Bayesian Analysis (2019) **TBA**, Number TBA, pp. 1–30

Hierarchical Species Sampling Models

Federico Bassetti*, Roberto Casarin†, and Luca Rossini‡§

Abstract. This paper introduces a general class of hierarchical nonparametric prior distributions which includes new hierarchical mixture priors such as the hierarchical Gnedin measures, and other well-known prior distributions such as the hierarchical Pitman-Yor and the hierarchical normalized random measures. The random probability measures are constructed by a hierarchy of generalized species sampling processes with possibly non-diffuse base measures. The proposed framework provides a probabilistic foundation for hierarchical random measures, and allows for studying their properties under the alternative assumptions of diffuse, atomic and mixed base measure. We show that hierarchical species sampling models have a Chinese Restaurants Franchise representation and can be used as prior distributions to undertake Bayesian nonparametric inference. We provide a general sampling method for posterior approximation which easily accounts for non-diffuse base measures such as spike-and-slab.

Keywords: Bayesian nonparametrics, generalized species sampling, Gibbs sampling, hierarchical random measures, spike-and-slab.

MSC 2010 subject classifications: 62G05, 62F15, 60G57, 60G09.

1 Introduction

Cluster structures in multiple groups of observations can be modelled by means of *hierarchical random probability measures* or *hierarchical processes* that allow for heterogenous clustering effects across groups and for sharing clusters among groups. As an effect of the heterogeneity, in these models the number of clusters in each group (marginal number of clusters) can differ, and due to cluster sharing, the number of clusters in the entire sample (total number of clusters) can be smaller than the sum of the marginal number of clusters. An important example of hierarchical random measure is the Hierarchical Dirichlet Process (HDP), introduced in the seminal paper of Teh et al. (2006). The HDP involves a simple Bayesian hierarchy where the common base measure for a set of Dirichlet processes is itself distributed according to a Dirichlet process. This means that the joint law of the random probability measures $(p_1, \ldots, p_I)$ is

$$\begin{aligned} p_i|p_0 &\overset{iid}{\sim} DP(\theta_1, p_0), \quad i = 1, \ldots, I, \\ p_0|H_0 &\sim DP(\theta_0, H_0), \end{aligned} \tag{1.1}$$

*Department of Mathematics, Polytechnic University of Milan, Via E. Bonardi 9, 20133, Milan, Italy, federico.bassetti@polimi.it

†Department of Economics, University Ca' Foscari of Venice, Cannaregio 873, 30121, Venezia, Italy, r.casarin@unive.it

‡Department of Econometrics and Operations Research, VU University Amsterdam, Amsterdam, The Netherlands, l.rossini@vu.nl

§Luca Rossini acknowledges financial support from the European Union Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 796902.

 https://doi.org/10.1214/19-BA1168

where $DP(\theta, p)$ denotes the Dirichlet process with base measure p and concentration parameter $\theta > 0$. Once the joint law of $(p_1, \ldots, p_I)$ has been specified, observations $[\xi_{i,j}]_{i=1,\ldots,I;j\geq 1}$ are assumed to be conditionally independent given $(p_1, \ldots, p_I)$ with

$$\xi_{i,j}|(p_1, \ldots, p_I) \overset{ind}{\sim} p_i, \ i = 1, \ldots, I \text{ and } j \geq 1.$$

Hierarchical processes are widely used as prior distributions in Bayesian nonparametric inference (see Teh and Jordan (2010) and reference therein), by assuming $\xi_{i,j}$ are latent variables describing the clustering structure of the data and the observations in the i-th group, $Y_{i,j}$, are conditionally independent given $\xi_{i,j}$ with

$$Y_{i,j}|\xi_{i,j} \overset{ind}{\sim} f(\cdot|\xi_{i,j}),$$

where f is a suitable kernel density.

In this paper, we introduce a new class of hierarchical random probability measures, called Hierarchical Species Sampling Model (HSSM), based on a hierarchy of species sampling models.

A *Species Sampling random probability* (SSrp) is defined as

$$p = \sum_{j\geq 1} \delta_{Z_j} q_j, \tag{1.2}$$

where $(Z_j)_{j\geq 1}$ and $(q_j)_{j\geq 1}$ are stochastically independent sequences, the atoms Z_j are i.i.d. with common distribution H_0 (base measure) and the non-negative weights $q_j \geq 0$ sum to one almost surely. By Kingman's theory on exchangeable partitions, any random sequence of positive weights such that $\sum_{j\geq 1} q_j \leq 1$ can be associated to an exchangeable random partition of the integers $(\Pi_n)_{n\geq 1}$. Moreover, the law of an exchangeable random partition $(\Pi_n)_{n\geq 1}$ is completely described by an *exchangeable partition probability function* (EPPF) $\mathfrak{q}_0$. Hence the law of the measure p defined in (1.2) is parametrized by $\mathfrak{q}_0$ and H_0, and it will be denoted by $SSrp(\mathfrak{q}_0, H_0)$.

A HSSM is a vector of random measures $(p_1, \ldots, p_I)$ with

$$\begin{aligned} p_i|p_0 &\overset{iid}{\sim} SSrp(\mathfrak{q}, p_0), \qquad i = 1, \ldots, I, \\ p_0 &\sim SSrp(\mathfrak{q}_0, H_0), \end{aligned} \tag{1.3}$$

where H_0 is a base measure and $\mathfrak{q}_0$ and $\mathfrak{q}$ are two EPPFs.

The proposed framework provides a general probabilistic foundation of both existing and novel hierarchical random measures, and relies on a convenient parametrization of the hierarchical process in terms of two EPPFs and a base measure. Our HSSM class includes the HDP, its generalizations given by the Hierarchical Pitman–Yor process (HPYP), see Teh (2006); Du et al. (2010); Lim et al. (2016); Camerlenghi et al. (2017) and the hierarchical normalized random measures with independent increments (HNRMI), first studied in Camerlenghi et al. (2018), Camerlenghi et al. (2019) and more recently in Argiento et al. (2019). Among the novel measures, we study hierarchical generalizations of Gnedin (Gnedin (2010)) and of finite mixture (e.g., Miller and Harrison (2018)) processes and asymmetric hierarchical constructions with p_0 and p_i of

different type (Du et al. (2010)). Another motivation for studying HSSMs relies on the introduction of non-diffuse base measures (e.g., the spike-and-slab prior of George and McCulloch (1993)) now widely used in Bayesian parametric (e.g., Castillo et al. (2015) and Rockova and George (2018)) and nonparametric (e.g., Kim et al. (2009), Canale et al. (2017)) inference.

We show that the arrays of observations from HSSMs have a Chinese Restaurant Franchise representation, that is appealing for the applications to Bayesian nonparametrics, since it sheds light on the clustering mechanism of the observations and suggests a simple and general sampling algorithm for posterior computations. The sampler can be used under both assumptions of diffuse and non-diffuse (e.g. spike-and-slab) base measure, whenever the EPPFs $\mathfrak{q}_0$ and $\mathfrak{q}$ are known explicitly.

By exploiting the properties of species sampling sequences, we are able to provide the finite sample distribution of the number of clusters for each group of observations and the total number of clusters for the hierarchy. We provide some new asymptotic results when the number of observations goes to infinity, thus extending to our general class of processes the asymptotic approximations given in Pitman (2006) and Camerlenghi et al. (2019) for species sampling and hierarchical normalized random measures, respectively.

The paper is organized as follows. Section 2 introduces exchangeable random partitions, generalized species sampling sequences and species sampling random probability measures. Section 3 defines hierarchical species sampling models and shows some useful properties for the applications to Bayesian nonparametric inference. Section 4 gives finite-sample and asymptotic distributions of the number of clusters under both assumptions of diffuse and non-diffuse base measure. A general Gibbs sampler for hierarchical species sampling mixtures is established in Section 5. Section 6 presents some simulation studies and a real data application.

2 Background Material

Our Hierarchical Species Sampling Models build on exchangeable random partitions and related processes, such as species sampling sequences and species sampling random probability measures. We review some of their definitions and properties, which will be used in the rest of the paper. Supplementary material (Bassetti et al., 2019a) provides further details, examples and some new results under the assumption of non-diffuse base measure.

2.1 Exchangeable Random Partitions

Exchangeable random partitions are used in a wide range of theoretical and applied problems in various fields, such as population genetics (Ewens, 1972; Kingman, 1980; Donnelly, 1986; Hoppe, 1984), combinatorics, algebra and number theory (Donnelly and Grimmett, 1993; Diaconis and Ram, 2012; Arratia et al., 2003), machine learning (Teh, 2006; Wood et al., 2009), psychology (Navarro et al., 2006), model-based clustering (Lau and Green, 2007; Müller and Quintana, 2010), and Bayesian nonparametrics (e.g., see Hjort et al. (2010) and references therein). For a comprehensive review see Pitman (2006).

A (set) partition π_n of $[n] := \{1, \ldots, n\}$ is an unordered collection $(\pi_{1,n}, \ldots, \pi_{k,n})$ of disjoint non-empty subsets (blocks) of $\{1, \ldots, n\}$ such that $\cup_{j=1}^{k} \pi_{j,n} = [n]$. A partition $\pi_n = [\pi_{1,n}, \pi_{2,n}, \ldots, \pi_{k,n}]$ has $|\pi_n| := k$ blocks (with $1 \leq |\pi_n| \leq n$) and we denote by $|\pi_{c,n}|$, the number of elements of the block $c = 1, \ldots, k$. We denote with $\mathcal{P}_n$ the collection of all partitions of $[n]$ and, given a partition, we list its blocks in ascending order of their smallest element. In other words, a partition $\pi_n \in \mathcal{P}_n$ is coded with elements *in order of appearance*.

A *random partition of* $\mathbb{N}$ is a sequence of random partitions, $\Pi = (\Pi_n)_n$, such that each element Π_n takes values in $\mathcal{P}_n$ and the restriction of Π_n to $\mathcal{P}_m$, $m < n$ is Π_m (*consistency property*). A random partition of $\mathbb{N}$ is said to be *exchangeable* if for every n the distribution of Π_n is invariant under the action of all permutations (acting on Π_n in the natural way).

Exchangeable random partitions are characterized by the fact that their distribution depends on Π_n only through its block size. A random partition on $\mathbb{N}$ is exchangeable if and only if its distribution can be written in terms of *exchangeable partition probability function* (EPPF). An EPPF is a symmetric function $\mathfrak{q}$ defined on the integers $(n_1, \ldots, n_k)$, with $\sum_{i=1}^{k} n_i = n$, that satisfies the additions rule $\mathfrak{q}(n_1, \ldots, n_k) = \sum_{j=1}^{k} \mathfrak{q}(n_1, \ldots, n_j + 1, \ldots, n_k) + \mathfrak{q}(n_1, \ldots, n_k, 1)$, (see Pitman (2006)). If $(\Pi_n)_n$ is an exchangeable random partition of $\mathbb{N}$, there exists an EPPF such that for every n and $\pi_n \in \mathcal{P}_n$

$$\mathbb{P}\{\Pi_n = \pi_n\} = \mathfrak{q}\left(|\pi_{1,n}|, \ldots, |\pi_{k,n}|\right), \tag{2.1}$$

where $k = |\pi_n|$. In other words, $\mathfrak{q}(n_1, \ldots, n_k)$ corresponds to the probability that Π_n is equal to any of the partitions of $[n]$ with k distinct blocks and block frequencies $(n_1, \ldots, n_k)$.

Given an EPPF $\mathfrak{q}$, one deduces the corresponding sequence of predictive distributions. Starting with $\Pi_1 = \{1\}$, given $\Pi_n = \pi_n$ (with $|\pi_n| = k$), the conditional probability of adding a new block (containing $n+1$) to Π_n is

$$\nu_n(|\pi_{1,n}|, \ldots, |\pi_{k,n}|) := \frac{\mathfrak{q}(|\pi_{1,n}|, \ldots, |\pi_{k,n}|, 1)}{\mathfrak{q}(|\pi_{1,n}|, \ldots, |\pi_{k,n}|)}; \tag{2.2}$$

while the conditional probability of adding $n+1$ to the c-th block of Π_n (for $c = 1, \ldots, k$) is

$$\omega_{n,c}(|\pi_{1,n}|, \ldots, |\pi_{k,n}|) := \frac{\mathfrak{q}(|\pi_{1,n}|, \ldots, |\pi_{c,n}| + 1, \ldots, |\pi_{k,n}|)}{\mathfrak{q}(|\pi_{1,n}|, \ldots, |\pi_{k,n}|)}. \tag{2.3}$$

An important class of exchangeable random partitions is the Gibbs-type partitions, introduced in Gnedin and Pitman (2005) and characterized by the EPPF

$$\mathfrak{q}(n_1, \ldots, n_k) := V_{n,k} \prod_{c=1}^{k} (1 - \sigma)_{n_c - 1},$$

where $(x)_n = x(x+1)\ldots(x+n-1)$ is the rising factorial (or Pochhammer's polynomial), $\sigma < 1$ and $V_{n,k}$ are positive real numbers such that $V_{1,1} = 1$ and

$$(n - \sigma k) V_{n+1,k} + V_{n+1,k+1} = V_{n,k}, \quad n \geq 1,\ 1 \leq k \leq n. \tag{2.4}$$

2.2 Species Sampling Models with General Base Measure

Kingman's theory of random partitions sets up a one-one correspondence (Kingman's correspondence) between EPPFs and distributions for decreasing sequences of random variables $(q_k^{\downarrow})_k$ with $q_i^{\downarrow} \geq 0$ and $\sum_i q_i^{\downarrow} \leq 1$ almost surely, by using the notion of random partition induced by a sequence of random variables. Let us recall that a sequence of random variables $(\zeta_n)_n$ induces a random partition on $\mathbb{N}$ by equivalence classes $i \sim j$ if and only if $\zeta_i = \zeta_j$.

