PhD THESIS DECLARATION

The undersigned

SURNAME *Battiston*

FIRST NAME *Marco*

PhD Registration Number *1422294*

Thesis title: *Gibbs-type priors for species sampling problems and feature models*

PhD in *Statistics*

Cycle *XXVIII*

Candidate's tutor *Professor Stefano Favaro*

Year of thesis defence *2017*

DECLARES

Under *his/her* responsibility:

1) that, according to Italian Republic Presidential Decree no. 445, 28th December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30th April 1999, to keep a copy of the thesis on deposit at the "Bilioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary embargo protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

3) that the Bocconi Library will file the thesis in its "Archivio istituzionale ad accesso aperto" (institutional registry) which permits online consultation of the complete text (except in cases of a temporary embargo);

4) that, in order to file the thesis at the Bocconi Library, the University requires that the thesis be submitted online by the candidate in unalterable format to Società NORMADEC (acting on behalf of the University), and that NORMADEC will indicate in each footnote the following information:

   - thesis *Gibbs-type priors for species sampling problems and feature models*;
   - by *Battiston Marco*;
   - defended at Università Commerciale "Luigi Bocconi" – Milano in *2017*;
   - the thesis is protected by the regulations governing copyright (Italian law no. 633, 22th April 1941 and subsequent modifications). The exception is the right of Università Commerciale "Luigi Bocconi" to reproduce the same for research and teaching purposes, quoting the source;

5) that the copy of the thesis submitted online to NORMADEC is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal, and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results, and is not subject to patent or protection registrations, and therefore not subject to an embargo;

Date *29/08/2016*

SURNAME *Battiston*
FIRST NAME *Marco*

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors Stefano Favaro and Yee Whye Teh for their help and support during the last two years spent working on this thesis.

I would also like to thank all people I had the chance to meet and discuss with in the last four years of my PhD. In particular, I am grateful to all professors and PhD students at Bocconi University, with whom I shared the first two years when I was attending courses. I am also indebted to the group of statisticians at Collegio Carlo Alberto, who I met during the third year. Finally, I want to thank all students, postdocs and researchers I met in Oxford in this last year of my PhD.

Abstract

Gibbs-type partitions are a wide class of distributions for random partitions introduced in Gnedin and Pitman [48]. They have become popular also in the Bayesian nonparametrics literature in the last few years following Lijoi et al. [78]. Within a Bayesian approach, Gibbs-type partitions can be used to model the random partitioning of observations in clustering and species sampling problems. The scope of this thesis is to provide both theoretical and methodological contributions on Gibbs-type partitions, with a particular focus on their application to species sampling problems. Under the usual setting of a species sampling model, in Chapter 3 we derive a characterization of Gibbs-type priors through conditions on the predictive distributions of the observables. This kind of characterizations are usually referred to as Johnson 'sufficientness' postulates and are helpful for prior elicitation. In the same chapter, we supply a Polya-type urn scheme representation for all Gibbs-type priors. Urn schemes provide an intuitive description of the data generating process and they are useful to build MCMC algorithm for posterior inference. In Chapter 4, we present a new methodology based on Gibbs-type priors for sample-size selection in species sampling models. The goal of this methodology is to determine the minimal size of additional samples needed to observe a prefixed amount a new species. This result is useful when sampling further observations is expensive and it is important to design cost-effective species inventories. In Chapter 5, we consider another problem of species discovery but now in a sequential and partially exchangeable setting. In this scenario, there are multiple populations available and we can choose at every time step from which population to collect further samples. We propose a sequential strategy with the goal of discovering as many species as possible. This strategy works for both abundance and incidence data and its performances are tested through simulations and in a real data example. Finally, in Chapter 6, we move from partitions to feature allocations, in which the assumptions of exhaustivity and disjointness of the blocks of a partition are relaxed. We study the class of Gibbs-type feature allocations, characterizing all its elements as mixtures of Indian Buffet Processes and Beta-Bernoulli models.

Chapter 1

Overview of the Thesis

Bayesian Nonparametric Statistics (BNP) is the branch of Bayesian statistics dealing with infinite dimensional parameters or problems in which the dimension of the unknown parameter increases with the sample size. Examples of infinite dimensional parameters are cumulative distribution functions, density functions, hazard and cumulative hazard functions, regression functions and Markov transition kernels. The simplest and most natural nonparametric problem is that with the unknown parameter P being the marginal distribution of observations sampled from an infinite exchangeable sequence. Within this framework, by imposing mild assumptions on the predictive distributions of the observations we recover the so called *species sampling models*, which will be reviewed in **Chapter 2** together with all preliminaries on BNP needed in the following chapters. The classical and simplest example of prior in a species sampling model is the *Dirichlet process*, introduced in the seminal work by Ferguson [40]. Following this pioneering contribution, many extensions of this process have been proposed in the literature, either trying to add flexibility or to accommodate some features of particular datasets in applied works. As recently discussed in De Blasi et al. [27], a class of priors which can be considered the most natural generalization of the Dirichlet process are the so called *Gibbs-type priors*, proposed in Gnedin and Pitman [48] and reviewed in some depth in Chapter 2. This class includes the majority of well-known priors for discrete distributions, like the Dirichlet and the Pitman–Yor processes. At the same time, it remains mathematically tractable. In this thesis, we will provide both theoretical and methodological results on Gibbs-type priors, with a particular focus on their usage within species sampling models.

In **Chapter 3**, we start studying theoretical properties and providing motivation for the usage of Gibbs-type priors within species sampling models. In the Bayesian approach, one of the main issue is the specification of the prior. This task is even more challenging in nonparametric contexts, where the parameter space is usually a function space. However, it is known from the De Finetti's representation theorem that the law of an infinite exchangeable sequence of observations is uniquely characterized by the prior dis-

tribution, and also the converse holds true. Hence, at least in principle, the statistician can add prior information through assumptions on the joint or predictive distributions of the observables, rather than directly on the prior. An early discussion of such a problem, although in a parametric framework, dates back to the seminal work by English philosopher W. E. Johnson, who introduced a noteworthy characterization for the predictive probabilities of the symmetric Dirichlet prior distribution. This is typically referred to as the Johnson's "sufficientness" postulate. An extension of this postulate was carried on to the nonparametric framework by Zabell, providing predictive conditions to characterize the Dirichlet, Zabell [125], and the Pitman–Yor processes, Zabell [129]. In Chapter 3, we further extend these results and we provide a predictive characterization of the Gibbs-type priors. Specifically, we will show that, given a sample $\mathbf{X}_{n,k} = (X_1, \ldots, X_n)$ from an infinite exchangeable sequence of random variables, displaying K_n distinct values in it, the following two predictive conditions imply a Gibbs-type prior:

1. the probability that the next observation X_{n+1} is different from those already observed in the sample is a function only of n and K_n;

2. the probability that X_{n+1} is equal to an already observed value depends only on n, K_n and the number of times this value has been previously observed.

As particular cases, we recover the two characterization theorems of Zabell. Indeed, we prove that the Pitman–Yor process is recovered when we drop the dependence on K_n in the probability of observing an old species, while the Dirichlet process is found when we also assume independence of K_n in the probability of discovering a new species. Finally, we provide a urn scheme representation for Gibbs-type priors, which generalizes the one and two parameters Hoppe urns, Hoppe [62].

Chapter 4 is devoted to a statistical application of Gibbs-type priors in species sampling models. Within this class of models, we study the issue of sample size selection, when the goal is discovering new species. Indeed, a difficult and important challenge in species sampling problems is the design of cost-effective species inventories. Such a design necessarily requires to measure the sampling effort in order to ensure an efficient allocation of time and financial resources along the inventory process. In Chapter 4, we introduce a collection of Bayesian nonparametric tools for estimating the minimum number of additional samples that are required to detect a preset amount of new species. Specifically, we consider a sample $\mathbf{X}_n = (X_1, \ldots, X_n)$, from an infinite exchangeable sequence of random variables $(X_i)_{i\in\mathbb{N}}$. From De Finetti's representation theorem, the joint law of the sample can be written in the following hierarchical form

$$\begin{aligned} X_i | P &\overset{iid}{\sim} P \quad i = 1, \ldots, n \\ P &\sim \mu, \end{aligned} \tag{1.1}$$

where P denotes the unknown probability law of X_i and μ is the de Finetti's measure governing the exchangeable sequence $(X_i)_{i\in\mathbb{N}}$, which, from a Bayesian point of view, is interpreted as the prior distribution of P. We choose μ to be a Gybbs-type prior and we compute the posterior point estimate of $T_\tau|\mathbf{X}_n = \min\{m \in \mathbb{N} : K_{n+m} - K_n = \tau\}$, i.e., the minimum number of additional observations needed to observed τ new species, and of similar quantities. This summary provides an effective and easily interpretable estimate of the additional cost that we have to incur in order to observe τ further species. The methodology is illustrated through the analysis of an Expressed Sequence Tags dataset in genomics.

In **Chapter 5**, we consider another problem of species discoveries, but now in a context of partial exchangeability. A samples $\mathbf{X_n} = (\mathbf{X}_{1,n_1}, \ldots, \mathbf{X}_{J,n_J})$ collected from J distinct populations are available, where $\mathbf{X}_{j,n_j} = (X_{j,1}, \ldots, X_{j,n_j})$ denotes the sample from the j-th population. We can think of these as samples of animals from different geographical regions. A priori, it is unknown which species of animals are present in each region and what are their frequencies. Species are shared among populations and each species can be present in more than one region with its frequency varying across populations. We assume the observations to be partially exchangeable. From De Finetti's representation theorem for partially exchangeable random variables, the joint law of the observables can be written in the following hierarchical form

$$\begin{aligned} X_{j,i}|P_j &\overset{iid}{\sim} P_j \qquad i = 1, \ldots, n_j,\ j = 1, \ldots, J, \\ (P_1, \ldots, P_J) &\sim \mu. \end{aligned} \tag{1.2}$$

The problem explored in Chapter 5 is how to sequentially sample these populations, in order to observe the greatest number of different species. This problem is relevant, for instance, in applications in genomics, ecology and biology. We adopt a Bayesian non-parametric approach and endow $(P_1, \ldots, P_J)$ with a Hierarchical Pitman–Yor prior. As a consequence of the hierarchical structure, the J unknown discrete probability measures share the same random support, that of their common random base measure. Given this prior choice, we propose a sequential rule that, at every time step, given the information available up to that point, selects the population from which to collect the next observation. Rather than simply picking the population with the highest posterior estimate of producing a new value, the proposed rule includes a Thompson sampling step to better balance the exploration-exploitation trade-off, inherent in any bandit problem. In addition, we propose an extension of the algorithm to deal with incidence data, where animals are sampled in groups. Performances of both algorithms are assessed through simulations and compared to three other strategies. Finally, we compare these algorithms using a dataset of species of trees, collected from different plots in South America.

In **Chapter 6** we consider a different kind of models, called feature models. This class of models is very popular and has received lot of attention in the last ten years,

particularly in the machine learning community. In these models, each observation X_i is a random vector of unknown, but finite, cardinality. Its values are called features. Any two observations X_i and X_j can have some of their features in common. Feature models are generalizations of species sampling and clustering problems. Indeed, these latter models are the particular case of feature models in which each X_i has only one feature with probability one. The most popular Bayesian model for feature models is the Indian Buffet Process (IBP), derived in Griffiths and Ghahramani [55] as a limit of a Beta–Bernoulli model. In Chapter 2, we recall this model, together with the notions of exchangeable feature allocations and of exchangeable feature allocation function. In Chapter 6 we proceed to characterize the class of exchangeable feature allocations assigning probability $V_{n,k}\prod_{l=1}^{k} W_{m_l}U_{n-m_l}$ to a feature allocation of n observations, displaying totally k distinct features and with $(m_1,\ldots,m_k)$ counts for each of these features. Each element of this class is parametrized by a countable matrix V and two sequences U and W of non-negative weights. Moreover, a consistency condition is imposed to guarantee that the distribution for feature allocations of $n-1$ individuals is recovered from that for n individuals, when the last individual is integrated out. In Theorem 6.1.1, we prove that the only members of this class satisfying the consistency condition are mixtures of the Indian Buffet Process over its γ parameter and mixtures of the Beta–Bernoulli model over its N parameter. Hence, we provide a characterization of these two models as the only, up to randomization of the parameters, consistent exchangeable feature allocations having the required Gibbs-type product form.

To sum up, the thesis is organized as follows. In Chapter 2, we collect some known preliminaries on BNP. In Chapter 3, we provide a predictive characterization and a urn scheme for Gibbs-type priors. In Chapter 4, we use these priors to solve the problem of sample size selection for discovering new species. In Chapter 5, we continue with the species discovery problem, but now in a both partially exchangeable and sequential context. We propose two sequential algorithms to deal with both abundance and incidence data. Finally, in Chapter 6, we move from partitions to feature models and we characterize the IBP process and the Beta–Bernoulli model as the only, up to randomization, feature allocations having a Gibb-type-like structure. Results in Chapter 2 are known, while the contents of chapters from 3 to 6 are novel.

Chapter 2

Preliminaries on Bayesian Nonparametrics

In this chapter, we collect some preliminaries on Bayesian Nonparametrics. The reader interested in a wider and more detailed treatment on BNP is referred to the books by Ghosh and Ramamoorthi [47], Hjort et al. [61] and Phadia [96]. In section 2.1, we briefly expose some basic notions, like exchangeability, partial exchageability and their De Finetti's representation theorems. In section 2.2, we introduce the Dirichlet and Pitman-Yor processes. In section 2.3, we present species sampling models. Section 2.4 deals with Gibbs-type priors. Section 2.5 describes the Hierarchical Pitman-Yor prior, which will be used in Chapter 5. Finally, in section 2.6 we review feature models, with two separated subsections dedicated to the Beta–Bernoulli model and the Indian Buffet Process.

2.1 Exchangeability, Partial Exchangeability and De Finetti's representation theorems

Let $\mathbf{X}_\infty = (X_i)_{i\in\mathbb{N}}$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, representing the ideally infinite sequence of observations. A probabilistic assumption that is common in many Bayesian applications is *exchangeability*. This assumption means that the law of $\mathbf{X}_\infty$ is invariant with respect to the group of finite permutations of the index set, i.e., $(X_1, \ldots, X_n)$ is equal in distribution to $(X_{\varrho(1)}, \ldots, X_{\varrho(n)})$ for all $n \in \mathbb{N}$ and for all permutations ϱ of the first n integers. From a statistical point of view, exchangeability corresponds to the assumption that the order in which observations arrive is irrelevant for making inference on their law. This assumption of order invariance is fundamental due to de Finetti's representation theorem, which, under mild assumptions on the spate space of $\mathbf{X}_\infty$, characterizes infinite exchangeable sequences as mixtures of independent and identically distributed (i.i.d.) sequences. We denote by $(\mathbb{X}, \mathcal{X})$ the state

space of each coordinate X_i and the regularity conditions needed in the theorem is that this space is a complete and separable metric space endowed with its Borel sigma algebra. Under this mild assumption, we can state the celebrated De Finetti's theorem,

Theorem 2.1.1 (De Finetti's representation theorem for exchangeable sequences)**.** *$\boldsymbol{X}_\infty$ is exchangeable if and only if (iff) there exists a probability measure μ such that for all $n \in \mathbb{N}$ for all $A = A_1 \times \ldots \times A_n \times \mathbb{X} \times \ldots$, with $A_i \in \mathcal{X}$ for all $i \leq n$,*

$$\mathbb{P}(\boldsymbol{X}_\infty \in A) = \int_{\mathcal{P}} \prod_{i=1}^{n} P(A_i)\mu(dP).$$

Moreover, μ is uniquely characterized by the law of $\boldsymbol{X}_\infty$.

In the statement of the theorem, μ is a probability measure on the space $\mathcal{P}$ of all probability measure on $\mathbb{X}$ and it usually referred as the *De Finetti's measure*. $\mathcal{P}$ is assumed to be endowed with the topology of weak convergence, which makes it separable and metrizable by a complete metric, and with the corresponding Borel sigma algebra. This theorem was initially proved by De Finetti [28] for binary variables and then extended to variables taking values on a complete and separable metric space by Hewitt and Savage [58]. From De Finetti's theorem, the exchangeability assumption can also be formulated in terms of conditional independence, i.e., the law of a sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ from an infinite exchangeable sequence can be rewritten in the following hierarchical form

$$\begin{aligned} X_i | P &\overset{iid}{\sim} P \quad i = 1, \ldots, n, \\ P &\sim \mu, \end{aligned} \tag{2.1}$$

where P is an unknown parameter vector, taking values on $\mathcal{P}$, and its probability law μ is interpreted in a Bayesian framework as its prior distribution. If μ is supported on subset of $\mathcal{P}$ isomorphic to a finite dimensional space, the inferential problem is called *parametric*. Otherwise it is called *nonparametric*.

In many applied contexts, the assumption of exchangeability may be too restrictive. An example in species sampling problems is when we sample animals from distinct regions. Even if it may still be reasonable to assume observations to be exchangeable within each region, it may not be the case when they are grouped all together. Indeed, there can be strong geographical or climatic differences, which affect the distribution of each region. An assumption weaker than exchangeability and which can represent well this kind of contexts is *partial exchangeability*. This condition implies that observations are exchangeable within groups but, in general, not across them. An array $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_J)$ composed by J infinite sequences of random variables is partially exchangeable if each of its elements $\mathbf{X}_i$ is exchangeable. A corresponding De Finetti's representation theorem for partially exchangeable sequences can be stated as follows,

Theorem 2.1.2 (De Finetti's representation theorem for partially exchangeable sequences)**.** *$\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_J)$ is partially exchangeable iff there exists a probability measure μ on $\mathcal{P}^J$ such that for all $(n_1, \ldots, n_J) \in \mathbb{N}^J$ and for all $A_j = A_{j,1} \times \ldots \times A_{j,n_j} \times \mathbb{X} \times \ldots$, with $A_{j,i} \in \mathcal{X}$ for all $j \leq J$ and $i \leq n_j$,*

$$P(\boldsymbol{X}_1 \in A_1, \ldots, \boldsymbol{X}_J \in A_J) = \int_{\mathcal{P}^J} \prod_{i=1}^{n_1} P_1(A_{J,i}) \cdots \prod_{i=1}^{n_J} P_J(A_{J,i}) \mu(d(P_1, \ldots, P_J)).$$

Examples of priors μ for partially exchangeable data are the Hierarchical Dirichlet Process, Teh et al. [112], and the Hiercarchical Pitman-Yor process, Teh [110]. Under this law, the elements of $(P_1, \ldots, P_J)$ are conditionally independent given another random probability measure P_0 on $\mathbb{X}$. In section 2.5, we will present in detail the Hierarchical Pitman-Yor process. The Hierarchical Dirichlet process corresponds to the particular case in which all α_j for $j \in \{0, \ldots, J\}$ are set equal to zero.

2.2 The Dirichlet and the Pitman–Yor Processes

Within the framework of (2.1), the classical example of nonparametric prior μ is the *Dirichlet process*, introduced in the seminal paper by Ferguson [40]. There are many constructions of this process, each of these can be taken as a definition. As in Ferguson [40], we define the Dirichlet process through a description of its finite dimensional distributions.

Definition 2.2.1 (Dirichlet Process)**.** Let m be a finite measure on $(\mathbb{X}, \mathcal{X})$. μ is a Dirichlet process if for each measurable partition $(A_1, \ldots, A_n)$ of $\mathbb{X}$, the random vector $(\mu(A_1), \ldots, \mu(A_n))$ has distribution $\mathrm{Dir}(m(A_1), \ldots, m(A_n))$, where Dir denotes the Dirichlet distribution, with density function with respect to the Lebesgue measure on $\mathbb{R}^{n-1}$ given by

$$f_{(\mu(A_1),\ldots,\mu(A_n))}(x_1, \ldots, x_n; m(A_1), \ldots, m(A_n)) = \frac{\prod_{i=1}^n \Gamma(m(A_i))}{\Gamma(\sum_{i=1}^n m(A_i))} \prod_{i=1}^n x_i^{m(A_i)-1},$$

on the open $(n-1)$-dimensional simplex $\Delta_{n-1} = \{(x_1, \ldots, x_n) \in \mathbb{R}^{n-1} : 0 \geq x_i \geq 1, \sum_{i=1}^n x_i = 1\}$ and 0 otherwise.

Other possible constructions of the Dirichlet process are by normalizing a gamma process (Ferguson [40]), by a stick-breaking construction (Sethraman [106]) or by a description of the Chinese Restaurant process (Aldous [3]). In the rest of this section, we will focus on the Pitman–Yor process and we will describe some of its properties and of its possible constructions. The corresponding counterparts for the Dirichlet process are recovered by setting $\alpha = 0$.

The Two Parameter Poisson-Dirichlet process introduced in Pitman and Yor [93] is a generalization of the Dirichlet process. See also Perman et al. [89], Pitman [91], and

Pitman [94]. Is is also known as *Pitman–Yor (PY) process*, especially in the machine learning literature. Like the Dirichlet process, this process is a probability measure on the space of probability measures on the sample space and it assigns probability one to the set of discrete distributions. It is parametrized by three parameters, (α, θ, P_0), where P_0, called *base distribution*, is a distribution on the sample space and α and θ are two scalars satisfying $0 \leq \alpha < 1$ and $\theta > -\alpha$, respectively called *concentration* and *mass parameters*. The Dirichlet process is recovered as a particular case when $\alpha = 0$, $\theta = m(\mathbb{X})$ and $P_0(\cdot) = \frac{m(\cdot)}{\theta}$.

The Pitman–Yor admits a *stick breaking representation*, akin to that of Dirichlet process. Specifically, a sample P from a Pitman–Yor process can be represented in the following form

$$P = \sum_{i=1}^{\infty} p_i \delta_{X_i^*},$$

where $(X_i^*)_{i\geq 1}$ is an i.i.d. sequence of random variables with marginal distribution P_0, and $(p_i)_{i\geq 1}$ is another sequence of random variables independent of $(X_i^*)_{i\geq 1}$ and having the following stick-breaking distribution

$$p_i = V_i \prod_{k=1}^{i-1}(1 - V_k),$$

where

$$V_i \overset{\text{ind}}{\sim} \text{beta}(1 - \alpha, \theta + i\alpha),$$

for all $i \geq 1$.

A description of the posterior distribution of the PY process was derived in Pitman [92]. Given a sample $\mathbf{X}_n = (X_1, \ldots, X_n)$, such that $X_i | P \overset{iid}{\sim} P$ for $1 \leq i \leq n$ and $P \sim \text{PY}(\alpha, \theta, P_0)$, the posterior distribution of P given $\mathbf{X}_n$ satisfies the following distributional equation

$$P|\mathbf{X}_n \overset{d}{=} \sum_{i=1}^{K_n} w_i \delta_{X_i^*} + w_0 \tilde{P}, \tag{2.2}$$

where K_n is the number of distinct values in the sample $\mathbf{X}_n$, denoted by $(X_1^*, \ldots, X_{K_n}^*)$ and having multiplicities $(n_1, \ldots, n_{K_n})$, $(w_0, w_1, \ldots, w_{K_n})$ is a random vector distributed according to a $\text{Dir}(\theta + K_n\alpha, n_1 - \alpha, \ldots, n_{K_n} - \alpha)$ and $\tilde{P}$ is distributed according to a PY process with parameters $(\alpha, \theta + K_n\alpha, P_0)$ and it is independent of $(w_0, w_1, \ldots, w_{K_n})$.

Using this posterior representation, it is also possible to derive the so called *Chinese Restaurant Representation* of the PY process, describing the conditional law of the observation X_{n+1} when the underlying random distribution P has been integrated out,

$$X_{n+1}|\mathbf{X}_n, \alpha, \theta, P_0 \sim \sum_{i=1}^{K_n} \frac{n_i - \alpha}{\theta + n} \delta_{X_i^*} + \frac{\theta + K_n\alpha}{\theta + n} P_0. \tag{2.3}$$

Following this predictive distribution, observation X_{n+1} is assigned to an old cluster with value X_i^* with probability proportional to $n_i - \alpha$ or it is sampled from P_0 and forms its own cluster with probability proportional to $\theta + K_n\alpha$.

The Chinese Restaurant Process can also be described using a *Polya urn scheme*, also called the *two parameters Hoppe urn*, Zabell [128]. This urn scheme generalizes the popular Blackwell-MacQueen Polya urn scheme for the Dirichlet process of Blackwell and MacQueen [10]. The two parameters Hoppe urn works as follows. We consider sampling balls from a urn, initially containing only a black ball of mass θ. Every time the black ball is sampled, we introduce in the urn a ball of new color of mass $1 - \alpha$, we put the black ball back into the urn and we reinforce its mass by α units. The new color is sampled from the base measure P_0, where with 'colors' we intend the possible values taken by the observations. Instead, if a color ball is picked, the reinforcement procedure works as in the classical Polya urn, i.e., the ball is returned into the urn together with an additional ball of the same color. It is easy to check that this sampling scheme give rise to the predictive distribution (2.3).

The interpretation of the *hyperparameters* of the PY are the following. The base distribution P_0 is the mean of the prior. From the stick breaking representation, it is clear that it is also the law of the locations of the support points of the random distribution P. The diversity parameter θ regulates the variance of the prior around the prior mean P_0. With high θ, the prior guess is strong, i.e., the variance of the prior is small. With small θ, we are more uncertain about possible values of P. A posteriori, a high θ implies more weight on the prior information, while a small θ more weight on the information provided by the sample. This fact can also be read from the Chinese Restaurant Representation. In fact, the weight of the prior mean is larger when θ is large. The concentration parameter α affects how the total mass of P is spread across its support points. When α is small, the prior samples with high probability distributions with few points of very high mass. With a value of α close to 1, distributions with the total mass evenly spread across many support points are more likely to be sampled. Indeed, from the Chinese Restaurant Representation, the probability that the next observation is different from all the previous ones is higher when α is large.

The Pitman–Yor process offers more flexibility than the Dirichlet process and it has desirable properties when dealing with species sampling problems. In particular, it offers a flexible predictive structure in which the probability of observing a new species depends not only on the sample size, like for example in Dirichlet process, but also on the number of distinct values observed in the sample. At the same time, it maintains mathematical tractability. Moreover, differently from the Dirichlet process, with a Pitman–Yor prior the number of distinct observations K_n grows following a power law behavior as the sample size increases, a feature common to many real world datasets, as observed in Mitzenmacher [88] and Goldwater et al. [49]. The PY process is a useful prior when the

population size is large but unknown and the number of species in the population is also unknown. However, it is not an adequate prior when the goal is estimating the total number of species in the population. Indeed, from the stick-breaking representation, a sample from a PY process has an infinite number of atoms with probability one. Hence, a point estimate for the total number of species in the population is always infinite.

2.3 Species Sampling Models

Species sampling problems refer to a broad class of statistical problems in which samples are assumed to be drawn from a population of individuals, belonging to a possibly infinite number of species. Given a sample $\boldsymbol{X}_{n,k} = (X_1, \ldots, X_n)$ from this population, featuring $K_n = k$ species with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$, interest lies in estimating various population's features, as well as predicting characteristics of additional future samples. Quantities of interest are, among others, the number of unseen species in the population, the fraction of the population belonging to each species, the probability that further draws yield new species, the number of additional samples to achieve a desired level of coverage, the probability of replicating species with certain features, etc. Species sampling problems have originally appeared in ecology, and their importance has grown considerably in the recent years, driven by challenging applications arising from bioinformatics, computer science, genetics, linguistics, molecular biology, networking and data confidentiality, design of experiments, machine learning, etc. A broad range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, have been proposed for making inference on various species sampling problems. Important contributions are Good [50], Good and Toulmin [51], Efron and Thisted [34], Hill [59], Rasmussen and Starr [98], Mao [87], Lijoi et al. [76] and Barger and Bunge [7] to name a few.

In species sampling problems, we imagine the random sample $\boldsymbol{X}_{n,k} = (X_1, \ldots, X_n)$ to be sampled from a large population of individuals of various species, where X_i denotes the species of the i-th individual. The state space of the observations, denoted by $\mathbb{X}$, has to be interpreted as an arbitrary set of colors or tags, which are used to label distinct species. It is assumed that when a new species is observed its color is sampled from a diffuse distribution on $\mathbb{X}$, the base measure P_0. When these labels are irrelevant for prediction purposes, we can assume the predictive rule of the observations to have the following general structure

$$\mathbb{P}[X_1 \in \cdot] = P_0(\cdot) \tag{2.4}$$

and

$$\mathbb{P}[X_{n+1} \in \cdot \mid \boldsymbol{X}_{n,k}] = g(n, k, n_1, \ldots, n_k)P_0(\cdot) + \sum_{i=1}^{k} f_i(n, k, n_1, \ldots, n_k)\delta_{X_i^*}(\cdot), \tag{2.5}$$

for any $n \geq 1$, where g and f_i, for any $i \geq 1$, are non-negative functions such that $g(n, k, n_1, \ldots, n_k) + \sum_{1 \leq i \leq k} f_i(n, k, n_1, \ldots, n_k) = 1$ and $(X_1^*, \ldots, X_k^*)$ denote the k distinct values observed in $\boldsymbol{X}_{n,k}$. This predictive structure is a generalization of the Blackwell-MacQueen urn and of the two parameters Hoppe Urn. Indeed, in (2.5) we allow the weights attached to the base measure and to the delta functions of the already observed species to be generic functions of n, k and of the frequencies $(n_1, \ldots, n_k)$, hence not restricting them to have any specific shape.

Pitman [92] studies the class of probability laws having conditional distributions as in (2.4) and (2.5) under the further assumption that $\boldsymbol{X}_{n,k}$ is a sample from an infinite exchangeable sequence. The members of this class are called *species sampling sequences* and the following theorem, providing a description of them, is derived in Pitman [92].

Theorem 2.3.1 (Pitman [92], Proposition 11)**.** *Let $(X_i)_{i \in \mathbb{N}}$ be a species sampling sequence. Then,*

1. *as $n \to \infty$, (2.5) converges in total variation almost surely (a.s.) to a random measure*
$$\tilde{P} = \sum_i p_i^* \delta_{X_i^*} + (1 - \sum_i p_i^*) P_0,$$
where p_i^* is the asymptotic frequency of the i-th species to appear, X_i^*;

2. *the X_i^* are an i.i.d. sequence distributed according to P_0 and independently of the p_i^*;*

3. *$(X_i)_{i \in \mathbb{N}}$ is a sample from $\tilde{P}$, i.e., $X_i | \tilde{P} \overset{iid}{\sim} \tilde{P}$ for all i.*

Also, a converse of this theorem holds true. Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of random variables sampled from a random probability measure defined as

$$\tilde{P} = \sum_i p_i \delta_{\hat{X}_i} + (1 - \sum_i p_i) P_0, \tag{2.6}$$

where $(\hat{X}_i)_{i \in \mathbb{N}}$ and $(p_i)_{i \in \mathbb{N}}$ are two sequences of random variables, such that $p_i \geq 0$ and $\sum_{i \geq 1} p_i \leq 1$ a.s. and $\hat{X}_i$ are i.i.d. according to the diffuse distribution P_0 and independent of $(p_i)_{i \in \mathbb{N}}$. This kind of models are called *species sampling models*. If $\sum_{i \geq 1} p_i = 1$, then the species sampling model is said to be *proper* and in this case $(p_i^*)_{i \in \mathbb{N}}$ is a size-biased permutation of $(p_i)_{i \in \mathbb{N}}$. Instead, if the model is improper, then $\tilde{P}$ has also a continuous component of total mass $p_0 = 1 - \sum_{i \in \mathbb{N}} p_i$, usually referred as the Kingman's dust.

From a Bayesian nonparametric perspective, species sampling models rely on the specification of a prior distribution for the unknown species compositions $(p_i)_{i \geq 1}$ and a choice of the base measure P_0. The classical example is the Dirichlet process in which $(p_i)_{i \geq 1}$, when ordered in decreasing order, have a Poisson-Dirichlet distribution, introduced in

Kingman [71], while their size-biased permutation $(p_i^*)_{i\geq 1}$ has the stick-braking representation described above (for the PY process, with $\alpha = 0$) also called GEM(θ) distribution, from the contributions of Griffiths, Engel and McClosky. Similar results holds true for the PY process.

Another, apparently different, way to specify a prior distribution for $(p_i)_{i\geq 1}$ is by choosing a distribution for a random exchangeable partition of the natural numbers. A random partition $\Pi_{\mathbb{N}}$ of the natural numbers is a random variable taking values on the space of all possible partitions of $\mathbb{N}$ and it said to be exchangeable if its distribution is invariant with respect to all permutations of a finite number of integers. This means that, for all n and for all permutations $\varrho : \mathbb{N} \to \mathbb{N}$ such that all points greater than n are fixed points, a particular partition $\pi_{\mathbb{N}}$ of the natural numbers has the same probability as the partition obtained by applying ϱ to all elements in the blocks of $\pi_{\mathbb{N}}$. As a consequence of exchangeability, the vector of block sizes is sufficient for the distribution of $\Pi_{\mathbb{N}}$. Moreover, $\Pi_{\mathbb{N}}$ can be described by a sequences of consistent random partitions $(\Pi_n)_{n\in\mathbb{N}}$, where Π_n is a random partition of the first n integers and with consistent it is intended that Π_n is (a.s.) recovered from Π_{n+1} by removing the element $n+1$. See Pitman [95] for an excellent treatment on exchangeable random partitions.

A way of obtaining a random exchangeable partition of the natural numbers is by considering a sequence of exchangeable random variables $(X_n)_{n\in\mathbb{N}}$ and to generate the partition of $\mathbb{N}$ by grouping in the same block two integers i and j if and only if the corresponding random variables take the same value, i.e. if $X_i = X_j$. The following theorem by Kingman [72] shows that this construction is not so specific and indeed all random exchangeable partitions can be generated in this way. Kingman's theorem establishes a bijection between distributions of random exchangeable partitions of the natural numbers and distributions for the frequencies $(p_i)_{i\geq 1}$ of a species sampling model. Specifically, for a fixed (non-random) infinite vector $(p_i)_{i\geq 1}$ such that $\sum_{i\geq 1} p_i \leq 1$, let $p_0 = 1 - \sum_{i\geq 1} p_i$ (Kingman's dust) be added to it, so that the new vector $(p_i)_{i\geq 0}$ sums to one, $\sum_{i\geq 0} p_i = 1$. A random partition is said to be a *paintbox* if it is obtained by sampling a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ from $(p_i)_{i\geq 0}$ and grouping in the same blocks integers i and j if $X_i = X_j$, except when X_i (and so X_j) is sampled from p_0, in which case integer i (and so integer j) forms its own block. Kingman's representation theorem characterizes exhangeable partitions as mixtures of paintboxes, mixtured over $(p_i)_{i\geq 0}$.

Theorem 2.3.2 (Kingman [72])**.** *Let $\Pi_{\mathbb{N}}$ be an exchangeable partition of the natural numbers and let $\tilde{\Pi}_p$ denotes a paintbox partition obtained from a fixed ordered vector $p = (p_i)_{i\geq 1}$ such that $\sum_{i\geq 1} p_i \leq 1$. Then, there exists a probability measure μ such that, for all partition $\pi_{\mathbb{N}}$ of the natural number, it holds*

$$\mathbb{P}(\Pi_{\mathbb{N}} = \pi_{\mathbb{N}}) = \int_{\nabla} \mathbb{P}(\tilde{\Pi}_p = \pi_{\mathbb{N}})\mu(dp),$$

where μ is a unique probability measure on the infinite ordered simplex $\nabla = \{p \in [0,1]^\infty : 1 \leq p_1 \leq p_2 \leq \ldots \leq 0, \sum_{i\geq 1} p_i \leq 1\}$ and it is completely characterized by the law of $\Pi_{\mathbb{N}}$.

From this theorem, we can specify a species sampling model by choosing a base measure P_0 and either a prior μ for the frequencies $(p_i)_{i\geq 1}$ or equivalently a distribution for a sequence of exchangeable partitions of the natural numbers. An example of a wide class of such random partitions are the Gibbs-type partitions introduced in Gnedin and Pitman [48] and presented in the next section.

2.4 Gibbs-type priors

As recently discussed in De Blasi et al. [27], Gibbs-type species sampling models may be considered as the most "natural" generalization of the Dirichlet process. Indeed, apart of the well-known conjugacy of the Dirichlet process, Gibbs-type species sampling models share numerous properties that are appealing from both a theoretical and an applied point of view: they stand out in terms of mathematical tractability, which allows to study their distributional properties for finite sample sizes and asymptotically; they admit a simple and intuitive definition in terms of predictive probabilities; they are characterized by a flexible parameterization, thus including numerous interesting special cases, most of them still unexplored. All these properties have made the class of Gibbs-type priors a common choice in several contexts, such as in hierarchical mixture modeling, species sampling problems, feature and graph modeling, hidden Markov modeling, etc. In this section we briefly review Gibbs-type species sampling models, with emphasis towards their predictive probabilities and sampling properties. The reader is referred to the monographs by Pitman [95] and Bertoin [9] for a comprehensive account on Gibbs-type species sampling models, and to Lijoi and Prünster [80] and De Blasi et al. [27] for reviews on their use in Bayesian nonparametric statistics.

Among various possible definitions of Gibbs-type species sampling models, the most intuitive is given in terms of their predictive probabilities. These probabilities are of the form (2.5) for a suitable specification of g and f_i. Let $\boldsymbol{X}_{n,k} := (X_1, \ldots, X_n)$ denote a sample of size n from an arbitrary species sampling model P and featuring $K_n = k \leq n$ species, labelled by $(X_1^*, \ldots, X_{K_n}^*)$, with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$. Then P is a Gibbs-type species sampling model if

$$\mathbb{P}[X_1 \in \cdot] = P_0(\cdot) \tag{2.7}$$

and

$$\mathbb{P}[X_{n+1} \in \cdot \mid \boldsymbol{X}_{n,k}] = \frac{V_{n+1,k+1}}{V_{n,k}} P_0(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{i=1}^{k} (n_i - \alpha)\delta_{X_i^*}(\cdot), \tag{2.8}$$

for any $n \geq 1$, where $\alpha < 1$ and $(V_{n,k})_{1 \leq k \leq n, n \geq 1}$ are nonnegative weights with $V_{1,1} := 1$ and satisfying the triangular recursion $V_{n,k} = V_{n+1,k}(n - \alpha k) + V_{n+1,k+1}$. We observe that the class of Gibbs-type species sampling models generalizes the Dirichlet process by introducing the dependency on k in both the nonnegative functions g and f_i in (2.5), for any $i = 1, \ldots, k$.

A characterization of the de Finetti measure of the X_i's with distribution (2.7) and (2.8) was proposed in Gnedin and Pitman [48], and it relies on the notion of Poisson-Kingman model introduced in Pitman [94]. Specifically, for any $\alpha \in (0, 1)$ let $(J_i)_{i \geq 1}$ be decreasing ordered jumps of an α-stable subordinator Kingman [71], let $P_i = J_i / T_\alpha$ where $T_\alpha = \sum_{i \geq 1} J_i < +\infty$ almost surely, and let $\mathrm{PK}(\alpha; t)$ denote the conditional distribution of $(P_i)_{i \geq 1}$ given $T_\alpha = t$. In particular T_α is a positive α-stable random variable and we denote by f_α its density function. If $T_{\alpha,h}$ is a random variable with density function $f_{T_{\alpha,h}}(t) = h(t) f_\alpha(t)$, for any nonnegative function h, then an α-stable Poisson-Kingman model is defined as the discrete random probability measure $P_{\alpha,h} = \sum_{i \geq 1} P_{i,h} \delta_{X_i^*}$, where $(P_{i,h})_{i \geq 1}$ is distributed as $\int_0^{+\infty} \mathrm{PK}(\alpha; t) f_{T_{\alpha,h}}(t) dt$ and $(X_i^*)_{i \geq 1}$ are random variables, independent of $(P_{i,h})_{i \geq 1}$, and independent and identically distributed as P_0. According to Gnedin and Pitman [48], the predictive probabilities (2.8) are those of: i) the class of α-stable Poisson-Kingman models, for $\alpha \in (0, 1)$, ii) the Dirichlet process with (possibly) randomized mass parameter, for $\alpha = 0$; iii) the class of M-dimensional symmetric Dirichlet distributions, with M being a nonnegative discrete random variable, for $\alpha < 0$. In other terms, $\alpha < 0$ implies a finite number M of species in the population, whereas $\alpha \in [0, 1)$ implies infinite species.

