{|\boldsymbol{\ell}|}}\end{aligned}$$

where the second equality follows by the above mentioned identity and Φ stands for the EPPF of a PY process (6). Then the result follows. □

A.5 Proof of Theorem 3

The posterior characterization may be established through the posterior Laplace functional of $\tilde{\mu}_0$

$$\mathbb{E}\left[\mathrm{e}^{-\tilde{\mu}_0(f)} \,\middle|\, \boldsymbol{X}^{(n)}\right] = \lim_{\varepsilon \downarrow 0} \frac{\mathbb{E}\,\mathrm{e}^{-\tilde{\mu}_0(f)} \prod_{j=1}^k \tilde{p}^{n_j}(A_{j,\varepsilon})}{\mathbb{E}\prod_{j=1}^k \tilde{p}^{n_j}(A_{j,\varepsilon})} \tag{34}$$

for any measurable $f : \mathbb{X} \to \mathbb{R}^+$, where $A_{j,\varepsilon}$ is the same set used in the Proofs of Theorems 1– 2. The denominator is $M_{n_1,\ldots,n_k}(A_{1,\varepsilon} \times \cdots \times A_{k,\varepsilon})$, defined in the proof of Theorem 1 and, as $\varepsilon \downarrow 0$, equals

$$\begin{aligned}&\prod_{j=1}^k P_0(A_{j\varepsilon}) \sum_{\boldsymbol{\ell}} \sum_{\boldsymbol{q}} \Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\ell_1, \cdots, \ell_k) \\ &\qquad \times \prod_{j=1}^k \frac{1}{\ell_j!} \binom{n_j}{q_{j,1}, \cdots, q_{j,\ell_j}} \Phi_{|\boldsymbol{\ell}|}^{(n)}(q_{1,1}, \ldots, q_{1,\ell_1}, \ldots, q_{k,1}, \ldots, q_{k,\ell_k}) + \lambda_{k,\varepsilon}\end{aligned}$$

where $\lambda_{k,\varepsilon} = \mathrm{o}\Big(\prod_{j=1}^k P_0(A_{j,\varepsilon})\Big)$. The numerator in (34) may be evaluated in a similar way, in fact

$$\begin{aligned}\mathbb{E}\,\mathrm{e}^{-\tilde{\mu}_0(f)} \prod_{j=1}^k \tilde{p}^{n_j}(A_{j,\varepsilon}) &= \mathbb{E}\,\mathrm{e}^{-\tilde{\mu}_0(f)}\,\mathbb{E}\Big[\tilde{p}^{n_j}(A_{j,\varepsilon})\,\Big|\tilde{\mu}_0\Big] \\ &= \sum_{\boldsymbol{\ell}} \frac{c^{|\boldsymbol{\ell}|}}{\Gamma(n)} \int_0^\infty u^{n-1}\,\mathrm{e}^{-c\,\psi(u)} \prod_{j=1}^k \xi_{n_j,\ell_j}(u)\,\mathrm{d}u\Big(\mathbb{E}\mathrm{e}^{-\tilde{\mu}_0(f)} \prod_{j=1}^k \tilde{p}_0^{\ell_j}(A_{j,\varepsilon})\Big)\end{aligned}$$

Setting $\mathbb{X}_\varepsilon^* = \mathbb{X} \setminus (\cup_{j=1}^k A_{j,\varepsilon})$, as $\varepsilon \downarrow 0$ one can see that

$$\begin{aligned}\mathbb{E}\mathrm{e}^{-\tilde{\mu}_0(f)} \prod_{j=1}^k \tilde{p}_0^{\ell_j}(A_{j,\varepsilon}) &= \frac{1}{\Gamma(|\boldsymbol{\ell}|)} \int_0^\infty u^{|\boldsymbol{\ell}|-1}\mathbb{E}\mathrm{e}^{-\tilde{\mu}_0((f+u)\mathbb{1}_{\mathbb{X}_\varepsilon^*})}\Big(\prod_{j=1}^k \mathbb{E}\mathrm{e}^{-\tilde{\mu}_0((f+u)\mathbb{1}_{A_{j,\varepsilon}})}\,\tilde{\mu}_0^{\ell_j}(A_{j,\varepsilon})\Big)\,\mathrm{d}u \\ &= \frac{\prod_{j=1}^k c_0\,P_0(A_{j,\varepsilon})}{\Gamma(|\boldsymbol{\ell}|)} \int_0^\infty u^{|\boldsymbol{\ell}|-1}\,\mathrm{e}^{-c_0\,\psi_0(f+u)} \prod_{j=1}^k \tau_{\ell_j}(u+f(X_j^*))\,\mathrm{d}u + \lambda_{k,\varepsilon}\end{aligned}$$