If $\sum_i q_i^{\downarrow} = 1$ a.s. then Kingman's correspondence between EPPF and $(q_j^{\downarrow})_j$ can be defined as follows. Let $(U_j)_j$ be an i.i.d. sequence of uniform random variables on $(0,1)$ independent from $(q_j^{\downarrow})_j$ and let Π be the random partition induced by a sequence $(\theta_n)_n$ of conditionally i.i.d. random variables from $\sum_{j\geq 1} q_j \delta_{U_j}$ where $(q_j)_j$ is any (possibly random) permutation of $(q_j^{\downarrow})_j$. Then the EPPF in the Kingman's correspondence is the EPPF of Π. In point of fact, one can prove that

$$\mathfrak{q}(n_1, \ldots, n_k) = \sum_{j_1,\ldots,j_k} \mathbb{E}\left[\prod_{i=1}^{k} q_{j_i}^{n_i}\right], \tag{2.5}$$

where $(j_1, \ldots, j_k)$ ranges over all ordered k-tuples of distinct positive integers. See Equation (2.14) in Pitman (2006).

A *Species Sampling random probability* of parameters $\mathfrak{q}$ and H, in symbols $p \sim SSrp(\mathfrak{q}, H)$, is a random distribution

$$p = \sum_{j\geq 1} \delta_{Z_j} q_j, \tag{2.6}$$

where $(Z_j)_j$ are i.i.d. random variables on a Polish space $\mathbb{X}$ with possibly non-diffuse common distribution H and EPPF $\mathfrak{q}$ given in (2.5). Such random probability measures are sometimes called *species sampling models*. In this parametrization, $\mathfrak{q}$ takes into account only the law of $(q_j^{\downarrow})_j$ while H describes the law of the Z_js.

If H is diffuse, a sequence $(\xi_n)_n$ sampled from p in (2.6), i.e. with ξ_n conditionally i.i.d. (given p) with law $p \sim SSrp(\mathfrak{q}, H)$, is a Species Sampling Sequence as defined by Pitman (1996) (Proposition 13 in Pitman (1996)) and the EPPF of the partition induced by $(\xi_n)_n$ is exactly $\mathfrak{q}$. On the contrary, when H is not diffuse then $(\xi_n)_n$ is not a Species Sampling Sequence in the sense of Pitman (1996) and the EPPF of the induced partition is not $\mathfrak{q}$. Nevertheless, as shown in the next Proposition, there exists an augmented space $\mathbb{X} \times (0,1)$ and a latent partition related to $(\xi_n)_n$ with EPPF $\mathfrak{q}$.

Hereafter, for a general base measure H, we refer to $(\xi_n)_n$ as *generalized species sampling sequence*, $gSSS(\mathfrak{q}, H)$.

Proposition 1. *Let $(U_j)_j$ be an i.i.d. sequence of uniform random variables on $(0,1)$, $(Z_j)_j$ an i.i.d. sequence with possibly non-diffuse common distribution H and $(q_j)_j$ a sequence of positive numbers with $\sum_j q_j = 1$ a.s.. Assume that all the previous elements*

are independent and let $(\zeta_n)_n := (\xi_n, \theta_n)_n$ *be a sequence of random variables, with values in* $\mathbb{X} \times (0,1)$*, conditionally i.i.d. from* p' *given*

$$p' = \sum_{j \geq 1} \delta_{(Z_j, U_j)} q_j. \tag{2.7}$$

Then, the EPPF of the partition induced by $(\zeta_n)_n$ *is* $\mathfrak{q}$ *given in* (2.5) *and* $(\xi_n)_n$ *is a* $gSSS(\mathfrak{q}, H)$*.*

From the previous Proposition, it follows that the partition induced by $(\zeta_n)_n$ is in general finer than the partition induced by $(\xi_n)_n$, with the equality if H is diffuse. This result is essential in order to properly define and study hierarchical models of type (1.3), since the random measure p_0 in (1.3) is almost surely discrete and hence not diffuse. Further properties of the gSSS are proved in the supplementary material (Bassetti et al., 2019a), whereas further results are available in Sangalli (2006) for normalized random measures with independent increments. These properties are relevant to the comprehension of the implications of mixed based measures for Bayesian non-parametrics, especially for hierarchical prior constructions.

3 Hierarchical Species Sampling Models

We introduce *hierarchical species sampling models* (HSSMs), provide some examples and derive relevant properties.

3.1 HSSM Definition and Examples

In the following definition a hierarchy of species sampling random probabilities is used to build hierarchical species sampling models.

Definition 1. *Let* $\mathfrak{q}$ *and* $\mathfrak{q}_0$ *be two EPPFs and* H_0 *a probability distribution on the Polish space* $\mathbb{X}$*. A Hierarchical Species Sampling model,* $HSSM(\mathfrak{q}, \mathfrak{q}_0, H_0)$*, of parameters* $(\mathfrak{q}, \mathfrak{q}_0, H_0)$ *is a vector of random probably measures* $(p_0, p_1, \ldots, p_I)$ *such that*

$$\begin{aligned} p_i | p_0 &\overset{iid}{\sim} SSrp(\mathfrak{q}, p_0), \qquad i = 1, \ldots, I, \\ p_0 &\sim SSrp(\mathfrak{q}_0, H_0). \end{aligned}$$

An array $[\xi_{i,j}]_{i=1,\ldots,I, j \geq 1}$ *is sampled from* $HSSM(\mathfrak{q}, \mathfrak{q}_0, H_0)$ *if its elements are conditionally independent random variables given* $(p_1, \ldots, p_I)$ *with* $\xi_{i,j} | (p_1, \ldots, p_I) \overset{ind}{\sim} p_i$*, where* $i = 1, \ldots, I$ *and* $j \geq 1$*.*

By de Finetti's representation theorem it follows that the array $[\xi_{i,j}]_{i=1,\ldots,I, j \geq 1}$ is partially exchangeable (in the sense of de Finetti), i.e.

$$\{(\xi_{i,j})_{j \geq 1}\}_{i=1,\ldots,I} \overset{\mathcal{L}}{=} \{(\xi_{i,\sigma_i(j)})_{j \geq 1}\}_{i=1,\ldots,I}$$

for any choice of (finite) permutations $\sigma_1, \ldots, \sigma_I$ of the integers $\{1, \ldots, I\}$ (see e.g. Kallenberg (2006)).

Definition 1 is general and provides a probabilistic foundation for a wide class of hierarchical random models. The properties of the SSrp and of the gSSS, guarantee that the hierarchical random measures in Definition 1 are well defined also for non-diffuse (e.g., atomic or mixed) probability measures H_0.

The HSSM class in Definition 1 includes well-known (e.g., Teh et al. (2006), Teh (2006), Bacallado et al. (2017)) and new hierarchical processes, as shown in the following examples. We assume that the reader is familiar with basic non-parametric prior processes. A brief account to these topics is included in the supplementary material (Bassetti et al., 2019a).

Example 1 (Hierarchical Pitman-Yor process)**.** *Let $PYP(\sigma,\theta,H)$ denote a Pitman-Yor process of parameters σ and θ, where $0 \leq \sigma < 1$ and $\theta > -\sigma$ (see Pitman (1995); Pitman and Yor (1997)). A vector of dependent random measures $(p_1,\dots,p_I)$, with law characterized by the following hierarchical structure*

$$\begin{aligned} &p_i|p_0 \overset{iid}{\sim} PYP(\sigma_1,\theta_1,p_0), \quad i=1,\dots,I, \\ &p_0|H_0 \sim PYP(\sigma_0,\theta_0,H_0) \end{aligned} \tag{3.1}$$

is called Hierarchical Pitman-Yor Process*, $HPYP(\sigma_0,\theta_0,\sigma_1,\theta_1,H_0)$, of parameters $0 \leq \sigma_i \leq 1$, $-\sigma_i < \theta_i$, $i=0,1$ and H_0 (see Teh (2006); Du et al. (2010); Lim et al. (2016); Camerlenghi et al. (2019)). By Definition 1, a $HPYP(\sigma_0,\theta_0,\sigma_1,\theta_1,H_0)$ is then a $HSSM$ of parameters $(\mathfrak{q},\mathfrak{q}_0,H_0)$ where $\mathfrak{q}$ and $\mathfrak{q}_0$ are Pitman-Yor EPPFs of parameters (σ_1,θ_1) and (σ_0,θ_0), respectively.*

If $\sigma_0=\sigma_1=0$, one recovers the Hierarchical Dirichlet process, in symbols $HDP(\theta_0,\theta_1,H_0)$. It is also possible to define some mixed cases*, where one considers a DP in one of the two stages of the hierarchy and a PYP with strictly positive discount parameter in the other, that are: $HDPYP(\theta_0,\sigma_1,\theta_1,H_0) = HPYP(0,\theta_0,\sigma_1,\theta_1,H_0)$ and $HPYDP(\sigma_0,\theta_0,\theta_1,H_0) = HPYP(\sigma_0,\theta_0,0,\theta_1,H_0)$. For an example of $HDPYP$ see Dubey et al. (2014).*

Example 2 (Hierarchical homogeneous normalized random measures)**.** *Hierarchical homogeneous Normalized Random Measures (HNRMI) introduced in Camerlenghi et al. (2019) are defined by*

$$\begin{aligned} &p_i|p_0 \overset{iid}{\sim} NRMI(\theta_1,\eta_1,p_0), \quad i=1,\dots,I, \\ &p_0|H_0 \sim NRMI(\theta_0,\eta_0,H_0), \end{aligned}$$

where $NRMI(\theta,\eta,H)$ denotes a normalized homogeneous random measure with parameters (θ,η,H), where $\theta>0$, η is Lévy a measure on $\mathbb{R}^+$ (absolutely continuous with respect to the Lebesgue measure) and H a measure on $\mathbb{X}$. A NRMI is a SSrp and hence HNRMI are HSSM.

Our class of HSSM includes new hierarchical processes such as hierarchical mixtures of finite mixture processes and combinations of finite mixture processes and PYP.

Example 3 (Hierarchical mixture of finite mixture processes)**.** *Let $\rho=(\rho_m)_{m\geq 1}$ be a probability measure on $\{1,2,\dots\}$, $\sigma>0$ and H a probability measure on $\mathbb{X}$. A mixture*

of finite mixture process of parameters (σ, ρ, H), $MFMP(\sigma, \rho, H)$ *from now on, is a random probability measure*

$$p = \sum_{k=1}^{K} q_k \delta_{Z_k}, \tag{3.2}$$

where $K \sim \rho$, $(Z_k)_{k\geq 1} \overset{iid}{\sim} H$, *and* $(q_1, \ldots, q_K) \mid K \sim$ *Dirichlet*$_K(\sigma, \ldots, \sigma)$ *see Miller and Harrison (2018). These random probability measures turn out to be* $SSrp(\mathfrak{q}, H)$ *for a suitable Gibbs Type EPPF* $\mathfrak{q}$.

A Hierarchical MFMP with parameters σ_i, $\rho^{(i)} = (\rho_k^{(i)})_{k\geq 1}$, $i = 0, 1$ *and base measure* H_0, *is*

$$\begin{aligned} &p_i | p_0 \overset{iid}{\sim} MFMP(\sigma_1, \rho^{(1)}, p_0), \quad i = 1, \ldots, I, \\ &p_0 | H_0 \sim MFMP(\sigma_0, \rho^{(0)}, H_0). \end{aligned} \tag{3.3}$$

As a special case when $|\sigma_i| = 1$ *and for a suitable* $\rho^{(i)}$ $(i = 0, 1, \ldots)$, *one obtains the Hierarchical Gnedin Process with parameters* $(\gamma_0, \zeta_0, \gamma_1, \zeta_1, H_0)$, *denoted with* $HGP(\gamma_0, \zeta_0, \gamma_1, \zeta_1, H_0)$, *which is a hierarchical extension of the Gnedin Process. For further details see Examples S.2 and S.3 in the supplementary material (Bassetti et al., 2019a).*

Example 4 (Mixed Cases)**.** *A hierarchical Gnedin-Pitman-Yor process, denoted with* $HGPYP(\gamma_0, \zeta_0, \sigma_1, \theta_1, H_0)$, *is defined as*

$$\begin{aligned} &p_i | p_0 \overset{iid}{\sim} PYP(\sigma_1, \theta_1, p_0), \quad i = 1, \ldots, I, \\ &p_0 | H_0 \sim GP(\gamma_0, \zeta_0, H_0) \end{aligned} \tag{3.4}$$

where $GP(\gamma_0, \zeta_0, H_0)$ *is a Gnedin Process. The hierarchical Gnedin-Dirichlet process is then obtained as special case for* $\sigma_1 = 0$ *and denoted with* $HGDP(\gamma_0, \zeta_0, \theta_1, H_0)$. *Exchanging the role of* PYP *and* GP *in the above construction, one gets the* $HPYGP(\sigma_0, \theta_0, \gamma_1, \zeta_1, H_0)$.