The characterization in Gnedin and Pitman [48] implies that, for $\alpha \in (0, 1)$, Gibbs-type species sampling models are parameterized by a nonnegative function h. A representation of $V_{n,k}$ in terms of h was provided in Pitman [94]. Specifically, let $B_{a,b}$ be a Beta random variable with parameter (a, b) and, for any $c > 0$, let $S_{\alpha,c}$ be a random variable with density function $f_{S_{\alpha,c}}(s) = \Gamma(c\alpha + 1) s^{-\alpha c} f_\alpha(s) / \Gamma(c + 1)$, i.e., $S_{\alpha,c}$ is a polynomially tilted α-stable random variable. Then,

$$V_{n,k} = \frac{\alpha^k}{\Gamma(n - \alpha k)} \int_0^{+\infty} h(t) t^{-\alpha k} \int_0^1 p^{n-1-\alpha k} f_\alpha((1-p)t) dp dt,$$

i.e.,

$$V_{n,k} = \frac{\alpha^k \Gamma(k)}{\Gamma(n)} \mathbb{E}\left[h\left(\frac{S_{\alpha,k}}{B_{\alpha k, n-\alpha k}} \right) \right]. \tag{2.9}$$

See Chapter 4 of Pitman [95] for additional details. Note that, according to the representation (2.9), a Monte Carlo evaluation of $V_{n,k}$ can be easily performed by sampling from $B_{\alpha k, n-\alpha k}$ and $S_{\alpha,k}$. In this respect, an efficient rejection sampling scheme for polynomially tilted α-stable random variables has been proposed by Devroye [29].

Among Gibbs-type species sampling models with $\alpha \in (0, 1)$, the Pitman–Yor process certainly stands out. Another noteworthy example is the normalized generalized Gamma

process, introduced in Pitman [94] and further investigated in Bayesian nonparametrics, e.g., James [66], Lijoi et al. [78], Lijoi et al. [81] and James [67]. For any $\alpha \in (0,1)$ and $\theta > -\alpha$, the Pitman–Yor process is a Gibbs-type species sampling model with $h(t) = \alpha\Gamma(\theta)t^{-\theta}/\Gamma(\theta/\alpha)$. In particular, by replacing this function in (2.9), one obtains

$$V_{n,k} = \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)}{(\theta)_{n\uparrow}}, \tag{2.10}$$

where $(\theta)_{n\uparrow}$ denotes the ascending factorial, i.e., $(a)_{n\uparrow} := \prod_{0\leq i\leq n-1}(a+i)$ with the proviso $(a)_{0\uparrow} = 1$. For any $\alpha \in (0,1)$ and $\tau \geq 0$ the normalized generalized Gamma process is a Gibbs-type species sampling model with $h(t) = \exp\{\tau - \tau^{1/\alpha}t\}$. By replacing this function in (2.9),

$$V_{n,k} = \frac{\alpha^k \mathrm{e}^{\tau}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\tau^{1/\alpha})^i \Gamma\left(k - \frac{i}{\alpha}, \tau\right), \tag{2.11}$$

where $\Gamma(k - i/\alpha, \tau)$ denotes the incomplete Gamma function, i.e.,
$\Gamma(a,b) := \int_b^{+\infty} x^{a-1}\exp\{-x\}dx$. Note that (2.10) may be viewed as a suitable mixture of (2.11). Specifically, if $G_{\theta/\alpha,1}$ is a Gamma random variable with parameter $(\theta/\alpha, 1)$ then (2.10) coincides with (2.11) where τ is replaced by $G_{\theta/\alpha,1}$. In other terms, for $\theta > 0$ the Pitman–Yor may be viewed as hierarchical generalization of the normalized generalized Gamma process, with a Gamma prior over τ.

The predictive probabilities (2.8) lead to the distribution of the exchangeable random partition Π_n induced by a sample $(X_1, \ldots, X_n)$ from a Gibbs-type species sampling model. As mentioned in section 2.3, Π_n is an exchangeable partition and the vector $(n_1, \ldots, n_k)$ is sufficient for its distribution. Hence, we can denote by $p_k^{(n)}(n_1, \ldots, n_k)$ the probability of any particular partition π_n of the set $\{1, \ldots, n\}$ induced by $(X_1, \ldots, X_n)$, and featuring $K_n = k$ distinct blocks with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$, for any $n \geq 1$. Then, by a direct application of the predictive probabilities (2.8), one may easily verify that

$$\mathbb{P}(\Pi_n = \pi_n) = p_k^{(n)}(n_1, \ldots, n_k) = V_{n,k} \prod_{i=1}^{k} (1-\alpha)_{(n_i-1)\uparrow}. \tag{2.12}$$

Moreover, by marginalizing the joint distribution of K_n and $(N_{1,n}, \ldots, N_{K_n,n})$, $\mathbb{P}[K_n = k, (N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)] = (k!)^{-1}\binom{n}{n_1,\ldots,n_k} p_k^{(n)}(n_1, \ldots, n_k)$, with respect to the frequencies $(n_1, \ldots, n_k)$, one obtains

$$\mathbb{P}[K_n = k] = V_{n,k} \frac{\mathscr{C}(n,k;\alpha)}{\alpha^k},$$

where $\mathscr{C}(n,k;\alpha)$ denotes the generalized factorial coefficient, i.e.,

$$\mathscr{C}(n,k;a) := (k!)^{-1} \sum_{1\leq i\leq k} (-1)^i \binom{k}{i} (-ai)_{n\uparrow}. \tag{2.13}$$

As discussed in Pitman [95] and De Blasi et al. [27], the mathematical tractability of Gibbs-type species sampling models originates from the product form of the exchangeable partition probability function (2.12).

The parameter $\alpha < 0$ has an interpretable influence on the distribution of the exchangeable random partition Π_n with distribution (2.12). A first interpretation of α follows directly from the predictive probabilities (2.8). Indeed, $\alpha > 0$ acts an interesting reinforcement mechanism in the empirical part of the predictive probability (2.8). Note that the probability that X_{n+1} coincides with the species X_i^*, for any $i = 1, \ldots, k$, is a function of the frequency n_i and α. In particular, the ratio of the probabilities assigned to any pair of species (X_i^*, X_j^*) is $(n_i - \alpha)/(n_j - \alpha)$. If $\alpha \to 0$ this ratio reduces to the ratio of the frequencies of the two species, and therefore the coincidence probability is proportional to the frequency of the species. On the other hand if $\alpha > 0$ and $n_i > n_j$ then the ratio is an increasing function of α. Accordingly, as α increases the mass is reallocated from the species X_j^* to the species X_i^*. In other terms the sampling procedure tends to reinforce, among the observed species, those having higher frequencies. See De Blasi et al. [27] and references therein for a detailed discussion on such a reinforcement mechanism. If $\alpha < 0$, the reinforcement mechanism works in the opposite way in the sense that the coincidence probabilities are less than proportional to the species frequencies.

A further interpretation of the parameter α arises from the large n asymptotic behavior of the random variable K_n with distribution (2.4). This behaviour was first investigated by Korwar and Hollander [73] for the Dirichlet process, and then extended by Pitman [94] to the general framework of Gibbs-type species sampling model. See also Gnedin and Pitman [48] and Pitman [95] for details. The parameter α determines the rate at which K_n increases, as the sample size n increases. Three different rates may be identified for Gibbs-type species sampling models. Let

$$c_n(\alpha) := \begin{cases} 1 & \text{if } \alpha \in (-\infty, 0) \\ \log(n) & \text{if } \alpha = 0 \\ n^{\alpha} & \text{if } \alpha \in (0, 1), \end{cases}$$

for any $n \geq 1$. Then there exists a random variable S_α, positive and finite almost surely, such that

$$\frac{K_n}{c_n(\alpha)} \to S_\alpha \tag{2.14}$$

almost surely, as $n \to +\infty$. Using the terminology in Pitman [94], S_α is referred to as the α-diversity of the the Gibbs-type species sampling model. More precisely: i) for $\alpha \in (0, 1)$ the α-diversity coincides, in distribution, with $T_{\alpha,h}^{-\alpha}$; ii) for $\alpha = 0$ the α-diversity is a random variable degenerate at $\theta > 0$; iii) for $\alpha < 0$ the α-diversity coincides, in distribution, with the random number M of species in the population. Of course the larger α, the faster the rate of increase of K_n or, in other terms, the more new species are

generated from the sampling mechanism described in (2.8). The result 2.14 shows that for $\alpha \in (0,1)$, like for example for a PY process with $0 < \alpha < 1$, the number of distinct values K_n in a sample of size n, for large n, increases following a power-law behaviour.

Gibbs-type species sampling models have been extensively used in the context of Bayesian nonparametric inference for species sampling problems. See, e.g., Lijoi et al. [76], Favaro et al. [36], Favaro et al. [37] and Arbel et al. [4]. Species sampling problems represent probably the field in which the mathematical tractably of Gibbs-type species sampling models can be most appreciated. Indeed a plethora of posterior properties of Gibbs-type priors, for finite sample sizes and asymptotically, have been derived explicitly, thus providing fundamental tools for estimating population's features and predicting features of additional unobservable samples. Gibbs-type species sampling models have been also applied in mixture modeling, thus generalizing the seminal work by Lo [82]. See, e.g., Ishwaran and James [65], Lijoi et al. [75], Lijoi et al. [76], Favaro and Walker [39] and Lomeli et al. [84]. While maintaining the same computational tractability of the Dirichlet process mixture model, the availability of the additional parameter α allows for a better control of the clustering behaviour. Most recently, Gibbs-type species sampling models have been proposed for Bayesian nonparametric inference for ranked data in Caron and Fox [18], sparse exchangeable random graphs and networks in Caron and Fox [18] and Herlau [57], reversible Markov chains in Bacallado et al. [6], dynamic textual data in Chen et al. [22] and Chen et al. [23], and bipartite graphs in Caron [17].

2.5 The Hierarchical Pitman-Yor Model

The Hierarchical Pitman–Yor (HPY) model is a model for groups of partially exchangeable data introduced by Teh [110], as an extension of the Hierarchical Dirichlet Process of Teh et al. [112]. For a review of hierarchical models in Bayesian nonparametric statistics, the reader is referred to Teh and Jordan [113]. This is a useful model in problems in which multiple groups of data are available and where we wish to introduce probabilistic dependence across populations. In particular, it is an appropriate model when data incorporates a discrete variable of unknown cardinality. The discrete variable can either be at the observations level, as in our case in Chapter 5, or at the latent level, like when it parametrizes the distribution of continuous observations or when it works as a classification variable in mixture models settings.

Denoting with $\mathbf{X}_{n_{j\cdot}} = (X_{j1}, \ldots, X_{jn_{j\cdot}})$ the vector of observations from the j-th population, each one taking values on a measurable space $(\mathbb{X}, \mathcal{X})$, the HPY model is described

by the following hierarchical representation

$$\begin{aligned} X_{j,i}|P_j &\overset{iid}{\sim} P_j && i=1,\dots,n_{j\cdot\cdot}, \\ P_j|\alpha_j,\theta_j,P_0 &\overset{ind}{\sim} \mathrm{PY}\left(\alpha_j,\theta_j,P_0\right) && j=1,\dots,J, \\ P_0|\alpha_0,\theta_0,H &\sim \mathrm{PY}\left(\alpha_0,\theta_0,H\right), \end{aligned}$$

where $n_{j\cdot\cdot}$ is the number of observations from the j-th population, H is a fixed and diffuse probability measure and the $J+1$ couples of hyperparameters (α_j,θ_j) are chosen to satisfy the conditions $1>\alpha_j\geq 0$ and $\theta_j>-\alpha_j$ for all $j\in\{0,\dots,J\}$. Also, the hyperparameters α_j and θ_j are usually assumed to be unknown and endowed with priors.

In this model, observations from the j-th group, when conditioned to the realization of the unknown P_j, are independent and identically distributed with distribution P_j. Moreover, they are conditionally independent of observations from other populations. The P_j's are treated as random objects and endowed with Pitman–Yor priors with the same base measure P_0. This latter hyperparameter is not fixed by the modeller, but is considered as a random element to be inferred from data. Another Pitman–Yor prior is used as nonparametric distribution for P_0. Due to the almost sure discreteness of P_0, this recursive construction has the effect that the support of the P_j's is contained in that of P_0. As a consequence, all populations share the same random support of P_0.

The HPY admits a representation in terms of *Chinese Restaurant Franchise*. See Teh and Jordan [113] for a detailed description of this culinary metaphor. The Chinese Restaurant Franchise representation of the HPY is summarized by the following predictive distributions for the observables and for the cluster values in population j

$$X_{j,i+1}|X_{j,1},\dots,X_{j,i},\alpha_j,\theta_j,P_0\sim\sum_{t=1}^{m_{j\cdot}}\frac{n_{jt\cdot}-\alpha_j}{\theta_j+n_{j\cdot\cdot}}\delta_{X^*_{j,t}}+\frac{\theta_j+m_{j\cdot}\alpha_j}{\theta_j+n_{j\cdot\cdot}}P_0 \tag{2.15}$$

and

$$X^*_{j,m_{j\cdot}+1}|X^*_{1,1},\dots,X^*_{J,m_{J\cdot}},\alpha_0,\theta_0,H\sim\sum_{k=1}^{K}\frac{m_{\cdot k}-\alpha_0}{\theta_0+m_{\cdot\cdot}}\delta_{X^{**}_k}+\frac{\theta_0+K\alpha_0}{\theta_0+m_{\cdot\cdot}}H, \tag{2.16}$$

where K stands for total number of distinct values $(X^{**}_1,\dots,X^{**}_K)$ observed in the joined sample containing observations from all populations, n_{jtk} denotes the number of observations in population j, belonging to cluster t and having value X^{**}_k, while m_{jk} is the number of clusters in population j with value X^{**}_k, $(X^*_{j,1},\dots,X^*_{j,m_{j\cdot}})$ are the values of the $m_{j\cdot}$ clusters in population j. As in Teh and Jordan [113], dots in the indexes denote that we are summing over that index, e.g. $m_{j\cdot}$ is the total number of clusters in population j.

Formula (2.15) is the Chinese Restaurant Representation of P_j. The new observation $X_{j,i+1}$ belongs to an old cluster $X^*_{j,t}$ with probability proportional to $n_{jt\cdot}-\alpha_j$ or it forms a new cluster and it is sampled from P_0 with probability proportional to $\theta_j+m_{j\cdot}\alpha_j$. The

sequence of cluster values is sampled from P_0, which, being distributed as a PY process, also admits a Chinese Restaurant Representation, summarized by Formula (2.16). The new cluster in population j has the same value as one already observed in the joined sample with probability proportional to $m_{\cdot\cdot} - K\alpha_0$ or it has a new value, sampled from H, with probability proportional to $\theta_0 + K\alpha_0$. The hyperparameters (α_j, θ_j) have the same interpretation as in the Pitman-Yor case. Instead, the hyperparameters (α_0, θ_0) regulate the total number and the sharing of cluster values among populations. If θ_0 is low, the average total number of different cluster values K will be very low and, when a new sample from P_0 is needed, it will coincide with high probability with an already observed one. If α_0 is high, the sharing of cluster values across populations is low, while, with α_0 small, there is a small set of popular cluster values which are seen many times across all populations.

2.6 Feature models

Feature allocations are popular models in machine learning. A feature allocation is a generalization of a partition in which the assumptions of exhaustivity and disjointness of the blocks of a partition are relaxed. Indeed, in a feature allocation each element can belong to more than one block or not belong to any of them, with the only proviso that each element must belong to at most a finite number of blocks. A random feature allocation of the set $\{1, \ldots, n\}$ is simply a random variable taking values on the space of all possible features allocations of $\{1, \ldots, n\}$. Many of the properties of random partitions that we presented in section 2.3 have counterparts in the feature allocations context. In particular, in this section we will recall the definitions of consistent, exchangeable and ordered feature allocation and of Exchangeable Feature Probability Function (EFPF).

In feature models, we consider a set of n individuals, each one displaying a (possibly empty) set of features. Specifically, let $(\mathcal{X}, \mathcal{B})$ be measurable space, representing the collection of all possible features. Each individual is described by a random finite subset X_i of $\mathcal{X}$, collecting his features. Each feature $x \in \mathcal{X}$ can be shared by many individuals. Given a set of n individuals, a *feature allocation* describes the sharing of features among these individuals. A way of describing this sharing is to associate to each of the k points in $\cup_{1 \leq i \leq n} X_i$ a subset of $[n] := \{1, \ldots, n\}$, summarizing the individuals having that particular feature. We denote by $(f_{n,1}, \ldots, f_{n,k})$ the subsets of $[n]$ representing each of the k features and by $(m_1, \ldots, m_k)$ the cardinalities of these sets.

A feature allocation is *exchangeable* when its distribution is invariant under permutation of the indexes of the individual, i.e. the feature allocation induced by the random sets $(X_1, \ldots, X_n)$ is equal in distribution to that induced by $(X_{\varrho(1)}, \ldots, X_{\varrho(n)})$, for all permutation ϱ of $[n]$. Moreover, as pointed out in Broderick et al. [12], it is usually conve-

nient to assign an order to the k features present in a feature allocation of n individuals. A way of achieving this purpose is drawing k values from a continuous distribution and ordering the k features accordingly. The resulting feature allocation is said to be a *randomly ordered feature allocation*. In Broderick et al. [12], the authors study the class of exchangeable randomly ordered feature allocations admitting as a sufficient statistics the vector $(m_1, \ldots, m_k)$, i.e., the class of randomly ordered exchangeable feature allocations of the form

$$P(f_{n,1}, \ldots, f_{n,k}) = \pi_n(m_1, \ldots, m_k), \tag{2.17}$$

for a symmetric function π_n, called an *exchangeable feature probability punction* (EFPF), Broderick et al. [12].

When dealing with random exchangeable feature allocations, we also require consistency conditions that guarantee that the distribution of a feature allocation of n individuals coincides with that of $n-1$ individuals, when the last individual is integrated out. When considering randomly ordered exchangeable feature allocations with EFPF, this consistency notion specializes to the condition

$$\pi_n\left(m_1, \ldots, m_k\right) = \sum_{j=0}^{\infty} \binom{k+j}{j} \sum_{\underline{z} \in \{0,1\}^k} \pi_{n+1}(m_1 + z_1, \ldots, m_k + z_k, \underbrace{1, \ldots, 1}_{j}). \tag{2.18}$$

Feature allocations satisfying this condition are said to be *consisent*.

The most remarkable example of exchangeable consistent feature allocation with EFPF is the *Indian Buffet Process* (IBP), initially introduced in Griffiths and Ghahramani [55], in its one parameter version, and then extended to its two, Ghahramani et al. [46], and three parameters versions, Teh and Gorür [111]. The EFPF of a 3-parameter (γ, θ, α) IBP has the following form

$$\frac{1}{k!}\Big(\frac{\gamma}{(\theta+1)_{n-1\uparrow}}\Big)^k \exp\Big(-\sum_{i=1}^{n} \gamma \frac{(\alpha+\theta)_{i-1\uparrow}}{(1+\theta)_{i-1\uparrow}}\Big) \prod_{l=1}^{k} (1-\alpha)_{m_l-1\uparrow} (\theta+\alpha)_{n-m_l\uparrow}, \tag{2.19}$$

where $(x)_{m\uparrow}$ denotes the rising factorial and the parameters must satisfy the conditions $\gamma > 0$, $0 \leq \alpha < 1$, and $\theta > -\alpha$. The 2-parameter IBP is recovered when α is set equal to zero, and the 1 parameter IBP when we also impose $\theta = 1$. In the subsection 2.6.2, we provide a concise description of this process.

The IBP is derived as the limit of a Beta–Bernoulli model, in Griffiths and Ghahramani [55]. This latter model is the counterpart of the IBP when the set of all possible features $\mathcal{X}$ has finite cardinality, N. The EFPF of a Beta–Bernoulli model with parameters (N, α, θ) is

$$\binom{N}{k} \left(\frac{-\alpha}{(\theta+\alpha)_{n\uparrow}}\right)^k \left(\frac{(\theta+\alpha)_{n\uparrow}}{(\theta)_{n\uparrow}}\right)^N \prod_{i=1}^{k} (1-\alpha)_{m_i-1\uparrow} (\theta+\alpha)_{n-m_i\uparrow}, \tag{2.20}$$

where $\alpha < 0$ and $\theta > -\alpha$. In the next subsection, we provide a brief description of the Beta–Bernoulli model and a derivation of its EFPF.

2.6.1 The Beta–Bernoulli model

The *Beta–Bernoulli* is described by considering a finite space of features, which can numbered using the integers in $[N]$, where N is the cardinality of the feature set. To each feature we associate a random number q_j with distribution $\text{Beta}(\eta_1, \eta_2)$. Each individual X_i possesses feature j with probability q_j. Let $Z_{i,j}$ be a binary random variable denoting the presence or absence of feature j in individual i. Then, the Beta–Bernoulli model can be written as

$$Z_{i,j}|q_j \sim \text{Bernoulli}(p_j) \qquad i = 1, \dots, n;\ j = 1, \dots, N; \tag{2.21}$$

$$q_j|\eta_1, \eta_2 \sim \text{Beta}(\eta_1, \eta_2) \qquad j = 1, \dots, N. \tag{2.22}$$

The conditional probability that $Z = (Z_{i,j})_{i \leq n, j \leq N}$ is equal to $z = (z_{i,j})_{i \leq n, j \leq N}$ given $q = (q_1, \dots, q_N)$ is

$$\mathbb{P}(Z = z|q) = \prod_{j=1}^{N} \prod_{i=1}^{n} \text{Bernoulli}(z_{i,j}|q_j).$$

Integrating q out, we obtain the probability mass function of Z

$$\mathbb{P}(Z = z) = \left(\frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)} \right)^N \prod_{i=1}^{N} \frac{\Gamma(m_i + \eta_1)\Gamma(n - m_i + \eta_2)}{\Gamma(n + \eta_1 + \eta_2)},$$

where $m_i = \sum_{j=1}^{n} z_{i,j}$. If $(m_1, \dots, m_N)$ has k non-zero entries, this probability becomes

$$\Big(\frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)\Gamma(n + \eta_1 + \eta_2)} \Big)^N (\Gamma(\eta_1)\Gamma(n + \eta_2))^{N-k} \prod_{i=1}^{k} \Gamma(m_i + \eta_1)\Gamma(n - m_i + \eta_2). \tag{2.23}$$

Finally, taking into account all $\binom{N}{k}$ possible uniform orderings of the $N - k$ features not possessed by any individual, which give rise to the same uniformly ordered feature allocation, we obtain

$$\binom{N}{k} \left(\frac{\Gamma(\eta_1 + \eta_2)}{\Gamma(\eta_1)\Gamma(\eta_2)\Gamma(n + \eta_1 + \eta_2)} \right)^N (\Gamma(\eta_1)\Gamma(n + \eta_2))^{N-k} \prod_{i=1}^{k} \Gamma(m_i + \eta_1)\Gamma(n - m_i + \eta_2),$$

which can be be rewritten as in formula (2.20), by using rising factorials and by changing the parametrization to $\alpha = -\eta_1$ and $\theta = \eta_2 + \eta_1$, with $\alpha < 0$ and $\theta > -\alpha$.

2.6.2 The Indian Buffet Process

The Indian Buffet Process was initially derived in Griffiths and Ghahramani [55] in its 1 parameter version, as a limit of the Beta–Bernoulli model (2.23). In this subsection, we briefly present its 3 parameter version introduced in Teh and Gorür [111]. Rather than specifying a hierarchical model of the form (2.21) for the IBP, we describe only a simple

predictive construction of the $Z_{i,j}$. We remark that a representation of the form (2.21) is available also for the IBP. It was derived in Thibaux and Jordan [114] and it relies on notions of Completely Random Measures, which are out of the scope of our work. This description relies on the fact that the X_is of an exchangeable feature allocations are exchangeable random variables and consequently they have a De Finetti's measure. Thibaux and Jordan [114] shows that the underlying de Finetti's measure of the IBP is the Beta process, introduced by Hjort [60] in a context of survival analysis. For a good review of the IBP, its possible derivations and its applications in machine learning, the reader is refered to Griffiths and Ghahramani [56].

In the Indian Buffet Process the number of features $[N]$ is assumed to be infinity. As in (2.21), each feature has its own random parameter q_j which is distributed as a degenerate Beta distribution. The rows of the matrix $(Z_{i,j})_{i\leq n, j\in\mathbb{N}}$ are now infinite vectors. We describe how to generate a binary matrix $(z_{i,j})_{i\leq n, j\in\mathbb{N}}$ distributed according to a IBP, by describing the conditional probability of row $Z_{i,\cdot}$ given $(Z_{1,\cdot}, \ldots, Z_{i-1,\cdot})$. The generative process of the rows of $(Z_{i,j})_{i\leq n, j\in\mathbb{N}}$ works as follows:

1. Sample a random variable K_1 from a Poisson distribution of parameter γ. Construct the first row $Z_{1,\cdot}$ by setting $Z_{1,j} = 1$ for $j = 1, \ldots, K_1$ and $Z_{1,j} = 0$ for $j > K_1$;

2. Given that the first $i-1$ rows $(Z_{1,\cdot}, \ldots, Z_{i-1,\cdot})$ have K_{i-1} non-zero columns (i.e., columns with at least one 1 entry), generate the next row $Z_{i,\cdot}$ in the following way:

   (a) for j in $1, \ldots, K_{i-1}$, sample $Z_{i,j}$ from a Benoulli distribution with parameter $\frac{m_j^{(i-1)} - \alpha}{\theta + i - 1}$, where $m_j^{(i-1)} = \sum_{l=1}^{i-1} Z_{l,j}$;

   (b) sample a random variable K_i^+ from a Poisson distribution with parameter $\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+i)} \frac{\Gamma(\theta+\alpha-1+i)}{\Gamma(\theta+\alpha)}$. Set $Z_{i,j} = 1$ for $j = K_{i-1} + 1, \ldots, K_{i-1} + K_i^+$ and $Z_{i,j} = 0$ for $j > K_{i-1} + K_i^+$.

By using this predictive rule and multiplying for a combinatorial coefficient accounting for all possible random reordering of the columns giving rise to the same feature allocations (see e.g. Griffiths and Ghahramani [56] or Broderick et al. [12] for details), the probability of randomly ordered exchangeable feature allocation corresponding to a matrix $(Z_{i,j})_{i\leq n, j\in\mathbb{N}}$ generated in this way and with n rows, k non-zero columns and $(m_1, \ldots, m_k)$ 1 entries in each of these k columns, is

$$\frac{1}{k!}\left(\gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+\alpha)}\right)^k \exp\left(-\sum_{i=1}^{n} \gamma \frac{\Gamma(\theta+1)}{\Gamma(\theta+i)} \frac{\Gamma(\theta+\alpha-1+i)}{\Gamma(\theta+\alpha)}\right) \cdot \left[\prod_{l=1}^{k} \frac{\Gamma(m_l-\alpha)}{\Gamma(1-\alpha)} \frac{\Gamma(\theta+n-m_l+\alpha)}{\Gamma(\theta+n)}\right],$$

which can be easily rewritten as in formula (2.19) by using factorial coefficients.

Chapter 3

Sufficientness Postulate and Urn Scheme for Gibbs-type Priors

3.1 Motivation

In the Bayesian nonparametric framework for species sampling problems, we consider a random vector $\boldsymbol{X}_{n,k} = (X_1, \ldots, X_n)$ of observations to be sampled from an exchangeable sequence directed by P, namely

$$\begin{aligned} X_i \,|\, P &\overset{iid}{\sim} P, \qquad i = 1, \ldots, n, \\ P &\sim \mu. \end{aligned} \tag{3.1}$$

An important issue consists in selecting the prior distribution μ. Of course one approach is to select μ by appealing to prior information about P, and then attempt to incorporate this information into μ. This is often a difficult task for nonparametric priors, since P is an infinite dimensional object. Alternatively, we may select μ by assuming that the predictive probabilities

$$\mathbb{P}[X_1 \in \cdot] = P_0(\cdot) \tag{3.2}$$

and

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, \boldsymbol{X}_{n,k}] = g(n, k, n_1, \ldots, n_k) P_0(\cdot) + \sum_{i=1}^{k} f_i(n, k, n_1, \ldots, n_k) \delta_{X_i^*}(\cdot), \tag{3.3}$$

obey or exhibit some characteristic or property. Indeed in practical applications it may be that the form of the functions g and f_i may be an adequate description of our current state of knowledge. An early discussion of this alternative approach, although in a parametric framework, dates back to the seminal work by English philosopher W. E. Johnson. Specifically, assuming $T < +\infty$ possible species that are known and equiprobable prior to observations, Johnson [69] characterized the T-dimensional symmetric Dirichlet distribution as the unique prior for which g depends only on n and T and f_i depends only on

n, n_i and T. As a direct consequence of the parametric assumption that $T < +\infty$, of course, $g = 0$ for all $k \geq T$. Using the terminology in Good [52], this characterization of the Dirichlet prior is referred to as the Johnson's "sufficientness" postulate. We refer to Zabell [125] and Zabell [127] for a review of the work by Johnson [69]. See also the monograph by Zabell [129] for a more comprehensive account on sufficientness, exchangeability and predictive probabilities.

Following these initial contributions, some generalizations of the Johnson's postulate that arise by removing the parametric assumption of a prespecified number $T < +\infty$ of possible species in the population have been proposed. In particular, Regazzini [100], and later on Lo [83], first provided a nonparametric counterpart of the Johnson's postulate. Indeed, under the assumption of an infinite number of species in the population, they showed that the Dirichlet process is the unique species sampling model for which g depends only on n, and f_i depends only on n and n_i. An extension of this nonparametric sufficientness postulate was presented in Zabell [128], and it characterizes the Pitman–Yor process as the unique species sampling model for which g depends only on n and k, and f_i depends only on n and n_i. In this chapter, we revisit and discuss the seminal work of Zabell [128] within the more recent framework of Gibbs-type species sampling models. Such a framework, which includes the Dirichlet process and the Pitman–Yor process as special cases, suggests the formulation of a novel nonparametric sufficientness postulate in which g depends only on n and k, and f_i depends only on n, k and n_i. Our discussion also includes an intuitive Pólya-like urn scheme for describing the predictive probabilities of Gibbs-type species sampling models. We show how the sufficientness postulate of Zabell [128], as well as of Regazzini [100], may be rephrased in terms of this Pólya-like urn scheme.

The chapter is structured as follows. Section 3.2 reviews the sufficientness postulate introduced by Zabell [128]. In Section 3.3, we present and prove the extension of this result to the more general framework of Gibbs-type species sampling models. In Section 3.4 we provide a Polya-like urn scheme representation for the Gibbs-type models. In Section 3.5, we show how the characterization theorems for the Dirichlet and the Pitman–Yor processes by Zabell [129] can be recovered and rephrased in terms of the proposed urn scheme. Finally, Section 3.6 contains a discussion of future work.

3.2 Review of "suffecientness" postulates

With "sufficientness" postulates we intend a class of theorems that characterize (classes of) priors through conditions on the law of observations. Imposing these conditions is then sufficient to characterize the prior, but, to distinguish from the usual probabilistic notion of sufficiency, Good [52] coined the new word "sufficientness". The first result of this kind

is the characterization, in a context of exchangeability, of the symmetric Dirichlet prior in a work of Johnson. This result is even more remarkable because it appeared in 1924, before the appearance of the notion of exchangeability and of De Finetti's Representation Theorem. Indeed, Johnson already introduced the notion of exchangeability, which he called "permutation postulate". Following this early work, a series of other "sufficientness" postulates have been derived. These results usually rely on conditions on the predictive distributions of the observables, which completely specify the law of the entire process of observations, thanks to the Ionescu-Tulcea's theorem. We remark that equivalent results could be stated using conditions on the finite dimensional distributions of the process, relying on Kolmogorov's Extension Theorem. In this section, we recall Zabell's characterization of the PY process and we briefly mention a few others "sufficientness" results.

A generalization of the *Johnson's sufficientness postulate* to the PY process prior was first discussed in Zabell [128]. Specifically, let P be an arbitrary species sampling model with predictive probabilities (2.5), and consider the following assumptions:

A1) $\mathbb{P}[\Pi_n = \pi_n] > 0$ for all the partitions π_n of $\{1, \ldots, n\}$, that is no scenario is deemed, a priori, to be impossible;

A2) $g(n, k, n_1, \ldots, n_k) = g(n, k)$, that is the probability of observing a new species depends only on n and k;

A3) $f_i(n, k, n_1, \ldots, n_k) = f(n, n_i)$, that is the probability of observing the species X_i^* depends only on n and n_i.

Zabell [128] showed that if just these three assumptions are imposed, then there exist three parameters $\alpha \in (0, 1)$, $\theta > -\alpha$ and $c_n \geq 0$ such that

i) if $k \geq 2$ then

$$g(n, k) = \frac{\theta + k\alpha}{\theta + n}; \qquad f(n, n_i) = \frac{n_i - \alpha}{\theta + n}; \tag{3.4}$$

ii) if $k = 1$ then

$$g(n, k) = \frac{\theta + \alpha}{\theta + n} - c_n; \qquad f(n, n) = \frac{n - \alpha}{\theta + n} + c_n. \tag{3.5}$$

In other words, if a species sampling model satisfies A1)-A3), then the functions g and f_i in the predictive probabilities (3.3) must have the expressions (3.4) and (3.5). The sufficientness postulate of Zabell [128] may be viewed as a nonparametric counterpart of the classical Johnson's postulate, in the sense that it removes the assumption of a prespecified number $T < +\infty$ of possible species in the population.

As detailed discussed in the work of Zabell [128], the parameter c_n refers to the prior probability of observing a unique species in an infinite sequence of trials. In particular

one has $c_n = 0$ under the following additional assumption: A4) $K_n \to +\infty$ almost surely as $n \to +\infty$, that is the number of species in the population is infinite. Therefore if $\boldsymbol{X}_{n,k}$ is a sample of size n from a species sampling model featuring $K_n = k \leq n$ species, labelled by $(X_1^*, \ldots, X_{K_n}^*)$, with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$, then under A1)-A4) one has

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, \boldsymbol{X}_{n,k}] = \frac{\theta + k\alpha}{\theta + n} P_0(\cdot) + \frac{1}{\theta + n} \sum_{i=1}^{k} (n_i - \alpha) \delta_{X_i^*}(\cdot), \tag{3.6}$$

for any $n \geq 1$, which are the precisely the predictive probabilities of the Pitman–Yor process.

Other examples of sufficientness results under the assumption of exchangeability are Walker and Muliere [122] and Walker and Muliere [123], which respectively characterize the Polya tree distribution and the class of Neutral to Right priors of Doksum [32]. See Fortini and Petrone [42] for a review of predictive characterization theorems of priors in a setting of exchangeability. We remark that similar sufficientness results can be stated also outside the context of exchangeability. Indeed, there are many representation theorems extending De Finetti's Representation theorem to more general contexts. These theorems provide mixture representations of classes of probability laws invariant under groups of transformations, e.g. for exchangeable sequences the group of finite permutations of the natural numbers. For an abstract treatment of these integral representation theorems we refer to Dynkin [33]. For a simpler treatment of the same topic see Diaconis and Freedman [31]. An example of these generalizations of De Finetti's Theorem is a result by Diaconis and Freedman [30] which characterizes recurrent Markov exchangeable sequences with countable state space as mixtures of Markov Chains. Examples of sufficientness results within this Markovian framework are Rolles [102] and Bacallado et al. [6]. For a review of predictive characterizations of priors for Markov exchangeable sequences see Fortini and Petrone [43] and references therein.

3.3 Sufficientness postulate for Gibbs-type priors

Accordingly to the characterization proposed by Zabell [128], the Pitman–Yor process is the unique species sampling model for which the function g depends only on n and k, and the function f_i depends only on n and n_i, for any $i = 1, \ldots, k$. See also the monograph by Zabell [129] and references therein for a detailed account. For any index $\alpha \in (0, 1)$ the predictive probabilities (2.8) of a Gibbs-type species sampling model generalize the predictive probabilities (3.6) of the Pitman–Yor process by introducing the dependency on k in the function f. In particular one may rephrase the sufficientness postulate of Zabell [128] as follows: for any index $\alpha \in (0, 1)$ the Pitman–Yor process is the only Gibbs-type

species sampling model for which the ratio $V_{n+1,k}/V_{n,k}$ simplifies in such a way to remove the dependency on the number k of observed species. Note that the normalized generalized Gamma process, whose predictive probabilities are expressed in terms of the $V_{n,k}$'s in (2.11), provides a representative example of a Gibbs-type species sampling model for which such a simplification does not occur. The predictive probabilities of Gibbs-type species sampling models thus suggest for the formalization of a more general nonparametric sufficientness postulate, where the original assumption A3) of Zabell's result is replaced by the assumption $f_i(n, k, n_1, \ldots, n_k) = f(n, k, n_i)$, that is the probability of observing the species X_i^* depends only on n, k and n_i, for any $i = 1, \ldots, k$. Along lines similar to Zabell [128] the following generalized nonparametric sufficientness can be proved.

Theorem 3.3.1. *Let P be an arbitrary species sampling model with predictive probabilities of the form displayed in* (2.5)*, and let us consider the following four assumptions:*

A1) $\mathbb{P}[\Pi_n = \pi_n] > 0$ *for all the partitions* π_n *of the set* $\{1, \ldots, n\}$*;*

A2) $g(n, k, n_1, \ldots, n_k) = g(n, k)$*;*

A3) $f_i(n, k, n_1, \ldots, n_k) = f(n, k, n_i)$ *for any* $i = 1, \ldots, k$*;*

A4) $K_n \to +\infty$ *almost surely as* $n \to +\infty$*.*

Under A1)-A4) there exists a parameter $\alpha \in (0, 1)$ *and a collection of nonnegative weights* $(V_{n,k})_{1 \leq k \leq n, n \geq 1}$ *with* $V_{1,1} = 1$ *and* $V_{n,k} = V_{n+1,k}(n - \alpha k) + V_{n+1,k+1}$ *such that*

$$g(n, k) = \frac{V_{n+1,k+1}}{V_{n,k}} \tag{3.7}$$

and

$$f(n, k, n_i) = \frac{V_{n+1,k}}{V_{n,k}}(n_i - \alpha), \tag{3.8}$$

for any $i = 1, \ldots, k$*. In other terms, if a species sampling model P satisfies the assumptions A1)-A4) then P is a Gibbs-type species sampling model with parameter* $\alpha \in (0, 1)$*.*

As we pointed out in the Section 3.1, the nonparametric sufficientness postulate of Zabell [128] extended an analogous postulate of Regazzini [100] by including the dependency on k in the function g. Theorem 3.3.1 provides an even more general framework by including the dependency on k in both the functions g and f_i, for any $i = 1, \ldots, k$, while maintaining the same structure with respect to the dependency on the frequencies $(n_1, \ldots, n_k)$. The proof of this new postulate is along lines similar to that in Zabell [128], and it consists in verifying the following statements

i) the function $f(n, k, n_i)$ is a linear with respect to n_i, for any $n \geq 1$, $1 \leq k \leq n$, i.e., there exist parameters $a_{n,k}$ and $b_{n,k}$ such that $f(n, k, m) = a_{n,k} + b_{n,k}m$;

ii) the parameter $b_{n,k}$ is different from zero, for any $n \geq 1$ and $1 \leq k \leq n$, in order to allow for normalizing $f(n,k,n_i)$ and $g(n,k) = 1 - \sum_{1 \leq i \leq k} f(n,k,n_i)$ with respect to $b_{n,k}$;

iii) the normalized parameter $a_{n,k}$, which is denoted by $r_{n,k} = -a_{n,k}/b_{n,k}$, does not depend on n and k, for any $n \geq 1$ and $1 \leq k \leq n$, and in particular $r_{n,k} \in (0,1)$.

iv) the new parametrization $V_{n,k}$ is introduced and it is shown to satisfy the Gibbs-type recursion.

It is worth pointing out that the assumption A4) has been imposed only for the sake of simplicity in our exposition. Indeed, similarly to Zabell [128], an even more general sufficientness postulate can be derived by removing A4) and including an additional parameter related to the prior probability that only one species is observed in an infinite sequence of trials. Of course this would lead to a class of generalized Gibbs-type species sampling models.