and as a consequence we obtain

$$\begin{aligned}\mathbb{E}\,\mathrm{e}^{-\tilde{\mu}_0(f)} \prod_{j=1}^k \tilde{p}^{n_j}(A_{j,\varepsilon}) &= \prod_{j=1}^k P_0(A_{j,\varepsilon}) \\ &\quad\times \sum_{\boldsymbol{\ell},\boldsymbol{q}} \prod_{j=1}^k \frac{1}{\ell_j!}\binom{n_j}{q_{j,1},\cdots,q_{j,\ell_j}}\,\Phi_{|\boldsymbol{\ell}|}(q_{1,1},\ldots,q_{1,\ell_1},\ldots,q_{k,1},\ldots,q_{k,\ell_k}) \\ &\quad\times \frac{c_0^{|\boldsymbol{\ell}|}}{\Gamma(|\boldsymbol{\ell}|)} \int_0^\infty u^{|\boldsymbol{\ell}|-1}\,\mathrm{e}^{-c_0\,\psi_0(f+u)} \prod_{j=1}^k \tau_{\ell_j}(u+f(X_j^*))\,\mathrm{d}u + \lambda_{k,\varepsilon}.\end{aligned}$$

Conditioning on the tables $\boldsymbol{T}^{(n)}$, the posterior Laplace functional coincides with

$$\mathbb{E}\Big[\mathrm{e}^{-\tilde{\mu}_0(f)}\,\Big|\,\boldsymbol{X}^{(n)},\boldsymbol{T}^{(n)}\Big] = \frac{\int_0^\infty u^{|\boldsymbol{\ell}|-1}\,\mathrm{e}^{-c_0\,\psi_0(f+u)}\,\prod_{j=1}^k \tau_{\ell_j}(u+f(X_j^*))\,\mathrm{d}u}{\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\ell_1,\ldots,\ell_k)}$$

and the result follows, observing that the normalizing constant of the density $f_0(\,\cdot\,|\boldsymbol{X}^{(n)},\boldsymbol{T}^{(n)})$ in (19) amounts to be $\Phi_{k,0}^{(|\boldsymbol{\ell}|)}(\ell_1,\ldots,\ell_k)$. □

A.6 Proof of Theorem 4

The proof follows the similar arguments as that of Theorem 3. The posterior Laplace functional of $\tilde{\mu}$, conditional on the observations $\boldsymbol{X}^{(n)}$ and the tables $\boldsymbol{T}^{(n)}$, may be expressed as

$$\mathbb{E}\Big[\mathrm{e}^{-\tilde{\mu}(f)}\,\Big|\,\boldsymbol{X}^{(n)},\boldsymbol{T}^{(n)}\Big] = \mathbb{E}\Big[\,\mathbb{E}\Big[\mathrm{e}^{-\tilde{\mu}(f)}\,\Big|\,\boldsymbol{X}^{(n)},\boldsymbol{T}^{(n)},\tilde{\mu}_0\Big]\,\Big|\,\boldsymbol{X}^{(n)},\boldsymbol{T}^{(n)}\Big]$$

for any measurable function $f : \mathbb{X} \to \mathbb{R}^+$. Then, we try to calculate

$$\mathbb{E}\Big[\mathrm{e}^{-\tilde{\mu}(f)}\,\Big|\,\boldsymbol{X}^{(n)},\boldsymbol{T}^{(n)},\tilde{\mu}_0\Big] = \lim_{\varepsilon\downarrow 0} \frac{\mathbb{E}\Big[\mathrm{e}^{-\tilde{\mu}(f)}\,\prod_{j=1}^k \tilde{p}^{n_j}(A_{j,\varepsilon})\,\Big|\,\boldsymbol{T}^{(n)},\tilde{\mu}_0\Big]}{\mathbb{E}\Big[\prod_{j=1}^k \tilde{p}^{n_j}(A_{j,\varepsilon})\,\Big|\,\boldsymbol{T}^{(n)},\tilde{\mu}_0\Big]} \tag{35}$$