3.2 HSSM and Chinese Restaurant Franchising Representation

The next proposition gives the marginal law of an array sampled from a HSSM. When $\pi_n = [\pi_{1,n}, \ldots, \pi_{k,n}]$ is a partition of $[n]$ and $\mathfrak{q}$ an EPPF, we will write $\mathfrak{q}(\pi_n)$ in place of $\mathfrak{q}(|\pi_{1,n}|, \ldots, |\pi_{k,n}|)$.

Proposition 2. *Let* $[\xi_{i,j}]_{i=1,\ldots,I, j\geq 1}$ *be sampled from* $HSSM(\mathfrak{q}, \mathfrak{q}_0, H_0)$, *then for every vector of integer numbers* $(n_1, \ldots, n_I)$ *and every collection of Borel sets* $\{A_{i,j} : i = 1, \ldots, I, j = 1, \ldots, n_i\}$ *it holds*

$$\begin{aligned} &\mathbb{P}\{\xi_{i,j} \in A_{i,j} \ i = 1, \ldots, I, j = 1, \ldots, n_i\} \\ &= \sum_{\pi_{n_1}^{(1)} \in \mathcal{P}_{n_1}, \ldots, \pi_{n_I}^{(I)} \in \mathcal{P}_{n_I}} \prod_{i=1}^{I} \mathfrak{q}\left(\pi_{n_i}^{(i)}\right) \mathbb{E}\left[\prod_{i=1}^{I} \prod_{c=1}^{|\pi_{n_i}^{(i)}|} \tilde{p}\left(\cap_{j \in \pi_{c,n_i}^{(i)}} A_{i,j}\right)\right], \end{aligned} \tag{3.5}$$

with $\tilde{p} \sim SSrp(\mathfrak{q}_0, H_0)$.

Starting from Proposition 2 we show that an array sampled from a HSSM has a Chinese Restaurant Franchise representation. Such representation is very useful because it leads to a generative interpretation of the nonparametric-priors in the HSSM class, and naturally allows for posterior simulation procedures (see Section 5).

In the Chinese Restaurant Franchise metaphor, observations are attributed to "customers", identified by the indices (i, j), and groups are described as "restaurants" $(i = 1, \dots, I)$. In each "restaurant", "customers" are clustered according to "tables", which are then clustered at the second hierarchy level by means of "dishes". Observations are clustered across restaurants at the second level of the clustering process, when dishes are associated to tables. One can think that the first customer sitting at each table chooses a dish from a common menu and this dish is shared by all other customers who join the same table afterwards.

The first level of the clustering process, acting within each group, is driven by independent random partitions $\Pi^{(1)}, \dots, \Pi^{(I)}$ with EPPF $\mathfrak{q}$. The second level, acting between groups, is driven by a random partition $\Pi^{(0)}$ with EPPF $\mathfrak{q}_0$.

Given $n_1, \dots, n_I$ integer numbers, we introduce the following set of observations:

$$\mathcal{O} := \{\xi_{i,j} : j = 1, \dots, n_i; i = 1, \dots I\},$$

and denote with $\mathscr{C}_j(\Pi)$ the random index of the block of the random partition Π that contains j, that is

$$\mathscr{C}_j(\Pi) = c \quad \text{if } j \in \Pi_{c,j}. \tag{3.6}$$

Theorem 1. *If $[\xi_{i,j}]_{i=1,\dots,I, j\geq 1}$ is a sample from a $HSSM(\mathfrak{q}, \mathfrak{q}_0, H_0)$, then $\mathcal{O}$ and $\{\phi_{d^*_{i,j}} : j = 1, \dots, n_i; i = 1, \dots I\}$ have the same laws, where*

$$d^*_{i,j} := \mathscr{C}_{\mathscr{D}(i,c_{i,j})}\left(\Pi^{(0)}\right), \quad \mathscr{D}(i,c) := \sum_{i'=1}^{i-1} |\Pi^{(i')}_{n_{i'}}| + c, \quad c_{i,j} := \mathscr{C}_j\left(\Pi^{(i)}\right),$$

$(\phi_n)_n$ is a sequence of i.i.d. random variables with distribution H_0, $\Pi^{(1)}, \dots, \Pi^{(I)}$ are i.i.d. exchangeable partitions with EPPF $\mathfrak{q}$ and $\Pi^{(0)}$ is an exchangeable partition with EPPF $\mathfrak{q}_0$. All the previous random variables are independent.

Since $d^*_{i,j} = d_{i,c_{i,j}}$ for $d_{i,c} := \mathscr{C}_{\mathscr{D}(i,c)}\left(\Pi^{(0)}\right)$, then the construction in Theorem 1 can be summarized by the following hierarchical structure

$$\begin{aligned}
&\xi_{i,j} = \phi_{d_{i,c_{i,j}}},\\
&d_{i,c} = \mathscr{C}_{\mathscr{D}(i,c)}\left(\Pi^{(0)}\right),\ c_{i,j} = \mathscr{C}_j\left(\Pi^{(i)}\right),\\
&\phi_n \overset{i.i.d.}{\sim} H_0,\\
&\left(\Pi^{(0)}, \Pi^{(1)}, \dots, \Pi^{(I)}\right) \sim \mathfrak{q}_0 \otimes \mathfrak{q} \otimes \cdots \otimes \mathfrak{q},
\end{aligned} \tag{3.7}$$

where, following the Chinese Restaurant Franchise metaphor (see Figure 1), $c_{i,j}$ is the table at which the j-th "customer" of the "restaurant" i sits, $d_{i,c}$ is the index of the

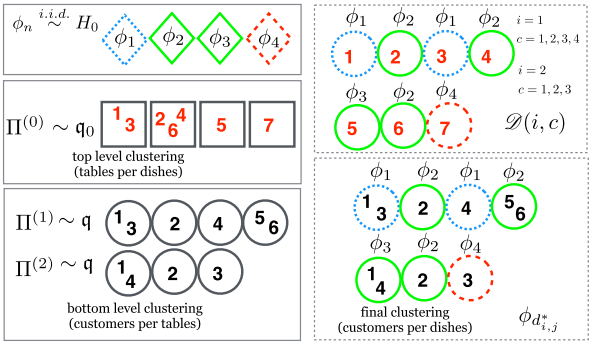


Figure 1: Illustration of the $HSSM(\mathfrak{q}, \mathfrak{q}_0, H_0)$ clustering process given in Theorem 1. We assume two groups (restaurants), $I = 2$, with $n_1 = 6$ and $n_2 = 4$ observations (customers) each. Top-left: Samples (dishes) ϕ_n from the non-diffuse base measure. Dishes have the same colour and line type if they take the same values. Mid-left: Indexes $\mathscr{D}(i, c)$ (from 1 to 7 in lexicographical order) of the tables which share the same dish. Boxes represent the blocks of the random partition at the top of the hierarchy. Bottom-left: Observations (customers) allocated by $c_{i,j}$ to each table (circles) in the group-specific random partitions. Top-right: Table lexicographical ordering and dishes assigned to the tables by the top level partition. Bottom-right: observations clustering implied by the joint tables and dishes allocation $d^*_{i,j}$.

"dish" served at table c in the restaurant i and $d^*_{i,j}$ is the index of the "dish" served to the j-th customer of the i-th restaurant.

A special case of Theorem 1 has been independently proved in Proposition 2 of Argiento et al. (2019) for HNRMI. Theorem 1 can also be used to describe in a recursive way the array $\mathcal{O}$. Having in mind the Chinese Restaurant Franchise, we shall denote with n_{icd} the number of customers in restaurant i seated at table c and being served dish d and with m_{id} the number of tables in the restaurant i serving dish d. We denote with dots the marginal counts. Thus, $n_{i\cdot d}$ is the number of customers in restaurant i being served dish d, $m_{i\cdot}$ is the number of tables in restaurant i, $n_{i\cdot\cdot}$ is the number of customers in restaurant i (i.e. the n_i observations), and $m_{\cdot\cdot}$ is the number of tables.

Finally, let $\omega_{n,k}$ and ν_n be the weights of the predictive distribution of the random partitions $\Pi^{(i)}$ $(i = 1, \ldots, I)$ with EPPF $\mathfrak{q}$ (see Section 2.1). Also, let $\tilde{\omega}_{n,k}$ and $\tilde{\nu}_n$ be the weights of the predictive distribution of the random partitions $\Pi^{(0)}$ with EPPF $\mathfrak{q}_0$ defined analogously by using $\mathfrak{q}_0$ in place of $\mathfrak{q}$. We can sample $\{\xi_{i,j}; j = 1, \ldots, n_i, i = 1, \ldots I\}$ starting with $i = 1$, $m_{1\cdot} = 1$, $n_{11\cdot} = 1$, $D = 1$, $m_{\cdot 1} = 1$ and $\xi_{1,1} = \xi^*_{1,1} = \phi_1 \sim H_0$ and then iterating, for $i = 1, \ldots, I$, the following steps:

(S1) for $t = 2, \ldots, n_{i\cdot\cdot}$, sample $\xi_{i,t}$ given $\xi_{i,1}, \ldots, \xi_{i,t-1}$ and $k := m_{i\cdot}$ from $G^*_{it}(\cdot) + \nu_t(n_{i1\cdot}, \ldots, n_{ik\cdot})G_{it}(\cdot)$ where

$$G_{it}(\cdot) = \tilde{G}_{it}(\cdot) + \tilde{\nu}_{m_{\cdot\cdot}}(m_{\cdot 1}, \ldots, m_{\cdot D})H_0(\cdot),$$

$$G^*_{it}(\cdot) = \sum_{c=1}^{k} \omega_{t,c}(n_{i1\cdot}, \ldots, n_{ik\cdot})\delta_{\xi^*_{i,c}}(\cdot),$$
$$\tilde{G}_{it}(\cdot) = \sum_{d=1}^{D} \tilde{\omega}_{m_{\cdot\cdot}d}(m_{\cdot 1}, \ldots, m_{\cdot D})\delta_{\phi_d}(\cdot).$$

(S2) If $\xi_{i,t}$ is sampled from G^*_{it}, then we set $\xi_{i,t} = \xi^*_{i,c}$ and let $c_{it} = c$ for the chosen c, we leave $m_{i\cdot}$ the same and set $n_{ic\cdot} := n_{ic\cdot} + 1$, while, if $\xi_{i,t}$ is sampled from G_{it}, then we set $m_{i\cdot} := m_{i\cdot} + 1$, $\xi_{i,t} = \xi^*_{i,m_{i\cdot}}$ and $c_{it} = m_{i\cdot}$. If $\xi_{i,t}$ is sampled from $\tilde{G}_{it}$, we set $\xi^*_{i,c} = \phi_d$, let $d_{ic} = d$ for chosen d and increment $m_{\cdot c}$ by one. If $\xi_{i,t}$ is sampled from H_0, then we increment D by one and set $\phi_D = \xi_{it}$, $\xi^*_{i,c} = \xi_{i,t}$ and $d_{ic} = D$. In both cases, we increment $m_{\cdot\cdot}$ by one.

(S3) Having sampled $\xi_{i,t}$ with $t = n_{i\cdot\cdot}$ in the previous Step, set $i := i + 1$, $m_{i\cdot} = 1$, $n_{i1\cdot} = 1$ take $\xi_{i,1} = \xi^*_{i,1}$ where $\xi^*_{i,1}$ is sampled from G_{it}. Update D, $m_{\cdot c}$, d_{ic} and $m_{\cdot\cdot}$ as described in the previous Step.

Remark 1. *The Chinese Restaurant Franchise representation and the Pólya Urn sampler in (S1)–(S3) are deduced directly from the latent partition representation given in Theorem 1, with no additional assumptions on H_0 and without resorting to the expression of the distribution of the partition induced by the observations. This expression can be derived for HSSM as a side result of our combinatorial framework and includes Theorem 3 and 4 of Camerlenghi et al. (2019) as special cases when the HSSM is a HNRMI. Since the derivation of this law is not a central result of the paper, it is given in the supplementary material (Bassetti et al., 2019a).*

4 Cluster Sizes Distributions

We study the distribution of the number of clusters in each group of observations (i.e., the number of distinct dishes served in the restaurant i), as well as the global number of clusters (i.e. the total number of distinct dishes in the restaurant franchise).