Throughout the rest of this section, we are going to prove Theorem 3.3.1. We denote by $\boldsymbol{X}_{n,k}$ a sample of size n featuring $K_n = k \leq n$ species, labelled by $X_1^*, \ldots, X_{K_n}^*$, with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$. In the next proposition, we start by showing that the assumptions A1)-A4) imply that $f_i(n,k,n_i)$ is linear in n_i, for any $n \geq 1$, $2 \leq k \leq n$. The special case $k = 1$ will be considered separately later on.

Proposition 3.3.2. *For all $n \geq 1$, $2 \leq k \leq n$ and $1 \leq m \leq n-k+1$ there exist constants $a_{n,k}$ and $b_{n,k}$ such that $f(n,k,m) = a_{n,k} + b_{n,k}m$.*

Proof. It $n \leq 3$ it is immediate that we can set the constants $a_{n,k}$ and $b_{n,k}$ in such a way that the function $f(n,k,m)$ is linear with respect to m. Therefore, let us consider the case $n \geq 4$. If $k = n$ then $m = 1$ and we can choose $a_{n,n}$ and $b_{n,n}$ in such a way that $f(n,n,1) = a_{n,n} + b_{n,n}$. Furthermore, if $k = n-1$ then $1 \leq m \leq 2$ and $a_{n,n-1}$ and $b_{n,n-1}$ are the solutions of the following system of equations

$$f(n,n-1,1) = a_{n,n-1} + b_{n,n-1}$$

$$f(n,n-1,2) = a_{n,n-1} + b_{n,n-1}2.$$

If $3 \leq k \leq n-2$ then $1 \leq m \leq n-k+1$ and we can prove the linearity of the function $f(n,k,m)$ by choosing $a_{n,k}$, $b_{n,k}$ as the solutions of the following system of equations

$$f(n,k,1) = a_{n,k} + b_{n,k}$$

$$f(n,k,n-k+1) = a_{n,k} + b_{n,k}(n-k+1),$$

and then showing that $f(n,k,m)$ has constant increments, namely $f(n,k,m+1) - f(n,k,m) = f(n,k,m) - f(n,k,m-1)$ for any $1 < m < n-k+1$. Let us consider

$(n_1, n_2, \ldots) = (2, 2, \ldots)$ with $k-2$ groups with a total of $n_3 + \ldots + n_k = n-4$ elements. Then,

$$f(n,k,2) + f(n,k,2) + \sum_{i=1}^{k} f(n,k,n_i) + g(n,k) = 1$$

and

$$f(n,k,1) + f(n,k,3) + \sum_{i=1}^{k} f(n,k,n_i) + g(n,k) = 1.$$

By equating these identities we obtain $f(n,k,3) - f(n,k,2) = f(n,k,2) - f(n,k,1)$. Along the same lines, for the partition $(n_1, n_2, \ldots) = (j, 2, \ldots)$ for any $3 \leq j \leq n-k$ we obtain $f(n,k,n_i+1) - f(n,k,n_i) = f(n,k,n_i) - f(n,k,n_i-1)$ for any $1 < n_i < n-k+1$. Finally, if $k=2$ by exchangeability $f(n,2,n_i)g(n+1,2) = g(n,2)f(n+1,3,n_i)$. Therefore $f(n,2,n_i)$ is linear since by the assumption A4) $g(n+1,2) > 0$. □

Lemma 3.3.3. *For all $n \geq 1$ and $2 \leq k \leq n$, $b_{n,k} \neq 0$.*

Proof. Since $k > 1$, for $n = 1$ there is nothing to prove since we are assuming $k > 1$. For $n = 2$ and $k = 2$, $b_{2,2}$ can be arbitrarily selected in such a way to be different from 0. Now, let us assume that $b_{n,k} = 0$ for some $n \geq 3$ and $2 \leq k \leq n$. Given $\boldsymbol{X}_{n,k}$ consider the following events: i) we observe the species X_i^* at the $(n+1)$-th draw and a new species at the $(n+2)$-th draw; ii) we observe a new species at the $(n+1)$-th draw and the species X_i^* at the $(n+2)$-th draw. By exchangeability these two events have the same probability, and from Proposition (3.3.2),

$$(1 - ka_{n,k})(a_{n+1,k+1} + b_{n+1,k+1}n_i) = a_{n,k}(1 - ka_{n+1,k} - b_{n+1,k}(n+1)),$$

that is

$$(1 - ka_{n,k})b_{n+1,k+1}n_i = a_{n,k}(1 - ka_{n+1,k} - b_{n+1,k}(n+1)) - (1 - ka_{n,k})a_{n+1,k+1}.$$

Therefore $(1-ka_{n,k})b_{n+1,k+1}n_i$ is constant as a function of n_i in the range $1 \leq n_i < n-k+1$. It follows that if $n \geq 3$, so that both $n_i = 1$ and $n_i = 2$ are possible, then either $1 - ka_{n,k} = 0$ or $b_{n+1,k+1} = 0$. If $1 - ka_{n,k} = 0$ then $a_{n,k} = k^{-1}$, which implies

$$0 = \frac{1}{k}(1 - ka_{n+1,k} - b_{n+1,k}(n+1)) = \frac{1}{k}g(n+1,k),$$

where, from A4), $g(n+1,k) > 0$. Hence $b_{n+1,k+1} = 0$. Furthermore, given $\boldsymbol{X}_{n,k}$ consider the following events: i) we observe the species X_i^* at the $(n+1)$-th draw and the species X_j^* at the $(n+2)$-th draw; ii) we observe the species X_j^* at the $(n+1)$-th draw and the species X_i^* at the $(n+2)$-th draw. By exchangeability, one has

$$a_{n,k}(a_{n+1,k} + b_{n+1,k}n_i) = a_{n,k}(a_{n+1,k} + b_{n+1,k}n_j).$$

Therefore either $a_{n,k} = 0$ or $b_{n+1,k} = 0$. If $a_{n,k} = 0$ then $g(n,k) = 1$ against the assumption A4), hence it must be $b_{n+1,k} = 0$. Therefore, if $b_{n,k} = 0$ then it must be $b_{n+1,k} = b_{n+1,k+1} = 0$. If $b_{n,k} = 0$, then also must be $b_{n-1,k-1} = 0$. Indeed, one has

$$(a_{n-1,k-1} + b_{n-1,k-1}n_i)g(n,k) = g(n-1,k-1)(a_{n,k} + b_{n,k}n_i)$$

and

$$b_{n-1,k-1}n_i = \frac{g(n-1,k-1)}{g(n,k-1)}a_{n,k} - a_{n-1,k-1}.$$

Furthermore, by similar arguments, if $b_{n,k} = 0$ then also must be $b_{n-1,k} = 0$. Indeed, one has

$$(a_{n-1,k} + b_{n-1,k}n_i)(a_{n,k} + b_{n,k}n_j) = (a_{n-1,k} + b_{n-1,k}n_j)(a_{n,k} + b_{n,k}n_i)$$

and

$$(a_{n-1,k} + b_{n-1,k}n_i)a_{n,k} = (a_{n-1,k} + b_{n-1,k}n_j)a_{n,k}$$

with $a_{n,k} \neq 0$, otherwise $g(n,k) = 1$ which is against the assumption A4). Accordingly $b_{n-1,k}n_i = b_{n-1,k}n_j$ which implies $b_{n-1,k} = 0$. Accordingly, if $b_{n,k} = 0$, then all $b_{m,l} = 0$ for any $n \geq 2$ and $2 \leq k \leq n$. Arguing as before we see that $(1 - ka_{n,k})a_{n+1,k+1} = a_{n,k}(1 - ka_{n+1,k})$. In particular this double recursion admits a solution with all elements between 0 and 1 only if the initial condition $(a_{n,2})_{n\geq 2}$ is constant after some n^*, which in turns implies that for n large all $a_{n,k}$ must be constant. Let us denote by a the common value of $a_{n,k}$ for n large. Since $a > 0$ it follows that $na > 1$ for large n. But this is impossible because $1 - an = g(n,n) > 0$. □

By a direct application of Proposition 3.3.2 and Lemma 3.3.3, we can define a normalized version of $f(n,k,n_i) = a_{n,k} + b_{n,k}n_i$ with respect to $b_{n,k}$, for any $n \geq 1$ and $2 \leq k \leq n$. This is because we showed that $b_{n,k} \neq 0$. Furthermore, due to the fact that $g(n,k) = 1 - \sum_{1\leq i\leq k} f(n,k,n_i)$, a normalization of $g(n,k)$ is also obtained. In particular, for any $n \geq 1$ and $2 \leq k \leq n$, let $r_{n,k} = -a_{n,k}/b_{n,k}$. Then, we can write

$$f(n,k,n_i) = b_{n,k}(n_i - r_{n,k})$$

and

$$g(n,k) = 1 - \sum_{i=1}^{k} f(n,k,n_i) = 1 - b_{n,k}(n - kr_{n,k}).$$

In the next lemma we show that the parameter $r_{n,k}$ is constant, independent of both n and k, and it is greater than 0 and less or equal to 1, for any $n \geq 1$ and $2 \leq k \leq n$.

Lemma 3.3.4. *For all $n \geq 1$ and $2 \leq k \leq n$, $r_{n,k}$ is equal to constant value $\alpha \in [0,1)$.*

Proof. We start by showing that $r_{n,k}$ is constant with respect to variations of $n \geq 1$ and $2 \leq k \leq n$. Then we complete the proof by showing that $r_{n,k} \in [0,1)$. Let $n \geq 3$ and

consider the following events: i) we observe the species X_i^* at the $(n+1)$-th draw and the species X_j^* at the $(n+2)$-th draw; ii) we observe the species X_j^* at the $(n+1)$-th draw and the species X_i^* at the $(n+2)$-th draw. By exchangeability these two events have the same probability. In particular we can write

$$b_{n,k}(n_i - r_{n,k})b_{n+1,k}(n_j - r_{n+1,k}) = b_{n,k}(n_j - r_{n,k})b_{n+1,k}(n_i - r_{n+1,k}).$$

Hence $(n_i - r_{n,k})(n_j - r_{n+1,k}) = (n_j - r_{n,k})(n_i - r_{n+1,k})$, which implies that $n_i(r_{n,k} - r_{n+1,k}) = n_j(r_{n,k} - r_{n+1,k})$. In particular, since $n \geq 3$ both $n_i = 1$ and $n_j = 2$ are possible. Therefore $r_{n,k} = r_{n+1,k}$, namely $r_{n,k} = r_k$ for any $n \geq k$. Now, consider the following events: we observe the species X_i^* at the $(n+1)$-th draw and a new species at the $(n+2)$-th draw; ii) we observe a new species at the $(n+1)$-th draw and the species X_i^* at the $(n+2)$-th draw. As before, by exchangeability

$$\begin{aligned} &b_{n,k}(n_i - r_k)(1 - b_{n+1,k}(n+1) + kb_{n+1,k}r_k) \\ &\quad = (1 - b_{n,k}n + kb_{n,k}r_k)b_{n+1,k+1}(n_i - r_{k+1}) \end{aligned}$$

i.e.,

$$\begin{aligned} &n_i(b_{n,k}(1 - b_{n+1,k}(n+1) + kb_{n+1,k}r_k) - (1 - b_{n,k}n + kb_{n,k}r_k)b_{n+1,k+1}) \\ &\quad = b_{n,k}r_k(1 - b_{n+1,k}(n+1) + kb_{n+1,k}r_k) - r_{k+1}(1 - b_{n,k}n + kb_{n,k}r_k)b_{n+1,k+1}. \end{aligned}$$

Note that, since the right-hand side does not depend on the frequency n_i, then it must be

$$n_i(b_{n,k}(1 - b_{n+1,k}(n+1) + kb_{n+1,k}r_k) - (1 - b_{n,k}n + kb_{n,k}r_k)b_{n+1,k+1}) = 0 \tag{3.9}$$

and in turns

$$b_{n,k}r_k(1 - b_{n+1,k}(n+1) + kb_{n+1,k}r_k) - r_{k+1}(1 - b_{n,k}n + kb_{n,k}r_k)b_{n+1,k+1} = 0. \tag{3.10}$$

Now let us consider Equation (3.9) and Equation (3.10). In particular these equations lead to

$$\frac{b_{n+1,k+1}}{b_{n,k}} = \frac{(1 - b_{n+1}(n+1) + kb_{n+1,k}r_k)}{(1 - b_{n,k}n + kb_{n,k}r_k)} \tag{3.11}$$

and

$$\frac{r_{k+1}}{r_k} = \frac{b_{n,k}(1 - b_{n+1,k}(n+1) + kb_{n+1,k}r_k)}{b_{n+1,k+1}(1 - b_{n,k}n + kb_{n,k}r_k)}, \tag{3.12}$$

respectively. By combining Equation (3.11) with Equation (3.12) we obtain $r_{k+1}/r_k = 1$. Accordingly, r_k does not depend on k, and we denote by α the value of r_k. Finally, since $0 < (n - \alpha k)/((g(n,k)/b_{n,k}) + n - \alpha k) < 1$, then for any $n \geq 3$ and $3 \leq k \leq n$ it can be easily verified that $\alpha \in (0,1)$. For $n = 2$ we can simply choose $a_{2,2}$ and $b_{2,2}$ satisfying $-a_{2,2}/b_{2,2} = \alpha$ and $f(2,2,1) = a_{2,2} + b_{2,2}$. This completes the proof. □

Note that so far we considered $n \geq 1$ and $2 \leq l \leq n$. For the case $k = 1$ the only possible valued for f is $f(n, 1, n)$. In order to have a common notation it is convenient to define also the additional parameter $b_{n,1} = f(n, 1, n)/(n - \alpha)$. We can now introduce, for any $n \geq 1$ and $1 \leq k \leq n$, the new parameters defined as follows

$$V_{n,k} = \prod_{i=k}^{n-1} b_{i,k} \prod_{j=1}^{k-1} (1 - j b_{j,j}(1 - \alpha))$$

with the proviso $\prod_{m \leq l \leq m-1} = 1$ for any $m \geq 1$, which also implies that $V_{1,1} = 1$. Observe that $b_{n,k} = V_{n+1,k}/V_{n,k}$ and $f(n, k, n_i)$ and $g(n, k)$ can be written as follows

$$f(n, k, n_i) = \frac{V_{n+1,k}}{V_{n,k}}(n_i - \alpha), \tag{3.13}$$

and

$$g(n, k) = 1 - \frac{V_{n+1,k}}{V_{n,k}}(n - k\alpha). \tag{3.14}$$

We conclude by showing that for any $n \geq 1$ and $1 \leq k \leq n$ the $V_{n,k}$'s, with $V_{1,1} = 1$, satisfies the triangular recursion $V_{n,k} = V_{n+1,k}(n - \alpha k) + V_{n+1,k+1}$.

Proposition 3.3.5. *For all $n \in \mathbb{N}$ and for all $k \leq n$, the $V_{n,k}$ must satisfy the recursion $V_{n,k} = V_{n+1,k+1} + V_{n+1,k}(n - k\alpha)$.*

Proof. Fix k. We show by induction on n that the property holds $\forall n \geq k$.
For $n = k$, the property holds by simple algebraic calculations. Assume it holds for $n+1$, we show it holds for $n + 2$.
Let us consider the following events: i) we observe a new species at the $(n+1)$-th trial and the species X_i^* at the $(n+2)$-th trial; ii) we observe the species X_i^* at the $(n+1)$-th trial and a new species at the $(n+2)$-th trial. By means of the exchangeability assumption, the conditional probabilities of these events are the same,

$$\begin{aligned} &\left(1 - \frac{V_{n+1,k}}{V_{n,k}}(n - k\alpha)\right)\left(\frac{V_{n+2,k+1}}{V_{n+1,k+1}}(n_i - \alpha)\right) \\ &= \left(\frac{V_{n+1,k}}{V_{n,k}}(n_i - \alpha)\right)\left(1 - \frac{V_{n+2,k}}{V_{n+1,k}}(n + 1 - k\alpha)\right), \end{aligned} \tag{3.15}$$

which implies

$$\frac{V_{n+2,k+1}}{V_{n+1,k+1}} = \frac{V_{n+1,k} - V_{n+2,k}(n + 1 - \alpha k)}{V_{n,k} - V_{n+1,k}(n - \alpha k)} \tag{3.16}$$

and the two denominators are equal by the induction hypothesis. □

Finally, from (3.13) and (3.14) and using the recursion of Proposition 3.3.5, we obtain the predictive of the Gibbs-type priors

$$f(n,k,n_i) = \frac{V_{n+1,k}}{V_{n,k}}(n_i - \alpha) \tag{3.17}$$

$$g(n,k) = \frac{V_{n+1,k+1}}{V_{n,k}} \tag{3.18}$$

3.4 Urn scheme for Gibbs-type priors

Similarly to the Pitman–Yor process, one can derive an intuitive urn scheme that describes the predictive probabilities of Gibbs-type species sampling model with $\alpha \in (0,1)$. Let $(V_{n,k})_{1\leq k\leq n, n\geq 1}$ be a collection of nonnegative weights such that $V_{1,1} = 1$ and $V_{n,k} = V_{n+1,k}(n-\alpha k) + V_{n+1,k+1}$. We consider an urn that initially contains only a black ball with an arbitrary weight. Balls are drawn successively from the urn with probabilities proportional to their weights, and the drawing mechanism is described by the following Pólya-like urn scheme. Assuming that at the i-th draw the black ball has weight M, and that there are k distinct colors in the urn with weights $M_1, \ldots, M_k$, respectively, at the $(i+1)$-th draw:

i) if we pick a black ball, then it is returned to the urn together with a black ball of weight

$$B^*_{i+1} = M\frac{V_{i+2,k+2}V_{i+1,k}}{V_{i+2,k+1}V_{i+1,k+1}} - M, \tag{3.19}$$

and a ball of a new color with weight

$$A^*_{i+1} = (1-\alpha)M\frac{V_{i+1,k}}{V_{i+1,k+1}}; \tag{3.20}$$

ii) if we pick a non-black ball, then it is returned to the urn together with a black ball of weight

$$\tilde{B}_{i+1} = M\frac{V_{i+2,k+1}V_{i+1,k}}{V_{i+2,k}V_{i+1,k+1}} - M, \tag{3.21}$$

and an additional ball of the same color with weight

$$\tilde{A}_{i+1} = M\frac{V_{i+1,k}}{V_{i+1,k+1}}. \tag{3.22}$$

In the next theorem we show that the Pólya-like urn urn scheme described by i) and ii) provides samples from a Gibbs-type species sampling model with parameter $\alpha \in (0,1)$. Accordingly, such an urn scheme extends the two parameter Hoppe urn in Zabell [128] for describing the predictive probabilities of the Pitman–Yor process.

Theorem 3.4.1. *Let K_n be the number of non-black distinct colors after n draws from the above stated Pólya-like urn scheme, and let $N_{i,n}$ be the number of balls of color C_i, for any $i = 1, \ldots, K_n$. The distribution of K_n and $(N_{1,n}, \ldots, N_{K_n,n})$ coincides with the distribution of the random partition of $\{1, \ldots, n\}$ induces by a sample of size n from a Gibbs-type species sampling model with parameter $\alpha \in (0, 1)$.*

Proof. We start by proving (3.19) and (3.20). Let $\boldsymbol{X}_{n+1,k+1}$ be a sample from a Gibbs-type species sampling model with $\alpha \in (0, 1)$ and featuring $K_{n+1} = k + 1$ species with frequencies $(N_{1,n+1}, \ldots, N_{K_{n+1},n+1} N_{K_{n+1}+1,n+1}) = (n_1, \ldots, n_k, 1)$. Assume that after the n-th draw the black ball has weight M, and that there are k distinct colors, labelled by $C_1, \ldots, C_k$, with weights $M_1, \ldots, M_k$, respectively. We denote by M^* the updated weight of the black ball, and by M^*_{k+1} the weight of the ball with color C_{k+1}. Then,

i)
$$\mathbb{P}[X_{n+1} \text{ is a new species } | \boldsymbol{X}_{n,k}] = \frac{V_{n+1,k+1}}{V_{n,k}} = \frac{M}{M + \sum_{1 \le j \le k} M_j} \tag{3.23}$$

ii)
$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a new species } | \boldsymbol{X}_{n+1,k+1}] \\ &= \frac{V_{n+2,k+2}}{V_{n+1,k+1}} = \frac{M^*}{M^* + \sum_{1 \le j \le k} M_j + M^*_{k+1}}; \end{aligned} \tag{3.24}$$

iii) for any $j = 1, \ldots, k$
$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a species of type } C_j \,|\, \boldsymbol{X}_{n+1,k+1}] \\ &= \frac{V_{n+2,k+1}}{V_{n+1,k+1}}(n_j - \alpha) = \frac{M_j}{M^* + \sum_{1 \le j \le k} M_j + M^*_{k+1}}; \end{aligned} \tag{3.25}$$

iv)
$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a species of type } C_{k+1} \,|\, \boldsymbol{X}_{n+1,k+1}] \\ &= \frac{V_{n+2,k+1}}{V_{n+1,k+1}}(1 - \alpha) = \frac{M^*_{k+1}}{M^* + \sum_{1 \le j \le k} M_j + M^*_{k+1}}. \end{aligned} \tag{3.26}$$

According to (3.23), and because $V_{n,k} = V_{n+1,k}(n - \alpha k) + V_{n+1,k+1}$, one has $\sum_{1 \le j \le k} M_j = M(n - \alpha k) V_{n+1,k} / V_{n+1,k+1}$. By means of simple algebraic manipulations,

$$M^* = \frac{V_{n+2,k+2}}{(n + 1 - \alpha(k+1)) V_{n+2,k+1}} \left(M^*_{k+1} + M \frac{(n - \alpha k) V_{n+1,k}}{V_{n+1,k+1}} \right) \tag{3.27}$$

and

$$M^*_{k+1} = (1 - \alpha) M \frac{V_{n+1,k}}{V_{n+1,k+1}}. \tag{3.28}$$

Then, by combining the identity (3.27) with the identity (3.28), we obtain an expression for M^*. Specifically, $M^* = MV_{n+2,k+2}V_{n+1,k}/V_{n+2,k+1}V_{n+1,k+1}$. Finally, it can be easily verified that M^* and (3.28) satisfy (3.25) for any $j = 1, \ldots, k$. Accordingly, M^* and (3.28) are the solutions of the system of equations defined by (3.24), (3.25) and (3.26).

The proof of (3.21) and (3.22) is obtained by using the same arguments applied for (3.19) and (3.20). Specifically, let $\boldsymbol{X}_{n+1,k}$ be a sample from a Gibbs-type species sampling model with $\alpha \in (0,1)$ and featuring $K_{n+1} = k$ species with frequencies $(N_{1,n+1}, \ldots, N_{i,n+1}, \ldots, N_{K_{n+1},n+1}) = (n_1, \ldots, n_i + 1, \ldots, n_k)$. Assume that after the n-th draw the black ball has weight M, and that there are k distinct colors, labelled by $C_1, \ldots, C_k$, with weights $M_1, \ldots, M_k$, respectively. We denote by $\tilde{M}$ the updated weight of the black ball, and by $\tilde{M}_i$ the updated weight of the ball of color C_i. Then,

i) for any $1 \leq j \leq k$,

$$\begin{aligned} &\mathbb{P}[X_{n+1} \text{ is a species of type } C_j \,|\, \boldsymbol{X}_{n,k}] \\ &= \frac{V_{n+1,k}}{V_{n,k}}(n_j - \alpha) = \frac{M_j}{M + \sum_{1\leq j\leq k} M_j}, \end{aligned} \tag{3.29}$$

ii)

$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a new species } \,|\, \boldsymbol{X}_{n+1,k}] \\ &= \frac{V_{n+2,k+1}}{V_{n+1,k}} = \frac{\tilde{M}}{\tilde{M} + \tilde{M}_i + \sum_{1\leq j\neq i\leq k} M_j}; \end{aligned} \tag{3.30}$$

iii) for $1 \leq j \neq i \leq k$

$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a species of type } C_j \,|\, \boldsymbol{X}_{n+1,k}] \\ &= \frac{V_{n+2,k}}{V_{n+1,k}}(n_j - \alpha) = \frac{M_j}{\tilde{M} + \tilde{M}_i + \sum_{1\leq j\neq i\leq k} M_j}; \end{aligned} \tag{3.31}$$

iv)

$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a species of type } C_i \,|\, \boldsymbol{X}_{n+1,k}] \\ &= \frac{V_{n+2,k}}{V_{n+1,k}}(n_i + 1 - \alpha) = \frac{\tilde{M}_i}{\tilde{M} + \tilde{M}_i + \sum_{1\leq j\neq i\leq k} M_j}. \end{aligned} \tag{3.32}$$

According to (3.29), and because $V_{n,k} = V_{n+1,k}(n - \alpha k) + V_{n+1,k+1}$, one has $M_i = M(n_i - \alpha)V_{n+1,k}/V_{n+1,k+1}$. By simple algebraic manipulations, identities (3.30) and (3.32) lead to

$$\tilde{M} = \frac{V_{n+2,k+1}}{(n+1-\alpha k)V_{n+2,k}} \left(\tilde{M}_i + M(n - n_i - \alpha(k-1)) \frac{V_{n+1,k}}{V_{n+1,k+1}} \right) \tag{3.33}$$

and

$$\tilde{M}_i = (n_i + 1 - \alpha) M \frac{V_{n+1,k}}{V_{n+1,k+1}}. \tag{3.34}$$

Then, by combining the identity (3.33) with the identity (3.34), we can easily obtain an expression for $\tilde{M}$. Specifically, $\tilde{M} = M V_{n+2,k+1} V_{n+1,k} / V_{n+2,k} V_{n+1,k+1}$. Finally, it can be easily verified that $\tilde{M}$ and (3.34) satisfies (3.25) for any $j = 1, \ldots, k$. Accordingly, $\tilde{M}$ and (3.34) is the solution of the system of equations defined by (3.30), (3.31) and (3.32). □

3.5 Recovering the DP and the PY cases

The two parameter Hoppe Urn of Zabell [128] is recovered from (3.19), (3.20), (3.21) and (3.22) by setting $V_{n,k}$ of the form (2.10) and $M = \theta + k\alpha$. Note that under this assumptions the black ball is updated only when the black ball is drawn. Differently, in the above urn scheme the weight of the black ball is updated when the black ball is drawn (3.19) and also when a non-black ball is drawn (3.21). According to (3.19) and (3.21), in order to update the black ball only when the black ball is drawn we must assume

$$\frac{V_{n+2,k+2} V_{n+1,k}}{V_{n+2,k+1} V_{n+1,k+1}} \neq 1 \tag{3.35}$$

and

$$\frac{V_{n+2,k+1} V_{n+1,k}}{V_{n+2,k} V_{n+1,k+1}} = 1. \tag{3.36}$$

By means of (2.8), it can be easily verified that the assumptions (3.35) and (3.36) are equivalent to

$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a new species } | \, \boldsymbol{X}_{n+1,k+1}] \\ &\quad \neq \mathbb{P}[X_{n+2} \text{ is a new species } | \, \boldsymbol{X}_{n+1,k}] \end{aligned} \tag{3.37}$$

and, for $i = 1, \ldots, k$,

$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a species of type } C_i \, | \, \boldsymbol{X}_{n+1,k+1}] \\ &\quad = \mathbb{P}[X_{n+2} \text{ is a species of type } C_i \, | \, \boldsymbol{X}_{n+1,k}], \end{aligned} \tag{3.38}$$

respectively. Note that, by the Johnson's postulate in Zabell [128], for any $\alpha \in (0, 1)$ and $\theta > -\alpha$ the Pitman–Yor process is the unique species sampling model for which (3.37) and (3.38) hold true. The next proposition is a straightforward consequence of these remarks. We provide an alternative proof of this proposition, which does rely on Zabell's characterization.

Proposition 3.5.1. *The Pitman–Yor process is the unique Gibbs-type species sampling model for which the assumptions* (3.35) *and* (3.36) *hold true.*

Proof. For any $\alpha \in (0,1)$ and $\theta > -\alpha$, we show that the weight $V_{n,k} = \prod_{0\leq i\leq k-1}(\theta + i\alpha)/(\theta)_{(n)}$ is the only solution of (3.35) and (3.36). This is precisely the form of the $V_{n,k}$'s which gives the predictive probabilities of the Pitman–Yor process. Recall the constraints $V_{1,1} = 1$ and $V_{n,k} = V_{n+1,k}(n-\alpha k)+V_{n+1,k+1}$. By means of simple algebraic manipulations, (3.36) leads to

$$\frac{V_{n+2,k+2}V_{n+1,k}}{V_{n+2,k+1}V_{n+1,k+1}} = 1 + \alpha\frac{V_{n+2,k}}{V_{n+2,k+1}}, \tag{3.39}$$

i.e.,

$$\frac{V_{n+2,k+2}}{V_{n+2,k+1}} - \frac{V_{n+2,k+1}}{V_{n+2,k}} = \alpha. \tag{3.40}$$

For any fixed n, let us define $g_n(k) = V_{n,k+1}/V_{n,k}$ as a function of k. Hence (3.40) becomes the difference equation $g_{n+2}(k+1) - g_{n+2}(k) = \alpha$. Given the initial condition $c_{n+2} = g_{n+2}(1)$, where c_{n+2} is an arbitrary positive constant, the difference equation has solution

$$g_{n+2}(k) = c_{n+2} + (k-1)\alpha. \tag{3.41}$$

We show that c_{n+2} does not depend on n. We start by showing that $c_{n+2} = c_{n+3}$. Let us consider the condition (3.41) with n and k replaced by $n+1$ and $k+1$, respectively, i.e.

$$\frac{V_{n+3,k+3}V_{n+2,k+1}}{V_{n+3,k+2}V_{n+2,k+2}} = 1 + \alpha\frac{V_{n+3,k+1}}{V_{n+3,k+2}},$$

that is

$$\frac{g_{n+3}(k+2)}{g_{n+2}(k+1)} = 1 + \frac{\alpha}{g_{n+3}(k+1)}. \tag{3.42}$$

By replacing (3.41) into (3.42) we obtain $c_{n+3}^2 + c_{n+3}(\alpha(k+1) - c_{n+2}) - c_{n+2}\alpha(k+1) = 0$. The solution in c_{n+3} is $c_{n+3} = c_{n+2}$. Along the same lines we obtain $c_{n+i} = c_{n+i+1} := c$ for any $i > 2$. Hence, (3.41) can be written as $g_{n+2}(k) = c + (k-1)\alpha$, that is

$$V_{n+2,k+1} = (c + (k-1)\alpha)V_{n+2,k}. \tag{3.43}$$

Given the initial condition $V_{n,1} = b_n$, the difference equation (3.43) has solution $V_{n+2,k} = b_{n+2}\prod_{0\leq i\leq k-1}(c - \alpha + i\alpha)$. Furthermore, one can write the difference equation $b_{n+3} = b_{n+2}/(n+2+c-\alpha)$. Since $V_{1,1} = 1$ we have the initial condition $b_1 = 1$ and the solution of the difference equation is $b_n = 1/(c-\alpha)_n$. The proof is completed by setting $\theta = c-\alpha$. Indeed recall that c is positive while the range of θ is $\theta > -\alpha$. □

Let us consider the scenario in which the weight of the black ball is not updated when the black ball is drawn, and it is not updated when a non-black ball is drawn. In the Pólya-like urn scheme by Zabell [128], this scenario is obtained by letting $\alpha \to 0$. According to (3.19) and (3.21), in order to never update the black ball we must assume

$$\frac{V_{n+2,k+2}V_{n+1,k}}{V_{n+2,k+1}V_{n+1,k+1}} = 1 \tag{3.44}$$

and

$$\frac{V_{n+2,k+1}V_{n+1,k}}{V_{n+2,k}V_{n+1,k+1}} = 1. \tag{3.45}$$

By means of (2.8), it can be easily verified that the assumptions (3.35) and (3.36) are equivalent to

$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a new species } \mid \boldsymbol{X}_{n+1,k+1}] \\ &\quad = \mathbb{P}[X_{n+2} \text{ is a new species } \mid \boldsymbol{X}_{n+1,k}] \end{aligned} \tag{3.46}$$

and, for $i = 1, \ldots, k$,

$$\begin{aligned} &\mathbb{P}[X_{n+2} \text{ is a species of type } C_i \mid \boldsymbol{X}_{n+1,k+1}] \\ &\quad = \mathbb{P}[X_{n+2} \text{ is a species of type } C_i \mid \boldsymbol{X}_{n+1,k}], \end{aligned} \tag{3.47}$$

respectively. As before, due to the Johnson's postulate in Zabell [128], the Dirichlet process is the unique species sampling model satisfying (3.46) and (3.47). The next propositions is a direct consequence of these remarks. We also provide an alternative proof of this proposition, which does not rely on Zabell's characterization.

Proposition 3.5.2. *The Dirichlet process is the unique Gibbs-type species sampling model for which the assumptions* (3.44) *and* (3.45) *hold true.*

Proof. Recall that $V_{1,1} = 1$ and $V_{n,k} = V_{n+1,k}(n - \alpha k) + V_{n+1,k+1}$. Let (3.36) hold true, namely we do not update the weight of the black ball when a non-black ball is drawn. By means of simple algebraic manipulations, (3.36) leads to

$$\frac{V_{n+2,k+2}V_{n+1,k}}{V_{n+2,k+1}V_{n+1,k+1}} = 1 + \alpha\frac{V_{n+2,k}}{V_{n+2,k+1}}, \tag{3.48}$$

that is

$$\frac{V_{n+2,k+2}V_{n+1,k}}{V_{n+2,k+1}V_{n+1,k+1}} = 1$$

if $\alpha = 0$. Now, let (3.44) hold true, namely we do not update the weight of the black ball when a black ball is drawn. By means of simple algebraic manipulations, (3.44) leads to

$$\frac{V_{n+2,k+1}V_{n+1,k}}{V_{n+2,k}V_{n+1,k+1}} = \frac{n+1-\alpha k}{n+1-\alpha(k+1)}, \tag{3.49}$$

that is

$$\frac{V_{n+2,k+1}V_{n+1,k}}{V_{n+2,k}V_{n+1,k+1}} = 1$$

if $\alpha = 0$. By combining identities (3.48) and (3.49) it follows that the case $\alpha = 0$ is the only case, within the our Pólya-like urn scheme for Gibbs-type species sampling models, for which the weight of the black ball is never updated. The proof is completed. □

In our discussion thus far, we pointed out a relationship between: i) the dependency on k of the ratio $V_{n+1,k+1}/V_{n,k}$ and $V_{n+1,k}/V_{n,k}$, which appear in the predictive probabilities (2.8); ii) the updates of the black ball in the above Pólya-like urn scheme. According to Proposition 3.5.1 the weight of the black ball is updated only when the black-ball is drawn if and only $V_{n+1,k+1}/V_{n,k}$ depends on k and $V_{n+1,k}/V_{n,k}$ does not depend on k. The opposite scenario consists in updating the weight of the black ball when a non-black ball is drawn, and not updating it when the black ball is drawn. According to (3.19) and (3.21), this scenario is obtained by assuming

$$\frac{V_{n+2,k+2}V_{n+1,k}}{V_{n+2,k+1}V_{n+1,k+1}} = 1 \tag{3.50}$$

and

$$\frac{V_{n+2,k+1}V_{n+1,k}}{V_{n+2,k}V_{n+1,k+1}} \neq 1. \tag{3.51}$$

In the next proposition we show that the assumptions (3.50) and (3.51) implies that $\alpha = 0$. In other terms, the assumptions (3.50) and (3.51) provide a trivial scenario, in the sense that they lead to the scenario in which the weight of the black ball is never updated.

Proposition 3.5.3. *The Dirichlet process is the unique Gibbs-type species sampling model for which the assumptions* (3.50) *and* (3.51) *hold true.*

Proof. Recall that $V_{1,1} = 1$ and $V_{n,k} = V_{n+1,k}(n - \alpha k) + V_{n+1,k+1}$. Let (3.50) and (3.51) hold true, namely we do not update the weight of the black ball when a black ball is drawn and we update the weight of the black ball when a non-black ball is drawn. By means of simple algebraic manipulations, Equation (3.50) leads to

$$\frac{V_{n+2,k+1}V_{n+1,k}}{V_{n+2,k}V_{n+1,k+1}} = \frac{n+1-\alpha k}{n+1-\alpha(k+1)},$$

i.e.,

$$\frac{V_{n+2,k+1}}{V_{n+2,k}} = \frac{V_{n+2,k+2}}{V_{n+2,k+1}}\frac{n+1-\alpha k}{n+1-\alpha(k+1)}. \tag{3.52}$$

Let $g_{n+2}(k) = V_{n+2,k+1}/V_{n+2,k}$. From (3.52) one has the equation $g_{n+2}(k) = g_{n+2}(k+1)((n+1-\alpha k)/(n+1-\alpha(k+1)))$, with final condition $g_{n+2}(n+1) = V_{n+2,n+2}/V_{n+2,n+1} = a_{n+2}$. If we set $a = V_{2,2}/V_{2,1} > 0$, the solution of this equation corresponds to

$$g_{n+2}(k) = a\prod_{i=k}^{n}\frac{n+1-i\alpha}{n+1-\alpha(i-1)}$$

i.e.,

$$\frac{V_{n+2,k+1}}{V_{n+2,k}} = a\prod_{i=k}^{n}\frac{n+1-i\alpha}{n+1-\alpha(i+1)}.$$

Therefore, according to (3.49), for any $n \geq 1$ and $k \leq n$ the following identity must be satisfied

$$a \prod_{i=k}^{n} \frac{n+1-i\alpha}{n+1-\alpha(i+1)} = a \prod_{i=k+1}^{n+1} \frac{n+2-i\alpha}{n+2-\alpha(i-1)}. \tag{3.53}$$

The identity (3.53) is satisfied if and only if $\alpha = 0$. In particular if $\alpha = 0$ then we obtain the difference equation $V_{n+2,k+1} = aV_{n+2,k}$ with initial condition $V_{n+2,1} = b_{n+2}$. The solution of this difference equation is $V_{n+2,k} = b_{n+2}a^{k-1}$. Furthermore, one has the difference equation $b_{n+1} = b_n/(n+a)$ with initial condition $b_1 = 1$. The solution of this difference equation is $b_n = 1/(a+1)_{(n-1)}$. Hence, $V_{n,k} = a^k/(a)_n$. □

3.6 Discussion and future work

In this chapter, we introduced a nonparametric counterpart of the celebrated "sufficientness" postulate by Johnson [69]. Our postulate provides a characterization of a class of species sampling priors for which the probability of discovering at the next trial a new species depends only on the sample size n and the number k of observed species. For $k \geq 3$ these priors are precisely the Gibbs-type species sampling priors introduced in Gnedin and Pitman [48]. The proposed characterization extends previous results introduced in Zabell [125] and Zabell [129] for the Dirichlet process and two parameter Poisson-Dirichlet process, respectively. Following the parallel with the "sufficientness" postulates for the Dirichlet process and the two parameter Poisson-Dirichlet process, we paired our postulate with a simple Pólya-like urn scheme for describing the predictive probabilities of Gibbs-type species sampling priors. Such a scheme provides a novel and intuitive interpretation of these predictive probabilities in terms of the updates of a sequence of balls drawn for a Pólya-like urn. We find this interpretation particularly useful in order to explain differences between the Dirichelet process, the two parameter Poisson-Dirichlet process, and the more general class of Gibbs-type species sampling models.