The denominator equals

$$\Big(\prod_{j=1}^{k}\tilde{p}_0^{\,\ell_j}(A_{j,\varepsilon})\Big)\,\frac{c^{|\boldsymbol{\ell}|}}{\Gamma(n)}\int_0^\infty u^{n-1}\,\mathrm{e}^{-c\,\psi(u)}\prod_{j=1}^{k}\prod_{t=1}^{\ell_j}\tau_{q_{j,t}}(u)\,\mathrm{d}u,$$

whereas, for the numerator we note that

$$\mathbb{E}\left[\mathrm{e}^{-\tilde{\mu}(f)}\prod_{j=1}^{k}\tilde{p}^{n_j}(A_{j,\varepsilon})\,\Big|\,\boldsymbol{T}^{(n)},\tilde{\mu}_0\right]=\mathbb{E}\Big[\mathrm{e}^{-\tilde{\mu}(f)}\prod_{j=1}^{k}\tilde{p}^{n_j}(A_{j,\varepsilon})\,\Big|\,\boldsymbol{T}^{(n)},\tilde{\mu}_0\Big]$$

$$=\Big(\prod_{j=1}^{k}\tilde{p}_0^{\,\ell_j}(A_{j,\varepsilon})\Big)\prod_{i=1}^{d}\frac{c^{|\boldsymbol{\ell}|}}{\Gamma(n)}\int_0^\infty u^{n-1}\,\mathrm{e}^{-c\,\tilde{\psi}(f+u)}\prod_{j=1}^{k}\prod_{t=1}^{\ell_j}\tau_{q_{j,t}}(u+f_i(X_j^*))\,\mathrm{d}u,$$

where we have put

$$\tilde{\psi}(f)=\int_{\mathbb{X}}\int_0^\infty[1-\mathrm{e}^{-sf(x)}],\rho(s)\,\mathrm{d}s\,\tilde{p}_0(\mathrm{d}x).$$

The right–hand–side of (35) boils down to

$$\frac{\int_0^\infty u^{n-1}\mathrm{e}^{-c\tilde{\psi}(f+u)}\prod_{j=1}^{k}\prod_{t=1}^{\ell_j}\tau_{q_{j,t}}(u+f(X_j^*))\mathrm{d}u}{\int_0^\infty u^{n-1}\mathrm{e}^{-c\psi(u)}\prod_{j=1}^{k}\prod_{t=1}^{\ell_j}\tau_{q_{j,t}}(u)\mathrm{d}u}.$$

which entails

$$\mathbb{E}\Big[\mathrm{e}^{-\tilde{\mu}(f)}\,\Big|\,\boldsymbol{X}^{(n)},\boldsymbol{T}^{(n)},U_n,\tilde{\mu}_0\Big]=\prod_{j=1}^{k}\prod_{t=1}^{\ell_j}\frac{\tau_{q_{j,t}}(U_n+f(X_j^*))}{\tau_{q_{j,t}}(U_n)}$$

$$\times\ \exp\left\{-c\int_{\mathbb{X}\times\mathbb{R}^+}(1-e^{-sf(x)})\mathrm{e}^{-sU_n}\,\rho(s)\,\mathrm{d}s\,\tilde{p}_0(\mathrm{d}x)\right\}$$

and the assertion follows. □

A.7 Proof of Theorem 7

Let $\tilde{\mu}_\sigma$ denote a σ–stable CRM and $\tilde{\mu}_{\sigma,\theta}$ is a random measure whose normalization yields a Pitman–Yor process with parameters (σ,θ). From Theorem 5, we have