Let us introduce a time index t to describe the customers arrival process. At time $t = 1, 2, \ldots$ and for each group i, $\mathcal{O}_{it}$ is the observation set and $n_i(t)$ is the number of elements in $\mathcal{O}_{it}$, i.e. the number of observations in the group i at time t. The collection of all the $n(t) := \sum_{i=1}^{I} n_i(t)$ observations at time t is $\mathcal{O}_t := \cup_{i=1}^{I} \mathcal{O}_{it}$. For example, if $n_i(t) = n_i(t-1) + 1$, with $n_i(1) = 1$, each group has one new observation between $t-1$ and t and hence the total number of observations at time t is $n(t) = It$. Different sampling rates can be assumed within our framework. For example $n_i(t) = tb_i$ for suitable integers b_i describes an asymmetric sampling scheme in which groups have different arrival rates, b_i.

We find the exact finite sample distribution of the number of clusters for given $n(t)$ and $n_i(t)$ when $t < \infty$. Some properties, such as the prior mean and variance, are discussed in order to provide some guidelines to setting HSSM parameters in the applications. We present some new asymptotic results when the number of observations goes

to infinity, such that $n(t)$ diverges to $+\infty$ as t goes to $+\infty$. The results extend existing asymptotic approximations for species sampling (Pitman (2006)) and for hierarchical normalized random measures (Camerlenghi et al. (2019)) to the general class of HSSMs. Finally, we provide a numerical study of the approximation accuracy.

4.1 Distribution of the Cluster Size Under the Prior

For every $i = 1, \ldots, I$, we define

$$K_{i,t} := |\Pi^{(i)}_{n_i(t)}|, \quad K_t := \sum_{i=1}^{I} |\Pi^{(i)}_{n_i(t)}|, \quad D_{i,t} = |\Pi^{(0)}_{K_{i,t}}|, \quad D_t = |\Pi^{(0)}_{K_t}|.$$

By Theorem 1, for every fixed t, the laws of $K_{i,t}$ and K_t are the same as the ones of the number of "active tables" in "restaurant" i and of the total number of "active tables" in the whole franchise, respectively. Analogously, the laws of D_t and $D_{i,t}$ are the same as the laws of the number of dishes served in the restaurant i and in the whole franchise, respectively. If H_0 is diffuse, then D_t and the number of distinct clusters in $\mathcal{O}_t$ have the same law and also $D_{i,t}$ and the number of clusters in the group i follow the same law.

The distributions of D_t and $D_{i,t}$ are derived in the following

Proposition 3. *For every $n \geq 1$ and $k = 1, \ldots, n$, we define $q_n(k) := \mathbb{P}\left\{|\Pi^{(i)}_n| = k\right\}$ and $q^{(0)}_n(k) := \mathbb{P}\left\{|\Pi^{(0)}_n| = k\right\}$. Then, for every $i = 1, \ldots, I$,*

$$\mathbb{P}\{D_{i,t} = k\} = \sum_{m=k}^{n_i(t)} q_{n_i(t)}(m) q^{(0)}_m(k), \quad k = 1, \ldots, n_i(t),$$

and

$$\mathbb{P}\{D_t = k\} = \sum_{m=\max(I,k)}^{n(t)} \left(\sum_{\substack{m_1+\cdots+m_I=m, \\ 1\leq m_i\leq n_i(t)}} \prod_{i=1}^{I} q_{n_i(t)}(m_i) \right) q^{(0)}_m(k), \quad k = 1, \ldots, n(t).$$

Moreover, for every $r > 0$

$$\mathbb{E}\left[D^r_{i,t}\right] = \sum_{m=1}^{n_i(t)} \mathbb{E}\left[|\Pi^{(0)}_m|^r\right] q_{n_i(t)}(m)$$

and

$$\mathbb{E}\left[D^r_t\right] = \sum_{\substack{m_1,\ldots,m_I: \\ 1\leq m_i\leq n_i(t)}} \mathbb{E}\left[\left|\Pi^{(0)}_{\sum_{i=1}^{I} m_i}\right|^r\right] \prod_{i=1}^{I} q_{n_i(t)}(m_i).$$

In particular, for every $i = 1, \ldots, I$, $n \geq 1$ and $k = 1, \ldots, n$,

$$q_n(k) = \sum_{(\lambda_1,\ldots,\lambda_n)\in\Lambda_{n,k}} \frac{n!}{\prod_{j=1}^{n}(\lambda_j!)(j!)^{\lambda_j}} \mathfrak{q}\left[[\lambda_1, \ldots, \lambda_n]\right],$$

where $\Lambda_{n,k}$ is the set of integers $(\lambda_1, \ldots, \lambda_n)$ such that $\sum_j \lambda_j = k$ and $\sum_j j\lambda_j = n$, $\mathfrak{q}[[\lambda_1, \ldots, \lambda_n]]$ is the common value of the symmetric function $\mathfrak{q}$ for all $n_1, \ldots, n_k$ with $|\{i : n_i = j\}| = \lambda_j$ for $j = 1, \ldots, n$ and $n!/(\prod_{j=1}^n (\lambda_j!)(j!)^{\lambda_j})$ is the number of partitions of $[n]$ with λ_j blocks of cardinality $j = 1, \ldots, n$ (see Equation (11) in Pitman (1995)). Similar expressions can be obtained for $q_n^{(0)}(k)$.

The results in Proposition 3 generalize to HSSM those in Theorem 5 of Camerlenghi et al. (2019) for HNRMI. Our proof relies on the hierarchical Species Sampling Sequence construction of HSSM processes (see Theorem 1) and does not require any knowledge of the partial exchangeable partition, whereas the proof in Camerlenghi et al. (2019) builds on the partial exchangeable partition function.

One of the advantages of our framework is that the gSSS properties allow us to easily derive the distribution of the number of clusters when H_0 is not diffuse. Indeed, it can be deduced by considering possible coalescences of latent clusters (due to ties in the i.i.d. sequence $(\phi_n)_n$ of Theorem 1) forming a true cluster. Let us denote with $\tilde{D}_t$ and $\tilde{D}_{i,t}$ the number of distinct clusters in $\mathcal{O}_t$ and $\mathcal{O}_{it}$, respectively.

Proposition 4. *Let $\tilde{H}_0(d|k)$ (for $1 \leq d \leq k$) be the probability of observing exactly d distinct values in the vector $(\phi_1, \ldots, \phi_k)$ where the ϕ_ns are i.i.d. H_0. Then,*

$$\mathbb{P}\{\tilde{D}_{i,t} = d\} = \sum_{k=d}^{n_i(t)} \tilde{H}_0(d|k)\mathbb{P}\{D_{i,t} = k\}$$

for $d = 1, \ldots, n_i(t)$. The probability of $\tilde{D}_t$ has the same expression as above with D_t in place of $D_{i,t}$ and $n(t)$ in place of $n_i(t)$. If H_0 is diffuse, then $\mathbb{P}\{\tilde{D}_{i,t} = d\} = \mathbb{P}\{D_{i,t} = d\}$ and $\mathbb{P}\{\tilde{D}_t = d\} = \mathbb{P}\{D_t = d\}$, for every $d \geq 1$.

The assumption of atomic base measures behind HDP and HPYP has been used in many studies, and some of its theoretical and computational implications have been investigated (e.g., see Nguyen (2016) and Sohn and Xing (2009)), whereas the implications of the use of mixed base measures are not yet well studied, especially in hierarchical constructions. In the following we state some new results for the case of a spike-and-slab base measure.

Proposition 5. *Assume that $H_0(dx) = a\delta_{x_0}(dx) + (1-a)H_C(dx)$, where $a \in (0,1)$, x_0 is a point of $\mathbb{X}$ and H_C is a diffuse measure on $\mathbb{X}$, then*

$$\mathbb{P}\{\tilde{D}_{i,t} = d\} = (1-a)^d \mathbb{P}\{D_{i,t} = d\} + \sum_{k=d}^{n_i(t)} \binom{k}{d-1} a^{k+1-d}(1-a)^{d-1}\mathbb{P}\{D_{i,t} = k\},$$

for $d = 1, \ldots, n_i(t)$. The probability of $\tilde{D}_t$ has the same expression as above with D_t in place of $D_{i,t}$ and n_t in place of $n_{i,t}$. Moreover,

$$\mathbb{E}[\tilde{D}_{i,t}] = 1 - \mathbb{E}[(1-a)^{D_{i,t}}] + (1-a)\mathbb{E}[D_{i,t}] \leq \mathbb{E}[D_{i,t}]$$

and $\mathbb{E}[\tilde{D}_t]$ has an analogous expression with $D_{i,t}$ replaced by D_t.

For a Gibbs-type EPPF with $\sigma > 0$, using results in Gnedin and Pitman (2005), we get

$$q_n(k) = V_{n,k} S_\sigma(n,k),$$

where $V_{n,k}$ satisfies the partial difference equation in (2.4) and $S_\sigma(n,k)$ is a generalized Stirling number of the first kind, defined as

$$S_\sigma(n,k) = \frac{1}{\sigma^k k!} \sum_{i=1}^{k} (-1)^i \binom{k}{i} (-i\sigma)_n,$$

for $\sigma \neq 0$ and $S_0(n,k) = |s(n,k)|$ for $\sigma = 0$, where $|s(n,k)|$ is the unsigned Stirling number of the first kind, see Pitman (2006). See De Blasi et al. (2015) for an up-to-date review of Gibbs-type prior processes.

For the hierarchical PY process the distribution $q_n(k)$ has closed-form expression

$$q_n(k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}} S_\sigma(n,k),$$

when $0 < \sigma < 1$ and $\theta > -\sigma$, whilst

$$q_n(k) = \frac{\theta^k \Gamma(\theta)}{\Gamma(\theta+n)} |s(n,k)|,$$

when $\sigma = 0$.

For the Gnedin model (Gnedin, 2010) the distribution $q_n(k)$ is

$$q_n(k) = \binom{n-1}{k-1} \frac{n!}{k!} \nu_{n,k}, \text{ with } \nu_{n,k} = \frac{(\gamma)_{n-k} \prod_{i=1}^{k-1}(i^2 - \gamma i + \zeta)}{\prod_{m=1}^{n-1}(m^2 + \gamma m + \zeta)}. \tag{4.1}$$

In the supplementary material (Bassetti et al., 2019b), we provide a graphical illustration of the prior distributions presented here above and a sensitivity analysis with respect to the prior parameters.

4.2 Asymptotic Distribution of the Cluster Size

An exchangeable random partition $(\Pi_n)_{n\geq 1}$ has asymptotic diversity S if

$$|\Pi_n|/c_n \to S \quad a.s. \tag{4.2}$$

for a positive random variable S and a suitable normalizing sequence $(c_n)_{n\geq 1}$. Asymptotic diversity generalizes the notion of σ-diversity, see Definition 3.10 in Pitman (2006). An exchangeable random partition $(\Pi_n)_{n\geq 1}$ has σ-diversity S if (4.2) holds with $c_n = n^\sigma$. For any Gibbs-type partition $(\Pi_n)_{n\geq 1}$, (4.2) holds with $c_n = 1$ if $\sigma < 0$, $c_n = \log(n)$ if $\sigma = 0$, and $c_n = n^\sigma$ if $\sigma > 0$ (see Section 6.1 of Pitman (2003) and Lemma 3.1 in Pitman (2006)).

In the following propositions, we use the (marginal) limiting behaviour (4.2) of the random partitions $\Pi_n^{(i)}$ $(i = 0, \dots, I)$, to obtain the asymptotic distribution of $D_{i,t}$ and D_t assuming $c_n = n^\sigma L(n)$, with L slowly-varying.

The first general result deals with HSSM where $\Pi_n = \Pi_n^{(i)}$ satisfies (4.2) for every $i = 1, \dots, I$ and $c_n \to +\infty$ and hence the cluster size $|\Pi_n^{(i)}|$ diverges to $+\infty$.