The Pólya-like urn schemes for the Dirichlet process and the two-parameter Poisson-Dirichlet process are often applied in hierarchical constructions. See, e.g, Teh et al. [112], Teh and Jordan [113] and references therein. In the most basic example of these hierarchical constructions a collection of J exchangeable sequences, say $(X_{1,i})_{i\geq 1}, \ldots, (X_{J,i})_{i\geq 1}$, are drawn from the urn scheme independently; however, when a black ball is drawn in each case, the new color is selected from a latent exchangeable sequence, which is shared by J exchangeable sequences. Accordingly, the generalization of the usual exchangeable

Bayesian framework to nonparametric hierarchical modelling can be described as follows

$$\begin{aligned} X_{j,i} \,|\, P_j &\overset{iid}{\sim} P_j, \qquad i = 1, \ldots, n_j, j = 1, \ldots, J \\ P_j \,|\, P_0 &\sim \mu_j, \\ P_0 &\sim \mu_0. \end{aligned}$$

In other terms, the directing probability measure of $(X_{1,i})_{i\geq 1}, \ldots, (X_{J,i})_{i\geq 1}$ can be represented by introducing a random probability measure P_0 followed by the conditional independence of the sequence $(X_{1,i})_{i\geq 1}, \ldots, (X_{J,i})_{i\geq 1}$, with directing probability measure $P_1, \ldots, P_J$, respectively, centered on P_0. In most cases it is convenient to model $P_1, \ldots, P_J$ with Gibbs-type species sampling priors with mean P_0. This hierarchical procedure introduces dependence between the random partitions generated by $(X_{1,i})_{i\geq 1}, \ldots, (X_{J,i})_{i\geq 1}$, which is desirable in many applications.

While hierarchical species sampling priors had a tremendous impact on several applied fields, it still remains difficult to guide a selection of the prior distribution with subjective arguments, such as the number of species and their variability across populations. On the other had it also remains challenging the tuning of these hierarchical construction to optimize the performances of the resulting tools, quantified by classification and prediction error metrics. Our hope, and a motivation for our work, is that "sufficientness" postulates and urn schemes contribute to a better understanding and interpretability of hierarchical constructions that builds and combines layers of exchangeable random partitions. In particular the study of Gibbs-type exchangeable random partitions has the potential of contributing to the critical evaluation of hierarchical constructions for data analysis. When, for example, heterogeneous populations, say in ecology of microbiome studies, are modeled using dependent random partitions embedded in hierarchical constructions, how can we use the imputed layers of partitions generated through Markov chain Monte Carlo algorithms or other approaches to evaluate the construction of the prior model? When can we say that the use of hierarchical species sampling priors appears appropriate? Which type of assumption can we leverage on to tackle this type of problems? The theoretical understanding of random partitions, and characterization results for classes of random partitions, appears appropriate tools to allow the statistical and machine learning communities to approach these problems.

Consideration of hierarchical models defined by layers of species sampling models, such as the hierarchal Dirichlet and Pitman–Yor processes, raises the interesting problem of whether there is a "sufficientness" postulate that characterize the resulting model. If we assume the hierarchical structure described in the previous paragraph, and if we condition on variables that determine the steps in $(X_{1,i})_{i\geq 1}, \ldots, (X_{J,i})_{i\geq 1}$ in which a black ball was drawn, then it is not difficult to formulate "sufficientness" conditions that characterize the exchangeable sequences. This would amount to a direct application of the results of

this chapter. However, this set of conditions is not in the spirit of the Johnson "sufficientness" postulate because, first, the variables that determine when a black ball was drawn are not observable since the species observed at those steps are not necessarily "new", and second, unlike the exchangeability of a random partition the hierarchical structure assumed does not have an apparent subjective motivation. We also note that model interpretability, in this case, is provided by the overall probability construction, rather than by characteristics of the joint distribution of dependent random partitions, which in most cases presents analytic expression that are far from trivial. With the exception of the correlations between the random probabilities $P_1, \ldots, P_J$, results to quantify and understand the degree of dependence among $(X_{1,i})_{i\geq 1}, \ldots, (X_{J,i})_{i\geq 1}$ remains limited.

Chapter 4

Sample-size estimation for finding unseen species

4.1 Motivation

A difficult and important challenge in species sampling problems is the design of cost-effective species inventories. This is accounted for by concrete applied problems where the sampling procedure is expensive and, therefore, further samples can be only motivated by the possibility of recording a certain amount of new species. Sustaining an acceptable level of cost-effectiveness requires to redirect sampling to more productive sites, methods or time period, as the probability of discovering a certain amount of new species in additional samples becomes unacceptably low. Hence, one can fix a possibly small threshold for the discovery probability such that the sampling procedure takes place until the estimated discovery probability becomes for the first time smaller than the fixed threshold. This procedure introduces a criterion for evaluating the effectiveness of further sampling, as well as a tool for assessing survey completeness. A full range of statistical approaches, parametric and nonparametric as well as frequentist and Bayesian, have been proposed for estimating the probability of discovering new species in additional samples. Important contributions are, e.g., Good [50], Good and Toulmin [51], Efron and Thisted [34], Rasmussen and Starr [98], Mao [87], Lijoi et al. [76], Zhang and Stern [131] and Barger and Bunge [7].

While estimates of the discovery probability provide a criterion for evaluating the effectiveness of further sampling, they do not indicate how much sampling effort would be necessary to discover new species. Measuring the sampling effort is crucial in the design of cost-effective species inventories, since it provides guidance for ensuring an efficient use of time and financial resources during the sampling process. In this chapter we introduce a Bayesian nonparametric estimator of the minimum number of additional samples required for discovering new species. Specifically, under the framework of Gibbs-

type priors introduced in Lijoi et al. [76] for modeling the unknown species composition $(p_i)_{i\geq 1}$, we derive an explicit expression of the posterior distribution, with respect to an initial observed sample, of the minimum number of additional samples required to detect a preset amount of new species. The corresponding Bayesian nonparametric estimator under squared loss function is then obtained as the posterior expectation. This estimator thus provides practitioners with a realistic and easily interpretable expectation of what differences, in terms of new observable species, may be detected with a reasonable amount of time and cost. As an illustrative example we focus on the Pitman–Yor prior, which stands out for its mathematical tractability and, hence, represents the natural candidate within the class of Gibbs-type priors. Under this prior assumption we apply the proposed estimator to the analysis of an Expressed Sequence Tags (EST) dataset in genomics.

4.2 Methodology

In this chapter, we assume the vector $\mathbf{X}_n = (X_1, \ldots, X_n)$ of observations to come from an infinite exchangeable sequence of random variables. From De Finetti's Representation Theorem, we can write the model as

$$\begin{aligned} X_i \,|\, P &\overset{iid}{\sim} P, \quad i = 1, \ldots, n, \\ P &\sim \mu, \end{aligned} \tag{4.1}$$

where P is assumed to be a discrete distribution and can be rewritten as $P = \sum_{i\geq 1} p_i \delta_{X_i^*}$, where $(p_i)_{i\geq 1}$ is a collection of nonnegative random weights such that $\sum_{i\geq 1} p_i = 1$ almost surely, and $(X_i^*)_{i\geq 1}$ are random locations, or random labels, independent of $(p_i)_{i\geq 1}$ and independent and identically distributed according to a nonatomic probability measure P_0. The distribution of $(p_i)_{i\geq 1}$ and $(X_i^*)_{i\geq 1}$ is encoded in μ, which is the prior distribution of P. Following Lijoi et al. [76], in this Chapter, we select μ to be a Gibbs-type prior. Throughout this chapter we will use the same notation introduced in the previous chapters for the number of distinct values K_n in $(X_1, \ldots, X_n)$ and for their frequencies $(N_{1,n}, \ldots, N_{K_n,n})$.

Given the collection of exchangeable observations $(X_i)_{i\geq 1}$, directed by a Gibbs-type prior, hence having predictive probabilities of the form (2.8), we introduce a new sequence of random variables $(B_i)_{i\geq 1}$ defined as follows: $B_1 = 1$ and $B_i = \prod_{1\leq j\leq i-1} \mathbb{1}_{\{X_i \neq X_j\}}$ for any $i \geq 2$. In other terms $B_i = 1$ if X_i is a species that does not coincide with any of the species in $(X_1, \ldots, X_{i-1})$, whereas $B_i = 0$ if X_i is a species that coincides with some of the species in $(X_1, \ldots, X_{i-1})$. The binary sequence $(B_i)_{i\geq 1}$ thus provides information on the positions, or times, where new species appears for the first time in the sampling process $(X_i)_{i\geq 1}$. The distribution of $(B_i)_{i\geq 1}$ is determined by the conditional distribution of B_i given $(B_1, \ldots, B_{i-1})$. Specifically, if $b(n) := \sum_{1\leq i\leq n} b_i$ then by direct application of

(2.8) one has

$$\mathbb{P}[B_{n+1} = b \,|\, (B_1, \ldots, B_n) = (b_1, \ldots, b_n)] = \begin{cases} \frac{V_{n+1,b(n)+1}}{V_{n,b(n)}} & \text{if } b = 1 \\ \frac{V_{n+1,b(n)}}{V_{n,b(n)}}(n - \alpha b(n)) & \text{if } b = 0, \end{cases} \tag{4.2}$$

for any $n \geq 1$. Now, let us assume that the first $n \geq 1$ samples in $(X_i)_{i\geq 1}$ have been observed, and they features $K_n = k \leq n$ species with corresponding frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$. If

$$T_i = \min\{t : \sum_{j=1}^{t} B_{n+j} = k + i\}$$

then $W_0 = T_1 \geq 1$ is the waiting time for discovering, in additional samples, the first species that does not coincide with any of the species in the initial observed sample of size n. In general, the random variable $W_i = T_{i+1} - T_i \geq 1$ is the waiting time for discovering the $(i+1)$th new species in additional samples, for any $i \geq 0$. In the next lemma we derive an explicit expression for the conditional distribution of $(W_0, \ldots, W_\tau)$ given the observable sample $(X_1, \ldots, X_n)$, for any $\tau \geq 0$. This expression follows by a straightforward repeated application of the conditional probability (4.2).

Lemma 4.2.1. *Under the framework of Gibbs-type priors, let $(X_1, \ldots, X_n)$ be a random sample featuring $K_n = k$ species with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$. If $w(\tau) := \sum_{0\leq i\leq \tau} w_i$ then*

$$\begin{aligned} &\mathbb{P}[W_0 = w_0, \ldots, W_\tau = w_\tau \,|\, X_1, \ldots, X_n] \\ &= \frac{V_{n+w(\tau),k+\tau+1}}{V_{n,k}} \prod_{i=0}^{\tau} (n + w(i-1) - (k+i)\alpha)_{(w_i-1)\uparrow} \end{aligned} \tag{4.3}$$

The conditional distribution (4.3) takes on the natural interpretation of the joint posterior distribution, with respect to an initial observed sample $(X_1, \ldots, X_n)$, of the waiting times for discovering $\tau + 1$ new species in additional samples. Note that, as a direct consequence of the Gibbs-type prior assumption, K_n is a sufficient statistics for the posterior distribution (4.3). See De Blasi et al. [27] for details. Posterior distributions of various statistics of $(W_0, \ldots, W_\tau)$ may be obtained by a direct application of Lemma 4.2.1. Hereafter we consider two statistics of interest in the context of designing cost-effective species inventories:

i) the total waiting time $W(\tau) = \sum_{0\leq i\leq \tau} W_i \geq \tau + 1$ for discovering $\tau + 1$ new species in additional samples;

ii) the conditional waiting time $W_\tau \,|\, W(\tau - 1) \geq 1$ for discovering the $(\tau + 1)$th new species in additional samples given the total waiting time of the previously discovered new species.

For any $n \geq 1$, $1 \leq k \leq n$, $\alpha \in [0,1)$ and $\gamma < 0$ we denote by $\mathscr{C}(n,k;\alpha,\gamma)$ the (n,k)th non-centered generalized factorial coefficient, i.e. $\mathscr{C}(n,k;\alpha,\gamma) := (k!)^{-1}\sum_{0\leq i\leq k}(-1)^i\binom{k}{i}(-i\alpha-\gamma)_{n\uparrow}$ with the proviso $\mathscr{C}(0,0;\alpha,\gamma) := 1$, $\mathscr{C}(n,0;\alpha,\gamma) := 0$ for any $n \geq 1$. Furthermore, recall that $\lim_{\alpha\to 0}\alpha^{-k}\mathscr{C}(n,k;\alpha,\gamma) = |s(n,k;\gamma)|$ where $|s(n,k;\gamma)| := \sum_{k\leq i\leq n}\binom{n}{i}|s(i,k)|(-\gamma)_{(n-i)\uparrow}$ with $|s(i,k)|$ being the signless Stirling number of the first kind. See, e.g., Charalambides [21] for details. In the next theorem we provide an explicit and simple expression for the posterior distributions of $W(\tau)$ and $W_\tau \,|\, W(\tau-1)$, for any index $\tau \geq 0$. Note that, by definition, for $\tau = 0$ these two statistics coincide.

Theorem 4.2.2. *Under the framework of Gibbs-type priors, let $(X_1,\ldots,X_n)$ be a random sample featuring $K_n = k$ species with corresponding frequencies $(N_{1,n},\ldots,N_{K_n,n}) = (n_1,\ldots,n_k)$. Then*

i)

$$\mathbb{P}[W(\tau) = w \,|\, X_1,\ldots,X_n] = \frac{V_{n+w,k+\tau+1}}{V_{n,k}}\alpha^{-\tau}\mathscr{C}(w-1,\tau;\alpha,-n+k\alpha); \tag{4.4}$$

ii)

$$\begin{aligned}&\mathbb{P}[W_\tau = w_\tau \,|\, W(\tau-1) = w(\tau-1), X_1,\ldots,X_n] \\ &\quad = \frac{V_{n+w(\tau-1)+w_\tau,k+\tau+1}}{V_{n+w(\tau-1),k+\tau}}(n+w(\tau-1)-(k+\tau)\alpha)_{(w_\tau-1)\uparrow}.\end{aligned} \tag{4.5}$$

Expected values of the posterior distributions (4.4) and (4.5) provide Bayesian nonparametric estimators, with respect to a squared loss functions, of the statistics $W(\tau)$ and $W_\tau \,|\, W_{\tau-1}$, respectively. In order to compute explicitly these estimators one needs to specify an expression for the nonnative weight $V_{n,k}$. As an illustrative example we compute the Bayesian nonparametric estimators of $W(\tau)$ and $W_\tau \,|\, W_{\tau-1}$ under the assumption of the Pitman–Yor prior, i.e., $V_{n,k} = \prod_{1\leq i\leq k}(\theta+i\alpha)/(\theta)_{n\uparrow}$, for any $\alpha \in [0,1)$ and $\theta > -\alpha$. With regards to the posterior distribution (4.4), under the assumption of the Pitman–Yor prior one obtains

$$\mathbb{P}[W(\tau) = w \,|\, X_1,\ldots,X_n] = \frac{\alpha(k+\theta/\alpha)_{(\tau+1)\uparrow}}{(\theta+n)_{(w)\uparrow}}\mathscr{C}(w-1,\tau;\alpha,-n+k\alpha) \tag{4.6}$$

and

$$\begin{aligned}\hat{W}(\tau) &:= \mathbb{E}[W(\tau) \,|\, X_1,\ldots,X_n] \\ &= \begin{cases} 1-n-\theta+\frac{(\theta+n-1)(k+\theta/\alpha)_{(\tau+1)\uparrow}}{(k+(\theta-1)/\alpha)_{(\tau+1)\uparrow}} & \text{if } \theta+k\alpha > 1 \\ +\infty & \text{otherwise.}\end{cases}\end{aligned} \tag{4.7}$$

The Bayesian nonparametric estimator $\hat{W}(\tau)$ can be easily evaluated for any $\tau \geq 0$, and corresponding credible intervals may be determined via Monte Carlo sampling from the

posterior distribution (4.6). Of particular interest is $\hat{W}(0) = (\theta + n - 1)/(\theta + k\alpha - 1)$, which provides an estimate of the minimum number of additional samples required for detecting the first new species, given k species has been observed in the first n samples. This, in turns, introduces a direct measure of the sampling effort, in terms of time and cost, for finding a new species in additional samples.

Along similar lines one obtains an explicit expression for the Bayesian nonparametric estimator of $W_\tau \,|\, W(\tau-1)$. Under the Pitman–Yor prior, due to the product form of $V_{n,k}$, the posterior distribution of $W_\tau \,|\, W(\tau-1)$ may be expressed in terms of a mixture of negative Binomial distributions. This representation provides a useful tool for sampling from the posterior distribution of $W_\tau \,|\, W(\tau-1)$ and, hence, from the posterior distribution (4.6) with $\tau = 0$. Let $Z_{p,r}$ be a random variable distributed according to a negative Binomial distribution with success probability p and stopping parameter r, and let $f_{B_{a,b}}$ be the density function of a Beta random variable with parameter (a, b). Under the Pitman–Yor prior one obtains

$$\begin{aligned} &\mathbb{P}[W_\tau = w_\tau \,|\, W(\tau-1) = w(\tau-1), X_1, \ldots, X_n] \\ &\quad = (\theta + (\tau+k)\alpha) \frac{(n - (k+\tau)\alpha + w(\tau-1))_{(w_\tau - 1)\uparrow}}{(\theta + n + w(\tau-1))_{(w_\tau)\uparrow}} \\ &\quad = \int_0^1 \mathbb{P}[Z_{x,1} = w_\tau - 1] f_{B_{n-(k+\tau)\alpha + w(\tau-1), \theta + (k+\tau)\alpha}}(x) dx \end{aligned} \tag{4.8}$$

and

$$\begin{aligned} \hat{W}_\tau &:= \mathbb{E}[W_\tau \,|\, W(\tau-1) = w(\tau-1), X_1, \ldots, X_n] \\ &= \begin{cases} \frac{\theta + n + w(\tau-1) - 1}{\theta + (k+\tau)\alpha - 1} & \text{if } \theta + k\alpha > 1 \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{4.9}$$

Similarly to the total waiting time, the Bayesian nonparametric estimator $\hat{W}_\tau$ can be easily evaluated for any $\tau \geq 0$. Furthermore, corresponding credible intervals are determined via Monte Carlo sampling from the posterior distribution (4.8); this only requires to sample from a negative Binomial random variable and a Beta random variable. The estimator $\hat{W}_\tau$ provides information of how the wanting time for discovering new species increases as the number of discovered species increases, that is as τ increases. One may define $\hat{R}_{\tau,c} := (\hat{W}_\tau - \hat{W}_{\tau-c})/\hat{W}_{\tau-c}$, for some $c < \tau$, in order to measure the growth rate of the sampling effort for finding a new species in additional samples.

4.3 Illustration with Expressed Sequence Tags

Expressed Sequence Tags (ESTs) were introduced in Adams et al. [1], and they play a fundamental role for gene discovery and for characterizing expressed genes from a given

organism. ESTs are generated by sequencing randomly isolated gene transcripts that have been converted into cDNA. The resulting transcript sequences and their abundances are the main focus of interest providing the identification and level of expression of genes. As pointed out in Susko and Roger [108], sequencing is an expensive procedure, and typically it require to perform difficult "normalization" protocols on cDNA libraries before large numbers of ESTs are gathered from an organism. Hence, for an efficient use of time and financial resources, the decision of proceed with sequencing has to balance carefully between the probability of discovering new genes in further sequencing and the involved costs of such a sequencing. This decision is necessarily based on the estimation of the sampling effort for discovering new genes. The dataset we analyze consists of ESTs samples obtained from a cDNA library from Naegleria gruberi cells grown under different culture conditions, aerobic and anaerobic. See Susko and Roger [108] for details. Let m_i be the number of genes with frequency i. The Naegleria gruberi aerobic library consists of $n = 959$ ESTs with $k = 473$ genes and $m_i = 346, 57, 19, 12, 9, 5, 4, 2, 4, 5, 4, 1, 1, 1, 1, 1, 1,$ for $i = \{1, 2, \ldots, 12\} \cup \{16, 17, 18\} \cup \{27\} \cup \{55\}$. The Naegleria gruberi anaerobic library consists of $n = 969$ ESTs with $k = 631$ genes and $m_i = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1,$ for $i \in \{1, 2, \ldots, 13\}$.

We assume a Pitman–Yor prior. In order to implement the posterior distributions displayed in (4.6) and (4.8), as well as the corresponding estimators, the first issue to be faced is the specification of the parameter (α, θ). Hereafter, following the approach of Lijoi et al. [76], we resort to an empirical Bayes procedure. Specifically let $(X_1, \ldots, X_n)$ be a sample from a Pitman–Yor process and featuring $K_n = k$ species with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$. The empirical Bayes procedure consists in choosing θ and α that maximize the distribution of the sample $(X_1, \ldots, X_n)$. This corresponds to setting $(\alpha, \theta) = (\hat{\alpha}, \hat{\theta})$, where

$$(\hat{\alpha}, \hat{\theta}) = \arg\max_{(\alpha,\theta)} \left\{ \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)}{(\theta)_{n\uparrow}} \prod_{i=1}^{k} (1-\alpha)_{(n_i-1)\uparrow} \right\}. \tag{4.10}$$

Clearly one could also specify a prior distribution on the parameter (α, θ), and then seek a full Bayesian inference. However, as discussed in Lijoi et al. [76], there are no relevant differences between the fully Bayes approach and the empirical Bayes approach. Indeed it can be verified that for sufficiently large datasets, like the Naegleria gruberi libraries, the posterior distribution of (α, θ) is highly concentrated. See also De Blasi et al. [27] for details. The application of the empirical Bayes approach (4.10) lead to the following estimates for (α, θ): $(\hat{\alpha}, \hat{\theta}) = (0.669, 46.241)$ for the Naegleria gruberi aerobic library and $(\hat{\alpha}, \hat{\theta}) = (0.656, 155.408)$ for the Naegleria gruberi anaerobic library.

We first focus on the Naegleria gruberi aerobic library. Table 4.1 shows the Bayesian nonparametric estimates, as well as the corresponding 95% and 99% credible intervals, of the minimum number $W(\tau)$ of additional samples that are required to detect $\tau + 1$

new genes for $\tau = 0, 20, 40, 60, 80, 100$. Setting $n = 959$, $k = 473$, $\alpha = 0.669$ and $\theta = 46.241$, the estimates $\hat{W}(\tau)$ are obtained by a direct application of (4.7), whereas credible intervals are obtained by a Monte Carlo evaluation of the quantiles of the posterior distribution (4.6). Specifically, random variates from (4.6) can be easily obtained by means of the inverse transform sampling and using the fact that the factorial coefficient $\mathscr{C}(w-1, \tau; \alpha, -n+k\alpha)$ satisfies the following triangular recursion:

$$\begin{aligned}&\mathscr{C}(w-1, \tau; \alpha, -n+k\alpha)\\ &\quad = (w-2-\alpha\tau+n-k\alpha)\mathscr{C}(w-2, \tau; \alpha, -n+k\alpha) + \alpha\mathscr{C}(w-2, \tau-1; \alpha, -n+k\alpha).\end{aligned}$$

Estimates $\hat{W}(\tau)$ in Table 4.1 provide explicit measures of the sampling effort for the Naegleria gruberi aerobic library, and they can be directly applied in the process of designing cost-effective sequencing procedures. Table 4.1 also shows estimates of the minimum number $W_\tau \,|\, W(\tau)$ of additional samples that are required to detect the $(\tau+1)$th genes given the number $W(\tau)$ of samples used for discovering the previously genes. Estimates $\hat{W}_\tau$ and corresponding credible intervals are obtained from (4.9) and (4.8), respectively, where $w(\tau)$ is replaced by the estimates given by (4.7).

Table 4.1: Real data example: Naegleria gruberi aerobic library.

	$W(\tau)$			$W_{\tau+1} \,|\, W(\tau)$			
τ	$\hat{W}(\tau)$	95% C. I.	99% C. I.	$\hat{W}_{\tau+1}$	95% C. I.	99% C. I.	$\hat{R}_{\tau+1,20}$
0	2.7766	$(1, 9)$	$(1, 13)$	2.7792	$(1, 9)$	$(1, 13)$	–
20	58.8385	$(41, 82)$	$(36, 91)$	2.8294	$(1, 9)$	$(1, 13)$	0.0181
40	115.8966	$(88, 149)$	$(82, 161)$	2.8787	$(1, 9)$	$(1, 13)$	0.0174
60	173.9335	$(139, 215)$	$(130, 230)$	2.9272	$(1, 9)$	$(1, 13)$	0.0168
80	232.9325	$(191, 282)$	$(180, 299)$	2.9749	$(1, 10)$	$(1, 13)$	0.0163
100	292.8781	$(244, 349)$	$(231, 369)$	3.0218	$(1, 10)$	$(1, 14)$	0.0158

Figure 4.1 shows the estimates $\hat{W}_{\text{Aer}}$ and $\hat{W}_{\text{Anaer}}$ of $W(\tau)$ for the Naegleria gruberi aerobic library and for the Naegleria gruberi anaerobic library, respectively. Corresponding 95% credible intervals are also displayed. For each library we set the maximum value of τ as the observed number k of genes in the library, i.e., $\tau = 473$ for the aerobic library with sample size $n = 959$, and $\tau = 631$ for the anaerobic library with sample size $n = 969$. The estimated numbers of additional samples that are required to detect $\tau = 473$ and

$\tau = 631$ new genes are 1564.6353 and 1462.9275, respectively. In general, for any $\tau \geq 0$, the behaviour of the estimated sampling effort is apparent from Figure 4.1, in the sense that the anaerobic library presents a sampling effort that is lower than the aerobic library. In other terms the anaerobic library systematically produces more new genes along the sequencing process. This is in agreement with the analysis originally proposed in Susko and Roger [108] by means of the Good-Toulmin estimator for discovery probabilities.

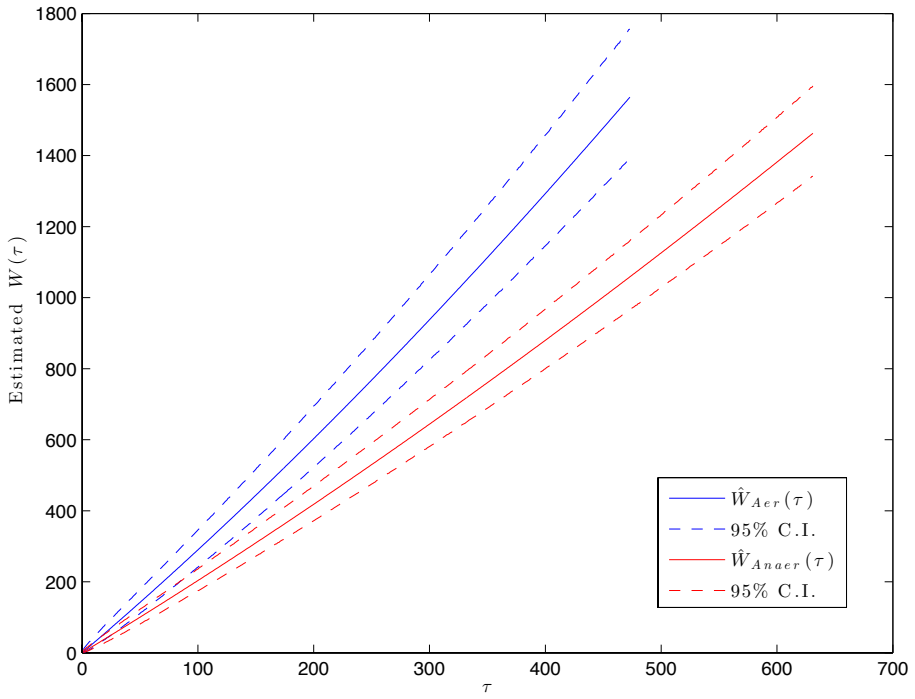

Figure 4.1: Real data example: Naegleria gruberi aerobic (blue) and anaerobic (red) libraries.

4.4 Discussion and future work

In this chapter we have introduced a Bayesian nonparametric analysis of the problem of designing cost-effective species inventories. This methodology is based on the use of Gibbs-type priors and provides estimates of the minimum number of additional samples required to detect a preset amount of new species. This information is useful to design species inventories and, at the same time, is easily interpretable and implementable by practitioners and researchers from other fields outside statistics. The proposed strategy is non-sequential, i.e. we assume a fixed sample of observations and, given this sample, our

approach directly estimates the sampling effort for discovering additional new species. A different approach which we think is worth exploring is a sequential one, in which observations arrive one at a time. Within this framework, we may proceed with the search for new species for as long as we wish, but there is a fixed cost, denoted by c, associated with each selection. At every time step we can decide to stop the search if the probability of discovering new species in additional samples becomes unacceptably low. One can therefore set a low threshold for the discovery probability, as a function of the fixed cost associated with each draw, and then proceed with the sampling process until the estimated discovery probability becomes first smaller than the fixed threshold. This sequential approach was first formalized in Rasmussen and Starr [98].

Rasmussen and Starr [98] studies this sequential problem by introducing an utility function h defined on the integers and such that, for each $k \in \mathbb{N}$, $h(k)$ is the gain of observing k distinct species. The pay-off of terminating the search after n observations have been collected is $w(n) = h(K_n) - cn$, and the goal is to find the stopping time s maximizing the expected pay-off, namely $\mathbb{E}[w(s)]$. Under the mild regularity condition that h is non-decreasing and concave, and under the assumption that the species composition $(p_i)_{i\geq 1}$ is known, Rasmussen and Starr [98] found the optimal solution of the above sequential problem. Let $\mathbf{X}_n$ be a sample from a population of individuals $(X_i)_{i\geq 1}$ belonging to an (ideally) infinite number of species $(X_i^*)_{i\geq 1}$ with known species proportions $(p_i)_{i\geq 1}$, and let $u(n) = \sum_{i\geq 1} p_i \mathbb{1}_{\{X_i^* \notin \mathbf{X}_n\}}$ be the so-called missing mass. Rasmussen and Starr [98] proved the optimality of the stopping rule defined as

$$s_{opt} = \inf\{n \geq 0 : [h(K_n + 1) - h(K_n)]u(n) \leq c\}. \tag{4.11}$$

Their proof strongly relies on the fact that monotonicity of $u(n)$ implies monotonicity of the stopping rule s_{opt}.

However, in species sampling problem, the p_i's are unknown and hence the stopping rule (4.11) is not computable because $u(n)$ is not observable. In order to overcome this drawback, Rasmussen and Starr [98] proposed an adaptive strategy in which the missing mass $u(n)$ in (4.11) is replaced by an estimator of it. As an estimator of the missing mass they considered the Good-Turing estimator, i.e.,

$$\tilde{u}(n) = \frac{1}{n} \sum_{i=1}^{K_n} \mathbb{1}_{\{N_{i,n}=1\}} \tag{4.12}$$

with $N_{i,n}$ being the frequency of the species X_i^* in $\mathbf{X}_n$. Replacing $u(n)$ in (4.11) with the estimator $\tilde{u}(n)$ has the side effect that the resulting adaptive stopping rule is no more monotone. The monotonicity of a stopping rule is a desired property for proving its optimality. See the monograph by Chow et al. [24] for a comprehensive discussion. Due to the non-monotonicity of their adaptive stopping rule, Rasmussen and Starr [98]

restricted themselves to show through empirical evidence that it compares well with the optimal one, for various choices of $(p_i)_{i\geq 1}$.

A Bayesian nonparametric counterpart of the adaptive stopping rule in Rasmussen and Starr [98] can obtained by replacing $u(n)$ in (4.11) with its Bayesian nonparametric counterpart. Specifically, under the model (4.1), with μ being a Gibbs-type prior, $\mathbf{X}_n$ a sample of observations and $A_n = \mathbb{X} - \{X_1^*, \ldots, X_{K_n}^*\}$ the set of unseen species, then the Bayesian nonparametric estimator of the probability of discovering new species at the $(n+1)$-the draw is

$$\mathbb{P}[X_{n+1} \in A_n \,|\, \mathbf{X}_n] = \mathrm{E}[P(A_n) \,|\, \mathbf{X}_n] = \frac{V_{n+1,k+1}}{V_{n,k}}. \tag{4.13}$$

The probability (4.13) provides a Bayesian nonparametric counterpart of the Good-Turing estimator (4.12).

In particular, under the special case of a Pitman–Yor prior with $\alpha \in [0,1)$ and $\theta > -\alpha$, the random variable $P(A_n) \,|\, \mathbf{X}_n$ is distributed according to a Beta distribution with parameter $(\theta+\alpha k, n-\alpha k)$, and hence $\mathbb{P}[X_{n+1} \in A_n \,|\, \mathbf{X}_n] = \mathrm{E}[P(A_n) \,|\, \mathbf{X}_n] = (\theta+\alpha k)/(\theta+n)$. Hence, a Bayesian nonparametric adaptive stopping based on the two parameter Poisson-Dirichlet prior is going to have the following form,

$$s_{bnp} = \inf\{n \geq 0 : [h(K_n+1) - h(K_n)]\frac{(\theta+\alpha k)}{(\theta+n)} \leq c\}. \tag{4.14}$$

A possible direction of research is to study Bayesian nonparametric procedures for sequential problems. Up to our knowledge, this field of research has not been explored yet. However, for the specific problem described in this section, we must remark that, as in the adaptive rule of Rasmussen and Starr [98], the stopping time (4.14) is not monotone. This fact can complicate the derivation of theoretical results and proving the optimality of the adaptive rule (4.14) can be a very challenging problem.

Chapter 5

Multi-armed bandit for species discovery

5.1 Description of the problem and proposed solution

In this chapter, we consider a different setting of the species sampling problem, in which J distinct populations are available and we can choose in a sequential manner from which of these populations to collect further samples. Specifically, let $(P_1, \ldots, P_J)$ denote J populations of animals of distinct regions. Each population P_j is assumed to contain a large number of species of animals, but both the represented species and their frequencies are unknown a priori. Also, the J populations are allowed to share the same species of animals and each species can have different frequencies in distinct regions. In this chapter we consider the problem of sequentially sampling these populations with the goal of maximizing the number of distinct species observed. In ecology and biology this problem arises naturally when different environments are explored in search of new species. In order for the exploration to remain cost-effective, redirecting it to a different environment is necessary whenever the probability of discovering a new species at the next draw becomes unacceptably low in the current environment. Similarly, in genetics the goal is to increase the number of genetic variants one expect to discover. See, e.g., Ionita-Laza et al. [63] and Ionita-Laza and Laird [64] where it has been shown that combining data from multiple populations in a discovery study increase the number of genetic variants identified relative to studies on single populations. Other applications arise in electrical engineering, in the context of security analysis of electrical power systems, and in software engineer, in the context of bugs discovery. See, e.g, Fonteneau-Belmudes et al. [41], Bubeck et al. [15] and references therein.

The framework of our species sampling problem resembles that of stochastic multi-

armed bandit (MAB) problems. These are problems in reinforcement learning which can be described using a gambling metaphor. We imagine a gambler facing J slot machines ("one armed bandit" is the colloquial term for a slot machine in American slang) with different unknown reward distribution functions. At every step, given the history of plays and realized rewards, the player can choose on which machine to play next and he will receive a random reward from the distribution of that arm. In the bandit literature the most common formalization of the problem is that of independent rewards from J unknown distributions. In this setting the two most popular sequential strategies are the Upper Confidence Bound (UCB) algorithm, introduced in Lai and Robbins [74] and further developed by Auer et al. [5], and the Thompson sampling (TS), proposed by Thompson [121]. The former solves the famous exploration-exploitation trade-off inherent in any bandit problem by constructing deterministic upper bounds for the expected reward in each arm, and then playing the arm with the highest bound value. The latter is a Bayesian algorithm which assigns priors to the unknown parameters and plays an arm according to its posterior probability of being the best one.

The problem analysed in this chapter can be traced back to a similar bandit formalization. In our setting, at every time step, a reward of one unit is received if a new species is observed and zero otherwise. Hence, rewards are Bernoulli distributed, but, differently from the usual MAB setting, in our case they are not assumed to be independent. Indeed every time a new species is observed the probability of observing another one in the next steps decreases. We propose a sequential rule that, given the information available up to a point, select the population from which to collect the next observation, with the goal of discovering as many distinct species as possible. Our sequential rule is made of two elements: a Bayesian nonparametric procedure for the estimation of the $(P_1, \ldots, P_J)$, seen as random discrete probability measures on a suitable space of species of animals, together with a TS strategy for the sequential choice of the best arm.

We consider a sample of observations $\mathbf{X_n} = (\mathbf{X}_{n_{1\cdot\cdot}}, \ldots, \mathbf{X}_{n_{J\cdot\cdot}})$ from the J populations, where $\mathbf{X}_{n_{j\cdot\cdot}} = (X_{j1}, \ldots, X_{jn_{j\cdot\cdot}})$ denotes the vector of observations from the j-th population, taking values on a measurable space $(\mathbb{X}, \mathcal{X})$, and with $n_{j\cdot\cdot}$ being the number of observations collected in this population. We assume the populations to be partially exchangeable. Hence, from De Finetti's Theorem for partially exchangeable sequences, their joint law can written in the following hierarchical form

$$\begin{aligned} X_{j,i}|P_j &\overset{iid}{\sim} P_j \qquad i = 1, \ldots, n_{j\cdot\cdot} \quad j = 1, \ldots, J \\ (P_1, \ldots, P_J) &\sim \mu, \end{aligned}$$

for a probability measure μ on $\mathcal{P}^J$, where $\mathcal{P}$ is the space of all distribution functions on $\mathbb{X}$. We choose a Hierarchical Pitman-Yor process as μ, i.e. $(P_1, \ldots, P_J)$ has the following

nonparametric prior

$$P_j|\alpha_j,\theta_j,P_0 \overset{ind}{\sim} \mathrm{PY}\left(\alpha_j,\theta_j,P_0\right) \qquad\qquad j=1,\ldots,J,$$
$$P_0|\alpha_0,\theta_0,H \sim \mathrm{PY}\left(\alpha_0,\theta_0,H\right).$$

for a fixed and diffuse probability measure H and $J+1$ couples of hyperparameters (α_j,θ_j) satisfying the conditions $1>\alpha_j\geq 0$ and $\theta_j>-\alpha_j$, for all $j\in\{0,\ldots,J\}$. This prior choice induces a prior also for the J mean parameters of the Bernoulli reward distributions. Given the induced prior and given a set of data from $(P_1,\ldots,P_J)$, we derive the corresponding posterior, which is then used to implement the TS strategy. We refer to the proposed strategy as HPY-TS. In addition, we also propose an extension of it, to deal with incidence data, in which animals are collected in groups.

We must also remark the connection of our strategy to adaptive sampling techniques. These are modifications of stratified random sampling, see e.g. Cochran [25] chapter 5, in which the choice of the sampling units are not fixed prior to making observations, but units are sequentially chosen, depending on previously observed values of some variable of interest. Theoretical advantages of adaptive selection designs were already pointed out in Basu [8] and Zacks [130] and can be remarkable, particularly when dealing with rare or elusive species. The first successful attempt in proposing an adaptive sampling procedure is Thompson [115], who proposes the adaptive cluster sampling. With this technique, biologists search for rare species of interest nearby locations on which the species was previously observed. Extensions and refinements of his work can be found in some his following works, e.g. Thompson [116], Thompson [118] and Thompson [117]. Good references for adaptive sampling techniques are Thompson [119] and Thompson and Seber [120]. Our algorithms have a similar flavor, but rather than focusing on re-observing a particular rare species, the goal is now to detect new ones.

We assess the performances of the proposed HPY-TS algorithms through simulations and using a dataset from biology. We compare the HPY-TS algorithms to other three strategies: an Oracle strategy in which the composition of the $(P_1,\ldots,P_J)$ are known; a Uniform strategy that selects at every step a population uniformly at random; a rule recently proposed in Bubeck et al. [15] based on the Good-Turing missing mass estimator introduced in Good [50]. Our simulation study considers different scenarios, by varying the level of heterogeneity in species variety among populations. Simulated results show that the HPY-TS performs better than the Uniform and the strategy of Bubeck et al. [15] in all scenarios, discovering more new species both in the abundance and in the incidence case. These results suggests also that the HPY-TS algorithms are robust to changes in the level of heterogeneity in species variety across the J populations, without the need of tuning parameters to regulate the exploration rate. We also compare the algorithms using a dataset of species of trees, collected in different plots in South America, analyzed in Pyke et al. [97] and Condit et al. [26].

To sum up, the rest of this chapter is organized as follows. In Section 5.2, we provide a brief account of the MAB problem and of TS in particular. Section 5.3 introduces the HPY-TS algorithm in the case of abundance and incidence data. This section includes also a subsection about computational issues to implement the algorithm. In Section 5.4, we present the results of simulations and the real data example. Finally, a discussion section about future work and an appendix containing further information about the simulated study close the chapter.