$$\mathbb{E}\Big[\mathrm{e}^{-\tilde{\eta}_0^*(f)}\Big]=\frac{\sigma_0}{\Gamma\Big(\frac{\theta_0}{\sigma_0}+k\Big)}\int_0^\infty v^{\theta_0+k\sigma_0-1}\,\mathrm{e}^{-v^{\sigma_0}}\,\mathrm{e}^{-\int_{\mathbb{X}}[(v+f(x))^{\sigma_0}-v^{\sigma_0}]\,P_0(\mathrm{d}x)}\,\mathrm{d}v$$

$$=\frac{\sigma_0}{\Gamma\Big(\frac{\theta_0}{\sigma_0}+k\Big)}\int_0^\infty v^{\theta_0+k\sigma_0-1}\,\Big(\mathbb{E}\,\mathrm{e}^{-v\tilde{\mu}_{\sigma_0}(\mathbb{X})-\tilde{\mu}_{\sigma_0}(f)}\Big)\,\mathrm{d}v$$

$$=\frac{\sigma_0\,\Gamma(\theta_0+k\sigma_0)}{\Gamma\Big(\frac{\theta_0}{\sigma_0}+k\Big)}\,\Big(\mathbb{E}\,\mathrm{e}^{-\tilde{\mu}_{\sigma_0}(f)}\{\tilde{\mu}_{\sigma_0}(\mathbb{X})\}^{-\theta_0-k\sigma_0}\Big)=\mathbb{E}\,\mathrm{e}^{-\tilde{\mu}_{\sigma_0,\theta_0+k\sigma_0}(f)}$$

Hence, one can conclude that $\eta_0^*\overset{\mathrm{d}}{=}\tilde{\mu}_{\sigma_0,\theta_0+k\sigma_0}$ and

$$\tilde{p}_0\,|\,(\boldsymbol{X}^{(n)},\boldsymbol{T}^{(n)},V_0)\ \overset{\mathrm{d}}{=}\ \sum_{j=1}^{k}W_{j,V_0}\,\delta_{X_j^*}+W_{k+1,V_0}\,\tilde{p}_{\sigma_0,\theta_0+k\sigma_0}$$

where $\tilde{p}_{\sigma_0,\theta_0+k\sigma_0} \sim \mathrm{PY}(\sigma_0, \theta_0 + k\sigma_0; P_0)$. Moreover $W_{j,V_0} = I_j/(\eta_0^*(\mathbb{X}) + \sum_{i=1}^k I_i)$, for any $j = 1, \ldots, k$, and $W_{k+1,V_0} = \eta_0^*(\mathbb{X})/(\eta_0^*(\mathbb{X}) + \sum_{i=1}^k I_i)$. Let us denote by f_v the density of the vector $(W_{1,V_0}, \ldots, W_{k,V_0})$ on the k–dimensional simplex Δ_k, then we want to determine

$$f(w_1, \ldots, w_k) = \int_0^\infty f_v(w_1, \ldots, w_k) \, \frac{\sigma_0}{\Gamma\left(\frac{\theta_0}{\sigma_0} + k\right)} \, v^{\theta_0 + k\sigma_0 - 1} \, \mathrm{e}^{-v^{\sigma_0}} \, \mathrm{d}v.$$

Denoted by h_v the density function of $\eta_0^*(\mathbb{X})$ and by independence, the vector $(I_1, \ldots, I_k, \eta_0^*(\mathbb{X}))$ has density given by

$$f_v^*(x_1, \ldots, x_k, t) = h_v(t) \, \frac{v^{|\boldsymbol{\ell}| - k\sigma_0}}{\prod_{j=1}^k \Gamma(\ell_j - \sigma_0)} \, \mathrm{e}^{-v \sum_{i=1}^k x_i} \, \prod_{j=1}^k x_j^{\ell_j - \sigma_0 - 1}.$$