Proposition 6. *Assume that $\Pi^{(0)}$ and $\Pi^{(i)}$ (for $i = 1, \dots, I$) are independent exchangeable random partitions such that $|\Pi_n^{(0)}|/a_n$ ($|\Pi_n^{(i)}|/b_n$ for $i = 1, \dots, I$, respectively) converges almost surely to a strictly positive random variable $D_\infty^{(0)}$ ($D_\infty^{(i)}$, respectively) for suitable diverging sequences a_n and b_n. Moreover assume that $a_n = n^{\sigma_0} L_0(n)$ and $b_n = n^{\sigma_1} L_1(n)$, with $\sigma_i \geq 0$ and L_i is a slowly varying function, $i = 0, 1$, and set $d_n := a_{b_n} = n^{\sigma_0 \sigma_1} L_0(n^{\sigma_1} L_1(n))$.*

(i) If $\lim_{t \to +\infty} n_i(t) = +\infty$ for some i, then for $t \to +\infty$

$$\frac{D_{i,t}}{d_{n_i(t)}} \to D_\infty^{(0)} \left(D_\infty^{(i)} \right)^{\sigma_0} \quad a.s.$$

(ii) If $\lim_{t \to +\infty} n_i(t) = +\infty$ and $n_i(t)/n(t) \to \ w_i > 0$ for every $i = 1, \dots, I$ then for $t \to +\infty$

$$\frac{D_t}{d_{n(t)}} \to D_\infty^{(0)} \left(\sum_{i=1}^{I} w_i^{\sigma_1} D_\infty^{(i)} \right)^{\sigma_0} \quad a.s.$$

Remark 2. *Part (ii) extends to HSSM with different group sizes, $n_i(t)$, the results in Theorem 7 of Camerlenghi et al. (2019) for HNRMI with groups of equal size. Both part (i) and (ii) provide deterministic scaling of diversities, in the spirit of Pitman (2006), and differently from Camerlenghi et al. (2019) where a random scaling is obtained.*

Remark 3. *Combining Propositions 4 and 6 one can obtain similar asymptotic results also for $\tilde{D}_{i,t}$ and $\tilde{D}_t$. For instance, one can prove that, under the same assumptions of Proposition 4, if $H_0(dx) = a \sum_{i=1}^{M} \beta_i \delta_{x_i}(dx) + (1-a) H_C(dx)$ with $0 < a < 1$, $\sum_i \beta_i = 1$ and H_C diffuse (as in the spike-and-slab case), for $t \to +\infty$ one has*

$$\frac{\tilde{D}_{i,t}}{d_{n_i(t)}} \to (1-a) D_\infty^{(0)} \left(D_\infty^{(i)} \right)^{\sigma_0} \quad a.s.$$

and

$$\frac{\tilde{D}_t}{d_{n(t)}} \to (1-a) D_\infty^{(0)} \left(\sum_{i=1}^{I} w_i^{\sigma_1} D_\infty^{(i)} \right)^{\sigma_0} \quad a.s.$$

The second general result describes the asymptotic behaviour of $D_{i,t}$ and D_t in presence of random partitions for which $c_n = 1$ for every n.

Proposition 7. *Assume that $\Pi^{(0)}$ and $\Pi^{(i)}$, $i = 1, \dots, I$ are independent exchangeable random partitions and that $\lim_{t \to \infty} n_i(t) = +\infty$ for every $i = 1, \dots, I$.*

(i) If $|\Pi_n^{(i)}|$ *converges a.s. to a positive random variable* K_i *as* $n \to +\infty$*, then for every* $k \geq 1$

$$\lim_{t\to+\infty} \mathbb{P}\{D_{i,t} = k\} = \sum_{m\geq k} \mathbb{P}\{K_i = m\}\, q_m^{(0)}(k),$$

and

$$\lim_{t\to+\infty} \mathbb{P}\{D_t = k\} = \sum_{m\geq \max(I,k)} \sum_{\substack{m_1+\cdots+m_I=m,\\ 1\leq m_i}} q_m^{(0)}(k) \prod_{i=1}^{I} \mathbb{P}\{K_i = m_i\}.$$

(ii) If $|\Pi_n^{(i)}|/b_n$ *converges a.s. to a strictly positive random variable* $D_\infty^{(i)}$ *for a suitable diverging sequences* b_n *and* $|\Pi_n^{(0)}|$ *converges a.s. to a positive random variable* K_0 *as* $n \to +\infty$*, then, for every* $k \geq 1$*,*

$$\lim_{t\to+\infty} \mathbb{P}\{D_t = k\} = \lim_{t\to+\infty} \mathbb{P}\{D_{i,t} = k\} = \mathbb{P}\{K_0 = k\}.$$

Starting from Propositions 6 and 7, analytic expressions for the asymptotic distributions of $D_{i,t}$ and D_t can be deduced for some special HSSMs.

As an example, consider the HGP and the HPYGP in Examples 3 and 4. If $(\Pi_n)_n$ is a Gnedin's partition, then $|\Pi_n|$ converges almost surely to a random variable K (see Gnedin (2010) and Example S.3 in the supplementary material (Bassetti et al., 2019a)) and the asymptotic behaviour of the number of clusters can be derived from Proposition 7 as stated here below.

Proposition 8. *In a* $HGP(\gamma_0, \zeta_0, \gamma_1, \zeta_1, H_0)$*, one has*

$$\lim_{t\to+\infty} \mathbb{P}\{D_{i,t} = k\} = \frac{c_{\gamma_1,\zeta_1}}{k!} \left(\prod_{i=1}^{k-1} (i^2 - \gamma_0 i + \zeta_0) \right) \times \sum_{m\geq k} \frac{(\gamma_0)_{m-k}}{(k-1)!(m-k)!} \prod_{j=1}^{m-1} \frac{(j^2 - \gamma_1 j + \zeta_1)}{(j^2 + \gamma_0 j + \zeta_0)}$$

with

$$c_{\gamma_1,\zeta_1} = \frac{\Gamma(1 + (\gamma_1 + \sqrt{\gamma_1^2 - 4\zeta_1})/2)\Gamma(1 + (\gamma_1 - \sqrt{\gamma_1^2 - 4\zeta_1})/2)}{\Gamma(\gamma_1)}.$$

In contrast, in a $HPYGP(\sigma_0, \theta_0, \gamma_1, \zeta_1, H_0)$*,*

$$\lim_{t\to+\infty} \mathbb{P}\{D_t = m\} = \lim_{t\to+\infty} \mathbb{P}\{D_{i,t} = m\} = c_{\gamma_1,\zeta_1} \frac{\prod_{l=1}^{m-1} (l^2 - \gamma l + \zeta)}{m!(m-1)!}.$$

For HPYPs one can derive explicit asymptotic distributions using the previous general results. From now on, $(\Pi_n)_n \sim PY(\sigma, \theta)$ denotes a random partition with Pitman-Yor EPPF of parameter (σ, θ). If $(\Pi_n)_n \sim PY(\sigma, \theta)$ with $0 < \sigma < 1$ and $\theta > -\sigma$,

then $|\Pi_n|/n^\sigma$ converges almost surely and in L^p (for every $p>0$) to a strictly positive random variable $S_{\sigma,\theta}$ with density

$$g_{\sigma,\theta}(s) := \frac{\Gamma(\theta+1)}{\Gamma(\frac{\theta}{\sigma}+1)} s^{\theta/\sigma} g_\sigma(s), \quad s>0, \tag{4.3}$$

where g_σ is the type-2 Mittag-Leffler density, i.e. the unique density such that

$$\int_0^{+\infty} x^p g_\sigma(x)dx = \frac{\Gamma(p+1)}{\Gamma(p\sigma+1)}. \tag{4.4}$$

See Theorem 3.8 in Pitman (2006). Moreover, if $\sigma=0$, we have that $|\Pi_n|/\log(n)$ converges almost surely and in L^p for every $p>0$ to $\theta>0$.

On the basis of these results, Proposition 6 can be specialized for the case of HPYPs and convergence in L^p obtained.

Proposition 9. *Assume that* $\Pi^{(0)} \sim PY(\sigma_0,\theta_0)$ *and* $\Pi^{(i)} \sim PY(\sigma_1,\theta_1)$ *(for* $i=1,\dots,I$*) with* $\sigma_0,\sigma_1 \geq 0$*. Then (i) and (ii) of Proposition 6 hold a.s. and in* L^p*,* $p>0$*, with the following specifications:*

(i) for $HPYP(\sigma_0,\theta_0;\sigma_1,\theta_1,H_0)$ *with* $\sigma_0,\sigma_1>0$*,* $d_n = n^{\sigma_0\sigma_1}$ *and*

$$D_{i,\infty} \overset{\mathcal{L}}{=} S_{\sigma_0,\theta_0}\left(S^{(i)}_{\sigma_1,\theta_1}\right)^{\sigma_0}, \quad D_\infty \overset{\mathcal{L}}{=} S_{\sigma_0,\theta_0}\left(\sum_{i=1}^I w_i^{\sigma_1} S^{(i)}_{\sigma_1,\theta_1}\right)^{\sigma_0},$$

with $S_{\sigma_0,\theta_0}, S^{(1)}_{\sigma_1,\theta_1},\dots,S^{(I)}_{\sigma_1,\theta_1}$ *independent random variables with densities* g_{σ_0,θ_0} *and* g_{σ_1,θ_1}*, respectively;*

(ii) for $HPYDP(\sigma_0,\theta_0;\theta_1,H_0)$ *with* $\sigma_0>0$*,* $d_n=\log(n)^{\sigma_0}$ *and*

$$D_{i,\infty} \overset{\mathcal{L}}{=} D_\infty \overset{\mathcal{L}}{=} S_{\sigma_0,\theta_0}\theta_1^{\sigma_0},$$

with S_{σ_0,θ_0} *random variable with density* g_{σ_0,θ_0}*;*

(iii) for $HDPYP(\theta_0;\sigma_1,\theta_1,H_0)$ *with* $\sigma_1>0$*,* $d_n=\sigma_1\log(n)$ *and* $D_{i,\infty}=D_\infty=\theta_0$*;*

(iv) for $HDP(\theta_0;\theta_1,H_0)$*,* $d_n=\log(\log(n))$ *and* $D_{i,\infty}=D_\infty=\theta_0$*.*

Proposition 9 can be used for approximating the moments (e.g., expectation and variance) of the number of clusters as stated in the following

Corollary 2. *Let* $x_n \simeq y_n$ *if and only if* $\lim_{n\to+\infty} x_n/y_n = 1$*, then under the same assumptions of Proposition 9, for every* $r>0$*:*

(i) for $HPYP(\sigma_0,\theta_0,\sigma_1,\theta_1)$ *with* $\sigma_0,\sigma_1>0$*:*

$$\mathbb{E}\left[D^r_{i,t}\right] \simeq n_i(t)^{r\sigma_0\sigma_1} \frac{\Gamma(\theta_0+1)}{\Gamma\left(\frac{\theta_0}{\sigma_0}+1\right)} \frac{\Gamma\left(r+\frac{\theta_0}{\sigma_0}+1\right)}{\Gamma(\theta_0+r\sigma_0+1)} \frac{\Gamma(\theta_1+1)}{\Gamma\left(\frac{\theta_1}{\sigma_1}+1\right)} \frac{\Gamma\left(r\sigma_0+\frac{\theta_1}{\sigma_1}+1\right)}{\Gamma(\theta_1+r\sigma_0\sigma_1+1)};$$

(ii) for $HPYDP(\theta_0, \sigma_0; \theta_1)$ *with* $\sigma_0 > 0$*:*

$$\mathbb{E}\left[D_{i,t}^r\right] \simeq (\log(n_i(t)))^{r\sigma_0} \theta_1^{r\sigma_0} \frac{\Gamma(\theta_0+1)}{\Gamma\left(\frac{\theta_0}{\sigma_0}+1\right)} \frac{\Gamma\left(r+\frac{\theta_0}{\sigma_0}+1\right)}{\Gamma(\theta_0+r\sigma_0+1)};$$

(iii) for $HDPYP(\theta_0, \sigma_1, \theta_1)$ *with* $\sigma_1 > 0$*:* $\mathbb{E}\left[D_{i,t}^r\right] \simeq \log(n_i(t))^r (\sigma_1\theta_0)^r$*;*

(iv) for $HDP(\theta_0, \theta_1)$*:* $\mathbb{E}\left[D_{i,t}^r\right] \simeq \theta_0^r \log(\log(n_i(t)))^r$.

In Figure 2, we compare exact and asymptotic values (see Proposition 3 and Corollary 2, respectively) of the expected marginal number of clusters for the HSSMs in the PY family: $HDP(\theta_0; \theta_1)$, $HDPYP(\theta_0; \sigma_1, \theta_1)$, $HPYP(\sigma_0, \theta_0; \sigma_1, \theta_1)$ and $HPYDP(\theta_0, \sigma_0; \theta_1)$ (different rows of Figure 2). For each HSSM we consider $n_i(t)$ increasing from 1 to 500 and different parameter settings (different columns and lines). For the HDP the exact value (dashed lines) is well approximated by the asymptotic one (solid line) for all sample sizes $n_i(t)$, and different values of θ_i (gray and blacks lines in the left and right plots of panel (i)). For the HPYP, the results in panel (ii) show that there are larger differences when θ_i, $i = 0, 1$ are large and σ_0 and σ_1 are close to zero (left plot). The approximation is good for small θ_i (right plot) and improves slowly with increasing $n_i(t)$ for smaller σ_i (gray lines in the right plot). In the panels (iii) and (iv) for HDPYP and HPYDP, there exist parameter settings where the asymptotic approximation is not satisfactory and is not improving when $n_i(t)$ increases.