5.2 The Multi-Armed Bandit Problem and Thompson Sampling

A MAB problem is a sequential allocation problem under limited information. We imagine J slot machines to be available and, at every time step, a decision about which machine to play next has to be made. The goal is to maximize the expected pay-off. Inherent to this decision problem is the exploration-exploitation trade-off between exploiting machines that gave high profits in the past or exploring the ones not played yet. Here we focus on stochastic MABs only. An updated and detailed review on them is Bubeck and Cesa-Bianchi [13]. In the stochastic formalization, the J slot machines have unknown reward distributions functions and, at every time step, a draw from the distribution of the chosen machine is collected. A strategy is a sequential rule that, given the history up to that point, chooses the next arm to play. To evaluate its performances, its expected total reward is usually compared with that of an 'Oracle' strategy, the strategy that chooses the arm with the highest expected payoff, when uncertainness about the reward distributions is removed. The difference from their total rewards is termed regret. The goal is to find strategies that minimize the expected regret.

Two popular strategies have been shown to effectively address the stochastic bandit problem: the UCB algorithm and TS. The UCB algorithm was initially suggested by Lai and Robbins [74] and further developed by Auer et al. [5]. This algorithm constructs a deterministic upper confidence bound for the expected reward of each arm and then plays the arm with highest bound. This algorithm has good theoretical guarantees for the i.i.d. case: Auer et al. [5] proved that its expected regret matches, up to a constant factor, the lower bound of Lai and Robbins [74]. This is a lower bound for the expected regret of any strategy satisfying mild conditions, in the i.i.d. context. TS was initially proposed by Thompson [121] as a randomized Bayesian algorithm to minimize regret in a clinical trial setting. The idea is to assume a prior for the unknown parameters in the distributions of each arm and, at every time step, play an arm according to its posterior probability of being the best one. Its most popular application is for Bernoulli bandits. In this setting, a Bernoulli distribution with unknown parameter is set as reward distribution for each

arm, and the unknown reward means are endowed with a Beta prior distributions. TS thus consists in sampling a draw from each of these J Beta distributed posteriors and then play the arm with the highest realization.

Even though it was proposed eighty years ago, TS has attracted attention only recently. Several recent studies have empirically demonstrated the efficacy of TS. Chapelle and Li [20] have empirically demonstrated that TS achieves regret comparable to the lower bound of Lai and Robbins [74]. In addition, the algorithm is more robust to delayed or batched feedback than other methods. Chapelle and Li [20] also show that TS performs equally or better of popular methods, such as UCB algorithms, in applications like display advertising and news article recommendation. Other empirical works on TS or randomized probability matching algorithm (to which TS belongs) are Granmo [53], Scott [105] and May and Leslie [86]. Theoretical investigations of the TS are in Agrawal and Goyal [2], Kaufmann et al. [70], Russo and Van Roy [103] and Szabo and Tran-Thanh [109]. A recent promising theoretical result for TS is in Russo and Van Roy [104], where the authors provide Baysian regret bounds for a broad range of on-line optimization algorithms, with TS being a particular case. Their bounds are derived using an information theoretic approach and they depend on the entropy of the prior distribution of the optimal arm. Their results are an improvement and a generalization of previous theoretical studies on TS.

5.3 HPY-TS algorithm

The problem of sequential species discovery in presence of many populations, can be traced back to a stochastic bandit problem by regarding a discovery as a unitary reward. We can think of each population to produce a random reward when sampled: one if a new value is observed, zero otherwise. Hence, the reward distribution of each population is Bernoulli as in the Bernoulli bandit problem, mentioned above. However, differently from this latter problem, in the species framework the J unknown means are not constant. Instead, they are non-increasing functions of the number of draws, because, every time a new species is observed, the remaining missing mass will be lower in following time steps. In subsection 3.1, we will derive the distribution the joint posterior of these J Bernoulli means and we will introduce the HPY-TS algorithm in the case of abundance data. In subsection 3.2, we propose an extension of this algorithm to deal with incidence data.

5.3.1 Abundance Data

In the abundance data scenario, a single animal is observed at a time. Given a model choice, TS draws values for each population from their posterior distributions of being the best arm. As already pointed out, in the species problem with many popula-

tions, this posterior is the joint distribution of the J random probabilities of observing a new value in each arm, given all observations. This joint distribution is derived in Proposition 5.3.1, in the case of a HPY prior for $(P_1, \ldots, P_J)$. In particular, denoting with $\mathbf{X}_{n_{j\cdot}} = \left(X_{j1}, \ldots, X_{jn_{j\cdot}}\right)^T$ the vector of observations from population j, with $\mathbf{X_n} = (\mathbf{X}_{n_{1\cdot}}, \ldots, \mathbf{X}_{n_{J\cdot}})$ the joint sample (the array containing observations from all populations) and with $A = \{x \in \mathbb{X} : x \notin \mathbf{X_n}\}$ the set of possible new species, what is needed is the joint distribution of

$$(P_1(A), \ldots, P_J(A))|\mathbf{X_n}, \sigma_1, \ldots, \alpha_j, \theta_1, \ldots, \theta_J, \alpha_0, \theta_0, H.$$

For ease of notation, from now on we omit the reference to the hyperparameters of the HPY, $\alpha_j, \theta_j, \alpha_0, \theta_0, H$ when conditioning on them. The density of this joint distribution is provided in the following proposition, whose proof is postponed to Appendix A. In its statement, we adopt the notation for table counts and distinct values previously introduced for the Chinese Franchise Representation of the HPY. Also, beta $(p|a, b)$ stands for a beta density function with parameters a and b, evaluated at p.

Proposition 5.3.1. *Let $\mathbf{X_n}$ denote the joined sample from a HPY and let $A = \{x \in \mathcal{X} : x \notin \mathbf{X_n}\}$. Then, $(P_1(A), \ldots, P_J(A))|\mathbf{X_n}$ admits the following multivariate density*

$$f_{(P_1(A),\ldots,P_J(A))|\mathbf{X_n}}(p_1, \ldots p_J) = \int_0^1 \prod_{j=1}^{J} f_j(p_j|\beta_0, m_{j\cdot}, n_{j\cdot}) \cdot f_0(\beta_0|K, m_{\cdot\cdot})\, d\beta_0,$$

where

$$f_j(p_j|\beta_0, m_{j\cdot}, n_{j\cdot}) = beta(p_j|(\theta_j + m_{j\cdot}\alpha_j) \cdot \beta_0, (\theta_j + m_{j\cdot}\alpha_j) \cdot (1 - \beta_0) + n_{j\cdot} - \alpha_j \cdot m_{j\cdot})$$

and

$$f_0(\beta_0|K, m_{\cdot\cdot}) = beta(\beta_0|\theta_0 + K\alpha_0, m_{\cdot\cdot} - \alpha_0 K).$$

Proof. From formula (2.16), the franchise-wide distinct values $(X_1^{**}, \ldots, X_K^{**})$ are governed by P_0 and $P_0 \sim \text{PY}(\alpha_0, \theta_0, H)$. Using formula (2.2), the posterior distribution of P_0, given the observations, satisfies the distributional equation

$$P_0|\mathbf{X_n} \stackrel{d}{=} \sum_{k=1}^{K} \beta_k \cdot \delta_{X_k^{**}} + \beta_0 \cdot P_0',$$

where

$$P_0'|\mathbf{X_n} \sim \text{PY}(\alpha_0, \theta_0 + K\alpha_0, H)$$
$$\beta|\mathbf{X_n} = (\beta_0, \ldots, \beta_K)|\mathbf{X_n} \sim \text{Dir}(\theta_0 + K\alpha_0, m_{\cdot 1} - \alpha_0, \ldots, m_{\cdot K} - \alpha_0).$$

Similarly, from formula (2.15), we can apply formula (2.2) to P_j to find a distributional equation for P_j, conditionally on P_0 and the data. Also, using the distributional equation for the posterior of P_0, we find the following distributional equation for P_j

$$P_j|\beta, P_0^{'}, \mathbf{X_n} \stackrel{d}{=} \sum_{k=1}^{K} \pi_{j,k} \cdot \delta_{X_k^{**}} + \pi_{j,0} \cdot P_j^{'}, \tag{5.1}$$

where

$$\begin{gathered} P_j^{'}|P_0^{'}, \mathbf{X_n} \sim \mathrm{PY}(\alpha_j, (\theta_j + m_{j\cdot}\alpha_j) \cdot \beta_0, P_0^{'}) \\ (\pi_{j,0}, \ldots, \pi_{j,K}) \,|\beta, \mathbf{X_n} \sim \mathrm{Dir}((\theta_j + m_{j\cdot}\alpha_j) \cdot \beta_0, (\theta_j + m_{j\cdot}\alpha_j) \cdot \beta_1 + n_{j\cdot 1} - \alpha_j \cdot m_{j1}, \ldots \\ \ldots, (\theta_j + m_{j\cdot}\alpha_j) \cdot \beta_K + n_{j\cdot K} - \alpha_j \cdot m_{jK}). \end{gathered}$$

So, the distribution of $P_j(A)\,|\mathbf{Y_n}, P_0$ satisfies

$$P_j(A)\,|\beta, P_0^{'}, \mathbf{X_n} \stackrel{d}{=} \sum_{k=1}^{K} \pi_{j,k} \cdot \delta_{X_k^{**}}(A) + \pi_{j,0} \cdot P_j^{'}(A),$$

for all $j \in \{1, \ldots, J\}$, which implies

$$P_j(A)\,|\beta_0, \mathbf{X_n} \sim \mathrm{beta}\left((\theta_j + m_{j\cdot}\alpha_j) \cdot \beta_0, (\theta_j + m_{j\cdot}\alpha_j) \cdot (1-\beta_0) + n_{j\cdot\cdot} - \alpha_j \cdot m_{j\cdot}\right),$$

where we made use of the following facts:

1. $\delta_{X_k^{**}}(A) = 0 \;\forall k = 1, \ldots, K$: since $(X_1^{**}, \ldots, X_K^{**}) = A^c$.
2. $P_j^{'}(A) \stackrel{as}{=} 1$: $P_j^{'}$ can be rewritten as $P_j^{'} = \sum_{i\geq 1} \psi_i \cdot \delta_{Y_i}$ for some weights $\{\psi_i\}_{i\geq 1}$ and atoms $\{Y_i\}_{i\geq 1} \stackrel{iid}{\sim} H$. Then, $\mathbb{P}\left(\cap_{i\geq 1}\{Y_i \in A\}\right) = \prod_{i\geq 1}\mathbb{P}(Y_i \in A) = \prod_{i\geq 1} 1 = 1$, since H is diffuse and A^c is a finite set of points. Finally, $\mathbb{P}\left(\cap_{i\geq 1}\{Y_i \in A\}\right) = 1 \Rightarrow P_j^{'}(A) \stackrel{as}{=} 1$.
3. $\pi_{j,0}|\beta_0, \mathbf{X_n} \sim \mathrm{beta}\left((\theta_j + m_{j\cdot}\alpha_j)\,\beta_0, (\theta_j + m_{j\cdot}\alpha_j)\,(1-\beta_0) + n_{j\cdot\cdot} - \alpha_j m_{j\cdot}\right)$: by the aggregation property of Dirichlet distribution.

Also, since we are conditioning on P_0 (through $\beta, P_0^{'}$), $P_j(A)\,|\beta_0, \mathbf{X_n}$ is independent of $P_i(A)\,|\beta_0, \mathbf{X_n}$ $\forall i, j \in \{1, \ldots J\}$, $i \neq j$. Hence, their joint distribution is simply the product of the marginals. The last step is to integrate β_0 out to find the joint density

$$f_{(P_1(A),\ldots,P_J(A))|\mathbf{X_n}}(p_1, \ldots, p_J) = \int_0^1 \prod_{j=1}^{J} f_{P_j(A)|\beta_0,\mathbf{X_n}}(p_j) \cdot dF_{\beta_0}(\beta_0),$$

where the distribution of β_0 is another beta (again by aggregation of Dirichlet distribution). So, $(P_1(A), \ldots, P_J(A))|\mathbf{X_n}$ admits a density as stated. □

The following corollary provides a Bayesian nonparametric point estimate of the missing mass for each populations. This result follows by a direct appellation of Proposition 5.3.1.

Corollary 5.3.2. *Under squared loss function, the Bayesian nonparametric point estimate for the probability of discovering a new value in population j,given the joined sample $\mathbf{X_n}$, is*

$$\mathbb{E}\left[P_j\left(A\right)|\mathbf{X_n}\right] = \left(\frac{\theta_j + m_{j\cdot}\alpha_j}{\theta_j + n_{j\cdot\cdot}}\right)\left(\frac{\theta_0 + K\alpha_0}{\theta_0 + m_{\cdot\cdot}}\right).$$

With the distribution of Proposition 5.3.1 at hand, HPY-TS prescribes to sample a draw from it and to select the population with the highest realized value. This strategy outperforms the greedy one that selects the arm with the highest posterior point estimate, $j^{greedy} = \text{argmax}\{\mathbb{E}\left[P_j\left(A\right)|\mathbf{X_n}\right] : j \in \{1,\dots,J\}\}$, since it better balances the exploration step. Intuitively, suppose to have only a few observations, with an unlucky sample, from a 'winning' arm (a population with a very high species variety), resulting in a low point estimate for its missing mass. This population will not be chosen by the greedy strategy, which only exploits arms with good past behavior. Whereas, with HPY-TS strategy, the posterior distribution of the missing mass of this population will have high variability, due to the small sample size. This implies a positive probability for that arm to be chosen, if its Thompson draw results in a high value. The HPY-TS strategy for abundance data is summarized in Algorithm 1.

Algorithm 1: (*HPY-TS - Abundance Data*)
for *i in 1:additional sample* **do**
  draw $\beta_0 \sim \text{beta}\left(\theta_0 + K\alpha_0, m_{\cdot\cdot} - \alpha_0 K\right)$;
  for *j in 1:J* **do**
    draw $p_j \sim \text{beta}\left(\left(\theta_j + m_{j\cdot}\alpha_j\right)\beta_0, \left(\theta_j + m_{j\cdot}\alpha_j\right)\left(1-\beta_0\right) + n_{j\cdot\cdot} - \alpha_j m_{j\cdot}\right)$;
  end
  Compute $j^* = \text{argmax}\{p_j : j \in \{1,\dots J\}\}$;
  Sample the next observation from population j^*;
  Update table counts and estimates of the HPY hyperparameters;
end

Note that in Algorithm 1 the parameters of the beta distributions depend on the counts in the Chinese Franchise Representation of the HPY process. In particular, they depend on the number of observations for each population $(n_{j\cdot\cdot} : j \in \{1,\dots,J\})$, the number of clusters in each population $(m_{j\cdot} : j \in \{1,\dots,J\})$ and the total number of distinct species observed in the joint sample, K. We must remark that the collection clusters counts, $(m_{j\cdot} : j \in \{1,\dots,J\})$, are latent variables, namely they are not directly observed. Hence,

if an initial sample is available, we must estimate these components before running the algorithm. In subsection 5.3.3, we describe a MCMC procedure to handle this problem, together with the problem of inferring the hyperparameters $((\alpha_j, \theta_j) : j \in \{0, \dots, J\})$, in case they are treated as unknown components.

5.3.2 Incidence Data

There are applications where we cannot sample an animal at a time. Instead, multiple individuals are jointly collected in the sample. In these situations an extension of Algorithm 1 is needed. Suppose that, as in the case of abundance data, at every time step we can choose the next population to sample from, but now, instead of one animal, a collection l animals are observed from that population. In such a context, what must be maximized is the expected number of new distinct values observed in a additional sample of size l. In particular, given the array of data $\mathbf{X_n}$, let us denote by $K_j^{(l)}|\mathbf{X_n}$ the random number of new distinct species observed in a new sample of size l, collected from population j. With new distinct species, we mean species that are observed in the additional sample, but which were not previously observed in any of the J populations. In such a context, the reward distribution for arm j is the distribution of $\mathbb{E}[K_j^{(l)}|\mathbf{X_n}]$. Note that $\mathbb{E}[K_j^{(l)}|\mathbf{X_n}]$ is a random variable (since P_j is random), but if conditioned to $P_j|\mathbf{X_n}$, it becomes a number. Another remark is that, when $l = 1$, we are back to the abundance case. In fact, $\mathbb{E}[K_j^{(1)}|\mathbf{X_n}] = \mathbb{E}[I\left(X_{n_{j\cdot\cdot}+1} \in A\right)|\mathbf{X_n}] = P_j\left(A\right)|\mathbf{X_n}$, where I is the indicator function and $A = \{x \in \mathbb{X} : x \notin \mathbf{X_n}\}$. In Proposition 5.3.3, we derive the distribution of $(\mathbb{E}[K_1^{(l)}|\mathbf{X_n}], \dots, \mathbb{E}[K_J^{(l)}|\mathbf{X_n}])$.

Proposition 5.3.3. *Conditionally to*

$$\beta_0|\mathbf{X_n} \sim beta\left(\beta_0|\theta_0 + K\alpha_0, m_{\cdot\cdot} - \alpha_0 K\right)$$

and to $P_j\left(A\right)|\mathbf{X_n}, \beta_0 = p_j$, *where*

$$P_j\left(A\right)|\mathbf{X_n}, \beta_0 \sim beta\left(p_j|\left(\theta_j + m_{j\cdot}\alpha_j\right)\beta_0, \left(\theta_j + m_{j\cdot}\alpha_j\right)\left(1-\beta_0\right) + n_{j\cdot\cdot} - \alpha_j m_{j\cdot}\right),$$

$\mathbb{E}[K_j^{(l)}|\mathbf{X_n}]$ *is a constant, independent of the other arms and* $\mathbb{E}[K_j^{(l)}|\mathbf{X_n}, \beta_0, p_j]$ *can be computed as*

$$\sum_{k=0}^{l} k \cdot \sum_{i=k}^{l} \binom{l}{i} \cdot p_j^i \cdot (1-p_j)^{l-i} \sum_{\tilde{m}=k}^{i} F\left(\tilde{m}, k, \alpha_0, \theta_0 + K\alpha_0\right) \cdot F\left(i, \tilde{m}, \alpha_j, \left(\theta_j + m_{j\cdot}\right)\beta_0\right),$$

where the function $F\left(n, k, \alpha, \theta\right)$ *is the probability of having* k *distinct values in a sample of size* n *sampled from a Pitman-Yor process with hyperparameters* (α, θ), *which, from formula* (2.4), *has form*

$$F\left(n, k, \alpha, \theta\right) = \frac{\prod_{r=1}^{k-1}\left(\theta + r\cdot\alpha\right)}{\alpha^k\left(\theta+1\right)_{n-1}} \mathscr{C}(n, k; \alpha),$$

where $\mathscr{C}(n,k;\alpha)$ denotes the generalized factorial coefficient.

Proof. Using the distributional equation (5.1) for the posterior of P_j and working conditionally on $\beta_0|\mathbf{X_n} \sim beta\,(\theta_0 + K\alpha_0, m_{\cdot\cdot} - \alpha_0 K)$, we compute $P(K_j^{(l)} = k|\mathbf{X_n}, \beta_0)$.
From the distributional equation, we know that, given

$$\pi_{j,0}|\beta_0, \mathbf{X_n} \sim \text{beta}\left((\theta_j + m_{j\cdot}\alpha_j)\cdot\beta_0, (\theta_j + m_{j\cdot}\alpha_j)\cdot(1-\beta_0) + n_{j\cdot\cdot} - \alpha_j\cdot m_{j\cdot}\right),$$

an observation $X_{n_{j\cdot\cdot}+i}$ with $i = 1,\dots,l$ does not coincide with any of the K distinct species (in the joint sample) with probability $\pi_{j,0}$. To have $K_j^{(l)} = k$, at least k of the l data $X_{n_{j\cdot\cdot}+1},\dots,X_{n_{j\cdot\cdot}+l}$ must be allocated to the k new distinct species that have not previously observed. Hence,

$$\mathbb{P}(K_j^{(l)} = k|\mathbf{X_n}, \beta_0, \pi_{j,0}) = \sum_{i=k}^{l}\binom{l}{i}\cdot {\pi_{j,0}}^i\cdot(1-\pi_{j,0})^{l-i}\cdot\mathbb{P}\left(K_i = k|\beta_0\right),$$

where K_i is now the number of distinct species in a sample of size i generated by a $\text{PY}(\alpha_j, (\theta_j + m_{j\cdot}\alpha_j)\cdot\beta_0, P_0')$, where $P_0' \sim \text{PY}\left(\alpha_0, \theta_0 + K\alpha_0, H\right)$.
We need to find $\mathbb{P}\left(K_i = k|\beta_0\right)$. Using the Chinese Franchise Representation and the result by Gnedin and Pitman [48], denoting by M_i the number of tables, we have that, for $\tilde{m} = 1,\dots,i$, $\mathbb{P}(M_i = \tilde{m}) = F(i,\tilde{m},\alpha_j,(\theta_j + m_{j\cdot}\alpha_j)\,\beta_0)$. Moreover, conditionally on $M_i = \tilde{m}$, for $k = 1,\dots,\tilde{m}$, $\mathbb{P}(K_i = k|M_i = \tilde{m}) = F(\tilde{m},k,\alpha_0,\theta_0 + K\alpha_0)$. Finally, $\mathbb{P}(K_j^{(l)} = k|\mathbf{X_n},\beta_0,\pi_{j,0})$ can be computed as

$$\sum_{i=k}^{l}\binom{l}{i}\cdot{\pi_{j,0}}^i\cdot(1-\pi_{j,0})^{l-i}\sum_{\tilde{m}=k}^{i}F\left(\tilde{m},k,\alpha_0,\theta_0+K\alpha_0\right)\cdot F\left(i,\tilde{m},\alpha_j,(\theta_j+m_{j\cdot}\alpha_j)\,\beta_0\right).$$

The conditional mean $\mathbb{E}(K_j^{(l)}|\mathbf{X_n},\beta_0,\pi_{j,0})$ is found by averaging over $\{0,\dots,l\}$ and, being constant, they are trivially independent among arms. Hence, the joint distribution of $(\mathbb{E}(K_1^{(l)}|\mathbf{X_n}),\dots,\mathbb{E}(K_J^{(l)}|\mathbf{X_n}))$ is found by integrating $\beta_0,(\pi_{j,0} : j \in \{1,\dots J\})$ out from the product of these J conditional (constant) distributions. □

HPY-TS for incidence data, prescribes to sample a draw from the joint distribution of $\{\mathbb{E}[K_1^{(l)}|\mathbf{X_n}],\dots,\mathbb{E}[K_J^{(l)}|\mathbf{X_n}]\}$ and to choose the population corresponding to the highest value. A schematic description of the algorithm for incidence data is summarized in Algorithm 2.

Algorithm 2: (*HPY-TS - Incidence Data*)
for *i in 1: number of new samples* **do**
  fix l equal to its posterior estimate ;
  draw $\beta_0 \sim \text{beta}\,(\theta_0 + K\alpha_0, m_{\cdot\cdot} - \alpha_0 K)$;
  for *j in 1:J* **do**
    draw $p_j \sim \text{beta}\,((\theta_j + m_{j\cdot}\alpha_j)\cdot\beta_0, (\theta_j + m_{j\cdot}\alpha_j)\cdot(1-\beta_0) + n_{j\cdot\cdot} - \alpha_j \cdot m_{j\cdot})$;
    compute $\mathbb{E}[K_j^{(l)}|\mathbf{X_n}, \beta_0, p_j]$ as in Proposition 5.3.3 ;
  end
  Compute $j^* = \text{argmax}\{\mathbb{E}[K_j^{(l)}|\mathbf{X_n}, \beta_0, p_j] : j \in \{1, \ldots, J\}\}$;
  Sample the next group of observations from population j^*;
  Update table counts and estimates of the HPY hyperparameters;
end

A remark is that, up to now, we considered the additional sample size l as fixed. However, in some applications it could not be the case. For instance, if we capture animals using traps, we could not know in advance how many animals will be captured in the next step. In these circumstances, l is unknown and must be subjected to inference too. For example, we can assume independence of the rest of the model, adopt a simple parametric model for l and, at every time step, use its posterior point estimate to compute $\mathbb{E}[K_j^{(l)}|\mathbf{X_n}]$.

5.3.3 Implementation issues

The number of clusters in each population $\mathbf{m}_J = (m_{j\cdot} : j \in \{1, \ldots, J\})$ appearing in the parametrization of the beta distributions in both Algorithms 1 and 2 of Subsection 5.3.1 and 5.3.2, are latent variables. In the next paragraph we describe a simple MCMC scheme to estimate them in case an initial sample is available. The MCMC algorithm directly follows from paragraph 5.1 of Teh et al. [112]. Moreover if the hyperparameters of the HPY model are unknown, they must added to the MCMC sampler too, as outlined in a separated paragraph.

MCMC for $\mathbf{m}_J$: In principle a Gibbs sampler to estimate $\mathbf{m}_J$ should sequentially draw samples from the full conditionals $\pi(m_{j\cdot}|m_{1\cdot}, \ldots, m_{j-1\cdot}, m_{j+1\cdot}, \ldots, m_{J\cdot}, \mathbf{X_n})$. However both the joint $\pi(m_{1\cdot}, \ldots, m_{J\cdot}|\mathbf{X_n})$ and the full conditional posterior distributions are difficult combinatorial objects and cannot be derived in closed form. A possible solution is a Gibbs sampler that, rather than directly updating $m_{j\cdot}|m_{-j\cdot}, \mathbf{X_n}$, updates the cluster allocations $(t_{ji} : i \in \{1, \ldots, n_{j\cdot\cdot}\})$ and then computes $m_{j\cdot}|m_{-j\cdot}, \mathbf{X_n}$. As in Teh et al. [112], the cluster allocation variable t_{ji} specifies the cluster to which the i-th observation of population j belongs. Let $\mathbf{t}_{-jp}^{(i-1)}$ denote the array of cluster allocations after iteration

$i-1$ of the sampler and with the p-th observation of the j-th population removed. Then $t_{jp}^{(i)}|(\mathbf{X_n}, \mathbf{t}_{-jp}^{(i-1)})$ is proportional to

$$\sum_{t:\psi_{jt}=\psi_{jt_{ji}}} \frac{n_{jt\cdot} - I\left(t = t_{jp}^{(i-1)}\right) - \sigma}{\theta + n_{j\cdot\cdot} - 1}\delta_t + \frac{\theta + m_{j\cdot}^{(i-1,p-1)}\sigma}{\theta + n_{j\cdot\cdot} - 1}\frac{m_{\cdot k_{jp}}^{(i-1,p-1)} - \alpha_0}{\theta_0 + m_{\cdot\cdot}^{(i-1,p-1)}}\delta\Big(m_{j\cdot}^{(i-1,p-1)} + 1\Big),$$

where $m_{j\cdot}^{(i-1,p-1)}$ denotes the number of clusters in population j at the i-th iteration after having updated the first $p-1$ cluster allocations of that population, ψ_{jt} is a classification variable that tells us the species of the observations in the t-th cluster in population j and k_{jp} is the species of the observations in the p-th cluster in population j. If $n_{jt_{jp}\cdot} = 1$ (i.e. the observation is forming its own cluster), before updating $t_{jp}^{(i)}$ we must remove its cluster and subtract one to all the m's. The updated value for $m_{j\cdot}^{(i)}$ can also be taken as the highest $t_{jp}^{(i)}$ for $p \in \{1, \ldots, n_{j\cdot\cdot}\}$, rather than the number of distinct values in the $t_{jp}^{(i)}$.

The algorithm is time expensive because at every iteration it re-samples the cluster allocations of all populations and of all observations. However, we experienced that a good choice of the starting value makes the chain converge to its stationary distribution in just a few iterations. We suggest to run a Chinese Franchise given the data to find the initial point for cluster allocations to start the Gibbs sampler.

When the HPY-TS algorithm is run, the vector $\mathbf{m}_J$ can be updated by allocating new observations to either old or new clusters using the Chinese Restaurant Franchise. If the observation is new, it forms a new cluster. If it is old, say of type X_k^{**}, then the corresponding observation either will form a new cluster with probability proportional to $((m_{\cdot k} - \alpha_0)/(\theta_0 + m_{\cdot\cdot}))(\theta_j + m_{j\cdot}\alpha_j)/((\theta_j + n_{j\cdot\cdot}))$ or it will join an existing cluster (with dish X_k^{**}) with probability proportional to $(n_{j\cdot k} - m_{jk}\alpha_j)/(\theta_j + n_{j\cdot\cdot})$.

HPY Hyperparameters: If the hyperparameters are considered as unknown, they must be included in the Gibbs sampler for the cluster sizes. Assuming independent priors for hyperparameters of different Pitman-Yor processes, the full conditional distributions can be derived from

$$\pi(\alpha_0, \theta_0|(m_{jk} : j \in \{1, \ldots, J\}, k \in \{1, \ldots, K\}), (\alpha_j, \theta_j : j \in \{1, \ldots, J\}), \mathbf{X_n}) =$$

$$= \pi(\alpha_0, \theta_0|m_{\cdot\cdot}, K) \propto \frac{\Gamma(\frac{\theta_0}{\alpha_0} + K)\Gamma(\theta_0)\mathscr{C}(m_{\cdot\cdot}, K, \alpha_0)}{\Gamma(\frac{\theta_0}{\alpha_0})\Gamma(\theta_0 + m_{\cdot\cdot})}\pi^{prior}(\alpha_0, \theta_0)$$

and, for each couple $((\alpha_j, \theta_j) : j \in \{1, \ldots J\})$, from

$$\pi(\alpha_j, \theta_j|(m_{jk} : j \in \{1, \ldots, J\}, k \in \{1, \ldots, K\}), \sigma_{-j}, \theta_{-j}, \alpha_0, \theta_0, \mathbf{X_n}) =$$

$$= \pi(\alpha_j, \theta_j|n_{j\cdot\cdot}, m_{j\cdot}) \propto \frac{\Gamma(\frac{\theta_j}{\alpha_j} + m_{j\cdot})\Gamma(\theta_j)\mathscr{C}(n_{j\cdot\cdot}, m_{j\cdot}, \alpha_j)}{\Gamma(\frac{\theta_j}{\alpha_j})\Gamma(\theta_j + n_{j\cdot\cdot})}\pi^{prior}(\alpha_j, \theta_j).$$

5.4 Applications

5.4.1 Simulated Results

In the following simulations, the *true distribution* of each arm is supported on a subset of size 2500, randomly chosen from a total number of 3000 possible species, hence allowing for a partial sharing of the supports. These J distributions are assumed to follow *Zipf laws*. The mass assigned to the k-th most common species in population j, is

$$g_j\left(k; s_j\right) = \frac{1/k^{s_j}}{\sum_{n=1}^{2500}\left(1/n^{s_j}\right)},$$

where $s_j > 1$ is a real parameter controlling how the total mass is spread along the support points. When s_j is large, the total mass is concentrated on a few points and the ordered masses steeply decrease toward zero. As s_j approaches 1, the total mass is more spread, with many points of high mass.
In the bandit context, an arm with low parameter s_j can be viewed as a 'winning arm', an arm with high species variety. Whereas, a high value for s_j implies that, after the few very common species have been discovered, the discovery probability for that arm will be very close to zero.

The three *competing strategies*, used as a term of comparison, are the following:

- an *Oracle strategy*: this strategy knows the $(P_1, \ldots, P_J)$ that generates the data. Hence, uncertainty on the underlying data generating process is removed and, at every time step, this strategy selects the arm with the highest missing mass, so maximizing the probability of observing a new value in the next observation.

- a *Uniform strategy*: this strategy, at every time step, picks an arm uniformly at random, i.e. every arm has probability $1/J$ of being played next. Another similar strategy can be a deterministic strategy that cycles through the experts, i.e., at time t, it draws from population $(t \ mod\,[J])$.

- a UCB algorithm, recently proposed by Bubeck et al. [15], based on the Good and Turing missing mass estimator, derived in Good [50]. We refer to this algorithm as *Good-Turing strategy* for simplicity. It is an UCB-like algorithm introduced to solve the issue of security analysis of a power system. This algorithm uses an adaptation of Good and Turing missing mass estimator of Good [50] to produce a point estimate of the probability of observing a new item, in each arm. Then, it constructs a deterministic upper bound for this estimate, inversely proportional to the number of times that that arm has been played. The chosen arm is the one with the highest upper bound. More precisely, the adapted Good and Turing estimator counts the number of items with frequency one in joined sample that has been observed in

arm j and divides it by the number of plays of that arm. This ratio is the point estimator of the missing mass in arm j. The upper bound is constructed by summing $C \cdot (\log(4n)/n_{j\cdot\cdot})^{1/2}$ to the point estimate, where $n = \sum_{j=1}^{J} n_{j\cdot\cdot}$ is the total number of plays and C is a tuning parameter to be fixed.

We consider three different *scenarios*, corresponding to different levels of heterogeneity or homogeneity in species variety across arms. Heterogeneity in species variety depends on how different the parameters of the Zipf laws are across arms. When heterogeneity is high, 'winning' and 'losing' arms emerge. Winning arms are those with high species variety (with a low Zipf parameter), while the losing ones (those with high Zipf parameters) are those in which the mass will be concentrated on just a few dominating species. In presence of heterogeneity, a good strategy must be able to detect winning arms soon and play them only. Whereas, in presence of homogeneity, there will not be 'winning' arms and all arms will have similar probabilities of producing new species. In this case, a strategy must be able not to get stuck exploiting only a few arms, but to carefully explore all of them. In our simulations, we fix $J = 8$ and consider the following three scenarios:

1. *Pure Exploitation*, Zipf parameters=(1.3,1.3,2,2,2,2,2,2): in this scenario, there are two 'winning' arms. A good strategy should be able to intensively exploit these two arms, without exploring much the other six suboptimal arms.

2. *Pure Exploration*, Zipf parameters=(1.3,1.3,1.3,1.3,1.3,1.3,2,2): in this scenario, the majority of arms are equally profitable. A good strategy should not get stuck exploiting just a few of them, but continue to explore all the six good arms.

3. *Exploration plus Exploitation*, Zipf parameters=(1.3,1.3,1.3,1.3,2,2,2,2): in this scenario, there are four good arms and four bad ones. A good strategy should adequately balance exploitation and exploration, by stopping to play the four suboptimal arms soon, but continuing to play all the other four.

Figure 5.1, 5.2 and 5.3 report the results of simulations in the three scenarios just described for both abundance and incidence data. Each figure displays the average number of species discovered by the four algorithms, as a function of the additional samples observed. In particular, the results are averages of 60 runs. For each run, we assume an initial sample of 30 observations per arm to be available and collect further 300 observations, following the four possible strategies. In the abundance case, new observations arrive one at a time. In the incidence one, they arrive as 30 bunches of size 10. The hyperparameters of the HPY are endowed with priors, $\alpha_0, \ldots, \alpha_j \overset{iid}{\sim} \text{beta}\,(1,2)$ and $\theta_0, \ldots, \theta_J \overset{iid}{\sim} \exp(1)$, and then estimated using the MCMC described in Subsection 5.3.3. In the appendix of this chapter, we also provide Tables containing the weights given to each arm by the four algorithms in the first 20 simulations of each scenario, i.e. the number of times each arm

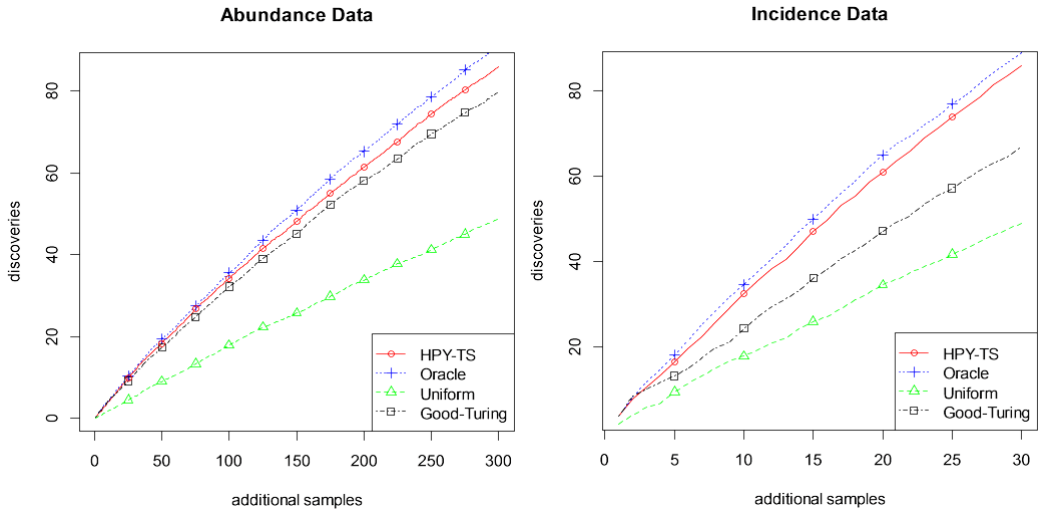

Figure 5.1: Simulated results: Pure Exploitation Scenario.

has been chosen by the four algorithms in each simulation.

In the simulations, the HPY-TS algorithm performs well in all scenarios, discovering fewer new species than the Oracle strategy, but more than the Uniform and the Good-Turing strategy. Figure 5.1, 5.2 and 5.3 show how these latter strategies seem to balance the exploration-exploration trade-off worse than HPY-TS. They perform relatively well only in the two extreme cases of pure exploration or pure exploitation, Figure 5.1 and 5.2. This guess is strongly confirmed by looking at the Tables in the appendix providing the weights of each arm. On the one hand, the Good-Turing strategy does too much exploitation. It selects the arm that seems the most profitable at initial time point and exploits it only, without exploring the others. This behaviour is evident by looking at the Tables with the weights. The algorithm performs well only in the pure exploitation scenario, Figure 5.1, in which exploiting just one arm is a profitable strategy. However, this strategy becomes suboptimal in presence of more 'winning' arms, as displayed in Figure 5.2 and 5.3. On the other hand, as expected, the Uniform strategy does too much exploration. It continues to play all arms, irrespectively of their past behaviors. Its performances are very poor, except in the extreme scenario of pure exploration, Figure 5.2. Instead, the HPY-TS algorithm seems to be robust to changes in species variety across arms. In all scenarios, it performs well, standing behind only to the Oracle strategy. In particular, in the intermediate scenario, Figure 5.3, its results are very close to the Oracle's ones, while in the extreme cases, Figure 5.1 and 5.2, it is still as good as or better than both the Uniform and the Good-Turing strategies.

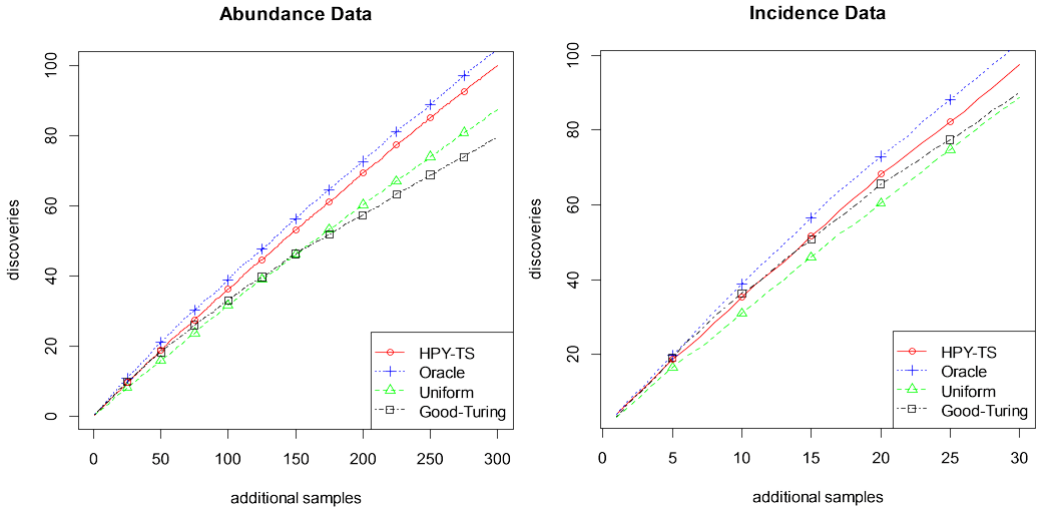

Figure 5.2: Simulated results: Pure Exploration Scenario.