The density function of $(W_{1,V_0}, \ldots, W_{k,V_0}, W_{V_0})$ follows by the simple transformation $W_{V_0} = \sum_{i=1}^k I_j + \eta_0^*(\mathbb{X})$:

$$f_v(w_1, \ldots, w_k, w) = \frac{v^{|\boldsymbol{\ell}| - k\sigma_0} \, \prod_{j=1}^k w_j^{\ell_j - \sigma_0 - 1}}{\prod_{j=1}^k \Gamma(\ell_j - \sigma_0)} w^{|\boldsymbol{\ell}| - k\sigma_0} \, \mathrm{e}^{-vw|\boldsymbol{w}|} \, h_v(w(1 - |\boldsymbol{w}|))$$

where $|\boldsymbol{w}| = \sum_{i=1}^k w_i$. From this, an expression for the density of $(W_{1,V_0}, \ldots, W_{k,V_0})$ easily follows and it turns out to be

$$f_v(w_1, \ldots, w_k) = \frac{v^{|\boldsymbol{\ell}| - k\sigma_0} \, \prod_{j=1}^k w_j^{\ell_j - \sigma_0 - 1}}{\prod_{j=1}^k \Gamma(\ell_j - \sigma_0)} \, \frac{1}{(1 - |\boldsymbol{w}|)^{|\boldsymbol{\ell}| - k\sigma_0 + 1}} \left(\mathbb{E}(\eta_0^*(\mathbb{X}))^{|\boldsymbol{\ell}| - k\sigma_0} \, \mathrm{e}^{-\frac{v|\boldsymbol{w}|}{1 - |\boldsymbol{w}|} \eta_0^*(\mathbb{X})} \right).$$

Since η_0^* is a generalized gamma CRM with parameters (σ_0, V_0) and base measure P_0, its probability distributions $\mathbb{P}^*$ is absolutely continuous with respect to the probability distribution $\mathbb{P}_{\sigma_0}$ of a σ_0–stable CRM and

$$\frac{\mathrm{d}\mathbb{P}^*}{\mathrm{d}\mathbb{P}_{\sigma_0}}(m) = \exp\{-vm(\mathbb{X}) + v^{\sigma_0}\},$$

then

$$\mathbb{E}(\eta_0^*(\mathbb{X}))^{|\boldsymbol{\ell}| - k\sigma_0} \, \mathrm{e}^{-\frac{v|\boldsymbol{w}|}{1 - |\boldsymbol{w}|} \eta_0^*(\mathbb{X})} = \mathrm{e}^{v^{\sigma_0}} \, \mathbb{E}\left(\tilde{\mu}_{\sigma_0}(\mathbb{X})\right)^{|\boldsymbol{\ell}| - k\sigma_0} \, \mathrm{e}^{-\frac{v}{1 - |\boldsymbol{w}|} \tilde{\mu}_{\sigma_0}(\mathbb{X})}$$

In view of this, one can now marginalize f_v with respect to v and and obtain a density of $(W_1, \ldots, W_k)$. Indeed, a straightforward calculation leads to

$$f(w_1, \ldots, w_k) = \frac{\Gamma(\theta_0 + |\boldsymbol{\ell}|)}{\Gamma(\theta_0 + k\sigma_0) \prod_{j=1}^k \Gamma(\ell_j - \sigma_0)} \, (1 - |\boldsymbol{w}|)^{\theta_0 + k\sigma_0 - 1} \, \prod_{j=1}^k w_j^{\ell_j - \sigma_0 - 1}$$

and this completes the proof of the posterior characterization of $\tilde{p}_0$. The representation (24) may be proved in a similar fashion. □

Acknowledgements

A. Lijoi and I. Prünster are supported by the European Research Council (ERC), StG "N-BNP" 306406 and by MIUR, PRIN Project 2015SNS29B.

References

Airoldi, E.M., Costa, T., Bassetti, F., Leisen, F. & Guindani, M. (2014). Generalized species sampling priors with latent beta reinforcements. *J. Amer. Statist. Assoc.* **109**, 1466–1480.

Bassetti, F., Crimaldi, I. & Leisen, F. (2010). Conditionally identically distributed species sampling sequences. *Advances in Applied Probability* **42**, 433–459.

Bertoin, J. (2006). *Random fragmentation and coagulation processes.* Cambridge University Press, Cambridge.