Our numerical results point out that the asymptotic approximation for both PY and HPY lacks of accuracy for some parameters settings. Thus, the exact formula for the number of clusters should be used in the applications when calibrating the parameters of the process.

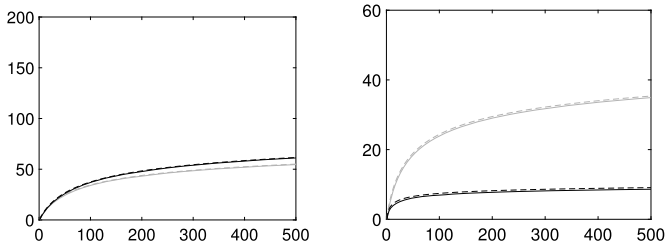
5 Chinese Restaurant Franchise Sampler

Random measures and hierarchical random measures are widely used in Bayesian nonparametric inference (see Hjort et al. (2010) for an introduction) as prior distributions for the parameters of a given density function. In this context a further stage is added to the hierarchical structure of Equation (3.7) involving an observation model

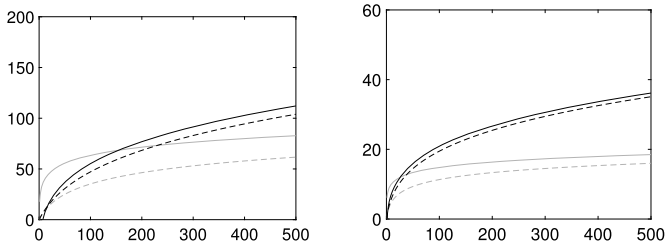
$$Y_{i,j}|\xi_{i,j} \overset{ind}{\sim} f(\cdot|\xi_{i,j}),$$

where f is a suitable kernel density.

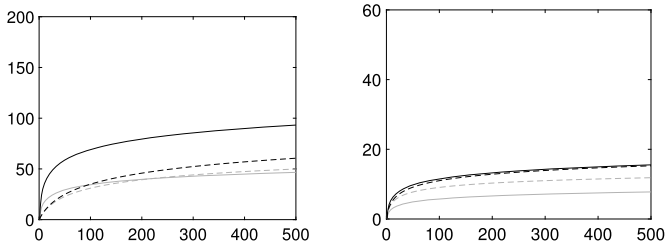
The resulting model is an infinite mixture, which is the object of the Bayesian inference. In this framework, the posterior distribution is usually not tractable and Gibbs sampling is used to approximate the posterior quantities of interest. There are two main classes of samplers for posterior approximation in Bayesian nonparametrics: marginal (see Escobar (1994) and Escobar and West (1995)) and conditional (Walker (2007), Papaspiliopoulos and Roberts (2008), Kalli et al. (2011)) samplers. See also



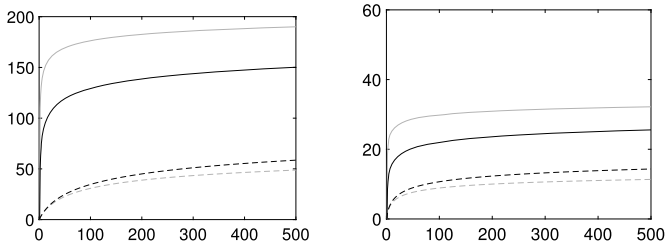
(i) HDP with $\theta_0 = \theta_1 = 43.3$ (left, gray), $\theta_0 = \theta_1 = 50$ (left, black), $\theta_0 = \theta_1 = 25$ (right, gray) and $\theta_0 = \theta_1 = 5$ (right, black).



(ii) HPYP with $(\theta_0, \sigma_0) = (\theta_1, \sigma_1) = (29.9, 0.25)$ (left, gray), $(\theta_0, \sigma_0) = (\theta_1, \sigma_1) = (29.9, 0.5)$ (left, black), $(\theta_0, \sigma_0) = (\theta_1, \sigma_1) = (5, 0.25)$ (right, gray) and $(\theta_0, \sigma_0) = (\theta_1, \sigma_1) = (5, 0.5)$ (right, black).



(iii) HDPYP with $\theta_0 = 30$, $(\theta_1, \sigma_1) = (30, 0.25)$ (left, gray), $\theta_0 = 30$, $(\theta_1, \sigma_1) = (30, 0.5)$ (left, black), $\theta_0 = 5$, $(\theta_1, \sigma_1) = (5, 0.25)$ (right, gray) and $\theta_0 = 5$, $(\theta_1, \sigma_1) = (5, 0.5)$ (right, black).



(iv) HPYDP with $(\theta_0, \sigma_0) = (30, 0.25)$, $\theta_1 = 30$ (left, gray), $(\theta_0, \sigma_0) = (30, 0.5)$, $\theta_1 = 30$ (left, black), $(\theta_0, \sigma_0) = (5, 0.25)$, $\theta_1 = 5$ (right, gray) and $(\theta_0, \sigma_0) = (5, 0.5)$, $\theta_1 = 5$ (right, black).

Figure 2: Exact (dashed lines) and asymptotic (solid lines) expected marginal number of clusters $E(D_{i,t})$ when $n_i(t) = 1, \ldots, 500$ for different HSSMs.

Favaro and Teh (2013) for an up-to-date review. In this section, we extend the marginal sampler for HDP mixture (see Teh et al. (2006), Teh (2006) and Teh and Jordan (2010)), to our general class of HSSMs. We present the sampler for the case kernel and base measure are conjugate. When this assumption is not satisfied our sampling method can be easily modified following the auxiliary variable sampler of Neal (2000) and Favaro and Teh (2013).

Following the notation in Section 3.2, we consider the data structure

$$\begin{aligned} Y_{i,j}, c_{i,j} : i \in \mathcal{J}, \text{ and } j = 1, \ldots, n_{i\cdot\cdot}, \\ d_{i,c} : i \in \mathcal{J}, \text{ and } c = 1, \ldots, m_{i\cdot}, \\ \phi_d : d \in \mathcal{D}, \end{aligned}$$

where $Y_{i,j}$ is the j-th observation in the i-th group, $n_{i\cdot\cdot} = n_i$ is the total number of observations in the i-th group, and $\mathcal{J} = \{1, \ldots, I\}$ is the set of group indexes. The latent variable $c_{i,j}$ denotes the table at which the j-th "customer" of "restaurant" i sits and $d_{i,c}$ the index of the "dish" served at table c in restaurant i. The random variables ϕ_d are the "dishes" and $\mathcal{D} = \{d : d = d_{i,c} \text{ for some } i \in \mathcal{J} \text{ and } c \in \{1, \ldots, m_{i\cdot}\}\}$ is the set of indexes of the served dishes.

Let us assume that the distribution H of the atoms ϕ_ds has density h and the observations $Y_{i,j}$ have a kernel density $f(\cdot|\cdot)$, then our hierarchical infinite mixture model is

$$Y_{i,j}|\boldsymbol{\phi}, \boldsymbol{c}, \boldsymbol{d} \overset{ind}{\sim} f\left(\cdot|\phi_{d_{i,c_{i,j}}}\right), \quad \boldsymbol{\phi}|\boldsymbol{c}, \boldsymbol{d} \overset{i.i.d.}{\sim} h(\cdot), \quad [\boldsymbol{c}, \boldsymbol{d}] \sim HSSM,$$

where

$$\begin{aligned} \boldsymbol{c} &= [\boldsymbol{c}_i : i \in \mathcal{J}], \quad \text{with } \boldsymbol{c}_i = [c_{i,j} : j = 1, \ldots, n_{i\cdot\cdot}], \\ \boldsymbol{d} &= [d_{i,c} : i \in \mathcal{J}, c = 1, \ldots, m_{i\cdot}], \quad \boldsymbol{\phi} = [\phi_d : d \in \mathcal{D}], \end{aligned}$$

and, with a slight abuse of notation, we write $[\boldsymbol{c}, \boldsymbol{d}] \sim HSSM$ in order to denote the distribution of the labels $[\boldsymbol{c}, \boldsymbol{d}]$ obtained from a $HSSM$ as in (3.7). If we define

$$d^*_{i,j} = d_{i,c_{i,j}} \quad \text{and} \quad \boldsymbol{d}^* = [d^*_{i,j} : i \in \mathcal{J}, j = 1, \ldots, n_{i\cdot\cdot}],$$

then $[\boldsymbol{c}, \boldsymbol{d}]$ and $[\boldsymbol{c}, \boldsymbol{d}^*]$ contain the same amount of information, indeed $\boldsymbol{d}^*$ is a function of $\boldsymbol{d}$ and $\boldsymbol{c}$, while $\boldsymbol{d}$ is a function of $\boldsymbol{d}^*$ and $\boldsymbol{c}$. From now on, we denote with $\boldsymbol{Y} = [Y_{i,j} : i \in \mathcal{J}, j = 1, \ldots, n_{i\cdot\cdot}]$ the set of observations.

If f and H are conjugate, the *Chinese Restaurant Franchise Sampler* of Teh et al. (2006) can be generalized and a new sampler can be obtained for our class of models.

Denote with the superscript $^{\neg ij}$ the counts and sets in which the customer j in the restaurant i is removed and, analogously, with $^{\neg ic}$ the counts and sets in which all the customers in the table c of the restaurant i are removed. We denote with $p(X)$ the density of the random variable X.

The proposed Gibbs sampler simulates iteratively the elements of $\boldsymbol{c}$ and $\boldsymbol{d}$ from their full conditional distributions, where the latent variables ϕ_d are integrated out

analytically. In sampling the latent variable $\boldsymbol{c}$, we need to sample jointly $[\boldsymbol{c}, \boldsymbol{d}^*]$ and, since $\boldsymbol{d}$ is a function of $[\boldsymbol{c}, \boldsymbol{d}^*]$, this also gives a sample for $\boldsymbol{d}$. In order to improve the mixing we re-sample $\boldsymbol{d}$ given $\boldsymbol{c}$ in a second step. In summary, the sampler iterates for $i = 1, \ldots, I$ according to the following steps:

(i) sample $[c_{i,j}, d^*_{i,j}]$ from $p(c_{i,j}, d^*_{i,j} | \boldsymbol{Y}, \boldsymbol{c}^{\neg ij}, \boldsymbol{d}^{*\neg ij})$ (see Equation (S.32) in the supplementary material (Bassetti et al., 2019a)), for $j = 1, \ldots, n_{i\cdot\cdot}$;

(ii) (re)-sample $d_{i,c}$ from $p(d_{i,c} | \boldsymbol{Y}, \boldsymbol{c}, \boldsymbol{d}^{\neg ic})$ (see Equation (S.34) in the supplementary material (Bassetti et al., 2019a)), for $c = 1, \ldots, m_{i\cdot}$.

A detailed description of the Gibbs sampler is given in the supplementary material (Bassetti et al., 2019a).

6 Illustrations

6.1 Simulation Experiments

We compare some of the HSSMs described in Section 3 on synthetic data generated under different assumptions on the true model. In the first experimental setting, we consider three groups of observations sampled from three-component normal mixtures with common mixture components, but different mixture probabilities:

$$\begin{aligned} Y_{1j} &\overset{iid}{\sim} 0.3\mathcal{N}(-5,1) + 0.3\mathcal{N}(0,1) + 0.4\mathcal{N}(5,1), \ j = 1, \ldots, 100, \\ Y_{2j} &\overset{iid}{\sim} 0.3\mathcal{N}(-5,1) + 0.7\mathcal{N}(0,1), \ j = 1, \ldots, 50, \\ Y_{3j} &\overset{iid}{\sim} 0.8\mathcal{N}(-5,1) + 0.1\mathcal{N}(0,1) + 0.1\mathcal{N}(5,1), \ j = 1, \ldots, 50. \end{aligned}$$

The parameters of the different prior processes are chosen such that the marginal expected number of clusters is $\mathbb{E}(D_{i,t}) = 5$ and its variance is between 1.97 and 3.53 assuming $n_i(t) = n_i = 50$ with $t = 1$ for $i = 1, \ldots, 3$.

In the second and third experimental settings, we consider ten groups of observations from two- and three-component normal mixtures respectively with one common component across groups. In the second experiment, we assume

$$Y_{ij} \overset{iid}{\sim} 0.7\mathcal{N}(-5,1) + 0.3\mathcal{N}(-4+i,1), \ j = 1, \ldots, n_i(t)$$

with $n_i(t)$ increasing from 5 to 100 with $t = 1$ and for $i = 1, \ldots, 10$. In the third setting, we assume a smaller weight for the common component and larger number of group specific components:

$$Y_{ij} \overset{iid}{\sim} 0.2\mathcal{N}(-5,1) + 0.4\mathcal{N}(-6-i,1) + 0.4\mathcal{N}(-4+i,1), \ j = 1, \ldots, 20.$$

The parameters of the prior processes are chosen such that the marginal expected value is $\mathbb{E}(D_{i,t}) = 10$ and the variance is between 4.37 and 6.53 assuming $n_i(t) = 20$ with $t = 1$ for $i = 1, \ldots, 10$.