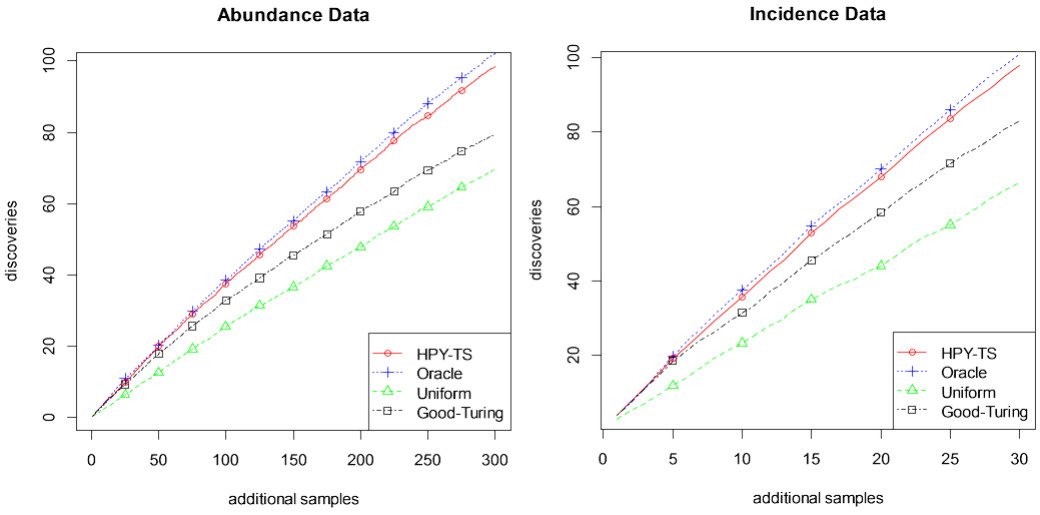

Figure 5.3: Simulated results: Exploitation plus Exploration Scenario.

5.4.2 Illustration using species of trees in South America

In this section, we compare the HPY-TS algorithm with the three competing alternatives previously described, using a dataset of species of trees, collected in South America. This dataset was studied in Pyke et al. [97] and Condit et al. [26] and contains species of trees observed in 100 plots near the Panama Canal, in Ecuador's Yasuní National Park and in Perú's Manu Biosphere Reserve. All plots were in terra firme forests and, in each of these plots, trees of $\geq$ 10-cm stem diameter were tagged, measured and sorted to morphospecies. A total of 41688 trees have been censored, with a total of 802 distinct species observed. The dataset is freely downloadable in the supplementary on-line section of Condit et al. [26].

In Condit et al. [26], this dataset has been used to study β-diversity in tropical forest trees. This is a measure of how species composition changes with distance. This notion was firstly introduced by Whittaker [124] together with the terms α and γ-diversities to describe species variety in a landscape. In particular, the total species variety (γ-diversity) can be viewed as the product of the mean species diversity in the habitat level (α-diversity) times the differentiation among habitats (β-diversity). α-diversity has been studied in many locations, in particular a high α-diversity has been amply documented for tropical forests under consideration. Using this dataset, Condit et al. [26] studies their β-diversity by comparing the actual data with those predicted by a neutral model in which habitat is uniform and only dispersal and speciation influence species turnover. The result of their study is that the data is inconsistent with the neutral model and a high level of β-diversity is observed. In particular, they show that β-diversity is higher in Panama than in western Amazonia.

We use the same dataset of counts of species to test the performances of the two HPY-TS algorithms against the three alternative strategies. Differently from Condit et al. [26] in which the analysis is static, in our context it is a sequential one. This problem is related to that of detecting rare or elusive species studied in the adaptive sampling literature, started with Thompson [115]. The difference with their analysis is that we do not have in advance a particular rare species we want to observe again, but we want to discover new ones. However, the HPY-TS algorithm can be easily modified to deal with the former problem. If we have a sample containing a particular rare species, labeled by X_k^{**}, that we are interested to re-observe, we can modify Proposition 5.3.1 and Algorithm 1, by computing $(P_1(\{X_k^{**}\}), \ldots, P_J(\{X_k^{**}\}))|\mathbf{X_n}$. This joint distribution can be derived in the same manner as in the proof of Proposition 5.3.1 and is still a mixture of a product of j beta distributions, with mixing measure another beta distribution, but with different parameters, depending also on table counts relative to the label X_k^{**}.

Another possible adaptation of the algorithm is to the case in which we want to maximize the sum of the distinct values observed in each location. In this problem,

Table 5.1: Real Data Example: Sorensen Index

Aggr.Plots	BCI	P	S	C
BCI	1	0.97	0.358	0.44
P	0.97	1	0.196	0.267
S	0.358	0.196	1	0.134
C	0.44	0.267	0.134	1

Table 5.2: Real Data Example: Shannon and Simpson Indeces

Aggr.Plots	Shannon	Simpson
BCI	4.27	0.974
P	5.25	0.988
S	3.412	0.936
C	3.953	0.97

the target function is affected when we observe a species that we have not observed in that location before, irrespectively of the fact that that species has already been observed it in other location. Proposition 5.3.1 and Algorithm 1 can be easily adapted to this problem. Denoting by $A_j = \left\{x \in \mathbb{X} : x \notin \mathbf{X}_{n_{j\cdot\cdot}}\right\}$ the set of species not observed in population j and following the same steps of the proof of Proposition 5.3.1, we derive the posterior density of $\{P_1(A_1), \ldots, P_J(A_j)\}|\mathbf{X_n}$ as a mixture of the product of J beta densities with parameters of the $j-th$ factor being $(\beta_0(\theta_j + m_{j\cdot}\alpha_j) + \sum_{k:X_k^{**}\notin\mathbf{X}_{n_{j\cdot\cdot}}}((\theta_j + m_{j\cdot}\alpha_j)\beta_k + n_{j\cdot k} - \alpha_j m_{jk}), \sum_{k:X_k^{**}\in\mathbf{X}_{n_{j\cdot\cdot}}}((\theta_j + m_{j\cdot}\alpha_j)\beta_k + n_{j\cdot k} - \alpha_j m_{jk}))$, and with mixing measure for $(\beta_0, \ldots, \beta_K)$ being a Dirichlet distribution of parameters $(\theta_0 + K\alpha_0, m_{\cdot 1} - \alpha_0, \ldots, m_{\cdot K} - \alpha_0)$. A HPY-TS algorithm can easily be implemented by substituting the joint posterior of Proposition 5.3.1 with this posterior density in Algorithm 1.

In order to test HPY-TS algorithm, we aggregated the 100 individual plots into 4 bigger groups, according to spatial location. In particular, we joined columns in the dataset with code starting with BCI, P, S and C. In Table 5.1, we computed the Sorensen similarity index. This similarity index measures the fraction of species shared in two plots and is computed as $2A/(2A + B + C)$, where A is the number of species shared between plots and B and C are the number of species unique to each plot. As a measure of similarity, the Sorensen index has the feature that it weights all species equally. Also, we computed in Table 5.2, the Shannon and Simpson indexes for the four aggregated plots. These two indexes measure species variety in a location and, given a (finite or infinite) discrete probability vector $\mathbf{p} = (p_1, p_2, \ldots)$, are computed as $S_{Shan}(\mathbf{p}) = -\sum_j p_j \log(p_j)$ and $S_{Simp}(\mathbf{p}) = 1 - \sum_j p_j^2$. Both indexes suggest a high species variety in the four plots.

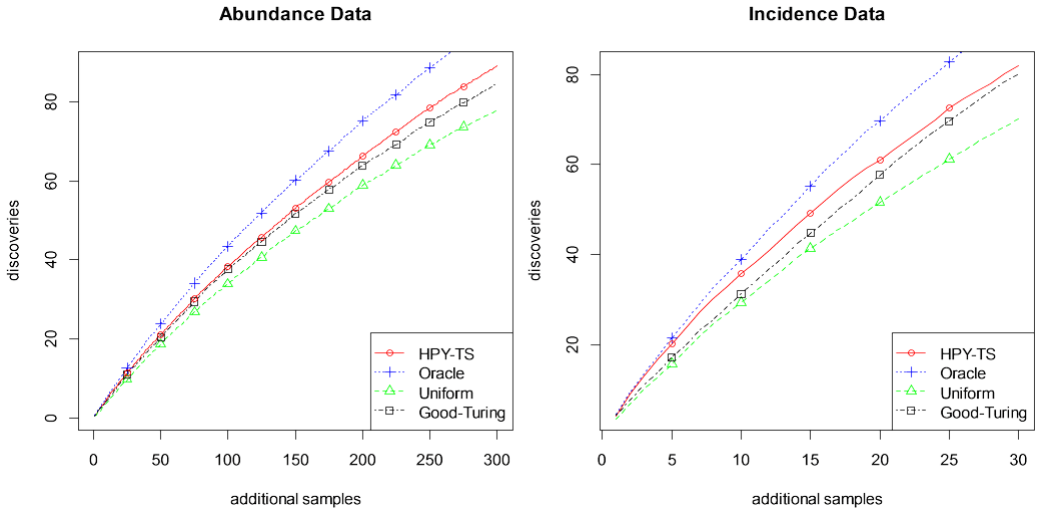

Figure 5.4: Real data example: Species of trees in South America

Given the four aggregated plots, we used their empirical distributions as data generating process. As for simulated results, we consider a small sample of 30 observations for each arm as initial sample and then we let the four competing algorithms choose where to collect further observations. These 300 additional observations are sampled one at a time in the abundance data case and as 30 bunches of size 10 when dealing with incidence data. The average results over 60 runs are displayed in Figure 5.4. Tables 5.3 and 5.4 report the weights given to each region by the four algorithms in the first 20 runs.

Figure 5.4 shows how the Oracle strategy outperforms the three other algorithms. The HPY-TS algorithm performs slightly better than the Good-Turing and better than the Uniform strategy, discovering on average more new species both in the abundance and in the incidence case. As shown in Condit et al. [26], the level of species variety in the four regions is very different, with the plots of Panama having a species variety much higher compared to the other three locations. Therefore this example is a quite extreme scenario and it is similar to scenario 1 of the simulated results, what we called pure exploitation scenario, with the arm P being the 'winning' one. As in the simulated case in this scenario results of the HPY-TS algorithms and Good-Turing are not too different, with the HPY-TS discovering on average just a few species more than the Good-Turing strategy. However, from the tables with the weights we note that the behaviours of the two algorithms are completely different. On the one side, the HPY-TS does exploration until it starts sampling only from region P. On the other side, the Good-Turing picks one arm at the beginning and exploits it only. Hence, in some simulations the Good-Turing algorithm selects the suboptimal arm BCI and remains stuck exploiting it, without doing any exploration and without realizing that indeed the best arm is P. In presence of more 'winning' arms, this greedy behaviour would turn out to be much less profitable than

Table 5.3: Real Data Example: Species of trees in South America. Abundance data.

	HPY-TS				Good-Turing			
Runs/Aggr.Plots	BCI	P	S	C	BCI	P	S	C
1	58	129	41	72	294	3	2	1
2	45	182	13	60	0	299	0	1
3	41	211	23	25	20	273	3	4
4	67	148	20	65	3	292	1	4
5	76	104	51	69	1	296	2	1
6	47	170	14	69	2	296	1	1
7	39	210	8	43	1	297	1	1
8	44	202	4	50	3	289	3	5
9	74	148	1	77	300	0	0	0
10	72	131	26	71	291	5	2	2
11	36	164	13	87	0	0	0	300
12	20	272	0	8	3	292	2	3
13	54	76	65	105	1	294	1	4
14	73	117	11	99	5	288	1	6
15	35	138	65	62	2	289	4	5
16	39	197	20	44	0	300	0	0
17	40	168	38	54	3	294	1	2
18	83	62	50	105	296	1	2	1
19	48	230	1	21	0	0	0	300
20	39	186	20	55	1	1	0	298
	Uniform				Oracle			
Runs/Aggr.Plots	BCI	P	S	C	BCI	P	S	C
1	75	73	64	88	0	300	0	0
2	72	85	69	74	0	300	0	0
3	93	65	69	73	0	284	0	16
4	86	73	68	73	0	300	0	0
5	73	79	73	75	0	298	0	2
6	78	92	70	60	5	285	0	10
7	81	60	83	76	0	278	0	22
8	84	64	88	64	0	300	0	0
9	60	91	58	91	0	284	0	16
10	77	81	54	88	0	293	0	7
11	76	68	80	76	0	287	0	13
12	78	79	63	80	0	299	0	1
13	76	76	65	83	0	291	0	9
14	68	75	80	77	0	265	35	0
15	64	76	75	85	0	300	0	0
16	76	76	73	75	0	300	0	0
17	89	79	65	67	0	300	0	0
18	74	69	79	78	0	300	0	0
19	72	68	85	75	0	286	9	5
20	77	81	75	67	0	300	0	0

Table 5.4: Real Data Example: Species of trees in South America. Incidence data.

	HPY-TS				Good-Turing			
Runs/Aggr.Plots	BCI	P	S	C	BCI	P	S	C
1	8	13	2	7	5	16	4	5
2	7	18	1	4	3	23	2	2
3	7	17	0	6	5	18	3	4
4	7	23	0	0	4	20	3	3
5	4	20	0	6	11	9	4	6
6	5	17	0	8	7	16	4	3
7	7	17	0	6	5	14	4	7
8	8	16	0	6	7	11	5	7
9	6	17	2	5	23	3	2	2
10	5	9	3	13	1	27	1	1
11	4	15	4	7	3	19	4	4
12	7	14	3	6	2	23	1	4
13	12	13	1	4	2	24	2	2
14	4	20	3	3	8	14	4	4
15	8	15	3	4	4	22	2	2
16	7	17	2	4	5	15	3	7
17	10	15	2	3	6	14	5	5
18	3	20	1	6	8	14	3	5
19	0	24	0	6	4	15	4	7
20	6	17	1	6	2	25	1	2
	Uniform				Oracle			
Runs/Aggr.Plots	BCI	P	S	C	BCI	P	S	C
1	5	12	9	4	0	28	0	2
2	9	8	7	6	0	30	0	0
3	7	7	9	7	1	29	0	0
4	7	5	9	9	0	30	0	0
5	8	8	10	4	0	30	0	0
6	8	10	5	7	0	30	0	0
7	5	9	7	9	0	30	0	0
8	3	13	6	8	1	29	0	0
9	7	6	10	7	0	30	0	0
10	9	8	7	6	0	30	0	0
11	9	8	7	6	0	30	0	0
12	7	11	6	6	0	30	0	0
13	7	7	7	9	0	30	0	0
14	7	5	9	9	0	29	0	1
15	8	13	5	4	0	29	0	1
16	5	9	8	8	0	30	0	0
17	10	4	5	11	0	30	0	0
18	3	11	7	9	2	28	0	0
19	7	6	10	7	0	30	0	0
20	6	6	9	9	0	29	0	1

in this example. Moreover, we remark that a big advantage of the HPY-TS algorithm is that it does not require any tuning parameter to regulate the exploration rate, but it balances exploration and exploitation automatically. This feature is very important because the right exploration rate depends on the level of heterogeneity in species variety in the populations, which is usually not known in advance. The HPY-TS seems robust to changes in the level of heterogeneity among populations and it is able to correctly balance the level of exploration and exploitation. Finally, as expected, the Uniform strategy performs worse than all the other strategies. However, the differences in performances among algorithms are now less remarked than in the simulated results, since, having only four arms rather than eight, the probability of picking suboptimal arms is lower in this context.

5.5 Discussion and future work

In this chapter we have introduced a new methodology to choose where to allocate resources when J distinct locations are available. This procedure works sequentially, suggesting at every time step from which location to collect the next sample. This sample can be composed of one observation only or by a group of them and the group sizes can be either fixed or random. For both cases we have provided HPY-TS algorithms, based on a joint use of tools from the Bayesian nonparametric and the multi-armed bandit literature. In particular a HPY is used to estimate the unknown cumulative distribution functions and TS for the sequential allocation problem. Up to our knowledge, this is the first instance that such tools have been used together, particularly with the aim of discovering items from multiple populations. Results from simulated and real data are good, showing that the HPY-TS algorithms are competitive or better than other strategies already proposed, both when dealing with abundance and with incidence data. These good empirical results encourage to continue the research in this direction, also in consideration of the wide applicability of the algorithm in a variety of fields, like ecology, biology or genetics.

What is surely missing is a theoretical analysis of the algorithm, which assures, not just empirically, its good properties. We think the next step is to analytically study the behavior of the HPY-TS algorithm, trying to provide a finite time bound for its regret. However, our context seems to be more challenging than that of the classical multi-armed bandit problem, due to the dependence of rewards both across time and populations. We think it could be helpful to substitute TS with an UCB strategy, in which the upper bound for the missing mass of each arm is constructed as a credible interval around the point estimate of Corollary 5.3.2, using the joint posterior distribution derived in Proposition 5.3.1. A concentration inequality providing an upper bound for the probability that the true missing mass of that arm is outside the credible interval, would then be very helpful

to prove a finite time regret bound for the new HPY-UCB strategy.

Another possible direction of research is to improve the proposed models by including spatial and covariate dependence. Indeed, the HPY model assumes exchangeability of the J distributions $(P_1, \ldots, P_J)$. In applied settings, this assumption may not be adequate, because the marginal distribution of each P_j could be different from place to place, depending on spatial or environmental factors. A possible extension of the proposed strategy can be to use dependent spatial models. In particular, in Bayesian nonparametric literature, a set of spatial models have been proposed as particular cases of the general Dependent Dirichlet Processes (DDPs) of MacEachern [85]. The DDPs are extensions of the Dirichlet Process to account for spatial or temporal dependence, but the same kind of generalization can be applied to any nonparametric prior admitting a stick breaking representation. The dependence on time or location is introduced by indexing the weights, the locations or both by a temporal or spatial variable. The most popular bayesian nonparametric spatial models are the spatial dependent Dirichlet process in Gelfand et al. [44], and its generalizations in Duan et al. [35] and Gelfand et al. [45], the hybrid Dirichlet mixture model of Petrone et al. [90], the order-based dependent Dirichlet process of Griffin and Steel [54], and the spatial kernel stick-breaking prior of Reich and Fuentes [101]. For a concise description of all of these models, the reader is referred to Steel and Fuentes [107], section 11.2.

5.6 Appendix - Tables of weights

The tables in the next pages report the weights given to each arm by the four algorithms in the simulation study. We consider the behaviour of the HPY-TS algorithm and of the three other competing strategies in the 3 scenarios described in Section 5.4.1. For each scenario we run the four algorithms for 60 times. We report here only the results of the first 20 simulations. Each row in the tables is the result of one simulation. The columns correspond to the possible arms. Finally, we repeat the same simulations both for abundance and for incidence data.

Table 5.5: Simulations: Pure Exploitation. Abundance data.

	HPY-TS								Good-Turing							
Runs/Zipf	1.3	1.3	2	2	2	2	2	2	1.3	1.3	2	2	2	2	2	2
1	157	137	4	1	1	0	0	0	3	292	1	1	1	1	1	0
2	147	121	3	5	1	4	13	6	282	11	1	1	1	1	2	1
3	168	125	0	0	0	4	2	1	2	298	0	0	0	0	0	0
4	10	279	2	0	6	0	2	1	5	269	4	4	4	5	4	5
5	216	66	0	0	5	3	0	10	293	1	1	1	1	1	1	1
6	134	150	0	1	0	12	0	3	1	298	0	0	0	1	0	0
7	147	100	1	9	1	35	0	7	291	2	1	1	1	2	1	1
8	137	161	0	2	0	0	0	0	290	5	1	1	0	1	1	1
9	146	149	1	0	0	4	0	0	299	1	0	0	0	0	0	0
10	134	111	8	3	4	6	13	21	288	5	1	1	1	1	1	2
11	78	201	0	0	19	0	1	1	8	279	2	2	3	2	2	2
12	167	132	0	0	0	0	1	0	290	4	1	1	1	1	1	1
13	118	177	0	0	2	0	3	0	275	13	2	2	2	2	2	2
14	85	203	0	1	5	1	4	1	2	291	1	1	2	1	1	1
15	84	192	4	0	7	10	2	1	4	283	2	2	2	3	2	2
16	137	97	53	1	0	5	5	2	1	296	1	0	0	1	0	1
17	166	87	0	2	4	31	5	5	290	2	1	1	1	2	2	1
18	227	58	1	2	2	6	2	2	295	1	0	1	1	1	0	1
19	154	102	2	11	6	2	16	7	293	1	1	1	1	1	1	1
20	60	223	8	0	1	3	5	0	7	282	2	1	2	2	2	2
	Uniform								Oracle							
Runs/Zipf	1.3	1.3	2	2	2	2	2	2	1.3	1.3	2	2	2	2	2	2
1	31	48	34	39	40	32	38	38	139	161	0	0	0	0	0	0
2	45	39	33	43	38	26	39	37	178	122	0	0	0	0	0	0
3	40	39	41	33	41	31	34	41	147	153	0	0	0	0	0	0
4	51	35	40	28	32	27	48	39	157	143	0	0	0	0	0	0
5	40	40	41	36	29	42	30	42	112	188	0	0	0	0	0	0
6	26	45	37	42	39	37	36	38	176	124	0	0	0	0	0	0
7	38	38	56	37	31	29	30	41	126	174	0	0	0	0	0	0
8	29	38	38	40	43	31	47	34	197	103	0	0	0	0	0	0
9	39	30	37	37	47	43	35	32	179	121	0	0	0	0	0	0
10	36	31	34	38	31	39	46	45	120	180	0	0	0	0	0	0
11	36	36	38	37	34	39	33	47	165	135	0	0	0	0	0	0
12	35	34	26	36	37	42	45	45	183	115	0	0	0	0	0	2
13	44	28	43	43	33	33	29	47	154	146	0	0	0	0	0	0
14	40	23	34	32	39	46	45	41	149	151	0	0	0	0	0	0
15	34	28	35	31	37	43	54	38	130	170	0	0	0	0	0	0
16	22	37	35	38	40	47	40	41	147	153	0	0	0	0	0	0
17	46	36	39	39	40	46	28	26	139	161	0	0	0	0	0	0
18	41	39	43	33	38	35	35	36	154	146	0	0	0	0	0	0
19	37	37	39	37	45	36	30	39	122	178	0	0	0	0	0	0
20	41	31	43	43	35	35	40	32	166	134	0	0	0	0	0	0

Table 5.6: Simulations: Pure Exploitation. Incidence data.

	HPY-TS								Good-Turing							
Runs/Zipf	1.3	1.3	2	2	2	2	2	2	1.3	1.3	2	2	2	2	2	2
1	14	16	0	0	0	0	0	0	5	9	2	2	3	3	3	3
2	15	14	0	0	0	1	0	0	3	15	1	2	2	2	2	3
3	22	6	0	0	0	0	2	0	7	6	3	3	2	3	3	3
4	9	19	0	2	0	0	0	0	24	1	1	1	1	0	1	1
5	11	19	0	0	0	0	0	0	14	3	2	2	3	2	2	2
6	10	20	0	0	0	0	0	0	14	4	1	2	3	2	2	2
7	16	13	0	0	0	0	0	1	4	13	2	2	2	3	2	2
8	17	12	0	1	0	0	0	0	6	11	3	2	2	2	2	2
9	16	14	0	0	0	0	0	0	18	2	2	2	1	2	1	2
10	13	16	0	0	0	0	1	0	11	5	2	3	2	2	2	3
11	11	19	0	0	0	0	0	0	6	9	2	3	3	2	2	3
12	19	11	0	0	0	0	0	0	9	6	3	2	3	2	3	2
13	11	19	0	0	0	0	0	0	6	8	3	2	2	3	3	3
14	18	12	0	0	0	0	0	0	3	20	1	1	2	1	1	1
15	19	10	0	0	0	0	1	0	18	4	2	2	1	1	1	1
16	18	12	0	0	0	0	0	0	3	17	2	1	1	2	2	2
17	18	12	0	0	0	0	0	0	5	11	2	3	2	3	2	2
18	12	13	1	2	0	1	0	1	8	5	3	3	2	3	2	4
19	18	12	0	0	0	0	0	0	6	5	3	3	4	4	3	2
20	15	14	0	0	1	0	0	0	18	2	2	1	2	2	1	2
	Uniform								Oracle							
Runs/Zipf	1.3	1.3	2	2	2	2	2	2	1.3	1.3	2	2	2	2	2	2
1	4	5	6	3	3	0	7	2	15	15	0	0	0	0	0	0
2	4	4	3	3	2	6	4	4	15	15	0	0	0	0	0	0
3	2	4	6	4	3	5	2	4	13	17	0	0	0	0	0	0
4	4	8	3	2	3	2	4	4	20	10	0	0	0	0	0	0
5	5	2	1	3	5	3	6	5	12	18	0	0	0	0	0	0
6	3	10	1	2	4	4	3	3	12	17	1	0	0	0	0	0
7	3	2	2	4	3	4	6	6	15	15	0	0	0	0	0	0
8	7	5	4	0	2	2	4	6	12	18	0	0	0	0	0	0
9	2	2	9	1	2	5	3	6	13	17	0	0	0	0	0	0
10	2	6	4	2	7	5	2	2	16	14	0	0	0	0	0	0
11	2	3	5	4	2	6	5	3	17	13	0	0	0	0	0	0
12	3	3	5	3	5	4	5	2	15	15	0	0	0	0	0	0
13	7	4	4	9	3	1	1	1	19	11	0	0	0	0	0	0
14	7	2	2	5	0	3	4	7	20	10	0	0	0	0	0	0
15	3	5	2	2	3	2	7	6	12	18	0	0	0	0	0	0
16	2	1	1	4	7	4	7	4	15	15	0	0	0	0	0	0
17	3	4	5	2	1	4	7	4	14	16	0	0	0	0	0	0
18	4	6	2	2	3	6	4	3	12	18	0	0	0	0	0	0
19	7	2	4	3	3	3	4	4	15	15	0	0	0	0	0	0
20	2	6	1	4	4	2	6	5	11	19	0	0	0	0	0	0

Table 5.7: Simulations: Pure Exploration. Abundance data.

	HPY-TS								Good-Turing							
Runs/Zipf	1.3	1.3	1.3	1.3	1.3	1.3	2	2	1.3	1.3	1.3	1.3	1.3	1.3	2	2
1	43	47	41	68	30	71	0	0	2	4	1	1	1	290	1	0
2	1	61	37	15	100	86	0	0	0	2	295	1	1	1	0	0
3	11	40	148	57	40	3	0	1	2	287	4	2	2	1	1	1
4	46	114	15	47	64	14	0	0	298	0	0	0	2	0	0	0
5	4	112	1	43	37	103	0	0	2	286	1	3	2	5	0	1
6	56	47	48	21	100	28	0	0	2	2	289	1	5	1	0	0
7	42	53	9	67	98	31	0	0	6	2	1	3	286	2	0	0
8	58	25	31	51	34	101	0	0	285	2	7	3	1	1	1	0
9	36	68	107	37	1	51	0	0	1	3	293	1	1	1	0	0
10	36	35	44	68	41	76	0	0	2	2	1	2	280	12	1	0
11	8	64	82	37	58	51	0	0	0	0	298	1	1	0	0	0
12	44	60	53	71	35	35	0	2	3	1	1	1	2	291	0	1
13	2	83	73	26	57	59	0	0	1	294	1	1	1	1	1	0
14	53	34	29	90	61	33	0	0	0	0	1	0	299	0	0	0
15	12	89	64	44	29	62	0	0	1	293	2	3	0	1	0	0
16	53	47	44	83	30	43	0	0	1	0	0	4	0	295	0	0
17	23	77	114	23	10	53	0	0	295	2	0	0	1	2	0	0
18	54	15	118	101	10	2	0	0	292	1	3	1	1	1	1	0
19	43	75	2	38	78	64	0	0	1	1	0	0	3	295	0	0
20	12	74	78	62	48	26	0	0	1	1	294	1	2	1	0	0
	Uniform								Oracle							
Runs/Zipf	1.3	1.3	1.3	1.3	1.3	1.3	2	2	1.3	1.3	1.3	1.3	1.3	1.3	2	2
1	31	43	31	34	38	49	30	44	83	41	30	58	54	34	0	0
2	45	24	29	36	32	40	50	44	73	41	74	77	5	30	0	0
3	43	47	32	36	38	28	35	41	38	31	61	78	48	44	0	0
4	27	44	37	45	41	30	41	35	46	30	60	42	54	68	0	0
5	33	40	49	44	34	30	35	35	52	45	42	49	35	77	0	0
6	39	36	42	40	33	36	33	41	75	57	75	36	10	47	0	0
7	31	36	43	54	30	42	30	34	38	60	108	19	51	24	0	0
8	49	32	31	36	39	42	35	36	37	72	19	86	53	33	0	0
9	33	39	42	36	36	40	34	40	53	52	51	17	69	58	0	0
10	40	33	30	33	40	45	39	40	55	73	47	36	61	28	0	0
11	27	39	47	41	43	34	30	39	68	63	31	68	17	53	0	0
12	29	29	50	44	35	39	31	43	43	39	73	61	65	19	0	0
13	46	42	40	24	47	38	26	37	55	60	45	38	50	52	0	0
14	45	37	38	44	37	30	39	30	58	83	31	52	39	37	0	0
15	36	32	38	32	35	41	48	38	51	31	35	33	60	90	0	0
16	40	29	30	56	31	41	43	30	65	49	67	84	25	10	0	0
17	49	36	34	35	46	38	29	33	30	74	45	62	47	42	0	0
18	34	42	31	39	36	39	40	39	60	44	20	55	64	57	0	0
19	33	38	40	40	37	36	47	29	94	42	49	21	45	49	0	0
20	35	32	41	40	31	43	37	41	58	33	56	37	68	48	0	0

Table 5.8: Simulations: Pure Exploration. Incidence data.

	HPY-TS								Good-Turing							
Runs/Zipf	1.3	1.3	1.3	1.3	1.3	1.3	2	2	1.3	1.3	1.3	1.3	1.3	1.3	2	2
1	0	14	5	6	2	3	0	0	2	2	5	15	2	2	1	1
2	6	8	5	5	6	0	0	0	2	5	15	2	2	2	1	1
3	1	3	4	7	7	8	0	0	3	1	19	1	1	3	1	1
4	7	4	0	5	4	10	0	0	5	2	2	2	8	9	1	1
5	6	8	7	7	2	0	0	0	2	7	6	7	2	3	2	1
6	1	16	2	5	2	4	0	0	1	1	1	25	1	1	0	0
7	10	1	6	1	9	3	0	0	1	1	24	1	1	2	0	0
8	5	12	1	2	2	8	0	0	3	1	1	1	2	21	0	1
9	7	5	10	2	0	6	0	0	4	4	2	1	1	16	1	1
10	3	0	15	0	6	6	0	0	1	1	3	3	18	2	1	1
11	6	3	2	6	5	8	0	0	20	1	2	2	1	2	1	1
12	3	0	7	2	6	12	0	0	1	1	25	1	1	1	0	0
13	8	2	0	6	7	7	0	0	9	3	2	6	3	5	1	1
14	2	4	7	0	11	6	0	0	0	0	28	0	1	1	0	0
15	5	7	7	0	5	6	0	0	1	1	1	1	25	1	0	0
16	1	1	2	12	9	5	0	0	3	3	2	3	9	6	2	2
17	5	9	9	1	4	2	0	0	16	2	3	2	3	2	1	1
18	0	5	4	10	4	7	0	0	2	2	2	18	2	2	1	1
19	4	3	2	7	14	0	0	0	2	3	2	17	2	2	1	1
20	1	0	8	8	5	8	0	0	3	1	3	5	15	1	1	1
	Uniform								Oracle							
Runs/Zipf	1.3	1.3	1.3	1.3	1.3	1.3	2	2	1.3	1.3	1.3	1.3	1.3	1.3	2	2
1	4	4	3	5	4	3	3	4	6	5	1	6	7	5	0	0
2	3	4	4	3	6	4	3	3	5	8	4	4	4	5	0	0
3	4	4	3	5	3	5	2	4	6	10	3	7	2	2	0	0
4	7	0	4	3	6	5	2	3	4	2	5	6	9	4	0	0
5	3	5	3	2	3	6	3	5	3	4	4	3	10	6	0	0
6	4	6	3	3	8	2	1	3	5	5	5	5	6	4	0	0
7	3	3	1	3	5	7	6	2	8	4	3	3	4	8	0	0
8	3	7	5	2	4	3	1	5	3	4	3	6	6	8	0	0
9	3	4	2	7	5	4	2	3	3	7	6	4	6	4	0	0
10	6	3	2	3	6	2	6	2	5	5	7	4	4	5	0	0
11	5	4	5	1	6	2	4	3	3	8	6	4	2	7	0	0
12	4	5	1	7	3	6	2	2	5	4	7	5	4	5	0	0
13	4	7	5	0	1	4	2	7	6	3	4	4	3	10	0	0
14	2	6	5	5	2	3	2	5	2	5	3	6	6	8	0	0
15	6	2	4	5	3	2	6	2	1	5	7	6	4	7	0	0
16	1	1	4	8	1	5	7	3	3	3	4	3	10	7	0	0
17	2	6	3	7	2	1	6	3	3	11	3	4	5	4	0	0
18	2	3	4	6	4	4	3	4	5	8	5	3	5	4	0	0
19	2	3	5	4	4	5	5	2	3	5	10	4	4	4	0	0
20	2	4	3	6	6	4	1	4	4	6	2	9	6	3	0	0

Table 5.9: Simulations: Exploration-Exploitation. Abundance data.

	HPY-TS								Good-Turing							
Runs/Zipf	1.3	1.3	1.3	1.3	2	2	2	2	1.3	1.3	1.3	1.3	2	2	2	2
1	10	159	74	57	0	0	0	0	0	0	300	0	0	0	0	0
2	81	76	63	80	0	0	0	0	297	1	1	1	0	0	0	0
3	72	60	85	83	0	0	0	0	1	2	4	292	1	0	0	0
4	72	30	129	69	0	0	0	0	1	1	2	294	1	0	0	1
5	107	42	59	90	0	0	2	0	298	0	1	1	0	0	0	0
6	63	105	112	20	0	0	0	0	298	2	0	0	0	0	0	0
7	92	86	42	78	0	1	1	0	4	2	3	290	0	0	1	0
8	79	120	27	73	0	0	1	0	293	5	1	1	0	0	0	0
9	61	49	108	81	0	0	0	1	296	1	1	2	0	0	0	0
10	100	35	57	108	0	0	0	0	300	0	0	0	0	0	0	0
11	21	164	65	50	0	0	0	0	0	0	0	300	0	0	0	0
12	29	54	111	100	5	0	0	1	5	4	13	272	2	2	1	1
13	104	97	25	71	0	3	0	0	3	293	1	1	0	1	0	1
14	20	95	91	86	0	2	6	0	1	2	5	288	1	1	1	1
15	15	153	81	50	0	0	0	1	1	4	2	293	0	0	0	0
16	82	70	89	50	5	0	4	0	0	0	299	1	0	0	0	0
17	3	82	133	81	0	0	0	1	2	4	4	286	1	1	1	1
18	36	92	120	52	0	0	0	0	9	2	286	2	1	0	0	0
19	70	125	77	28	0	0	0	0	1	296	1	1	0	0	1	0
20	47	67	102	84	0	0	0	0	1	1	1	295	1	0	1	0
	Uniform								Oracle							
Runs/Zipf	1.3	1.3	1.3	1.3	2	2	2	2	1.3	1.3	1.3	1.3	2	2	2	2
1	37	35	36	54	36	29	38	35	68	62	82	88	0	0	0	0
2	39	41	28	42	44	30	47	29	33	94	109	64	0	0	0	0
3	42	28	26	48	35	46	33	42	46	103	45	106	0	0	0	0
4	37	39	35	38	41	21	48	41	52	81	66	101	0	0	0	0
5	32	34	41	37	38	28	50	40	55	31	146	68	0	0	0	0
6	35	38	32	35	46	43	36	35	87	63	74	76	0	0	0	0
7	45	33	37	34	45	24	38	44	97	48	61	94	0	0	0	0
8	41	32	36	37	33	32	44	45	62	55	78	105	0	0	0	0
9	30	41	37	47	43	33	33	36	88	107	57	48	0	0	0	0
10	31	44	36	33	44	32	43	37	64	86	97	53	0	0	0	0
11	34	29	41	36	41	39	34	46	46	121	52	81	0	0	0	0
12	28	33	37	43	44	46	32	37	66	81	90	63	0	0	0	0
13	36	49	31	33	36	34	44	37	47	81	73	99	0	0	0	0
14	38	56	38	33	30	39	31	35	61	52	102	85	0	0	0	0
15	45	42	32	28	34	35	39	45	87	97	38	78	0	0	0	0
16	29	44	37	41	36	31	32	50	100	54	55	91	0	0	0	0
17	40	30	33	41	43	44	29	40	76	58	97	69	0	0	0	0
18	37	47	35	34	44	37	35	31	140	72	22	66	0	0	0	0
19	44	28	33	35	42	47	29	42	88	70	54	88	0	0	0	0
20	39	36	29	40	38	45	38	35	66	53	86	95	0	0	0	0

Table 5.10: Simulations: Exploration-Exploitation. Incidence data.

	HPY-TS								Good-Turing							
Runs/Zipf	1.3	1.3	1.3	1.3	2	2	2	2	1.3	1.3	1.3	1.3	2	2	2	2
1	5	5	6	14	0	0	0	0	3	8	4	9	2	1	2	1
2	5	14	7	4	0	0	0	0	18	2	2	4	1	1	1	1
3	6	6	9	9	0	0	0	0	2	6	2	16	1	1	1	1
4	9	8	1	12	0	0	0	0	3	3	1	18	2	1	1	1
5	1	6	8	15	0	0	0	0	3	3	5	12	2	1	2	2
6	1	9	4	16	0	0	0	0	3	16	3	3	2	1	1	1
7	4	9	9	8	0	0	0	0	2	1	2	21	1	1	1	1
8	5	6	12	7	0	0	0	0	7	2	14	3	1	1	1	1
9	6	7	9	8	0	0	0	0	27	1	1	1	0	0	0	0
10	4	6	8	12	0	0	0	0	4	3	2	16	1	2	1	1
11	0	12	4	14	0	0	0	0	1	2	19	4	1	1	1	1
12	0	13	5	12	0	0	0	0	2	3	2	19	1	1	1	1
13	3	9	12	6	0	0	0	0	4	19	1	2	1	1	1	1
14	7	7	7	9	0	0	0	0	2	21	1	2	1	1	1	1
15	7	6	8	9	0	0	0	0	1	1	27	1	0	0	0	0
16	8	10	5	7	0	0	0	0	18	2	5	1	1	1	1	1
17	7	9	8	6	0	0	0	0	3	11	5	5	1	2	2	1
18	8	6	9	7	0	0	0	0	17	2	2	3	2	1	1	2
19	3	10	7	10	0	0	0	0	3	6	3	11	2	1	2	2
20	4	16	5	5	0	0	0	0	4	3	2	17	1	1	1	1
	Uniform								Oracle							
Runs/Zipf	1.3	1.3	1.3	1.3	2	2	2	2	1.3	1.3	1.3	1.3	2	2	2	2
1	7	2	0	1	7	7	2	4	8	8	7	7	0	0	0	0
2	5	2	4	2	4	3	5	5	7	7	11	5	0	0	0	0
3	4	7	3	3	2	5	6	0	7	5	12	6	0	0	0	0
4	6	4	4	1	2	3	6	4	7	3	10	10	0	0	0	0
5	4	3	4	6	3	3	3	4	8	8	9	5	0	0	0	0
6	4	2	2	6	5	6	4	1	10	6	8	6	0	0	0	0
7	1	3	2	5	9	4	5	1	10	5	10	5	0	0	0	0
8	2	5	2	0	3	8	5	5	5	11	6	8	0	0	0	0
9	4	4	3	5	4	5	2	3	5	8	10	7	0	0	0	0
10	7	4	3	1	2	4	3	6	5	9	8	8	0	0	0	0
11	5	3	7	4	2	3	3	3	14	3	6	7	0	0	0	0
12	4	3	2	3	7	4	2	5	7	6	9	8	0	0	0	0
13	5	4	2	2	7	1	6	3	5	4	13	8	0	0	0	0
14	1	7	4	4	3	4	4	3	6	14	4	6	0	0	0	0
15	4	5	4	1	1	8	5	2	6	6	11	7	0	0	0	0
16	6	4	5	2	2	4	5	2	4	7	7	12	0	0	0	0
17	4	4	3	8	3	4	1	3	5	8	6	11	0	0	0	0
18	6	1	5	1	4	7	4	2	6	7	7	10	0	0	0	0
19	4	3	4	4	7	1	3	4	6	6	10	8	0	0	0	0
20	6	7	7	4	2	2	1	1	6	8	7	9	0	0	0	0

Chapter 6

Gibbs-type structure for feature models

6.1 Description of the problem and of the main theorem

In this chapter we deal with random feature allocations. As described in section 2.6, these models are generalizations of partitions in which the assumptions of exhaustivity and disjointness of the blocks are relaxed. In this chapter, we will proceed to study the counterpart of Gibbs-type partitions for feature allocations, hence introducing the class of *Gibbs-type feature allocations*.