Camerlenghi, F., Lijoi, A., Orbanz, P. & Prünster, I. (2017). Distribution theory for hierarchical processes. *Ann. Statist.*, https://doi.org/10.1214/17-AOS1678

Charalambides, C.A. (2002). *Enumerative combinatorics*, Chapman & Hall, Boca Raton, FL.

Dahl, D.B., Day, R. & Tsai, J. (2017). Random partition distribution indexed by pairwise information. *J. Amer. Statist. Assoc.* **112**, 721-732.

Daley, D.J. & Vere–Jones, D. (2008). *An introduction to the theory of point processes. Volume II*, Springer, New York.

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

Favaro, S., Prünster, I. & Walker, S.G. (2011). On a class of random probability measures with general predictive structure. *Scand. J. Statist.* **38**, 359–376.

Favaro, S., Lijoi, A. & Prünster, I. (2012). A new estimator of the discovery probability. *Biometrics*, **68**, 1188–1196.

Favaro, S., Lijoi, A. & Prünster, I. (2013). Conditional formulae for Gibbs–type exchangeable random partitions. *Ann. Appl. Probab.*, **23**, 1721–1754.

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

Ferguson, T.S. & Klass, M.J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Statist.* **43**, 1634–1643.

Fuentes-García, R., Mena, R.H. & Walker, S.G. (2010). A probability for classification based on the Dirichlet process mixture model. *J. Classification* **27**, 389–403.

Gnedin, A.V. & Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **325**, 83–102.

Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Good, I.J. & Toulmin, G.H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, **43**, 45–63.

Hardy M. (2006). Combinatorics of partial derivatives. *Electron. J. Combin.* **13,** 1–13

Hjort, N.L., Holmes, C.C., Müller, P. & Walker, S.G., eds. (2010). *Bayesian Nonparametrics.* Cambridge University Press, Cambridge.

James, L.F., Lijoi, A. & Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.* **36**, 76–97.

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

Kingman, J.F.C. (1982). The coalescent. *Stochastic Process. Appl.* **13**, 235–248.

Kingman, J.F.C. (1993). *Poisson processes.* Oxford University Press, Oxford.

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

Lijoi, A., Mena, R.H. & Prünster, I. (2007). Bayesian nonparametric estimation of the probability of discovering a new species *Biometrika*. **94**, 769–786.

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

Mao, C.X. (2004). Predicting the conditional probability of discovering a new class. *J. Amer. Statist. Assoc.*, **99**, 1108–1118.

Müller, P., Quintana, F. & Rosner, G.L. (2011). A product partition model with regression on covariates. *J. Comput. Graph. Statist.* **20**, 260–278.

Müller, P., Quintana, F.A., Jara, A. & Hanson, T. (2015). *Bayesian Nonparametric Data Analysis.* Springer, New York.

Navarrete, C.A. & Quintana, F.A. (2011) Similarity analysis in Bayesian random partition models. *Comput. Statist. Data Anal.* **55**, 97–109.

Nguyen, X. (2016). Borrowing strength in hierarchical Bayes: posterior concentration of the Dirichlet base measure. *Bernoulli*, **22**, 1535–1571.

Phadia, E.G. (2013). *Prior processes and their applications.* Springer, New York.

Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Probab.* **27**, 1870–1902.

Pitman, J. (2006). *Combinatorial stochastic processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875, Springer, New York.

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

Quintana, F.A., Müller, P. & Papoila, A.L. (2015). Cluster-specific variable selection for product partition models. *Scand. J. Statist.* **42**, 1065–1077.

Regazzini, E., Lijoi, A., & Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31**, 560–585.

Teh, Y.W., Jordan, M.I., Beal, M.J. & Blei, D.M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566–1581.

Teh, Y.W. & Jordan, M.I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics* (Hjort, N.L., Holmes, C.C. Müller, P., Walker, S.G. Eds.), pp. 158–207. Cambridge University Press, Cambridge.

Wood, F., Gasthaus, J., Archambeau, C., James, L.F. & Teh, Y.W. (2011). The sequence memoizer. *Communications ACM* **54**, 91–98.