For each setting we generate 50 independent datasets and run the marginal sampler described in Section 5 with $6,000$ iterations to approximate the posterior predictive distribution and the posterior distribution of the clustering variables $\boldsymbol{c}$ and $\boldsymbol{d}$. We discard the first $1,000$ iterations of each run. All inferences are averaged over the 50 independent runs.

We compare the models by evaluating their co-clustering errors and predictive abilities (see Favaro and Teh (2013) and Dahl (2006)). We denote with $\tilde{\mathbf{d}}^{(m)} = (d^{(m)}_{1,c_{11}}, \ldots, d^{(m)}_{1,c_{1n_1}}, \ldots, d^{(m)}_{I,c_{I1}}, \ldots, d^{(m)}_{I,c_{In_I}})$, the vector of allocation variables for all the observations, sampled at the Gibbs iteration $m = 1, \ldots, M$, where M is the number of Gibbs iterations. The co-clustering matrix of posterior pairwise probabilities of joint classification is estimated by:

$$P_{lk} = \frac{1}{M} \sum_{m=1}^{M} \delta_{\left\{\tilde{d}^{(m)}_l\right\}} \left(\tilde{d}^{(m)}_k\right) \qquad l, k = 1, \ldots, n_{\cdots}.$$

Let $\tilde{\boldsymbol{d}}_0$ be the true value of the allocation vector $\tilde{\mathbf{d}}$. The co-clustering error can be measured as the average L_1 distance between the true pairwise co-clustering matrix, $\delta_{\{d_{0l}\}}(d_{0k})$ and the estimated co-clustering probability matrix, P_{lk}, i.e.:

$$CN = \frac{1}{n^2_{\cdots}} \sum_{l=1}^{n_{\cdots}} \sum_{k=1}^{n_{\cdots}} |\delta_{\{d_{0l}\}}(d_{0k}) - P_{lk}|. \tag{6.1}$$

The following alternative measure can be defined by using the Hamming norm and the estimated co-clustering matrix, $\mathbb{I}(P_{lk} > 0.5)$:

$$CN^* = \frac{1}{n^2_{\cdots}} \sum_{l=1}^{n_{\cdots}} \sum_{k=1}^{n_{\cdots}} |\delta_{\{d_{0l}\}}(d_{0k}) - \mathbb{I}(P_{lk} > 0.5)|. \tag{6.2}$$

Both accuracy measures CN and CN^* attain 0 in absence of co-clustering error and 1 when co-clustering is mispredicted.

The L_1 distance between the true group-specific densities, $f(Y_{i,n_i+1})$ and the corresponding posterior predictive densities, $p(Y_{i,n_i+1}|\mathbf{Y})$, can be used to define the predictive score:

$$SC = \frac{1}{I} \sum_{i=1}^{I} \int |f(Y_{i,n_i+1}) - p(Y_{i,n_i+1}|\mathbf{Y})| \, dY_{i,n_i+1}.$$

Finally, we consider the posterior median ($\widehat{q_{0.5}(D)}$) and variance ($\widehat{V(D)}$) of the total number of clusters D.

The results in Table 1 point out similar co-clustering accuracy across HSSMs and experiments. In the first and second experimental settings, HPYP and HDPYP have significantly small co-clustering errors, CN and CN^*. As regard the predictive score SC, the seven HSSMs behave similarly in the three restaurants experiment (panel a), whereas in the two-components experiment the HDPYP performs slightly better with respect to

	HDP	HPYP	HGP	HDPYP	HPYDP	HGDP	HGPYP
(a) Three Restaurants – 3-component normal mixtures							
CN	0.0975	0.0829	0.1220	0.0668	0.0888	0.1018	0.0982
CN^*	0.0073	0.0056	0.0311	0.0053	0.0057	0.0079	0.0070
SC	0.5732	0.5556	0.6121	0.5368	0.5651	0.5872	0.5917
$\widehat{q_{0.5}(D)}$	7	7	6.7	5	6.96	6.04	6
$\widehat{V(D)}$	3.3365	4.5520	2.5166	2.1800	4.2211	2.3580	2.3509
(b) Ten Restaurants – 2-component normal mixtures							
CN	0.3120	0.2570	0.4115	0.2674	0.2825	0.4003	0.3967
CN^*	0.1870	0.1598	0.5558	0.1568	0.1617	0.5508	0.5462
SC	2.2666	2.2217	2.2657	2.1186	2.1612	2.2855	2.3054
$\widehat{q_{0.5}(D)}$	19.14	19	14.98	15.34	22	14.14	14.04
$\widehat{V(D)}$	9.3477	11.8293	6.4858	6.5009	13.2239	6.0336	5.8835
(c) Ten Restaurants – 3-component normal mixtures							
CN	0.3111	0.3124	0.3125	0.3048	0.3175	0.2977	0.2977
CN^*	0.3141	0.3152	0.3261	0.3009	0.3280	0.2921	0.2918
SC	5.2192	4.9978	5.3573	4.8874	4.6848	4.5712	4.6250
$\widehat{q_{0.5}(D)}$	24	26	22.12	20.24	29	15	15
$\widehat{V(D)}$	9.2830	13.5746	12.2650	6.9006	14.2011	5.2881	4.9160

Table 1: Model accuracy for seven HSSMs in three experimental settings (panel (a), (b) and (c)) using different measures: co-clustering norm (CN), threshold co-clustering norm (CN^*), predictive score (SC), posterior median ($\widehat{q_{0.5}(D)}$) and variance ($\widehat{V(D)}$) of the number of clusters. The accuracy has been estimated with 50 independent Markov Chain Monte Carlo (MCMC) experiments. Each experiment consists in 6000 MCMC iterations.

the other HSSMs. In presence of large heterogeneity across restaurants (third setting), the HGPYP is performing best following the co-clustering norm and the predictive score measures. A comparison between HPYP and HGPYP shows that these results do not depend on the number of observations and can be explained by a better fitting of tails and dispersion of the group-specific densities provided by the HGPYP. For illustrative purposes, we provide in Figure 3 a comparison of the log-predictive scores of the two models for an increasing number of observations.

In the first setting, the posterior number of clusters, $\widehat{q_{0.5}(D)}$, for all the HSSMs (panel (a) in Table 1) is significantly close to the true value, that corresponds to 3 mixture components. Increasing the number of restaurants (second and third settings), the HPYP tends to have extra clusters causing larger posterior median and variance of the number of clusters ($\widehat{q_{0.5}(D)}$ and $\widehat{V(D)}$ in Table 1). Conversely, the HGPYP have a smaller dispersion of the number of clusters with respect to the HPYP.

The results for the third experiment suggest that HGPYP performs better when groups of observations are heterogeneous. Also increasing the number of observations, HGPYP provides a consistent estimate of the true number of components (Figure 3).

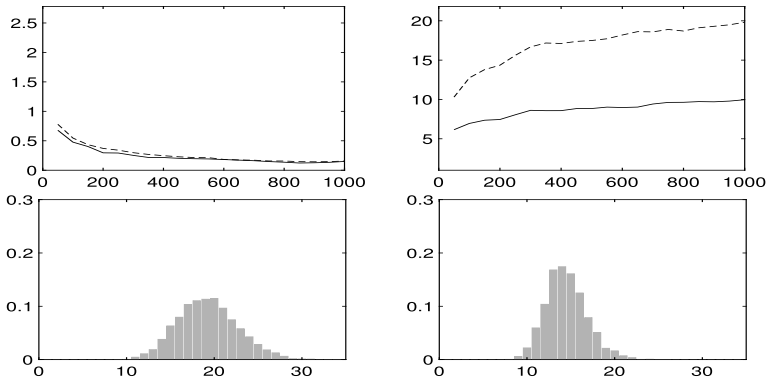


Figure 3: Top-left: Log-posterior predictive score for the right tail (above the 97.5% quantile of the true distribution). Top-right: posterior mean when the number of customers increases for HGPYP (solid) and HPYP (dashed). Bottom: posterior number of clusters for the HPYP (left) and HGPYP (right). In this setting the true number of clusters is 11.

In conclusion, our experiments indicate that using the Pitman-Yor process at some stage of the hierarchy may lead to a better accuracy. The HDPYP did reasonably well in all our experiments in line with previous findings on hierarchical Dirichlet and Pitman-Yor processes for topic models (see Du et al. (2010)). Also, using Gnedin process at the top of the hierarchy might lead to a better accuracy when groups of observations are heterogeneous. Moreover, when the researcher is interested in a consistent estimate of the number of components, HGPYP should be preferred. Further details and results are in the supplementary material (Bassetti et al., 2019b).

6.2 Real Data Application

Bayesian nonparametrics is used in economic time series modelling to capture observation clustering effects (e.g., see Hirano, 2002; Griffin and Steel, 2011; Bassetti et al., 2014; Kalli and Griffin, 2018; Billio et al., 2019). In this paper, we consider the industrial production index, an important indicator of macroeconomic activity used in business cycle analysis (see Stock and Watson (2002)). One of the most relevant issues in this field concerns the classification of observations by allowing for different parameter values in periods (called regimes) of recession and expansion.

The data has been previously analysed by Bassetti et al. (2014) and contains the seasonally and working day adjusted industrial production indexes (IPI) at a monthly frequency from April 1971 to January 2011 for both United States (US) and European Union (EU). We generate autoregressive-filtered IPI quarterly growth rates by calculating the residuals of a vector autoregressive model of order 4.

We follow a Bayesian nonparametric approach based on HSSM prior for the estimation of the number of regimes or structural breaks. Based on the simulation results, we

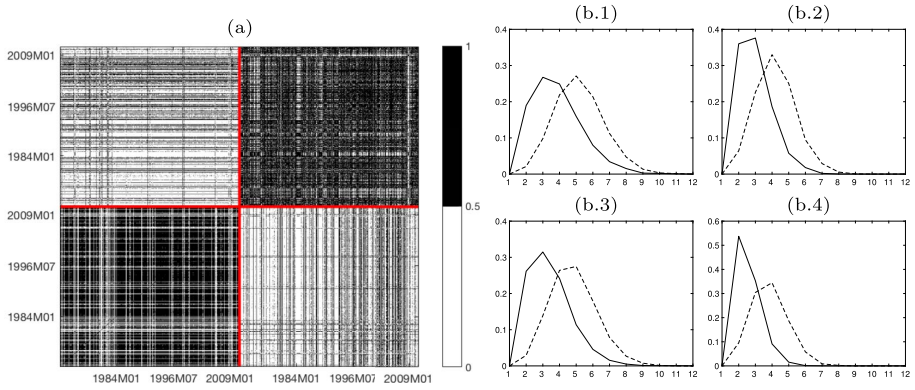


Figure 4: (a) Co-clustering matrix for the US (bottom left block) and EU (top right block) business cycles and cross-co-clustering (main diagonal blocks) between US and EU for the HPYP. (b) Posterior number of clusters. Total (b.1), marginal for US (b.2) and EU (b.3) and common (b.4) for the HPYP (solid line) and for the HGPYP (dashed line).

focus on the HPYP, with hyperparameters $(\theta_0, \sigma_0) = (1.2, 0.2)$ and $(\theta_1, \sigma_1) = (2, 0.2)$, and on the HGPYP, with hyperparameters $(\gamma_0, \zeta_0) = (14.7, 130)$ and $(\theta_1, \sigma_1) = (2, 0.23)$, such that the prior mean of the number of clusters is 5. The main results of the nonparametric inference can be summarized through the implied data clustering (panel (a) of Figure 4) and the marginal, total and common posterior number of clusters (panel (b)).

One of the most striking feature of the co-clustering is that in the first and second block of the minor diagonal there are vertical and horizontal black lines. They correspond to observations of a country, which belong to the same cluster that is the same phase of the business cycle.

Another feature that motivates the use of HSSMs is given by the black horizontal and vertical lines in the two main diagonal blocks. They correspond to observations of the two countries allocated to common clusters. The appearance of the posterior total number of clusters (see panel b.1) suggests that at least three clusters should be used in a joint modelling of the US and EU business cycle. The larger dispersion of the marginal number of cluster for EU (b.3) with respect to US (b.2) confirms the evidence in Bassetti et al. (2014) of a larger heterogeneity in the EU cycle. Finally, we found evidence (panel b.4) of common clusters of observations between EU and US business cycles.

Supplementary Material

Supplementary material A to Hierarchical Species Sampling Models
(DOI: 10.1214/19-BA1168SUPPA; .pdf). This document contains the derivations of the results of the paper and a detailed analysis of the generalized species sampling (with a general base measure). It also describes the Chinese Restaurant Franchise Sampler for Hierarchical Species Sampling Mixtures.