To derive the Gibb-type partitions, Gnedin and Pitman [48] started from the distribution of the random partition generated by sampling observations from a PY process and considering the corresponding partition of the integers $\{1, \dots, n\}$ induced by this sample. This distribution is described by the Ewens–Pitman formula, which is a generalisation of the famous Ewens sampling formula. In particular, a random partition Π_n of the first n integers having this distribution with parameters (α, θ) assigns to a partition π_n, displaying k block with block sizes $(n_1, \dots, n_k)$, probability

$$\mathbb{P}(\Pi_n = \pi_n) = \frac{\prod_{i=1}^{k-1}(\theta + i\alpha)}{(\theta+1)_{n-1\uparrow}} \prod_{j=1}^{k} (1-\alpha)_{n_j-1\uparrow}. \tag{6.1}$$

Starting from this distribution, Gnedin and Pitman [48] considers a larger class of distributions, parametrized by a triangular array $V = (V_{n,k} : (n,k) \in \mathbb{N} \times \mathbb{N}, k \leq n)$ and a sequence $W = (W_j : j \in \mathbb{N})$ of non-negative weights. In particular, they consider the class of distributions for random partitions of $\mathbb{N}$ such that the restriction of the partition to

the first n integers, Π_n, assigns probability

$$\mathbb{P}(\Pi_n = \pi_n) = V_{n,k} \prod_{j=1}^{k} W_{n_j} \tag{6.2}$$

to a partition π_n, with k blocks and block sizes $(n_1, \ldots, n_k)$, hence maintaining the same product form as in (6.2). By imposing a further constraint to guarantee the resulting sequence $(\Pi_n)_{n\in\mathbb{N}}$ to be consistent (see section 2.3), Gnedin and Pitman [48] characterizes all elements of this class of distributions, which are termed *Gibbs-type partitions*.

In the feature contexts, the two most popular models are the Indian Buffet Process and the Beta–Bernoulli model. We recall their EFPF from section 2.6. The EFPF of a 3-parameter (γ, θ, α) IBP has form

$$\frac{1}{k!}\Big(\frac{\gamma}{(\theta+1)_{n-1\uparrow}}\Big)^k \exp\Big(-\sum_{i=1}^{n} \gamma \frac{(\alpha+\theta)_{i-1\uparrow}}{(1+\theta)_{i-1\uparrow}}\Big) \prod_{l=1}^{k} (1-\alpha)_{m_l-1\uparrow} (\theta+\alpha)_{n-m_l\uparrow}, \tag{6.3}$$

while the EFPF of a (N, α, θ) Beta–Bernoulli model is

$$\binom{N}{k} \left(\frac{-\alpha}{(\theta+\alpha)_{n\uparrow}}\right)^k \left(\frac{(\theta+\alpha)_{n\uparrow}}{(\theta)_{n\uparrow}}\right)^N \prod_{i=1}^{k} (1-\alpha)_{m_i-1\uparrow} (\theta+\alpha)_{n-m_i\uparrow}. \tag{6.4}$$

We can notice that both these formulas have a product structure. Specifically, in both formulas there is a first component depending on the data only through the number of individuals n and the number of features k displayed among these n individuals and then there is a product of k factors, each one depending on the data only through the cardinality of the i-th feature, m_i.

Motivated by the work of Gnedin and Pitman [48] in the partition context and by the similar product structure in the EFPF of the IBP and the Beta–Bernoulli model, in this chapter we consider the class of distributions for consistent exchangeable feature allocations F_n assigning to a feature allocation f_n of the first n integer having k blocks and block sizes $(m_1, \ldots, m_k)$, probability

$$\mathbb{P}(F_n = f_n) = V_{n,k} \prod_{l=1}^{k} W_{m_l} U_{n-m_l}, \tag{6.5}$$

for an infinite array $V = (V_{n,k} : (n,k) \in \mathbb{N} \times \mathbb{N}_0)$ and two sequences $W = (W_j : j \in \mathbb{N})$ and $U = (U_j : j \in \mathbb{N}_0)$ of non-negative weights, where $\mathbb{N}$ denotes the set of positive natural numbers and $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$.

In the feature context, we show that the Indian Buffet Process (IBP) and the Beta–Bernoulli are the only consistent exchangeable feature allocations with form (6.5), up to randomization of their γ and N parameters respectively. Imposing consistency and exchangeability imply that the two sequences of weights, W and U, must have the same

form as in the IBP and in the Beta–Bernoulli, for two constants α and θ, satisfying $\alpha \leq 1$ and $\theta > -\alpha$. In addition, V must satisfy a recursion with coefficients depending on α and θ and the set of solutions of this recursion forms convex set. For each fixed α and θ, we describe the extreme points of this convex set. Their form only depends on the value of α. For $0 < \alpha < 1$, the set of extreme points coincide with the family of V of a 3-parameter IBP. For $\alpha = 0$, this set is still continuous and coincides with the family of V of the 2-parameter IBP. For $\alpha < 0$ the set of extreme points is countably infinite and each extreme point corresponds to the V of a Beta–Bernoulli model. To sum up, we will prove the following theorem.

Theorem 6.1.1. *A consistent exchangeable feature allocation has EFPF of the form* (6.5) *iff, for some $\alpha < 1$ and $\theta > -\alpha$, $W_m = (1-\alpha)_{m-1\uparrow}$ and $U_m = (\theta+\alpha)_{m\uparrow}$ and the elements of V satisfies the recursion*

$$V_{n,k} = \sum_{j=0}^{\infty} \binom{k+j}{j} \left((\theta+\alpha)_{n\uparrow}\right)^j (\theta+n)^k V_{n+1,k+j}. \tag{6.6}$$

Moreover, for fixed (α, θ), the set of solutions of (6.6) *are*

1. *for $0 < \alpha < 1$, mixtures over γ of the V of a 3-parameter IBP;*
2. *for $\alpha = 0$, mixtures over γ of the V of a 2-parameter IBP;*
3. *for $\alpha < 0$, mixtures over N of the V of a Beta–Bernoulli model with N features.*

We will prove the first part of Theorem 6.1.1 in section 6.2, in which we characterize U and W and find the recursion (6.6) for V. In section 6.3, we describe how to derive the extreme solutions of this recursion. Finally, in subsection 6.4, we study the three cases $0 < \alpha < 1$, $\alpha = 0$, and $\alpha < 0$. Some simple facts and technical lemmas are postponed to an appendix at the end of the chapter.

6.2 Characterization of W and U

The problem we want to solve is to describe all distributions for exchangeable feature allocations with EFPF (6.5) subject to the consistency constraint (2.18), which becomes

$$V_{n,k} \prod_{i=1}^{k} W_{m_l} U_{n-m_l} = \sum_{j=0}^{\infty} \binom{k+j}{j} U_n^j W_1^j \sum_{\underline{z} \in \{0,1\}^k} V_{n+1,k+j} \prod_{i=1}^{k} W_{m_i+z_i} U_{n+1-m_i-z_i}, \tag{6.7}$$

for all $n \in \mathbb{N}$, $k \in \mathbb{N}_0$, and for all $m_i \leq n$ with $i \leq k$. We start by noting that the representation (6.5) is not unique. Specifically, we can scale the weights in the following ways, for $\kappa > 0$, and obtain the same distribution:

1. $\tilde{V}_{n,k} = \kappa^{-k} V_{n,k}$ and $\tilde{W}_j = \kappa W_j$;
2. $\tilde{V}_{n,k} = \kappa^{-k} V_{n,k}$ and $\tilde{U}_j = \kappa U_j$;
3. $\tilde{V}_{n,k} = \kappa^{-nk} V_{n,k}$, $\tilde{W}_j = \kappa^j W_j$ and $\tilde{U}_j = \kappa^j U_j$;
4. $\tilde{V}_{n,k} = \kappa^{-k(n-1)} V_{n,k}$, $\tilde{W}_j = \kappa^{j-1} W_j$ and $\tilde{U}_j = \kappa^j U_j$.

By imposing $W_1 = 1$, we avoid the first ambiguity and with $U_0 = 1$ we fix the second one. These conditions also exclude the third ambiguity, but do not exclude the last one, which will be fixed following Proposition 6.2.1.

The following Proposition shows that W and U must have a form akin to that of the IBP and V is constrained to satisfy a particular recursion. In the statement of Proposition 6.2.1, $(x)_{n\uparrow\tau}$ denotes the generalized rising factorial, defined as $(x)_{n\uparrow\tau} = x(x+\tau)\dots(x+\tau n)$.

Proposition 6.2.1. *The weights V, W, and U, with the normalizations $W_1 = U_0 = 1$, define a consistent random feature allocation of form (6.5) iff for some $a, b > 0$ and $\tau \geq 0$*
(i) $W_m = (a)_{m-1\uparrow\tau}$, for all $m \in \mathbb{N}$;
(ii) $U_m = (b)_{m\uparrow\tau}$, for all $m \in \mathbb{N}_0$;
(iii) For all $(n,k) \in \mathbb{N} \times \mathbb{N}_0$, V satisfies

$$\begin{aligned} \sum_{j\geq 0} V_{1,j} &= 1, \\ V_{n,k} &= \sum_{j=0}^{\infty} \binom{k+j}{j} \left((b)_{n\uparrow\tau}\right)^j (a+b+\tau(n-1))^k V_{n+1,k+j}. \end{aligned} \tag{6.8}$$

Proof. The consistent exchangeable feature allocation with no features with probability one can be represented as in (6.5), with $V_{n,0} = 1$ for all $n \in \mathbb{N}$ and $V_{n,k} = 0$ for $k \geq 1$. The consistency condition (6.7) for $k = 1$ gives

$$V_{n,1} W_{m_1} U_{n-m_1} = \sum_{j=0}^{\infty} (j+1) V_{n+1,j+1} U_n^j \left(W_{m_1+1} U_{n-m_1} + W_{m_1} U_{n+1-m_1}\right).$$

This condition implies that, for all $n \in \mathbb{N}$ and for all $m_1 \leq n$,

$$\frac{W_{m_1+1}}{W_{m_1}} + \frac{U_{n+1-m_1}}{U_{n-m_1}} = \frac{V_{n,1}}{\sum_{j=0}^{\infty} (j+1) V_{n+1,j+1} U_n^j}. \tag{6.9}$$

Since the right hand side of (6.9) does not depend on m_1, it follows that, for all n and for all $i, j \leq n$,

$$\frac{W_{i+1}}{W_i} - \frac{W_{j+1}}{W_j} = \frac{U_{n+1-j}}{U_{n-j}} - \frac{U_{n+1-i}}{U_{n-i}}.$$

In particular, considering $n = 2$, $i = 2$, and $j = 1$, we find

$$\frac{W_3}{W_2} - W_2 = \frac{U_2}{U_1} - U_1 =: \tau.$$

For $n > 1$, $i = n$, and $j = n - 1$, we also obtain

$$\frac{W_{n+1}}{W_n} - \frac{W_n}{W_{n-1}} = \frac{U_2}{U_1} - U_1 = \tau,$$

which implies, for all $n > 1$,

$$\frac{W_{n+1}}{W_n} = \tau\,(n-1) + W_2,$$

hence $W_n = (W_2)_{n-1\uparrow\tau}$. In a similar manner, we consider $n > 1$, $i = 2$, and $j = 1$ and obtain

$$\frac{U_n}{U_{n-1}} - \frac{U_{n-1}}{U_{n-2}} = \frac{W_3}{W_2} - W_2 = \tau.$$

As before, this formula implies $U_n = (U_1)_{n\uparrow\tau}$. The recursion (6.8) follows, by rewriting (6.7) as

$$V_{n,k} = \sum_{j=0}^{\infty} \binom{k+j}{j} U_n^j V_{n+1,k+j} \sum_{\underline{z}\in\{0,1\}^k} \prod_{i=1}^{k} \frac{W_{m_i+z_i}}{W_{m_l}} \frac{U_{n+1-m_i-z_i}}{U_{n-m_l}},$$

and by noticing that

$$\sum_{\underline{z}\in\{0,1\}^k} \prod_{i=1}^{k} \frac{W_{m_i+z_i}}{W_{m_l}} \frac{U_{n+1-m_i-z_i}}{U_{n-m_l}} = \left(U_1 + W_2 + \tau\,(n-1)\right)^k.$$

Also, $\sum_{j=0}^{\infty} V_{1,j} = 1$ comes from $\sum_{j=0}^{\infty} V_{1,j} W_1^j = 1$ and $W_1 = 1$. Finally, the reverse implication easily follows by checking that the probability distribution with form (6.5) and V, W and U as in the statement of the proposition satisfies the consistency condition (6.7). □

The last possible rescaling can now be fixed by imposing $\tau = 1$. Indeed, let W a sequence of weights parametrizing a feature allocation of form (6.5). From Proposition 6.2.1, $W_j = (a)_{j-1\uparrow\tau}$ for some $a > 0$ and some $\tau \geq 0$. If we consider the rescaling $\tilde{W}_j = \kappa^{j-1} W_j$, from $\tilde{W}_j = \kappa^{j-1}\,(a)_{j-1\uparrow\tau} = (\kappa a)_{j-1\uparrow\kappa\tau}$ and Proposition 6.2.1, we obtain $\tilde{W}_j = (\tilde{a})_{j-1\uparrow\tilde{\tau}}$, where $\tilde{a} = \kappa a$ and $\tilde{\tau} = \kappa\tau$. Therefore, by imposing $\tau = 1$, we avoid the last ambiguity on the rescaling since κ must be equal to 1 and $\tilde{W} = W$.

We now introduce the parametrization, $\alpha := 1 - a$ and $\theta := a + b$, for $\alpha < 1$ and $\theta > -\alpha$. Then, $W_{m_l} = (1-\alpha)_{m_l-1\uparrow}$ and $U_{n-m_l} = (\theta+\alpha)_{n-m_l\uparrow}$, which matches the form of the W and U for the IBP.

6.3 General tools to derive the extreme V

Let $\mathcal{V}_{\alpha,\theta}$ be the set of those elements $V \in \mathbb{R}_+^{\mathbb{N}\times\mathbb{N}_0}$ satisfying (6.8). Endow this set with the smallest σ-algebra $\mathcal{B}_{\mathcal{V}}$ that makes the maps $V \mapsto V_{n,k}$ measurable and define the

barycenter V^μ of each measure μ on $\mathcal{B}_\mathcal{V}$ as the pointwise average,

$$V^\mu_{n,k} = \int_{\mathcal{V}_{\alpha,\theta}} V_{n,k}\, \mu(\mathrm{d}V). \tag{6.10}$$

It is easy to check that $\mathcal{V}_{\alpha,\theta}$ is a convex set, i.e., for all probability measures μ on $\mathcal{B}_\mathcal{V}$, $V^\mu \in \mathcal{V}_{\alpha,\theta}$ (see Appendix 6.6.1). The goal of this section is to check that this set is also a simplex and to describe its extreme elements.

Given a measurable space of functions with the convex structure just defined, Dynkin [33] describes a general theory which can be applied to show that this set is also a simplex and to determine its extreme points. Similar results have been studied or rediscovered by many works, see references in Gnedin and Pitman [48]. To apply the results of Dynkin [33] to our problem, we will follow the same strategy used by Gnedin and Pitman [48]. Rather than studying $\mathcal{V}_{\alpha,\theta}$ directly, we consider another space, isomorphic to $\mathcal{V}_{\alpha,\theta}$ and easier to study, and we find its extreme points applying the results by Dynkin [33].

Let $(\mathbb{N}_0^\infty, \mathcal{C}(\mathbb{N}_0^\infty))$ be the infinite product space of $\mathbb{N}_0$, endowed with its cylinder σ-algebra. To each $V \in \mathcal{V}_{\alpha,\theta}$ we associate a Markov law, P_V, on this space. Specifically, writing $K_n : \mathbb{N}_0^\infty \to \mathbb{N}_0$ for the n-th coordinate projection on the product space, the Markov law associated to V has the initial distribution given by

$$P_V\left(K_1 = j\right) = V_{1,j}, \tag{6.11}$$

and transition probabilities

$$P_V\left(K_{n+1} = j + k | K_n = k\right) = \binom{k+j}{j} \left((\alpha+\theta)_{n\uparrow}\right)^j (\theta+n)^k \frac{V_{n+1,k+j}}{V_{n,k}}, \tag{6.12}$$

if $j \geq 0$ and 0 otherwise. Let $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}} = \{P_V : V \in \mathcal{V}_{\alpha,\theta}\}$ be the set of Markov laws. The map $T : \mathcal{V}_{\alpha,\theta} \to \mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$, defined by $T(V) = P_V$ is a convex isomorphism (see Appendix 6.6.1 for a proof). Hence, if P_V is extreme in $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$, so is V in $\mathcal{V}_{\alpha,\theta}$. We now describe how to find the extreme points of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$. Before that, we remark that, given an EFPF with form (6.5) parametrized by V, it is straightforward to show that K_n corresponds to the number of features in the corresponding random feature allocation of n individuals, i.e., K_n is the cardinality of $\bigcup_{1\leq i\leq n} X_i$.

As we will see from Proposition 6.3.1, for every $n \in \mathbb{N}$, $\mathcal{F}_n = \sigma(K_n, K_{n+1}, \ldots)$ is a sufficient σ-algebra for $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$. Hence, for each $n \in \mathbb{N}$, there exists a common $\mathcal{F}_n$-measurable regular conditional probability $Q_n : \mathbb{N}_0^\infty \times \mathcal{C}(\mathbb{N}_0^\infty) \to [0,1]$ for $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ given $\mathcal{F}_n$, such that, for all $P_V \in \mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ and $A \in \mathcal{C}(\mathbb{N}_0^\infty)$,

$$Q_n(\omega, A) = P_V((K_m)_{m\in\mathbb{N}} \in A | \mathcal{F}_n)(\omega), \tag{6.13}$$

for P_V-almost all $\omega \in \mathbb{N}_0^\infty$. In order to avoid having to repeat uninteresting measure-theoretic details, when $A' \in \sigma(K_1, \ldots, K_n)$, we will take advantage of the Markov property

of $(K_m)_{m\in\mathbb{N}}$ to assume that

$$Q_n(\omega, A') = P((K_m)_{m\in\mathbb{N}} \in A'|\mathcal{F}_n)(\omega) = P((K_m)_{m\in\mathbb{N}} \in A'|K_n)(\omega) \tag{6.14}$$

for all $\omega \in \mathbb{N}_0^\infty$, where we have dropped the V from the notation P_V in order to highlight the independence of the cotransition probabilities under P_V from V itself. This is justified because the equality holds for all $P \in \mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$.

Associated to each Markov kernel Q_n, there is a Markov operator Π_n given by

$$\Pi_n f(\omega) = \int f(\omega') Q_n(\omega, \mathrm{d}\omega'), \tag{6.15}$$

for all f bounded $\mathcal{F}_n$-measurable real functions. Henceforth, for every σ-algebra $\mathcal{F}$, we will simply write $f \in \mathcal{F}$ to denote that f is bounded and $\mathcal{F}$-measurable. The sequence $(\mathcal{F}_n, \Pi_n)_{n\in\mathbb{N}}$ forms a *specification* in $(\mathbb{N}_0^\infty, \mathcal{C}(\mathbb{N}_0^\infty))$ (see Appendix 6.6.1 for a proof). We can apply Theorem 5.1 of Dynkin [33], which states that $(\Pi_n)_{n\in\mathbb{N}}$ is an asymptotically H-sufficient statistic, which in turn means (see also Section 4.4 of Dynkin [33]) that, for all P_V that are extreme,

$$P_V(\{\omega \in \mathbb{N}_0^\infty : \forall f \in \mathcal{C}(\mathbb{N}_0^\infty),\ \lim_{n\to\infty} \Pi_n f(\omega) = \textstyle\int f \,\mathrm{d}P_V\}) = 1. \tag{6.16}$$

A path $\omega \in \mathbb{N}_0^\infty$ induces a Markov law $P_V \in \mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ and is said to be *regular* iff for all $f \in \mathcal{C}(\mathbb{N}_0^\infty)$, $\lim_{n\to\infty} \Pi_n f(\omega) = \int f \,\mathrm{d}P_V$. The set of point in $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ that are induced by regular paths is called the *maximal boundary*. The set of extreme points, also called the *minimal boundary*, is the subset of the maximal boundary, corresponding to those points P_V that also satisfy (6.16), i.e., they assign probability 1 to the set of regular paths inducing them.

In our context, to identify the maximal boundary, it is enough to check (6.16) for all functions $f \in \mathcal{C}(\mathbb{N}_0^\infty)$ that are indicators of cylinder sets of the form $K_n^{-1}\{k\}$ for $n \in \mathbb{N}$ and $k \in \mathbb{N}_0$. That is, the elements belonging to the maximal boundary are those $P_{\bar{V}} \in \mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ such that, for some $\omega \in \mathbb{N}_0^\infty$,

$$\lim_{m\to\infty} P(K_n = k|\mathcal{F}_m)(\omega) = \lim_{m\to\infty} P(K_n = k|K_m)(\omega) = P_{\bar{V}}(K_n = k),$$

for all $(n, k) \in \mathbb{N}\times\mathbb{N}_0$. To find the extremes measures of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$, we compute the cotransition (backwards) probabilities of $(K_n)_{n\in\mathbb{N}}$.

Proposition 6.3.1. *The cotransition probabilities are*

$$P(K_n = k|K_m = l) = \frac{d_{n,k}^{m,l}}{d^{m,l}} d^{n,k}, \tag{6.17}$$

for $n < m$ and $k \leq l$, while the distribution of K_n under $P_V \in \mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ is

$$P_V(K_n = k) = V_{n,k} d^{n,k}, \tag{6.18}$$

where

$$d^{m,l} = ((\theta+1)_{m-1\uparrow} + \sum_{j=1}^{m-1}(\theta+\alpha)_{m-j\uparrow}(\theta+1+m-j)_{j-1\uparrow})^l, \tag{6.19}$$

and

$$d^{m,l}_{n,k} = \binom{l}{k}((\theta+n)_{m-n\uparrow})^k(\sum_{j=1}^{m-n}(\alpha+\theta)_{m-j\uparrow}(\theta+m-1)_{j-1\downarrow})^{l-k}. \tag{6.20}$$

Proof. First, note that for $m > n$ and $l \geq k$ $P_V(K_m = l|K_n = k) = V_{m,l}d^{m,l}_{n,k}$, for a function $d^{m,l}_{n,k}$ independent of V. Indeed, from (6.12), the probability of a path $(k_{n+1}, k_{n+2}, \ldots, k_{m-1}, l)$ depends only on the last $V_{m,l}$. Summing over all possible paths from $K_n = k$ to $K_m = l$, we see that $P_V(K_m = l|K_n = k)$ must be of the form $V_{m,l}d^{m,l}_{n,k}$. In addition, by considering $P_V(K_m = l) = \sum_{i=0}^{l} P_V(K_m = l|K_1 = i) \cdot P_V(K_1 = i)$, $P_V(K_m = l)$ must be of the form $V_{m,l}d^{m,l}$. Also, from

$$P_V\left(K_m = l\right) = \sum_{j=0}^{l} P_V\left(K_m = l|K_{m-1} = j\right) \cdot P_V\left(K_{m-1} = j\right),$$

it follows that, for $l > 2$, the function $d^{m,l}$ must satisfy

$$V_{m,l}d^{m,l} = \sum_{j=0}^{l} \frac{V_{m,l}}{V_{m-1,j}}\tbinom{l}{l-j}((\alpha+\theta)_{m-1\uparrow})^{l-j}\left(\theta+m-1\right)^j V_{m-1,j}d^{m-1,j},$$

which gives the following the recursion

$$d^{m,l} = \sum_{j=0}^{l} \tbinom{l}{l-j}((\alpha+\theta)_{m-1\uparrow})^{l-j}\left(\theta+m-1\right)^j d^{m-1,j}.$$

Substituting $d^{m-1,j}$, we find

$$\begin{aligned} d^{m,l} = &\sum_{j=0}^{l} \tbinom{l}{l-j}((\alpha+\theta)_{m-1\uparrow})^{l-j}\left(\theta+m-1\right)^j \cdot \\ &\cdot \sum_{i=0}^{j} \binom{j}{j-i}((\alpha+\theta)_{m-2\uparrow})^{j-i}\left(\theta+m-2\right)^i d^{m-2,i}. \end{aligned}$$

Grouping together all coefficient multiplying $d^{m-2,k}$ on the right hand side $(0 \leq k \leq l)$, we find

$$
\begin{aligned}
d^{m,l}_{n-2,k} = & \binom{l}{l-k}\left((\alpha+\theta)_{m-1\uparrow}\right)^{l-k}(\theta+m-1)^{k}(\theta+m-2)^{k} \\
& + \binom{l}{l-k-1}\left((\alpha+\theta)_{m-1\uparrow}\right)^{l-k-1}(\theta+m-1)^{k+1}\left((\alpha+\theta)_{m-2\uparrow}\right)(\theta+m-2)^{k} \\
& + \binom{l}{l-k-2}\left((\alpha+\theta)_{m-1\uparrow}\right)^{l-k-2}(\theta+m-1)^{k+2}\binom{k+2}{2}\left((\alpha+\theta)_{m-2\uparrow}\right)^{2}(\theta+m-2)^{k} \\
& \quad \vdots \\
& + \left((\alpha+\theta)_{m-1\uparrow}\right)(\theta+m-1)^{l-1}\binom{l-1}{l-1-k}\left((\alpha+\theta)_{m-2\uparrow}\right)^{l-1-k}(\theta+m-2)^{k} \\
& + (\theta+m-1)^{l}\binom{l}{l-k}\left((\alpha+\theta)_{m-2\uparrow}\right)^{l-k}(\theta+m-2)^{k} \\
= & \left((\theta+m-1)(\theta+m-2)\right)^{k}\left((\alpha+\theta)_{m-1\uparrow}+(\theta+m-1)(\alpha+\theta)_{m-2\uparrow}\right)^{l-k}\binom{l}{l-k}.
\end{aligned}
$$

So, the recursion for $d^{m,l}$ becomes

$$
\begin{aligned}
d^{m,l} = & \sum_{j=0}^{l}\binom{l}{l-j}\left((\theta+m-1)(\theta+m-2)\right)^{j}\cdot \\
& \cdot\left((\alpha+\theta)_{m-1\uparrow}+(\theta+m-1)(\alpha+\theta)_{m-2\uparrow}\right)^{l-j}d^{m-2,j}.
\end{aligned}
$$

In the same manner, we find

$$
\begin{aligned}
d^{m,l}_{m-3,k} = & \left((\theta+m-1)(\theta+m-2)(\theta+m-3)\right)^{k}\binom{l}{l-k}\cdot \\
& \cdot\left((\alpha+\theta)_{m-1\uparrow}+(\theta+m-1)(\alpha+\theta)_{m-2\uparrow}+(\theta+m-1)(\theta+m-2)(\alpha+\theta)_{m-3\uparrow}\right)^{l-k}.
\end{aligned}
$$

Finally, we obtain,

$$
\begin{aligned}
d^{m,l}_{n,k} &= \binom{l}{l-k}((\theta+m-1)_{m-n\downarrow})^{k}(\sum_{j=1}^{m-n}(\alpha+\theta)_{m-j\uparrow}(\theta+m-1)_{j-1\downarrow})^{l-k} \\
&= \binom{l}{k}((\theta+n)_{m-n\uparrow})^{k}(\sum_{j=1}^{m-n}(\alpha+\theta)_{m-j\uparrow}(\theta+m-1)_{j-1\downarrow})^{l-k}.
\end{aligned}
$$

In addition,

$$
\begin{aligned}
d^{m,l} &= \sum_{i=0}^{l}d^{m,l}_{1,i} = \sum_{i=0}^{l}\binom{l}{i}((\theta+1)_{m-1\uparrow})^{i}(\sum_{j=1}^{m-1}(\alpha+\theta)_{m-j\uparrow}(\theta+m-1)_{j-1\downarrow})^{l-i} \\
&= ((\theta+1)_{m-1\uparrow}+\sum_{j=1}^{m-1}(\alpha+\theta)_{m-j\uparrow}(\theta+1+m-j)_{j-1\uparrow})^{l}.
\end{aligned}
$$

□

Note that the cotransition probabilities are independent of V.

6.4 Characterization of $V_{n,k}$

In this section, we study the three cases $0 < \alpha < 1$, $\alpha = 0$, and $\alpha < 0$ separately. Recall that a path $\omega = (\omega_1, \omega_2, \ldots) \in \mathbb{N}_0^\infty$ is regular and induces $\bar{V} \in \mathcal{V}_{\alpha,\theta}$ if the limit

$$\lim_{m\to\infty} P(K_n = k | K_m = \omega_m) = \lim_{m\to\infty} \frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} d^{n,k} = \bar{V}_{n,k} d^{n,k} \tag{6.21}$$

exists for all (n,k). In this case, $P_{\bar{V}}$ belongs to the maximal boundary of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$. If $P_{\bar{V}}$ also assigns probability one to the set of regular paths inducing it, then $P_{\bar{V}}$ is extreme.

6.4.1 Case $0 < \alpha < 1$

For (α, θ) fixed, s.t. $0 < \alpha < 1$ and $\theta > -\alpha$, let $V^{3IBP,\alpha,\theta}(\gamma)$ be the V of the 3-parameter IBP, defined as

$$\begin{aligned} V_{n,k}^{3IBP,\alpha,\theta}(\gamma) &= \frac{1}{k!}\Big(\frac{\gamma}{(\theta+1)_{n-1\uparrow}}\Big)^k \exp\Big(-\sum_{i=1}^{n} \gamma \frac{(\alpha+\theta)_{i-1\uparrow}}{(1+\theta)_{i-1\uparrow}}\Big) \\ &= \frac{1}{k!}\Big(\frac{\gamma}{(\theta+1)_{n-1\uparrow}}\Big)^k \exp\Big(-\gamma\Big(\frac{\Gamma(\theta+1)\Gamma(\alpha+\theta+n)}{\alpha\Gamma(\alpha+\theta)\Gamma(\theta+n)} - \frac{\theta}{\alpha}\Big)\Big), \end{aligned}$$

for all $\gamma \geq 0$. Define $\mathcal{P}_{V^{3IBP,\alpha,\theta}} = \{P_{V^{3IBP,\alpha,\theta}(\gamma)} \in \mathcal{P}_{\mathcal{V}_{\alpha,\theta}} : \gamma \geq 0\}$.

Proposition 6.4.1. *Let $0 < \alpha < 1$ and $\theta > -\alpha$.*

a) The elements of the set $\mathcal{P}_{V^{3IBP,\alpha,\theta}}$ belong to the maximal boundary of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ and they are induced by those paths $w \in \mathbb{N}_0^\infty$ s.t. $\frac{w_m}{m^\alpha} \to c$, where $c = \frac{\gamma\Gamma(\theta+1)}{\alpha\Gamma(\alpha+\theta)}$;

b) The elements of $\mathcal{P}_{V^{3IBP,\alpha,\theta}}$ also belong to the minimal boundary of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$, i.e., they are extreme points of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$;

c) The elements of $\mathcal{P}_{V^{3IBP,\alpha,\theta}}$ are the only extreme points, i.e., $\mathcal{P}_{V^{3IBP,\alpha,\theta}}$ coincides with the maximal and the minimal boundary.

Proof. a) We must check that

$$\lim_{\substack{m\to\infty \\ \frac{w_m}{m^\alpha}\to c}} \frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} = V_{n,k}^{3IBP,\alpha,\theta}\Big(\frac{c\alpha\Gamma(\alpha+\theta)}{\Gamma(\theta+1)}\Big). \tag{6.22}$$

From Proposition 6.3.1,

$$\begin{aligned}
\frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} &= \binom{\omega_m}{k} \frac{\left[(\theta+m-1)_{m-n\downarrow}\right]^k \left[\sum_{i=1}^{m-n}(\alpha+\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow}\right]^{\omega_m-k}}{\left[(\theta+m-1)_{m-1\downarrow}+\sum_{i=1}^{m-1}(\alpha+\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow}\right]^{\omega_m}} \\
&= \binom{\omega_m}{k}\left[\frac{(\theta+m-1)_{m-n\downarrow}}{\sum_{i=1}^{m-n}(\alpha+\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow}}\right]^k \\
&\quad \cdot \left[\frac{\sum_{i=1}^{m-n}(\alpha+\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow}}{(\theta+m-1)_{m-1\downarrow}+\sum_{i=1}^{m-1}(\alpha+\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow}}\right]^{\omega_m} \\
&= \binom{\omega_m}{k}\left[\frac{\frac{\Gamma(\theta+m)}{\Gamma(\theta+n)}}{\frac{\Gamma(\alpha+\theta+m)}{\alpha\cdot\Gamma(\alpha+\theta)}-\frac{\Gamma(\theta+m)\cdot\Gamma(\alpha+\theta+n)}{\alpha\cdot\Gamma(\alpha+\theta)\cdot\Gamma(\theta+n)}}\right]^k \\
&\quad \cdot \left[\frac{\frac{\Gamma(\alpha+\theta+m)}{\alpha\cdot\Gamma(\alpha+\theta)}-\frac{\Gamma(\theta+m)\cdot\Gamma(\alpha+\theta+n)}{\alpha\cdot\Gamma(\alpha+\theta)\cdot\Gamma(\theta+n)}}{\frac{\Gamma(\theta+m)}{\Gamma(\theta+1)}+\frac{\Gamma(\alpha+\theta+m)}{\alpha\cdot\Gamma(\alpha+\theta)}-\frac{\Gamma(\theta+m)\cdot\Gamma(\alpha+\theta+1)}{\alpha\cdot\Gamma(\alpha+\theta)\cdot\Gamma(\theta+1)}}\right]^{\omega_m},
\end{aligned} \tag{6.23}$$

where the third equality follows from the identity

$$\sum_{i=1}^{m-n}(\alpha+\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow} = \frac{\Gamma(\alpha+\theta+m)}{\alpha\Gamma(\alpha+\theta)} - \frac{\Gamma(\theta+m)\Gamma(\alpha+\theta+n)}{\alpha\Gamma(\alpha+\theta)\Gamma(\theta+n)},$$

which itself arises from rewriting the sum as a difference of two infinite hypergeometric series evaluated at 1 and applying the Gauss theorem for hypergeometric series.

Using the asymptotic equivalence $\Gamma(m+\delta)\approx\Gamma(m)m^\delta$ and limit $(m+\theta)^\alpha - m^\alpha \to 0$, the Stirling formula $\binom{\omega_m}{k} \approx \frac{1}{k!}\omega_m^k$ for the binomial coefficient, and then the limit $\frac{\omega_m}{m^\alpha}\to c$, the first line of (6.23) can be simplified to yield a limiting form:

$$\begin{aligned}
&\binom{\omega_m}{k}\left[\frac{\frac{\Gamma(\theta+m)}{\Gamma(\theta+n)}}{\frac{\Gamma(\alpha+\theta+m)}{\alpha\cdot\Gamma(\alpha+\theta)}-\frac{\Gamma(\theta+m)\cdot\Gamma(\alpha+\theta+n)}{\alpha\cdot\Gamma(\alpha+\theta)\cdot\Gamma(\theta+n)}}\right]^k \\
&\approx \binom{\omega_m}{k}\left[\frac{\frac{1}{\Gamma(\theta+n)}}{\frac{m^\alpha}{\alpha\Gamma(\alpha+\theta)}-\frac{\Gamma(\alpha+\theta+n)}{\alpha\Gamma(\alpha+\theta)\Gamma(\theta+n)}}\right]^k \\
&\approx \frac{1}{k!}\left[\omega_m m^{-\alpha}\cdot\frac{\frac{1}{\Gamma(\theta+n)}}{\frac{1}{\alpha\Gamma(\alpha+\theta)}-\frac{\Gamma(\alpha+\theta+n)}{m^\alpha\alpha\Gamma(\alpha+\theta)\Gamma(\theta+n)}}\right]^k \\
&\approx \frac{1}{k!}\left[c\cdot\frac{\frac{1}{\Gamma(\theta+n)}}{\frac{1}{\alpha\Gamma(\alpha+\theta)}-\frac{\Gamma(\alpha+\theta+n)}{m^\alpha\alpha\Gamma(\alpha+\theta)\Gamma(\theta+n)}}\right]^k \to \frac{1}{k!}\left(c\frac{\alpha\Gamma(\alpha+\theta)}{\Gamma(\theta+n)}\right)^k.
\end{aligned} \tag{6.24}$$

Similarly, the second line of (6.23) can be simplified by Lemma 6.6.4, using the asymptotic equivalence $\Gamma(m+\delta)\approx\Gamma(m)m^\delta$ and the limits $(m+\theta)^\alpha - m^\alpha \to 0$ and

$\frac{\omega_m}{m^\alpha} \to c$, to yield

$$\begin{aligned}&\left[\frac{\frac{\Gamma(\alpha+\theta+m)}{\alpha\cdot\Gamma(\alpha+\theta)} - \frac{\Gamma(\theta+m)\cdot\Gamma(\alpha+\theta+n)}{\alpha\cdot\Gamma(\alpha+\theta)\cdot\Gamma(\theta+n)}}{\frac{\Gamma(\theta+m)}{\Gamma(\theta+1)} + \frac{\Gamma(\alpha+\theta+m)}{\alpha\cdot\Gamma(\alpha+\theta)} - \frac{\Gamma(\theta+m)\cdot\Gamma(\alpha+\theta+1)}{\alpha\cdot\Gamma(\alpha+\theta)\cdot\Gamma(\theta+1)}}\right]^{\omega_m} \\ &\approx \left[\frac{m^\alpha - \frac{\Gamma(\alpha+\theta+n)}{\Gamma(\theta+n)}}{m^\alpha - \frac{\theta\Gamma(\alpha+\theta)}{\Gamma(\theta+1)}}\right]^{cm^\alpha} \to \exp\left\{c\left(\frac{\theta\Gamma(\alpha+\theta)}{\Gamma(\theta+1)} - \frac{\Gamma(\alpha+\theta+n)}{\Gamma(\theta+n)}\right)\right\}.\end{aligned} \tag{6.25}$$

Substituting back into (6.23), we obtain the V of the 3-parameter IBP.

b) From Theorem 4 of Berti et al. [11], it follows that

$$P_{V^{3IBP,\alpha,\theta}(\frac{c\alpha\Gamma(\alpha+\theta)}{\Gamma(\theta+1)})}\left(\frac{K_n}{n^\alpha} \to c\right) = 1.$$

c) To check that the only regular paths are those paths $\omega \in \mathbb{N}_0^{\mathbb{N}}$ such that $\frac{w_m}{m^\alpha} \to c$ for some $c \geq 0$, suppose otherwise; i.e., let $\omega \in \mathbb{N}_0^{\mathbb{N}}$ be a regular path, but assume $\frac{w_m}{m^\alpha}$ does not converge to some finite $c \geq 0$. If $(\frac{w_m}{m^\alpha})_{m\in\mathbb{N}}$ has at least two distinct subsequential limits, then, from the proof of part (a), we see that $d_{n,k}^{m,\omega_m}/d^{m,\omega_m}$ has at least two distinct subsequential limits, a contradiction, and so $\frac{w_m}{m^\alpha} \to \infty$. But then it follows from equations (6.23), (6.24), and (6.25); the asymptotic equivalence $\Gamma(m+\delta) \approx \Gamma(m)m^\delta$ and limit $(m+\theta)^\alpha - m^\alpha \to 0$; and finally an application of Lemma 6.6.5 that $\frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} \to 0$ as $m \to \infty$ for every $k \in \mathbb{N}_0$. As these limits must define a probability distribution, this is a contradiction, completing the proof.

□

In Proposition 6.4.1, the case $\gamma = 0$ corresponds to the degenerate feature allocation with no features with probability one, corresponding to $V_{n,0} = 1$ and $V_{n,k} = 0$ for all $n \in \mathbb{N}$ and $k \geq 1$. This solution is induced by the path $\omega_m = 0$ for all $m \in \mathbb{N}$, which has probability one under this degenerate law.

6.4.2 Case $\alpha = 0$

For θ fixed and positive, the V of the 2-parameter IBP are of the form

$$\begin{aligned}V_{n,k}^{2IBP,\theta}(\gamma) &= \frac{1}{k!}\left(\frac{\gamma}{(\theta+1)_{n-1\uparrow}}\right)^k \exp\left(-\sum_{i=1}^n \gamma\frac{(\theta)_{i-1\uparrow}}{(1+\theta)_{i-1\uparrow}}\right) \\ &= \frac{1}{k!}\left(\frac{\gamma}{(\theta+1)_{n-1\uparrow}}\right)^k \exp\left(-\gamma\sum_{i=1}^n \frac{\theta}{\theta+i-1}\right),\end{aligned}$$

with the convention that, when $\gamma = 0$, we recover the degenerate feature allocation with no features. Define $\mathcal{P}_{V^{2IBP,\theta}} = \{P_{V^{2IBP,\theta}(\gamma)} \in \mathcal{P}_{\mathcal{V}_{0,\theta}} : \gamma \geq 0\}$.

Proposition 6.4.2. *Let $\alpha = 0$ and $\theta > 0$.*

a) The elements of the set $\mathcal{P}_{V^{2IBP,\theta}}$ belong the maximal boundary of $\mathcal{P}_{\mathcal{V}_{0,\theta}}$ and they are induced by paths $w \in \mathbb{N}_0^\infty$ s.t. $\frac{w_m}{\log(m)} \to \gamma$;

b) The elements of $\mathcal{P}_{V^{2IBP,\theta}}$ also belong to the minimal boundary of $\mathcal{P}_{\mathcal{V}_{0,\theta}}$, i.e., they are extreme points of $\mathcal{P}_{\mathcal{V}_{0,\theta}}$;

c) The elements of $\mathcal{P}_{V^{2IBP,\theta}}$ are the only extreme points, i.e., $\mathcal{P}_{V^{3IBP,\theta}}$ coincides with the maximal and the minimal boundary.

Proof. a) We must check that

$$\lim_{\substack{m \to \infty \\ \frac{w_m}{\log(m)} \to \gamma}} \frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} = V_{n,k}^{2IBP,\theta}(\gamma). \tag{6.26}$$

From Proposition 6.3.1,

$$\begin{aligned}\frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} &= \frac{\binom{\omega_m}{k}\left[(\theta+m-1)_{m-n\downarrow}\right]^k\left[\sum_{i=1}^{m-n}(\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow}\right]^{l-k}}{\left[(\theta+m-1)_{m-1\downarrow}+\sum_{i=1}^{m-1}(\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow}\right]^{\omega_m}} \\ &= \binom{\omega_m}{k}\left[\frac{\frac{\Gamma(\theta+m)}{\Gamma(\theta+n)}}{\frac{\Gamma(\theta+m)}{\Gamma(\theta)}\sum_{i=1}^{m-n}\frac{1}{\theta+m-i}}\right]^k\left[\frac{\frac{\Gamma(\theta+m)}{\Gamma(\theta)}\sum_{i=1}^{m-n}\frac{1}{\theta+m-i}}{\frac{\Gamma(\theta+m)}{\Gamma(\theta)}\sum_{i=1}^{m-1}\frac{1}{\theta+m-i}+\frac{\Gamma(\theta+m)}{\Gamma(\theta+1)}}\right]^{\omega_m},\end{aligned}$$

where the second equality follows from the identity

$$\sum_{i=1}^{m-n}(\theta)_{m-i\uparrow}(\theta+m-1)_{i-1\downarrow} = (\theta)_{m\uparrow}\sum_{i=1}^{m-n}\frac{1}{\theta+m-i}.$$

Using the Stirling formula for the binomial coefficient, $\binom{\omega_m}{k} \approx \frac{1}{k!}\omega_m^k$, and the identity $\sum_{i=1}^{m}\frac{1}{\theta+m-i} = \sum_{i=1}^{m-n}\frac{1}{\theta+m-i} + \sum_{i=1}^{n}\frac{1}{\theta+n-i}$, we have

$$\begin{aligned}\frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} &\approx \frac{1}{k!}\left[\omega_m\frac{\frac{\Gamma(\theta)}{\Gamma(\theta+n)}}{\sum_{i=1}^{m-n}\frac{1}{\theta+m-i}}\right]^k\left[\frac{\sum_{i=1}^{m-n}\frac{1}{\theta+m-i}}{\sum_{i=1}^{m}\frac{1}{\theta+m-i}}\right]^{\omega_m} \\ &\approx \frac{1}{k!}\left[\frac{\omega_m}{\sum_{i=1}^{m-n}\frac{1}{\theta+m-i}}\frac{\Gamma(\theta)}{\Gamma(\theta+n)}\right]^k\left[\frac{\sum_{i=1}^{m-n}\frac{1}{\theta+m-i}}{\sum_{i=1}^{m-n}\frac{1}{\theta+m-i}+\sum_{i=1}^{n}\frac{1}{\theta+n-i}}\right]^{\omega_m}.\end{aligned}$$

Therefore, by Lemma 6.6.4 and the fact that $\log(m) \approx \sum_{i=1}^{m-n}\frac{1}{\theta+m-i}$ and $\omega_m/\log(m) \to \gamma$, we have

$$\begin{aligned}\frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} &\approx \frac{1}{k!}\left[\gamma\frac{\Gamma(\theta)}{\Gamma(\theta+n)}\right]^k\left[\frac{\log(m)}{\log(m)+\sum_{i=1}^{n}\frac{1}{\theta+n-i}}\right]^{\gamma\log(m)} \\ &\to \frac{1}{k!}\left[\gamma\frac{\Gamma(\theta)}{\Gamma(\theta+n)}\right]^k\exp\Big(-\gamma\sum_{i=1}^{n}\frac{1}{\theta+n-i}\Big),\end{aligned}$$

as $m \to \infty$, recovering the V of the 2-parameter IBP.

b) This also follows from Theorem 4 of Berti et al. [11], which establish that

$$P_{V^{2IBP,\theta}(\gamma)}\Big(\frac{K_n}{\log(n)} \to \gamma\Big) = 1.$$

c) To check that the only regular paths are those paths $\omega \in \mathbb{N}_0^{\mathbb{N}}$ such that $\frac{\omega_m}{\log(m)} \to \gamma$ for some $\gamma \geq 0$, we can repeat the same argument as in the proof of Proposition 6.4.1, part (c). First, we note that if $(\frac{w_m}{\log(m)})_{m\in\mathbb{N}}$ has at least two distinct subsequential limits, then ω cannot be regular, because the proof of point (a) shows that there will be two distinct induced laws, a contradiction. If $\frac{\omega_m}{\log(m)} \to \infty$ as $m \to \infty$, then it follows again from Lemma 6.6.5 that $\frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} \to 0$ as $m \to \infty$ for all $k \in \mathbb{N}_0$, a contradiction.

□

6.4.3 Case $\alpha < 0$

From formula (2.20), we see that the Beta–Bernoulli is of form (6.5), with V of the form

$$V_{n,k}^{BB,\alpha,\theta}(N) = \frac{\binom{N}{k}\big(\frac{-\alpha\Gamma(\theta+\alpha)}{\Gamma(\theta+\alpha+n)}\big)^k}{\big(\frac{\Gamma(\theta+\alpha)\Gamma(\theta+n)}{\Gamma(\theta+\alpha+n)\Gamma(\theta)}\big)^N}. \tag{6.27}$$

for all $N \in \mathbb{N}$. As before, when $N = 0$, we consider the feature allocation with no feature with probability one. Define $\mathcal{P}_{V^{BB,\alpha,\theta}} = \{P_{V^{BB,\alpha,\theta}(N)} \in \mathcal{P}_{\mathcal{V}_{\alpha,\theta}} : N \in \mathbb{N}_0\}$.

Proposition 6.4.3. *Let $\alpha < 0$ and $\theta > -\alpha$.*

a) The elements of the set $\mathcal{P}_{V^{BB,\alpha,\theta}}$ belong the maximal boundary of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ and they are induced by paths $w \in \mathbb{N}_0^\infty$ s.t. $w_m \to N$;

b) The elements of $\mathcal{P}_{V^{BB,\alpha,\theta}}$ also belong to the minimal boundary of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$, i.e., they are extreme points of $\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$;

c) The elements of $\mathcal{P}_{V^{BB,\alpha,\theta}}$ are the only extreme points, i.e., $\mathcal{P}_{V^{BB,\alpha,\theta}}$ coincides with the maximal and the minimal boundary.

Proof. a) We must check that

$$\lim_{\substack{m\to\infty \\ w_m\to N}} \frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} = V_{n,k}^{BB,\alpha,\theta}(N). \tag{6.28}$$

Starting with Proposition 6.3.1 and following similar steps as for the case $0 < \alpha < 1$, we obtain the approximation

$$\frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} \approx \binom{\omega_m}{k}\left(\frac{\frac{1}{\Gamma(\theta+n)}}{\frac{m^\alpha}{\alpha\Gamma(\alpha+\theta)} - \frac{\Gamma(\alpha+\theta+n)}{\alpha\Gamma(\alpha+\theta)\Gamma(\theta+n)}}\right)^k \left(\frac{m^\alpha - \frac{\Gamma(\alpha+\theta+n)}{\Gamma(\theta+n)}}{m^\alpha - \frac{\theta\Gamma(\alpha+\theta)}{\Gamma(\theta+1)}}\right)^{\omega_m}, \tag{6.29}$$

assuming $\omega_m \to N$ and $\alpha < 0$. Taking the limit as $m \to \infty$, we obtain

$$\frac{d_{n,k}^{m,\omega_m}}{d^{m,\omega_m}} \to \binom{N}{k} \left[\frac{-\alpha\Gamma(\alpha+\theta)}{\Gamma(\alpha+\theta+n)}\right]^k \left[\frac{\Gamma(\alpha+\theta+n)\,\Gamma(\theta)}{\Gamma(\alpha+\theta)\,\Gamma(\theta+n)}\right]^N .$$

b) This follows since under a Beta–Bernoulli model with N features, $K_m \to N$ a.s. Indeed, the probability of each feature, q_j, is a.s. strictly positive, being Beta distributed. The probability of this feature having all zeros in a m-individuals allocation is $(1-q_k)^m$, which tends to zero as $m \to \infty$.

c) By the a.s. monotonicity of regular paths, as $m \to \infty$, the number of features ω_m either diverges or converges to a finite (integer) limit. The divergent paths cannot be regular for $\alpha < 0$, because for these paths, (6.29) diverges as $m \to \infty$. Hence, the only regular paths are those of part (a). .

□

6.5 Discussion and future work

In this chapter, we have considered the class of consistent exchangeable feature allocations with EFPF of the form (6.5). While this is a tractable family, the only elements of this class are mixtures over γ of the 2 and 3-parameter IBP or mixtures over N of the Beta–Bernoulli model. From both an applied and theoretical perspective, it would be of interest to have larger but still tractable classes of exchangeable feature allocations. Finding new tractable priors for feature models is still an active area of research. Some extensions of the IBP have recently been proposed in Berti et al. [11] and James [68]. In particular, James [68] describes a very general framework to construct priors and to do Bayesian inference for (sparse) matrices. These matrices need not to be binary and indeed his theory work for matrices with general entries. Analytical computations are handled using an adaptation of the Poisson Partition Calculus to feature models. Another possible direction of research would be to study a more general class than (6.5), with form $V_{n,k}\prod_{l=1}^{k} W_{n,m_l}$, for a triangular array $W = (W_{n,k} : n \in \mathbb{N}, 0 \leq k \leq n)$. However, a characterization of W in this case would seem to be much more complicated than Proposition 6.2.1.

6.6 Appendix

In this appendix section, we collect proofs of some facts mentioned in this chapter. These proofs are generally quite simple and have been moved to this separate section to improve readability of the chapter.

6.6.1 Some Facts

Proposition 6.6.1. *$\mathcal{V}_{\alpha,\theta}$ is a convex set.*

Proof. We want to show that $\mathcal{V}_{\alpha,\theta}$ is a convex set, i.e., for all probability measures μ on $\mathcal{B}_{\mathcal{V}}$, $V^{\mu} \in \mathcal{V}_{\alpha,\theta}$. We have

$$\begin{aligned}
V^{\mu}_{n,k} &= \int_{\mathcal{V}_{\alpha,\theta}} V_{n,k}\mu(\mathrm{d}V) \\
&= \int_{\mathcal{V}_{\alpha,\theta}} \sum_{j=0}^{\infty} \binom{k+j}{j} ((\alpha+\theta)_{n\uparrow})^j (\theta+n)^k V_{n+1,k+j}\mu(\mathrm{d}V) \\
&= \sum_{j=0}^{\infty} \binom{k+j}{j} ((\alpha+\theta)_{n\uparrow})^j (\theta+n)^k \int_{\mathcal{V}_{\alpha,\theta}} V_{n+1,k+j}\mu(\mathrm{d}V) \\
&= \sum_{j=0}^{\infty} \binom{k+j}{j} ((\alpha+\theta)_{n\uparrow})^j (\theta+n)^k V^{\mu}_{n+1,k+j},
\end{aligned}$$

for all (n,k), where the first and last equality follow from the definition of barycenter, and the second from the monotone convergence theorem. In a similar manner,

$$\begin{aligned}
\sum_{j=0}^{\infty} V^{\mu}_{1,j} &= \sum_{j=0}^{\infty} \int_{\mathcal{V}_{\alpha,\theta}} V_{1,j}\mu(\mathrm{d}V) \\
&= \int_{\mathcal{V}_{\alpha,\theta}} \sum_{j=0}^{\infty} V_{1,j}\mu(\mathrm{d}V) \\
&= \int_{\mathcal{V}_{\alpha,\theta}} 1\mu(\mathrm{d}V) = 1.
\end{aligned}$$

□

Proposition 6.6.2. *$T(V) = P_V$ is an isomorphism between convex sets.*

Proof. According to Dynkin [33, pg 706], the map $T(V) = P_V$ is a convex isomorphism if T is invertible and T and T^{-1} are measurable and preserve the convex structure. T is 1-1 from Proposition 6.3.1 and it is onto by construction. We prove T is measurable and preserves the convex structure, proving that the same is true for T^{-1} can be done in similar way.

$\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}$ is endowed with the smallest σ-algebra $\mathcal{B}_{\mathcal{P}}$ that makes the maps $P_V \mapsto P_V(A)$ measurable for all $A \in \mathcal{C}(\mathbb{N}_0^{\infty})$. A generator of this σ-algebra is composed by sets $\{P_V \in \mathcal{P}_{\mathcal{V}_{\alpha,\theta}} : P_V(K_n = k) \leq x\}$ for $(n,k) \in \mathbb{N} \times \mathbb{N}_0$ and $x \in [0,1]$. The inverse image under T of this set is $\{V \in \mathcal{V}_{\alpha,\theta} : V_{n,k}d_{n,k} \leq x\}$ (see Proposition 6.3.1 for the definition of $d_{n,k}$), which lies in $\mathcal{B}_{\mathcal{V}}$. Hence, T is measurable.

T preserves the convex structure if, for every measure μ on $\mathcal{B}_{\mathcal{V}}$, we have $T(V^{\mu}) = P^{\mu'}$, where μ' the push-forward measure of μ on $\mathcal{B}_{\mathcal{P}}$ (i.e., $\mu' = \mu \circ T^{-1}$), and $P^{\mu'}$ is the barycenter of μ', defined as

$$P^{\mu'}(A) = \int_{\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}} P(A)\mu'(\mathrm{d}P), \tag{6.30}$$

for all $A \in \mathcal{C}(\mathbb{N}_0^{\infty})$. Using the change of variable formula, it is easy to check that T preserves the convex structure. Indeed, considering cylinder sets of the form $K_n^{-1}\{k\}$ for $n \in \mathbb{N}$ and $k \in \mathbb{N}_0$, we have

$$\begin{aligned} P^{\mu'}(K_n = k) &= \int_{\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}} P(K_n = k)\mu'(\mathrm{d}P) = \int_{\mathcal{P}_{\mathcal{V}_{\alpha,\theta}}} P(K_n = k)\mu \circ T^{-1}(\mathrm{d}P) \\ &= \int_{\mathcal{V}_{\alpha,\theta}} d_{n,k} V_{n,k} \mu(\mathrm{d}V) = d_{n,k} V_{n,k}^{\mu}. \end{aligned}$$

Hence, $T(V^{\mu}) = P^{\mu'}$. □

Proposition 6.6.3. $(\mathcal{F}_n, \Pi_n)_{n\in\mathbb{N}}$ *forms a specification in* $(\mathbb{N}_0^{\infty}, \mathcal{C}(\mathbb{N}_0^{\infty}))$.

Proof. According to Dynkin [33], section 5.1, given a directed set L and a measurable space $(\Lambda, \mathcal{F})$, a *specification* on this is space $(\mathcal{F}_\Lambda, \Pi_\Lambda)_{\Lambda\in L}$ is a family of sub-σ-algebras and Markov operators satisfying

(i) $\mathcal{F}_{\Lambda'} \subseteq \mathcal{F}_\Lambda$, if $\Lambda' \succeq \Lambda$;

(ii) $\Pi_{\Lambda'}\Pi_\Lambda = \Pi_{\Lambda'}$, if $\Lambda' \succeq \Lambda$;

(iii) $\Pi_\Lambda f \in \mathcal{F}_\Lambda$, for all $f \in \mathcal{F}$;

(iv) $\Pi_\Lambda f = f$, for all $f \in \mathcal{F}_\Lambda$.

In our context, with $L = \mathbb{N}$ and the sub-σ-algebras and Markov operators defined in section 3.1, formula (6.15), (i), (ii), and (iv) follow immediately. To check (ii), it is enough to check for indicators of measurable sets. In particular, for $f = \mathbb{1}_A$, with $A \in \mathcal{C}(\mathbb{N}_0^{\infty})$, we must check

$$\int_{\mathbb{N}_0^{\infty}} Q_n(\omega', A)Q_{n+1}(\omega, \mathrm{d}\omega') = Q_{n+1}(\omega, A).$$

Indeed, it is enough to check this condition for a thin cylinder A of the form $K_1^{-1}\{k_1\} \cap K_2^{-1}\{k_2\} \cap \ldots K_m^{-1}\{k_m\}$ for $m > n+1$ and $k_i \in \mathbb{N}_0$ for all $i \leq m$,

$$
\begin{aligned}
&\int_{\mathbb{N}_0^\infty} Q_n(\omega', A) Q_{n+1}(\omega, \mathrm{d}\omega') \\
&\quad = \int_{\mathbb{N}_0^\infty} P(K_1 = k_1, \dots, K_m = k_m | \mathcal{F}_n)(\omega') P((K_l)_{l\in\mathbb{N}} \in \mathrm{d}\omega' | \mathcal{F}_{n+1})(\omega) \\
&\quad = \int_{\mathbb{N}_0^\infty} P(K_1 = k_1, \dots, K_{n-1} = k_{n-1} | K_n = k_n) \\
&\qquad\qquad \cdot \mathbb{1}(\omega'_n = k_n) \dots \mathbb{1}(\omega'_m = k_m) P((K_l)_{l\in\mathbb{N}} \in \mathrm{d}\omega' | \mathcal{F}_{n+1})(\omega) \\
&\quad = P(K_1 = k_1, \dots, K_{n-1} = k_{n-1} | K_n = k_n) \\
&\qquad\qquad \cdot \int_{\mathbb{N}_0^\infty} \mathbb{1}(\omega'_n = k_n) \dots \mathbb{1}(\omega'_m = k_m) P((K_l)_{l\in\mathbb{N}} \in \mathrm{d}\omega' | \mathcal{F}_{n+1})(\omega) \\
&\quad = P(K_1 = k_1, \dots, K_{n-1} = k_{n-1} | K_n = k_n) P(K_n = k_n | \mathcal{F}_{n+1})(\omega) \\
&\qquad\qquad \cdot \mathbb{1}(\omega_{n+1} = k_{n+1}) \dots \mathbb{1}(\omega_m = k_m) \\
&\quad = P(K_1 = k_1, \dots, K_m = k_m | \mathcal{F}_{n+1})(\omega) = Q_{n+1}(\omega, A).
\end{aligned}
$$

□

6.6.2 Technical Lemmas

We state here two technical results about asymptotic equivalence of functions which are used in the proofs of 6.4.1 and 6.4.2. Write $f \approx g$ to denote that $f(m)/g(m) \to 1$ as $m \to \infty$, and note that $f_i \approx g_i$ implies $f_1 + f_2 \approx f_1 + g_2 \approx g_1 + g_2$ and $f_1 f_2 \approx f_1 g_2 \approx g_1 g_2$. In general, it does not hold that $f_1^{f_2} \approx g_1^{g_2}$. The following two results characterize special cases:

Lemma 6.6.4. *Let $g(m) \to \infty$ as $m \to \infty$, let $f \approx g$ and $h/g \to c$ for some constant $c \geq 0$. Then, for every $p, q \in \mathbb{R}$, we have*

$$
\left(\frac{f(m) - p}{f(m) - q}\right)^{h(m)} \approx \left(\frac{g(m) - p}{g(m) - q}\right)^{c\, g(m)} \to e^{c\,(q-p)} \tag{6.31}
$$

Proof. We prove only the first equivalence, because the limiting exponential form is well known. Taking logarithms, we have

$$
\log \left[\left(\frac{f(m) - p}{f(m) - q}\right)^{h(m)} \left(\frac{g(m) - q}{g(m) - p}\right)^{c\, g(m)} \right] \tag{6.32}
$$

$$
= c\, g(m)\, \log \frac{(f(m) - p)\,(g(m) - q)}{(f(m) - q)\,(g(m) - p)} \tag{6.33}
$$

$$
+ \left(h(m) - c\, g(m)\right) \log \frac{f(m) - p}{f(m) - q}. \tag{6.34}
$$

The arguments to the logarithms can we written as

$$\frac{(f(m)-p)\,(g(m)-q)}{(f(m)-q)\,(g(m)-p)} = 1 + \frac{(f(m)-g(m))\,(p-q)}{(f(m)-q)(g(m)-p)} \tag{6.35}$$

and

$$\frac{f(m)-p}{f(m)-q} = 1 + \frac{q-p}{f(m)-q}. \tag{6.36}$$

Using the fact that $z(m) \to 0$ implies $\log(1+z(m)) \approx z(m)$, and that both terms (6.35) and (6.36) converge to one, it follows that

$$(6.32) \approx c\,g(m)\frac{(f(m)-g(m))\,(p-q)}{(f(m)-q)(g(m)-p)} + (h(m)-c\,g(m))\,\frac{q-p}{f(m)-q}. \tag{6.37}$$

It is straightforward to show that the r.h.s. converges to 0. □

Lemma 6.6.5. *Let $f(m) \to \infty$ as $m \to \infty$, let $g(m)/f(m) \to \infty$ as $m \to \infty$, and let $h \approx f$. For every $p > q$,*

$$\left(\frac{g(m)}{f(m)}\right)^k \left(\frac{h(m)-p}{h(m)-q}\right)^{g(m)} \to 0 \qquad \text{as } m \to \infty.$$

Proof. Taking logarithms

$$\begin{aligned} & k \log \frac{g(m)}{f(m)} + g(m) \log\Big\{1 + \frac{q-p}{h(m)-q}\Big\} \\ & \approx k \log \frac{g(m)}{f(m)} + (q-p)\frac{g(m)}{h(m)-q} \\ & \approx k \log \frac{g(m)}{f(m)} + (q-p)\frac{g(m)}{f(m)} \to -\infty \end{aligned}$$

as $m \to \infty$, completing the proof. □

Bibliography

[1] Adams, M. et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

[2] Agrawal, S. and Goyal, N. (2012). Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*.

[3] Aldous, D. (1985). Exchangeability and related topics. Ecole d'Eté de Probabilités de Saint-Flour, XIII 1983, 1-198.

[4] Arbel, J., Favaro, S., Nipoti, B. and Teh, Y.W. (2015). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Preprint arXiv:1506.04915*.

[5] Auer, P., Cesa-Bianchi, N. and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, **47**(2), 235-256.

[6] Bacallado, S., Favaro, S. and Trippa, L. (2013). Bayesian nonparametric analysis of reversible Markov chains. *Annals of Statistics*, **41**, 870–896.

[7] Barger, K. and Bunge, J. (2010). Objective Bayesian estimation of the number of species. *Bayesian Analysis*, **5**, 619–639.

[8] Basu, D. (1969). Role of the Sufficiency and Likelihood Principles in Sample Survey Theory. *Sankhya, Series A*, **31**, 441-454.

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

[10] Blackwell, D. and MacQueen, J.B. (1973). Ferguson Distributions Via Polya Urn Schemes. *Annals of Statistics*, **1**, 353–355.

[11] Berti, P., Crimaldi, I., Pratelli, L. and Rigo, P. (2015). Central limit theorems for an Indian buffet model with random weights. *Annals of Applied Probability*, **25**(2), 523–547.

[12] Broderick, T., Pitman, J. and Jordan, M.I. (2013). Feature Allocations, Probability Functions, and Paintboxes. *Bayesian Analysis*, **8**(4), 801–836.

[13] Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, **5**(1), 1-122.

[14] Bubeck, S., Cesa-Bianchi, N. and Lugosi, G. (2013). Bandits with heavy tail. *IEEE Transactions on Information Theory*, **59**, 7711-7717.

[15] Bubeck, S., Ernst, D. and Garivier, A. (2013). Optimal discovery with probabilistic expert advice: finite time analysis and macroscopic optimality. *Journal of Machine Learning Research*, **14**, 601-623.

[16] Bunge, J., Willis, A. and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statisitcs and its Applications*, **1**, 427–445.

[17] Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. In *Advances in Neural Information Processing Systems*.

[18] Caron, F. and Fox, E.B. (2015). Sparse graphs with exchangeable random measures. *Preprint ArXiv:1401.1137*.

[19] Caron, F., Teh, Y.W. and Murphy, T.B. (2014). Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes. *Annals of Applied Statistics*, **8**, 1145–1181.

[20] Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson sampling. In *Neural Information Processing Systems*.

[21] Charalambides, C. A. (2005). *Combinatorial methods in discrete distributions*. Wiley, New York.

[22] Chen, C., Ding, N. and Buntine, W (2012). Dependent hierarchical normalized random measures for dynamic topic modeling. In *International Conference in Machine Learning*.

[23] Chen, C., Rao, V.A., Buntine, W. and Teh, Y.W. (2013). Dependent normalized random measures. In *International Conference in Machine Learning*.

[24] Chow, Y.S., Robbins, H., and Siegmund, D. (1971). *Great expectations: the theory of optimal stopping*. Houghton Mifflin, Boston.

[25] Cochran, W.G.(1977). *Sampling Techniques*. 3rd ed. Wiley, New York.

[26] Condit, R. et al. (2002). Beta-Diversity in Tropical Forest Trees. *Science*, **295**, 666-669.

[27] De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 212–229.

[28] De Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della Reale Academia Nazionale dei Lincei, Serie 6*, **4**, 251-299.

[29] Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation*, **4**, 1-18.

[30] Diaconis, P. and Freedman, D. (1980). De Finetti's Theorem for Markov Chains. *Annals of Probability*, **8**, 115-130.

[31] Diaconis, P. and Freedman, D. (1982). Partial exchangeability and sufficiency. In *Statistics Applications and New Directions; Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Sankhya A*, J. K. Ghosh and J. Roy, Eds., pg 205-236. Indian Statistical Institute.

[32] Doksum, K. (1974). Tailfree and Neutral Random Probabilities and Their Posterior Distributions. *Annals of Probability*, **2**, 183-201.

[33] Dynkin, E.B. (1978). Sufficient Statistics and Extreme Points. *Annals of Probability*, **6**(5), 705–730.

[34] Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**, 435–447.

[35] Duan, J.A., Guindani, M. and Gelfand, A.E. (2007). Generalized spatial Dirichlet process models. *Biometrika*, **77**, 1-11.

[36] Favaro, S., Lijoi, A., Mena, R.H. and Prünster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society, Series B*, **71**, 993–1008.

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

[38] Favaro, S., Lijoi, A. and Prünster, I. (2012b). Asymptotics for a Bayesian nonparametric estimator of species variety. *Bernoulli*, **18**(4), 1267-1283.

[39] Favaro, S and Walker, S.G. (2013). Slice sampling σ-stable Poisson-Kingman mixture models. *Journal of Computational and Graphical Statistics*, **22**, 830–847.

[40] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

[41] Fonteneau-Belmudes, F., Ernst, D., Druet, P., Panciatici, P. and Wehenkel, L. (2010). Consequence driven decomposition of large-scale power system security analysis. In *Proceedings of the 2010 IREP Symposium*.

[42] Fortini, S. and Petrone, S. (2012). Predictive construction of priors in Bayesian nonparametrics. *Brazilian Journal of Probability and Statistics*, **26**, 423-449.

[43] Fortini, S. and Petrone, S. (2016). Predictive Characterization of Mixtures of Markov Chains. *Bernoulli, to appear*.

[44] Gelfand, A.E., Kottas, A. and MacEachern S.N. (2005). Bayesian nonparametric spatial modelling with Dirichlet processes mixing. *Journal of the American Statistical Association*, **100**, 1021–1035.

[45] Gelfand, A.E., Guindani, M. and Petrone, S. (2007). Bayesian nonparametric modeling for spatial data analysis using Dirichlet processes. In *Bayesian Statistics* **8**, (eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West). Oxford University Press, Oxford.

[46] Ghahramani, Z., Griffiths, T.L. and Sollich, P. (2007). Bayesian nonparametric latent feature models. In *Bayesian Statistics 8*. Oxford University Press, Oxford.

[47] Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. Springer, New York.

[48] Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Science*, **138**, 5674–5685.

[49] Goldwater, S., Griffiths, T., and Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems*, **18**. MIT Press, Cambridge, MA.

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

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

[52] Good, I.J. (1965). *The estimation of probabilities: an essay on modern Bayesian methods.* MIT press, Cambridge, MA.

[53] Granmo, O.C. (2010). Solving two-armed bernoulli bandit problems using a Bayesian learning automaton. *International Journal of Intelligent Computing and Cybernetics*, **3**(2), 207-234.

[54] Griffin, J.E. and Steel, M.F.J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, **101**, 179–194.

[55] Griffiths, T.L. and Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. *In Advances in Neural Information Processing Systems.*

[56] Griffiths, T.L. and Ghahramani, Z. (2011). The Indian Buffet Process: An Introduction and Review. *Journal of Machine Learning Research*, **12**, 1185–1224.

[57] Herlau, T. (2015). Completely random measures for modelling block-structured sparse networks. *Preprint arXiv:1507.02925.*

[58] Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, **80**, 470-501.

[59] Hill, B.M. (1979). Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *Journal of the American Statistical Association*, **74**, 668–673.

[60] Hjort, N.L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics*, **18**, 1259-1294.

[61] Hjort, N.L., Holmes, C., Müller, P. and Walker, S.G. (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge.

[62] Hoppe, F.H. (1984). Pólya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology*, **20**, 91–94.

[63] Ionita-Laza, I., Lange, C. and Laird, N.M. (2009). Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, **106**, 5008-5013.

[64] Ionita-Laza, I. and Laird, N.M. (2010). On the pptimal design of genetic variant discovery studies. *Statistical Applications in Genetics and Molecular Biology*, **9**, article no. 33.

[65] Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick- breaking priors. *Journal of the American Statistical Association*, **96**, 161–173.

[66] James, L.F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. *Preprint arXiv:math/0205093*.

[67] James, L.F. (2013). Stick-breaking PG(α, ζ)-generalized Gamma processes. *Preprint arXiv:1308.6570*.

[68] James, L. (2016). Bayesian Poisson Calculus for Latent Feature Modeling via Generalized Indian Buffet Process Priors. *Annals of Statistics*, to appear.

[69] Johnson, W.E. (1932). Probability: the deductive and inductive problems. *Mind*, **41**, 409–423.

[70] Kaufmann, E., Korda, N. and Munos, R. (2012). Thompson Sampling: An Optimal Finite-Time Analysis. *The International Conference on Algorithmic Learnig Theory*.

[71] Kingman, J.F.C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society, Series B*, **37**(1), 1–22.

[72] Kingman, J.F.C. (1978). The representation of partiton structure. *Journal of the London Mathematical Society*, **18**, 374–380.

[73] Korwar, R.M. and Hollander, M. (1973). Contribution to the theory of Dirichlet processes. *Annals of Probability*, **1**, 705–711.

[74] Lai, T.L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, **6**, 4–22.

[75] Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modelling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, **100**, 1278–1291.

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

[77] Lijoi, A., Mena, R.H. and Prünster, I. (2007b). A Bayesian nonparametric method for prediction in EST analysis. *BMC Bioinformatics*, **8**, article no. 339.

[78] Lijoi, A., Mena, R.H. and Prünster, I. (2007c). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society, Series B*, **69**, 769–786.

[79] Lijoi, A., Mena, R.H. and Prünster, I. (2008). A Bayesian nonparametric approach for comparing clustering structures in EST libraries. *Journal of Computational Biology*, **15**, 1315–1327.

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

[81] Lijoi, A., Prünster, I. and Walker, S.G. (2008). Investigating nonparametric priors with Gibbs structure. *Statistica Sinica*, **18**, 1653–1668.

[82] Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. *Annals of Statistics*, **12**, 351–357.

[83] Lo, A.Y. (1991). A characterization of the Dirichlet process. *Statistics and Probability Letters*, **12**, 185–187.

[84] Lomeli, M., Favaro, S and Teh, Y.W. (2015). A marginal sampler for σ-stable Poisson-Kingman mixture models. *Journal of Computational and Graphical Statistics*, to appear.

[85] MacEachern, S.N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*. American Statistical Association, Alexandria.

[86] May, B.C. and Leslie, D.S. (2011). Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:02, Statistics Group, Department of Mathematics, University of Bristol.

[87] Mao, C.X. (2004). Prediction of the conditional probability of discovering a new class. *Journal of the American Statistical Association*, **99**, 1108–1118.

[88] Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, **1**(2), 226-251.

[89] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields*. **92**, 21–39.

[90] Petrone, S., Guindani, M. and Gelfand, A.E. (2009). Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society, series B*, **71**, 755–782.

[91] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, **102**, 145–158.

[92] Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In *Statistics, Probability and Game Theory*, Ferguson, T.S., Shapley, L.S. and MacQueen, J.B. Eds., Institute of Mathematical Statistics.

[93] Pitman, J. and Yor, M. (1997). The two parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, **25**, 855–900.

[94] Pitman, J. (2003). Poisson-Kingman partitions. In *Science and Statistics: A Festschrift for Terry Speed*, Goldstein, D.R. Eds. Institute of Mathematical Statistics.

[95] Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté de Probabilités de Saint-Flour XXXII. Lecture notes in mathematics, Springer, New York.

[96] Phadia, E.G. (2013). *Prior Processes and their Applications*. Springer, New York.

[97] Pyke, C.R., Condit, R., Aguilar, S. and Lao, S. (2001). Floristic composition across a climatic gradient in a neotropical lowland forest. *Journal of Vegetation Science*, **12**, 553–566.

[98] Rasmussen, S.L. and Starr, N. (1979). Optimal and adaptive stopping in the search for new species. *Journal of the American Statistical Association*, **74**, 661–667.

[99] Ravel, J. et al. (2011). Vaginal microbiome of reproductive-age women. *Proceeding of the National Academy of Sciences*, **108**, 4680–4687.

[100] Regazzini, E. (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilià. *Giornale dell'Istituto Italiano degli Attuari*, **41**, 77–89.

[101] Reich, B. and Fuentes, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Annals of Applied Statistics*, **1**, 249–264.

[102] Rolles, S. (2003). How edge-reinforced random walk arises naturally. *Probability Theory and Related Fields*, **126**, 243–260.

[103] Russo, D. and Van Roy, B. (2014). Learning to Optimize Via Posterior Sampling. *Mathematics of Operations Research*, **39**, 1221–1243.

[104] Russo, D. and Van Roy, B. (2016). An Information-Theoretic Analysis of Thompson Sampling. *Journal of Machine Learning Research*, to appear.

[105] Scott, S. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, **26**, 639–658.

[106] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.

[107] Steel, M.F.J. and Fuentes, M. (2010). Non-Gaussian and Nonparametric Models for Continuous Spatial Data. In *Handbook of Spatial Statistics* (eds. A.E. Gelfand, P.J. Diggle, M. Fuentes and P.Guttorp), pp. 149-167. CRC Press.

[108] Susko, E. and Roger, A.J. (2004). Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys. *Bioinformatics*, **20**, 2279–2287.

[109] Szabo, B. and Tran-Thanh, L. (2015). Finite-Time Concentration Inequalities for the Posteriori and Regret Analysis of Bayesian Online Learning Algorithms. Working Paper.

[110] 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 44th Annual Meeting of the Association for Computational Linguistics*, 985-92. Association for Computational Linguistics, Morristown, NJ.

[111] Teh, Y.W. and Gorür, D. (2009). Indian buffet processes with power-law behavior. In *Advances in Neural Information Processing Systems*. Vancouver, Canada.

[112] Teh, Y.W., Jordan, M.I., Beal, M.J. and Blei, D.M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

[113] Teh, Y.W. and Jordan, M.I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*, Hjort, N.L., Holmes, C.C. Müller, P. and Walker, S.G. Eds., Cambridge University Press, Cambridge.

[114] Thibaux, R. and Jordan, M.I. (2007). Hierarchical Beta processes and the Indian buffet process. In *Eleventh International Conference on Artificial Intelligence and Statistics*.

[115] Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, **85**, 1050–1059.

[116] Thompson, S.K. (1991). Stratified Adaptive Cluster Sampling. *Biometrika*, **78**, 389–397.

[117] Thompson, S.K. (1991). Adaptive Cluster Sampling: Designs with Primary and Secondary Units. *Biometrics*, **47**(3), 1103–1115.

[118] Thompson, S.K. (1992). An Adaptive Procedure for Sampling Animal Populations. *Biometrics*, **48**(4), 1195–1199.

[119] Thompson, S.K. (2002). *Sampling*. 2nd ed. Wiley, New York.

[120] Thompson, S.K. and Seber, G.A.F. (1996). *Adaptive Sampling*. Wiley, New York.

[121] Thompson, W.R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**(3-4), 285–294.

[122] Walker, S.G. and Muliere, P. (1997). A characterisation of Polya tree distributions. *Statistics and Probability Letters*, **31**, 163–168.

[123] Walker, S.G. and Muliere, P. (1999). A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate. *Annals of Statistics*, **27**, 589–599.

[124] Whittaker, R.H. (1960). Vegetation of the Siskiyou Mountains. *Ecological Monographs*, **30**, 279–338.

[125] Zabell, S.L. (1982). W. E. Johnson's "sufficientness" postulate. *Annals of Statistics*, **10**, 1090–1099.

[126] Zabell, S.L. (1985). Characterizing Markov exchangeable sequences. *Journal of Theoretical Probability*, **8**, 175–178.

[127] Zabell, S.L. (1992). Predicting the unpredictable. *Synthese*, **90**, 205–232.

[128] Zabell, S.L. (1997). The continuum of inductive methods revisited. In *The cosmos of science: essays in exploration*, Earman, J. and Norton, J.D. Eds. Universty of Pittsburgh Press, PA.

[129] Zabell, S.L. (2005). The continuum of inductive methods revisited. In *Symmetry and its discontents: essays on the history of inductive probability*. Cambridge Univ. Press, New York.

[130] Zacks, S. (1969). Bayes Sequential Designs of Fixed Size Samples From Finite Populations. *Journal of the American Statistical Association*, **64**, 1342–1349.

[131] Zhang, H. and Stern, H. (2009). Sample size calculation for finding unseen species. *Bayesian Analysis*, **4**, 763–792.