Supplementary material B to Hierarchical Species Sampling Models (DOI: 10.1214/19-BA1168SUPPB; .pdf). This document provides further numerical illustrations and robustness checks.

References

Argiento, R., Cremaschi, A., and Vannucci, M. (2019). "Hierarchical Normalized Completely Random Measures to Cluster Grouped Data." *Journal of the American Statistical Association*, 1–43. doi: https://doi.org/10.1080/01621459.2019.1594833. 2, 10

Arratia, R., Barbour, A. D., and S., T. (2003). *Logarithmic combinatorial structures: a probabilistic approach*. European Mathematical Society. MR2032426. doi: https://doi.org/10.4171/000. 3

Bacallado, S., Battiston, M., Favaro, S., and Trippa, L. (2017). "Sufficientness Postulates for Gibbs-Type Priors and Hierarchical Generalizations." *Statistical Science*, 32(4): 487–500. MR3730518. doi: https://doi.org/10.1214/17-STS619. 7

Bassetti, F., Casarin, R., and Leisen, F. (2014). "Beta-product dependent Pitman-Yor processes for Bayesian inference." *Journal of Econometrics*, 180(1): 49–72. MR3188911. doi: https://doi.org/10.1016/j.jeconom.2014.01.007. 24, 25

Bassetti, F., Casarin, R., Rossini, L. (2019a). "Supplementary Material A to Hierarchical Species Sampling Models." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1168SUPPA. 3, 6, 7, 8, 11, 16, 21

Bassetti, F., Casarin, R., Rossini, L. (2019b). "Supplementary Material B to Hierarchical Species Sampling Models." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1168SUPPB. 14, 24

Billio, M., Casarin, R., and Rossini, L. (2019). "Bayesian nonparametric sparse VAR models." *Journal of Econometrics*, 212: 97–115. URL http://www.sciencedirect.com/science/article/pii/S0304407619300776. MR3994009. doi: https://doi.org/10.1016/j.jeconom.2019.04.022. 24

Camerlenghi, F., Lijoi, A., Orbanz, P., and Pruenster, I. (2019). "Distribution theory for hierarchical processes." *Annals of Statistics*, 47(1): 67–92. MR3909927. doi: https://doi.org/10.1214/17-AOS1678. 2, 3, 7, 11, 12, 13, 15

Camerlenghi, F., Lijoi, A., and Prünster, I. (2017). "Bayesian prediction with multiple-samples information." *Journal of Multivariate Analysis*, 156: 18–28. URL http://www.sciencedirect.com/science/article/pii/S0047259X17300568. MR3624682. doi: https://doi.org/10.1016/j.jmva.2017.01.010. 2

Camerlenghi, F., Lijoi, A., and Prünster, I. (2018). "Bayesian nonparametric inference beyond the Gibbs-type framework." *Scandinavian Journal of Statistics*, 45(4): 1062–1091. MR3884900. doi: https://doi.org/10.1111/sjos.12334. 2

Canale, A., Lijoi, A., Nipoti, B., and Prünster, I. (2017). "On the Pitman–Yor pro-

cess with spike and slab base measure." *Biometrika*, 104(3): 681–697. MR3694590. doi: https://doi.org/10.1093/biomet/asx041. 3

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *Annals of Statistics*, 43(5): 1986–2018. URL https://projecteuclid.org:443/euclid.aos/1438606851. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 3

Dahl, D. B. (2006). "Model-based clustering for expression data via a Dirichlet process mixture model." In Do, K.-A., Müller, P. P., and Vannucci, M. (eds.), *Bayesian Inference for Gene Expression and Proteomics*, 201–218. Cambridge University Press. MR2706330. 22

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prunster, I., and Ruggiero, M. (2015). "Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?" *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(2): 212–229. 14

Diaconis, P. and Ram, A. (2012). "A probabilistic interpretation of the Macdonald polynomials." *Annals of Probability*, 40(5): 1861–1896. MR3025704. doi: https://doi.org/10.1214/11-AOP674. 3

Donnelly, P. (1986). "Partition structures, Pólya urns, the Ewens sampling formula, and the ages of alleles." *Theoretical Population Biology*, 30(2): 271–288. MR0865115. doi: https://doi.org/10.1016/0040-5809(86)90037-7. 3

Donnelly, P. and Grimmett, G. (1993). "On the asymptotic distribution of large prime factors." *Journal of the London Mathematical Society (2)*, 47(3): 395–404. MR1214904. doi: https://doi.org/10.1112/jlms/s2-47.3.395. 3

Du, L., Buntine, W., and Jin, H. (2010). "A segmented topic model based on the two-parameter Poisson-Dirichlet process." *Machine Learning*, 81(1): 5–19. MR3108170. doi: https://doi.org/10.1007/s10994-010-5197-4. 2, 3, 7, 24

Dubey, A., Williamson, S., and Xing, E. (2014). "Parallel Markov chain Monte Carlo for Pitman-Yor mixture models." In *Uncertainty in Artificial Intelligence – Proceedings of the 30th Conference, UAI 2014*, 142–151. 7

Escobar, M. (1994). "Estimating normal means with a Dirichlet process prior." *Journal of the American Statistical Association*, 89(425): 268–277. MR1266299. 18

Escobar, M. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 18

Ewens, W. J. (1972). "The sampling theory of selectively neutral alleles." *Theoretical Population Biology*, 3: 87–112; erratum, ibid. 3 (1972), 240; erratum, ibid. 3 (1972), 376. MR0325177. doi: https://doi.org/10.1016/0040-5809(72)90035-4. 3

Favaro, S. and Teh, Y. W. (2013). "MCMC for Normalized Random Measure Mixture Models." *Statistical Science*, 28(3): 335–359. MR3135536. doi: https://doi.org/10.1214/13-STS422. 20, 22

George, E. I. and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88(423): 881–889. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476353 3

Gnedin, A. (2010). "A species sampling model with finitely many types." *Electronic Communications in Probability*, 15(8): 79–88. MR2606505. doi: https://doi.org/10.1214/ECP.v15-1532. 2, 14, 16

Gnedin, A. and Pitman, J. (2006). "Exchangeable Gibbs partitions and Stirling triangles." *Journal of Mathematical Sciences*, 138(3): 5674–5685. MR2160320. doi: https://doi.org/10.1007/s10958-006-0335-z. 4, 14

Griffin, J. E. and Steel, M. F. J. (2011). "Stick-breaking autoregressive processes." *Journal of Econometrics*, 162(2): 383–396. MR2795625. doi: https://doi.org/10.1016/j.jeconom.2011.03.001. 24

Hirano, K. (2002). "Semiparametric Bayesian Inference in autoregressive panel data models." *Econometrica*, 70(2): 781–799. MR1913831. doi: https://doi.org/10.1111/1468-0262.00305. 24

Hjort, N. L., Homes, C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press. MR2722988. doi: https://doi.org/10.1017/CBO9780511802478.002. 3, 18

Hoppe, F. M. (1984). "Pólya-like urns and the Ewens' sampling formula." *Journal of Mathematical Biology*, 20(1): 91–94. MR0758915. doi: https://doi.org/10.1007/BF00275863. 3

Kallenberg, O. (2006). *Probabilistic Symmetries and Invariance Principles*. Springer-Verlag New York. MR2161313. 6

Kalli, M. and Griffin, J. E. (2018). "Bayesian nonparametric vector autoregressive models." *Journal of Econometrics*, 203(2): 267–282. URL http://www.sciencedirect.com/science/article/pii/S0304407617302415. MR3770826. doi: https://doi.org/10.1016/j.jeconom.2017.11.009. 24

Kalli, M., Griffin, J. E., and Walker, S. (2011). "Slice sampling mixture models." *Statistics and Computing*, 21(1): 93–105. MR2746606. doi: https://doi.org/10.1007/s11222-009-9150-y. 18

Kim, S., Dahl, D. B., and Vannucci, M. (2009). "Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models." *Bayesian Analysis*, 4(4): 707–732. MR2570085. doi: https://doi.org/10.1214/09-BA426. 3

Kingman, J. F. C. (1980). *Mathematics of genetic diversity*, volume 34 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa. MR0591166. 3

Lau, J. W. and Green, P. J. (2007). "Bayesian Model-Based Clustering Procedures." *Journal of Computational and Graphical Statistics*, 16(3): 526–558. MR2351079. doi: https://doi.org/10.1198/106186007X238855. 3

Lim, K. W., Buntine, W., Chen, C., and Du, L. (2016). "Nonparametric Bayesian topic modelling with the hierarchical Pitman-Yor processes." *International Journal of Approximate Reasoning*, 78(C): 172–191. MR3543880. doi: https://doi.org/10.1016/j.ijar.2016.07.007. 2, 7

Miller, J. and Harrison, M. (2018). "Mixture models with a Prior on the number of components." *Journal of the American Statistical Association*, 113(521): 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 2, 8

Müller, P. and Quintana, F. (2010). "Random partition models with regression on covariates." *Journal of Statistical Planning and Inference*, 140(10): 2801–2808. MR2651966. doi: https://doi.org/10.1016/j.jspi.2010.03.002. 3

Navarro, D. J., Griffiths, T. L., Steyvers, M., and Lee, M. D. (2006). "Modeling individual differences using Dirichlet processes." *Journal of Mathematical Psychology*, 50(2): 101–122. MR2215141. doi: https://doi.org/10.1016/j.jmp.2005.11.006. 3

Neal, R. (2000). "Markov Chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 20

Nguyen, X. (2016). "Borrowing strengh in hierarchical Bayes: Posterior concentration of the Dirichlet base measure." *Bernoulli*, 22(3): 1535–1571. MR3474825. doi: https://doi.org/10.3150/15-BEJ703. 13

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models." *Biometrika*, 95(1): 169–186. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 18

Pitman, J. (1995). "Exchangeable and partially exchangeable random partitions." *Probability Theory and Related Fields*, 102(2): 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 7, 13

Pitman, J. (1996). "Some developments of the Blackwell-MacQueen urn scheme." In *Statistics, probability and game theory*, volume 30 of *IMS Lecture Notes—Monograph Series*, 245–267. Institute of Mathematical Statistics, Hayward, CA. MR1481784. doi: https://doi.org/10.1214/lnms/1215453576. 5

Pitman, J. (2003). "Poisson-Kingman partitions." In *Statistics and science: a Festschrift for Terry Speed*, volume 40 of *IMS Lecture Notes—Monograph Series*, 1–34. Institute of Mathematical Statistics, Beachwood, OH. MR2004330. doi: https://doi.org/10.1214/lnms/1215091133. 14

Pitman, J. (2006). *Combinatorial Stochastic Processes*, volume 1875. Springer-Verlag. MR2245368. 3, 4, 5, 12, 14, 15, 17

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 25(2): 855–900. MR1434129. doi: https://doi.org/10.1214/aop/1024404422. 7

Rockova, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the*

American Statistical Association, 113(521): 431–444. MR3803476. doi: https://doi.org/10.1080/01621459.2016.1260469. 3

Sangalli, L. M. (2006). "Some developments of the normalized random measures with independent increments." *Sankhyā*, 68(3): 461–487. MR2322195. 6

Sohn, K.-A. and Xing, E. P. (2009). "A hierarchical Dirichlet process mixture model for haplotype reconstruction from multi-population data." *The Annals of Applied Statistics*, 3(2): 791–821. MR2750682. doi: https://doi.org/10.1214/08-AOAS225. 13

Stock, J. H. and Watson, M. W. (2002). "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association*, 97(460): 1167–1179. MR1951271. doi: https://doi.org/10.1198/016214502388618960. 24

Teh, Y. and Jordan, M. I. (2010). "Hierarchical Bayesian nonparametric models with applications." In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. (eds.), *Bayesian Nonparametrics*. Cambridge University Press. MR2730663. 2, 20

Teh, Y. W. (2006). "A Hierarchical Bayesian Language Model Based on Pitman-Yor Processes." In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, 985–992. Stroudsburg, PA, USA: Association for Computational Linguistics. doi: https://doi.org/10.3115/1220175.1220299. 2, 3, 7, 20

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet processes." *Journal of the American Statistical Association*, 101(476): 1566–1581. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 1, 7, 20

Walker, S. G. (2007). "Sampling the Dirichlet Mixture Model with Slices." *Communications in Statistics – Simulation and Computation*, 36(1): 45–54. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 18

Wood, F., Archambeau, C., Gasthaus, J., James, L. F., and Teh, Y. W. (2009). "A Stochastic Memoizer for Sequence Data." In *International Conference on Machine Learning (ICML)*, volume 26, 1129–1136. 3

Acknowledgments

We thank for their useful comments on a previous version an Associate Editor, two Referees, Federico Camerlenghi and Antonio Lijoi and the participants of the Italian-French Statistics Workshop 2017, Venice. This research used the SCSCF multiprocessor cluster system at University Ca' Foscari of Venice. This paper is part of the research activities at the Venice Center in Economic and Risk Analytics for public policies (VERA) at Ca' Foscari University of Venice.